

How Good are Foundation Models in Longitudinal MRI Disease Progression Reasoning?

Wafa Al Ghallabi¹, Ritesh Thawkar¹, Sara Ghaboura¹, Omkar Thawakar¹, Numan Saeed¹, Dana Al Nuaimi², Ajnas Alkatheeri³, Salman Khan¹, and Fahad Shahbaz Khan^{1,4}

¹ Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
wafa.alghallabi@mbzuai.ac.ae

² Department of Health Abu Dhabi, Abu Dhabi, UAE

³ Fatima College of Health Sciences, Abu Dhabi, UAE

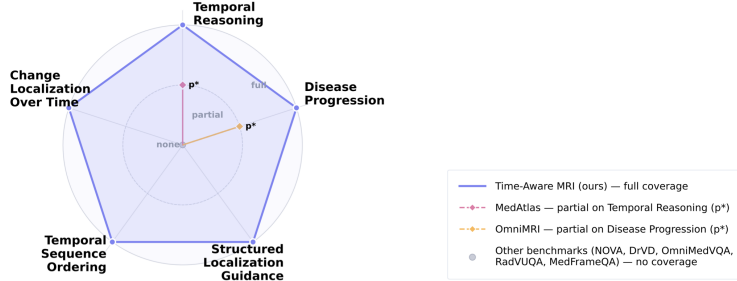
⁴ Linköping University, Linköping, Sweden

Abstract. Magnetic Resonance Imaging (MRI) interpretation is fundamental to clinical decision-making, requiring radiologists to integrate multi-view anatomical planes across sequential timepoints while precisely localizing interval changes. However, existing vision-language benchmarks remain confined to single-timepoint, single-view interpretation, failing to capture the temporal-spatial reasoning essential to radiologic practice. We introduce the *Time-Aware Multi-View MRI Benchmark*, an evaluation framework unifying multi-view anatomical input, temporal reasoning across longitudinal scans, and structured localization guidance. The benchmark comprises 3,920 expert-verified question-answer pairs derived from 890 patients across over 3,200 longitudinal MRI timepoints, drawn from seven clinical cohorts covering glioblastoma, neurodegeneration, vestibular schwannoma, and brain metastases, in open-ended, multiple-choice, and binary formats, requiring models to identify anatomical regions of maximal change, characterize progression across sequences and views, and provide structured guidance specifying boundaries, imaging features, and confounders. Experiments across 17 vision-language models reveal moderate temporal alignment but systematic failure on change direction recognition and volumetric quantification, while multi-view inputs improve spatial localization yet degrade temporal reasoning in compact architectures. Our benchmark establishes a systematic standard for evaluating progression tracking, interval change localization, and temporal ordering essential for clinical deployment. Code, evaluation splits, and dataset are available at <https://github.com/wafaAlghallabi/Time-Aware-MRI>

1 Introduction

Magnetic Resonance Imaging (MRI) serves as the cornerstone of longitudinal disease assessment, enabling clinicians to evaluate treatment response, detect recur-

Fig. 1: Comparison of reasoning dimension coverage across MRI benchmarks. Time-Aware Multi-View MRI benchmark achieves full coverage across all five tasks. MedAtlas [28] and OmniMRI [12] offer partial coverage (p^*) on Temporal Reasoning and Disease Progression, respectively. Remaining benchmarks, NOVA [5], DrVD-Bench [34], OmniMedVQA [14], RadVUQA [19], and MedFrameQA [31] provide no coverage across any dimension.



rence, and characterize the temporal trajectory of pathology. Clinical interpretation of follow-up MRI is inherently comparative and multi-dimensional. Radiologists systematically align current and prior studies across multiple anatomical planes (axial, coronal, sagittal), assess interval changes in lesion morphology and signal characteristics, and synthesize these observations into conclusions about disease progression [1,16]. This integrative reasoning, which combines spatial understanding across anatomical views, temporal comparison between timepoints, and precise localization of change, represents the cognitive foundation of radiologic decision-making. Despite this clinical reality, existing vision-language benchmarks evaluate models almost exclusively on static, single-timepoint interpretation.

