# Rad-Phi4-Vision-CXR: A Compact Multimodal Assistant for Versatile Radiology Workflows

**Mercy Prasanna Ranjit** *Microsoft Research India* MERANJIT@MICROSOFT.COM
**Anirban Porya** *Microsoft Research India* T-ANIPORYA@MICROSOFT.COM
**Shaury Srivastav** *Microsoft Research India* SHAURY.BTECH.EC19@IIITRANCHI.AC.IN
**Niharika Vadlamudi** *Microsoft Research India* T-NVADLAMUDI@MICROSOFT.COM
**Nikhilesh Chowdary Eathamukkala** *Microsoft Research India* B-NIKHILESHE@MICROSOFT.COM
**Shashank Udyavar** *Microsoft Research India* B-SUDYAVAR@MICROSOFT.COM
**Rahul Kumar** *Microsoft Research India* B-RAHULKUMAR@MICROSOFT.COM
**Tanuja Ganu** *Microsoft Research India* TAGANU@MICROSOFT.COM

## Abstract

The integration of artificial intelligence into radiology underscores the need for efficient models capable of supporting a wide range of clinical tasks. We introduce Rad-Phi4-Vision-CXR, a compact multimodal vision-language model designed to seamlessly integrate into radiology workflows for chest X-rays. It supports radiology report generation, fine-grained visual question answering (VQA) for abnormalities and tubes/lines (including presence and placement), and grounding capabilities for anatomies, pathologies, and medical devices. Beyond these tasks, we propose a capability for findings generation with causal exploration of radiology findings and differential diagnosis, enabling the model to affirm findings or rule out conditions, thereby enhancing its utility in clinical decision-making. Rad-Phi4-Vision-CXR achieves state-of-the-art performance on the ReXrank benchmark for report generation, VQA, and grounding. Its compact architecture provides a scalable, high-performance solution for AI-driven radiology.

**Keywords:** multimodal small language models, chest x-rays, radiology report generation, causal exploration.

**Data and Code Availability** We leverage the Interpret-CXR (Xu et al., 2024), CheXpert-Plus (Chambon et al., 2024), ReXGradient (Zhang et al., 2025b), ReXVQA (Pal et al., 2025) and other datasets which are publicly available for training and benchmarking. Additional dataset details can be found in Appendix D.

**Institutional Review Board (IRB)** Proposed use of public datasets was reviewed by home institution. Under policy, use of de-identified public datasets is classified as Not Human Subjects Research.

## 1. Introduction

The integration of artificial intelligence (AI) into radiology has revolutionized clinical practice by automating image analysis, enhancing diagnostic precision, and alleviating radiologist workload amidst escalating imaging demands and staffing shortages. (Nair et al., 2025). The heterogeneity of radiology workflows—from quality assurance to real-time diagnostic support—demands versatile AI models capable of adapting to diverse tasks without requiring multiple specialized systems. For instance, in an AI-enabled radiology report auditing workflow, the model must be capable of identifying clinically significant and insignificant errors in radiology reports, thereby enabling standardized scoring for quality improvement. In contrast, a radiology report writing assistant must synthesize clinical indications, patient history, and radiographic images to produce detailed, contextually grounded findings that elucidate disease manifestations. In post-operative settings, the model must excel at evaluating medical device placements, such as tubes and lines, to detect malpositions or complications critical to patient outcomes. This versatility is essential, as fragmented AI solutions increase integration complexity, elevate costs, and hinder scalability. A single, adaptable model that supports multiple tasks can streamline work-

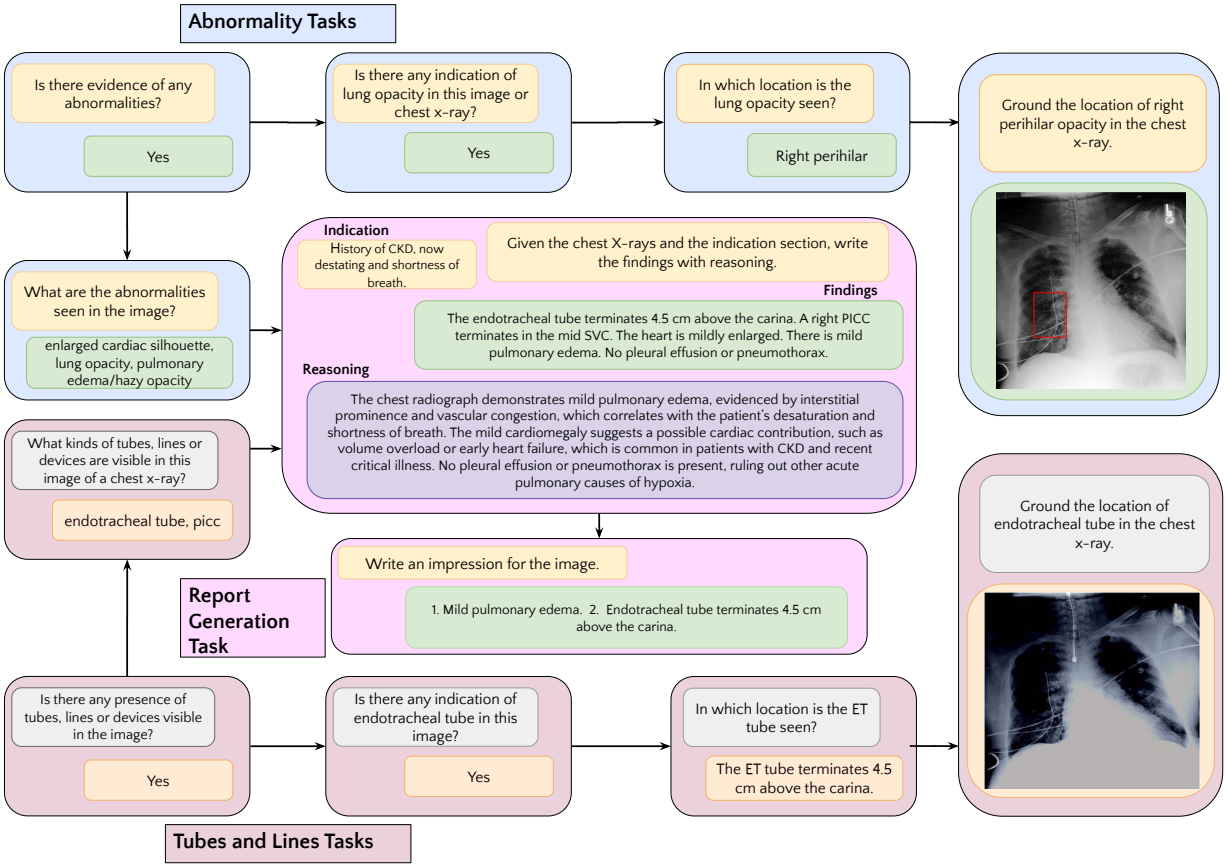



Figure 1: Illustrates the tasks supported by `Rad-Phi4-Vision-CXR` for chest X-ray interpretation. Arrows indicate reasoning at different levels of granularity and across multiple querying modes. Color coding visually distinguishes abnormality-related queries from those concerning tubes and lines.

flows, ensure consistency, and enhance operational efficiency. `Rad-Phi4-Vision-CXR` supports diverse tasks for Chest X-rays. Figure 1 illustrates tasks associated with a radiology reporting workflow.

Beyond versatility, integrating causal exploration, differential diagnosis, and localization capabilities into AI models is essential for producing reliable radiology findings and enhancing clinical decision-making. We define causal exploration as the process of expanding findings into structured reasoning that incorporates radiology knowledge, clinical context, and differential diagnosis, with the goal of supporting better clinical diagnosis. This involves:

- Evidence-based reasoning: Elaborating on radiographic findings (e.g., interstitial prominence, cardiomegaly, device positioning) and explaining their clinical significance.
- Differential diagnosis: Considering plausible alternative explanations for observed findings, especially when imaging appearances are non-specific or patient history is limited.
- Integration of radiology knowledge: Incorporating disease presentations, characteristic radiographic patterns, and clinical context provided in the indication/history section when available.
- Exclusionary reasoning: Highlighting the absence of key findings (e.g., no effusion, no pneumothorax) to rule out competing causes.

By combining these elements, causal exploration challenges models to move beyond surface-level text generation and instead produce clinically useful reasoning narratives that bridge descriptive findings

with diagnostic decision-making. For example, a radiology report assessing free air in the diaphragm could benefit from a causal exploration of findings stating: The presence of free air under the diaphragm is highly suggestive of a perforated viscus. This helps affirm or rule out conditions, facilitating differential diagnosis for efficient clinical decision-making. The `Hidden-Rad` benchmark pioneered the evaluation of causal reasoning in radiology reports (Choi et al., 2025), while `ReXVQA` established a benchmark for fine-grained visual question answering (VQA) aimed at supporting differential diagnosis (Pal et al., 2025). Our key contributions are:

- Radiology Findings Generation with Causal Exploration and Differential Diagnosis: The model generates comprehensive radiology reports augmented with causal analysis, synthesizing clinical indications, patient history, and radiographic findings to produce explainable outputs that affirm or rule out conditions, enhancing clinical decision-making.
- Grounding and Segmentation of Findings and Devices: The model supports precise grounding of abnormalities, anatomies, and pathologies, alongside segmentation of medical devices such as endotracheal tubes (ETT), central venous catheters (CVC), and nasogastric tubes (NGT).
- Fine-Grained Visual Question Answering (VQA): The model delivers strong performance on detailed VQA tasks, addressing questions about abnormality identification and tubes and lines evaluation.
- Comprehensive Benchmarking and Consistency Analysis: We achieve state-of-the-art performance across various benchmarks such as RexRank for report generation and VQA (Pal et al., 2025), PadChest GR benchmark (Castro et al., 2024) for grounding, and Hidden-Rad for causal exploration (Choi et al., 2025). Additionally, we introduce a novel evaluation of model consistency across tasks, verifying reliable performance when findings are queried via different tasks (e.g., report writing versus VQA), a critical step for trustworthy clinical deployment.

## 2. Related Work

Vision-Language Models (VLMs) have emerged as foundational tools for developing versatile radiology assistants, with successive models progressively expanding their capabilities to meet the demands of modern clinical practice.

VLM architectures typically couple an image encoder with a lightweight adapter—commonly a Multi-Layer Perceptron (MLP)—that connects visual features to a Large or Small Language Model (LLM/SLM) for downstream decoding across tasks such as report generation, visual grounding, question answering, and classification.

**Specialist Models** `MAIRA-2` (Bannur et al., 2024) advanced automated report generation for chest X-rays through grounded outputs, localizing findings at the image level to produce explainable and clinically useful drafts. It further demonstrated the value of additional context—e.g., clinical indications, prior scans, and lateral views—showing improved performance when such inputs are incorporated; architecturally, `MAIRA-2` employed a DINO-based image encoder with a 4-layer MLP adapter connected to a Vicuna-7B LLM. Building on task diversity, `CheXagent` (Chen et al., 2024) extends beyond report generation, demonstrating strong performance across chest X-ray interpretation tasks such as disease classification, fine-grained visual question answering (VQA), grounding, and report generation. It uses a SigLIP-based vision encoder and a two-layer MLP adapter connected to a Phi-2 SLM in a multi-stage training setup.

`RadVLM` (Deperrois et al., 2025) further expanded chest X-ray interpretation with multi-turn conversational abilities. It supports diverse tasks (report generation, grounding of anatomies and abnormalities, and VQA) and reports benefits from joint training across tasks.

**Generalist Models** Progress has also been made toward generalist models spanning multiple modalities and anatomies. `MedVersa` (Zhou et al., 2024) is a generalist foundation model that accepts multimodal inputs (CT, MRI, ultrasound, etc.) beyond chest X-rays and orchestrates tasks via an LLM while relying on specialized vision modules for grounding (bounding-box localization) and segmentation. The paper notes that broad task diversity—captioning, classification, VQA, detection, and segmentation—is most fully supported for chest X-rays, with other modalities primarily segmentation-focused; Complementing these, `MedGemma` (Sellergren et al., 2025) is a medical VLM that emphasizes clinical reasoning, biomedical knowledge extraction, report generation,

Table 1: Radiology Reporting Task Settings, Baselines and Benchmarks

| Task Settings | Frontal Img | Prior Img | Lateral Img | Generated Section | Evaluation Benchmarks | Baseline Models |
|---|---|---|---|---|---|---|
| Findings + Impression | ✓ | × | ✓ | Findings, Impression | MIMIC, CheXpert Plus, ReXGradient | RadVLM, MedVersa, MedGemma |
| Findings + Impression + Priors | ✓ | ✓ | × | Findings, Impression | MIMIC, CheXpert Plus, ReXGradient | RadVLM, MedVersa, MedGemma |
| Findings + Causal Exploration | ✓ | × | × | Findings, Causal Exploration | Hidden-RAD, MIMIC, ReXGradient | MedGemma, GPT-5-Chat |

Table 2: Visual Grounding & Segmentation Task Settings, Baselines and Benchmarks

| Task Settings | Evaluation Benchmarks | Baseline Models |
|---|---|---|
| Anatomical Grounding | Chest ImaGenome | RadVLM |
| Phrase Grounding | MS-CXR, PadChest-GR | RadVLM, AGPT-MDETR |
| Abnormality Grounding | NIH-CXR, PadChest-GR | RadVLM, AGPT-MDETR |
| Tubes and Lines Segmentation | RANZCR | - |
| Anatomical Segmentation | CheXmask | MedVersa |

and medical image classification for 2D images across radiology, histopathology, ophthalmology, dermatology, and text-only tasks. It does not support grounding or segmentation.

