CORTEX: A Structured Reasoning Benchmark for Trustworthy 3D Chest CT MLLMs

Hashmat Shadab Malik[1†], Anees Ur Rehman Hashmi[2†], Numan Saeed[1], Muzammal Naseer[3], Salman Khan[1], and Christoph Lippert[2]

[1] MBZUAI, Abu Dhabi, UAE
[2] Hasso Plattner Institute, Potsdam, Germany
[3] Khalifa University, Abu Dhabi, UAE

Abstract. Reasoning in multimodal large language models (MLLMs) has shown strong promise in medical imaging. However, this reasoning is usually free-form text judged only by its final answer, which makes it hard to interpret and verify, especially in 3D radiology, where a diagnosis should be traceable to the evidence in the scan. Existing chest computed tomography (CT) question-answering datasets make this harder by reducing expert radiology reports to answer-only pairs, dropping the reasoning that links findings to conclusions and omitting the patient history clinicians rely on. As a result, reasoning-capable 3D chest CT MLLMs remain out of reach as neither the structured supervision needed to train them nor the protocol needed to verify their reasoning yet exists. We introduce CORTEX (**C**linically **O**rganized **R**easoning and s**T**ructured **EX**planation), a structured reasoning benchmark for 3D chest CT. For each question, CORTEX restores the missing reasoning as a four-stage diagnostic trace that mirrors a radiologist's workflow (task understanding, visual observation, diagnostic reasoning, and answer synthesis). We generate these traces using frontier large language models with broad medical and general-domain knowledge, then filter and verify them with a stage-level evaluation protocol that combines automated rubric scoring with expert radiologist review. Crucially, both the reasoning structure and the evaluation rubrics are designed in close collaboration with clinicians. Built on CT-RATE, a large, publicly available chest CT dataset without reasoning annotations, CORTEX comprises 76,177 validated reasoning traces across open-ended VQA, closed-ended VQA, and report generation, providing both the structured supervision and the stage-level evaluation protocol needed to build and evaluate trustworthy reasoning models for 3D chest CT. Our dataset and codebase is publicly available at github.com/aneesurhashmi/cortex

1 Introduction

Large language models (LLMs) have reshaped natural language processing, achieving strong performance across diverse tasks through transformer scaling and instruction tuning [1]. Building on this success, MLLMs extend the LLMs by

† Equal contribution.

integrating it with visual encoders, enabling joint vision-language understanding [17,30]. The medical community has adapted this paradigm to clinical imaging, achieving strong performance on medical visual question answering (VQA) and radiology report generation [14,28,23]. Yet the dominant paradigm in medical MLLMs is direct response generation, where the model maps an input to an answer in a single step, with no explicit intermediate reasoning. While this suffices for simple recognition or retrieval, complex clinical interpretation that requires integrating heterogeneous visual evidence, reasoning over spatial relationships, and ruling out competing hypotheses exceeds what single-step generation can reliably support.

A common remedy is chain-of-thought (CoT) reasoning, in which a model produces intermediate reasoning steps in natural language before committing to a final answer [27]. CoT can be elicited at inference time through prompting, or taught at training time through supervision, where models are finetuned on examples that contain explicit reasoning traces [10]; both have since been extended to multimodal and medical settings [3,11,29]. In almost all of these, the reasoning is *free-form*, i.e., unconstrained natural language text with no prescribed structure. In medicine, this is a critical shortcoming rather than a stylistic one. When a radiologist reads a scan, we accept their diagnosis because they can articulate how specific imaging findings lead to it and defend that reasoning under scrutiny. A model's internal computation, by contrast, is opaque and its failure modes are difficult to anticipate. A correct answer reached through an implausible or incomplete path offers no clinical guarantee, since its correctness may be coincidental rather than systematic. Because free-form CoT imposes no constraints on how evidence is surveyed, how hypotheses are formed, or how conclusions are justified, its traces vary widely in structure, depth, and grounding. Even more problematic, holistic evaluation by a stronger model can prioritize fluency over correctness, rating a confident, articulate chain highly despite erroneous steps [22].

