Lost in Volume: The CT-SpatialVQA Benchmark for Evaluating Semantic-Spatial Understanding of 3D Medical Vision–Language Models

Mashrafi Monon[1](✉), Umaima Rahman[1,2], Asif Hanif[1], Numan Saeed[1], and Mohammad Yaqub[1]

[1] Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
[2] New York University Abu Dhabi, Abu Dhabi, UAE
mashrafi.monon@mbzuai.ac.ae

Abstract. Recent advances in 3D medical vision–language models have enabled joint reasoning over volumetric images and text, showing strong performance in medical visual question-answering (VQA) and report generation. Despite this progress, it remains unclear whether these models learn spatially grounded anatomy from 3D volumes or rely primarily on learned priors and language correlations. This uncertainty stems from the lack of systematic evaluation of semantic–spatial reasoning in volumetric medical VLMs for clinically reliable decision support. To address this gap, we introduce CT-SpatialVQA, a benchmark designed to evaluate semantic–spatial reasoning in 3D CT data. The benchmark comprises 9077 clinically grounded question-answer (QA) pairs derived directly from 1601 radiology reports and CT volumes, which are validated via a robust LLM-assisted pipeline with a 95% human consensus agreement rate. Our dataset, requires explicit anatomical localization, laterality awareness, structural comparison, and 3D inter-structure relational reasoning. We also introduce standardized evaluation protocol and benchmark eight 3D medical VLMs, finding severe degradation on semantic-spatial reasoning task, averaging 34% accuracy and often below random, highlighting the need for deeper integration of volumetric evidence for trustworthy clinical use. The code is available at: https://github.com/BioMedIA-MBZUAI/CT-SpatialVQA.

Keywords: 3D Vision-Language Models (VLMs) · Semantic-Spatial Reasoning · Visual Question Answering · CT · Large Language Model (LLM)

1 Introduction

On a typical day in radiology, a clinician reviews a chest CT to stage a suspected lung cancer. The scan contains hundreds of axial slices, each showing only a small part of the anatomy. As the radiologist scrolls through the volume, a small lesion appears and disappears near the mediastinum. To understand it, the clinician must mentally reconstruct the 3D structure: Is the lesion in front of or behind the great vessels? Does it extend to the pleura, or is it confined within the lung

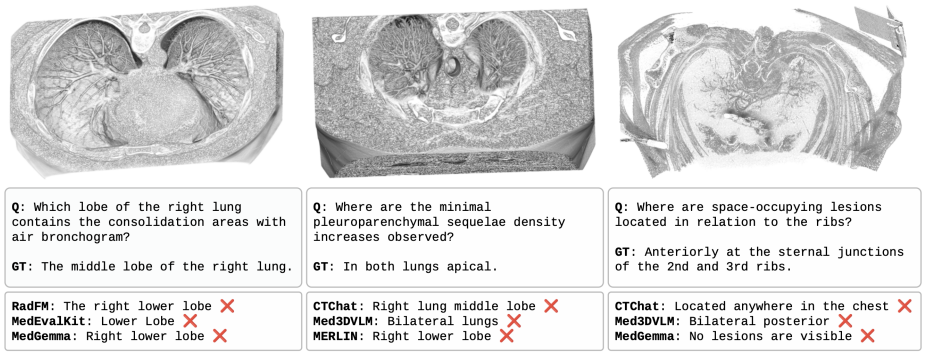


Fig. 1. Current 3D medical VLMs struggle to answer spatially grounded questions, highlighting limitations in spatial reasoning and raising concerns about their reliability for clinically meaningful use.

tissue? Answering these questions requires integrating information across many slices and forming a coherent 3D picture of the anatomy. These spatial judgments are not ancillary; instead, they directly govern diagnosis and procedural planning. Given this fundamental role of spatial reasoning in volumetric imaging, an important question arises: *Can vision-language models (VLMs) develop similar capabilities for understanding 3D spatial representations?*

