

MedThinkVQA

Medical Thinking with Multiple Images



Zonghai Yao^{12*}, Benlu Wang^{3*}, Yifan Zhang¹⁴, ..., Arman Cohan³, Hong Yu¹²⁴

¹ Center for Healthcare Organization and Implementation Research, VA Bedford Health Care

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

³ Yale NLP Lab, Department of Computer Science, Yale University

⁴ Miner School of Computer and Information Sciences, UMass Lowell



Real clinical diagnosis ≠ single-image QA

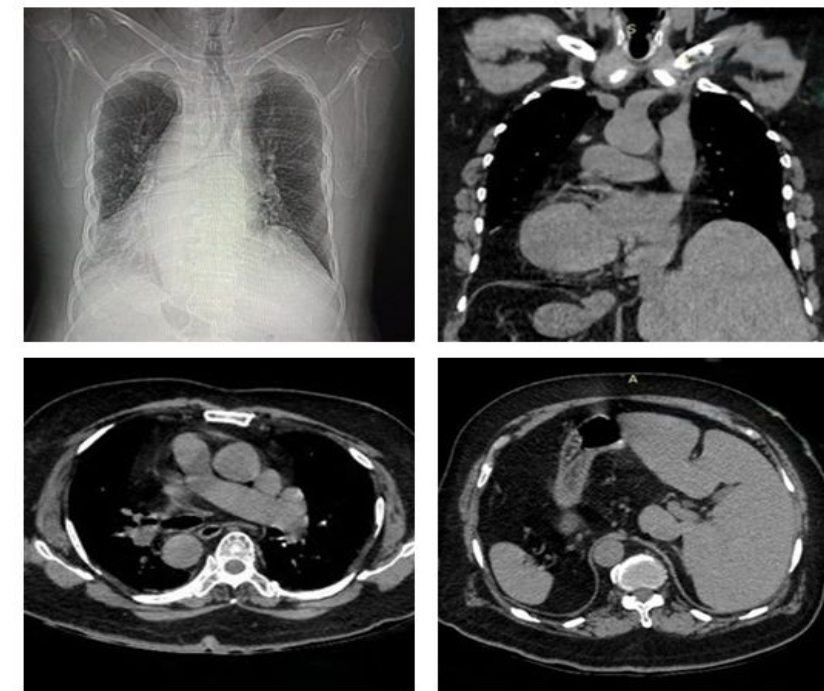
Real-world radiology cases come with many images across modalities and time, plus rich clinical histories. Clinicians first read each image, then integrate **cross-view evidence**, and finally reason through differentials before deciding on a diagnosis. Existing MedVQA / multimodal benchmarks mostly use **single images, auto-generated labels**, and **answer-only evaluation**, so they cannot test this stepwise “think-with-images” process.

Benchmark	# Case	Expert Annotation	Clinical Scenarios	# Img per Cas	Multi-Mod Imaging	Longitud Studies	Think-with-Images Intermediate Signals	Beyond-ACC Evaluation
VQA-Rad Lau et al.	451	X	X	0.45	X	X	X	X
VQA-Med Ben Abacha et al.	500	X	X	1.00	X	X	X	X
Path-VQA He et al.	6,719	X	X	0.13	X	X	X	X
SLAKE-En Liu et al.	1,061	X	X	0.09	X	X	X	X
PMC-VQA Zhang et al.	33,430	X	X	0.87	X	X	X	X
OmniMedVQA Hu et al.	127,995	X	X	0.92	X	X	X	X
GMAI-MMBench Chen et al.	21,281	X	X	1.00	X	X	X	X
GEMeX Liu et al.	1,605,575	X	X	1.00	X	X	X	X
Medical-Diff-VQA Hu et al.	700,703	X	X	1.23	X	X	X	X
MedFrameQA Yu et al.	2,851	X	X	3.24	X	X	X	X
ICG-CXR Ma et al.	11,439	X	X	2.00	X	X	X	X
MedRAX ¹ Fallahpour et al.	2,500	X	X	1.85	X	X	X	X
GEMeX-ThinkVG Liu et al.	206,071	X	X	1.00	X	X	X	X
MMMU (H & M) Yue et al.	1,752	✓	X	1.14	X	X	X	X
MMMU-Pro (H & M) Yue et al.	346	✓	X	1.25	X	X	X	X
S-Chain Le-Duc et al.	12,000	✓	X	1.00	X	X	X	X
MedXpertQA MM Zuo et al.	2,000	✓	✓	1.43	X	X	X	X
MedThinkVQA	10,067	✓	✓	6.68	✓	✓	✓	✓

- Most datasets: ≤ 1.4 images per case, often **single modality**, **no longitudinal studies**
- Few expert-annotated step traces; almost none with **per-image findings + case-level summary + teaching note**
- Evaluation is answer-accuracy only → cannot localize whether failure is due to image reading, cross-view fusion, or reasoning
- Need a benchmark that looks like real teaching cases and supervises the whole diagnostic process

MedThinkVQA dataset & coverage

Clinical Scenarios: One male patient aged 72 years complained of one-week chest pain and three-day exertional dyspnoea. There was no past history of cardiovascular, respiratory, or gastrointestinal disease ...



