NEURAL INFORMATION PROCESSING SYSTEMS

ZHEJIANG UNIVERSITY

# 3D-RAD: A Comprehensive 3D Radiology Med-VQA Dataset with Multi-Temporal Analysis and Diverse Diagnostic Tasks

**Xiaotang Gai[1,2,*], Jiaxiang Liu[1,3,*], Yichen Li[1,2,*], Zijie Meng[1,2,*], Jian Wu[2], Zuozhu Liu[1,2,#]**

*[1] ZJU-Angelalign R&D Center for Intelligence Healthcare, Zhejiang University, China*

*[2] Zhejiang Key Laboratory of Medical Imaging Artificial Intelligence, Zhejiang University, China*

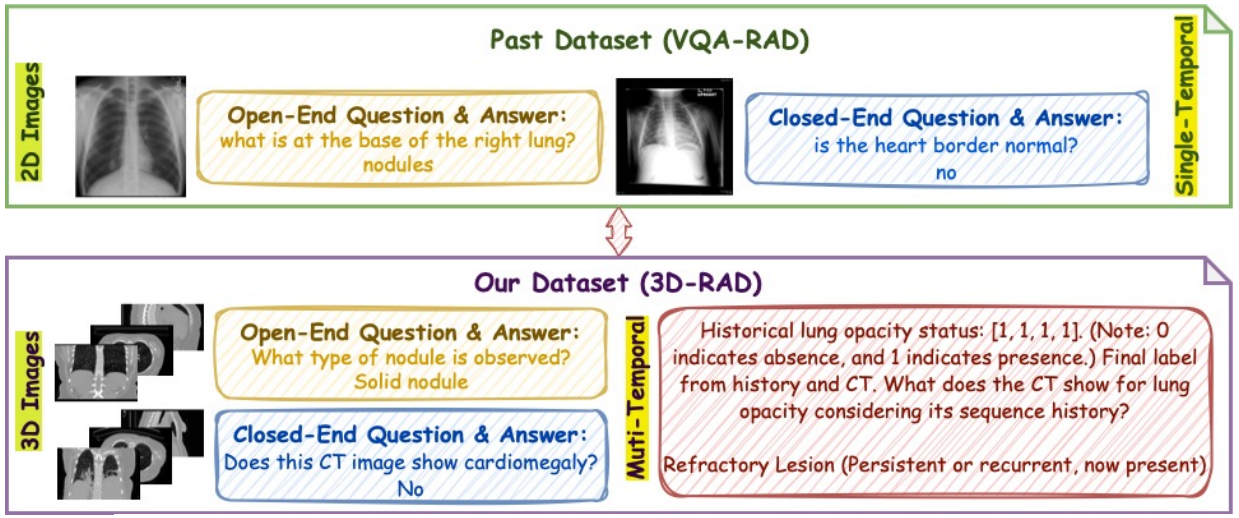
*[3] Guangdong Institute of Intelligence Science and Technology, Hengqin, Zhuhai, China*

**November 6, 2025**

ZJU-UIUC INSTITUTE

# Contents

# Related Work




**Past Dataset:**

2D Images & Open-End & Closed-End (Single-Temporal)

**Our Dataset:**

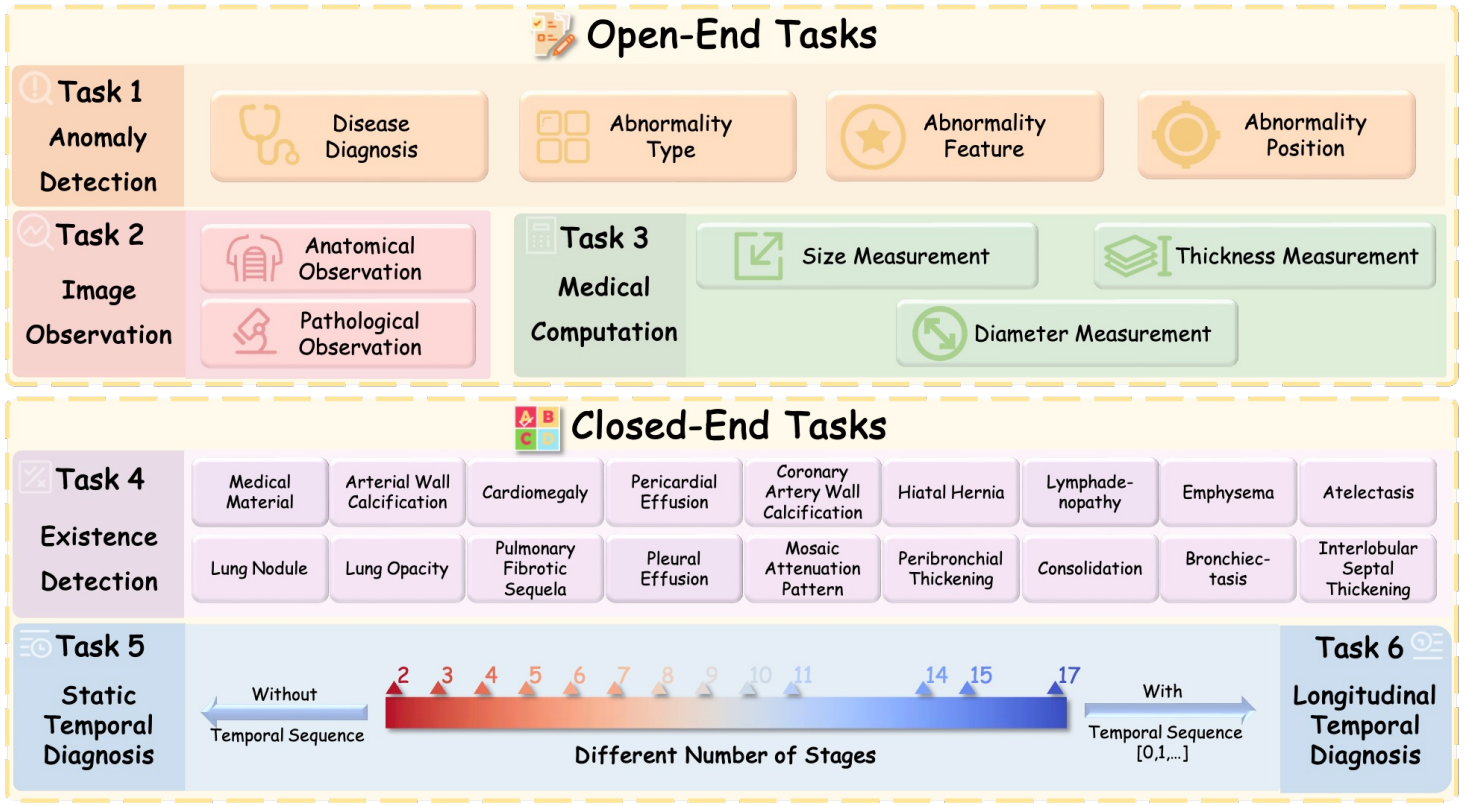
3D Images & Open-End & Closed-End (Muti-Temporal)

| Dataset | Modality | Dataset Scale | Tasks Covered | Temporal Reasoning | 3D Support | Quality Check |
|---|---|---|---|---|---|---|
| **VQA-RAD [11]** | 2D (X-ray, CT) | 315 images 3,515 QA | Modality, anatomy, abnormalities | ✗ | ✗ | Dual Annotation |
| **SLAKE [13]** | 2D | 642 images 14,028 QA | Function, anatomy | ✗ | ✗ | - |
| **PathVQA [14]** | 2D (Pathology) | 4,998 images 32,799 QA | Histopathology | ✗ | ✗ | Manual Validation |
| **VQA-Med [12]** | 2D | ~5K images | Classification, description | ✗ | ✗ | Manual Validation |
| **M3D-VQA [16]** | 3D (CT) | 120K QA | Slice finding, spatial reasoning, diagnosis | ✗ | ✓ | - |
| **3D-RAD (Ours)** | 3D (CT) | **16,188** images, **170K** QA across 6 tasks | Detection, quantification, diagnosis, **progression** | ✓ | ✓ | LLM Score + **Human** Validation + **Consistency Check** |

