

MultiViewDx: Evidence-Linked Multi-View Clinical Diagnosis

Junda Wang^{1,2} Zonghai Yao² Yujan Ting¹ Eric Z. Chen¹
Hieu Tran^{1,2} Hong Yu^{2,3} Weijing Huang¹ Terrence Chen¹

¹United Imaging Intelligence, MA, USA

²Manning College of Information and Computer Sciences, UMass Amherst, MA, USA

³Department of Medicine, University of Massachusetts Medical School, Worcester, MA, USA

Abstract

Medical multimodal large language models (MLLMs) can perform well on existing medical visual question answering (MedVQA) benchmarks, but their training data often does not match clinical diagnosis. Most supervision is organized around isolated images or short QA pairs, leaving two structures weakly specified: how evidence leads to a decision, and how views, series, modalities, and patient context from the same case are linked. We introduce MultiViewDx, a partly physician-validated multimodal instruction dataset for evidence-linked multi-view medical imaging diagnosis. MultiViewDx uses the clinical case as the supervision unit. It links imaging studies with patient context, normalizes heterogeneous reports into an evidence-linked workflow (evidence → findings → differential discussion → diagnosis), and uses a unified image-text retriever to constrain instruction synthesis to source-supported evidence. It covers X-ray, CT, MRI, ultrasound, histopathology, and other clinical visual sources. We fine-tune MultiViewDx-8B-AN and evaluate it on both existing MedVQA benchmarks and real-world case-based diagnostic reasoning. Across four MedVQA benchmarks, it achieves the best average accuracy among compared systems (79.0%), outperforming HuatuoGPT-Vision-34B (66.7%) and Claude3-Opus (55.7%). Beyond MedVQA, on JAMA Clinical Challenge cases, it receives the strongest overall rating under a physician-designed rubric for key clinical points, diagnostic inference, and evidence grounding. Controlled ablations and clinician evaluation show that both case-level multi-view organization and evidence-linked reasoning targets contribute to the gain.¹

¹Code and datasets are available at <https://github.com/believablewhat/MultiViewDx>.

1 Introduction

Medical imaging diagnosis in clinical practice is rarely a single-image, single-answer task. A clinician usually reviews a study as a set of related views, slices, or sequences, and interprets them with patient history, laboratory results, vital signs, medications, and other clinical signals. The one or two images shown in a case report, teaching example, or patient-facing explanation are selected for communication; they are not necessarily the full evidence used in care. When such material is converted into one-image QA examples, the task can become visual recognition or answer matching rather than case-level diagnosis.

Medical multimodal large language models (MLLMs) have improved medical image QA through biomedical image-text corpora and instruction tuning (Yan et al., 2023; Liu et al., 2024a; Jin et al., 2024; Yang et al., 2025; Li et al., 2024; Chen et al., 2024b; Yao and Yu, 2025). Recent systems such as LLaVA-Med, HuatuoGPT-Vision, MedTrinity, and Hulu-Med adapt general MLLMs to medical data (Li et al., 2024; Chen et al., 2024b; Xie et al., 2024; Jiang et al., 2025). However, much of the supervision still uses isolated image-caption pairs or short QA instances. We study *evidence-linked multi-view diagnosis*, where the clinical case is the training and evaluation unit. The model must identify findings across available views, slices, sequences, or modalities, connect them to patient context, compare differentials, and state a diagnosis or next step with explicit evidence support.

Current datasets often leave this structure underspecified in two ways. **Workflow-sparse supervision** omits the diagnostic path from evidence to findings, differential discussion, and final decision; image captions and short QA pairs usually state what is present, but not how it supports a clinical conclusion (Bustos et al., 2020; Irvin et al., 2019; Johnson et al., 2019; Ikezogwo et al., 2024).

From Isolated Image QA to Evidence-Linked Clinical Diagnosis

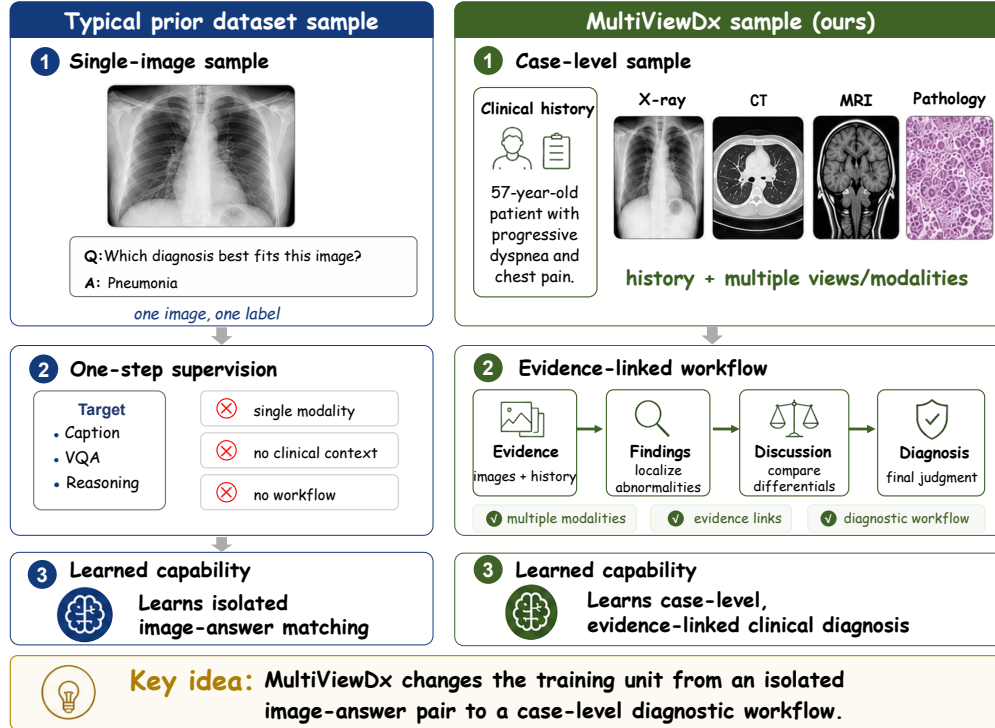


Figure 1: **From image-level QA to case-level diagnosis.** Left: common medical VQA or caption data use one displayed image and a short target answer. Right: MultiViewDx uses the clinical case as the training unit, linking patient context, multiple views or series, localized evidence, intermediate findings, differential discussion, and the final diagnosis or next step. The key change is the supervision unit: from an isolated image-answer pair to an evidence-linked diagnostic case.

Case-fragmented supervision treats views or series from the same study as independent samples, or drops them entirely. This simplification is understandable, because public reports and case presentations often keep only a few illustrative images. But it makes the data easier than clinical reasoning, since the original diagnostic process used a richer case-level evidence set (Xie et al., 2024; Pal et al., 2023; Dorfner et al., 2024).

This mismatch is hard to see if evaluation stops at existing MedVQA benchmarks. Benchmarks such as VQA-RAD, SLAKE, PathVQA, and PMC-VQA remain useful for measuring medical QA performance, but many items can be answered from a single image or a short visual-text association. They do not consistently require cross-view corroboration, evidence attribution, or integration of patient context. We therefore report both standard MedVQA performance and diagnosis-oriented evaluation on JAMA Clinical Challenge cases, using rubric-based scoring and clinician review.

To address these limitations, we introduce **Multi-ViewDx**, a data-centric instruction corpus constructed from real, de-identified clinical data with physician validation over sampled synthesized reasoning chains. MultiViewDx changes the supervision unit from an isolated image-answer pair to an evidence-linked clinical case. It links imaging studies to clinical context, preserves available view and series structure, and normalizes heterogeneous reports into a workflow of evidence, findings, differential discussion, and diagnosis or next step. A unified image-text retriever gathers source-supported image and context evidence before instruction synthesis. Each instance is designed to teach both the answer and the provenance of the key evidence.

We fine-tune MultiViewDx-8B-AN and test whether this data design changes model behavior. On VQA-RAD and SLAKE, MultiViewDx-8B-AN achieves the best average accuracy among compared models, showing strong performance on existing MedVQA benchmarks. On JAMA case-based diagnostic reasoning, it obtains the strongest

overall rating under a physician-designed rubric for key clinical points, diagnostic inference, and evidence grounding. The ablations tie these gains to the proposed design: removing case-level multi-view organization substantially reduces clinician-rated quality, while removing explicit reasoning targets further weakens clinician judgments. Physician validation of synthesized reasoning chains and clinician evaluation of model outputs provide checks beyond automatic answer matching.

Contributions.

- MultiViewDx, an evidence-linked multi-view instruction dataset that uses the clinical case as the supervision unit and aligns imaging views, patient context, and staged diagnostic fields
- a retrieval-constrained construction pipeline that preserves evidence provenance across views, regions, and clinical signals during instruction synthesis
- an evaluation and ablation protocol that combines existing MedVQA benchmarks with real-world case-based diagnostic reasoning, rubric scoring, and clinician evaluation to distinguish benchmark QA performance from diagnostic reasoning.

2 Related Work

Existing Multimodal Medical Datasets Medical multimodal datasets have evolved from report-paired corpora such as MIMIC-CXR-JPG (Johnson et al., 2019) to large image-caption collections like PMC-OA (Lin et al., 2023), case/VQA resources (PMC-CaseReport, PMC-VQA) (Wu et al., 2023; Zhang et al., 2023), instruction-style sets from PubMed imagery (LLaVA-Med VQA, PubMedVision) (Li et al., 2024; Chen et al., 2024b), structured annotation efforts (RadGenome-Chest CT) (Zhang et al., 2024), and multigranular pipelines such as MedTrinity (Xie et al., 2024), Hulu-Med Data (Jiang et al., 2025), Lingshu Data (Team et al., 2025) and HealthGPT Data (Lin et al., 2025). Despite this progress, much of the available supervision remains image-level: a displayed image is paired with a caption, a short answer, or a generated instruction. This creates workflow-sparse supervision because intermediate diagnostic fields and evidence provenance are not consistently represented. It also creates case-fragmented supervision because views, slices, se-

quences, and non-imaging signals from the same clinical study are often separated into independent examples. Recent resources that include multi-view inputs or richer clinical context, such as MedThinkVQA (Yao et al., 2025) and MedXpertQA (Zuo et al., 2025), are valuable for evaluation but remain too small to train large medical MLLMs at scale. MultiViewDx differs by organizing training data around evidence-linked clinical cases: it aligns multi-view studies with patient context, normalizes reports into staged diagnostic fields, and constrains instruction synthesis with retrieved source evidence.

Medical Multimodal Models Several efforts fine-tune general-purpose multimodal models on medical data and show strong gains. Med-Flamingo (Moor et al., 2023) adapts OpenFlamingo-9B (Chen et al., 2024a) with a medical corpus, and Med-PaLM (Tu et al., 2024) extends PaLM-E (Driess et al., 2023) using large-scale medical image-text pairs. Similarly, LLaVA-Med, Med-Gemini (Saab et al., 2024), HuatuoGPT-Vision (Chen et al., 2024b), HealthGPT (Lin et al., 2025), Lingshu (Team et al., 2025) and Hulu-Med (Jiang et al., 2025) leverage instruction tuning over curated PubMed-derived image-text data for medical QA. Our focus is complementary to model-architecture work. We do not propose a new fusion module; instead, we study how changing the supervision unit from isolated images to evidence-linked clinical cases affects both held-out MedVQA performance and case-based diagnostic reasoning.

3 MultiViewDx

MultiViewDx is built around a simple design choice: the clinical case, rather than the isolated image, is the basic training unit. For each case, we preserve two relations when they are available. The first is *vertical provenance*: evidence is linked to localized findings, differential discussion, and the final diagnosis or next step. The second is *horizontal study integrity*: related views, slices, sequences, modalities, and non-imaging clinical signals remain grouped under the same case. This design directly targets workflow-sparse supervision and case-fragmented supervision.

