MedThinkVQA
Medical Thinking with Multiple Images

Zonghai Yao[1,2]*, Benlu Wang[3]*, Yifan Zhang[1,4], …, Arman Cohan[3], Hong Yu[1,2,4]
[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

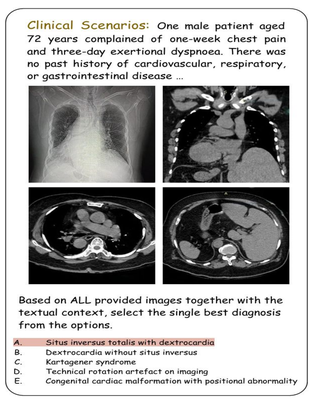
Real clinical diagnosis ≠ single-image QA

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	✗	✗	0.45	✗	✗	✗	✗
VQA-Med Ben Abacha et al.	500	✗	✗	1.00	✗	✗	✗	✗
Path-VQA He et al.	6,719	✗	✗	0.13	✗	✗	✗	✗
SLAKE-En Liu et al.	1,061	✗	✗	0.09	✗	✗	✗	✗
PMC-VQA Zhang et al.	33,430	✗	✗	0.87	✗	✗	✗	✗
OmniMedVQA Hu et al.	127,995	✗	✗	0.92	✗	✗	✗	✗
GMAI-MMBench Chen et al.	21,281	✗	✗	1.00	✗	✗	✗	✗
GEMeX Liu et al.	1,605,575	✗	✗	1.00	✗	✗	✓	✓
Medical-Diff-VQA Hu et al.	700,703	✗	✗	1.23	✗	✓	✗	✗
MedFrameQA Yu et al.	2,851	✗	✗	3.24	✗	✗	✓	✗
ICG-CXR Ma et al.	11,439	✗	✗	2.00	✗	✓	✗	✗
MedRAX[1] Fallahpour et al.	2,500	✗	✗	1.85	✗	✗	✗	✗
GEMeX-ThinkVG Liu et al.	206,071	✗	✗	1.00	✗	✗	✓	✓
MMMU (H & M) Yue et al.	1,752	✓	✗	1.14	✗	✗	✗	✗
MMMU-Pro (H & M) Yue et al.	346	✓	✗	1.25	✗	✗	✗	✗
S-Chain Le-Duc et al.	12,000	✓	✗	1.00	✗	✗	✓	✓
MedXpertQA MM Zuo et al.	2,000	✓	✓	1.43	✗	✗	✗	✗
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

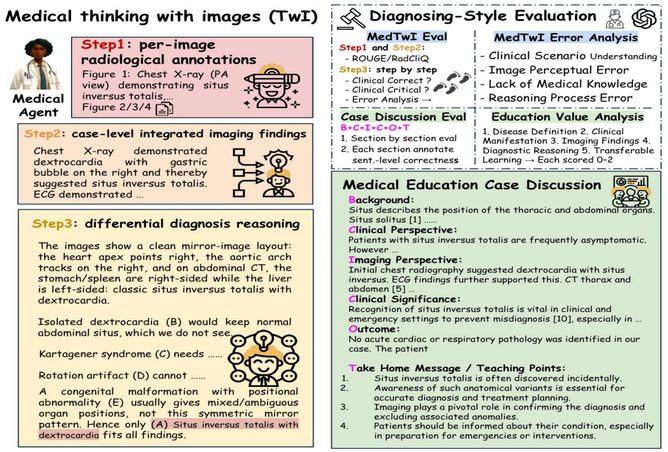


MedThinkVQA contains **9** aggregated imaging modalities. About **29.0%** of cases include **longitudinal follow-up studies**, and cases span **19 of 22** ICD-10 chapters.