**Comparison to training approaches adopted in prior work.** `Rad-Phi4-Vision-CXR` builds on this "encoder + adapter + LLM/SLM" design style adopted across `MAIRA-2`, `CheXagent`, `RadVLM`, and `MedVersa` (Bannur et al., 2024; Chen et al., 2024; Deperrois et al., 2025; Zhou et al., 2024), while extending task coverage and training recipes for chest X-ray interpretation.Our work does not propose new architectural components; rather, it consolidates best practices from the literature and extends task breadth to include segmentation of tubes and lines and causal exploration for decision support. We establish new baselines for segmenting tubes and lines—specifically endotracheal tubes (ETT), central venous catheters (CVC), and nasogastric tubes (NGT)—by leveraging LLM embeddings to guide object segmentation.

**Choice of Baselines** We selected a diverse set of baseline models to ensure a comprehensive evaluation. This included more generalist medical models such as `MedGemma` and `MedVersa`, which are designed to handle multiple modalities beyond X-rays, as well as chest-specific multi-task models like `RadVLM` and `CheXagent`. For reasoning-intensive tasks, we incorporate models recognized for strong reasoning capabilities, including `MedGemma` and `GPT-5-Chat`.

## 3. Methodology

### 3.1. Datasets and Tasks

A key component in developing our model is the construction of an instruction tuning dataset. To this end, we aggregate and process multiple publicly available chest X-ray (CXR) datasets that include diverse forms of supervision such as free-text reports, categorical labels, bounding boxes, and segmentation masks. We curate an instruction tuning dataset comprising nearly 7 million instances. Each instance consists of a frontal CXR image (with lateral or prior images when available) paired with a user–assistant interaction for the task at hand. The dataset is designed to support a broad range of tasks relevant to the radiology workflow, including radiology report generation, placement description of tubes and lines and grounding of abnormalities, anatomies and tubes and lines. The set of generative, visual grounding and visual question answering tasks incorporated into the dataset is summarized in Table 1, Table 2 and Table 3. Additional dataset details can be found in Appendix D. The train/test distribution of the datasets for different tasks is given in Table 19. The prompts associated with all the tasks are available in Appendix E.2.

**Radiology Report Generation** We leverage public datasets containing CXRs paired with de-identified reports. Since, radiology reports frequently reference prior examinations, earlier images are often expected as input. However, prior images are

Table 3: VQA Task Settings, Baselines and Benchmarks

| Task Settings | Evaluation Benchmarks | Baselines |
|---|---|---|
| Differential Diagnosis | ReXVQA | GPT-5-Chat, MedGemma |
| Geometric Information | ReXVQA | GPT-5-Chat, MedGemma |
| Negation Assessment | ReXVQA, Chest ImaGenome | GPT-5-Chat, MedGemma |
| Presence Assessment | ReXVQA, Chest ImaGenome, CXR-LT, NIH-CXR, Candid-PTX | ChexAgent, GPT-5-Chat, MedGemma |
| Location Assessment | ReXVQA, Chest ImaGenome | ChexAgent |
| Long-tail Abnormality Classification | Chest ImaGenome | ChexAgent |
| Tubes and Lines Presence | ReXVQA, Chest ImaGenome | - |
| Tubes and Lines Classification | Chest ImaGenome | - |
| Tubes and Lines Placement Description | MIMIC-CXR, ReXGradient, CheXpert-Plus | - |

not consistently included for all the records in the dataset. To mitigate this discrepancy, we employ the `Llama-3.3-70B` (Meta AI, 2024) model to automatically edit reports and remove explicit mentions of prior studies when prior images are not available.

**Visual grounding tasks** focuses on linking textual descriptions to specific image regions. It includes anatomical grounding, grounding of abnormalities - both phrases and labels and segmentation of both anatomical structures and tubes and lines.

**Visual Question Answering** Clinical VQA tasks focus on the assessment of presence, negation, and location of abnormalities as well as tubes and lines. They also encompass differential diagnosis generation, extraction of geometric information, long-tail abnormality classification, and the detection, classification, and placement description of tubes and lines.

**Causal Exploration** We leverage the causal exploration dataset, curated by radiologists from MIMIC-CXR reports and released as part of the Hidden-Rad Challenge - Hidden Causality Inclusion in Radiology Reports (Choi et al., 2025) for the causal exploration task. This task aims to enhance the interpretability of AI systems in radiology-related diagnostic reasoning by encouraging models to clearly state the rationale behind clinical interpretations. We reserve Hidden-Rad exclusively for evaluation. For training, we use a synthetic causal exploration dataset generated through few-shot prompt engineering with `Llama-3.3-70B` on MIMIC-CXR and ReXGradient, using prompts grounded in radiographic patterns, disease presentations, and clinical context. The prompt used is detailed in Appendix I. More details on the datasets can be found in Appendix D.

### 3.2. Model Architecture

The architecture of `Rad-Phi4-Vision-CXR` comprises a chest X-ray image encoder, a vision-language adapter, a language decoder, and modules for grounding and segmentation. For the image encoder, we use the pre-trained `SigLIP2-so400M-patch14-224` (Tschannen et al., 2025), a 400 million parameter model, which is further fine-tuned on CXR datasets to enhance radiology-specific understanding. The vision-language adapter is a two-layer MLP that aligns encoder embeddings with the language model. We leverage `Phi-4-mini-instruct` (Abdin et al., 2024) as the language decoder, a 3.8 billion parameter model. Figure 2 illustrates the architecture.

To support vision focused tasks, `Rad-Phi-4-Vision-CXR` incorporates two additional vision modules. The first is a lightweight detection module designed for grounding tasks. This module predicts both bounding box coordinates and confidence scores from the language model's hidden state embeddings. The hidden state embeddings from the language model are first projected into a lower-dimensional space through a feed-forward projector. From the projected embeddings, the embeddings corresponding to the `<DET>` token is extracted. From these the embeddings, two parallel heads operate: a bounding box regression head that outputs normalized coordinates for a fixed number of boxes, and a confidence head that assigns a confidence score to each box. Both heads are implemented as multi layer perceptrons with non-linear activations. The confidence score further indicates the reliability of the predicted coordinates. The second vision module is a segmentation module, implemented as a `2D U-Net` (Ronneberger et al., 2015). The encoder of the `U-Net` is initialized with `ResNet-34` weights pre-trained on ImageNet. The hidden states from the language model are

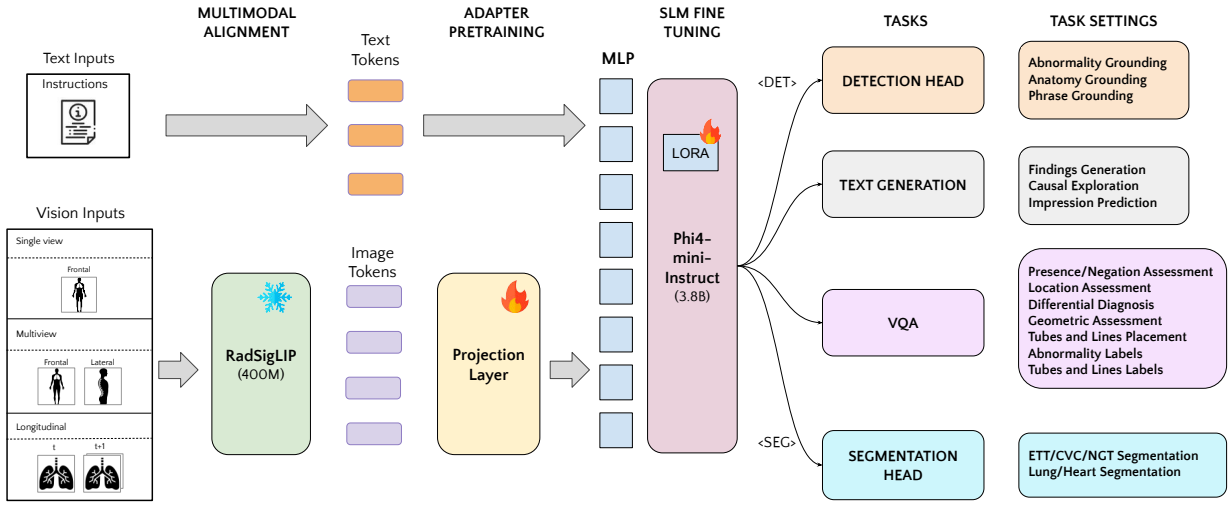



Figure 2: Rad-Phi4-Vision-CXR Architecture

projected into a lower-dimensional space and then the embedding corresponding to the `<SEG>` token is extracted. This gets integrated into the intermediate feature maps of the `U-Net` at the bottle-neck layer, thereby enriching the segmentation process with language-guided information.

The model is trained in three phases, first the image encoder is trained leveraging SigLip contrastive loss (Zhai et al., 2023). Then the vision-language adapter gets trained followed by the fine-tuning of the language model in LoRA (Hu et al., 2021) mode. Lastly, we train the segmentation module where we keep the language model and the encoder frozen, training only the `U-Net`. More details regarding the training can be found in Appendix A.1.

### 3.3. Experimental Design

We designed multiple experimental settings tailored to the different task types addressed in our work. For report generation task, we consider three settings. The first involves findings and impression generation without prior image and the second includes the prior image context. The third setting extends beyond direct report generation and explores causal explanation generation, where the model is expected to reason about potential causes underlying the findings. We evaluate on the ReXrank benchmark datasets. As baselines, we include `RadVLM`, `CheXagent`, `MedVersa` and `MedGemma`. The details of the settings can be found in Table 1. For causal exploration, we leverage the Hidden-Rad dataset which contains the ground truth causal explorations written by radiologists for MIMIC-CXR reports to benchmark the reasoning performance in comparison to `MedGemma` and `GPT-5-Chat`, as these models are known to perform strongly on reasoning-intensive tasks. We use a special LLM based evaluation design for measuring the reasoning performance leveraging the evaluation principles established in the Hidden-Rad challenge (Cho et al., 2024). More details of the evaluation design in Appendix E.

We investigate fine-grained visual question answering (VQA) across diverse tasks including image-level tasks (e.g., classifying images as normal/abnormal or detecting presence of devices like tubes and lines), disease-level and specific tube-level queries (e.g., querying specific diseases or tube types), long-tailed abnormality and tubes or lines classification, precise abnormality localization, and detailed tubes and lines placement description. Our approach handles negative scenarios, such as describing locations of absent diseases or non-existent tubes and lines. We benchmark against `CheXagent`, leveraging the Chest Imagenome dataset's long-tailed labels to evaluate abnormality presence and localisation. Additionally, we utilize the ReXVQA dataset, transforming its multiple-choice questions into an open-generation format for reasoning-driven responses, while retaining the closed MCQ setting for comparison with mod-

Table 4: Findings Generation Metrics - ReXrank Benchmark

| Dataset | Metric | MedVersa | RadVLM | MedGemma | CheXagent | Rad-Phi4-Vision-CXR (with priors) |
|---|---|---|---|---|---|---|
| CheXpert-Plus | BLEU-2 | 13.62 [12.74, 15.15] | 10.49 [9.21, 13.48] | 8.14 [7.65, 9.20] | 12.28 [12.31, 14.00] | **19.97 [17.04,22.59]** |
| | BertScore | 40.07 [38.73, 42.58] | 40.67 [39.28, 43.11] | 22.47 [20.90, 24.56] | 29.90 [29.96, 31.97] | **46.53 [42.93,50.57]** |
| | Rouge-L | 20.67 [20.00, 22.81] | 19.30 [17.99, 20.63] | 14.91 [14.63, 16.50] | 21.66 [20.74, 23.31] | **27.41 [25.08,29.52]** |
| | RadGraph-F1 | 21.43 [19.79, 22.99] | 21.14 [21.18, 22.68] | 12.50 [10.10, 14.40] | **28.3 [27.00, 29.20]** | 26.8 [24.9,28.8] |
| | CheXbert | 47.87 [43.74, 53.35] | 52.54 [45.63, 57.82] | 43.90 [38.50, 48.80] | 62.1 [58.8, 63.7] | **62.53[56.89,69.59]** |
| | GREEN | 20.34 | 21.44 | 13.30 | 25.70 | **30.02** |
| ReXGradient | BLEU-2 | 19.95 [19.84, 20.07] | 12.74 [12.58, 12.82] | 7.75 [7.68, 7.82] | 9.99 [9.89, 10.09] | **30.24 [29.75, 30.64]** |
| | BertScore | 47.40 [47.30, 47.54] | 41.28 [40.96, 41.29] | 26.61 [26.54, 26.73] | 29.18 [29.07, 29.31] | **52.20 [51.87, 52.54]** |
| | Rouge-L | 25.66 [25.68, 25.76] | 20.79 [20.58, 20.87] | 15.98 [15.90, 16.14] | 18.82 [18.63, 19.02] | **36.26 [35.81, 36.72]** |
| | RadGraph-F1 | 26.31 [26.11, 26.47] | 19.50 [19.34, 19.77] | 17.34 [17.27, 17.41] | 21.80 [21.50, 22.00] | **36.60 [35.90, 37.60]** |
| | CheXbert | 49.69 [49.79, 50.29] | 46.66 [46.34, 47.99] | 38.33 [37.79, 38.66] | 50.5 [50.2, 51.1] | **57.06 [56.29, 58.15]** |
| | GREEN | 47.01 | 35.31 | **48.06** | 34.50 | 47.83 |
| MIMIC | BLEU-2 | 15.49 [15.16, 15.72] | 7.79 [7.65, 8.10] | 9.49 [9.34, 9.65] | 14.30 [14.25, 14.48] | **22.38 [21.88, 22.40]** |
| | BertScore | 45.90 [45.75, 46.24] | 39.22 [39.02, 39.80] | 26.41 [26.03, 26.63] | 33.74 [33.60, 33.86] | **49.92 [49.48, 50.13]** |
| | Rouge-L | 23.68 [23.50, 23.85] | 17.99 [17.89, 18.36] | 16.10 [15.95, 16.19] | 22.84 [22.82, 23.04] | **29.50 [29.14, 29.69]** |
| | RadGraph-F1 | 24.36 [24.33, 25.03] | 16.76 [16.60, 16.84] | 12.49 [12.28, 12.65] | 25.80 [25.50, 26.10] | **27.70 [27.60, 28.50]** |
| | CheXbert | 48.51 [47.97, 49.56] | 41.29 [40.55, 41.54] | 44.46 [43.32, 45.02] | **56.6 [56.5, 57.3]** | 52.9 [52.70, 53.70] |
| | GREEN | 27.27 | 20.52 | 16.37 | 28.50 | **33.50** |