3.1 Data Collection

Data sources and study preservation. We collect data from sources that provide complemen-

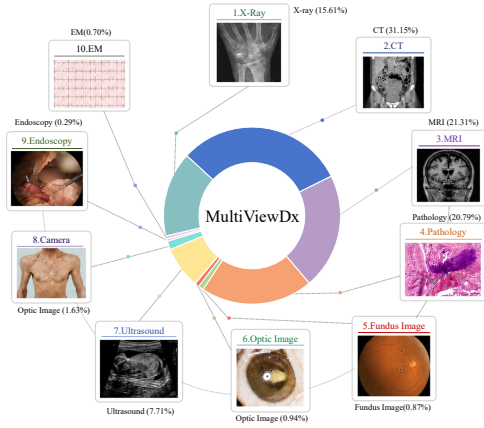


Figure 2: **Modality coverage of MultiViewDx.** MultiViewDx covers a broad range of medical imaging modalities and clinical visual sources, including X-ray, CT, MRI, histopathology, ultrasound, endoscopy, clinical photography, fundus/optic imaging, and electrophysiology. The central distribution summarizes the major modality composition of the dataset, with CT, MRI, histopathology, and X-ray forming the largest proportions. This diverse modality coverage supports training on multi-view and multi-modal clinical evidence rather than a single image type.

tary forms of medical supervision, including report-paired radiology data, pathology VQA, public biomedical image-text corpora, and expert-authored clinical case repositories. When a source contains study structure, such as AP/LAT radiographs, CT slices, MRI sequences, or ROI hints, we preserve this structure under a shared case identifier instead of converting each image into an unrelated example. Fig. 2 summarizes the resulting modality coverage, with CT, MRI, histopathology, and X-ray forming the largest portions, and additional coverage from ultrasound, endoscopy, clinical photography, fundus or optic imaging, and electrophysiology.

Image and slice selection. For 3D CT and MRI studies, we represent each study with a compact set of informative 2D slices. We retain clinically marked slices when available and uniformly sample additional slices along the study axis, with at most 20 slices per case. This cap is intended to preserve anatomical coverage and cross-slice continuity while reducing redundancy from densely adjacent slices under limited visual-context budgets. The all-slice comparison in Sec. 5.3 tests whether simply exposing more slices is sufficient.

Data Source The training corpus draws from DeepLesion, MIMIC-CXR-JPG, PadChest, Quilt, LLD-MMRI, and MAMA-MIA, along with benchmark training QA datasets such as VQA-RAD, PathVQA, PMC VQA, and SLAKE, covering multiple modalities like CT, MRI, X-ray and so on. Additionally, we expanded the dataset with data from Eurorad and Radiopaedia to include more diverse modalities as shown in table 1. To support bilingual instruction following, we uniformly sample 300k instances from the combined training corpus and translate their textual fields into Chinese. The translated examples preserve the source image, modality, annotation metadata, and provenance information when available.

Dataset	N	Mod	ROI	HA	Slice	Cap	Lic (short)
DeepLesion	24,821	CT	×	×	×	Y	CC-BY4
PadChest	150,730	XR	×	✓	—	Y	AGR-PadChest
Eurorad	691,370	Multi	✓	✓	✓	Y	CC-BY4
MIMIC-CXR-JPG	620,113	XR	×	✓	—	N	PN-CHDL1.5
LLD	30,390	MR	✓	×	✓	Y	AGR-LLD
MAMA-MIA	76,381	MR	✓	×	✓	Y	CC-BY-NC-SA4
PMC-VQA*	152,603	Multi	×	✓	—	Y	CC-BY-SA
Path-VQA	19,654	Path	×	✓	—	Y	CC-BY4
PMC-Instruct	619,606	Multi	×	✓	—	Y	OpenRAIL
Quilt	1,017,416	Histo	×	✓	—	Y	—
Radiopaedia	1,131,614	Multi	✓	✓	✓	N	AGR-RP
SLAKE	9,835	Multi	×	✓	—	Y	CC-BY4
VQA-RAD	1,798	Multi	×	✓	—	Y	CC-BY4
AMBOSS & JAMA	45,820	Mix	✓	✓	—	N	AGR-JAMA

Table 1: Unified data sources with annotation, caption availability, and license. N denotes processed sample records. For fragmented sources, the retained metadata support sample-level counts only; unique patient-, study-, and case-level counts are unavailable. Mod: CT=CT, XR=X-ray, MR=MRI, Path=pathology, Histo=histopathology, Multi=multi-modality, Mix=multi + text. ROI: ROI annotation. HA: human annotation. Slice: per-slice identifier. Cap: caption availability (Y/N). Lic (short): license type (CC-BY4, CC-BY-SA, CC-BY-NC4, CC-BY-NC-SA4, PN-CHDL1.5, OpenRAIL, AGR-* = dataset-specific agreement). “—” means not applicable or not provided. *PMC-VQA and the Stage 1 corpus share PMC/PMC-OA source provenance and may contain overlapping articles, images, or captions.

Human Annotated Workflow To supervise the end-to-end diagnostic workflow (**Workflow-sparse supervision**), we prioritize expert-labeled sources (e.g., Eurorad, Radiopaedia) that mirror real workflows (history → findings → differential → impression/plan). Because raw annotations vary in length and style, we standardize them into consistent workflow fields and align sentences to views/ROIs. For lengthy reports (e.g., Eurorad), we decompose them into (i) per-image findings, (ii) study-level synthesis, and (iii) discussion/decision text, regen-

erating only format (not content) to preserve expert intent. Only explicit figure, panel, view, slice, or ROI references create image-level links; all other statements remain at study level. Specific-image text maps to per-image findings, cross-image integration maps to study-level synthesis, and diagnosis, differential, impression, management, or next-step text maps to discussion/decision content. The result is study-level supervision that simultaneously (i) teaches stepwise reasoning and decisions (addresses Workflow-sparse supervision) and (ii) ties claims to specific views/slices and clinical context (supports Case-fragmented supervision). A concrete human-annotated case example is provided in Appendix C.1.

3.2 Data Construction Pipeline

Case preservation and normalization. For each source case, we first assign a case identifier and collect all available images, view or slice metadata, report text, ROI hints, and clinical signals. We normalize modality names, laterality, anatomy terms, and report sections into a shared schema. Near-duplicate cases are removed by matching source identifiers and by checking high-overlap image-text pairs. The normalized fields are clinical context, visual evidence, localized image findings, differential discussion, diagnosis or impression, next step, and evidence citations.

Retrieval-constrained synthesis. For incomplete cases, we use a unified image-text retriever to retrieve source-supported context before synthesis. The retriever indexes clinical guidelines and image-report cases, and returns a small set of relevant items, typically $k = 4$, using fused image-text similarity. GPT-4o then drafts captions, QA pairs, or diagnostic-workflow responses conditioned on the retrieved context, the source report fields, available ROI hints, and the study structure. The prompt requires every clinical claim to be supported by either a source finding, a retrieved case, a guideline statement, an ROI, a view or slice reference, or a non-imaging clinical signal. Unsupported details are not allowed to be introduced.

Unified instruction schema. The final instructions cover three formats. Caption-style instructions ask for concise descriptions of visible findings, including anatomy, abnormality, laterality, and view or slice information when available. VQA-style instructions pair an image or case context with a question and answer, optionally with

a short evidence-based explanation. Diagnosis-workflow instructions use case-level inputs, including multi-view images and clinical context, and produce evidence, image findings, differential discussion, diagnosis, and next step when available. Each instance stores machine-readable provenance fields, including study id, modality, series or view, ROI references, clinical signals, findings, differentials, diagnosis, next step, and evidence citations.

Negative and normal controls. To reduce shortcut learning, we add normal controls, absent-lesion distractors, and near-miss distractors. Normal controls require the model to justify the absence of acute findings. Absent-lesion distractors are plausible diseases from guidelines or similar cases but are not supported by the current case. Near-miss distractors refer to adjacent regions, alternate views, or similar-looking findings. For each negative item, the target response must explain why the distractor is not supported, such as an incorrect region, missing view evidence, or contradictory clinical signal.

Physician validation of synthesized reasoning chains. Because part of MultiViewDx is constructed through retrieval-augmented synthesis, we validate generated reasoning chains with physician review. We randomly sample 100 synthesized cases and ask three physicians with more than 10 years of clinical experience to independently review the original source case, associated imaging evidence, clinical context, and generated reasoning chain. Each case is labeled as fully consistent, inconsistent but clinically correct, or incorrect or misleading. A fully consistent case has reasoning and diagnosis supported by the source case and its evidence. An inconsistent but clinically correct case reaches an acceptable final conclusion but has incomplete or imprecise intermediate reasoning or evidence attribution. An incorrect or misleading case contains unsupported claims, contradicts source evidence, or could lead to an incorrect diagnostic interpretation. Under this protocol, 88% of sampled cases are fully consistent, 10% are inconsistent but clinically correct, and 2% are incorrect or misleading. We report Fleiss' $\kappa = 0.81$ for the three-way nominal labels.

Instruction formats. MultiViewDx contains three instruction formats. First, caption-style instructions use one image, one view, or a small set of views as input, and the target output is a concise

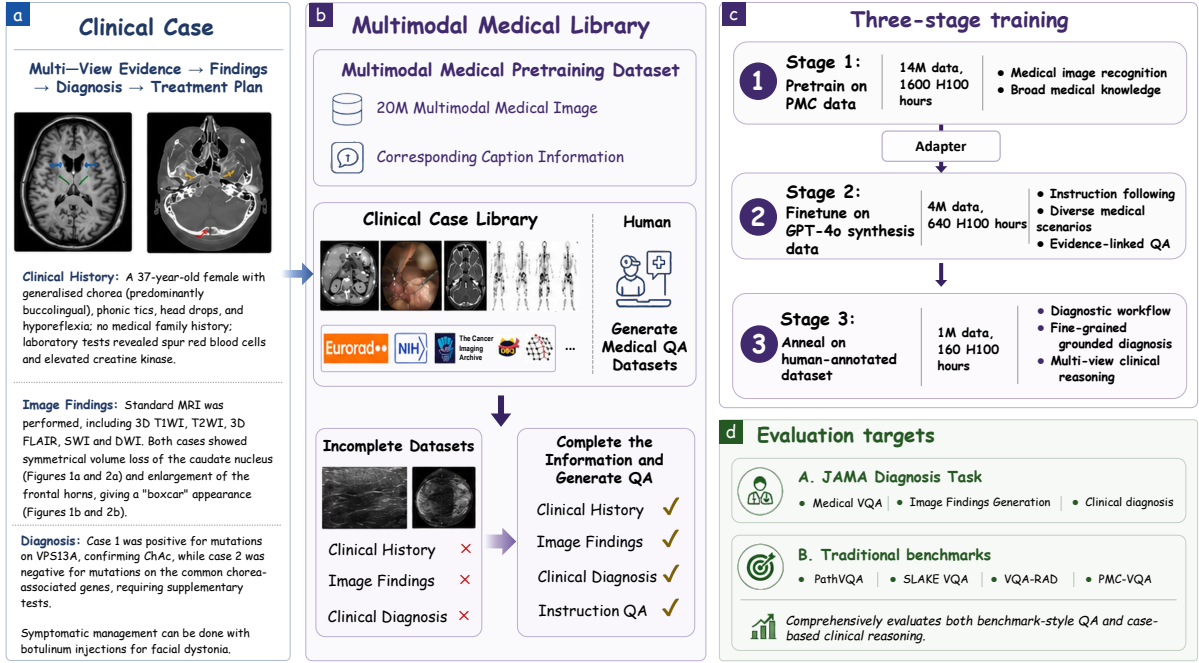


Figure 3: MultiViewDx data construction and training pipeline. The pipeline starts from clinical cases that contain multi-view evidence, image findings, diagnosis, and treatment-related information. These cases are organized into a multimodal medical library that combines large-scale image-caption pretraining data with a clinical case library. For incomplete sources, retrieval and GPT-4o synthesis are used to complete missing clinical history, image findings, diagnosis, and instruction QA. The model is then trained in three stages: pretraining on PMC image-text data, fine-tuning on GPT-4o-synthesized instruction data, and sequential fine-tuning on human-annotated clinical cases. The resulting model is evaluated on both real-world medical tasks and traditional medical VQA benchmarks.

description of visible findings, including anatomy, abnormality, laterality, and view or slice information when available. Second, VQA-style instructions use an image or case context together with a natural-language question as input, and the target output is the answer to the question, optionally with a short evidence-based explanation when the source annotation supports it. Third, diagnosis-workflow instructions use case-level inputs, including multi-view images and available clinical context, and the target output follows our evidence-linked workflow: evidence, image findings, discussion or differential diagnosis, final diagnosis, and when available, the next diagnostic or treatment-related step. These formats allow the same dataset to supervise image understanding, medical QA, and case-level diagnostic reasoning.