[1] Xiaotang Gai et al. "3D-RAD: A Comprehensive 3D Radiology Med-VQA Dataset with Multi-Temporal Analysis and Diverse Diagnostic Tasks." *arxiv 2506.11147*

# Contents

# 3D-RAD Dataset



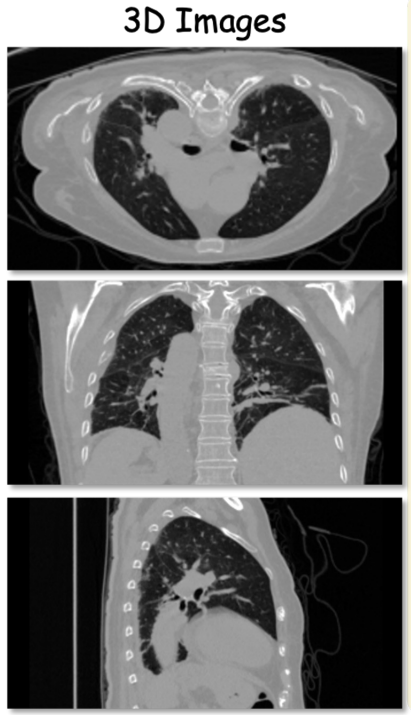

## Task Definition

- **Task 1:** Anomaly Detection
- **Task 2:** Image Observation
- **Task 3:** Medical Computation
- **Task 4:** Existence Detection
- **Task 5:** Static Temporal Diagnosis
- **Task 6:** Longitudinal Temporal Diagnosis

| Source | Category | Task 1 | Task 2 | Task 3 | Task 4 | Task 5 | Task 6 | Overall |
|---|---|---|---|---|---|---|---|---|
| 3D-RAD-Bench | Images | 1,858 | 980 | 656 | 1,304 | 169 | 169 | 2,662 |
| | Patients | 1,008 | 651 | 331 | 1,304 | 169 | 169 | 1,304 |
| | Q-A Pairs | 2,666 | 1,024 | 1,002 | 23,472 | 2,873 | 2,873 | 33,910 |
| 3D-RAD-T (Train) | Images | 5,670 | 2,045 | 1,148 | 5,565 | 774 | 774 | 13,526 |
| | Patients | 4,443 | 1,697 | 587 | 5,565 | 774 | 774 | 9,951 |
| | Q-A Pairs | 6,055 | 2,081 | 1,573 | 100,170 | 13,158 | 13,158 | 136,195 |

[1] Xiaotang Gai et al. "3D-RAD: A Comprehensive 3D Radiology Med-VQA Dataset with Multi-Temporal Analysis and Diverse Diagnostic Tasks." *arxiv 2506.11147*

# Meta Data




Findings_EN:
...The aortic arch calibration is 30 mm, slightly wider than normal. Pericardial thickening is observed.Thoracic esophageal calibration was normal and no significant tumoral wall thickening was detected. In the case, there is a lymph node of approximately 24x16 mm on the right in the central cervical central group. In the old review it is 18x12 mm...

Impressions_EN:
There is thickening of the pleura at posterobasal levels in both lungs, a slight smear-like pleural effusion on the right and fibroatelectatic density increases are observed...

Multi-Stage Labels (18) :
... Emphysema,Atelectasis,Lung nodule,Pulmonary fibrotic sequela, ...
First :0,0,0,1,0,0,1,0,0,1,0,0,0,0,0,0,0,0
Second:0,1,1,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0
Third :0,0,1,1,0,0,1,0,0,0,0,1,1,0,0,0,0,0
Fourth:0,1,0,1,0,0,1,0,0,0,1,1,1,0,0,0,0,0

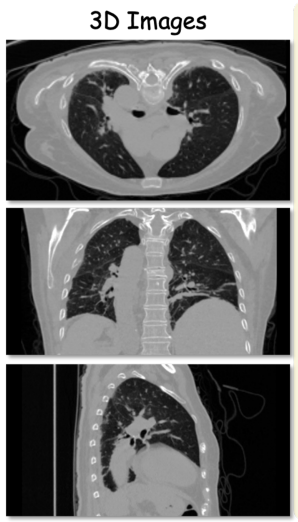
| | |
|---|---|
| Task1: Anomaly Detection | Train: 6,055 Bench:2,666 |
| Task2: Image Observation | Train: 2,081 Bench:1,024 |
| Task 3: Medical Computation | Train: 1,573 Bench: 1,002 |
| Task 4: Existence Detection | Train: 100,170 Bench: 23,472 |
| Task 5: Static Temporal Diagnosis | Train: 13,158 Bench: 2,873 |
| Task 6: Longitudinal Temporal Diagnosis | Train: 13,158 Bench: 2,873 |

- 3D-RAD is built upon **CT-RATE**, a large-scale 3D chest CT dataset paired with radiology reports, licensed under **CC BY-NC-SA**.
- Source data includes **Findings, Impressions, and Labels**. Labels include annotations across different stages of the same case. Tasks use different parts, as color-coded in the figure.
- We collect **16,188 CT scans** from **11,255 patients**, strictly separating training and benchmark sets by following the original training/validation split.

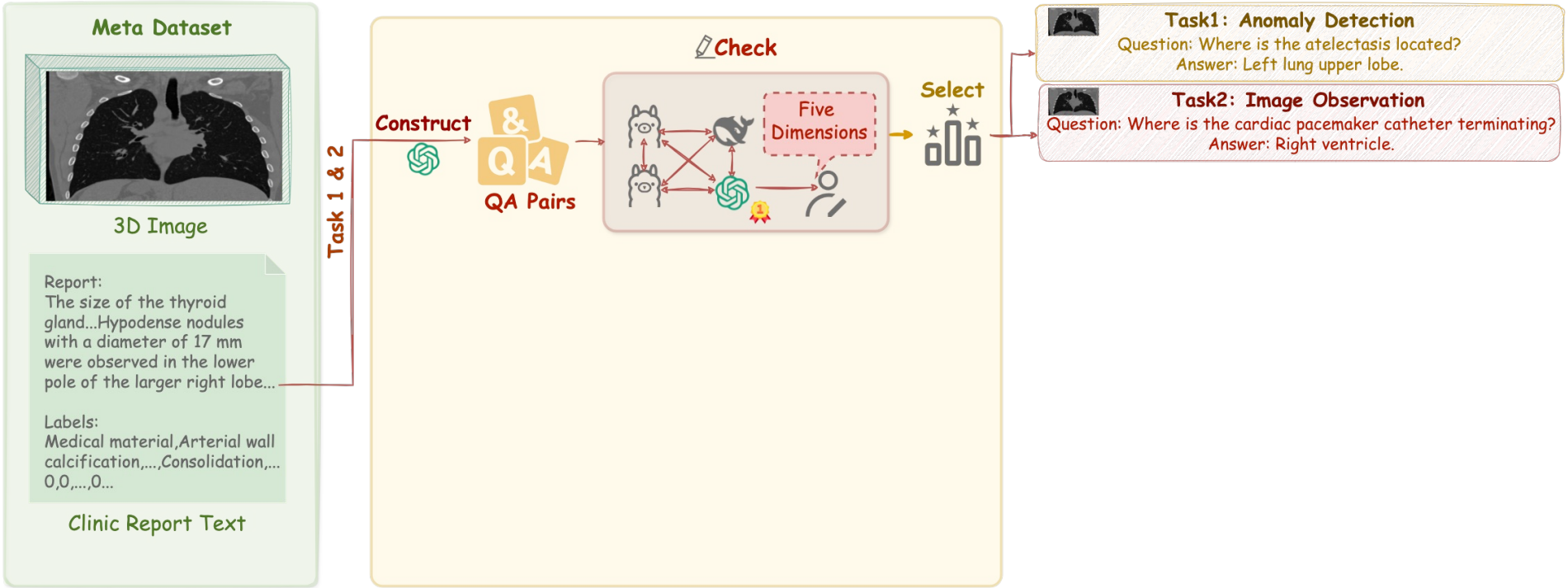
[1] Xiaotang Gai et al. "3D-RAD: A Comprehensive 3D Radiology Med-VQA Dataset with Multi-Temporal Analysis and Diverse Diagnostic Tasks." *arxiv 2506.11147*