Recent multimodal benchmarks have advanced medical image understanding but remain fundamentally limited to single-image analysis. Large-scale efforts such as OmniMedVQA [14], GMAI-MMBench [8], and PMC-VQA [33] aggregate extensive collections but process each image independently. Limited efforts have begun addressing temporal reasoning [13,29], but explicitly acknowledge that longitudinal tracking and multi-view anatomical understanding remain unexplored. Consequently, despite impressive accuracy on static medical question-answering [20], vision-language models (VLMs) perform poorly on temporal tasks [32], exposing fundamental training gaps. Furthermore, existing benchmarks lack structured localization guidance, including anatomical boundaries, imaging features, and confounders, which is essential for connecting semantic reasoning to spatial precision.

The disconnect between benchmark design and clinical requirements is compounded by demonstrated VLMs’ failures on spatial-temporal reasoning. Recent work highlights systematic failures in spatial localization and relative positioning [25], with reports often lacking explicit anatomical references [26]. VLMs processing isolated slices without multi-view context show over ten percent per-

formance degradation [30]. Existing work incorporating temporal information [4] lacks multi-view anatomical analysis and segmentation capabilities, while work achieving multi-view reasoning [11] demonstrates no temporal capability. Recent approaches [3] combining viewing angles with temporal context operate exclusively on 2D radiography, unable to process axial-coronal-sagittal planes essential to cross-sectional imaging.

To address these gaps, we introduce temporal localization reasoning tasks that require models to identify anatomical regions of maximal interval change across viewing planes and provide structured guidance specifying boundaries, imaging features, and confounders to exclude. These tasks extend visual question answering beyond categorical recognition to integrated spatial-temporal localization, testing the radiological workflow from change detection to actionable delineation essential for treatment planning and longitudinal monitoring. The benchmark targets five tasks: Temporal Reasoning, Disease Progression, Structured Localization Guidance, Temporal Sequence Ordering, and Change Localization Over Time. **Our contributions are three-fold:**

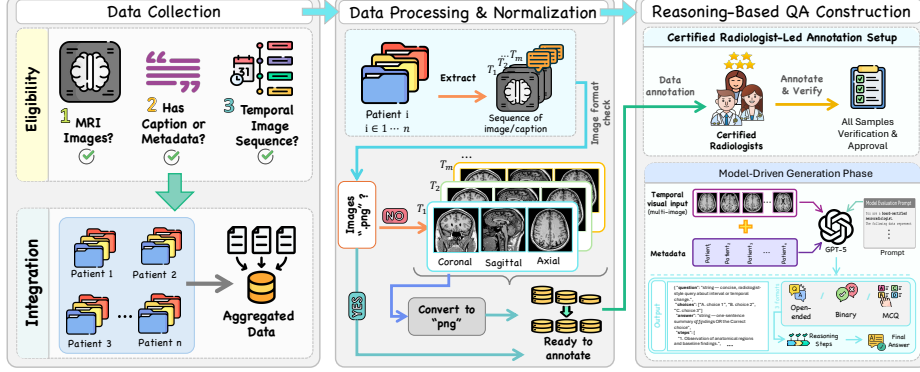
- **Novel Benchmark:** We present the Time-Aware Multi-View MRI Benchmark, a multi-disease evaluation framework for temporal-spatial reasoning in longitudinal neuroimaging. It integrates seven public cohorts (Glioblastoma, Alzheimer’s, Vestibular Schwannoma, Brain Metastases) harmonized under a unified temporal protocol.
- **Structured Clinical Tasks:** We introduce tasks providing temporally aligned multi-sequence MRI across multi-view planes (axial, coronal, sagittal). Unlike prior categorical question answering (QA), our tasks require models to perform interval change localization with structured guidance (identifying boundaries, features, and confounders).
- **Comprehensive Evaluation:** We explicitly measure capabilities including progression tracking and temporal sequence ordering, providing systematic evaluation of five integrated dimensions absent from existing benchmarks (see Fig. 1).