# 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

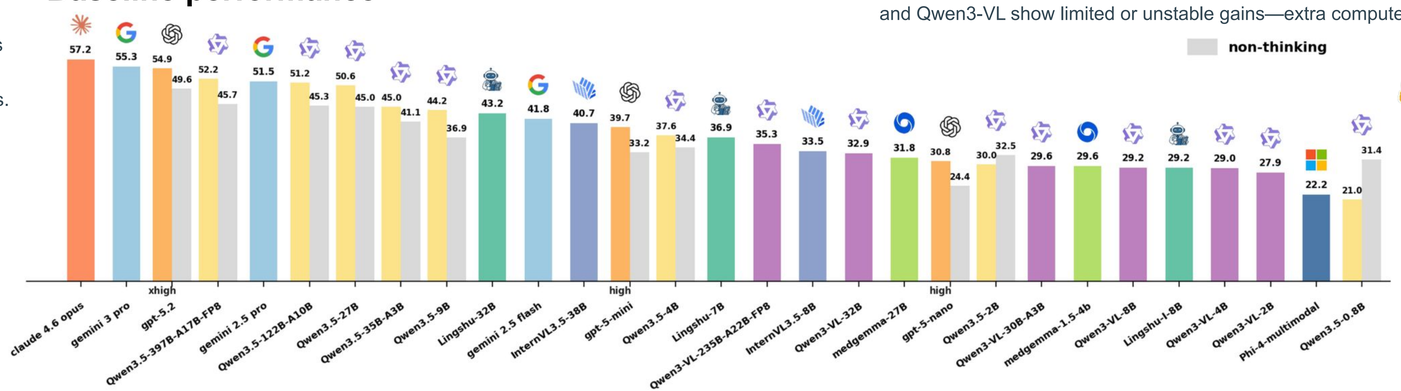
Split	# Cases	# Longitudinal	Share (%)
Train	9,347	2,703	28.9
Test	720	219	30.4
Overall	10,067	2,922	29.0

Think-with-Images: 3 supervised steps

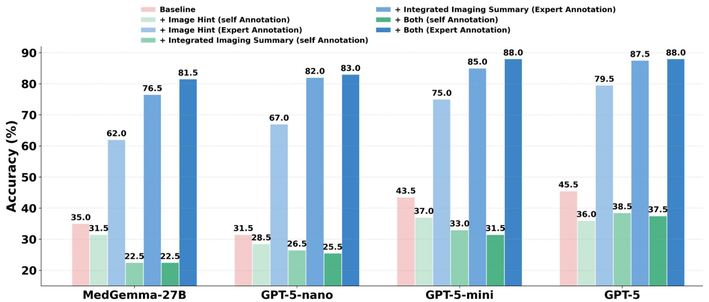


Step 1 – Per-image Findings: concise **expert** radiological statements per image.
Step 2 – Integrated Imaging Summary: **cross-view** fusion into a case-level summary.
Step 3 – DDx reasoning: option-wise elimination and final diagnosis, plus a case discussion with 5 sections (Background, Clinical perspective, Imaging perspective, Clinical significance/Outcome, Take-home messages)

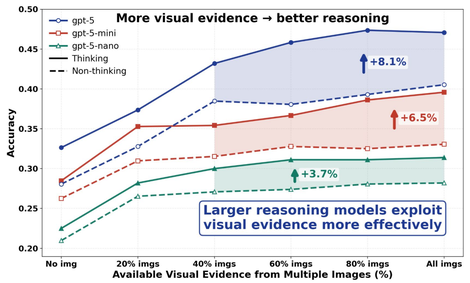
Baseline performance

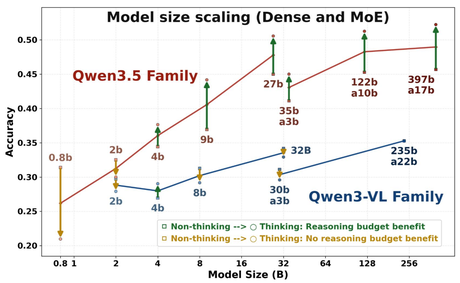


Expert imaging summaries unlock language ability



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.





More images → better performance, especially for larger reasoning models.
Larger models benefit more from extra reasoning compute. Qwen3.5 scales well (>2B), while smaller models and Qwen3-VL show limited or unstable gains—extra compute helps **only with strong visual grounding.**

Hugging Face