These problems are most acute in 3D radiology, where interpretation requires coherently integrating evidence across hundreds of interrelated slices that share spatial context and modality-specific patterns. Existing 3D medical MLLMs demonstrate volumetric understanding through report generation and VQA, but they either answer directly or produce unconstrained reasoning that lacks clinically meaningful structure [7,16,13,12,18], and no radiology specific protocol exists to verify such reasoning at the level of individual diagnostic steps. Furthermore, almost all existing CT VQA works omit patient history and reason for examination [31,9,16], even though these are integral to how clinicians reason and determine the correct differential. Without this context, a model must infer from the image alone, its output cannot be checked against the considerations a clinician would actually apply, and hallucination is encouraged wherever context is critical. Progress toward reasoning 3D CT MLLMs is thus constrained on two fronts; 1) the absence of structured reasoning supervision and 2) the absence of a reasoning specific evaluation protocol. To our knowledge, no radiology inspired structured reasoning chest CT MLLM yet exists. A key obstacle is that exist-

ing CT VQA datasets distill dense radiology reports into *answer-only* QA pairs, discarding the diagnostic reasoning that links findings to conclusions, the very supervision such a model would need.

Building on this insight, we introduce **CORTEX** (**C**linically **O**rganized **R**easoning and s**T**ructured **EX**planation), a structured reasoning benchmark for 3D chest CT that recovers the reasoning discarded by answer-only datasets. CORTEX recasts each sample into a four-stage trace mirroring the radiologist's hypothetico-deductive workflow: *task understanding*, *visual observation*, *diagnostic reasoning*, and *answer synthesis*, and reattaches the patient context that prior CT VQA benchmarks omit. We pair the dataset with a clinician-designed evaluation protocol that scores each reasoning stage rather than the final answer alone; here we use it to verify trace quality, and it can subsequently serve to evaluate reasoning models. Both the reasoning structure and the rubrics are designed in close collaboration with clinicians. We frame CORTEX as a concrete step toward future reasoning-capable 3D CT MLLMs: it supplies the structured supervision required to train such models and the stage-level protocol required to evaluate them, with model training and evaluation left to future work (Sec. 4). Our contributions are:

Key Contributions

1. **Radiologist-inspired reasoning framework.** We introduce **CORTEX**, which structures 3D chest CT interpretation into four clinically grounded stages: *task understanding*, *visual observation*, *diagnostic reasoning*, and *answer synthesis*, mirroring the radiologist's hypothetico-deductive workflow and constraining how evidence and hypotheses are handled.
2. **Clinical context grounding.** We integrate patient history and examination indications into both questions and reasoning traces, grounding diagnosis in the interaction between findings and presentation while reducing hallucination.
3. **A clinician-designed evaluation protocol.** Five stage-wise rubrics that score each diagnostic step rather than the final answer alone, used here to verify the benchmark and reusable to evaluate future reasoning models.
4. **A large structured reasoning dataset for 3D chest CT.** We construct a dataset pairing volumetric chest CT with multi-stage reasoning traces, comprising **76,177** validated traces (64,224 open-ended, 8,914 closed-ended, and 3,039 report generation) selected from over 1.9M generated candidates via successive rubric-based filtering.

2 Related Work

Reasoning in Medical MLLMs. CoT prompting [27] and its use as training supervision [10] have been adapted to medical MLLMs mainly to expose