2 Time-Aware Multi-View MRI Benchmark

2.1 Dataset Composition and Preprocessing

Our benchmark integrates seven longitudinal MRI cohorts selected for temporal depth and pathological diversity. The **Yale Brain Metastases Dataset** [7] provides serial post-treatment monitoring documenting response to surgery and radiotherapy. Three glioblastoma cohorts capture distinct disease phases: the **UCSF Post-Operative Glioma Dataset** [9] tracks early post-surgical evolution, the **UCSD-PTGBM Dataset** [10] includes rich molecular characterization (MGMT, IDH status) enabling correlation with imaging findings, and the **LUMIERE Dataset** [21] provides expert RANO-annotated longitudinal

Fig. 2: Time-Aware Multi-View MRI Benchmark pipeline: data selection, multi-view extraction, expert-guided question generation, and verification workflow.



scans across treatment cycles. The **OASIS-2** [18] and **ADNI** [15] cohorts enable assessment of neurodegenerative progression in aging and Alzheimer’s disease. The **Vestibular Schwannoma Multi-Center Dataset** [17] contributes benign tumor monitoring data with multi-year follow-up intervals. Across cohorts, available sequences include T1, T2, FLAIR, T1CE, DWI, and ADC, with 3-4 timepoints per patient and inter-scan intervals ranging from 4 months in post-treatment glioma cohorts to over 18 months in vestibular schwannoma follow-up.

Data Preprocessing and Quality Control. To ensure cross-cohort consistency and clinical-grade image quality, we implemented a multi-stage preprocessing pipeline. Serial scans were registered to baseline using ANTs [2] (deformable for most cohorts, rigid for geometrically consistent acquisitions). All volumes were reoriented to RAS+ neurological convention to ensure standardized anatomical presentation. For tumor cohorts, we employed tumor-aware slice extraction; segmentation masks identified tumor center of mass, and representative slices were extracted through the tumor centroid to guarantee lesion visibility in all views. N4 bias field correction [23] was applied where needed, followed by sequence-specific percentile-based intensity normalization. To preserve clinically meaningful hyper- and hypointense signal characteristics critical for pathology assessment (e.g., enhancing tumor margins on T1CE, FLAIR hyperintensities), normalization windows were calibrated per sequence: T1CE and post-contrast sequences used p_1 to $p_{99.5}$ to retain enhancement extremes, while T2 and FLAIR used p_2 to p_{98} with adaptive ceiling extension when bright lesion signal exceeded the upper bound. An automated quality control pipeline verified sufficient contrast, tumor visibility, and adequate tissue coverage. All registrations underwent dual expert review by two certified radiologists, with cases exhibiting motion artifacts, registration failures, or quality violations excluded. Temporal metadata including inter-scan intervals (Δt) and sequence identifiers were encoded to support time-aware reasoning.

2.2 Multi-View and Temporal Integration

Mainstream VLMs are constrained to 2D inputs [24], limiting their application to volumetric imaging despite emerging 3D architectures [27,6]. Radiologists, however, interpret MRI across orthogonal planes to triangulate findings [1]. To approximate this workflow while remaining compatible with deployed VLMs, we extract multi-view presentations (axial, coronal, sagittal) from registered volumes. Center slices across all available sequences (T1, T2, FLAIR, T1CE, DWI, ADC) yield 9–12 images per timepoint, and 3–4 serial timepoints per patient (months to years) enable longitudinal rather than single-scan evaluation.