Building on the success of 2D medical vision-language models, recent works have extended it to volumetric medical data and, in some cases, explicitly constructed 3D models designed to encode cross-slice context and spatial structure. This has led to the development of 3D medical VLMs, such as Med3DVLM [24], MERLIN [4], M3D [2], RadFM [23], CT-CLIP, and CT-CHAT [8]. While these developments raise expectations that 3D medical VLMs may support spatially grounded reasoning, we directly evaluate whether current models fulfill this promise. Through our study, we find that although 3D medical VLMs generate fluent descriptions and correctly answer many semantic or measurement-based questions, their performance deteriorates when tasks demand precise and consistent spatially grounded reasoning, as evident from Figure 1. This gap underscores a critical distinction: linguistic fluency does not imply true spatial understanding. Benchmarks such as DeepTumorVQA [5] primarily emphasize descriptive or measurement-focused queries and provide limited assessment of relational semantic-spatial reasoning. Further analyses of models including M3D [2] and RadFM [23] demonstrate performance degradation on queries requiring fine-grained 3D structural understanding or stable object–object relationships.

Most VLMs are pretrained on large-scale image text corpora derived from reports, where spatial relationships are weakly specified and often linguistically ambiguous. Recognizing these limitations, recent research in vision–language domain has begun to enhance spatial reasoning. Within the medical domain

[10,18,9,17], however, explicit and systematic investigation of spatial reasoning remains comparatively limited. For instance, recent work in 2D medical imaging [22] investigated spatial understanding, showing that models struggle with even simple positional questions such as "Is the left kidney below the stomach?" This issue is only compounded in 3D volumetric imaging, where spatial relationships must be inferred across depth and preserved globally, rather than within a single projection. Even when 3D backbones are used, the lack of explicit mechanisms for enforcing cross-slice consistency limits the models' ability to form a coherent representation of 3D space. While recent efforts like VILA-M3 [14], FILA [20], E3D-GPT [11], MS-VLM [12], and BrainMD [21] incorporate expert models or alternative supervision to strengthen anatomical discrimination, they still lack explicit geometric inductive biases and report limitations in precise spatial awareness. Spatial reasoning failures may be amplified in 3D settings, highlighting the need for systematic evaluation in volumetric medical imaging.

Through **CT-SpatialVQA**, a benchmark designed to evaluate semantic–spatial grounding in 3D medical VLMs, we identify a clear mismatch between current training objectives and the demands of clinical spatial reasoning. Thus, suggesting current training objectives and evaluations predominantly reward semantic fluency, while fine-grained spatial relationships, cross-slice consistency, and volumetric structure remain weakly enforced.

Contributions

1. **CT-SpatialVQA.** We introduce CT-SpatialVQA, a benchmark for systematically evaluating semantic-spatial reasoning in 3D medical VLMs using volumetric CT data and clinically grounded QA pairs.
2. **Spatially-Grounded QA Design.** We generate clinically grounded QA pairs from radiology reports through an LLM-based pipeline comprising generation, validation, and human auditing. The resulting questions require explicit anatomical localization, laterality awareness, structural comparison, and 3D inter-structure reasoning.
3. **Comprehensive Evaluation.** We evaluate eight 3D medical VLMs on CT-SpatialVQA, quantifying their semantic–spatial reasoning capabilities and identifying failure modes that reveal over-reliance on prior knowledge instead of volumetric visual grounding. We publicly release the CT-SpatialVQA dataset and evaluation code on github.

2 CT-SpatialVQA Benchmark

The CT-SpatialVQA benchmark is structured in two sequential phases: (1) spatially grounded QA pair generation from radiology reports, and (2) systematic evaluation of 3D medical VLMs using the generated QA pairs to assess semantic–spatial reasoning over volumetric CT data. Figure 2 provides an overview of