Table 5: Casual Exploration Performance on HiddenRad Dataset (evaluated by Llama 3.3)

| Model Name | Settings | Interpretation Accuracy (0–5) | Clinical Plausibility (0–5) | Completeness (0–5) | Evidence Linking (0–5) | Clarity & Conciseness (0–5) | Mean Score (0–25) |
|---|---|---|---|---|---|---|---|
| Rad-Phi4-Vision-CXR | **All Findings** | **3.28** | 3.70 | **2.74** | 3.38 | **3.89** | **16.99** |
| | **Matched Findings** | 4.02 | 4.54 | 3.36 | 4.14 | 4.77 | 20.84 |
| GPT-5-Chat | **All Findings** | 2.50 | 2.88 | 2.13 | 2.66 | 2.91 | 13.08 |
| | **Matched Findings** | 3.97 | 4.75 | 3.38 | 4.22 | 4.62 | 20.76 |
| MedGemma | **All Findings** | **3.28** | **3.71** | 2.73 | **3.39** | 3.88 | **16.99** |
| | **Matched Findings** | 4.04 | 4.56 | 3.35 | 4.17 | 4.77 | 20.88 |

els like `MedGemma` and `GPT-5-Chat`. Task details are outlined in Table 3.

For visual grounding tasks, we consider three settings: anatomical grounding, phrase grounding, and abnormality label grounding. In anatomical grounding, the goal is to localize 29 anatomical regions of the lungs using the Chest ImaGenome dataset. We compare with `RadVLM` as the baseline model.

In phrase grounding, the objective is to map medical phrases to image regions using the MS-CXR and PadChest-GR datasets, which provide phrase-level grounding annotations. We compare our approach against `RadVLM` and `AGPT-MDETR` (Zhang et al., 2025a), a phrase-grounding model built on the `MDETR` (Kamath et al., 2021) architecture, as baseline models. In abnormality grounding, the task is to localize abnormalities in chest X-rays using labels from PadChest-GR and NIH-CXR for benchmarking. NIH-CXR is used solely for evaluation, not training. Comparisons are made against `RadVLM` and `AGPT-MDETR` as baseline models. For segmentation tasks, we consider two settings: anatomical segmentation using the CheXmask dataset, benchmarked against `MedVersa`, and tubes-and-lines segmentation, focusing on endotracheal tubes (ETT), nasogastric tubes (NGT), and central venous catheters (CVC) using the RANZCR dataset, with no existing foundation model identified for comparison. Task details for grounding and segmentation are provided in Table 2.

## 4. Results

We evaluate `Rad-Phi4-Vision-CXR` on core radiology tasks, focusing on findings generation from chest X-rays and clinical indication. In Table 4, we report both lexical and clinical metrics. Overall, `Rad-Phi4-Vision-CXR` performs better than all the baseline models. Sample outputs for the findings generation task are provided in Appendix C.1. The

Table 6: ReXVQA Benchmark Results

| Task Settings | GPT-5-Chat | MedGemma | CheXagent | Rad-Phi4-Vision-CXR |
|---|---|---|---|---|
| Differential Diagnosis | 0.782 | 0.807 | 0.813 | **0.954** |
| Geometric Information | 0.725 | 0.649 | 0.673 | **0.833** |
| Negation Assessment | 0.888 | 0.906 | 0.843 | **0.988** |
| Presence Assessment | 0.755 | 0.785 | 0.768 | **0.972** |
| Location Assessment | 0.733 | 0.803 | 0.793 | **0.932** |
| **Overall Accuracy** | 0.808 | 0.835 | 0.806 | **0.969** |

Table 7: Visual Grounding Metrics

| Task Settings | Dataset | MedVersa | RadVLM | AGPT-MDETR | Rad-Phi4-Vision-CXR |
|---|---|---|---|---|---|
| Anatomical Grounding | Chest ImaGenome | - | mAP: 0.853<br>- | - | **mAP: 0.922**<br>**mIoU: 0.670** |
| Abnormality Grounding Phrases | MS-CXR | - | **mAP: 0.828**<br>mIoU: 0.531 | mAP: 0.814<br>**mIoU: 0.600** | mAP: 0.818<br>mIoU: 0.540 |
| | PadChest-GR | - | mAP: 0.443<br>mIoU: 0.288 | mAP: 0.161<br>mIoU: 0.141 | **mAP: 0.736**<br>**mIoU: 0.517** |
| Abnormality Grounding Labels | NIH-CXR | mAP: -<br>mIoU: 0.239 | **mAP: 0.578**<br>mIoU: 0.357 | mAP: 0.320<br>mIoU: 0.230 | mAP: 0.544<br>**mIoU: 0.363** |
| | PadChest-GR | -<br>- | mAP: 0.349<br>mIoU: 0.226 | mAP: 0.138<br>mIoU: 0.137 | **mAP: 0.675**<br>**mIoU: 0.460** |
| Anatomical Segmentation | CheXMask | DICE: 0.955<br>IoU: - | - | - | **DICE: 0.9773**<br>**IoU: 0.954** |
| Tubes & Lines Segmentation | RANZCR | - | - | - | **DICE: 0.790**<br>**IoU: 0.673** |

metrics for the impression generation task is provided in Table 9.

`Rad-Phi4-Vision-CXR` outperforms baseline models in both mAP and mIoU in the anatomical grounding task. For phrase grounding , it surpasses other models on the larger PadChest-GR dataset, while achieving metrics close to the state-of-the-art model on the MS-CXR dataset. In the abnormality grounding task, `Rad-Phi4-Vision-CXR` again outperforms baselines on the PadChest-GR dataset and close to the state-of-the-art model on NIH-CXR, and achieves a higher mIoU than others. The results are reported in Table 7. Few examples of the grounding task are given in Appendix C.6.

In the segmentation tasks, `Rad-Phi4-Vision-CXR` outperforms MedVersa on anatomical segmentation. For tubes and lines segmentation, no baseline comparisons are available in the LLM setup, so we directly report our results in Table 10 and provide a baseline with vision based unet setup in Appendix J.8. Few examples of segmentation task are given in Appendix C.5.

In visual question answering (VQA) tasks, `Rad-Phi4-Vision-CXR` outperforms the baseline model `CheXagent` across both long-tail abnormality classification and presence assessment. The corresponding metrics are reported in Table 8, and an example of the tubes and lines VQA task is provided in Appendix C.2. We further evaluate on the public ReXVQA test set, benchmarking against other baseline models. Rad-Phi4-Vision-CXR establishes superior performance, the results are summarized in Table 6. Sample outputs from the ReXVQA benchmark are included in Appendix C.4. The metrics for VQA tasks across other public datasets are summarized in Table 11.

We benchmark the causal exploration summary generation using the Hidden-Rad dataset (Choi et al., 2025) which contains causal exploration summary for 1219 radiology reports from the MIMIC dataset. We leverage the LLM based evaluation scheme defined in the Evaluation Design section in Appendix E. The LLM based evaluation is done using two models, `Llama 3.3` and `GPT-5-Chat`. The results are available in Table 5 and Table 12 respectively. As we do both findings generation and reasoning as a combined task. Both the ground truth reports and the model generated outputs consist of two structured components: a findings section and a causal exploration section. To ensure that the reasoning quality is assessed in a clinically meaningful context, we evaluate reasoning for correctly matched findings. If

a model predicts an incorrect finding, the reasoning associated with that incorrect prediction is also considered invalid and therefore penalized. `MedGemma` and `Rad-Phi4-Vision-CXR` had similar performance in this benchmark. `GPT-5-Chat` though had good reasoning capability, their overall score was low as the number of matched findings with the ground truth was less. We note that the results from both the LLM evaluations are very similar.

**Ablation Study** Our ablation studies highlight several key factors influencing model performance across tasks. First, synthetic data augmentation substantially improves causal exploration outcomes compared to both zero-shot and non-augmented training [J.1]. Second, sigmoid loss consistently outperforms standard contrastive loss for image-text retrieval across common and rare pathologies [J.2]. Third, specialized vision modules leveraging language embeddings deliver superior grounding accuracy for both abnormality labels and phrase-level grounding compared to direct LLM decoding [J.3] Fourth, Stage-1 pretraining with Chest ImaGenome VQA annotations enhances report generation metrics [J.4]. Fifth, LoRA-based fine-tuning yields better report generation performance than full fine-tuning [J.5]. Sixth, multi-task training combining VQA and report generation improves VQA accuracy relative to standalone VQA training [J.6]. Finally, incorporating prior images alongside the current image consistently boosts findings generation quality across all evaluation metrics [J.7].

**Consistency Analysis** We performed an LLM-based consistency evaluation for predictions across various tasks for the same record (example tasks as shown in Figure 1). Details of the consistency analysis evaluation design is available in Appendix E, and the results are presented in Table 23. The results indicate that 50% of records have a very high consistency score across tasks, another 47% of records have a consistency score of around 6 on a scale of 10, and only 3% of records have a lower score. Examples of a couple of such analyses are provided in Table 24. We concluded that radiology report generation and abnormality label classification are more reliable querying modes than abnormality-presence queries, as not all conditions have sufficient support in this query mode.

**Radiologist Evaluation** Radiologist evaluation of 208 model-predicted findings from 30 radiology reports showed that 92.31% of findings were correctly reported, with 4.80% false positives and 2.88% findings with inaccurate attributes, mainly in severity levels. However, 10.50% of clinically significant findings, such as nodular opacities and mediastinal widening, were missed, alongside 2.10% of insignificant findings like calcifications. The details of the evaluation is available in Appendix G. These results highlight the model's strong performance but underscore the need for improved detection of subtle but significant findings.

The qualitative evaluation of the causal exploration narrative found most causal explorations clinically sound, with 25 of 30 reports requiring no accuracy-related comments. The radiologist emphasized limiting causal reasoning for negative findings to cases where absence directly influences diagnosis and recommended incorporating more patient history or broader differentials when history is lacking. Comments about tubes and lines should be framed as suggestive rather than strictly causal, especially when describing complications or malposition. The details of the evaluation is available in Appendix G.

**GREEN Error Category Analysis** We analysed the model generated errors leveraging the GREEN metric (Ostmeier et al., 2024) across all baseline models. The results are available in Appendix K.1. We find that missing findings present in the reference report is the most dominant error, accounting for more than 60% of cases which concurs with the Radiologist findings as well.

**Compactness** `Rad-Phi4-Vision-CXR` requires approximately 16,358 MB of memory and achieves an average throughput of about 30 tokens per second. A comparison of memory usage and throughput with other models is provided in Table 25.

## 5. Conclusion

`Rad-Phi4-Vision-CXR` establishes a new benchmark in radiology AI by integrating report generation, causal exploration, fine-grained visual question answering, and precise detection and segmentation within a unified, compact framework. Future work will focus on addressing gaps related to missed findings, incorporating advanced reasoning capabilities to address cases that rely on measurements, and advancing self-verification workflows to further enhance reliability and diagnostic confidence.

## Acknowledgments

We thank Dr. R. Rajasekarapandian, Radiologist at RSP Speciality Hospital, Thirumayam, India, for reviewing the AI-generated radiology reports and causal exploration sections. His expert feedback and the time dedicated to this process greatly enhanced the clinical relevance of our work.

## References


Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. Phi-4 technical report, 2024. URL https://arxiv.org/abs/2412.08905.

Ayat Abedalla, Malak Abdullah, Mahmoud Al-Ayyoub, and Elhadj Benkhelifa. 2st-unet: 2-stage training model using u-net for pneumothorax segmentation in chest x-rays. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6. IEEE, 2020.

Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Pérez-García, Max Ilse, Daniel Coelho de Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anton Schwaighofer, et al. Mscxr-t: Learning to exploit temporal structure for biomedical vision-language processing, 2023.

Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, et al. Maira-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*, 2024.

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797, 2020.

Daniel C Castro, Aurelia Bustos, Shruthi Bannur, Stephanie L Hyland, Kenza Bouzid, Maria Teodora Wetscherek, Maria Dolores Sánchez-Valverde, Lara Jaques-Pérez, Lourdes Pérez-Rodríguez, Kenji Takeda, et al. Padchest-gr: A bilingual chest x-ray dataset for grounded radiology report generation. *arXiv preprint arXiv:2411.05085*, 2024.

Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P. Langlotz. Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats, 2024. URL https://arxiv.org/abs/2405.19538.

Zhihong Chen, Maya Varma, Justin Xu, Magdalini Paschali, Dave Van Veen, Andrew Johnston, Alaa Youssef, Louis Blankemeier, Christian Bluethgen, Stephan Altmayer, Jeya Maria Jose Valanarasu, Mohamed Siddig Eltayeb Muneer, Eduardo Pontes Reis, Joseph Paul Cohen, Cameron Olsen, Tanishq Mathew Abraham, Emily B. Tsai, Christopher F. Beaulieu, Jenia Jitsev, Sergios Gatidis, Jean-Benoit Delbrouck, Akshay S. Chaudhari, and Curtis P. Langlotz. A vision-language foundation model to enhance efficiency of chest x-ray interpretation, 2024. URL https://arxiv.org/abs/2401.12208.

Yousang Cho, Key-Sun Choi, and Hidden-Rad Organizing Committee. Hidden-rad evaluation scheme of ntcir-18, 2024. URL https://github.com/hidden-rad/Evaluation-Scheme-Experiment-.

Key-Sun Choi, Yousang Cho, and Hidden-Rad Organizing Committee. Overview of the ntcir-18 hidden-rad task: Hidden causality inclusion in radiology report generation. In *Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies*, 2025.

Jean-Benoit Delbrouck, Pierre Chambon, Zhihong Chen, Maya Varma, Andrew Johnston, Louis Blankemeier, Dave Van Veen, Tan Bui, Steven Truong, and Curtis Langlotz. Radgraph-xl: A large-scale expert-annotated dataset for entity and relation extraction from radiology reports. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12902–12915, 2024.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.

Nicolas Deperrois, Hidetoshi Matsuo, Samuel Ruipírez-Campillo, Moritz Vandenhirtz, Sonia Laguna, Alain Ryser, Koji Fujimoto, Mizuho Nishio, Thomas M. Sutter, Julia E. Vogt, Jonas Kluckert, Thomas Frauenfelder, Christian Blüthgen,

Farhad Nooralahzadeh, and Michael Krauthammer. Radvlm: A multitask conversational vision-language model for radiology. *arXiv preprint arXiv:2502.03333*, 2025. URL https://arxiv.org/abs/2502.03333.

Sijing Feng, Damian Azzollini, Ji Soo Kim, Cheng-Kai Jin, Simon P Gordon, Jason Yeoh, Eve Kim, Mina Han, Andrew Lee, Aakash Patel, et al. Curation of the candid-ptx dataset with free-text reports. *Radiology: Artificial Intelligence*, 3(6): e210136, 2021.

N. Gaggion, C. Mosquera, M. Aineseder, L. Mansilla, D. Milone, and E. Ferrante. CheXmask Database: a large-scale dataset of anatomical segmentation masks for chest x-ray images (version 0.1). PhysioNet, 2023. https://doi.org/10.13026/dx54-8351.

Gregory Holste, Song Wang, Ajay Jaiswal, Yuzhe Yang, Mingquan Lin, Yifan Peng, and Atlas Wang. Cxr-lt: Multi-label long-tailed classification on chest x-rays. *PhysioNet*, 5:19, 2023.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1780–1790, 2021.

Sai Kishan Kapavarapu. Nih-chest-x-rays-multi-label-image-classification.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

Meta AI. Llama 3.3 70b: Model card. https://github.com/meta-llama/llama-models/blob/main/models/llama3_3/MODEL_CARD.md, 2024. Accessed: 2025-09-07.

Arun Nair, Wilson Ong, Aric Lee, Naomi Wenxin Leow, Andrew Makmur, Yong Han Ting, You Jun Lee, Shao Jin Ong, Jonathan Jiong Hao Tan, Naresh Kumar, and James Thomas Patrick Decourcy Hallinan. Enhancing radiologist productivity with artificial intelligence in magnetic resonance imaging (mri): A narrative review. *Diagnostics*, 15(9):1146, 2025. doi: 10.3390/diagnostics15091146. URL https://www.mdpi.com/2075-4418/15/9/1146.

Ha Q Nguyen, Khanh Lam, Linh T Le, Hieu H Pham, Dat Q Tran, Dung B Nguyen, Dung D Le, Chi M Pham, Hang TT Tong, Diep H Dinh, et al. Vindr-cxr: An open dataset of chest x-rays with radiologists' annotations. *Scientific Data*, 9(1):429, 2022.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, et al. Green: Generative radiology report evaluation and error notation. *arXiv preprint arXiv:2405.03595*, 2024.

Ankit Pal, Jung-Oh Lee, Xiaoman Zhang, Malaikannan Sankarasubbu, Seunghyeon Roh, Won Jung Kim, Meesun Lee, and Pranav Rajpurkar. Rexvqa: A large-scale visual question answering benchmark for generalist chest x-ray understanding, 2025. URL https://arxiv.org/abs/2506.04353.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

Eduardo P Reis, Joselisa PQ de Paiva, Maria CB da Silva, Guilherme AS Ribeiro, Victor F Paiva, Lucas Bulgarelli, Henrique MH Lee, Paulo V Santos, Vanessa M Brito, Lucas TW Amaral, et al. Brax, brazilian labeled chest x-ray dataset. *Scientific Data*, 9(1):487, 2022.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. URL https://arxiv.org/abs/1505.04597.

Jarrel Seah, Jen, Maggie, Meng Law, Phil Culliton, and Sarah Dowd. Ranzcr clip - catheter and line position challenge. https://kaggle.com/competitions/ranzcr-clip-catheter-line-classification, 2020. Kaggle.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren. Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert. *arXiv preprint arXiv:2004.09167*, 2020.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. URL https://arxiv.org/abs/2502.14786.

Maria De La Iglesia Vayá, Jose Manuel Saborit, Joaquim Angel Montell, Antonio Pertusa, Aurelia Bustos, Miguel Cazorla, Joaquin Galant, Xavier Barber, Domingo Orozco-Beltrán, Francisco García-García, et al. Bimcv covid-19+: a large annotated dataset of rx and ct images from covid-19 patients. *arXiv preprint arXiv:2006.01174*, 2020.

Joy T Wu, Nkechinyere N Agu, Ismini Lourentzou, Arjun Sharma, Joseph A Paguio, Jasper S Yao, Edward C Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, et al. Chest imagenome dataset for clinical reasoning. *arXiv preprint arXiv:2108.00316*, 2021.

Justin Xu, Zhihong Chen, Andrew Johnston, Louis Blankemeier, Maya Varma, Jason Hom, William J. Collins, Ankit Modi, Robert Lloyd, Benjamin Hopkins, Curtis Langlotz, and Jean-Benoit Delbrouck. Overview of the first shared task on clinical text generation: RRG24 and "discharge me!". In Dina Demner-Fushman, Sophia Ananiadou, Makoto Miwa, Kirk Roberts, and Junichi Tsujii, editors, *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 85–98, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.bionlp-1.7. URL https://aclanthology.org/2024.bionlp-1.7/.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

Wenjun Zhang, Shakes Chandra, and Aaron Nicolson. Anatomical grounding pre-training for medical phrase grounding, 2025a. URL https://arxiv.org/abs/2502.16585.

Xiaoman Zhang, Julián N. Acosta, Josh Miller, Ouwen Huang, and Pranav Rajpurkar. Rexgradient-160k: A large-scale publicly available dataset of chest radiographs with free-text reports. In *arXiv:2505.00228v1*, 2025b.

Hong-Yu Zhou, Subathra Adithan, Julián N. Acosta, Eric J. Topol, and Pranav Rajpurkar. A generalist learner for multifaceted medical image interpretation. *arXiv preprint arXiv:2405.07988*, 2024. URL https://arxiv.org/abs/2405.07988.

# Appendix A. Experimental Setup

## A.1. Training Details

The training of `Rad-Phi-4-Vision-CXR` is conducted in multiple stages. In the first stage, we fine-tune the general-domain image encoder, `SigLIP-so400M-patch14-224` on CXR-specific datasets. We use paired chest X-rays and reports from Interpret-CXR and CheXpert Plus datasets to align image embeddings with textual report sentences. The training is performed in a contrastive learning setup using the sigmoid-based SigLIP loss (Zhai et al., 2023).

In the second stage, we train the vision-language adapter and the detection module and the rest of the model remains frozen. In the third stage, we fine-tune `Rad-Phi-4-Vision-CXR` using low-rank adaptation (LoRA) (Hu et al., 2021). The image encoder is frozen, while the vision-language adapter and the detection module remain trainable. Additionally, we enable training of the last layer of the language model to accommodate task-specific tokens. In the final stage, we train the segmentation module (2D U-Net) independently, while keeping all other components frozen. We observed that training the U-Net jointly with other components led to suboptimal performance, making isolated training essential.

In the second stage, we train the model for one epoch with a learning rate of 1e-4 and in the third stage it was trained for three epochs with a learning rate of 5e-4. In both the stages the batch size was maintained at 128. In the last phase, the U-Net is trained for 100 epochs with a batch size of 64 and a learning rate of 1e-4. In all the stages we use AdamW optimizer with cosine learning rate scheduling and linear warm-up.

We use two special tokens `<DET>` and `<SEG>` to help determine that a detection or segmentation task is involved. Also, the hidden state embeddings corresponding to these tokens are passed to the respective vision modules. This helps in the implementation of the grounding and segmentation tasks.

During training, we employ a combination of task-specific loss functions. For language model training, we use the standard Causal Language Modeling (CausalLM) loss, which is commonly adopted in auto-regressive language models.

For grounding tasks, we adopt a weighted multi-loss setup. Bounding box prediction is optimized using a combination of Generalized IoU (GIoU) loss (Rezatofighi et al., 2019), Mean IoU (mIoU) loss, and L1 regression loss, which together ensure accurate localization. We use focal loss to determine if the box qualifies for the phrase or label. For the segmentation task, we combine Dice loss, which measures the overlap between predicted and ground-truth masks, with focal loss.

# Appendix B. Evaluation Benchmarks

## B.1. Visual Question Answering Benchmark

Table 8: Chest Imagenome VQA Closed Task Metrics

| Task | ChexAgent | Rad-Phi4-Vision-CXR |
|---|---|---|
| **Long-tail Abnormality Classification** | F1: 73.51 [73.32, 73.94]<br>Precision: 71.00 [70.86, 71.47]<br>Recall: 76.21 [75.79, 76.60] | F1: **76.68** [76.22, 76.98]<br>Precision: **74.93** [74.39, 75.28]<br>Recall: **78.5** [78.1, 78.9] |
| **Presence Assessment** | F1: 0.879 [0.8779, 0.8812]<br>Precision: 0.879 [0.8779, 0.8812]<br>Recall: 0.879 [0.8779, 0.8812] | F1: **0.895** [0.893, 0.898]<br>Precision: **0.894** [0.89, 0.898]<br>Recall: **0.895** [0.891, 0.898] |
| **Tubes and Lines Classification** | – | F1: **87.54** [86.86, 88.16]<br>Precision: **87.17** [86.5, 87.78]<br>Recall: **87.9** [87.25, 88.5] |
| **Tubes and Lines Presence** | – | F1: **0.91** [0.90, 0.91]<br>Precision: **0.91** [0.90, 0.915]<br>Recall: **0.90** [0.90, 0.914] |

### B.2. Impression Metrics

Table 9: Impression Generation Metrics (with Priors)

| Dataset | BLEU-2 | BertScore | RadGraph | GREEN | ROUGE-L | CheXbert Micro-F1-14 |
|---|---|---|---|---|---|---|
| CheXpert-Plus | 13.42 [11.12, 15.08] | 50.68 [48.79, 52.44] | 0.2696 [0.2530, 0.2897] | 0.3217 | 32.31 [30.57, 34.08] | 0.5846 [0.5543, 0.6166] |
| MIMIC | 22.04 [20.45, 23.83] | 47.02 [45.49, 48.72] | 0.2890 [0.2703, 0.3042] | 0.3292 | 31.81 [30.48, 33.39] | 0.5405 [0.5167, 0.5564] |
| ReXGradient | 29.15 [28.78, 29.60] | 52.39 [52.16, 52.69] | 0.3283 [0.3176, 0.3364] | 0.4822 | 38.74 [38.43, 39.31] | 0.6061 [0.5974, 0.6158] |

### B.3. RANZCR Tubes and Lines Segmentation Benchmark

Table 10: Segmentation Performance Metrics (DICE Scores)

| Task | RadPhi4-Vision-CXR | MedVersa |
|---|---|---|
| **Anatomical Segmentation** | | |
| Heart | 0.965 | – |
| Left Lung | 0.979 | – |
| Right Lung | 0.984 | – |
| **Mean Anatomical** | **0.976** | 0.955 |
| **Tubes and Lines** | | |
| ETT | 0.800 | – |
| CVC | 0.791 | – |
| NGT | 0.782 | – |
| **Mean Tubes/Lines** | **0.791** | – |

### B.4. VQA Benchmark - Closed

Table 11: Closed VQA Benchmarks

| Dataset | Task | F1-Score | Precision | Recall |
|---|---|---|---|---|
| Brax | View Classification | 0.994 | 0.994 | 0.995 |
| CXR-LT | Abnormality Presence | 0.880 | 0.882 | 0.883 |
| NIH-CXR | Abnormality Presence | 0.796 | 0.796 | 0.796 |
| Candid-PTX | Abnormality Presence | 0.910 | 0.912 | 0.912 |