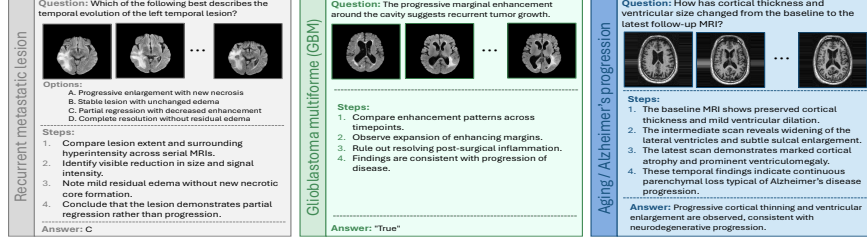
2.3 Question Generation and Expert Validation

We developed a hybrid annotation pipeline combining automated generation with radiologist oversight (Fig. 2). For each patient sequence, GPT-5 received a structured prompt encoding cohort identity, temporal metadata (Δt , sequence parameters), available sequences, and clinical annotations where available (RANO scores, MGMT/IDH status, tumor volumes). This metadata grounding ensures questions reference verifiable findings rather than free-form image descriptions. Candidates were produced in three formats: open-ended for Temporal Reasoning and Disease Progression, multiple-choice with clinically plausible distractors for Structured Localization Guidance and Change Localization Over Time, and binary for Temporal Sequence Ordering. Each candidate included the question, options, final answer, and intermediate reasoning steps.

Two board-certified radiologists independently verified diagnostic accuracy, temporal consistency, distractor plausibility, and clinical relevance. Dual-approved samples were retained; flagged or conflicting items were returned to the generation pipeline with correction notes (e.g., ambiguous temporal reference, implausible distractor) and re-reviewed, with persistent disagreements discarded. This process yielded a 72% acceptance rate and 3,920 expert-verified QA pairs.

Benchmark Statistics. The final benchmark comprises 3,920 QA pairs from 890 patients across over 3,200 timepoints (7.4 months mean follow-up), with disease distribution spanning glioblastoma (41%), brain metastases (31%), neurodegeneration (14%), and vestibular schwannoma (14%). Questions span five categories: (1) *Temporal Reasoning* (open-ended, 1,101 pairs): interval change identification, (2) *Disease Progression* (open-ended, 942): trajectory and treatment response, (3) *Structured Localization Guidance* (multiple-choice, 828): anatomical change regions with boundaries and imaging features, (4) *Temporal Sequence Ordering* (binary, 487): chronological reconstruction, and (5) *Change Localization Over Time* (multiple-choice, 562): maximal-change timepoints and locations. For localization tasks, structured guidance specifies boundaries, imaging features, and confounders to test full radiological workflow. Representative samples are shown in Fig. 3.

Fig. 3: Representative samples from the benchmark illustrating diverse pathologies and disease progression across serial timepoints (axial view).



3 Evaluation Framework

Evaluation Tasks: The *Time-Aware Multi-View MRI Benchmark* evaluates the model’s ability to reason about disease evolution across serial multi-sequence, multi-view MRI through five complementary tasks. *Temporal Reasoning* requires interpreting sequential scans across anatomical planes and imaging contrasts to answer clinically grounded open-ended questions about interval changes. *Disease Progression* assesses open-ended prediction of disease trajectory (growth, regression, stable). *Structured Localization Guidance* tests multiple-choice identification of anatomical regions exhibiting change, with distractors encoding structured descriptions of boundaries, imaging features, and confounders. *Temporal Sequence Ordering* evaluates binary chronological scan arrangement based on progression patterns. *Change Localization Over Time* requires multiple-choice identification of timepoints and spatial regions of maximal interval change. Together, these tasks capture complementary dimensions of longitudinal understanding, from spatial-temporal alignment and multi-sequence integration to biological interpretation.