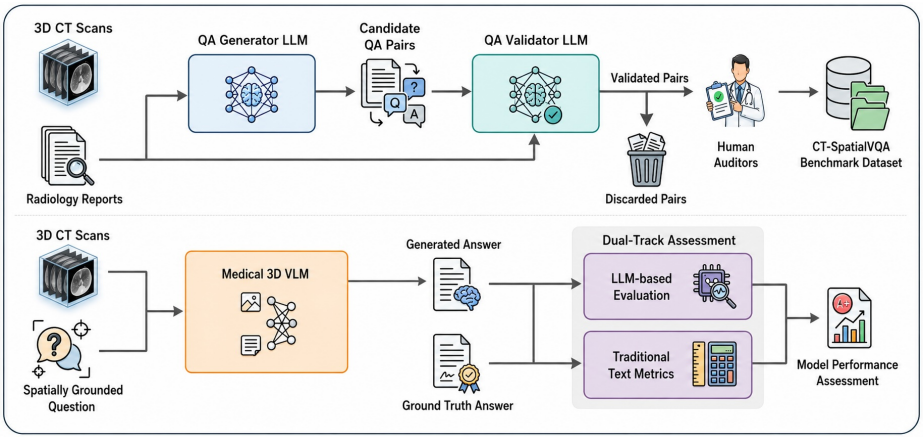


Fig. 2. An Overview of CT-SpatialVQA (*top*) The dataset generation pipeline uses paired radiology reports annotated by expert radiologists from 3D CT scans to prompt a "QA Generator LLM", synthesizing candidate spatially-grounded QA pairs. An independent "QA Validator LLM" filters the candidates, and a small subset of remaining pairs is then verified by human auditors to produce the final benchmark dataset. (*bottom*) The 3D medical VLM evaluation protocol tests a given model by inputting a 3D CT scan and a spatially-grounded question. The model's generated answer is compared to the ground truth using a dual-track assessment system, where LLM-based evaluation and traditional text metrics are conducted independently.

the dataset generation and evaluation phases.

Radiology Text Reports. The CT-SpatialVQA benchmark is constructed from the CT-RATE [8] dataset, which consists of volumetric (3D) CT scans paired with radiology reports. These reports are narrative text documents written by board-certified radiologists after reviewing CT volumes. Radiology reports encode rich spatial information describing anatomical structures and pathological findings, including their location, laterality, spatial extent, and relationships to surrounding structures. Spatial descriptions may be explicit (e.g., left lower lobe, bilateral basal atelectasis) or implicit (e.g., apical opacity, central lesion).

Spatially-Grounded Question–Answer Generation. To construct spatially-grounded QA pairs, each radiology report is provided to an LLM with a carefully designed prompt. The LLM is instructed to identify spatially grounded observations, generate questions that explicitly target semantic-spatial attributes, and provide concise answers derived directly from the report text, while avoiding diagnostic or severity-related content. QA generation focuses exclusively on six spatial dimensions: laterality (left/right/bilateral), longitudinal position (superior/inferior), anterior–posterior relations (front/back orientation), medial–lateral orientation (central/peripheral), adjacency or containment (within/adjacent to structures), and spatial extent or boundaries (confined to or extending across

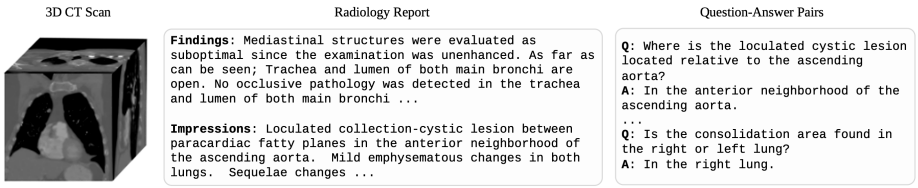


Fig. 3. An illustrative example of a 3D CT scan, its corresponding full radiology report, and representative QA pairs.

regions). For each identified spatial observation, the LLM generates one or more QA pairs as shown in Figure 3, each targeting a single spatial concept with a short, unambiguous answer grounded in the report. The resulting candidate QA pairs are subsequently passed to a validation stage to ensure correctness and reliability.