Human validation of synthesized reasoning chains. Because part of MultiViewDx is constructed through retrieval-augmented synthesis, we conduct a physician validation study to assess the reliability of the generated reasoning chains. We

randomly sample 100 cases from the synthesized portion of MultiViewDx and ask three physicians, each with more than 10 years of clinical experience, to independently review them. For each case, annotators are shown the original source case, the associated imaging evidence and clinical context, and the generated evidence-linked reasoning chain. We use three outcome categories. A case is marked as *fully consistent* if the generated reasoning and diagnosis are supported by the source case and its evidence. It is marked as *inconsistent but clinically correct* if the final diagnostic conclusion is clinically acceptable but some intermediate reasoning step or evidence attribution is incomplete or imprecise. It is marked as *incorrect or misleading* if the output contains unsupported clinical claims, contradicts the source evidence, or could lead to an incorrect diagnostic interpretation. Under this protocol, 88% of sampled cases are judged fully consistent, 10% are judged inconsistent but clinically correct, and 2% are identified as incorrect or misleading. We measure inter-annotator agreement using Fleiss’ κ , treating the three validation labels as nominal cate-

gories, and obtain $\kappa = 0.81$. These results suggest that the synthesized reasoning chains are generally reliable, while also motivating our use of expert-authored cases in the final sequential fine-tuning stage.

4 Training and Evaluation Pipeline

This section defines how MultiViewDx is used for training and evaluation. We first describe the three training stages and the controlled variants used in ablations. We then define the benchmark splits, baseline groups, and diagnostic-reasoning metrics. All results are reported in Sec. 5.

4.1 Training Pipeline

We train MultiViewDx-8B in three stages. Stage 1 performs medical image-text pretraining on 14M filtered PMC samples after removing corrupted items, non-medical examples, and captions with fewer than 20 words. Only the adapter is updated, using learning rate $1e-4$, image token length 512, DeepSpeed Stage 2, batch size 8, gradient accumulation 6, and 4 H100 GPUs for 420 hours. Stage 2 performs instruction tuning on GPT-4o-synthesized MultiViewDx instructions, updating all parameters with learning rate $2e-5$, DeepSpeed Stage 2, batch size 6, gradient accumulation 6, and 8 H100 GPUs for 80 hours. Stage 3 sequentially fine-tunes the model on human-annotated clinical cases with learning rate $1e-5$. The full pipeline uses approximately 2,400 H100 GPU-hours and an estimated $0.81-1.0 \times 10^{21}$ FLOPs.

Multi-image representation. MultiViewDx-8B is initialized from an LLaVA-8B model. Each image is independently encoded by the shared vision encoder and represented as a block of 512 visual tokens. For a multi-image case, visual blocks are serialized in case order with available modality, view, or slice identifiers in the accompanying text; standard one-dimensional sequence positions continue across the serialized visual and textual input. CT and MRI use the selected 2D slices described in Sec. 3.1, ordered by slice position when available.

Training variants. We use six variants to isolate the data and training components.

- **MultiViewDx-8B-20M** uses all available slices from 3D studies during training, and tests whether more visual context alone improves performance.

- **MultiViewDx-8B-Synthesis** uses Stage 1 PMC image-text pretraining followed by GPT-4o-synthesized instruction tuning with the slice-capped case construction and stops before the human-annotated Stage 3.
- **MultiViewDx-8B-Mix** combines GPT-4o-synthesized instructions and human-annotated clinical cases in a single SFT stage.
- **MultiViewDx-8B-AN** first trains on synthesized instructions and then sequentially fine-tunes on human-annotated clinical cases.
- **MultiViewDx-Non-Multi-View** keeps diagnostic-reasoning targets and human-annotated cases, but breaks case-level multi-view grouping into single-view or independently sampled examples.
- **MultiViewDx-NoReasoning** keeps case-level inputs and human-annotated data, but reduces target outputs to direct answers or diagnoses without the evidence-linked diagnostic trajectory.

The last two variants test case-fragmented supervision and workflow-sparse supervision. The Non-Multi-View ablation jointly changes case organization and the associated training distribution. Appendix B.8 illustrates the evidence-linking contrast through paired same-case outputs and reports an inference-time view-removal test.

4.2 Evaluation Pipeline

Training and evaluation splits. For VQA-RAD, SLAKE, PathVQA, and PMC-VQA, we follow the official dataset partitions. Only official training splits are used during fine-tuning, and all reported results use the corresponding official held-out test splits. For the JAMA Clinical Challenge, we construct a disjoint case-level split because the dataset does not provide the same predefined train/test split as the VQA benchmarks: 60% of cases are used for training and 40% are reserved for testing. No held-out MedVQA test instance or held-out JAMA case is used during training. MultiViewDx-8B-AN uses all four official benchmark training partitions. PMC-VQA* shares PMC/PMC-OA source provenance with the Stage 1 corpus.

Evaluation axes. We evaluate two axes. The first is target-supervised performance on official

held-out MedVQA test splits, measured by accuracy on VQA-RAD (Lau et al., 2018), SLAKE (Liu et al., 2021), PathVQA (He et al., 2020), and PMC-VQA (Zhang et al., 2023). The second is case-based diagnostic reasoning, measured on held-out JAMA Clinical Challenge cases that require integration of case context and imaging evidence.

For existing MedVQA benchmarks, we report answer accuracy on the official held-out test sets. For JAMA, we report close-ended accuracy when applicable, UMLS-F1 factuality, GPT-4o rubric scores, and clinician ratings. UMLS-F1 measures medical concept overlap using SciSpacy/UMLS, as described in Appendix A.1. The GPT-4o rubric is a blinded 0–2 scoring protocol designed with clinicians. It contains three axes: **Key Points**, which measures coverage of critical clinical elements; **Inference**, which measures the correctness and completeness of the diagnostic reasoning path; and **Evidence**, which measures whether claims are grounded in specific findings, imaging views or ROIs, and clinical signals. Overall is the mean of the three rubric axes. To reduce stylistic bias, outputs are style-normalized before judging, and the judge sees extracted gold summaries of Key Points, Inference, and Evidence rather than the full reference text.

We also conduct a separate clinician evaluation on 100 randomly sampled held-out JAMA cases. Three clinicians review each case and its gold rationale and assign one holistic rating to the model rationale while considering Key Points, Inference, and Evidence. The averaged and normalized rating is **Doctor Score**, our primary human evaluation of model rationales. GPT-4o and DeepSeek provide complementary automatic scores; all three evaluations rank MultiViewDx-AN first.

Knowledge and inference item classification.

We estimate the proportion of knowledge-oriented and inference-oriented items using a fixed GPT-4o judging prompt. For each benchmark item, the input to the judge includes the question, answer choices when available, and the reference answer. The judge assigns one of two labels. A question is labeled knowledge if it mainly requires recognition of a visual pattern, recall of a medical fact, or a short image-text association. A question is labeled inference if it requires combining multiple findings, using clinical context, comparing possible diagnoses, reasoning across views or modalities, or making a multi-step clinical judgment. We apply

the same prompt to all datasets to characterize the proportion of knowledge-oriented and inference-oriented items, as shown in Appendix Fig. 4. The full prompt is provided in Appendix B.5. These labels are used only for aggregate characterization of benchmark style, not as gold diagnostic labels for model evaluation.

5 Results

We organize results around four questions. First, does MultiViewDx retain strong performance on existing MedVQA benchmarks? Second, does it improve case-based diagnostic reasoning beyond MedVQA accuracy? Third, which components of the data design and training schedule matter? Fourth, do clinician judgments and qualitative cases support the automatic results? Additional baseline details, training configurations, and prompt templates are provided in Appendix B.1.

5.1 Performance on Existing MedVQA Benchmarks

Table 2 compares MultiViewDx-8B-AN with proprietary models, general-purpose multimodal VLMs, and medical multimodal VLMs on VQA-RAD, SLAKE, PathVQA, and PMC-VQA*. MultiViewDx-8B-AN achieves the best average accuracy among the compared systems, reaching 79.0%. It outperforms the strongest medical baseline Hulu-Med-7B (74.3%) and HuatuoGPT-Vision-34B (66.7%), and also exceeds strong general-purpose models such as Gemini-2.5-Flash (63.8%), GPT-4.1 (62.0%), and Claude3-Opus (55.7%). MultiViewDx-8B-AN is trained on all four official benchmark training partitions and evaluated on the corresponding held-out test sets. Appendix Tables 12–13 report model-by-model training exposure.

Fine-tuning on benchmark training sets. We additionally report a separate fine-tuning-on-training-set comparison in Table 3. In this setting, each model is fine-tuned on the corresponding benchmark training split before evaluation on the held-out test split. Every row uses benchmark-specific supervised adaptation; Appendix Table 13 records the corresponding training exposure.

5.2 Case-Based Diagnostic Reasoning Beyond MedVQA

Existing MedVQA benchmarks test important medical image QA skills, but they do not fully test

Model	VQA-RAD	SLAKE	PathVQA	PMC-VQA*	Avg.
Proprietary Models					
GPT-4.1	65.0	72.2	55.5	55.2	62.0
GPT-4o	61.0	71.2	55.5	49.7	59.3
Claude Sonnet 4	67.6	70.6	54.2	54.4	61.7
Gemini-2.5-Flash	68.5	75.8	55.4	55.4	63.8
GPT-4o-mini	45.9	59.0	37.9	33.3	44.0
Claude3-Opus	52.5	55.2	54.3	60.7	55.7
General-purpose Multimodal VLMs					
Qwen2.5VL-7B	63.2	66.8	44.1	51.9	56.5
Janus-Pro-7B	49.7	55.2	35.4	50.1	47.6
InternVL2.5-8B	59.4	69.0	42.1	51.3	55.5
InternVL3-8B	65.4	72.8	48.6	53.8	60.1
Llama3.2-11B	58.8	65.8	32.9	48.1	51.4
InternVL3-14B	66.3	72.8	48.0	54.1	60.3
Qwen2.5VL-32B	71.8	71.2	41.9	54.5	59.9
InternVL2.5-38B	61.4	70.3	46.9	57.2	59.0
InternVL3-38B	65.4	72.7	51.0	56.6	61.4
Qwen-VL-Chat	47.0	56.0	55.1	36.6	48.9
Yi-VL-34B	53.0	58.9	47.3	39.5	49.7
LLaVA-7B	52.6	57.9	47.9	35.5	48.5
LLaVA-13B	55.8	58.9	51.9	36.6	50.8
LLaVA-34B	58.6	67.3	59.1	44.4	57.4
LLaVA-8B	54.2	59.4	54.1	36.4	51.0
Medical Multimodal VLMs					
BiomedGPT	16.6	13.6	11.3	27.6	17.3
Med-R1-2B	39.0	54.5	15.3	47.4	39.0
MedVLM-R1-2B	49.2	56.3	36.0	48.8	47.6
HealthGPT-M3	56.8	70.8	55.4	55.4	59.6
BioMedix2-8B	55.7	54.1	34.6	41.8	46.5
MedGemma-4B-IT	72.3	78.2	48.1	49.2	62.0
HuatuoGPT-V-7B	67.6	68.1	44.8	53.1	58.4
Lingshu-7B	67.9	83.1	61.9	56.3	67.3
Med-Flamingo	45.4	43.5	54.7	23.3	41.7
RadFM	50.6	34.6	38.7	25.9	37.5
LLaVA-Med-7B	51.4	48.6	56.8	24.7	45.4
+ PubMedVision	63.8	74.5	59.9	52.7	62.7
HuatuoGPT-Vision-34B	68.1	76.9	63.5	58.2	66.7
Hulu-Med-7B	78.0	86.8	65.6	66.8	74.3
Our Model					
MultiViewDx-8B-AN	86.1	87.7	80.4	61.9	79.0

Table 2: Main comparison on official held-out medical VQA test splits. MultiViewDx-8B-AN uses the four official training partitions. Additional MultiViewDx variants appear in Table 5, and Appendix Table 12 reports exposure for every compared model. *PMC-VQA and Stage 1 share PMC/PMC-OA source provenance.

whether a model can integrate imaging evidence with patient context and justify a diagnostic decision. We therefore evaluate on held-out JAMA Clinical Challenge cases, which provide real-world clinical vignettes with case context and associated visual evidence. Table 4 reports close-ended accuracy, UMLS-F1 factuality, GPT-4o rubric scores for Key Points, Inference, and Evidence when available.