Evaluation Metrics: To ensure fair assessment across diverse temporal reasoning tasks, we employ three complementary measures: *Final Answer Accuracy*, *Reasoning Score (RS)*, and the *Time-Aware Composite (TAC)* metric. The choice of three complementary metrics reflects the heterogeneous nature of the five tasks: closed-form tasks (multiple-choice, binary) admit direct accuracy evaluation, while open-ended tasks require reasoning fidelity in addition to final correctness. Final Answer Accuracy captures end-to-end correctness across all task formats. RS, following LlamaV-o1 [22], jointly considers reasoning consistency, temporal alignment, and factual accuracy, and is applied to tasks with explicit reasoning step traces (Temporal Reasoning, Disease Progression). TAC provides unified stepwise temporal reasoning assessment for tasks requiring multi-timepoint comparison:

$$\begin{aligned} \text{TAC} = & 0.5 \times \text{TEDS} + 0.2 \times \text{Trend-F1} \\ & + 0.2 \times \text{SignAcc} + 0.1 \times \text{Coverage} \end{aligned} \quad (1)$$

Table 1: Model performance on the *Time-Aware Multi-View MRI Benchmark* across final accuracy, Reasoning Score (RS), and Time-Aware Composite (TAC) metric. Higher TAC indicates stronger temporal consistency and reasoning fidelity. Bold: best per column.

Model	Final Acc (%)	RS	TAC	TEDS	Trend F1	Sign Acc	Coverage	Chronology
<i>Closed-source models</i>								
o4-mini	32.18	6.68	0.753	0.832	0.548	0.681	0.908	0.918
GPT-4o	32.00	6.26	0.731	0.807	0.546	0.654	0.877	0.917
GPT-5.2	21.20	5.83	0.661	0.805	0.192	0.639	0.921	0.856
Gemini-2.5-Flash	23.57	5.83	0.692	0.780	0.477	0.596	0.875	0.825
Gemini-2.5-Pro	23.66	5.88	0.672	0.785	0.504	0.528	0.730	0.957
Gemini-3-Flash	22.30	5.17	0.577	0.764	0.216	0.470	0.575	1.000
Gemini-3-Pro	35.10	5.31	0.590	0.775	0.235	0.485	0.600	0.980
<i>Open-source models</i>								
InternVL3.5-Inst	35.15	6.68	0.800	0.870	0.631	0.740	0.903	0.951
Qwen3-VL-Plus-Thinking	28.38	6.54	0.733	0.812	0.558	0.659	0.835	0.830
Qwen3-VL-235B-Thinking	30.37	6.55	0.742	0.815	0.571	0.674	0.852	0.825
Qwen3-VL-8B-Inst	24.31	6.05	0.732	0.801	0.557	0.655	0.888	0.825
Llama-4-Scout-17B-Inst	28.47	5.78	0.708	0.810	0.485	0.601	0.860	0.870
Llama-4-Maverick-17B-Inst	26.84	5.85	0.690	0.779	0.505	0.574	0.846	0.661
MedGemma-27B-IT	19.13	5.04	0.602	0.696	0.280	0.523	0.936	0.645
MedGemma-1.5-4B-IT	21.80	4.81	0.587	0.706	0.262	0.472	0.873	0.749
MedGemma-4B-IT	23.50	4.58	0.572	0.717	0.245	0.421	0.809	0.854

TAC integrates: *TEDS* (alignment between predicted and reference change sequences, penalizing skipped or hallucinated intervals), *Trend-F1* (progression vs. regression detection), *Sign Accuracy* (change direction correctness), and *Coverage* (interval reasoning completeness). Additionally, *Chronology* evaluates temporal ordering preservation in model narratives. Weights were assigned to prioritize sequence alignment (TEDS) as the primary temporal fidelity measure, with equal contribution from change characterization components (Trend-F1, SignAcc), and a smaller weight for coverage as a completeness auxiliary. These metrics assess not only whether models detect change, but how coherently they reason about timing, direction, and sequence. TAC ranges from 0 to 1, with higher scores indicating stronger temporal consistency.

4 Experiments and Discussion

Experimental Setup: We evaluate 17 vision-language models spanning closed-source systems (GPT-4o, o4-mini, GPT-5.2, Gemini-2.5/3-Pro/Flash), open-source foundation models (Qwen3-VL, InternVL3.5, Llama-4), and medical specialized VLMs (MedGemma). All models are evaluated zero-shot using multi-sequence, multi-view MRI inputs to assess out-of-the-box temporal reasoning capabilities. Closed-source models are accessed via APIs; open-source models are deployed locally. The public release of training and evaluation splits supports future supervised adaptation studies on this benchmark.