Instruction Prompt for the QA Generator LLM

```
You are a medical AI assistant specialized in radiology and 3D spatial reasoning. Read
the 3D CT scan report below and generate 7-10 question-answer (QA) pairs that test a
vision-language model's understanding of spatial and anatomical relationships explicitly
described in the report.

Focus only on spatial facts such as:
- Laterality (left/right/bilateral)
- Longitudinal position (superior/inferior)
- Anterior-posterior relations (front/back orientation)
- Medial-lateral orientation (central/peripheral)
- Adjacency or containment (within/adjacent to structures)
- Spatial extent or boundaries (regional confinement or extension)

Guidelines:
- Use only information from the Findings and Impressions sections
- Do not include diagnostic, interpretive, or normality statements
- Questions must emphasize where, which side, above/below, or extent
- Answers must be strictly factual and directly derived from report

### Radiology Report
{Findings}{Impressions}
```

Spatially-Grounded Question–Answer Validation. To ensure correctness and proper grounding, each candidate QA pair is evaluated together with its corresponding full radiology report by a second, independent LLM acting as a validator. Providing both the generated QA pair and the original report enables the validator to explicitly verify that the answer is directly supported by the report text and that no information is introduced beyond what is stated. The validator further ensures that the question strictly concerns semantic-spatial attributes (e.g., anatomical location, laterality, relative position) and that the answer accurately reflects the spatial relationships described in the report. It also checks for hallucinations, unsupported inferences, ambiguity, and verbosity, ensuring concise and unambiguous responses. QA pairs that fail any of these criteria are discarded. This two-stage LLM pipeline, combining generation with

independent grounding and spatial-consistency validation, reduces hallucinations and improves spatial faithfulness.

Instruction Prompt for the QA Validator LLM

```
You are a radiology QA validator. You receive radiology report (Findings & Impressions)
and QA pairs created from them. Label each QA pair with two booleans.

- is_spatial: true only if answering the question requires spatial reasoning about
anatomical location, orientation, relative position, or laterality visible in the image
(e.g., which lung, which lobe, proximity, direction, comparison of organ sizes, location
of effusion). Pure presence/absence or textual metadata questions are false.
- is_relevant: true only if the question could be answered from the imaging-derived
Findings/Impressions (not from general knowledge or the wording of the text itself).
Questions that just ask which words appear in the report, cite textual phrases, or
hallucinate unseen details are false.

### Radiology Report
{Findings}{Impressions}

### QA Pairs
{question_answer_pairs}

For each QA pair, return {'is_spatial':, 'is_relevant':} as JSON only.
```

Spatially-Grounded Question–Answer Validation (Human Audit). To verify the LLM-based validator, we conduct a human audit on a random subset (12%) of validated QA pairs to assess spatial grounding, correctness, and consistency with the source reports. Trained reviewers with relevant backgrounds examine each QA pair together with the full radiology report for textual support, spatial specificity, correct interpretation, and clinical clarity. We observe 95% agreement with no systematic discrepancies, indicating that the validator reliably enforces semantic-spatial grounding and aligns closely with human judgment, supporting the robustness of our CT-SpatialVQA benchmark.

3D Medical VLM Evaluation. The finalized QA pairs are used to evaluate 3D medical VLMs. For each sample, a model is provided with a volumetric CT scan and a spatially-grounded question, and it generates an answer based on the 3D image content. Importantly, the corresponding radiology report is not provided during evaluation, ensuring that models cannot rely on textual cues and must instead reason directly from the volumetric data. Performance is assessed through LLM-based evaluation, where external LLMs first act as independent judges and subsequently as a jury to produce binary correctness decisions, alongside traditional text-based similarity metrics.

3 Experiment and Results

3D Medical VLMs. To ensure fairness and reproducibility, we restrict our experiments to models with fully accessible implementations and reproducible inference pipelines. We evaluate a diverse set of state-of-the-art 3D medical VLMs, including CT-Chat [8], MERLIN [4], Med3DVLM [24], M3D [2], RadFM [23],