MultiViewDx-AN achieves the best Accuracy (58.5%), the highest UMLS Factuality (0.23), and the best GPT-4 Overall score (1.29 ± 0.02). It also obtains the highest Key-Points score (1.28), strong Inference (1.32), and high Evidence (1.27). Among baselines, Claude3-Opus achieves the strongest Inference score (1.56) but weaker Evidence (0.67), while Huatuo-34B provides the strongest Evidence score among baselines (1.31) but lower Inference (1.06). Hulu-Med-7B and Lingshu-7B obtain relatively strong overall rubric scores (1.22 ± 0.03 and 1.18 ± 0.04), suggesting that medical-domain train-

Model	VQA-RAD	SLAKE	PathVQA	PMC-VQA*	Avg.
LLaVA-v1.5-LLAMA3-8B	63.3	68.9	85.2	50.3	66.9
LLaVA_Med-8B	66.3	69.5	90.7	52.7	69.8
HuatuoGPT-Vision-8B	68.9	84.1	93.0	57.3	75.8
MultiViewDx	88.3	91.1	92.7	88.6	90.2

Table 3: Benchmark-specific supervised adaptation results. Every row is fine-tuned on the corresponding official training partition and evaluated on the held-out test split. Appendix Table 13 summarizes the setting. *PMC-VQA shares the source provenance noted in Table 2.

Model	Acc.	UMLS	KP	Inf.	Evi.	Overall
C3-Opus	58.4	0.18	1.27	1.56	0.67	1.17 ± 0.04
G4o-mini	46.2	0.16	0.99	1.13	0.60	0.91 ± 0.06
H-7B	34.5	0.13	1.11	1.06	1.08	1.08 ± 0.03
H-34B	44.7	0.16	1.01	1.06	1.31	1.13 ± 0.05
MG-4B	45.8	0.19	1.06	1.10	1.21	1.14 ± 0.03
LS-7B	53.6	0.19	1.10	1.19	1.27	1.18 ± 0.04
HM-7B	54.6	0.18	1.15	1.22	1.26	1.22 ± 0.03
MVDX-AN	58.5	0.23	1.28	1.32	1.27	1.29 ± 0.02

Table 4: Results on the JAMA Clinical Challenge. Acc.=close-ended accuracy; UMLS=UMLS-F1 factuality; KP=Key Points; Inf.=Inference; Evi.=Evidence. Model abbreviations: C3-Opus=Claude3-Opus, G4o-mini=GPT-4o-mini, H-7B/H-34B=Huatuo-7B/34B, MG-4B=Med-Gemma-4B-IT, LS-7B=Lingshu-7B, HM-7B=Hulu-Med-7B, and MVDX-AN=MultiViewDx-AN. We additionally rescore a subset with DeepSeek in Appendix Table 7; the relative ranking is consistent.

ing improves evidence-oriented outputs, but still does not fully match MultiViewDx-AN on overall case-based diagnostic reasoning.

Fig. 4 further shows that JAMA contains a larger share of inference-oriented items than existing MedVQA benchmarks. We use it as a complementary diagnosis-oriented evaluation, not as a replacement for MedVQA accuracy.

Why fixing Case-fragmented supervision helps and why traditional metrics can mislead. Traditional VQA sets are dominated by *knowledge* items (SLAKE 78.1%, VQA-RAD 76.4%, PathVQA 69.2%), so models can score well by recalling single-view facts or template associations; JAMA contains far fewer knowledge items (44.9%) and instead requires *inference* from *multiple* views/modalities (Fig. 4). This mismatch explains why surface accuracy on VQA can overstate clinical readiness: those tests rarely require cross-view corroboration, ROI grounding, or contradiction checks. Case-level supervision supports triangulation across views, disambiguation through view-dependent cues, evidence attribution to specific views or ROIs, and

Variant	VQA-RAD	SLAKE	PathVQA	PMC-VQA*	Avg.
MultiViewDx-8B-20M	67.8	76.1	57.8	53.6	63.8
MultiViewDx-8B-Synthesis	69.2	77.2	63.6	58.4	67.1
MultiViewDx-8B-Mix	74.2	81.3	76.3	59.1	72.2
MultiViewDx-8B-AN	86.1	87.7	80.4	61.9	79.0

Table 5: Ablation of MultiViewDx training variants on standard medical VQA benchmarks. MultiViewDx-8B-20M uses all available 3D slices during training; Synthesis uses GPT-4o-synthesized instruction data with slice-capped construction; Mix combines synthesized and human-annotated cases in one SFT stage; AN first trains on synthesized data and then sequentially fine-tunes on human-annotated cases. *PMC-VQA shares the source provenance noted in Table 2.

study-level coherence. All values reported here match Table 4; Appendix Table 7 provides the DeepSeek sensitivity analysis. On 100 multi-view cases, removing the view containing the relevant image evidence while retaining the vignette and remaining inputs reduced accuracy to 31%. The result shows that view-specific evidence is essential during inference.

5.3 Ablations: Workflow, Case Organization, and Training Schedule

We use the variants defined in Sec. 4.1 to test which parts of MultiViewDx matter. The ablations are designed to answer four questions. First, does a compact slice-capped case representation work better than feeding all available 3D slices? Second, is sequential fine-tuning on human-annotated clinical cases more effective than mixing synthetic and human cases in one SFT stage? Third, does case-level multi-view organization matter after controlling for reasoning supervision and human cases? Fourth, does explicit evidence-linked reasoning supervision matter after controlling for multi-view inputs and human cases?

Training schedule ablation. Table 5 shows the effect of slice selection and training schedule on standard medical VQA benchmarks. MultiViewDx-8B-20M, which uses all available slices from 3D studies, obtains 63.8 average accuracy. Using slice-capped GPT-4o-synthesized instruction data improves the average score to 67.1, suggesting that densely adjacent slices do not necessarily help when they introduce redundant visual context. Mixing synthesized data with human-annotated clinical cases further improves the score to 72.2. The best result is obtained by MultiViewDx-8B-AN, which first trains on synthesized instruction data and then

Model	Syn.	MV	Reason.	HA	Doc.
C3-Opus	-	-	-	-	0.48
H-7B	✓	×	×	×	0.21
H-34B	✓	×	×	×	0.35
MG-4B	✓	×	×	×	0.40
LS-7B	✓	×	✓	×	0.44
HM-7B	✓	×	✓	×	0.50
MVDX-Syn	✓	×	✓	×	0.30
MVDX-NoMV	✓	×	✓	✓	0.29
MVDX-NoR	✓	✓	×	✓	0.11
MVDX-AN	✓	✓	✓	✓	0.53

Table 6: Component-level clinician evaluation on 100 held-out JAMA cases. Doctor Score (Doc.) is the averaged and normalized holistic rating from three clinicians, who consider Key Points, Inference, and Evidence with reference to the case and gold rationale. Syn.=synthetic or teacher-distilled instruction data; MV=case-level multi-view grouping; Reason.=evidence-linked diagnostic reasoning targets; HA=human-annotated clinical cases. Baseline checkmarks denote analogous training signals reported in the original papers.

sequentially fine-tunes on human-annotated cases, reaching 79.0 average accuracy. This suggests that human-annotated cases are most effective when used as a later expert-case alignment stage rather than simply mixed into one SFT stage.

Component-level clinician evaluation. Table 6 complements the automatic VQA ablation with clinician evaluation on JAMA. MultiViewDx-Synthesis uses synthetic instruction data and diagnosis-reasoning supervision but does not include case-level multi-view grouping or human-annotated case fine-tuning, obtaining a Doctor Score of 0.30. MultiViewDx-Non-Multi-View adds human-annotated cases but removes case-level multi-view organization, yielding a Doctor Score of 0.29. MultiViewDx-NoReasoning keeps multi-view inputs and human-annotated cases but removes explicit diagnosis-reasoning supervision, leading to the lowest Doctor Score of 0.11. MultiViewDx-AN includes all components and obtains the highest Doctor Score of 0.53. The NoMV-AN gap shows the value of case-level organization under the associated training distribution. Appendix Table 10 illustrates the evidence-linking contrast on the same case, and Appendix Table 11 reports the 100-case view-removal result.

5.4 Human Evaluation and Qualitative Analysis

The clinician evaluation in Table 6 assesses whether generated rationales are clinically acceptable beyond final-answer matching. MultiViewDx-AN obtains the highest Doctor Score (0.53), above HULU-Med-7B (0.50) and Claude3-Opus (0.48). The margin is modest, so we interpret it together with the ablations and qualitative cases rather than treating the score alone as sufficient evidence.

The qualitative JAMA case in Appendix Table 14 illustrates the same pattern. Appendix Table 10 additionally contrasts illustrative Full and Non-Multi-View outputs on the same case. The case requires linking bilateral optic nerve edema and macular edema with pregnancy, hypertension, and chronic kidney disease, then prioritizing blood pressure assessment to rule out malignant hypertension. MultiViewDx connects the visual findings with the patient context and explains why vital-sign assessment is the immediate next step. Claude3-Opus selects the correct action but gives a less evidence-specific rationale, while HuatuoGPT-Vision variants focus on diabetes-related testing and miss the urgent prioritization. The example suggests that MultiViewDx’s advantage lies in evidence grounding and clinical prioritization, not only in final answer selection.

5.5 Interpretation

The results support the two data-design claims in different ways. MedVQA gains show that workflow-aware, report-anchored supervision improves medical image QA performance on held-out test sets. The JAMA results and ablations provide the stronger test of diagnostic reasoning: removing case-level multi-view grouping weakens performance even when reasoning supervision and human cases remain, and removing evidence-linked reasoning targets causes the largest drop in clinician-rated quality. Thus, the main effect is not explained by parameter scale or by adding more images alone; it comes from organizing supervision around evidence-linked clinical cases.

6 Limitations and Ethical Considerations

MultiViewDx has several limitations. First, modality and anatomy coverage remain uneven because high-quality human-validated clinical data are scarce. Performance may therefore vary across body regions, acquisition protocols, institutions,

and patient populations that are underrepresented in the corpus. Second, our representation of 3D CT and MRI studies uses selected 2D slices rather than full volumetric reasoning. The slice-capped strategy is practical and empirically useful in our experiments, but it does not replace architectures designed for full 3D fusion. Third, part of the corpus is retrieval-synthesized. Although physician review finds a low rate of incorrect or misleading reasoning in the sampled cases, this does not eliminate all unsupported statements or subtle evidence-attribution errors. Fourth, clinician evaluation is conducted on a finite sample of held-out JAMA cases and should be expanded across specialties and institutions. Current validation includes a random, unstratified physician audit of 100 synthesized cases. Future audits will stratify by source, modality, diagnosis, and synthesis workflow. PMC-VQA and Stage 1 share PMC/PMC-OA provenance, and the retained PMC manifests provide sample-level counts rather than unique patient-, study-, or case-level counts. Future validation will quantify retrieval relevance and atomic claim entailment, clinically review the 300,000-instance Chinese subset, and calibrate automatic rubric scores against clinician ratings on matched outputs.

The dataset and model are intended for research on medical multimodal reasoning, not for autonomous diagnosis. Any clinical use would require prospective validation, institution-specific calibration, uncertainty reporting, bias and subgroup analysis, and human oversight. We also avoid using identifiable patient information and preserve only de-identified clinical context, but de-identification and data governance should be re-audited before any redistribution or deployment.