# Contents

# Task 1&2 Construction Pipeline



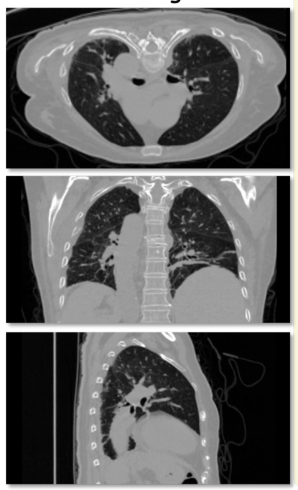

| | |
|---|---|
| Task1: Anomaly Detection | Train: 6,055 Bench:2,666 |
| Task2: Image Observation | Train: 2,081 Bench:1,024 |
| Task 3: Medical Computation | Train: 1,573 Bench: 1,002 |
| Task 4: Existence Detection | Train: 100,170 Bench: 23,472 |
| Task 5: Static Temporal Diagnosis | Train: 13,158 Bench: 2,873 |
| Task 6: Longitudinal Temporal Diagnosis | Train: 13,158 Bench: 2,873 |






[1] Xiaotang Gai et al. "3D-RAD: A Comprehensive 3D Radiology Med-VQA Dataset with Multi-Temporal Analysis and Diverse Diagnostic Tasks." *arxiv 2506.11147*

# Task 3 Construction Pipeline

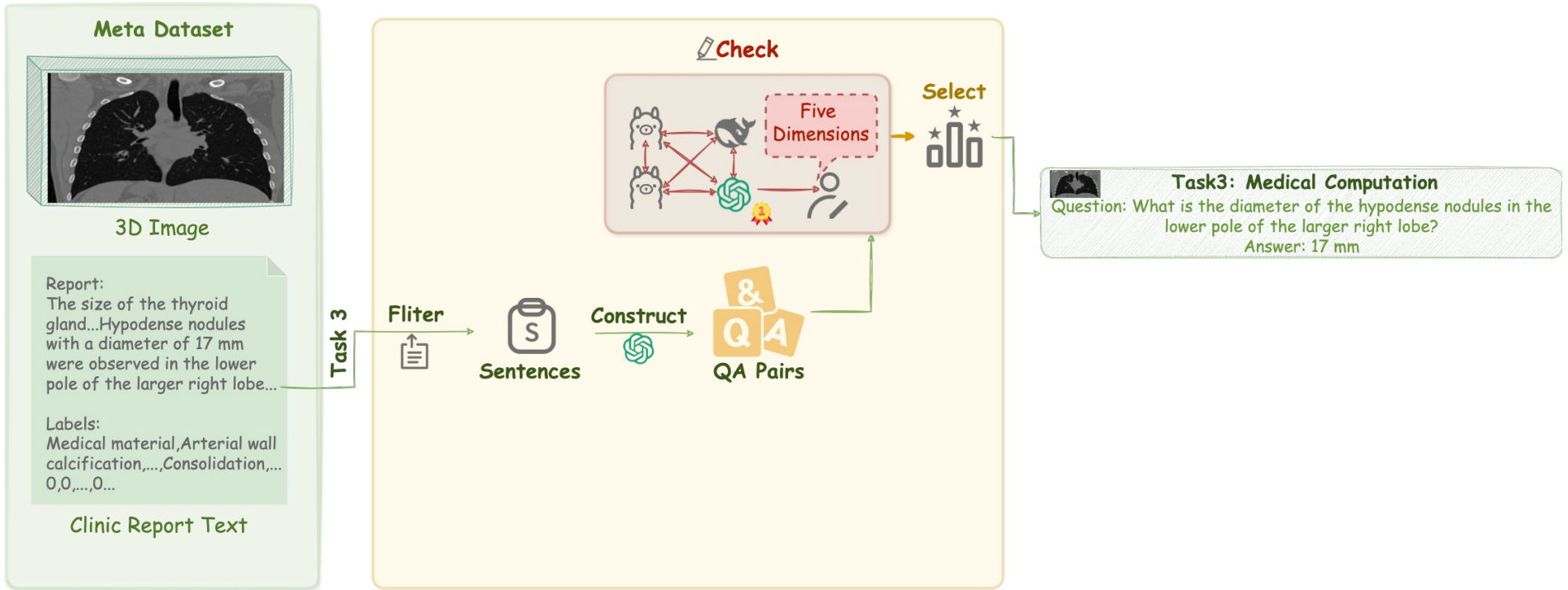



Findings_EN:
...The aortic arch calibration is 30 mm, slightly wider than normal. Pericardial thickening is observed.Thoracic esophageal calibration was normal and no significant tumoral wall thickening was detected. In the case, there is a lymph node of approximately 24x16 mm on the right in the central cervical central group. In the old review it is 18x12 mm...

Impressions_EN:
There is thickening of the pleura at posterobasal levels in both lungs, a slight smear-like pleural effusion on the right and fibroatelectatic density increases are observed...

Multi-Stage Labels (18) :
... Emphysema,Atelectasis,Lung nodule,Pulmonary fibrotic sequela, ...
First:0,0,0,1,0,0,1,0,0,1,0,0,0,0,0,0,0,0
Second:0,1,1,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0
Third:0,0,1,1,0,0,1,0,0,0,0,1,1,0,0,0,0,0
Fourth:0,1,0,1,0,0,1,0,0,0,1,1,1,0,0,0,0,0

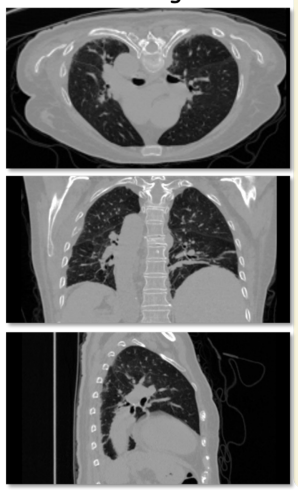
| | |
|---|---|
| Task1: Anomaly Detection | Train: 6,055 Bench:2,666 |
| Task2: Image Observation | Train: 2,081 Bench:1,024 |
| Task 3: Medical Computation | Train: 1,573 Bench: 1,002 |
| Task 4: Existence Detection | Train: 100,170 Bench: 23,472 |
| Task 5: Static Temporal Diagnosis | Train: 13,158 Bench: 2,873 |
| Task 6: Longitudinal Temporal Diagnosis | Train: 13,158 Bench: 2,873 |






[1] Xiaotang Gai et al. “3D-RAD: A Comprehensive 3D Radiology Med-VQA Dataset with Multi-Temporal Analysis and Diverse Diagnostic Tasks.” *arxiv 2506.11147*