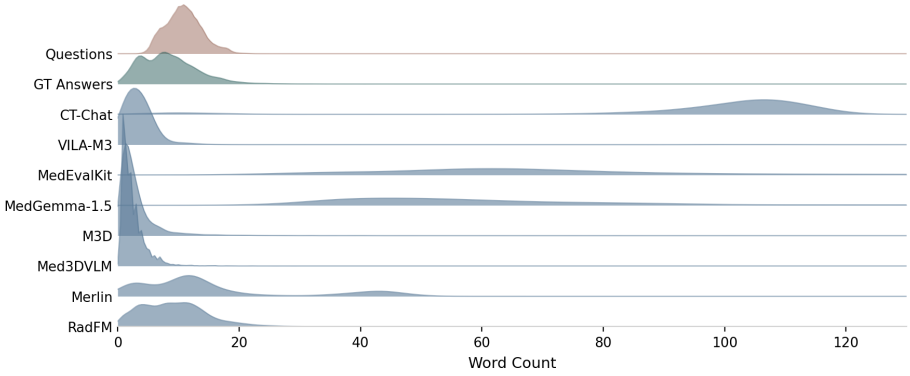


Fig. 4. Distribution of word counts in questions, ground-truth answers and predictions of 3D medical VLMs.

VILA-M3 [14], MedGemma-1.5 (4B) [7], and MedEvalKit [25].

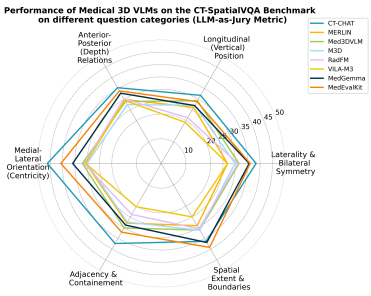
CT-SpatialVQA Dataset. We utilized the publicly available CT-RATE [8] dataset (test split), which contains 1,601 unique radiology reports paired with their corresponding 3D CT scans. Using `GPT-4o` as the Generator LLM and `Gemini 2.5 Flash` as the Validator LLM, we constructed a total of 9,077 spatially grounded question–answer (QA) pairs. The CT-SpatialVQA dataset contains an average of 5.67 QA pairs per report. The average length of the questions is 10.89 words, while the ground-truth answers have an average length of 8.45 words. Figure 4 illustrates the distribution of word counts for questions, ground-truth answers, and predictions generated by eight models. Most models tend to produce concise responses; however, CT-Chat, MedEvalKit and MedGemma generate comparatively longer answers.

Inference & Evaluation Metrics. All models are evaluated in a zero-shot setting following their official inference protocols for fairness. Volumetric CT inputs are processed using each model's prescribed preprocessing pipeline (e.g., slice sampling and resizing). We use both reference-based and LLM-based evaluation. GPT-4o [15], Gemini 2.5 Flash [6], and Qwen3 [1] act as independent binary judges, with aggregated decisions forming an LLM-as-Jury. Semantic alignment is measured via SBERT cosine similarity [19], and lexical overlap via BLEU [16], ROUGE [13], and METEOR [3].

Results and Discussion. Table 1 presents a comparison of various 3D medical VLMs on the CT-SpatialVQA dataset. The LLM-as-Jury evaluation indicates that all models struggle to accurately answer spatially grounded questions related to CT scans, with all models falling below 50% accuracy. Among the

Table 1. Performance of Medical 3D VLMs on the CT-SpatialVQA Benchmark. Note: The first four rows correspond to Accuracy (%).

MODELS → METRICS ↓	CT-CHAT	MERLIN	Med3DVLM	M3D	RadFM	VILA-M3	MedGemma	MedEvalKit	AVERAGE
LLM as Judge (Gemini)	44.68	29.85	34.55	33.04	33.36	29.49	39.30	39.96	35.53
LLM as Judge (GPT)	41.50	29.45	31.82	31.09	31.40	28.43	40.55	39.44	34.21
LLM as Judge (Qwen)	45.27	30.34	32.20	31.21	32.09	27.81	37.33	41.89	34.79
LLM as Jury	43.69	28.85	31.44	30.57	31.49	28.20	38.24	40.16	34.08
SBERT Cosine Similarity	0.562	0.503	0.399	0.412	0.585	0.471	0.581	0.579	0.512
BLEU	3.290	8.110	0.560	1.080	16.45	2.510	3.070	3.460	4.816
ROUGE-L	0.130	0.300	0.198	0.204	0.372	0.273	0.167	0.158	0.225
METEOR	0.230	0.256	0.068	0.077	0.311	0.114	0.237	0.252	0.193