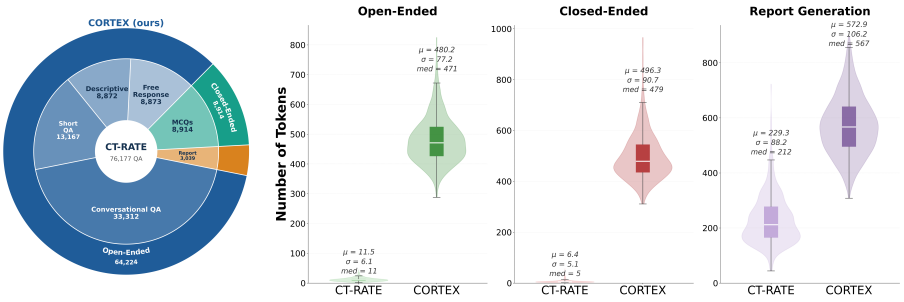

Fig. 1. CORTEX restructuring of the CT-RATE validation split. *Left:* original CT-RATE categories (inner ring) regrouped into three CORTEX VQA types (outer ring), spanning all 76,177 QA pairs. *Right:* per-type answer length in tokens for CT-RATE answers vs CORTEX reasoning traces, far longer due to multi-stage reasoning.

intermediate rationales. These rationales are predominantly free-form [32], and where structure exists it is generic: LLaVA-CoT and chain-of-steps [29,3] use broad stages, while 2D methods such as X-Ray-CoT [19], GEMeX-RMCoT [15], ViTAR [4], and Citrus-V [26] produce region-grounded or iterative rationales. Volumetric efforts remain largely unstructured: Med3D-R1 [12] emits free-form rationales for 3D abnormality diagnosis, and 3DReasonKnee [21] pairs knee MRI with expert rationales but imposes no staged schema. Across these, the reasoning trace is shaped by the model rather than the clinician's diagnostic workflow.

Reasoning Benchmarks in Medical Imaging. Existing 3D medical imaging benchmarks fall into two groups. **Outcome-level benchmarks** score only the final answer: CT-RATE [9] and RadGenome-ChestCT [31] provide large-scale 3D VQA and grounded reports, and tumor-centric [5] and spatial [18] VQA probe understanding through accuracy alone; a parallel modeling line (CT2Rep [7], BTB3D [8], COLIPRI [24], 3D-CT-GPT++ [2], DCP-PD [25], Argus [16]) advances tokenization, alignment, or grounding but is judged holistically, leaving no reasoning to verify. **Process-level benchmarks** are closest to ours but target different settings: DrVD-Bench [33] structures clinical reasoning hierarchically but works on 2D slices, and Med-StepBench [20] detects step-wise hallucinations in 3D oncological PET/CT via clinician-verified distractors. CORTEX instead pairs each 3D chest CT volume with structured, clinician-validated reasoning traces (incorporating patient history where available) and a five-rubric protocol that scores each diagnostic stage (task understanding, observation fidelity, hypothesis evaluation, reasoning logic, answer correctness), bringing stage-level verification to 3D chest CT reasoning.

3 Methodology

CORTEX augments an answer-only chest CT VQA corpus with structured, multi-stage clinical reasoning through a three step pipeline (Fig. 2). We (i) restructure and context-enrich the source data, (ii) generate reasoning traces by

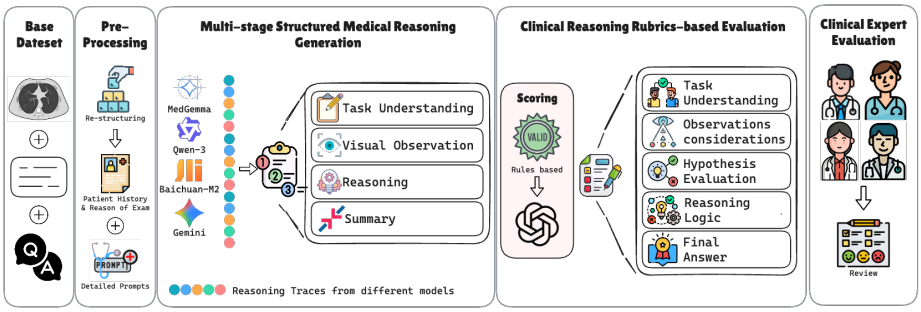

Fig. 2. CORTEX data generation pipeline. Preprocessing restructures the CT-RATE VQA pairs and reattaches clinical context. Using task-specific, clinician-designed prompts, knowledge-rich to frontier LLMs generate four-stage reasoning traces from the paired reports across multiple temperatures. A rule-based gate and clinician-designed rubrics then score and filter the traces, and the best-scoring trace is retained per sample. Clinicians independently review a random subset of the final dataset.