# Task 4&5&6 Construction Pipeline

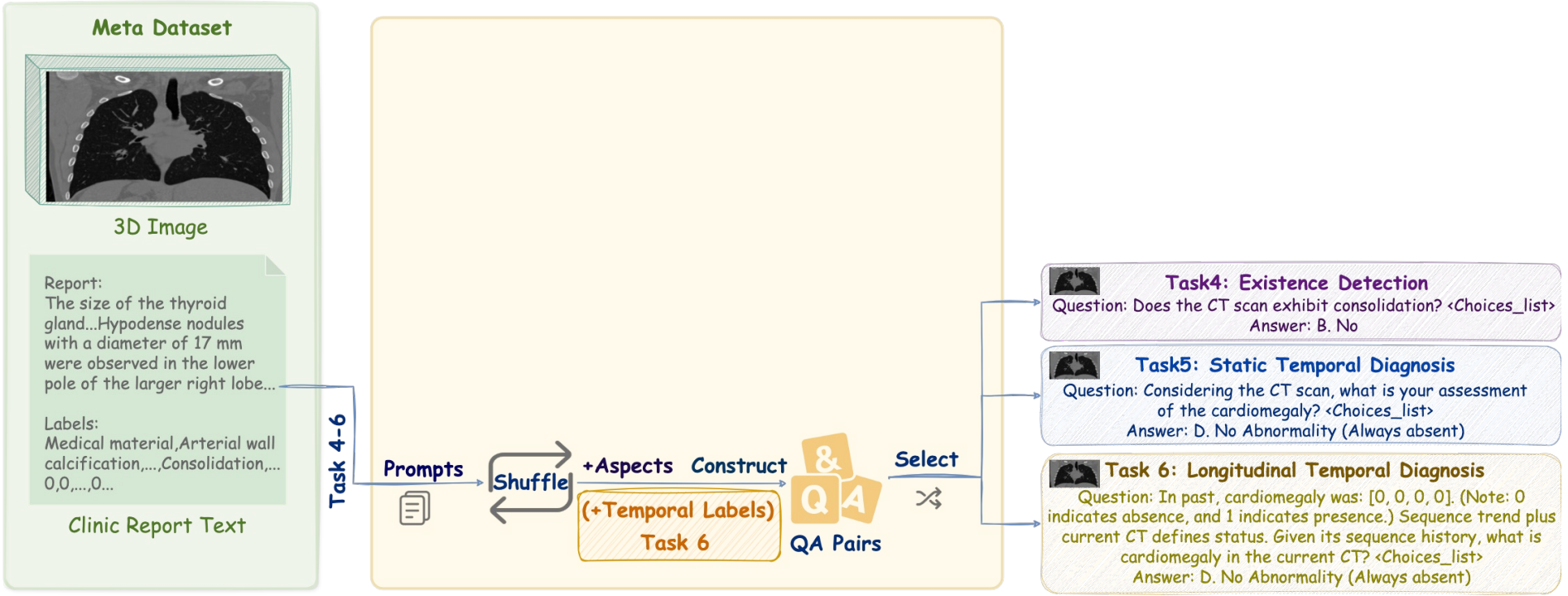



Findings_EN:
...The aortic arch calibration is 30 mm, slightly wider than normal. Pericardial thickening is observed.Thoracic esophageal calibration was normal and no significant tumoral wall thickening was detected. In the case, there is a lymph node of approximately 24x16 mm on the right in the central cervical central group. In the old review it is 18x12 mm...

Impressions_EN:
There is thickening of the pleura at posterobasal levels in both lungs, a slight smear-like pleural effusion on the right and fibroatelectatic density increases are observed...

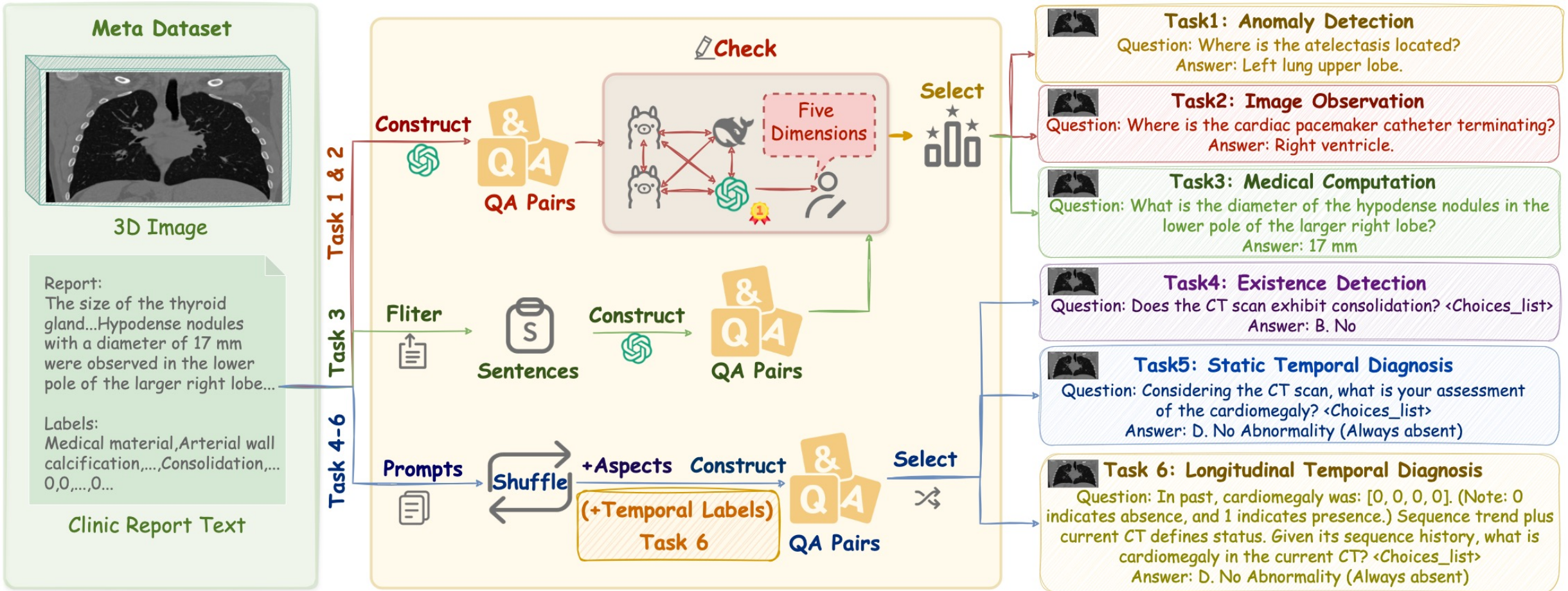
Multi-Stage Labels (18) :
... Emphysema,Atelectasis,Lung nodule,Pulmonary fibrotic sequela, ...
First :0,0,0,1,0,0,1,0,0,1,0,0,0,0,0,0,0,0
Second:0,1,1,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0
Third :0,0,1,1,0,0,1,0,0,0,0,1,1,0,0,0,0,0
Fourth:0,1,0,1,0,0,1,0,0,0,1,1,1,0,0,0,0,0

| | |
|---|---|
| Task1: Anomaly Detection | Train: 6,055 Bench:2,666 |
| Task2: Image Observation | Train: 2,081 Bench:1,024 |
| Task 3: Medical Computation | Train: 1,573 Bench: 1,002 |
| Task 4: Existence Detection | Train: 100,170 Bench: 23,472 |
| Task 5: Static Temporal Diagnosis | Train: 13,158 Bench: 2,873 |
| Task 6: Longitudinal Temporal Diagnosis | Train: 13,158 Bench: 2,873 |




[1] Xiaotang Gai et al. “3D-RAD: A Comprehensive 3D Radiology Med-VQA Dataset with Multi-Temporal Analysis and Diverse Diagnostic Tasks.” *arxiv 2506.11147*

# Construction Pipeline




Overview of Dataset Construction Pipeline



[1] Xiaotang Gai et al. “3D-RAD: A Comprehensive 3D Radiology Med-VQA Dataset with Multi-Temporal Analysis and Diverse Diagnostic Tasks.” *arxiv 2506.11147*

# Prompts and Check

- **Prompt Example**

**Task 1-2 Prompts**

**system_text** = """
You are a medical AI visual assistant that can analyze a single CT image. You receive the medical diagnosis report. The report describes multiple abnormal lesions in the image.
The task is to…These questions come from the following 6 aspects:
*1). Anatomical observation (based on Findings) 2). Pathological observation (based on Findings) 3). Abnormality type (based on Findings) 4). Abnormality feature (based on Findings) 5). Abnormality position (based on Findings) 6). Abnormality or normality diagnosis (based on Impressions)*"""