Category	Spatial Relevance
• Laterality & Bilateral Symmetry	• Consistent left-right anatomical grounding relative to the sagittal midline; prevents laterality inversion and side hallucination.
• Longitudinal (Vertical) Position	• Requires volumetric integration along the superior-inferior axis and enforces slice-level consistency across the 3D stack.
• Anterior-Posterior (Depth) Relations	• Preserves depth cues and front-back anatomical arrangement essential for compartment-level reasoning.
• Medial-Lateral Orientation (Centricity)	• Maintains stable central-peripheral localization and global anatomical reference frames.
• Adjacency & Containment	• Encodes topological relations, distinguishing touching from containment within anatomical boundaries.
• Spatial Extent & Boundaries	• Tracks spatial continuity, confinement, and compartment crossing to model disease spread accurately.

Fig. 5. (*left*) Radar plot showing model performance across six spatial reasoning categories in CT-SpatialVQA (LLM-as-Jury). (*right*). Spatial categories considered for our evaluation. Each category captures a clinically relevant spatial primitive required for preserving volumetric spatial integrity in medical VLMs.

evaluated models, CT-Chat (43.69%) performs best, followed by MedEvalKit (40.16%) as the second-best model, and MedGemma-1.5 (38.24%) in third place. In terms of semantic similarity and lexical overlap metrics, the model demonstrates relatively weak performance on the CT-SpatialVQA dataset. The average SBERT cosine similarity is 0.51, indicating limited semantic alignment between predicted and reference answers. Lexical-based metrics show similarly modest results, with an average BLEU score of 4.81, ROUGE-L (F1) of 0.22, and METEOR of 0.19. Overall, these scores suggest that the generated answers frequently diverge from the ground-truth responses both semantically and lexically, reflecting poor performance on spatial reasoning questions within the dataset.

Figure 5 further analyzes performance at the category level. Accuracy varies across spatial relation types, with sharper declines in vertically and depth-oriented reasoning compared to central–peripheral localization. This pattern suggests that performance varies across spatial dimensions, with lower accuracy on relations that require consistent volumetric representations.

4 Conclusion

In this work, we introduce **CT-SpatialVQA**, a clinically grounded benchmark designed to systematically evaluate semantic–spatial reasoning in volumetric CT imaging. CT-SpatialVQA reframes spatial grounding as an explicit and measurable requirement for 3D medical AI. By decomposing spatial reasoning into clinically meaningful primitives and exposing consistent model failures, we establish spatial grounding as a benchmarkable property rather than an assumed byproduct of multimodal scaling. While spatial reasoning limitations are known to affect vision–language models more broadly, our findings demonstrate how these weaknesses manifest in high-stakes 3D clinical settings, where geometric and topological precision are essential. The results highlight structural constraints in prevailing training paradigms that rely heavily on large-scale image–text supervision with limited spatial specificity. CT-SpatialVQA provides a diagnostic framework and research direction, highlighting the need for 3D medical VLMs to move beyond surface correlations toward architectures that encode volumetric structure and spatial consistency.

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

References

1. Alibaba Cloud Model Studio: Qwen-plus (qwen3 series) model listing. https://www.alibabacloud.com/help/en/model-studio/models, accessed 2026-02-24
2. Bai, F., Du, Y., Huang, T., Meng, M.Q.H., Zhao, B.: M3d: Advancing 3d medical image analysis with multi-modal large language models. arXiv preprint arXiv:2404.00578 (2024)
3. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
4. Blankemeier, L., Cohen, J.P., Kumar, A., Van Veen, D., Gardezi, S.J.S., Paschali, M., Chen, Z., Delbrouck, J.B., Reis, E., Truyts, C., et al.: Merlin: A vision language foundation model for 3d computed tomography. Research Square pp. rs–3 (2024)
5. Chen, Y., Xiao, W., Bassi, P.R., Zhou, X., Er, S., Hamamci, I.E., Zhou, Z., Yuille, A.: Are vision language models ready for clinical diagnosis? a 3d medical benchmark for tumor-centric visual question answering. arXiv preprint arXiv:2505.18915 (2025)
6. Google AI for Developers: Gemini 2.5 flash (model code: gemini-2.5-flash). https://ai.google.dev/gemini-api/docs/models/gemini-2.5-flash, accessed 2026-02-24
7. Google Research: Medgemma 1.5 model card (2026), https://huggingface.co/google/medgemma-1.5-4b-it, accessed: 2026-02-22
8. Hamamci, I.E., Er, S., Wang, C., Almas, F., Simsek, A.G., Esirgun, S.N., Dogan, I., Durugol, O.F., Hou, B., Shit, S., et al.: Generalist foundation models from a multimodal dataset for 3d computed tomography. Nature Biomedical Engineering pp. 1–19 (2026)