### B.5. HiddenRad Benchmark

Table 12: Casual Exploration Performance on HiddenRad Dataset (evaluated by GPT-5)

| Model Name | Settings | Interpretation Accuracy (0–5) | Clinical Plausibility (0–5) | Completeness (0–5) | Evidence Linking (0–5) | Clarity & Conciseness (0–5) | Mean Score (0–25) |
|---|---|---|---|---|---|---|---|
| Rad-Phi4-Vision-CXR | **Total Findings** | **3.28** | **3.71** | **2.74** | **3.39** | **3.90** | **17.01** |
| | **Matched Findings** | 4.02 | 4.54 | 3.35 | 4.15 | 4.78 | 20.84 |
| GPT-5-Chat | **Total Findings** | 2.32 | 2.66 | 1.99 | 2.47 | 2.73 | 12.17 |
| | **Matched Findings** | 3.87 | 4.45 | 3.32 | 4.13 | 4.57 | 20.34 |
| MedGemma | **Total Findings** | 3.09 | 3.45 | 2.60 | 3.23 | 3.64 | 16.01 |
| | **Matched Findings** | 4.07 | 4.54 | 3.42 | 4.24 | 4.78 | 21.05 |

## Appendix C. Rad-Phi4-Vision-CXR: Qualitative Results

### C.1. Findings Generation

Table 13: Report Generation + VQA.

| | **Ground Truth** | **Prediction** |
|---|---|---|
| Indication | Distended abdomen. | |
| Findings | There are **low lung volumes**. The **heart size is normal**. The **aorta remains tortuous**. There is **crowding of the bronchovascular structures**, but no overt pulmonary edema. **Linear opacities at the lung bases** likely reflect **atelectasis**. Possible trace left pleural effusion is present. **No pneumothorax**.**No free air** under the diaphragms. There is gaseous distention of the stomach. | The **lung volumes** are low. **The heart size is normal**. The **aorta is tortuous** and diffusely calcified. There is **crowding of the bronchovascular structures**. **Streaky opacities in the lung bases** likely **reflect atelectasis**. There is **no large pleural effusion or pneumothorax**. There is **no free air**. |
| Reasoning | The absence of **free air under the diaphragms rules out pneumoperitoneum** as a cause of abdominal distension. **Low lung volumes with bibasilar atelectasis are likely secondary to diaphragmatic elevation from abdominal distension**. The probable small left pleural effusion may be reactive or related to underlying systemic or abdominal processes. Gaseous distention of the stomach is consistent with the clinical presentation of abdominal distension. | The chest radiograph shows **low lung volumes with bibasilar streaky opacities consistent with atelectasis**, likely due to hypoventilation from abdominal distension. There is **no evidence of free intraperitoneal air, effectively ruling out pneumoperitoneum as a cause of the distended abdomen**. The normal heart size and absence of large pleural effusion or pneumothorax indicate no acute cardiopulmonary abnormality contributing to the presentation. |
| Abnormality List | **vascular congestion, lung opacity**, pleural effusion, tortuous aorta, linear/patchy atelectasis | **atelectasis, lung opacity, vascular congestion** |
| In which location is the atelectasis seen? | **bibasilar in the base of both the lungs** | **bibasilar in the base of both the lungs** |
| Is there evidence of any abnormalities? | **YES** | **YES** |
| Is there any indication of atelectasis in this image or chest x-ray? | **YES** | **YES** |
| Is there any indication of pneumothorax in this image or chest x-ray? | **NO** | **NO** |

## C.2. Tubes and Lines Generation

Table 14: Tubes and Lines Findings Generation + VQA.

| | **Ground Truth** | **Prediction** |
|---|---|---|
| Indication | ____F h/o CKD ____ diabetic nephropathy s/p DDRT in ____ p/w abdominal pain x1d, N/V, CT concerning for R-sided? ischemic colitis vs. typhlitis now desating/sob. Assess for pulmonary edema. | |
| Findings | **Right upper and lower lobe opacities are new since the prior day, with indistinctness of the pulmonary vessels, suggesting pulmonary edema**. However, concurrent pneumonia cannot be excluded, in the correct clinical setting. The right **PICC line terminates in the lower SVC**, and the **ET tube terminates 4.5 cm above the carina**. Unchanged calcified pulmonary granulomas in the left lung. **No pneumothorax**. Stable cardiomediastinal borders. | The **endotracheal tube terminates 4.5 cm above the carina**. A right **PICC terminates in the mid SVC**. The heart is mildly enlarged. There is **mild pulmonary edema**. **No pleural effusion or pneumothorax**. |
| Reasoning | **The new right upper and lower lobe opacities with vascular indistinctness are most consistent with pulmonary edema, correlating with the patient's acute desaturation and shortness of breath**. Given the focal right-sided distribution, superimposed pneumonia remains a consideration, especially in the context of recent abdominal infection risk (possible ischemic colitis/typhlitis) and immunosuppression post renal transplant. The absence of cardiomegaly and the acute onset suggest that non-cardiogenic causes of pulmonary edema (e.g., sepsis-related capillary leak) are possible. | The chest radiograph demonstrates **mild pulmonary edema, evidenced by interstitial prominence and vascular congestion, which correlates with the patient's desaturation and shortness of breath**. The mild cardiomegaly suggests a possible cardiac contribution, such as volume overload or early heart failure, which is common in patients with CKD and recent critical illness. No pleural effusion or pneumothorax is present, ruling out other acute pulmonary causes of hypoxia. |
| Tubes and Lines List | **endotracheal tube,picc** | **endotracheal tube,picc** |
| In which location is the ET Tube seen? | The ET tube **terminates 4.5 cm above the carina**. | The endotracheal tube **terminates 4.5 cm above the carina.** |
| Is there any presence of tubes, lines or devices visible in the image? | **YES** | **YES** |
| Is there any indication of endotracheal tube in this image? | **YES** | **YES** |
| Is there any indication of NG Tube in this image or chest x-ray? | **NO** | **NO** |

## C.3. Impact of Causal Reasoning on Findings Generation

Table 15: Improvements in Findings Generation with Causal Exploration

| | Ground Truth | Prediction(without Reasoning) | Prediction(with Reasoning) |
|---|---|---|---|
| **Example 1** | Interval removal of the left chest tube. A possible **left tiny pneumothorax** is seen. There is stable focal pleural thickening in the left upper lobe. There is a **small left pleural effusion** that has decreased. The **left basal opacity** has also decreased. A small amount of **subcutaneous emphysema** in the left chest wall. **The right lung remains clear**. | There is a **small left apical pneumothorax**. The cardiomediastinal silhouette is normal. There is a **left basilar opacity.** There is no right pleural effusion. There is no right pneumothorax. There is a **right lung base opacity**. There is a **small left pleural effusion**. The **pulmonary vasculature is normal** | The left chest tube has been removed. There is a **small left apical pneumothorax**. There is a **small left pleural effusion.** There is a small amount of **subcutaneous emphysema** along the left lateral chest wall. There is a small amount of **left basilar atelectasis. The right lung is clear**. The cardiomediastinal silhouette is within normal limits. |
| **Example 2** | There is **moderate cardiomegaly** which is new since ____. **Central pulmonary vessels are engorged**, and there is mild interstitial edema with a **large right pleural effusion**. A **trace left pleural effusion** is also present. There is no pneumothorax. | There is a **small left pleural effusion**. There is a **large right pleural effusion**. | The **heart is mildly enlarged**. There is a **moderate right pleural effusion** with overlying atelectasis. There is also a **small left pleural effusion**. There is mild **pulmonary vascular congestion**. |

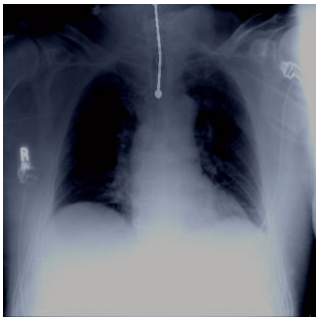
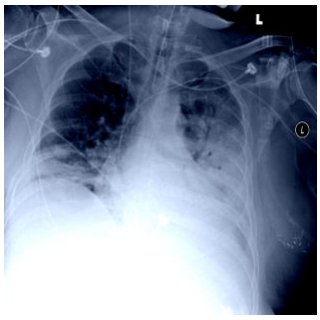
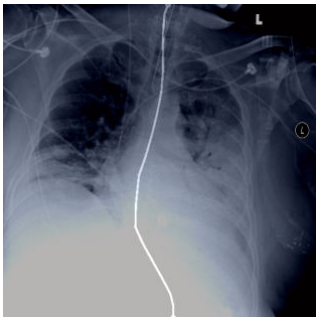
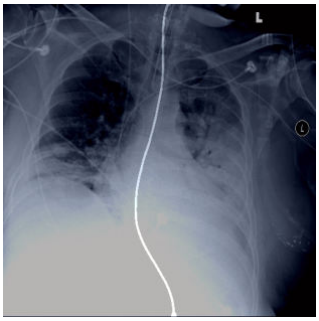
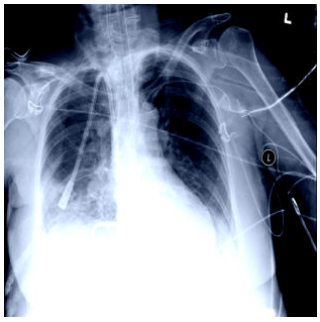
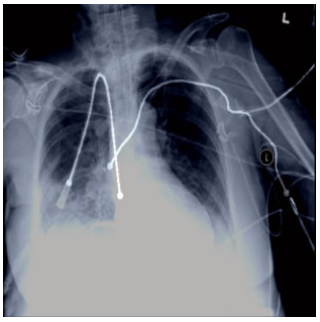
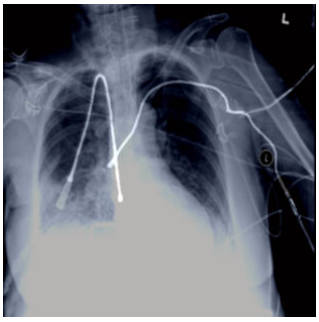
## C.4. ReXrank Visual Question Answering (VQA)

Table 16: Sample ReXVQA Outputs

| Question Type | Question | Predicted Answer | Actual Answer |
|---|---|---|---|
| **Geometric Information** | What is the position of the endotracheal tube as seen on the chest X-ray? | Answer: **B** – Endotracheal tube positioned **3.7 cm above the carina**, correct placement for ventilation safety. | Answer: **B** – Endotracheal tube positioned **3.7 cm above the carina**, appropriate placement. |
| **Location and Distribution** | Which ribs are affected by fractures on the left side as seen in the chest X-ray? | Answer: **B** – Fractures of the **fourth through seventh ribs on the left side**. | Answer: **B** – Mildly displaced fractures of the **fourth through seventh ribs** on the **left lateral** side. |
| **Negation Assessment** | What is the status of the visualized skeletal structures on this chest X-ray? | Answer: **B** – Skeletal structures **unremarkable**, indicating no abnormalities. | Answer: **B** – Skeletal structures **unremarkable**, indicating no acute abnormalities. |
| **Presence Assessment** | What lung finding is present on the chest X-ray? | Answer: **C** – **Bibasilar atelectasis**, collapse of lung tissue at the bases. | Answer: **C** – **Bibasilar atelectasis**, minimal, collapse of the lung bases. |
| **Differential Diagnosis** | Based on the chest X-ray images, what is the most likely diagnosis? | Answer: **D** – **No active cardiopulmonary disease**, no pneumonia, pulmonary edema, or other active disease. | Answer: **D** – **No active cardiopulmonary disease**, such as pneumonia or pulmonary edema. |

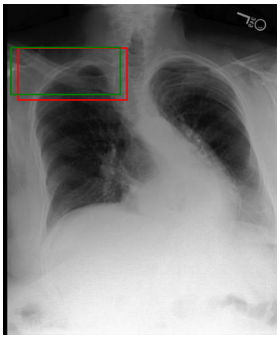
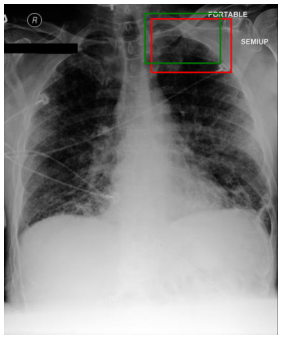
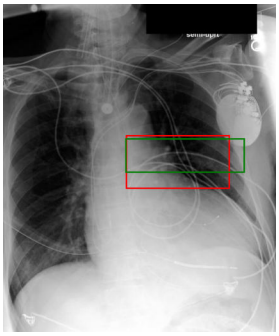
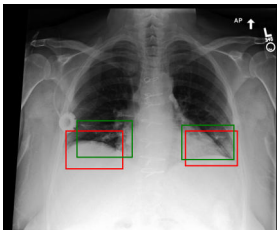
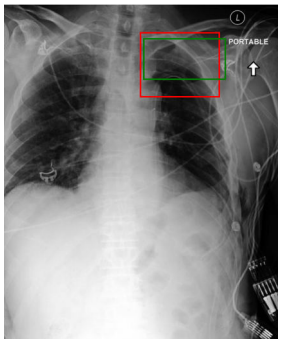
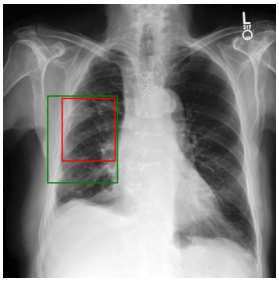
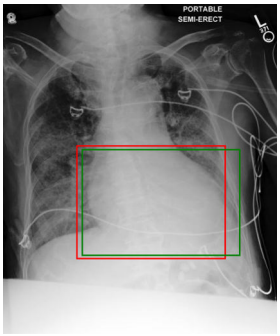
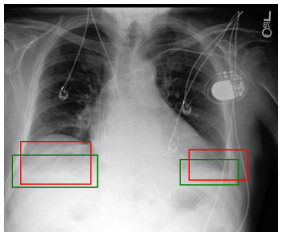
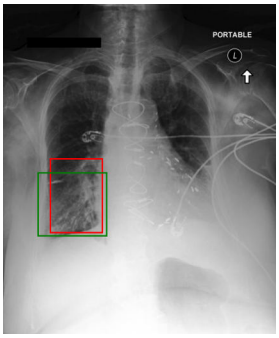
### C.5. Tubes and Lines Segmentation Tasks