Main Results Overview: Table 1 reports performance across final accuracy, RS, and TAC. Overall, models achieve moderate TAC (0.57–0.80) and chronol-

ogy scores (0.64–1.00) but exhibit weaker Trend-F1 (0.19–0.63) and Sign Accuracy (0.42–0.74), indicating temporal alignment is more reliable than characterizing change direction. InternVL3.5-Inst leads on both TAC (0.80) and final accuracy (35.15%), narrowly ahead of Gemini-3-Pro (35.10%). Gemini-3-Flash show lower accuracy (22.3%) despite a perfect chronology score (1.00), suggesting potential instruction-following limitations or challenges with multi-sequence medical imaging. Specialized medical models (MedGemma variants) underperform general-purpose VLMs on TAC, indicating that domain specialization alone does not confer temporal reasoning ability. Collectively, results indicate that current VLMs handle temporal ordering reliably but consistently fail on change-type recognition, the clinically most critical capability.

Multi-View Configuration Analysis: To assess whether multi-view presentation compensates for the absence of volumetric processing in current VLMs, we conducted controlled ablations on the UCSF-GBM subset (1,192 samples) comparing multi-view (axial + coronal + sagittal) versus axial-only inputs. We employed an agentic Resident-Attending workflow: a Resident agent first extracts structured spatial findings (lesion location, signal changes, interval comparisons) from multi-timepoint MRI inputs, which an Attending agent then integrates for final diagnostic classification. Fixing the Resident prompt across view configurations isolates the effect of input modality from prompt sensitivity. We evaluated representative models: GPT-4o, Gemini-2.5-Pro/Flash (closed-source) and Qwen3-VL-8B-Inst, MedGemma-4B-IT, InternVL3.5-Inst (open-source).

Results: Table 2 reports accuracy on four clinical tasks. Multi-view inputs improve progression localization across models (average +6.2 pp) but produce mixed effects on temporal ordering: closed-source models benefit (GPT-4o: +7.2 pp), whereas smaller open-source VLMs show degradation (Qwen3-VL-8B: -8.0 pp, MedGemma-4B: -5.8 pp), suggesting information overload in compact architectures. Open-source models achieve highest global accuracy (InternVL3.5: 44.2%, Qwen3-VL-8B: 43.6%) versus closed-source models (GPT-4o: 36.7%). Performance peaks on progression localization (97.3%) and temporal ordering (72.0%) but remains below 16% on change segmentation and quantification, indicating that protocol-dependent segmentation and volumetric estimation remain fundamental bottlenecks independent of spatial localization success.

Implications: Multi-view presentation improves spatial localization, demonstrating that orthogonal 2D planes provide complementary spatial context approaching clinical radiological workflows. However, view redundancy degrades temporal reasoning in smaller models, suggesting that adaptive view selection or attention mechanisms are needed to leverage multi-view inputs efficiently. The dichotomy between strong localization (97%) and failed volumetric quantification (<16%) reveals that current VLMs lack the geometric reasoning required for protocol-adherent RANO measurements despite accurate spatial identification. Future architectures should integrate explicit 3D geometric priors, cross-timepoint attention, and protocol-conditioned training to bridge this gap.

Table 2: Agentic workflow accuracy (%) on UCSF-GBM subset across four clinical tasks. Multi-view inputs improve spatial localization but reveal persistent limitations in volumetric reasoning. Bold: best per column.