applying task specific, clinician-designed prompts to frontier LLMs grounded in the paired radiology report, and (iii) filter the traces with automatic, clinician-designed verification, confirmed by a final expert review. Clinician feedback enters at three points: the generation templates, the verification rubrics, and the independent review of a final subset. Full clinician-designed prompts, for both reasoning-trace generation and rubric-based validation, along with qualitative samples from different tasks, are available in our codebase.

3.1 Dataset Preprocessing

We build on CT-RATE [9], one of the largest public chest CT corpora, using its validation split (1,304 patients, 3,039 examinations, 76,177 QA pairs). We denote this base corpus as:

$$\mathcal{D}_{\text{base}} = \left\{ (V_i, \mathcal{P}_i, Q_i, H_i, Y_i) \right\}_{i=1}^{N}, \qquad N = 76{,}177,$$

where, for sample i, V_i is the chest CT volume, $\mathcal{P}_i$ its paired report, Q_i the clinical question, H_i the optional clinical context (patient history or reason for examination, $H_i = \varnothing$ if absent), and Y_i the reference answer; volumes and reports are shared across an examination's QA pairs. Preprocessing assigns each sample a reasoning type κ_i and a context-enriched query $\bar{Q}_i$, which with $\mathcal{P}_i$ and Y_i form the inputs for reasoning-based generation (Sec. 3.2).

Reasoning-driven restructuring. CT-RATE's native categories (conversational, descriptive, free-response, short QA, MCQ, report generation) do not map to distinct reasoning demands. We relabel each sample with a map ϕ that sends its category to a reasoning type $\kappa_i \in \mathcal{K} = \{\text{OE}, \text{CE}, \text{RG}\}$: **open-ended** (OE; conversational, descriptive, free-response, short QA) requires free-text synthesis from question relevant findings, **closed-ended** (CE; MCQ) requires selection from

a fixed candidate set, and **report generation** (RG) requires holistic synthesis across volume. Fig. 1*(left)* shows this regrouping for all 76,177 QA pairs of the dataset.

Clinical context integration. Radiology exams are read within a clinical context that informs the diagnosis, yet most CT VQA benchmarks (including CT-RATE) omit it, even though their VQAs derive from the very reports that document it. In $\mathcal{D}_{\text{base}}$, 48.1% of cases (1,462 examinations) carry such context (e.g., *"myelofibrosis patient, ischemic heart disease"*). We append any available context to the question, $\bar{Q}_i = Q_i \oplus H_i$ (and $\bar{Q}_i = Q_i$ when $H_i = \varnothing$), where $\oplus$ denotes text concatenation.

3.2 Structured Reasoning Generation

For each sample we synthesize a reasoning trace R_i that derives Y_i from the visual evidence in V_i. We treat the report $\mathcal{P}_i$ as a faithful textual surrogate for the volume's visual content and draw the trace from model g at temperature τ,

$$R_i \sim p_g\big(\cdot \mid x_i; \tau\big), \qquad x_i = \pi_{\kappa_i}\big(\bar{Q}_i, Y_i, \mathcal{P}_i\big), \tag{1}$$

where π_{κ_i} is the clinician-designed, task-specific prompt template for type κ_i. We use a pool $\mathcal{G}$ of frontier open and closed-weight, medical and general-domain LLMs whose large scale pretraining encodes broad medical and general knowledge, sampled over a temperature set $\mathcal{T} \subset [0, 1]$ for complementary strengths. Each configuration $(g, \tau) \in \mathcal{G} \times \mathcal{T}$ yields one trace $R_i^{(g,\tau)}$ per sample.