**PROMPT** = """
Clinical text: { Findings: <Findings> Impressions: <Impressions>}
Please generate a set of exactly 6 clinical image-based question-answer pairs, strictly following these constraints: …
Desired format: …"""

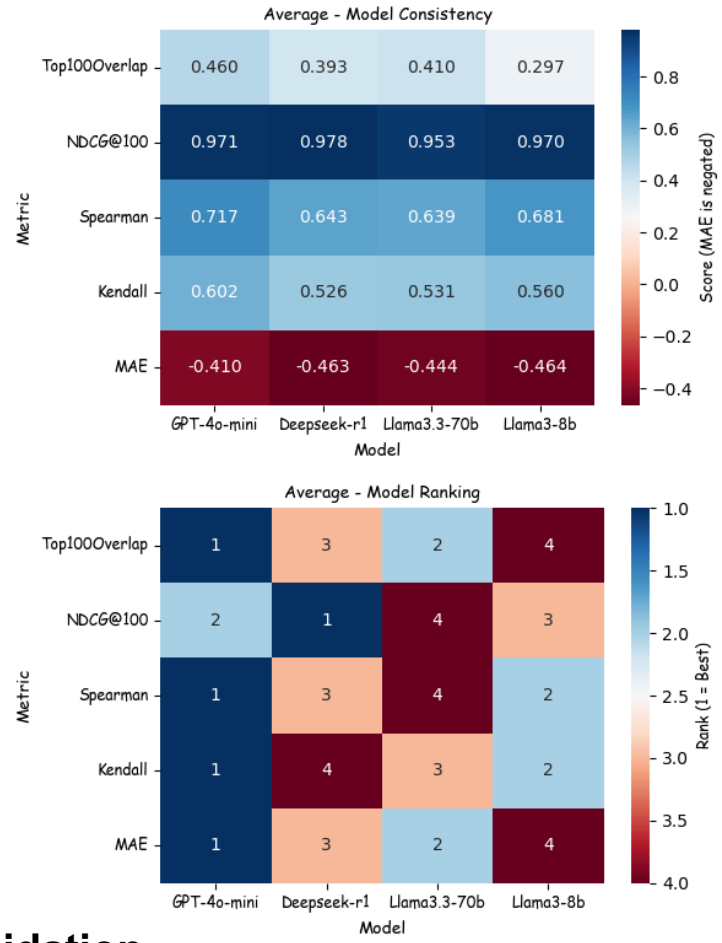
- **Check Prompt**

**Check**

**system_text** = """
You are a medical expert specializing in radiology, particularly in chest CT imaging. Your task is to score the following radiology question and answer pairs based on specific criteria…"""

**PROMPT** = """
Please score the following radiology visual question and answer pair, using the 5 dimensions below. Each should be scored from 1 (very poor) to 5 (excellent). Be strict, specific, and consistent in your evaluation.
Question: {question} Answer: {answer}
Scoring Dimensions:
*1. Visual Verifiability: Can this question be answered just by looking at the image, without requiring medical knowledge, inference, or external context? Does the answer also rely on the image for validation? …*
*2. Specificity & Clarity: Is the question precise and unambiguous? Is the answer also specific and clear, without ambiguity?...*
*3. Answer Appropriateness: Is the answer correct, medically appropriate, specific, and directly relevant to the question? Does it match the expected format/type? …*
*4. Q-A Alignment: Does the answer format/type match the question format/type? Is the answer logically aligned with the type of information requested in the question? …*
*5. Linguistic Quality: Are both the question and the answer grammatically correct, fluent, and easy to understand? Are there any issues with language clarity in either the question or the answer? …*
Example response format:…"""