Model	Global	Ch. Seg.	Ch. Quant.	Temp. Ord.	Prog. Loc.
Open-source models					
InternVL3.5-Inst	44.2	1.1	13.5	72.0	97.3
Qwen3-VL-8B-Inst	43.6	0.0	12.7	71.4	96.9
MedGemma-4B-IT	42.1	0.6	11.7	65.9	96.2
Closed-source models					
GPT-4o	36.7	0.9	12.7	42.8	95.1
Gemini-2.5-Flash	27.8	4.3	9.2	37.7	63.3
Gemini-2.5-Pro	27.7	15.7	15.5	44.2	37.4

5 Conclusion

We introduced the Time-Aware Multi-View MRI Benchmark for evaluating temporal reasoning in longitudinal MRI interpretation. Across 17 vision-language models, current systems achieve moderate temporal alignment but struggle with change characterization and volumetric reasoning, while multi-view inputs improve spatial localization at the cost of temporal coherence in compact architectures. Future work should prioritize native 3D architectures with explicit geometric priors and cross-timepoint attention to support longitudinal clinical decision-making.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. American College of Radiology: ACR practice parameter for communication of diagnostic imaging findings (2024)
2. Avants, B.B., Tustison, N.J., Song, G., Cook, P.A., Klein, A., Gee, J.C.: A reproducible evaluation of ANTs similarity metric performance in brain image registration. NeuroImage (2011)
3. Bannur, S., Bouzid, K., Castro, D.C., Schwaighofer, A., Thieme, A., et al.: MAIRA-2: Grounded radiology report generation. arXiv preprint arXiv:2406.04449 (2024)
4. Bannur, S., Hyland, S., Liu, Q., Pérez-García, F., et al.: Learning to exploit temporal structure for biomedical vision-language processing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
5. Bercea, C.I., Li, J., Raffler, P., et al.: NOVA: A benchmark for anomaly localization and clinical reasoning in brain MRI. arXiv preprint arXiv:2505.14064 (2025)
6. Blankemeier, L., Cohen, J.P., Kumar, A., Wentland, D., Moor, M., Reis, E., Rubin, N., Bluethgen, C., et al.: Merlin: A vision language foundation model for 3D computed tomography. arXiv preprint arXiv:2406.06512 (2024)

7. Chadha, S., Weiss, D., Janas, A., Ramakrishnan, D., et al.: Yale-brain-mets-longitudinal: Yale longitudinal dataset of brain metastases on mri with associated clinical data (2025)
8. Chen, P., Ye, J., Wang, G., Li, Y., et al.: GMAI-MMBench: A comprehensive multimodal evaluation benchmark towards general medical AI. arXiv preprint arXiv:2408.03361 (2024)
9. Fields, B.K.K., Calabrese, E., Mongan, J., Cha, S., et al.: The university of california san francisco adult longitudinal post-treatment diffuse glioma MRI dataset. *Radiology: Artificial Intelligence* (2024)
10. Gagnon, L., Gupta, D., Nguyen, U., Correia de Verdier, M., Saluja, R., et al.: The university of california san diego post-treatment glioblastoma (UCSD-PTGBM) annotated multimodal MRI dataset. *Scientific Data* (2026)
11. Ghouse, H., Behzad, M.: MOSAIC: A multi-view 2.5D organ slice selector with cross-attentional reasoning for anatomically-aware CT localization in medical organ segmentation. *Computer Vision and Image Understanding* (2025)
12. He, X., Rofena, A., Feng, R., et al.: OmniMRI: A unified vision-language foundation model for generalist MRI interpretation. arXiv preprint arXiv:2508.17524 (2025)
13. Hu, X., Gu, L., An, Q., Zhang, M., Liu, L., et al.: Medical-Diff-VQA: A large-scale medical dataset for difference visual question answering on chest x-ray images. *PhysioNet* (2025)
14. Hu, Y., Li, T., Lu, Q., Shao, W., He, J., Qiao, Y., Luo, P.: OmniMedVQA: A new large-scale comprehensive evaluation benchmark for medical LVLM. arXiv preprint arXiv:2402.09181 (2024)
15. Jack Jr., C.R., Bernstein, M.A., Fox, N.C., Thompson, P.M., et al.: The Alzheimer’s disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging* (2008)
16. Khosa, F., Krinsky, G., Macari, M., et al.: Managing incidental findings on abdominal and pelvic CT and MRI, part 2: White paper of the ACR incidental findings committee II on vascular findings. *Journal of the American College of Radiology* (2013)
17. Kujawa, A., Dorent, R., Wijethilake, N., et al.: Segmentation of vestibular schwannoma from magnetic resonance imaging: An annotated multi-center routine clinical dataset (Vestibular-Schwannoma-MC-RC) (2023)
18. Marcus, D.S., Fotenos, A.F., Csernansky, J.G., Morris, J.C., Buckner, R.L.: Open access series of imaging studies (OASIS): Longitudinal MRI data in nondemented and demented older adults. *Journal of Cognitive Neuroscience* (2010)
19. Nan, Y., Zhou, H., Xing, X., et al.: Beyond the hype: A dispassionate look at vision-language models in medical scenario. arXiv preprint arXiv:2408.08704 (2025)
20. Saab, K., Tu, T., Weng, W.H., Tanno, R., Stutz, D., Wulczyn, E., et al.: Capabilities of Gemini models in medicine. arXiv preprint arXiv:2404.18416 (2024)
21. Suter, Y., Knecht, U., Valenzuela, W., Notter, M., et al.: The LUMIERE dataset: Longitudinal glioblastoma MRI with expert RANO evaluation. *Scientific Data* (2022)
22. Thawakar, O., Dissanayake, D., More, K., Thawkar, R., et al.: LlamaV-o1: Rethinking step-by-step visual reasoning in LLMs. arXiv preprint arXiv:2501.06186 (2025)
23. Tustison, N.J., Avants, B.B., Cook, P.A., et al.: N4ITK: Improved N3 bias correction. *IEEE Transactions on Medical Imaging* (2010)
24. Wang, Y., Peng, J., Dai, Y., Jones, C., et al.: Enhancing vision-language models for medical imaging: Bridging the 3D gap with innovative slice selection. In: *Advances in Neural Information Processing Systems*. Curran Associates, Inc. (2024)