7 Conclusion

We identify two data limitations that make medical MLLMs appear stronger on MedVQA than on diagnosis-oriented cases: workflow-sparse supervision and case-fragmented supervision. MultiViewDx addresses these limitations by using evidence-linked clinical cases as the training unit, aligning multi-view imaging evidence with patient context and staged diagnostic fields. Fine-tuning on MultiViewDx improves performance on existing MedVQA held-out test sets and yields stronger performance on held-out JAMA case-based diagnostic reasoning. Controlled ablations and clinician evaluation further show that the gains depend

on both case-level multi-view organization and explicit evidence-linked reasoning targets.

Acknowledgments

We thank the anonymous reviewers and Action Editor Mihai Surdeanu for their constructive feedback.

References

- 01.AI, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yanpeng Li, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open foundation models by 01. AI. *arXiv preprint arXiv:2403.04652*.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.
- Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. 2020. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797.
- Delong Chen, Jianfeng Liu, Wenliang Dai, and Baoyuan Wang. 2024a. Visual instruction tuning with polite flamingo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17745–17753.
- Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, Guangjun Yu, Xiang Wan, and Benyou Wang. 2024b. HuatuoGPT-Vision, towards injecting medical visual knowledge into multimodal llms at scale. *arXiv preprint arXiv:2406.19280*.
- Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*.
- Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.
- Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhao Wang. 2024c. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.
- Felix J. Dornier, Amin Dada, Felix Busch, Marcus R. Makowski, Tianyu Han, Daniel Truhn, Jens Kleesiek, Madhumita Sushil, Jacqueline Lammert, Lisa C. Adams, and Keno K. Bressen. 2024. Biomedical large languages models seem not to be superior to generalist models on unseen medical data. *arXiv preprint arXiv:2408.13833*.
- Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor

- Mordatch, and Pete Florence. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.
- Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressemer. 2023. Medalpaca—an open-source collection of medical conversational AI models and training data. *arXiv preprint arXiv:2304.08247*.
- 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*.
- Wisdom Ikezogwo, Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. 2024. Quilt-1m: One million image-text pairs for histopathology. *Advances in neural information processing systems*, 36.
- Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silvana Ciurea-Ilcus, Christopher Chute, Henrik Marklund, Behzad Haghighi, Robyn L. Ball, Katie S. Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597.
- Songtao Jiang, Yuan Wang, Sibao Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, Jin Hao, Zijian Chen, Ruijia Wu, Tao Tang, Junhui Lv, Hongxia Xu, Hongwei Wang, Jun Xiao, Bin Feng, Fudong Zhu, Kenli Li, Weidi Xie, Jimeng Sun, Jian Wu, and Zuozhu Liu. 2025. Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. *arXiv preprint arXiv:2510.08668*.
- Qiao Jin, Fangyuan Chen, Yiliang Zhou, Ziyang Xu, Justin M. Cheung, Robert Chen, Ronald M. Summers, Justin F. Rousseau, Peiyun Ni, Marc J Landsman, Sally L. Baxter, Subhi J. Al’Aref, Yijia Li, Alex Chen, Josef A. Brejt, Michael F. Chiang, Yifan Peng, and Zhiyong Lu. 2024. Hidden flaws behind expert-level accuracy of GPT-4 vision in medicine. *arXiv preprint arXiv:2401.08396*.
- Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. 2019. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.
- Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, Yuheng Li, Konstantinos Psounis, and Xiaofeng Yang. 2025. Med-R1: Reinforcement learning for generalizable medical reasoning in vision-language models. *arXiv preprint arXiv:2503.13939*.
- 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.
- Chunyu 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.
- Chunyu Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2024. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.
- Tianwei Lin, Wenqiao Zhang, Sijing Li, Yuqian Yuan, Binhe Yu, Haoyuan Li, Wanggui He, Hao Jiang, Mengze Li, Xiaohui Song, Siliang Tang, Jun Xiao, Hui Lin, Yueting Zhuang, and Beng Chin Ooi. 2025. HealthGPT: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. *arXiv preprint arXiv:2502.09838*.
- Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. PMC-CLIP: Contrastive language-image pre-training using biomedical

- documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 525–536. Springer.
- 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.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024a. Visual instruction tuning. *Advances in neural information processing systems*, 36.
- Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, Lei Zhang, Jianfeng Gao, and Chunyuan Li. 2024b. LLaVA-Plus: Learning to use tools for creating multimodal agents. In *European conference on computer vision*, pages 126–142. Springer.
- Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The FLAN collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, pages 22631–22648. PMLR.
- Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Yushuai Wu, Mu Qiao, and Zaiqing Nie. 2023. BiomedGPT: Open multimodal generative pre-trained transformer for biomedicine. *arXiv preprint arXiv:2308.09442*.
- AI Meta. 2024. LLaMA 3.2: Revolutionizing edge AI and vision with open, customizable models. *Meta AI Blog*. Retrieved December, 20:2024.
- Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. 2023. Med-Flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR.
- Sahal Shaji Mullappilly, Mohammed Irfan Kurpath, Sara Pieri, Saeed Yahya Alseieri, Shanavas Cholakkal, Khaled Aldahmani, Fahad Khan, Rao Anwer, Salman Khan, Timothy Baldwin, and Hisham Cholakkal. 2024. BiMedix2: Bio-medical expert lmm for diverse medical modalities. *arXiv preprint arXiv:2412.07769*.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikanan Sankarasubbu. 2023. Med-halt: Medical domain hallucination test for large language models. *arXiv preprint arXiv:2307.15343*.
- Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. 2025. MedVLM-R1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 337–347. Springer.
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.
- Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, Juanma Zambrano Chaves, Szu-Yeu Hu, Mike Schaekermann, Aishwarya Kamath, Yong Cheng, David G. T. Barrett, Cathy Cheung, Basil Mustafa, Anil Palepu, Daniel McDuff, Le Hou, Tomer Golany, Luyang Liu, Jean baptiste Alayrac, Neil Houlsby, Nenad Tomasev, Jan Freyberg, Charles Lau, Jonas Kemp, Jeremy Lai, Shekoofeh Azizi, Kimberly Kanada, SiWai Man, Kavita Kulkarni, Ruoxi Sun, Siamak Shakeri, Luheng He, Ben Caine, Albert Webson, Natasha Latysheva, Melvin Johnson, Philip Mansfield, Jian Lu, Ehud Rivlin, Jesper Anderson, Bradley Green, Renee Wong, Jonathan Krause, Jonathon Shlens, Ewa Dominowska, S. M. Ali Eslami, Katherine Chou, Claire Cui, Oriol Vinyals, Koray Kavukcuoglu, James Manyika, Jeff Dean, Demis Hassabis, Yossi Matias, Dale Webster, Joelle Barral, Greg Corrado, Christopher Semturs, S. Sara Mahdavi, Juraj Gottweis, Alan Karthikesalingam, and Vivek Natarajan. 2024. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*.
- Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían

- Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, Lu Yang, Kejia Chen, Per Bjornsson, Shashir Reddy, Ryan Brush, Kenneth Philbrick, Mercy Asiedu, Ines Mezerreg, Howard Hu, Howard Yang, Richa Tiwari, Sunny Jansen, Preeti Singh, Yun Liu, Shekoofeh Azizi, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Riviere, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Elena Buchatskaya, Jean-Baptiste Alayrac, Dmitry Lepikhin, Vlad Feinberg, Sebastian Borgeaud, Alek Andreev, Cassidy Hardin, Robert Dadashi, Léonard Hussenot, Armand Joulin, Olivier Bachem, Yossi Matias, Katherine Chou, Avinatan Hassidim, Kavi Goel, Clement Farabet, Joelle Barral, Tris Warkentin, Jonathon Shlens, David Fleet, Victor Cotruta, Omar Sanseviero, Gus Martins, Phoebe Kirk, Anand Rao, Shravya Shetty, David F. Steiner, Can Kirmizibayrak, Rory Pilgrim, Daniel Golden, and Lin Yang. 2025. MedGemma technical report. *arXiv preprint arXiv:2507.05201*.
- LASA Team, Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, Yu Sun, Junao Shen, Chaojun Wang, Jie Tan, Deli Zhao, Tingyang Xu, Hao Zhang, and Yu Rong. 2025. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*.
- Hieu Tran, Zhichao Yang, Zonghai Yao, and Hong Yu. 2024. Bioinstruct: instruction tuning of large language models for biomedical natural language processing. *Journal of the American Medical Informatics Association*, 31(9):1821–1832.
- Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, Basil Mustafa, Aakanksha Chowdhery, Yun Liu, Simon Kornblith, David Fleet, Philip Mansfield, Sushant Prakash, Renee Wong, Sunny Virmani, Christopher Sem-turs, S. Sara Mahdavi, Bradley Green, Ewa Dominowska, Blaise Aguerre y Arcas, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Karan Singhal, Pete Florence, Alan Karthikesalingam, and Vivek Natarajan. 2024. Towards generalist biomedical AI. *NEJM AI*, 1(3):AIoa2300138.
- Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 13484–13508.
- Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. 2025. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *Nature Communications*, 16(1):7866.
- Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Towards generalist foundation model for radiology. *arXiv preprint arXiv:2308.02463*.
- Yunfei Xie, Ce Zhou, Lang Gao, Juncheng Wu, Xianhang Li, Hong-Yu Zhou, Sheng Liu, Lei Xing, James Zou, Cihang Xie, and Yuyin Zhou. 2024. MedTrinity-25M: A large-scale multimodal dataset with multigranular annotations for medicine. *arXiv preprint arXiv:2408.02900*.
- Zhiling Yan, Kai Zhang, Rong Zhou, Lifang He, Xiang Li, and Lichao Sun. 2023. Multimodal ChatGPT for medical applications: an experimental study of GPT-4V. *arXiv preprint arXiv:2310.19061*.
- Zhichao Yang, Zonghai Yao, Mahbuba Tasmin, Parth Vashisht, Won Seok Jang, Feiyun Ouyang, Beining Wang, David McManus, Dan Berlowitz, and Hong Yu. 2025. Unveiling GPT-4V’s hidden challenges behind high accuracy on USMLE questions: Observational study. *Journal of Medical Internet Research*, 27:e65146.
- Zonghai Yao, Benlu Wang, Yifan Zhang, Junda Wang, Iris Xia, Zhipeng Tang, Shuo Han, Feiyun Ouyang, Zhichao Yang, Arman Cohan, and Hong Yu. 2025. Medical thinking with images.
- Zonghai Yao and Hong Yu. 2025. A survey on LLM-based multi-agent AI hospital.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Jiayu Lei, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. RadGenome-Chest CT: A grounded vision-language dataset for chest CT analysis. *arXiv preprint arXiv:2404.16754*.

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*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2025. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

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 Evaluation Metrics and Dataset Statistics

Doctor Score is the primary human evaluation of model rationales. GPT-4o and DeepSeek provide complementary automatic analyses, and MultiViewDx-AN ranks highest under all three evaluation routes among the displayed systems. Style normalization reduces surface-form differences before automatic judging. Table 7 uses the DeepSeek score to compare model ordering on its own scale.

A.1 UMLS-F1 Factuality Metric

To evaluate the factual accuracy of LLM outputs, we leverage the UMLS concept overlap metric. The Unified Medical Language System (UMLS) (Bodenreider, 2004) enhances biomedical interoperability by unifying a comprehensive collection of biomedical terminologies, classification systems, and coding standards. By reconciling semantic variances and representational disparities across different biomedical repositories, UMLS facilitates standardized understanding.

We employ the Scispacy library² to identify and align medical named entities in texts with their corresponding UMLS concepts. Scispacy excels in entity recognition, enabling accurate association of named entities in LLM outputs with relevant UMLS concepts, a critical capability for assessing factual accuracy.

Our analytical process utilizes precision and recall metrics. Precision measures the proportion of shared concepts between the LLM output and the ground truth, indicating factual correctness. Recall assesses how well the LLM output covers the concepts present in the ground truth, reflecting the relevance of the information. Formally, given the concept sets from the ground truth (C_{ref}) and the LLM output (C_{gen}), precision and recall are calculated as:

$$\text{Precision} = \frac{|C_{\text{ref}} \cap C_{\text{gen}}|}{|C_{\text{gen}}|}, \quad (1)$$

$$\text{Recall} = \frac{|C_{\text{ref}} \cap C_{\text{gen}}|}{|C_{\text{ref}}|}. \quad (2)$$

The F1 score, derived from these precision and recall values, provides a balanced measure of the LLM output’s accuracy and relevance.