Table 17: Segmentation Task Visualization: Original Image, Ground Truth Mask & Predicted Mask.

| Task | Original | GT Mask | Predicted Mask |
|---|---|---|---|
| **ET Tube Segmentation** |  |  |  |
| **NG Tube Segmentation** |  |  |  |
| **Central Venous Catheter (CVC) Segmentation** |  |  |  |

### C.6. Visual Grounding Tasks

Table 18: Grounding Task Visualization with Confidence Scores

| Task | Right Collarbone [0.91] | Left Apical Zone [0.94] | Left Central Lung Region [0.37] |
|---|---|---|---|
| Anatomical Grounding |  |  |  |
| | **Atelectasis [L: 0.61, R: 0.57]** | **Pneumothorax [0.47]** | **Pneumonia [0.42]** |
| Abnormality Label Grounding |  |  |  |
| | **Cardiac Silhouette is Enlarged [0.51]** | **Bilateral Pleural Effusion [L:0.46, R:0.40]** | **Patchy Right Infrahilar Opacity [0.42]** |
| Phrase Grounding |  |  |  |

# Appendix D. Instruction Dataset Overview

## D.1. Instruction Dataset Description

For the findings and impression generation task, we utilize three primary resources: Interpret-CXR (Xu et al., 2024), a large-scale benchmark that integrates multiple datasets including MIMIC-CXR (Johnson et al., 2019), CheXpert (Irvin et al., 2019), PadChest(Bustos et al., 2020), BIMCV-COVID19 (Vayá et al., 2020) and OpenI(Demner-Fushman et al., 2016). Other than that, we also used CheXpert-Plus (Chambon et al., 2024) and ReXGradient (Zhang et al., 2025b) datasets.

For the segmentation task, we primarily focus on segmenting ETT, CVC, and NGT tubes using the RANZCR dataset (Seah et al., 2020). For anatomical segmentation, we use the CheXmask dataset (Gaggion et al., 2023), which targets three anatomical structures: the heart, left lung, and right lung. The Chest ImaGenome dataset (Wu et al., 2021) is employed for the anatomical grounding task. For phrase grounding, we utilize the MS-CXR dataset (Bannur et al., 2023) and the PadChest-GR dataset (Castro et al., 2024), while for abnormality grounding we additionally include the VinDR-CXR dataset (Nguyen et al., 2022).

For the visual question answering task, we leverage ReXVQA (Pal et al., 2025) and Chest ImaGenome (Wu et al., 2021) datasets. We also utilized some label datasets like, CXR-LT (Holste et al., 2023), SIIM-ACR (Abedalla et al., 2020), NIH-CXR (Kapavarapu), VinDR-CXR (Nguyen et al., 2022), Brax (Reis et al., 2022) and Candid-PTX (Feng et al., 2021) to create samples for abnormality presence questions. The train/test distribution of the datasets for different tasks is given in Table 19.

# Appendix E. Evaluation Overview

## E.1. Evaluation Design

**Findings and Impression generation** For findings and impression generation, we use a combination of lexical and clinical evaluation metrics. The lexical metrics include BLEU-2 (Papineni et al., 2002), ROUGE-L (Lin, 2004), and BERTScore (Zhang et al., 2020). The clinical metrics consist of RadGraph-F1 (Delbrouck et al., 2024), GREEN (Ostmeier et al., 2024), and CheXbert (Smit et al., 2020).

**Visual Question Answering (VQA)** For visual question answering (VQA), we distinguish between open-ended and closed settings. In the open-ended setting, where the model generates free-text responses, we report the same lexical metrics as in report generation: BLEU-2, ROUGE-L, and BERTScore. In the closed setting, we evaluate performance using precision, recall, and F1-score. To evaluate model performance across different experimental settings, we employ task-specific metrics.

**Grounding Tasks** We report both the mean Average Precision (mAP) and mean Intersection-over-Union (mIoU) to capture both detection accuracy and spatial localization quality. Finally, for segmentation tasks, we use the Dice Score, which measures the overlap between predicted and ground-truth masks.

**Causal Exploration** We follow an LLM (`Llama-3.3-70B-Instruct` (Meta AI, 2024)) based evaluation framework leveraging the evaluation principles established in the Hidden-Rad challenge evaluation design (Cho et al., 2024) and extending it further. The LLM-based evaluation framework assesses predicted radiology reports against ground-truth reports by matching clinically equivalent findings. For matched findings, predicted reasoning is scored from 1 to 5 with justifications on: *Accuracy of Interpretation*, ensuring correct analysis of all salient radiographic features; *Clinical Plausibility*, verifying appropriate and contextually relevant differential diagnoses; *Completeness*, confirming all findings and differentials are addressed; *Evidence Linking*, checking explicit connections between assertions and findings; and *Clarity and Conciseness*, evaluating precise, non-redundant language. The evaluation prompt is available in Appendix I. We got similar evaluation results on evaluation with `Llama-3.3-70B-Instruct` and `GPT-5-Chat`.

**Consistency Evaluation for Multi-Task Radiology Prediction Models** In the development of multi-task AI models for radiology, which simultaneously generate narrative reports, detect abnormalities, and identify tubes or lines, ensuring consistency across these tasks is critical for clinical reliability. Consistent

Table 19: Instruction Dataset Overview

| Task | Training Datasets | Sample Count | Test Datasets | Sample Count |
|---|---|---|---|---|
| Findings Generation | Interpret-CXR | 217,268 | MIMIC-CXR | 2,345 |
| | CheXpert-Plus | 49,600 | CheXpert-Plus | 62 |
| | ReXGradient | 99,425 | ReXGradient | 6,271 |
| Impression Generation | Interpret-CXR | 396,111 | MIMIC-CXR | 2,345 |
| | CheXpert-Plus | 190,747 | CheXpert-Plus | 200 |
| | ReXGradient | 99,425 | ReXGradient | 6,271 |
| Abnormality Classification | Chest ImaGenome | 236,414 | Chest ImaGenome | 3,386 |
| Abnormality Presence | Chest ImaGenome | 1,041,022 | Chest ImaGenome | 19,606 |
| | ReXVQA | 200,437 | ReXVQA | 14,602 |
| | NIH-CXR | 89,008 | NIH-CXR | 49,976 |
| | SIIM | 10,712 | SIIM | 1,377 |
| | Candid-PTX | 12,476 | Candid-PTX | 4,158 |
| | CXR-LT | 173,955 | | |
| Abnormality Negation | Chest ImaGenome | 52,229 | ReXVQA | 14,918 |
| | ReXVQA | 191,890 | | |
| Abnormality Location | Chest ImaGenome | 1,383,559 | Chest ImaGenome | 23,785 |
| | ReXVQA | 33,196 | ReXVQA | 2,394 |
| Geometric Information Assessment | ReXVQA | 2,418 | ReXVQA | 168 |
| Tubes and Lines Presence | Chest ImaGenome | 702,181 | Chest ImaGenome | 11,145 |
| Tubes and Lines Classification | Chest ImaGenome | 236,414 | Chest ImaGenome | 11,145 |
| Tubes and Lines Placement Description | Interpret-CXR | 138,726 | Interpret-CXR | 1,331 |
| | CheXpert-Plus | 32,278 | CheXpert-Plus | 33 |
| | ReXGradient | 27,757 | | |
| Anatomical Grounding | Chest ImaGenome | 166,496 | Chest ImaGenome | 47,388 |
| Phrase Grounding | MS-CXR | 815 | MS-CXR | 176 |
| | PadChest-GR | 4,335 | PadChest-GR | 1,238 |
| Abnormality Grounding | MS-CXR | 800 | NIH-CXR | 984 |
| | PadChest-GR | 4,499 | PadChest-GR | 1,279 |
| | VinDR-CXR | 16,089 | | |
| Anatomical Segmentation | CheXmask | 42,297 | CheXmask | 8,364 |
| Tubes and Lines Segmentation | RANZCR | 34,293 | RANZCR | 3,810 |
| Causal Exploration | MIMIC-CXR | 102,819 | Hidden-RAD | 1084 |
| | ReXGradient | 106,137 | | |

predictions across tasks enhance the model's dependability, making it suitable for integration into clinical workflows. Furthermore, this evaluation enables the design of a robust inference pipeline that filters predictions based on consistency and minimizes the risk of contradictory findings reaching clinicians, improving patient safety and supporting the adoption of AI in high-stakes medical applications.

The LLM (`Llama-3.3-70B-Instruct`) based evaluation mechanism assesses consistency across multiple prediction components in a radiology record. It identifies explicit contradictions, such as a narrative report denying a finding present in the exhaustive lists or query responses. It produces a JSON output with identified contradictions, a consistency score (0–10, where 10 indicates no contradictions), and detailed reasoning. The evaluation prompt is available in Appendix I.

## E.2. Evaluation Prompts

Table 20: Prompt templates for findings and impression generation tasks.

| Task | Task Setting | ChestImagenome Prompt |
|---|---|---|
| **Report Generation** | Findings Generation | Given the chest X-rays and the indication section, write the findings. |
| | Impression Generation | Given the chest X-rays and the indication section, write the impression. |
| | Findings + Reasoning Generation | Given the chest X-rays and the indication section, write the findings with reasoning. |

Table 21: Prompt templates for visual grounding and segmentation tasks.

| Task Setting | ChestImagenome Prompt |
|---|---|
| **Visual Grounding** | Ground the location of {label / phrase / region} in the chest X-ray. |
| **Segmentation** | Segment the region of {label / tube}. |

Table 22: Prompt templates for VQA tasks.

| Task | Task Setting | Chestimagenome Prompt | ReXVQA Prompt |
|---|---|---|---|
| **VQA** | Abnormality Presence (image-level) | Is there evidence of any abnormalities? | Does the image show any abnormalities? |
| | Abnormality Presence (by-finding) | Is <> seen in the image? | Is <> present in the image? |
| | Abnormality Location | In which location is the <> seen? | Where can the <> be observed? |
| | Abnormality Classes | List the abnormalities seen in the image. | What findings can be identified in this image? |
| | Tubes/Lines Presence (image-level) | Is there any presence of tubes, lines, or devices visible in the chest X-ray? | Are there any medical devices present in this chest X-ray? |
| | Tubes/Lines Presence (tube-type) | Is there any indication of an <> in this chest X-ray? | Is a <> visible in this chest X-ray? |
| | Tubes/Lines Location | In which location is the <> seen? | Where is the <> located in the chest X-ray? |
| | Tubes/Lines Type | Could you please describe the types of tubes, lines or devices displayed in this chest X-ray? | Which types of medical devices are visible? |
| | Tubes/Lines Placement Description | Write the placement description of tubes, lines or devices if seen in the provided chest X-ray. | Where are the tubes and lines located in the chest X-ray? |

## Appendix F. Consistency Analysis Results

Table 23: Rad-Phi4-Vision-CXR Cross-task Consistency Analysis Score

| Values | Low(0-3) | Medium(4-7) | High(8-10) | Total |
|---|---|---|---|---|
| Score | 2.76 | 5.88 | 9.46 | 7.54/10.0 |
| #Records | 142 | 1454 | 1618 | 3214 |

### F.1. Consistency Analysis Outputs

We performed a detailed qualitative analysis of the consistency outputs by categorizing them into four groups: (1) structured query vs. structured query (abnormality-related), (2) structured query vs. structured query (tubes and lines-related), (3) predicted report vs. structured query (abnormality-related), and (4) predicted report vs. structured query (tubes and lines-related). Here, structured queries refer to closed-setting queries, such as classification labels or queries for specific abnormalities/devices, while the predicted report refers to the findings generated by our model. Our analysis showed that many predicted report vs. structured query inconsistencies reflected natural reporting variation rather than true errors. For example, a report might mention an additional device not captured in the structured list, since radiology reports often include more detail than what is exhaustively labeled. In one case, the predicted report described a "prosthetic aortic valve," while the classification of tubes and lines stated "no tubes, lines or devices are present," as this device was not included on the classification labels. In contrast, most valid inconsistencies arose in structured query vs. structured query comparisons. For example, in one record, the specific abnormality query (checking explicitly for pulmonary edema/hazy opacity) was marked as "No," even though the abnormality was present in the list of findings. From this analysis, we concluded that the generation of radiology reports and the classification of abnormalities labels are more reliable querying modes than abnormality-presence queries, as not all conditions have sufficient support in this query mode.