Four-stage reasoning structure. We require each trace to follow a four-stage decomposition that makes the diagnostic process explicit and auditable rather than free-form. The design follows the hypothetico-deductive model of diagnosis [6] and was shaped through structured interviews with our clinician panel, mirroring the radiologist's workflow. Each trace factorizes into ordered stages $\mathcal{S} = (\text{TU}, \text{VO}, \text{DR}, \text{AS})$, each conditioned on its predecessors:

$$p_g\big(R_i \mid x_i; \tau\big) = \prod_{s \in \mathcal{S}} p_g\big(R_i^s \mid x_i, R_i^{\prec s}; \tau\big), \qquad R_i = \big(R_i^{\text{TU}}, R_i^{\text{VO}}, R_i^{\text{DR}}, R_i^{\text{AS}}\big). \tag{2}$$

(TU) Task understanding restates the question $\bar{Q}_i$, the study type, and any clinical context, anchoring the task before any visual reasoning. *(VO) Visual observation* narrates, in the first person, the volume's findings as recorded in $\mathcal{P}_i$, organized by anatomy. Observations are restricted to question relevant anatomies for OE/CE and listed exhaustively for RG, and anatomy absent from $\mathcal{P}_i$ yields nothing rather than fabrication. *(DR) Diagnostic reasoning* forms candidate hypotheses (the given options for CE, or hypotheses induced from $(\bar{Q}_i, R_i^{\text{VO}})$ otherwise) and accepts or rejects each using evidence cited from R_i^{VO}, yielding an auditable decision path. *(AS) Answer synthesis* closes with a concise rationale and a final answer consistent with Y_i. The stages are delimited by tags

⟨TASK⟩, ⟨OBSERVATION⟩, ⟨REASON⟩, ⟨ANSWER⟩, making traces training-ready and enabling structural filtering. The resulting traces are substantially longer than the original CT-RATE answers, reflecting the explicit multi-stage reasoning the templates elicit (Fig. 1*(right)*).

Prompt design. We instantiate one task-specific template π_κ per type, designed from clinician feedback before large scale generation. The clinician panel reviews sample traces and templates (flagging anatomical inconsistencies, implausible hypotheses, or missing instructions), and we revise until the reasoning meets the intended structure at a satisfactory clinical standard. Only finalized templates run over the full corpus, so the reasoning is clinically faithful from the outset.

3.3 Quality Verification and Trace Selection

A rule-based gate Φ_{rule} first retains only traces whose four-stage tags appear in the correct order. Surviving traces are scored by an LLM judge J on five clinician-designed, stage-wise rubrics: task understanding, observation fidelity, hypothesis evaluation, reasoning logic, and answer correctness. Each is scored on a 1–10 scale with written justifications, localizing where a trace fails rather than collapsing its quality into a single number.

Empirically, two rubrics (task understanding and answer correctness) consistently saturate at the top of the scale (8–10), while the other three vary across traces. We treat these two as a hard gate $\mathcal{C}_{\text{gate}}$, keeping a trace only if it scores in the 8–10 band on both, and rank the survivors by the mean of the remaining rubrics $\mathcal{C}_{\text{rank}}$ (observation fidelity, hypothesis evaluation, reasoning logic), $\bar{q}(R) = \frac{1}{3}\sum_{j\in\mathcal{C}_{\text{rank}}} q_j(R)$. Selection is performed *independently per sample*: for each i, we collect every configuration $(g,\tau) \in \mathcal{G}\times\mathcal{T}$ whose trace passes the rule gate, clears the hard gate, and meets a quality threshold θ, and keep the single highest-ranked trace among them:

$$\mathcal{V}_i = \Big\{ R_i^{(g,\tau)} : (g,\tau)\in\mathcal{G}\times\mathcal{T},\ \Phi_{\text{rule}}(R_i^{(g,\tau)}) = 1, \\ \min_{j\in\mathcal{C}_{\text{gate}}} q_j(R_i^{(g,\tau)}) \geq 8,\ \bar{q}(R_i^{(g,\tau)}) \geq \theta \Big\}, \tag{3}$$

$$R_i^\star = \arg\max_{R\in\mathcal{V}_i} \bar{q}(R), \qquad \mathcal{D} = \big\{ R_i^\star \big\}_{i=1}^N. \tag{4}$$

This keeps, for every question, the best-scoring trace across all models and temperatures rather than committing to a single global configuration.