Based on ALL provided images together with the textual context, select the single best diagnosis from the options.

- A. Situs inversus totalis with dextrocardia
- B. Dextrocardia without situs inversus
- C. Kartagener syndrome
- D. Technical rotation artefact on imaging
- E. Congenital cardiac malformation with positional abnormality

MedThinkVQA contains **9** aggregated imaging modalities. About **29.0%** of cases include **longitudinal follow-up studies**, and cases span **19** of **22** ICD-10 chapters.

# Modalities	Train cases	Train (%)	Test cases	Test (%)
1	3,093	33.09	165	22.92
2	3,656	39.11	292	40.56
3	1,875	20.06	166	23.06
4	592	6.33	74	10.28
5	109	1.17	22	3.06
6	16	0.17	1	0.14
7	6	0.06	0	0.00
Total	9,347	100.00	720	100.00

Split	# Cases	# Longitudinal	Share (%)
Train	9,347	2,703	28.9
Test	720	219	30.4
Overall	10,067	2,922	29.0

Think-with-Images: 3 supervised steps

From per-image findings to Ddx and teaching discussion

Medical thinking with images (Twi)

Step1: per-image radiological annotations

Figure 1: Chest X-ray (PA view) demonstrating situs inversus totalis...
Figure 2/3/4

Step2: case-level integrated imaging findings

Chest X-ray demonstrated dextrocardia with gastric bubble on the right and thereby suggested situs inversus totalis. ECG demonstrated ...

Step3: differential diagnosis reasoning

The images show a clean mirror-image layout: the heart apex points right, the aortic arch tracks on the right, and on abdominal CT, the stomach/spleen are right-sided while the liver is left-sided: classic situs inversus totalis with dextrocardia.

Isolated dextrocardia (B) would keep normal abdominal situs, which we do not see.

Kartagener syndrome (C) needs

Rotation artifact (D) cannot

A congenital malformation with positional abnormality (E) usually gives mixed/ambiguous organ positions, not this symmetric mirror pattern. Hence only (A) Situs inversus totalis with dextrocardia fits all findings.

Diagnosing-Style Evaluation

MedTwi Eval

Step1 and Step2:

- ROUGE/RadCLIQ
- Image Perceptual Error
- Clinical Correct?
- Clinical Critical?
- Error Analysis

Step3: step by step

- Clinical Scenario Understanding
- Image Perceptual Error
- Lack of Medical Knowledge
- Reasoning Process Error

Case Discussion Eval

1. Section by section eval
2. Each section annotate sent-level correctness

Education Value Analysis

1. Disease Definition 2. Clinical Manifestation 3. Imaging Findings 4. Diagnostic Reasoning 5. Transferable Learning → Each scored 0-2

Medical Education Case Discussion

Background:
Situs inversus [1]

Clinical Perspective:
Patients with situs inversus totalis are frequently asymptomatic. However:

Imaging Perspective:
Initial chest radiography suggested dextrocardia with situs inversus. ECG findings further supported this. CT thorax and abdomen [5]

Clinical Significance:
Recognition of situs inversus totalis is vital in clinical and emergency settings to prevent misdiagnosis [10], especially in ...

Outcome:
No acute cardiac or respiratory pathology was identified in our case. The patient

Take Home Message / Teaching Points:

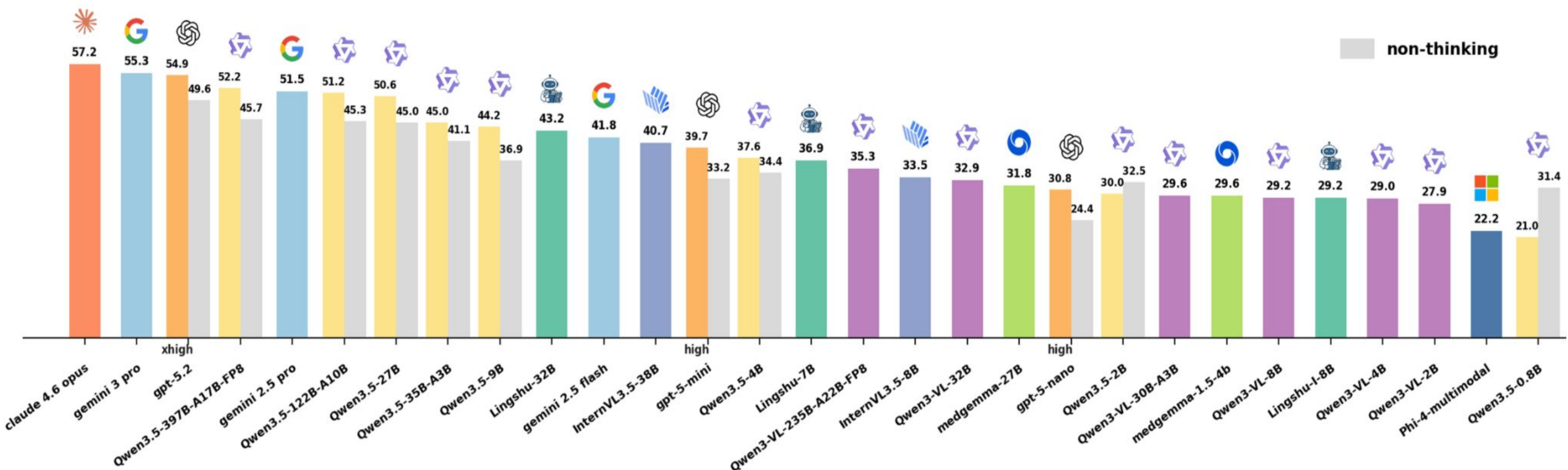
- Situs inversus totalis is often discovered incidentally.
- Awareness of such anatomical variants is essential for accurate diagnosis and treatment planning.
- Imaging plays a pivotal role in confirming the diagnosis and excluding associated anomalies.
- Patients should be informed about their condition, especially in preparation for emergencies or interventions.