- **Consistency Check**



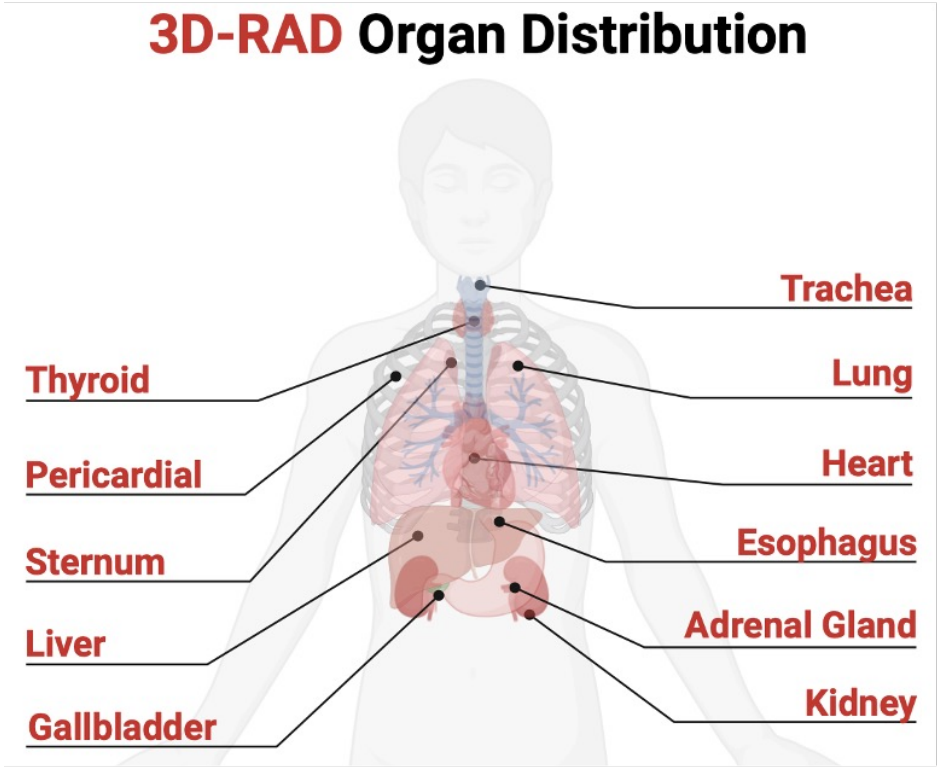

- **Human Validation**

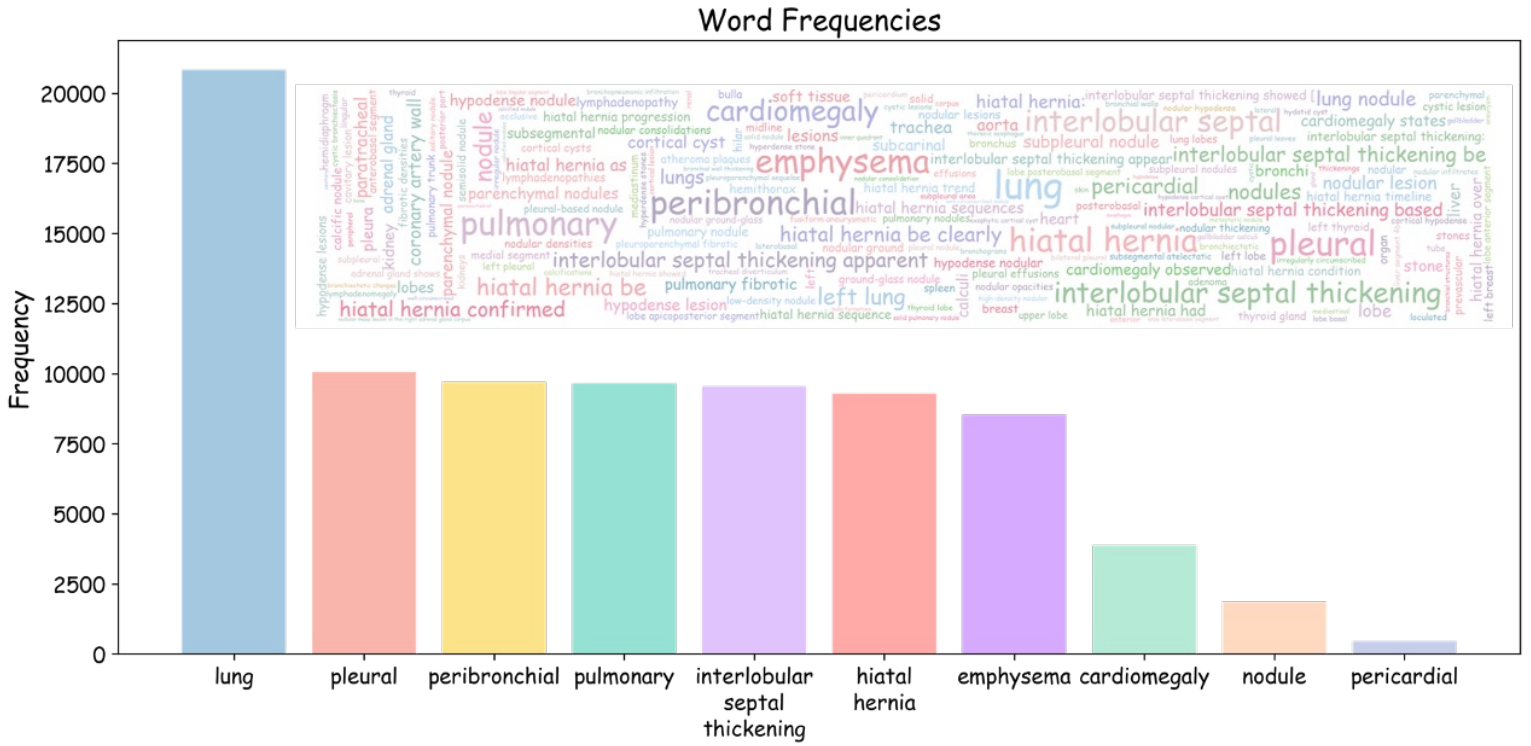
The overall agreement rate is **91.17%** (547/600), which rises to **96.17%** (577/600) after excluding samples with low score.

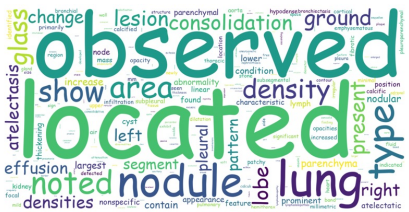
[1] Xiaotang Gai et al. "3D-RAD: A Comprehensive 3D Radiology Med-VQA Dataset with Multi-Temporal Analysis and Diverse Diagnostic Tasks." *arxiv 2506.11147*

# Contents

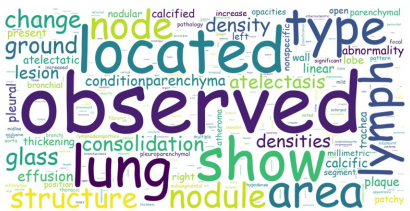
# Overview of 3D-RAD



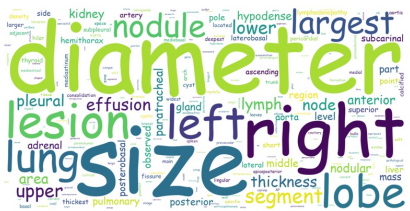



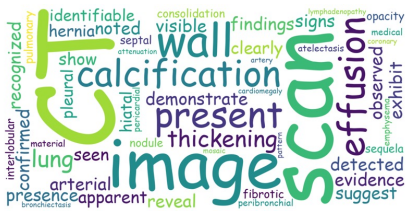

Task 1



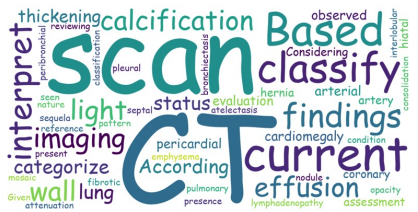

Task 2



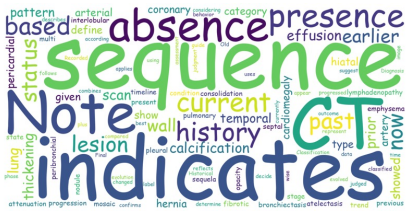

Task 3




Task 4




Task 5




Task 6



[1] Xiaotang Gai et al. “3D-RAD: A Comprehensive 3D Radiology Med-VQA Dataset with Multi-Temporal Analysis and Diverse Diagnostic Tasks.” *arxiv 2506.11147*

# Contents

# Finetuned Results

| Task | Metric | M3D (Llama2-7B) | M3D-RAD (Llama2-7B) | M3D (Phi3-4B) | M3D-RAD (Phi3-4B) |
|---|---|---|---|---|---|
| Anomaly Detection | BLEU | 9.10 | 25.25+16.15 | 15.06 | **33.28**+18.22 |
| | Rouge | 18.64 | 33.76+15.12 | 23.19 | **42.45**+19.26 |
| | BERTScore | 86.07 | 89.16+3.09 | 87.11 | **90.72**+3.61 |
| Image Observation | BLEU | 10.69 | 31.28 +20.59 | 16.31 | **39.66**+23.35 |
| | Rouge | 20.82 | 39.12 +18.30 | 23.19 | **50.52**+27.33 |
| | BERTScore | 86.61 | 90.00 +3.39 | 86.92 | **92.19**+5.27 |
| Medical Computation | BLEU | 15.95 | 30.54+14.59 | 2.55 | **33.52**+30.97 |
| | Rouge | 23.24 | 36.06+12.82 | 5.63 | **36.46**+30.83 |
| | BERTScore | 91.50 | 94.65+3.15 | 85.74 | **94.86**+9.12 |
| Existence Detection | Accuracy | 18.00 | 81.09+63.09 | 40.25 | **82.43**+42.18 |
| Static Temporal Diagnosis | Accuracy | 25.47 | **51.20**+25.73 | 25.40 | 49.30+23.90 |
| Longitudinal Temporal Diagnosis | Accuracy | 24.17 | **74.78**+50.61 | 24.31 | 74.77+50.46 |

- Fine-tuning **consistently improves** both small and large M3D-RAD variants across all tasks, demonstrating our dataset's effectiveness.
- For the newly introduced multi-phase Tasks 5 and 6, baseline accuracies were low (~20%) but rose to **over 70% after fine-tuning**, revealing significant gains.
- However, performance still **lags behind traditional single-phase tasks**, indicating that temporal cues help but don't fully bridge the reasoning gap.

[1] Xiaotang Gai et al. "3D-RAD: A Comprehensive 3D Radiology Med-VQA Dataset with Multi-Temporal Analysis and Diverse Diagnostic Tasks." *arxiv 2506.11147*