3.4 Dataset Generation

We instantiate the framework with five frontier LLMs (MedGemma 27B, Qwen-3 32B, Baichuan-M2 32B, Gemini-3-Flash, and Gemini-3.1-Flash), each sampled at five temperatures (0.0, 0.2, 0.5, 0.7, 1.0), producing 1,605,600 open-ended, 222,850 closed-ended, and 75,975 report-generation candidate traces. We score

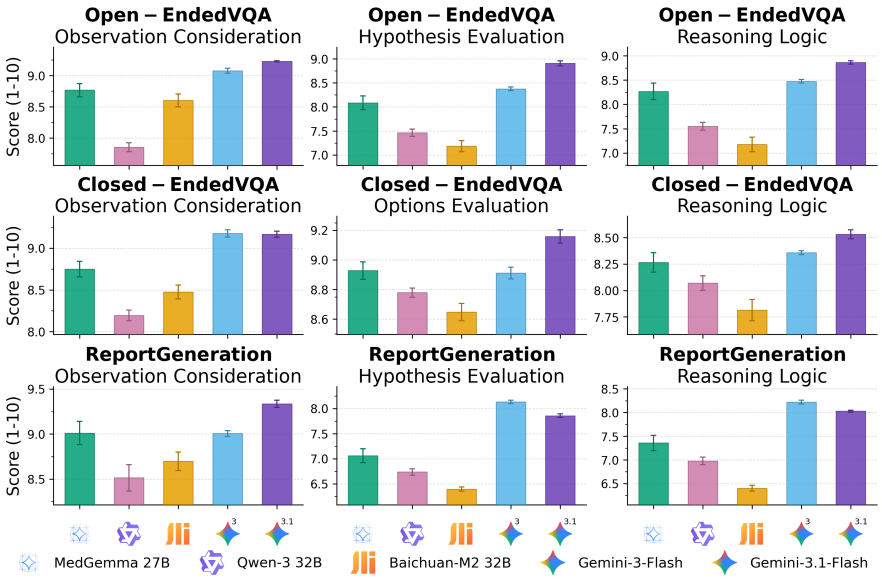

Fig. 3. Rubric-based evaluation of reasoning traces across frontier models. Each bar is a model's mean score (1–10) on the three ranking rubrics (observation fidelity, hypothesis evaluation, reasoning logic), averaged over the validation set and the five temperatures (0.0, 0.2, 0.5, 0.7, 1.0). The two gate rubrics (task understanding and answer correctness) are omitted, as all models score uniformly high. These averages characterize overall model behavior and are not used for the per-sample trace selection.

these with GPT-5.4-mini as judge J and apply the verification and per-sample selection of Sec. 3.3, retaining one trace per question. Fig. 3 reports each generator's mean rubric scores, averaged over the validation set. The resulting dataset $\mathcal{D}$ comprises 64,224 open-ended, 8,914 closed-ended, and 3,039 report-generation reasoning traces, with each subset divided into training, validation, and test splits in an 80:10:10 ratio. Finally, three board-certified radiologists independently review a random subset of $\mathcal{D}$ (1,000 questions), labeling each reasoning trace correct or incorrect; inter-rater agreement on correctness reaches **93**%, corroborating the automated rubric-based validation.