Step 1 – Per-image Findings: concise expert radiological statements per image.

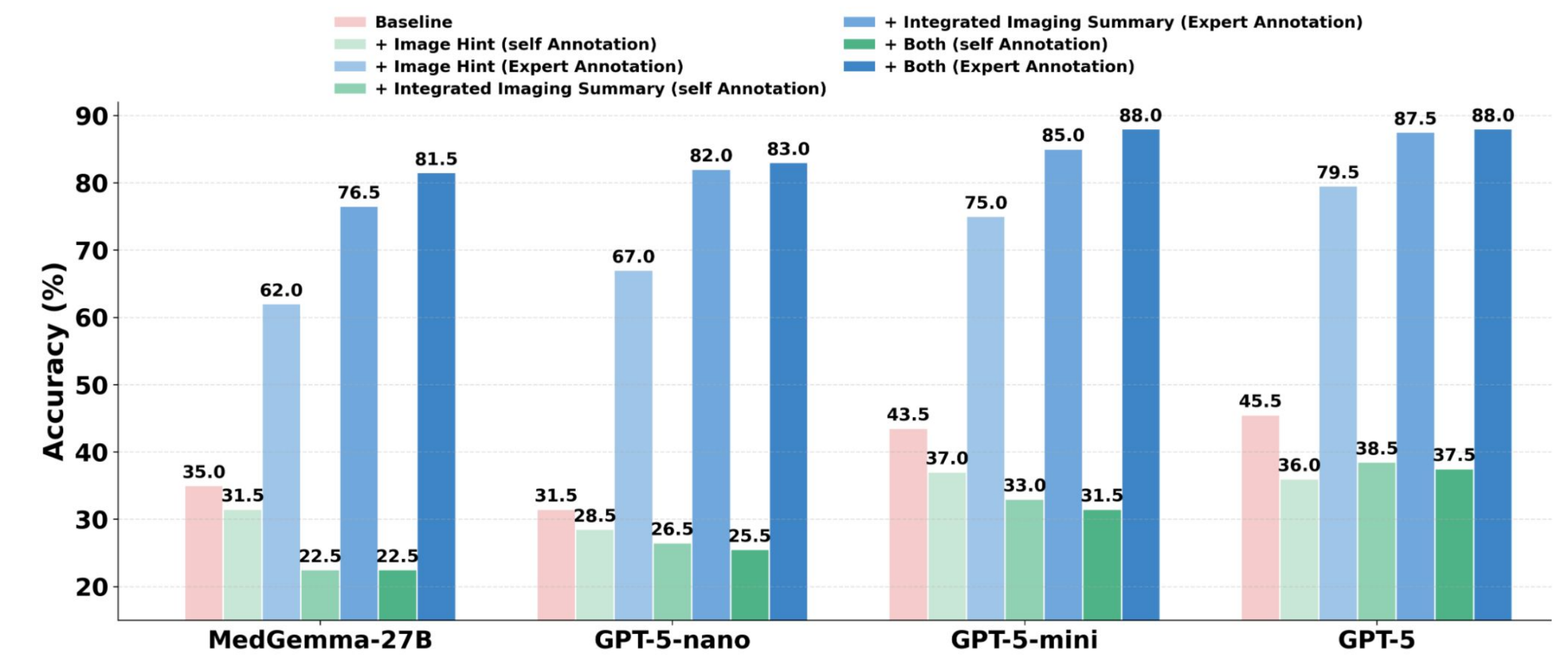
Step 2 – Integrated Imaging Summary: cross-view fusion into a case-level summary.

Step 3 – Ddx reasoning: option-wise elimination and final diagnosis, plus a case discussion with 5 sections (Background, Clinical perspective, Imaging perspective, Clinical significance/Outcome, Take-home messages)

Baseline performance



Expert imaging summaries unlock language ability



Providing the expert integrated imaging summary boosts accuracy by **+41–50** points (up to 2.6× relative gains), while adding only caption-like hints gives smaller benefits. When models first generate their **own** summaries and then condition on them, gains are negative, often missing laterality or key findings.