25. Wolf, D., Hillenhausen, H., Taskin, B., Bäuerle, A., Beer, M., et al.: Your other left! vision-language models fail to identify relative positions in medical images. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. Springer (2025)
26. Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Towards generalist foundation model for radiology by leveraging web-scale 2D&3D medical data. *Nature Communications* (2025)
27. Xin, Y., Ates, G.C., Gong, K., Shao, W.: Med3DVLM: An efficient vision-language model for 3D medical image analysis. *IEEE Journal of Biomedical and Health Informatics* (2025)
28. Xu, R., Huang, Z., Wei, Y., et al.: MedAtlas: Evaluating LLMs for multi-round, multi-task medical reasoning across diverse imaging modalities and clinical text. *arXiv preprint arXiv:2508.10947* (2025)
29. Yang, X., Miao, J., Yuan, Y., Wang, J., Dou, Q., Li, J., Heng, P.A.: Med-MIM: Medical large vision language models with multi-image visual ability. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer (2025)
30. Yang, Y., Wang, Z.Y., Liu, Q., Sun, S., Wang, K., Chellappa, R., Zhou, Z., et al.: Medical world model: Generative simulation of tumor evolution for treatment planning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2025)
31. Yu, S., Wang, H., Wu, J., Xie, C., Zhou, Y.: MedFrameQA: A multi-image medical VQA benchmark for clinical reasoning. *arXiv preprint arXiv:2505.16964* (2025)
32. Zhang, J., Gu, J.C., Hu, W., Zhou, Y., Piramuthu, R., Peng, N.: TemMed-Bench: Evaluating temporal medical image reasoning in vision-language models. *arXiv preprint arXiv:2509.25143* (2025)
33. Zhang, X., Wu, C., Zhao, Z., Lin, W., Zhang, Y., et al.: PMC-VQA: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415* (2023)
34. Zhou, T., Xu, Y., Zhu, Y., et al.: DrVD-Bench: Do vision-language models reason like human doctors in medical image diagnosis? *arXiv preprint arXiv:2505.24173* (2025)