²Using the Scispacy *en_core_sci_lg* model

Model	Acc.	UMLS-F	GPT4-O	DS-O
Claude3-Opus	58.4	0.18	1.17	2.31
GPT-4o-mini	46.2	0.16	0.91	1.95
Huatuo-7B	34.5	0.13	1.08	2.06
Huatuo-34B	44.7	0.16	1.13	2.24
MultiViewDx	41.2	0.11	0.78	1.86
MultiViewDx-AN	58.5	0.23	1.29	2.55

Table 7: Auxiliary automatic-evaluator sensitivity analysis. Acc.=Accuracy; UMLS-F=UMLS-based factuality; GPT4-O=GPT-4o Overall; DS-O=DeepSeek Overall. MultiViewDx-AN ranks highest under both automatic evaluators among the displayed rows; DS-O is interpreted on the DeepSeek scale.

A.2 JAMA Specialty Distribution

	Count
Oto-HNS	513
Ophth	466
Derm	368
JAMA	328
JN Learn	299
Surg	133
Oncol	105
Cardiol	92
Neurol	61
Pediatr	60
Psych	6

Table 8: Distribution across JAMA specialty journals. Abbr.: Oto-HNS=JAMA Otolaryngology–Head & Neck Surgery, Ophth=JAMA Ophthalmology, Derm=JAMA Dermatology, Surg=JAMA Surgery, Oncol=JAMA Oncology, Cardiol=JAMA Cardiology, Neurol=JAMA Neurology, Pediatr=JAMA Pediatrics, Psych=JAMA Psychiatry, JN Learn=JN Learning.

B Experiment Settings and Data Construction Examples

B.1 Experiment and Implementation Details

This section provides the experimental details that are summarized in Sec. 4. The main text focuses on the training and evaluation pipeline in Fig. 3; here we provide the full baseline list, benchmark split details, training configuration, and variant definitions used in the results.

B.2 Baseline Models

We organize baselines into three groups: proprietary multimodal models, general-purpose open multimodal VLMs, and medical multimodal VLMs. This grouping matches the main benchmark comparison in Table 2.

Proprietary models. We include strong closed-source baselines that are commonly used as upper bounds for general multimodal reasoning, in-

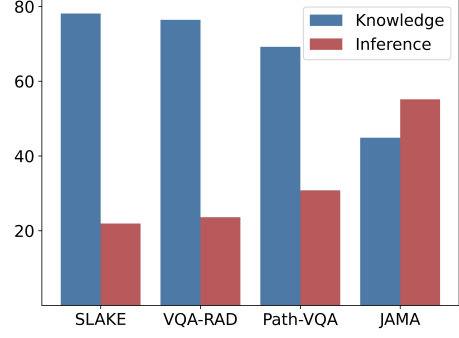


Figure 4: This figure illustrates the proportion of questions assessing knowledge and inference in the Slake, VQA-RAD, Path-VQA, and JAMA Clinical Challenge datasets.

cluding GPT-4.1, GPT-4o, GPT-4o-mini, Claude3-Opus, Claude Sonnet 4, and Gemini-2.5-Flash.

General-purpose multimodal VLMs. We benchmark a diverse set of open VLMs across model families and scales, including Qwen2.5VL-7B/32B (Bai et al., 2025), Janus-Pro-7B (Chen et al., 2025), InternVL2.5-8B/38B (Chen et al., 2024c), InternVL3-8B/14B/38B (Zhu et al., 2025), Llama3.2-11B (Meta, 2024), Qwen-VL-Chat, Yi-VL-34B (01.AI et al., 2024), and LLaVA variants including LLaVA-7B/8B/13B/34B (Liu et al., 2024b).

Medical multimodal VLMs. We compare against representative medical-domain models, including BiomedGPT (Luo et al., 2023), Med-R1-2B (Lai et al., 2025), MedVLM-R1-2B (Pan et al., 2025), HealthGPT-M3 (Lin et al., 2025), BioMedix2-8B (Mullappilly et al., 2024), MedGemma-4B-IT (Sellersgren et al., 2025), HuatuoGPT-Vision-7B/34B (Chen et al., 2024b), Lingshu-7B (Team et al., 2025), Hulu-Med-7B (Jiang et al., 2025), Med-Flamingo (Moor et al., 2023), RadFM (Wu et al., 2025), LLaVA-Med (Li et al., 2023), and LLaVA-Med variants trained with PubMedVision-style data.

B.3 Benchmark Splits

For VQA-RAD, SLAKE, PathVQA, and PMC-VQA, we follow the official dataset partitions. Only the official training split of each dataset is used during fine-tuning, and evaluation is conducted exclusively on the corresponding official held-out test split. Thus, although these datasets contribute training samples to MultiViewDx fine-tuning, their test samples are not used during training. For the

JAMA Clinical Challenge, which does not provide the same standard train/test split as the VQA benchmarks, we construct a disjoint case-level split. We use 60% of cases for training and reserve the remaining 40% exclusively for testing. All JAMA accuracy, UMLS-F1, GPT-4o rubric, and Doctor Score results are reported on this held-out 40% test split. PMC-VQA* follows its official QA split and shares PMC/PMC-OA source provenance with Stage 1.

B.4 Training Configuration and Compute

Table 9 summarizes the training configuration used for MultiViewDx-8B. The compute values are approximate and are reported to clarify the scale of training.

We adopt the three-stage training pipeline shown in Fig. 3. For pre-training (Radford et al., 2018; Chen et al., 2023), we filter the original PMC data by removing captions with fewer than 20 words, yielding 14M samples, and train with a learning rate $1e-4$, image token length 512, DeepSpeed Stage 2 (batch size 8, gradient accumulation 6), updating only the adapter. This stage runs on 4 H100 GPUs for 420 hours. For instruction tuning (Longpre et al., 2023; Wang et al., 2023; Tran et al., 2024; Han et al., 2023), we fine-tune on MultiViewDx with learning rate $2e-5$ under DeepSpeed Stage 2 (batch size 6, gradient accumulation 6), updating full parameters, and train on 8 H100 GPUs for 100 hours. Overall, the full training pipeline uses approximately 2,400 H100 GPU-hours and has an estimated compute cost on the order of $0.81-1.0 \times 10^{21}$ FLOPs. Finally, we apply the third stage in Fig. 3: sequential fine-tuning on human-annotated clinical cases with learning rate $1e-5$, using additional GPT-4o augmentation described in Appendix B.6. This sequential fine-tuning stage is designed to further align the model with expert-authored diagnostic workflow, fine-grained evidence grounding, and case-level multi-view reasoning.

B.5 Prompt for Knowledge/Inference Classification

We use GPT-4o to classify each benchmark item as knowledge-oriented or inference-oriented.

You are classifying medical QA examples into two categories.

Input: Question: [QUESTION]

Answer choices: [CHOICES, if available]

Reference answer: [REFERENCE ANSWER]

Label the example as one of the following:

Knowledge: the question can be answered mainly by recognizing a visual pattern, recalling a medical fact, or matching a short image-text association. It does not require combining multiple evidence sources or multi-step clinical reasoning.

Inference: the question requires combining multiple findings, using clinical context, comparing differential diagnoses, reasoning across views or modalities, or making a multi-step clinical judgment.

Return only one label: Knowledge or Inference.

B.6 Template Prompt

Generate Instruction Data. We construct instruction data with both closed-ended and open-ended formats. For close-ended datasets (e.g., PMC-VQA, AMBOSS, JAMA train, SLAKE train, VQA-RAD train, PathVQA), we generate only answer options. For open-ended tasks (mainly from JAMA), we additionally require step-by-step reasoning. We use GPT-4o to generate question-answer pairs from each image and its augmented caption (3-10 QAPs per caption), and we constrain generation to be grounded in the provided caption and image evidence (including ROI-level cues when available). Full prompt templates are shown in Github.

Evaluation Pipeline Prompt. For close-ended QA we report accuracy. For open-ended diagnostic reasoning, we use two stricter metrics: UMLS-based factuality and a GPT-4o rubric along **Key Points**, **Inference**, and **Evidence**. To reduce style bias, we first normalize model outputs into a consistent style, then ask the judge to score each aspect independently based on extracted summaries. Prompt details are provided in our Github.

B.7 JAMA Case Study: Clinical Prioritization

Table 14 presents a qualitative JAMA case study used to analyze diagnostic reasoning behavior across models. The case requires more than recognizing ophthalmic image findings: the model must

Stage	Data	Samples	LR	Updated Params.	GPU Setting	Approx. Compute
Stage 1: Medical pretraining	Filtered PMC image-text data	14M	1e−4	Adapter only	4 H100 GPUs	~1,600 H100 hours
Stage 2: Instruction tuning	GPT-4o-synthesized MultiViewDx data	4M	2e−5	Full model	8 H100 GPUs	~640 H100 hours
Stage 3: Sequential fine-tuning	Human-annotated clinical cases	1M	1e−5	Full model	8 H100 GPUs	~160 H100 hours
Total	–	–	–	–	–	~2,400 H100 hours

Table 9: Training configuration and approximate compute cost for MultiViewDx-8B. Stage 1 provides broad medical image-text grounding; Stage 2 adds instruction following and evidence-linked QA from GPT-4o-synthesized data; Stage 3 sequentially fine-tunes on human-annotated clinical cases. The overall estimated compute is on the order of $0.81\text{--}1.0 \times 10^{21}$ FLOPs.

combine bilateral optic nerve edema, pregnancy, hypertension, and chronic kidney disease to prioritize immediate vital-sign assessment for possible malignant hypertension. This example supports the main-paper discussion that clinical diagnostic reasoning is not fully captured by answer accuracy alone and that models may fail by over-weighting a salient but less urgent disease prior, such as diabetes, instead of selecting the clinically urgent next step.

B.8 Full-versus-Non-Multi-View Ablation Case

Table 10 presents illustrative simulated outputs on the same JAMA Clinical Challenge case in Table 14.

The full-model response connects bilateral fundus findings, OCT evidence, and systemic risk factors before selecting the urgent next step. The Non-Multi-View response recognizes local retinal abnormalities but anchors on diabetes, omits the cross-view clinical link, and selects the wrong action.

B.9 Benchmark-Training Exposure Matrix

Table 12 records documented benchmark-training exposure for every model in Table 2: ✓ indicates use of the official training partition or a rewritten derivative, × indicates no reported use in the public training recipe, and “?” indicates undisclosed checkpoint-level exposure.

C Additional Data Construction Examples

This section contains the JAMA case and paired ablation outputs, the view-removal result, benchmark-exposure tables, a human-annotated case in the MultiViewDx format, and a retrieval-guided synthesis example.

C.1 Human-Annotated Case Normalization

Table 15 shows an expert-authored clinical case after being organized into the MultiViewDx format. The example illustrates the information fields used in our human-annotated branch, including image caption, clinical history, image findings, discussion, and instruction-style supervision. This supports the data construction section in the main paper by showing how expert-authored cases provide evidence-linked supervision rather than only image-answer pairs.

C.2 Retriever-Guided Synthetic Construction

Table 16 provides an example of the retrieval-guided synthesis process. Without retrieved context, GPT-4o produces a generic image description and remains uncertain about the target structure. With retrieved clinical cases, the generation is conditioned on related medical evidence and becomes more specific to the cardiac finding. This example supports the data construction section by illustrating why retrieval is used in the GPT-4o synthesis branch: it provides contextual medical evidence for generating more grounded instruction data.