Table 24: Consistency Analysis Examples

| **Predictions** | **Consistency Score** | **Confidence Assessment** | **Explanation** |
|---|---|---|---|
| **predicted_answer:** There is a small left pleural effusion. The mediastinal contours are normal. The hilar contours are normal. There is a small right pleural effusion. There is a diffuse interstitial abnormality. There is no pneumothorax. The pulmonary vasculature is normal. The heart size is normal.<br>**abnormality_presence**: Yes<br>**tubes_lines_devices_presence**: No<br>**abnormalities**: lung opacity, pleural effusion, pulmonary edema/hazy opacity<br>**abnormality_question**: Is there any indication of lung opacity in this image or chest X-ray?<br>**abnormality_answer**: Yes<br>**tubes_lines_devices**: no tubes, lines or devices are present. | 10 | High | First, the **predicted_answer** describes small left and right pleural effusions and a diffuse interstitial abnormality, which are abnormalities, so **abnormality_presence**='Yes' is consistent. It also states 'no pneumothorax', which is not listed in the abnormalities list, so there is no contradiction. The abnormalities list includes 'lung opacity, pleural effusion, pulmonary edema/hazy opacity', and the **predicted_answer** does not explicitly deny any of these; it actually supports pleural effusion and diffuse interstitial abnormality (which could correspond to pulmonary edema/hazy opacity) without contradiction. The **abnormality_question** is about lung opacity, with **abnormality_label_presence**='Yes', and 'lung opacity' is indeed in the abnormalities list, so this is consistent. For tubes/lines/devices, **tubes_lines_devices_presence**='No' and **tubes_lines_devices**='no tubes, lines or devices are present' match, and the **predicted_answer** does not mention any tubes or lines, so there is no contradiction. All fields are internally consistent with no explicit mismatches. |
| **predicted_answer:** There is a small left apical pneumothorax. A right subclavian central venous catheter is in place. There is a small left pleural effusion. A tracheostomy tube is in place. There is a small right pleural effusion. There is a small left apical pneumothorax. There is a small right apical pneumothorax.<br>**abnormality_presence**: 'Yes'<br>**tubes_lines_devices_presence**: 'Yes'<br>**abnormalities**: 'consolidation, lung opacity, pleural effusion, subcutaneous air'<br>**abnormality_question**: 'Is there any indication of lung opacity in this image or chest x-ray?'<br>**abnormality_label_presence**: 'Yes'<br>**tubes_lines_devices**: 'chest tube'<br>**tubes_lines_devices_question**: 'Is there any indication of chest tube in this image or chest x-ray?'<br>**tubes_lines_devices_label_presence**: 'Yes' | 6 | Medium | The **predicted_answer** is generally consistent with **abnormality_presence** ('Yes') and **tubes_lines_devices_presence** ('Yes'), as it describes multiple abnormalities and devices. There is no explicit denial of any abnormality or device that is listed in the exhaustive lists. However, there is a notable inconsistency: the exhaustive **tubes_lines_devices** list and the specific **tubes_lines_devices_label_presence** both indicate a chest tube is present, but the **predicted_answer** does not mention a chest tube at all, instead listing other devices (central venous catheter, tracheostomy tube). This is a direct contradiction because the chest tube is confirmed present in the labels but absent in the narrative. Additionally, the abnormalities list does not include pneumothorax, which is described in the **predicted_answer**, indicating a mismatch in labeling approaches, though not a direct contradiction under the given rules. These issues reduce the consistency score to 6, with medium confidence. |

## Appendix G. Radiologist Evaluation

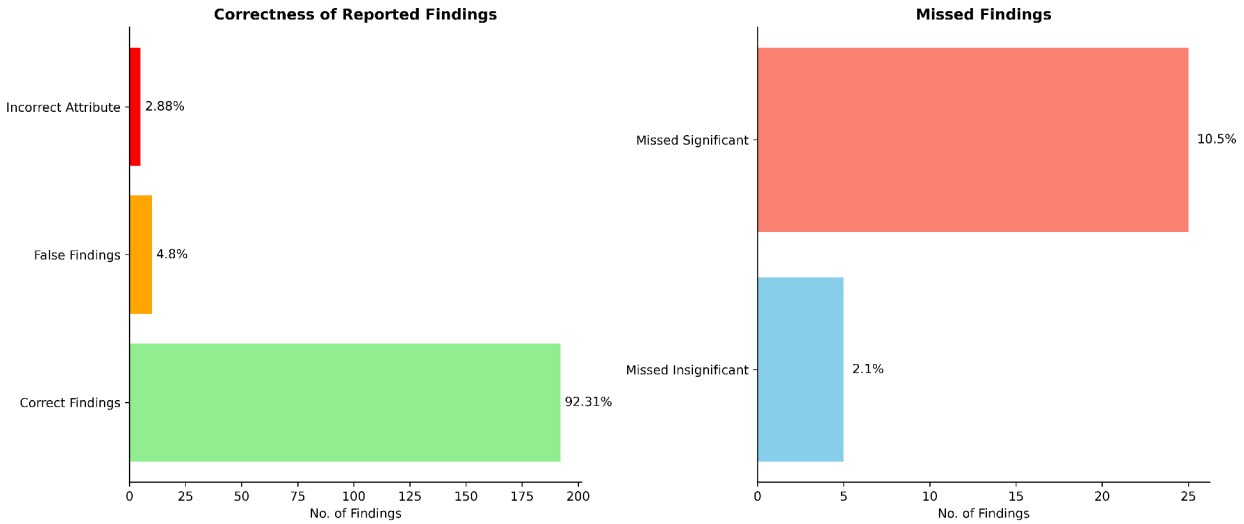



Figure 3: Findings from Radiologist Evaluation

We performed an evaluation of 208 model predicted findings from 30 radiology reports with a board-certified radiologist. The analysis focused on missed findings (classified as clinically significant or insignificant) and the correctness of reported findings (categorized as correct, false, or with incorrect attributes).

The model demonstrated strong overall performance in correctness, with 92.31% of findings accurately reported. However, 4.80% were false positives, such as unwarranted mentions of mild pleural effusions, pneumothorax, or misidentification of eventration of the hemidiaphragm as hiatal hernia in rotated films. An additional 2.88% involved correct findings but with inaccurate attributes, particularly severity levels for pleural effusions (e.g., small vs. moderate or moderate vs. large).

Regarding completeness, 10.50% of clinically significant findings were missed, including cardiomegaly, hyperinflated lungs, mediastinal shift, dilated pulmonary arteries, thoracic scoliosis, nodular opacities, left upper lobe cavitation, fissural opacification, increased bronchovascular marking, and lower zone haziness; 2.10% of misses were insignificant, limited to non-critical details like calcifications and sternal wires. These insights suggest opportunities for model refinement in detection of subtle but significant abnormalities to enhance diagnostic reliability.

**Causal Exploration Qualitative Evaluation** The radiologist also performed a qualitative assessment of the causal exploration summaries accompanying the radiology reports. The evaluation focused on two aspects: clinical soundness and usefulness. Overall, the causal explorations were judged to be clinically sound, with 25 of 30 reports requiring no comments regarding accuracy. The detailed feedback is summarized below:

- **Tubes and Lines** Some explorations linked pleural effusions to recent interventions. The radiologist noted that such associations are of limited value, except in contexts such as unresolved fluid retention or pneumothorax. For example, a useful prediction was: "The imaging shows stable right pneumothorax despite the presence of a right chest tube, with the tube kinked, which may impair drainage." In contrast, other statements were considered less helpful. For instance, the comment "The presence of an endotracheal tube, central venous catheter, and pacemaker indicates a critically ill patient" was considered non-informative. Similarly, when incorrect tube placement was described (e.g., "The endotracheal tube's position in the right mainstem bronchus would preferentially ventilate the right lung, potentially leading to left lung collapse if not corrected"), the radiologist recommended that the language be framed as suggestive rather than strictly causal.

- **Negative Findings** The radiologist emphasized that causal explorations of negative findings are generally not useful unless the absence of a finding directly contributes to diagnostic reasoning.
- **Patient History** To enhance clinical utility, causal explorations should incorporate more patient history beyond the clinical indication provided. Where history is unavailable, the radiologist recommended offering a broader differential diagnosis that considers multiple possible clinical contexts.
- **Differential Diagnosis** The radiologist suggested that differential diagnoses should be more detailed, particularly when patient history is limited. For example, in the case of persistent left lower lobe atelectasis without pleural effusion or pneumothorax, the causal exploration stated: "The persistent left lower lobe atelectasis is most consistent with chronic airway obstruction (e.g., mucus plugging, endobronchial lesion, or extrinsic compression) or post-inflammatory scarring. The absence of effusion, pneumothorax, or cardiomegaly makes compressive or passive causes unlikely." While accurate, the radiologist noted that a more elaborate exploration would be valuable.
- **Incidental Findings:** In several cases, the causal explorations highlighted findings as "incidental" such as the presence of a tracheostomy tube or a tortuous aorta. The radiologist advised that such findings should not be labeled incidental, and in many cases causal explorations need not address them at all.

## Appendix H. Memory Usage and Throughput Comparison

Table 25: Model Memory Footprint and Throughput Comparison

| Model | Memory Footprint (MB) | Throughput (tokens/sec) |
|---|---|---|
| Rad-Phi4-Vision-CXR | 16,358 | 33 |
| MedVersa | 26,197 | 17 |
| RADVLM | 30,635 | 25 |
| MedGemma | 16,403 | 23 |
| CheXagent | 11,981 | 20 |

## Appendix I. Additional Prompts

In this section, we describe the Llama-3.3-70B (Meta AI, 2024) prompts employed for dataset construction and evaluation.

## Consistency Analysis Prompt

**Task:** You are an expert radiology evaluator tasked with assessing the consistency of predictions across multiple tasks for a single radiology record.

**Input Format (JSON fields):**

- "question": Original query for generating findings
- "predicted_answer": Predicted radiology report
- "abnormality_presence": Yes/No (any abnormality)
- "tubes_lines_devices_presence": Yes/No (tubes/lines/devices present)
- "abnormalities": List of abnormalities (comma-separated)
- "abnormality_question": Specific abnormality query
- "abnormality_label_presence": Yes/No (answer to query)
- "tubes_lines_devices": List of detected tubes/lines/devices
- "tubes_lines_devices_question": Specific tube/line query
- "tubes_lines_devices_label_presence": Yes/No (answer to query)

**Evaluation Focus:**

- Identify **contradictions** among the 7 prediction types.
- Ignore missed findings or completeness.
- Use **direct label matching only** (e.g., "consolidation" $\neq$ "opacity").
- Example: If `predicted_answer` says "no pneumothorax" but `abnormality_label_presence = yes`, that is a contradiction.

**Steps to Follow:**

1. Extract key findings from each field using direct labels.
2. Cross-check consistency across:
   - `predicted_answer` vs. presence flags.
   - `abnormalities` vs. `abnormality_question` + label.
   - `tubes_lines_devices` vs. `tubes_lines_devices_question` + label.
3. List explicit contradictions only.
4. Assign a **consistency score (0–10)** and a **confidence level:** High (8–10), Medium (4–7), Low (0–3).

**Output Format (JSON):**

```
{
  "contradictions": ["List of contradictions or 'None'"],
  "consistency_score": integer,
  "confidence_assessment": "High/Medium/Low",
  "reasoning": "Detailed explanation of analysis"
}
```

## Causal Exploration Prompt

You are a highly skilled radiology assistant with expertise in diagnostic reasoning and differential diagnosis. Your role is to assist in the interpretation of radiology reports by integrating imaging findings with the patient's clinical presentation, symptoms, and medical history. Using your in-depth knowledge of radiographic patterns, imaging features of various diseases, and their underlying pathophysiology, your task is to explore potential diagnoses and causal relationships. When analyzing the radiology report, consider the clinical indications for why the imaging was ordered and provide a concise, evidence-based analysis that connects the radiologic findings to the primary clinical concern. Your response should focus strictly on the imaging findings and their relevance to the clinical indication. If abnormalities are present, prioritize the most likely diagnosis and briefly address key differentials. If findings are normal, confirm whether the clinical indication is ruled out based on the imaging. Keep your reasoning brief, clear, and directly relevant to the primary diagnosis. Do not include any suggestion. Output only in the format given in the examples below:

**Radiology Report:** INDICATION: $___$-year-old man with a history of end-stage renal disease, status post kidney transplant, presents to the clinic with increasing fatigue and dyspnea on exertion and chest congestion. Rule out pulmonary edema. COMPARISON: Preop chest radiograph, ___. PA AND LATERAL CHEST RADIOGRAPH: The cardiac, mediastinal, and hilar contours are unchanged. Pleural thickening within both lung bases is unchanged from the prior examinations. Opacification in the right lower lung medial base is consistent with right lower lobe pneumonia. Findings were discussed with Dr. ___ at 16:31 on ___ via telephone.

REASONING: The opacification in the right lower lung medial base suggests right lower lobe pneumonia, which is consistent with the increased density area seen in the radiograph. Additionally, pleural thickening may indicate pleurisy, contributing to chest congestion and dyspnea on exertion. The findings are located in the right lower lobe and pleural regions, correlating with the noted symptoms and radiographic observations.