9. Imam, R., Marew, R., Yaqub, M.: On the robustness of medical vision-language models: Are they truly generalizable? In: Annual Conference on Medical Image Understanding and Analysis. pp. 233–256. Springer (2025)
10. Imam, R., Wang, H., Mahapatra, D., Yaqub, M.: T3: Test-time model merging in vlms for zero-shot medical imaging analysis. arXiv preprint arXiv:2510.27265 (2025)
11. Lai, H., Jiang, Z., Yao, Q., Wang, R., He, Z., Tao, X., Wei, W., Lv, W., Zhou, S.K.: E3d-gpt: enhanced 3d visual foundation for medical vision-language model. arXiv preprint arXiv:2410.14200 (2024)
12. Lee, C., Park, S., Shin, C.I., Choi, W.H., Park, H.J., Lee, J.E., Ye, J.C.: Read like a radiologist: efficient vision-language model for 3d medical imaging interpretation. arXiv preprint arXiv:2412.13558 (2024)
13. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
14. Nath, V., Li, W., Yang, D., Myronenko, A., Zheng, M., Lu, Y., Liu, Z., Yin, H., Law, Y.M., Tang, Y., et al.: Vila-m3: Enhancing vision-language models with medical expert knowledge. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14788–14798 (2025)
15. OpenAI: Gpt-4o model documentation. https://platform.openai.com/docs/models/gpt-4o, accessed 2026-02-24
16. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
17. Rahman, U., Imam, R., Yaqub, M., Amor, B.B., Mahapatra, D.: Can language-guided unsupervised adaptation improve medical image classification using unpaired images and texts? In: 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2025)
18. Rahman, U., Imam, R., Yaqub, M., Mahapatra, D.: Decoupling clinical and class-agnostic features for reliable few-shot adaptation under shift. In: International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging. pp. 123–133. Springer (2025)
19. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). pp. 3982–3992 (2019)
20. Shui, Z., Zhang, J., Cao, W., Wang, S., Guo, R., Lu, L., Yang, L., Ye, X., Liang, T., Zhang, Q., et al.: Large-scale and fine-grained vision-language pre-training for enhanced ct image understanding. arXiv preprint arXiv:2501.14548 (2025)
21. Wang, Y., Dai, Y., Jones, C., Sair, H., Shen, J., Loizou, N., Hsu, W.C., Imami, M., Jiao, Z., Zhang, P., et al.: Enhancing vision-language models for medical imaging: bridging the 3d gap with innovative slice selection. Advances in Neural Information Processing Systems **37**, 99947–99964 (2024)
22. Wolf, D., Hillenhagen, H., Taskin, B., Bäuerle, A., Beer, M., Götz, M., Ropinski, T.: Your other left! vision-language models fail to identify relative positions in medical images (2025), https://arxiv.org/abs/2508.00549
23. Wu, C., Zhang, X., Zhang, Y., Hui, H., Wang, Y., Xie, W.: Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. Nature Communications **16**(1), 7866 (2025)
24. 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)

25. Xu, W., Chan, H.P., Li, L., Aljunied, M., Yuan, R., Wang, J., Xiao, C., Chen, G., Liu, C., Li, Z., et al.: Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. arXiv preprint arXiv:2506.07044 (2025)