Model	Simulated generated output on the same input	Observable result
MultiViewDx-AN (full multi-view)	Answer: C. Measure heart rate, respiratory rate, and blood pressure. The right- and left-eye fundus photographs show bilateral optic-disc edema, more pronounced in the left eye with associated nerve-fiber-layer hemorrhage. OCT additionally shows center-involving macular edema in both eyes. In a 15-week pregnant patient with hypertension and chronic kidney disease, this bilateral fundus-plus-OCT pattern is concerning for hypertensive end-organ injury rather than isolated progression of diabetic retinopathy. The immediate next step is to check vital signs and blood pressure to evaluate for malignant hypertension with papillopathy.	Correct. Explicitly links both fundus views, bilateral OCT findings, and systemic risk factors before selecting the urgent next step.
MultiViewDx-NoMV (non-multi-view)	Answer: B. Determine blood glucose level and perform glycated hemoglobin testing. The patient has type 1 diabetes and proliferative diabetic retinopathy. The retinal hemorrhages and bilateral macular edema are compatible with worsening diabetic retinopathy during pregnancy. Blood glucose and HbA1c should therefore be assessed to determine whether poor glycemic control is causing the visual decline and to guide diabetes management.	Incorrect. Recognizes plausible local abnormalities but anchors on diabetes and fails to connect the bilateral optic-disc pattern and OCT findings with hypertension, kidney disease, and pregnancy.

Table 10: Illustrative simulated outputs from the full and Non-Multi-View settings on the JAMA case in Table 14.

Inference condition	Cases	Accuracy
Evidence-bearing view removed; vignette and remaining inputs retained	100	31%

Table 11: Inference-time removal of the evidence-bearing view on 100 multi-view cases. Accuracy decreases to 31%.

Table 12: Benchmark-training exposure for every model reported in Table 2.

Model	VQA-RAD	SLAKE	Path VQA	PMC-VQA*	Documented training stage / exposure basis
Medical multimodal VLMs					
BiomedGPT	NAN	NAN	NAN	NAN	The paper reports task-specific fine-tuning on VQA-RAD, SLAKE, and PathVQA; the Table 2 checkpoint is unspecified.
Med-R1-2B	×	×	×	×	GRPO training uses an open-access OmniMedVQA split; the four official training partitions are not listed.
MedVLM-R1-2B	NAN	NAN	NAN	NAN	GRPO uses a pooled set containing all four benchmarks; per-source split provenance is undisclosed.
HealthGPT-M3	✓	✓	✓	×	Stage 3 visual instruction fine-tuning includes VQA-RAD, SLAKE, and PathVQA; PMC-VQA is not listed.
BioMedix2-8B	✓	✓	✓	×	Multimodal instruction alignment uses restructured VQA-RAD, SLAKE, and PathVQA training data.
MedGemma-4B-IT	✓	✓	×	×	Medical post-training data explicitly include VQA-RAD and SLAKE; PathVQA and PMC-VQA are not listed.
HuatuoGPT-V-7B	×	×	×	×	PubMedVision instruction tuning followed by zero-shot benchmark evaluation.
Lingshu-7B	✓	✓	✓	✓	Medical multimodal instruction tuning includes all four datasets.
Med-Flamingo	×	×	×	×	Continued medical pretraining; the evaluation recipe lists no target-split training.
RadFM	✓	✓	×	✓	Visual-instruction training includes VQA-RAD, SLAKE, and PMC-VQA; PathVQA is not listed.
LLaVA-Med-7B	×	×	×	×	The released PubMed-derived checkpoint is evaluated zero-shot.
+ PubMedVision	×	×	×	×	PubMedVision medical instruction tuning followed by zero-shot evaluation.
HuatuoGPT-Vision-34B	×	×	×	×	PubMedVision medical instruction tuning followed by zero-shot evaluation.
Hulu-Med-7B	✓	✓	✓	✓	Mixed-modality instruction tuning includes the four benchmark training sets or rewritten derivatives.
MultiViewDx-8B-AN	✓	✓	✓	✓	Stage 3 human-annotated fine-tuning uses all four official training partitions.

*The PMC-VQA column records official QA training-partition exposure; Sec. 4.2 describes the shared PMC/PMC-OA provenance.

Model	VQA-RAD	SLAKE	Path VQA	PMC-VQA*	Training stage in the Table 3 setting
LLaVA-v1.5-LLAMA3-8B	✓	✓	✓	✓	Benchmark-specific supervised fine-tuning on each official training split.
LLaVA_Med-8B	✓	✓	✓	✓	Benchmark-specific supervised fine-tuning on each official training split.
HuatuoGPT-Vision-8B	✓	✓	✓	✓	Benchmark-specific supervised fine-tuning on each official training split.
MultiViewDx	✓	✓	✓	✓	Benchmark-specific supervised adaptation on the corresponding official training split.

Table 13: Benchmark-training exposure for every model in Table 3.

Table 14: JAMA case study illustrating a clinical prioritization error.

Clinical Prioritization Error

Question. A woman in her mid-20s presented with subacute bilateral vision loss that was worse in the left eye. Her medical history was remarkable for type 1 diabetes diagnosed at 16 years of age and proliferative diabetic retinopathy in both eyes that had been treated with panretinal photocoagulation 7 years earlier. She had undergone pars plana vitrectomy with endolaser to treat a tractional retinal detachment in her right eye 2 years before this presentation. She also had a history of hypertension and chronic kidney disease, and she was 15 weeks into pregnancy. Visual acuity was 20/50 OD and 20/100 OS. Intraocular pressure was normal bilaterally, and no relative afferent pupillary defect was detected. Findings of an anterior segment examination were normal. The patient was in no apparent distress and denied any headache, chest pain, or focal weakness. Ophthalmoscopic examination (Figure) revealed mild optic nerve head edema that was greater in the left eye than the right eye with associated nerve fiber layer hemorrhage in the left eye. Nerve fiber layer infarctions, dot and blot hemorrhages, and lesions caused by panretinal photocoagulation also were seen bilaterally. Optical coherence tomography showed macular edema that involved the center of the macula in both eyes (Figure, inset).



Choices:

- (A) Obtain a fluorescein angiogram;
- (B) Determine blood glucose level and perform glycated hemoglobin test;
- (C) Measure heart rate, respiratory rate, and blood pressure;
- (D) Perform immediate computed tomography of the head.

Correct Answer: (C) Measure heart rate, respiratory rate, and blood pressure.

Clinical Diagnosis: Malignant hypertension with papillopathy.

Diagnostic Reason: Malignant hypertension with papillopathy C. Measure heart rate, respiratory rate, and blood pressure The patient was found to have hypertension, with a blood pressure of 195/110 mm Hg. Heart and respiratory rates were normal. Measurement of the arterial blood pressure may be performed rapidly in the clinic with a sphygmomanometer and is essential to rule out malignant hypertension, which is a potentially life-threatening cause of vision loss. Although the differential diagnosis for bilateral optic nerve edema is broad, workup should always include assessment of blood pressure when appropriate, because a hypertensive emergency (also known as malignant hypertension) may cause substantial morbidity or mortality if not diagnosed and treated promptly. Findings may include macular star, macular edema, serous retinal detachment, intraretinal hemorrhage, and optic disc edema with or without associated hemorrhage^{1,2}. Optic nerve head edema may occur with systolic blood pressures as low as 160 mm Hg, with the median onset occurring at 188 mm Hg.³

The macular edema associated with hypertensive retinopathy may be distributed more nasally, as was seen in this patient.⁴ This patient had mild optic nerve edema despite high systemic blood pressure and substantial macular edema. This less-pronounced optic nerve edema likely was attributable to optic nerve atrophy at baseline. Proliferative diabetic retinopathy and panretinal photocoagulation can be associated with optic atrophy, and atrophic optic nerves tend to become less edematous than healthy optic nerves.⁵⁻⁷ Regarding the other choices above, a fluorescein angiogram (choice A) would be expected to show leakage from the optic nerve and macula, but such findings are already available from the optical coherence tomography, which showed intraretinal and subretinal fluid. Although assessment of serologic levels of glucose and glycated hemoglobin (choice B) is important in the management of diabetic retinopathy, results from such tests are not immediately available and have less bearing on the immediate management of the case (this patient's glycated hemoglobin level was 10.5% [to convert to a proportion of total hemoglobin, multiply by 0.01]). Performing computed tomography (choice D) would be a reasonable later step, especially if the patient's blood pressure was found to be normal, to rule out an intracranial mass. In addition to hypertension, other causes of optic nerve head edema exist. Papilledema (when optic nerve head edema is secondary to increased intracranial pressure) is often bilateral and may be caused by intracranial mass lesions, meningitis, cerebral venous thrombosis, or intracranial hypertension, or it may be idiopathic. Bilateral optic nerve head edema also may be caused by diabetic papillopathy (which is a rare cause of optic disc edema, but which is bilateral in up to 50% of cases^{8,9}) and toxic optic neuropathies. Optic disc drusen are frequently bilateral and may cause so-called pseudopapilledema.¹⁰ Severe hypertension resulting in acute retinopathy and optic nerve head edema is considered a hypertensive emergency because of its association with end-organ damage, namely, damage to the retina and optic nerve. This case serves as a reminder that, although the differential diagnosis for optic nerve head edema is broad, immediately assessing the patient's blood pressure to evaluate for hypertensive emergency in the eye clinic may be critical, because this treatable condition could lead to substantial morbidity or mortality if diagnosis is delayed. The patient was admitted to the medicine service for management of hypertension. At an ophthalmology follow-up examination 3 months later, her blood pressure was 128/73 mm Hg and her visual acuity was 20/40 OD and 20/60 OS, close to her visual acuity before this acute event. On examination, the optic nerve head edema had resolved, and optical coherence tomography showed resolution of macular edema.

Prediction Rationale

MultiViewDx. C. Measure heart rate, respiratory rate, and blood pressure. Malignant hypertension with papillopathy C. Measure heart rate, respiratory rate, and blood pressure The key to the correct diagnosis is the presence of bilateral optic neuropathy in a pregnant woman. The patient's history of hypertension and chronic kidney disease, along with the findings of bilateral optic neuropathies, suggested a systemic etiology. A fluorescein angiogram (choice A) would not be the next step because the patient was asymptomatic and had no signs of retinal ischemia. A glycated hemoglobin test (choice B) would not be indicated because the patient had no signs of diabetic retinopathy. A computed tomography scan of the head (choice D) would not be appropriate because the patient had a normal examination and no signs of intracranial hemorrhage. Besides hypertension, other causes of optic nerve head edema exist. Papilledema, which is optic nerve head edema secondary to increased intracranial pressure, is often bilateral and may result from intracranial mass lesions, meningitis, cerebral venous thrombosis, intracranial hypertension, or it may be idiopathic. Bilateral optic nerve head edema can also be caused by diabetic papillopathy—a rare cause of optic disc edema that is bilateral in up to 50% of cases¹—and toxic optic neuropathies. Optic disc drusen, which are frequently bilateral, may lead to so-called pseudopapilledema. Severe hypertension causing acute retinopathy and optic nerve head edema is considered a hypertensive emergency due to its association with end-organ damage, specifically affecting the retina and optic nerve. This case emphasizes that, although the differential diagnosis for optic nerve head edema is extensive, promptly assessing the patient's blood pressure in the eye clinic to evaluate for a hypertensive emergency is critical. Immediate evaluation is crucial because this treatable condition could lead to significant morbidity or mortality if diagnosis is delayed. The patient was admitted to the medical service for hypertension management. At an ophthalmology follow-up examination three months later, her blood pressure was 128/73 mm Hg, and her visual acuity improved to 20/40 in the right eye (OD) and 20/60 in the left eye (OS), nearing her visual acuity before this acute event. Upon examination, the optic nerve head edema had resolved, and optical coherence tomography showed a resolution of the macular edema.

Baseline Error Pattern

Claude3-Opus. This case presents several concerning features that make **measuring vital signs the most urgent next step: Bilateral optic nerve edema (worse in left eye)** in a **young pregnant woman** with:

- **Type 1 diabetes**
- **Hypertension**
- **Chronic kidney disease**
- **Active pregnancy (15 weeks)**

The key concern here is **POSTERIOR REVERSIBLE ENCEPHALOPATHY SYNDROME (PRES)**, which is:

- **A neurological condition characterized by vasogenic edema**
- **Commonly associated with severe hypertension**
- **More common in pregnant women**
- **Can present with visual symptoms and optic nerve edema**
- **Can be life-threatening if not treated promptly**

Why other options are less appropriate:

- **A (Fluorescein angiogram): Contraindicated in pregnancy**
- **B (Blood glucose/HbA1c): While important for diabetic management, not the most urgent concern given the presentation**
- **D (Head CT): While eventually needed if PRES is suspected, vital signs should be checked first as they'll guide immediate management**

The vital signs are crucial because:

- They will help determine if this is a hypertensive emergency requiring immediate intervention.
- They're non-invasive and provide immediate critical information

The findings will guide the urgency and direction of further workup

- If severely elevated blood pressure is found, immediate treatment could prevent serious complications for both mother and fetus
- This represents a potential medical emergency where checking vital signs is the most appropriate first step to guide immediate management decisions.