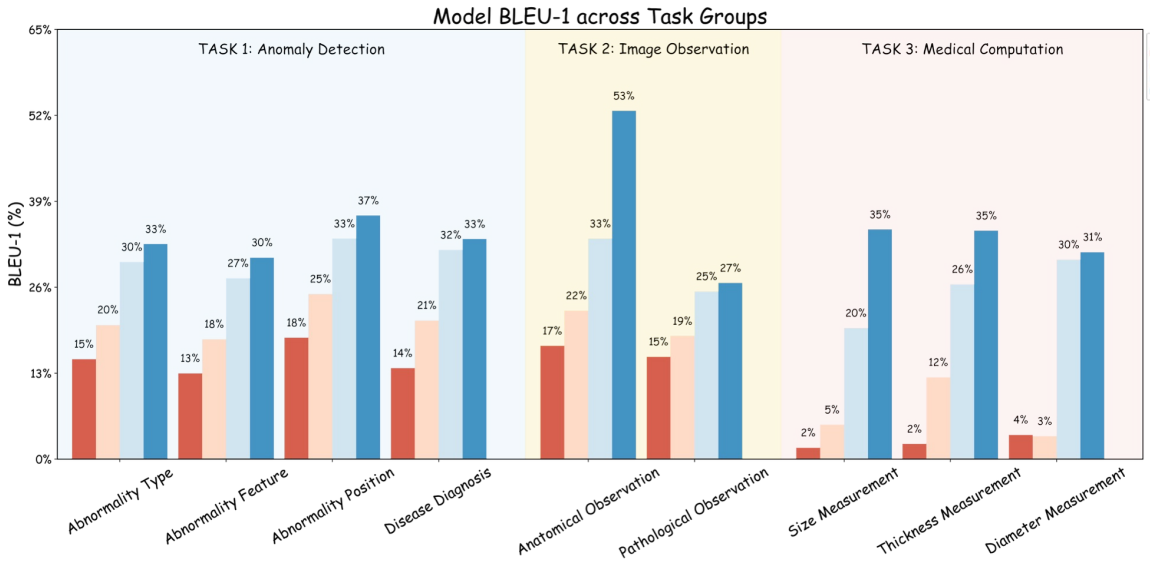
# Zero-Shot Results

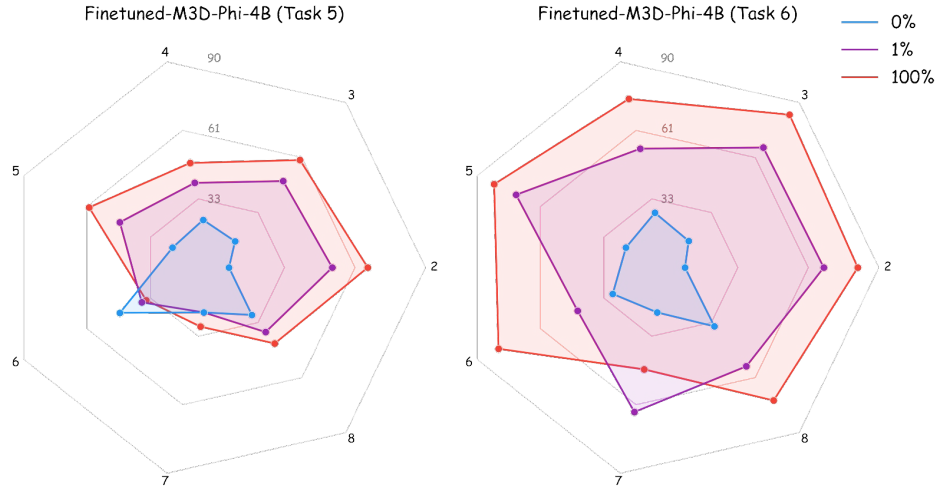
| Task | Metric | RadFM | M3D (Llama2-7B) | M3D (Phi3-4B) | OmniV (Qwen2.5-1.5B) |
|---|---|---|---|---|---|
| Anomaly Detection | BLEU | 11.00 | 9.10 | **15.06** | 13.47 |
| | Rouge | 17.62 | 18.64 | 23.19 | **25.72** |
| | BERTScore | 86.76 | 86.07 | 87.11 | **88.21** |
| Image Observation | BLEU | 13.48 | 10.69 | 16.31 | **16.42** |
| | Rouge | 19.14 | 20.82 | 23.19 | **26.69** |
| | BERTScore | 87.16 | 86.61 | 86.92 | **88.29** |
| Medical Computation | BLEU | 3.34 | **15.95** | 2.55 | 2.52 |
| | Rouge | 6.62 | **23.24** | 5.63 | 7.88 |
| | BERTScore | 86.85 | **91.50** | 85.74 | 85.66 |
| Existence Detection | Accuracy | 29.20 | 18.00 | **40.25** | 28.66 |
| Static Temporal Diagnosis | Accuracy | **44.11** | 25.47 | 25.40 | 22.96 |
| Longitudinal Temporal Diagnosis | Accuracy | **42.99** | 24.17 | 24.31 | 24.23 |

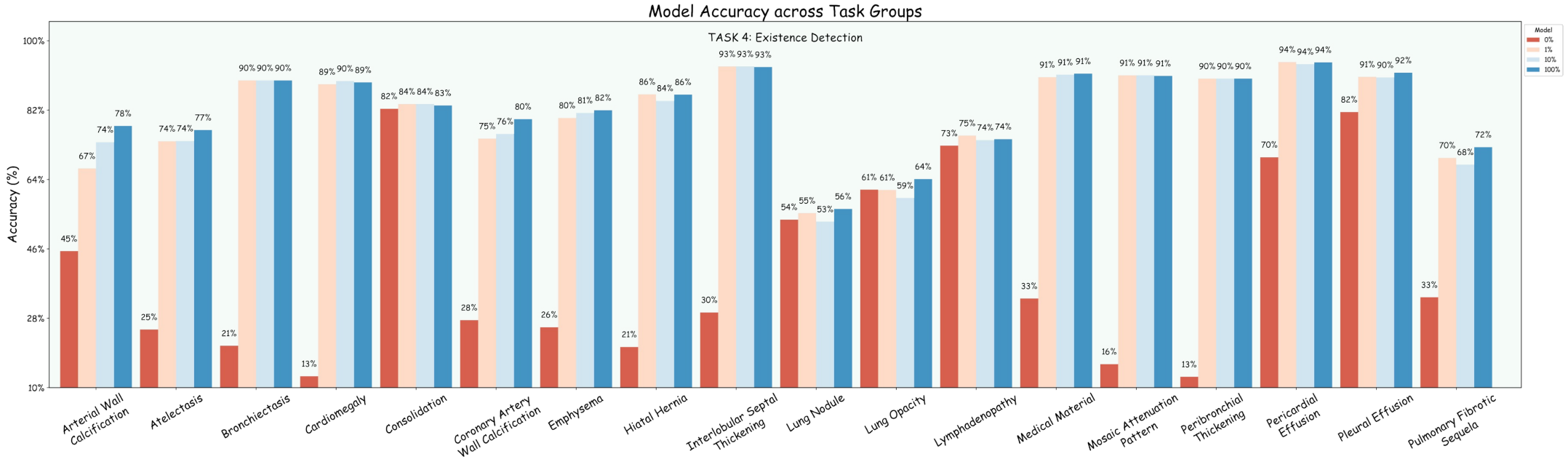
- On standard single-temporal tasks (Task 1, 2, 4), **M3D-4B and OmniV** showed the best performance.
- **M3D-7B** outperformed others on Task 3, suggesting stronger numerical reasoning abilities.
- For our newly proposed multi-temporal tasks, **RadFM** achieved the highest scores, indicating better generalization to temporally structured QA.
- These results reveal task-specific strengths across models, but none perform consistently well on all clinical tasks.This highlights **the need for comprehensive benchmarks covering diverse reasoning types**, including computation, visualization, and multi-temporal inference.

[1] Xiaotang Gai et al. “3D-RAD: A Comprehensive 3D Radiology Med-VQA Dataset with Multi-Temporal Analysis and Diverse Diagnostic Tasks.” *arxiv 2506.11147*