4 Discussion and Outlook

We introduced CORTEX, a structured reasoning benchmark for 3D chest CT that pairs each volume with a structured four-stage reasoning trace, reattaches the clinical context prior CT VQA benchmarks discard, and verifies quality through a clinician-designed five-rubric validation and independent expert review. To our knowledge, it is the first resource to make chest CT reasoning explicit, structured, and verifiable at the level of individual diagnostic stages rather than the final answer alone. The scope of this work is the construction and validation of this benchmark: at present to our knowledge, reasoning-capable 3D

chest CT MLLMs are underexplored, and progress is held back by two missing resources, structured reasoning supervision and a reasoning-specific evaluation protocol, both of which CORTEX supplies. Although CORTEX relies on radiology reports as clinically faithful surrogates for the underlying CT volumes, future work can strengthen this paradigm by generating and validating reasoning traces directly from the image data through expert image-grounded annotation, enabling richer visual evidence and reducing dependence on report content alone. Building on it, our next step is to apply the same generation and verification pipeline at scale to produce a large reasoning-trace training set, and to use it to train the first reasoning-capable 3D chest CT MLLMs, evaluated stage by stage with the present protocol, paving the way toward trustworthy reasoning in 3D radiology.

References

1. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
2. Chen, H., et al.: 3d-ct-gpt++: Enhancing 3d radiology report generation with direct preference optimization and large vision-language models. arXiv preprint (2024)
3. Chen, H., Lou, X., Feng, X., Huang, K., Wang, X.: Unveiling chain of step reasoning for vision-language models with fine-grained rewards. arXiv preprint arXiv:2509.19003 (2025)
4. Chen, K., Rui, S., Jiang, Y., Wu, J., Zheng, Q., Song, C., Wang, X., Zhou, M., Liu, M.: Think twice to see more: Iterative visual reasoning in medical vlms. arXiv preprint arXiv:2510.10052 (2025)
5. Chen, Q., Zhou, X., Liu, C., Chen, H., Li, W., Jiang, Z., Huang, Z., Zhao, Y., Yu, D., He, J., et al.: Scaling tumor segmentation: Best lessons from real and synthetic data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 24001–24013 (2025)
6. Elstein, A.S., Shulman, L.S., Sprafka, S.A.: Medical problem solving: An analysis of clinical reasoning. Harvard University Press (1978)
7. Hamamci, I.E., Er, S., Menze, B.: Ct2rep: Automated radiology report generation for 3d medical imaging. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 476–486. Springer (2024)
8. Hamamci, I.E., Er, S., Shit, S., Reynaud, H., Yang, D., Guo, P., Edgar, M., Xu, D., Kainz, B., Menze, B.: Better tokens for better 3d: Advancing vision-language modeling in 3d medical imaging. Advances in Neural Information Processing Systems (2025)
9. 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.: Developing generalist foundation models from a multimodal dataset for 3d computed tomography. arXiv preprint arXiv:2403.17834 (2024)
10. Hsieh, C.Y., Li, C.L., Yeh, C.K., Nakhost, H., Fujii, Y., Ratner, A., Krishna, R., Lee, C.Y., Pfister, T.: Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In: Findings of the Association for Computational Linguistics: ACL 2023. pp. 8003–8017 (2023)