HuatuoGPT-Vision-34B. The patient, who has a history of **type 1 diabetes** and is currently **15 weeks pregnant**, presents with **optic nerve head edema, macular edema, and hemorrhages in both eyes**; therefore, it is crucial to **assess her blood glucose levels and perform a glycated hemoglobin (HbA1c) test to evaluate and manage her glycemic control status**, as this evaluation is essential to **identify any acute changes or complications related to her pregnancy and existing diabetes**, since poor glycemic control can lead to rapid progression of diabetic retinopathy, exacerbate ocular conditions, and pose significant risks to both maternal and fetal health—including the potential for congenital anomalies, preeclampsia, and worsening of chronic kidney disease—thus, **timely assessment and adjustment of her diabetes management plan are vital to prevent further vision loss and ensure the well-being of both the mother and the developing fetus.**

HuatuoGPT-Vision-7B. B. Determine blood glucose level and perform glycated hemoglobin test Reasoning: The patient has a history of **type 1 diabetes**, which is known to cause **diabetic retinopathy leading to vision loss** as seen in the images provided (**optic nerve head edema with hemorrhage**). The presence of **macular edema** also suggests **ongoing or poorly controlled diabetes** affecting her eyes. **A fluorescein angiogram would be useful for assessing vascular abnormalities but may not directly address the underlying issue related to poor glycemic control.** Measuring heart rate, respiratory rate, and blood pressure are important general assessments but do not specifically relate to the ophthalmic findings presented here. Immediate computed tomography of the head might be considered if there was suspicion of other complications like intracranial pathology; however, it's less relevant given the current clinical context focused on eye examination results linked to diabetes management issues.

GPT-o-mini. Correct Answer: C. Measure heart rate, respiratory rate, and blood pressure. Reason: Measuring blood pressure is critical in this scenario to identify malignant hypertension , a potentially life-threatening condition that can lead to vision loss through optic nerve and retinal damage . Immediate assessment of blood pressure helps rule out hypertensive emergencies , allowing for timely intervention .	
Expert Analysis	
Model	Answer and Main Behavior
MultiViewDx	Correct: (C). Correctly links bilateral optic nerve edema, pregnancy, hypertension, and chronic kidney disease to the need for immediate vital-sign assessment and blood pressure measurement.
Claude3-Opus	Correct: (C). Correctly prioritizes vital signs, but frames the concern more broadly as a hypertensive neurologic emergency.
GPT-4o-mini	Correct: (C). Selects the correct action but provides a short rationale with limited supporting evidence.
HuatuoGPT-Vision-34B	Incorrect: (B). Focuses on diabetes and glycemic control, missing the more urgent need to rule out hypertensive emergency.
HuatuoGPT-Vision-7B	Incorrect: (B). Attributes the eye findings mainly to diabetic retinopathy and fails to prioritize blood pressure assessment.
MultiViewDx and Claude3-Opus accurately recognized the urgency of measuring vital signs to assess for malignant hypertension in this patient, who presents with bilateral optic nerve edema and a history of hypertension and chronic kidney disease during pregnancy. Their reasoning reflects a proper understanding of the need for immediate intervention to prevent serious complications. On the other hand, Huatuo-GPT-Vision-34B and Huatuo-GPT-Vision-7B focused on assessing glycemic control by selecting to determine blood glucose levels and perform a glycated hemoglobin test. While managing diabetes is important, they failed to prioritize the immediate life-threatening condition suggested by the patient's symptoms, thus overlooking the critical need to rule out a hypertensive emergency. GPT-4o-mini could get the correct answer and some key points but lack lots of detail evidence to prove it.	

Table 15: Human-annotated clinical case normalized into the Multi-ViewDx evidence-linked schema.

Human Annotated Sample Case	
	
Image Caption: 1. Sagittal T2-weighted image shows a large high signal intensity cystic mass (red arrow) with a nodular, low signal intensity component (yellow arrow). Normal left ovary with follicles (white arrow) is seen posteriorly – negative beak sign. 2. Sagittal T2-weighted image shows a large high signal intensity cystic mass (red arrow) arising from right ovary (white arrow) – positive beak sign. There is another small, nodular, low signal intensity component (yellow arrow). 3. Axial T1-weighted image shows a large cystic mass (red arrow). The lesion has parts of low signal intensity (yellow arrow) and its content is slightly hyperintense (asterisk). 4. Axial fat-suppressed post-contrast T1-weighted image shows wall enhancement (red arrow) and solid component enhancement (yellow arrow) and its content is hypointense (asterisk). 5. Diffusion-weighted MR image (b1000) shows hyperintensity of the solid component (yellow arrow). 6. Axial ADC map from diffusion-weighted MR image (Fig. 1e) demonstrates marked hypointensity of the solid component (yellow arrow), in keeping with dense cellularity of the lesion.	
Clinical History: A 21-year-old G0P0 woman with no medical history was referred to our institution for a sonographically detected cystic right adnexal mass. She has a history of pelvic discomfort without other complaints. Physical examination was normal. Laboratory findings were also normal except for an elevated CA 125 65.2 U/mL (normal <35.0).	

Image Findings: MRI examination revealed a cystic tumour arising from the right ovary with 7.5 cm. On T2-weighted images, the signal intensity of the cyst content was high and two small nodular peripheral solid components were detected, adhering to its internal wall, with low signal (Fig. 1a, b). The normal left ovary was present with follicles (Fig. 1a). On pre-contrast T1-weighted images, the mass exhibited slightly high signal intensity (Fig. 1c). On contrast-enhanced fat-suppressed T1-weighted images, wall enhancement and solid component enhancement were detected (Fig. 1d). Finally, the ADC map (Fig. 1f) from diffusion-weighted image (Fig. 1e) demonstrates marked hypointensity of the solid component, in keeping with its dense cellularity. Surgical excision was proposed and accepted by the patient. The histopathological investigation revealed a typical ovarian serous borderline tumour.

Discussion: Borderline ovarian tumours are uncommon ovarian neoplasms, intermediate between benign and malignant types, corresponding to 5% of all epithelial ovarian tumours. [1, 2] Serous borderline tumour represents the most common type of borderline tumours arising in the ovary, and typically, it is confined to the adnexa and presents an indolent course. [3] However, up to 6.8% of these tumours can progress to low grade serous carcinoma. [3] Serous borderline tumours are divided into typical (90%) and borderline tumours with micro-papillary patterns (5%–10%). [4] These neoplasms usually present as bilateral adnexal masses with more proliferation of papillary projections than do benign cystadenomas, they are often seen in younger patients, and laboratory findings show the serum CA-125 level mildly elevated. [2, 3, 5, 6] The peak age of presentation is 45 years. [5] Small tumours usually do not cause symptoms and are often detected as an incidental finding on sonography. [7] Larger or more advanced neoplasms might cause pain or pelvic discomfort. The diagnosis of this type of tumour is based on histopathological examination. As they are staged using the same ovarian cancer staging of malignant ovarian neoplasms [5], MRI plays a crucial role in this evaluation. There are no pathognomonic imaging features of borderline ovarian tumours and they frequently resemble a benign lesion. Nevertheless, some pathologic and MR imaging studies suggested that large papillary projections are highly suggestive of borderline or malignant neoplasms. [1, 8] The preferred treatment is bilateral oophorectomy. Fertility sparing procedures such as unilateral oophorectomy or cystectomy are related with a higher incidence of recurrence when compared with bilateral oophorectomy (10-20% vs. approximately 5% for radical surgery). [3, 4] Young age (<30 years) has also been found to be a risk factor for recurrent disease in patients treated conservatively. As in this case, patients should be instructed to complete child-bearing within a short time frame following initial diagnosis, due to the high risk of recurrence. [3] The most important feature in predicting an adverse outcome in serous borderline tumours is the presence of serous lesions involving the peritoneum (so-called invasive implants). [3, 4] Tumours without invasive implants, as in this case, have been considered to be nonaggressive tumours, associated with a good prognosis and high survival rate. Follow-up using ultrasound imaging is mandatory, with particular attention paid to the remaining ovary in conservatively treated young patients. [4] Written informed patient consent for publication has been obtained.

MultiViewDx(only a sample for one image):

User:What type of imaging technique is shown in Fig. 1a of the provided image?

GPT:The image shows a T2-weighted MRI scan.

User:Where is the cystic tumour located according to the image?

GPT:The cystic tumour is arising from the right ovary

User:What is the size of the cystic tumour?

GPT:The cystic tumour is 7.5 cm in size.

User:What does the red arrow in the image indicate?

GPT:The red arrow indicates the normal left ovary with follicles.

User:What does the yellow arrow in the image point to?

User:The yellow arrow points to one of the small nodular peripheral solid components adhering to the internal wall of the cystic tumour.

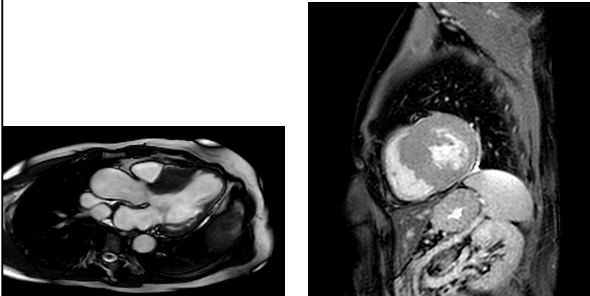
User:What characteristic of the cyst content is shown by the high signal intensity in the T2-weighted image?

User:The high signal intensity indicates that the cyst content is fluid-filled.

User:What is shown by the asterisk (*) in the image?

User:The asterisk (*) indicates the cystic tumour arising from the right ovary.

Table 16: Retriever-guided synthesis example comparing GPT-4o output with and without retrieved clinical context.

Retriever-Guided Example	
	Prompt: Generate the Image description based on the image. GPT-4o: The image is a sagittal CT scan, likely of the thoracic or abdominal region. The scan shows various soft tissue structures, including major organs. The blue arrow points toward a specific structure, possibly a blood vessel or part of an organ. This could indicate an abnormality such as a mass, an aortic aneurysm, or an area of vessel narrowing or dissection. In the background, major anatomical landmarks like the heart, aorta, and vertebral column are visible, as well as some portions of the liver or spleen. The areas of high contrast indicate dense structures, such as bones, while softer tissues are seen in varying shades of gray, highlighting different densities. This scan could be used to evaluate conditions like vascular abnormalities, tumors, or other pathologies in the chest or upper abdominal region. Further clinical information would be necessary to confirm the exact diagnosis.
Prompt(with retriever)	
	Prompt: Generate the Image description based on the image. Here are some medical guidelines and Clinical cases for you. Case Discussion: A case of asymmetric hypertrophic cardiomyopathy with mild subaortic stenosis and partial systolic anterior motion of the mitral valve. Since the pressure gradient obtained by phase contrast imaging was rather low, echocardiographic reevaluation and/or follow-up was recommended. The case also demonstrates a typical pattern of myocardial
fibrosis in the basal and midventricular septal segments at the anterior and posterior right ventricular insertion points.	
GPT-4o: The image is a coronal oblique CT scan of the thoracic region, focused on the heart and its surrounding structures. The blue arrow indicates a subaortic ventricular septal defect (VSD), measuring approximately 2.7 cm. This congenital heart defect refers to an abnormal opening in the interventricular septum (the wall separating the left and right ventricles of the heart), located just below the aortic valve. The surrounding anatomy includes parts of the heart chambers, likely showing the left and right ventricles, as well as nearby vascular structures such as the aorta. The defect appears to allow abnormal communication between the ventricles, which could lead to altered blood flow dynamics, potentially causing symptoms such as shortness of breath, fatigue, or heart failure if left untreated.	