# Varying Training Set Sizes



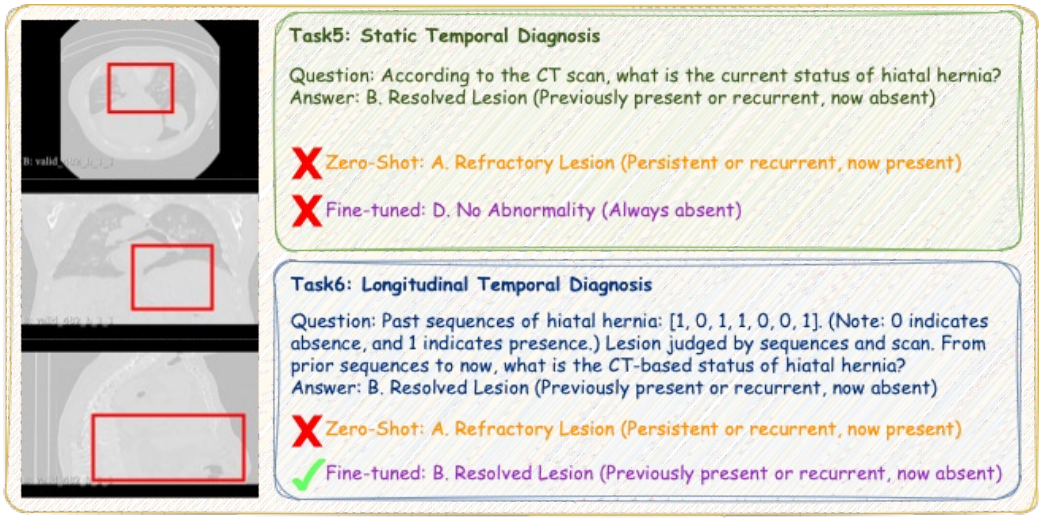



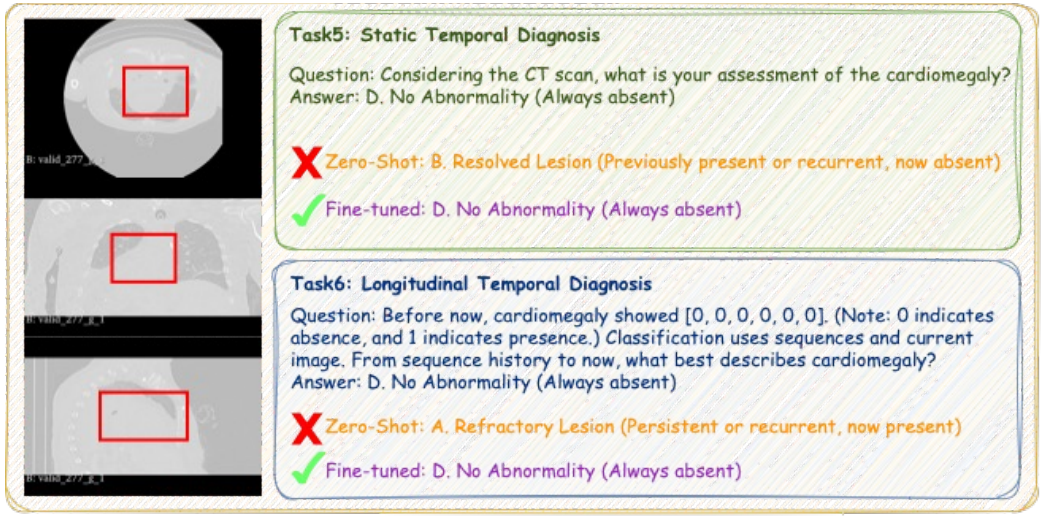

We observe a consistent performance gain across all tasks **as the training size increases,** demonstrating the data efficiency of our benchmark.

[1] Xiaotang Gai et al. "3D-RAD: A Comprehensive 3D Radiology Med-VQA Dataset with Multi-Temporal Analysis and Diverse Diagnostic Tasks." *arxiv 2506.11147*

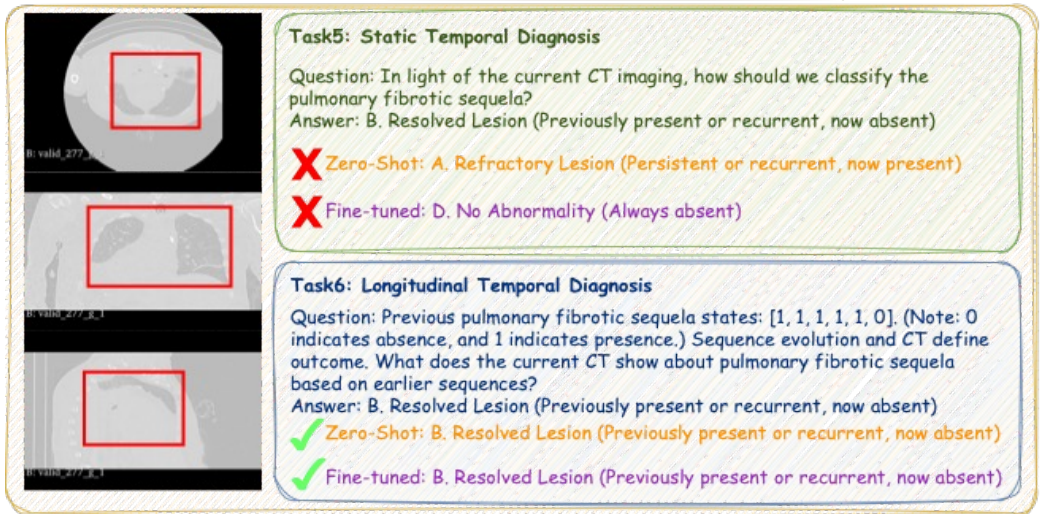
# Failure Case Visualization



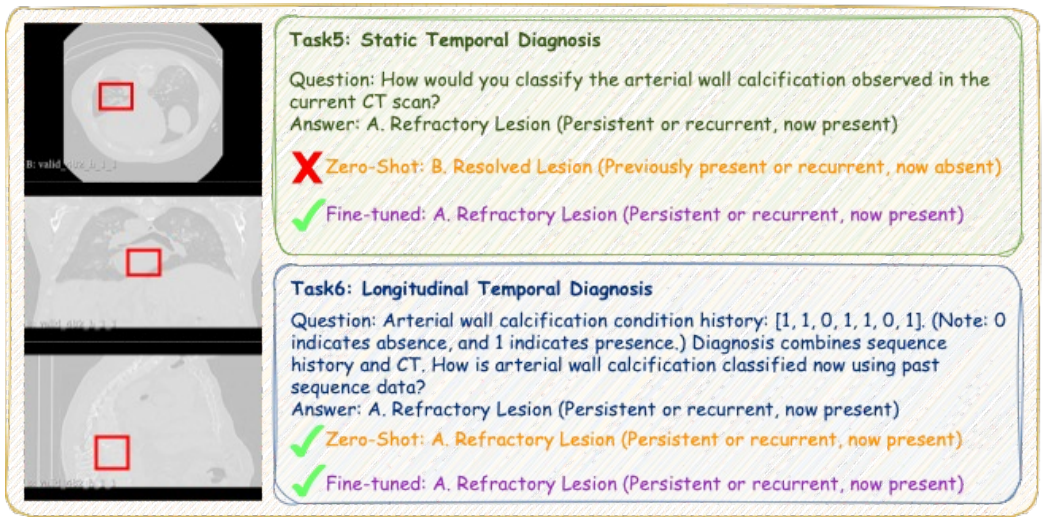

**Case 1:** Both Task 5 and Task 6 failed in zero-shot, and fine-tuning only improved Task 6.




**Case 2:** Zero-shot failed on both tasks; fine-tuning successfully corrected both.




**Case 3:** Task 5 failed in both conditions; Task 6 succeeded without fine-tuning.




**Case 4:** Task 5 failed in zero-shot but succeeded after fine-tuning; Task 6 succeeded in both.

[1] Xiaotang Gai et al. “3D-RAD: A Comprehensive 3D Radiology Med-VQA Dataset with Multi-Temporal Analysis and Diverse Diagnostic Tasks.” *arxiv 2506.11147*

NEURAL INFORMATION PROCESSING SYSTEMS

ZHEJIANG UNIVERSITY

GitHub

# Thank you!

Paper

**Xiaotang Gai[1,2,*], Jiaxiang Liu[1,3,*], Yichen Li[1,2,*], Zijie Meng[1,2,*], Jian Wu[2], Zuozhu Liu[1,2,#]**

*[1] ZJU-Angelalign R&D Center for Intelligence Healthcare, Zhejiang University, China*

*[2] Zhejiang Key Laboratory of Medical Imaging Artificial Intelligence, Zhejiang University, China*

*[3] Guangdong Institute of Intelligence Science and Technology, Hengqin, Zhuhai, China*

ZJU-UIUC INSTITUTE