11. Ke, F., Hsu, J., Cai, Z., Ma, Z., Zheng, X., Wu, X., Huang, S., Wang, W., Haghighi, P.D., Haffari, G., et al.: Explain before you answer: A survey on compositional visual reasoning. arXiv preprint arXiv:2508.17298 (2025)
12. Lai, H., Jiang, Z., Zhang, K., Yao, Q., Wang, R., He, Z., Tao, X., Wei, W., Zhou, S.K.: Med3d-r1: Incentivizing clinical reasoning in 3d medical vision-language models for abnormality diagnosis. arXiv preprint arXiv:2602.01200 (2026)
13. Li, C.Y., Chang, K.J., Yang, C.F., Wu, H.Y., Chen, W., Bansal, H., Chen, L., Yang, Y.P., Chen, Y.C., Chen, S.P., et al.: Towards a holistic framework for multimodal llm in 3d brain ct radiology report generation. Nature Communications **16**(1), 2258 (2025)
14. Li, C., Wong, C., Zhang, S., Usuyama, N., et al.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems **36** (2024)
15. Liu, B., et al.: Gemex-rmcot: An enhanced med-vqa dataset for region-aware multimodal chain-of-thought reasoning. In: arXiv preprint arXiv:2506.17939 (2025)
16. Liu, C., Wan, Z., Wang, Y., Shen, H., Wang, H., Zheng, K., Zhang, M., Arcucci, R.: Argus: benchmarking and enhancing vision-language models for 3d radiology report generation. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 16448–16460 (2025)
17. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
18. Monon, M., Rahman, U., Hanif, A., Saeed, N., Yaqub, M.: Lost in volume: The ct-spatialvqa benchmark for evaluating semantic-spatial understanding of 3d medical vision-language models. arXiv preprint arXiv:2605.08787 (2026)
19. Ng, C., Sun, L., Tang, S.: X-ray-cot: Interpretable chest x-ray diagnosis with vision-language models via chain-of-thought reasoning. arXiv preprint arXiv:2508.12455 (2025)
20. Nguyen, M.K., Le, D.L., Jafari, A.R., Nguyen, T.D., Son, M.H., Thong, M.H., Nguyen, Q.H., Nguyen, T.T., Farahbakhsh, R., Crespi, N., et al.: Med-stepbench: A hierarchical reasoning framework for evaluating hallucinations in medical vision-language models. arXiv preprint arXiv:2605.10002 (2026)
21. Sambara, S., Kim, S.E., Zhang, X., Luo, L., Johri, S., Baharoon, M., Ro, D.H., Rajpurkar, P.: 3dreasonknee: Advancing grounded reasoning in medical vision language models. In: Biocomputing 2026: Proceedings of the Pacific Symposium. pp. 99–113. World Scientific (2025)
22. Tu, M., Ni, S., Bi, K.: How long reasoning chains influence llms' judgment of answer factuality. arXiv preprint arXiv:2604.06756 (2026)
23. Tu, T., Azizi, S., Driess, D., Schaekermann, M., Amin, M., Chang, P.C., Carroll, A., Lau, C., Tanno, R., Ktena, I., et al.: Towards generalist biomedical ai. Nejm Ai **1**(3), AIoa2300138 (2024)
24. Wald, T., Hamamci, I.E., Gao, Y., Bond-Taylor, S., Sharma, H., Ilse, M., Lo, C., Melnichenko, O., Schwaighofer, A., Codella, N.C., et al.: Comprehensive language-image pre-training for 3d medical image understanding. arXiv preprint arXiv:2510.15042 (2025)
25. Wang, C., Dai, W., Liu, H., Li, W., Batmanghelich, K.: Enhancing fine-grained spatial grounding in 3d ct report generation via discriminative guidance. arXiv preprint arXiv:2604.10437 (2026)
26. Wang, G., et al.: Citrus-v: Advancing medical foundation models with unified medical image grounding for clinical reasoning. arXiv preprint arXiv:2509.19090 (2025)

27. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)
28. 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)
29. Xu, G., Jin, P., Wu, Z., Li, H., Song, Y., Sun, L., Yuan, L.: Llava-cot: Let vision language models reason step-by-step. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2087–2098 (2025)
30. Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., Chen, E.: A survey on multimodal large language models. National Science Review **11**(12), nwae403 (2024)
31. Zhang, X., Wu, C., Zhao, Z., Lei, J., Zhang, Y., Wang, Y., Xie, W.: Radgenome-chest ct: A grounded vision-language dataset for chest ct analysis. arXiv preprint arXiv:2404.16754 (2024)
32. Zhang, Z., Zhang, A., Li, M., Smola, A.: Multimodal chain-of-thought reasoning in language models. Transactions on Machine Learning Research (2023)
33. Zhou, T., Zhu, Y., Xiao, C., Bian, H., Wei, L., Zhang, X., et al.: Drvd-bench: Do vision-language models reason like human doctors in medical image diagnosis? Advances in Neural Information Processing Systems **38** (2026)