**Causal Exploration Evaluation Prompt**

Task: Evaluate two radiology reports|one Ground Truth and one Predicted|each containing Findings and Reasoning sections. The Predicted Reasoning is derived from its own Findings. Evaluate the Predicted Reasoning for findings that match the Ground-Truth Findings. If no findings in the Predicted Report match the Ground-Truth Findings, assign a score of 0 to all criteria and note this in the justifications. Ignore any management recommendations or non-radiographic causes in either report.

**Ground-Truth Report:**
ground_truth_findings: <ground truth findings>
ground_truth_reasoning: <ground truth reasoning>

**Predicted Report:**
predicted_findings: <predicted findings>
predicted_reasoning: <predicted reasoning>

**Tasks:**
Identify findings in the Predicted Report that are clinically equivalent to the Ground-Truth Findings.Use the following guidelines:

- Consider synonyms or related terms (e.g., "opacity" vs "consolidation").
- Account for anatomical proximity (e.g., "left perihilar" vs "left lower lobe").
- Focus on core radiographic findings; ignore minor differences.
- If uncertain, prioritize semantic similarity.

If no findings match, assign all criteria a score of 0. Otherwise, score the Predicted Reasoning on a scale of 1 to 5 with one-sentence justifications. Provide a comparative analysis against the Ground-Truth Reasoning.

**Criteria Definitions:**

- Accuracy of Interpretation: Correct interpretation of matched findings.
- Clinical Plausibility: Reasonable differentials and conclusions.
- Completeness: Addresses all matched findings.
- Evidence Linking: Connects reasoning to findings.
- Clarity and Conciseness: Clear, precise, and non-redundant.

## Appendix J. Ablation Study

### J.1. Data Augmentation - Causal Exploration Task

We measured the impact of data augmentation on the causal exploration task using the Hidden-Rad dataset.

Zero-shot baseline: We evaluated the `Rad-Phi4-Vision-CXR` model, which was not trained for causal reasoning, in a zero-shot setting with reasoning prompts. The model performed poorly, confirming its limited inherent reasoning capability.

Training without augmentation: We trained the `Rad-Phi4-Vision-CXR` model solely on the original Hidden-Rad dataset (without synthetic augmentation) and evaluated it on the same test set. While this led to some improvement, performance remained constrained due to the limited size of the dataset.

Training with synthetic augmentation: We trained the model using the synthetic dataset and evaluated it on the Hidden-Rad test set. This setup achieved the best performance, demonstrating that synthetic augmentation significantly improved results.

Table 26: Impact of Data Augmentation - Hidden-Rad Causal Reasoning Benchmark

| Setting | Interpretation Accuracy (0–5) | Clinical Plausibility (0–5) | Completeness (0–5) | Evidence Linking (0–5) | Clarity & Conciseness (0–5) | Mean Total Score (0–25) |
|---|---|---|---|---|---|---|
| Zero Shot | 0.83 | 0.91 | 0.87 | 0.71 | 0.58 | 3.90 |
| Without Data Augmentation | 1.24 | 1.30 | 1.03 | 1.28 | 1.22 | 6.07 |
| With Data Augmentation | **3.29** | **3.56** | **2.79** | **3.38** | **3.71** | **16.74** |

### J.2. Image to Text Retrieval Performance

Several prior works (CheXagent, RadVLM, MedGemma) also employed sigmoid loss for image encoder pre-training with radiology images and reports, but did not present ablations. Our ablation results demonstrates that sigmoid loss consistently outperforms standard contrastive loss in most pathologies for the Image-Text

Retrieval task. The common pathologies (first 5) in the table are sourced from the CheXpert dataset and the rarer pathologies are sourced from various different datasets like CXR-LT, VinDR-CXR and SIIM-CXR.

Table 27: Image-Text Retrieval Performance - CLIP vs SigLiP Loss

| Disease | Rad-CLIP Recall@3 | Rad-CLIP Recall@5 | Rad-CLIP Recall@15 | Rad-SigLip Recall@3 | Rad-SigLip Recall@5 | Rad-SigLip Recall@15 |
|---|---|---|---|---|---|---|
| Atelectasis | 0.764 | 0.859 | 0.955 | **0.792** | **0.887** | **0.983** |
| Edema | 0.741 | 0.800 | 0.847 | **0.800** | **0.894** | **0.941** |
| Cardiomegaly | **0.240** | **0.268** | 0.391 | 0.177 | **0.268** | **0.417** |
| Consolidation | 0.485 | 0.542 | 0.771 | **0.571** | **0.657** | **0.971** |
| Pleural Effusion | **0.653** | 0.775 | 0.916 | 0.650 | **0.783** | **0.933** |
| Hernia | 0.527 | 0.593 | 0.693 | **0.551** | **0.615** | **0.707** |
| Pneumoperitoneum | 0.535 | 0.571 | 0.707 | **0.564** | **0.621** | **0.750** |
| Pneumomediastinum | 0.558 | 0.631 | 0.754 | **0.586** | **0.670** | **0.787** |
| Mediastinal Shift | **0.750** | **0.800** | 0.950 | **0.750** | **0.800** | **1.000** |
| Pneumothorax | **0.998** | **0.998** | **1.000** | **0.998** | **0.998** | **1.000** |

### J.3. Grounding Performance - Abnormality Labels and Phrases

We favor specialized vision modules for tasks such as segmentation and grounding (bounding box detection), which leverage language model embeddings for guidance rather than relying on language models to directly decode grounding outputs. Ablation results show that embedding-driven grounding modules outperform direct language model decoding across multiple grounding settings, including grounding of abnormality labels and phrases across different datasets. Both settings leveraged the same grounding dataset for training.

Table 28: Grounding Performance of Abnormality Phrases

| Setting | MS_CXR (mAP) | MS_CXR (mIoU) | PadChest (mAP) | PadChest (mIoU) |
|---|---|---|---|---|
| Direct LLM decoding | 0.424 | 0.312 | 0.367 | 0.257 |
| Specialized vision module | **0.818** | **0.540** | **0.720** | **0.478** |

Table 29: Grounding Performance of Abnormality Labels

| Setting | NIH (mAP) | NIH (mIoU) | PadChest (mAP) | PadChest (mIoU) |
|---|---|---|---|---|
| Direct LLM decoding | 0.233 | 0.207 | 0.252 | 0.219 |
| Specialized vision module | **0.544** | **0.363** | **0.675** | **0.460** |

### J.4. Influence of Stage 1 Training

We find that including Stage-1 pretraining improves report generation. Specifically, leveraging the annotations of the Chest ImaGenome dataset for the VQA task during Stage 1 leads to measurable gains. Our ablation confirms that Stage-1 VQA pretraining enhances report generation performance.

Table 30: Report Generation Performance with and without Stage-1 Pre-training

| Setting | Dataset | BLEU-2 | BERT-Score | ROUGE-L | RadGraph-F1 |
|---|---|---|---|---|---|
| With stage-1 pre-training | CheXpert-Plus | **0.17 [0.15, 0.19]** | **0.42 [0.39, 0.45]** | **0.23 [0.22, 0.25]** | **0.26 [0.24, 0.28]** |
| Without stage-1 pre-training | CheXpert-Plus | 0.17 [0.01, 0.47] | 0.35 [0.11, 0.60] | 0.21 [0.04, 0.45] | 0.19 [0.02, 0.45] |
| With stage-1 pre-training | MIMIC | **0.16 [0.16, 0.17]** | **0.44 [0.43, 0.44]** | **0.23 [0.22, 0.23]** | **0.24 [0.23, 0.24]** |
| Without stage-1 pre-training | MIMIC | 0.14 [0.01, 0.38] | 0.37 [0.12, 0.60] | 0.23 [0.07, 0.42] | 0.17 [0.00, 0.43] |

### J.5. LLM Training Strategy

Consistent with previous work, we observe that LoRA-based fine-tuning performs better than full fine-tuning. Our ablation demonstrates improved report generation performance when using LoRA training.

Table 31: Findings Generation - LLM Full Finetuning vs LoRA Finetuning

| Setting | Dataset | BLEU-2 | BERT-Score | ROUGE-L | RadGraph-F1 |
|---|---|---|---|---|---|
| LLM LoRA Finetuning | CheXpert-Plus | **0.17 [0.15, 0.19]** | **0.42 [0.39, 0.45]** | **0.23 [0.22, 0.25]** | **0.26 [0.24, 0.28]** |
| LLM Full Finetuning | CheXpert-Plus | 0.17 [0.01, 0.38] | 0.38 [0.19, 0.50] | 0.21 [0.05, 0.39] | 0.17 [0.03, 0.37] |
| LLM LoRA Finetuning | MIMIC | **0.16 [0.16, 0.17]** | **0.44 [0.43, 0.44]** | **0.23 [0.22, 0.23]** | **0.24 [0.23, 0.24]** |
| LLM Full Finetuning | MIMIC | 0.14 [0.01, 0.35] | 0.38 [0.18, 0.55] | 0.23 [0.09, 0.38] | 0.16 [0.02, 0.38] |

### J.6. Multi-task training vs Standalone training - ReXVQA benchmark

Training VQA jointly with report generation improves VQA outcomes. Our ablation comparing standalone VQA training with multi-task VQA + report generation shows improved metrics in the multi-task setup.

Table 32: Multi-task Training vs Standalone Training - ReXVQA Benchmark

| Setting | Overall Accuracy | Differential Diagnosis | Geometric Information | Location Assessment | Negation Assessment | Presence Assessment |
|---|---|---|---|---|---|---|
| VQA + Report generation training | **0.986** | **0.974** | **0.941** | **0.971** | **0.996** | **0.985** |
| VQA standalone training | 0.969 | 0.954 | 0.833 | 0.932 | 0.988 | 0.972 |

### J.7. Inclusion of Prior Images

We compared the performance of Rad-Phi4-Vision-CXR with and without prior images, in addition to the current image, for the findings generation task. The results show that incorporating priors consistently improves performance across all metrics.

Table 33: Ablation: Rad-Phi4-Vision-CXR (Without vs With Priors) - Findings Generation Metrics

| Setting | BLEU-2 | BERT-Score | RadGraph-F1 | CheXbert | GREEN |
|---|---|---|---|---|---|
| **ReXGradient (without priors)** | 21.45 [20.97, 22.19] | 44.69 [44.31, 45.34] | 27.60 [27.10, 28.30] | 49.04 [47.85, 49.58] | 39.48 |
| **ReXGradient (with priors)** | **30.24 [29.75, 30.64]** | **52.20 [51.87, 52.54]** | **36.60 [35.90, 37.60]** | **57.06 [56.29, 58.15]** | **47.83** |
| **MIMIC (without priors)** | 16.83 [16.45, 17.45] | 44.40 [44.06, 44.94] | 23.60 [23.20, 23.90] | 52.30 [51.37, 52.70] | 30.03 |
| **MIMIC (with priors)** | **22.38 [21.88, 22.40]** | **49.92 [49.48, 50.13]** | **27.70 [27.60, 28.50]** | **52.90 [52.70, 53.70]** | **33.50** |

### J.8. Tubes and Lines Segmentation Baseline

Table 34 presents a comparison of segmentation performance on the RANZCR CLiP dataset between Rad-Phi4-Vision-CXR and a U-Net baseline (ResNet34 pretrained). Unlike the U-Net baseline, which is a pure vision model, Rad-Phi4-Vision-CXR integrates segmentation with prompt-based embeddings, making it the first baseline for tubes and lines segmentation using LLM prompting. Results show that Rad-Phi4-Vision-CXR achieves superior performance for ETT segmentation, while U-Net performs better on CVC and NGT, indicating promising results and opportunities for further improvement.

Table 34: Performance on RANZCR CLiP Dataset (Dice Scores): Tubes and Lines Segmentation

| Model Name | ETT | CVC | NGT | Mean Tubes/Lines |
|---|---|---|---|---|
| Rad-Phi4-Vision-CXR | **0.905** | 0.791 | 0.782 | 0.791 |
| U-Net (ResNet34 Pretrained) | 0.800 | **0.868** | **0.899** | **0.855** |

## Appendix K. GREEN Error Category Analysis

### K.1. ReXGradient Dataset

Table 35: GREEN Score and Error Category Distribution

| Error Category | Medversa | RadVLM | Rad-Phi4-Vision-CXR | ChexAgent | MedGemma |
|---|---|---|---|---|---|
| Avg GREEN Score | 0.47 | 0.36 | 0.49 | 0.39 | 0.46 |
| (a) False report of a finding | 27.7% | 21.1% | 26.2% | 26.4% | 25.6% |
| (b) Missing finding present in reference | **62.3%** | **71.6%** | **63.1%** | **65.2%** | **64.7%** |
| (c) Misidentification of location/position | 3.6% | 2.1% | 4.1% | 3.4% | 2.2% |
| (d) Misassessment of severity | 3.4% | 3.1% | 3.8% | 3.0% | 2.8% |
| (e) Mentioning comparison not in reference | 1.9% | 0.9% | 1.7% | 1.0% | 3.4% |
| (f) Omitting comparison from prior study | 1.1% | 1.1% | 1.1% | 1.0% | 1.2% |

Across all baseline models, the GREEN score in Table 35 indicates overall reporting quality. Consistent with radiologist observations, the predominant failure mode is missing findings present in the reference report, accounting for over 60% of errors across all models, while misidentification of location/position, misassessment of severity, and comparison-related discrepancies (both mentioning comparisons not in the reference and omitting comparisons from prior studies) are each below 5% for all models. These results suggest that the principal opportunity for improving report quality is completeness—ensuring that clinically relevant findings in the reference are systematically captured.