GIV-CXR: Densely Grounded, Visually Interpretable, Chest X-Ray Question Answering Dataset

Anonymous ACL submission

Abstract

Modern medical community seeks precise, multimodal interpretability. People want to explicitly connect image regions to diagnostic outcomes and reason using natural language. Large Multimodal Models (LMMs) are rapidly advancing open domain vision-language reasoning, yet progress in *medical* visual question answering (Med-VQA) remains limited by two persistent bottlenecks: the scarcity of large-scale **region-grounded** supervision and the high cost of continuous **radiologist oversight**. We present an automated Chest X-ray Med-VQA generation-validation pipeline and a grounded Chest X-ray (CXR) dataset GIV-CXR built on top of the Chest ImaGenome dataset. The automated pipeline incorporates LMMs based question-answers generation and validation scaling grounded data generation while preserving clinical reliability. Prompts incorporating domain experts insights regulate question-answer generation ensuring clinical regulation and Large language Models (LLMs) evaluators bring in the reliability from model generated Question-answers. GIV-CXR is a large scale dataset embibing 20,534 images from Chest ImaGenome , annotated over 81,257 bounding boxes, resulting in 354,293 question-answer pairs. The prompts used to generate the QA pairs are designed strategically to imbibe in-depth reasoning for efficient grounding. Standard MLMs underperformed on a sampled test set highlighting the lack of grounding capabilities of the models. On fine-tuning the LMMs on our dataset, the models demonstrate significantly better reasoning and grounding enhancing their interpretability. We will release the resources along with a detailed instructions and ethical use guidelines upon acceptance.

1 Introduction

Image-based methods have become a reliable approach in medical diagnostics, offering clinicians an accurate and non-invasive window into disease progression and treatment response (Bae et al., 2023). Recent advancements in text-image aligned spaces (Radford et al., 2021; Wu et al., 2021) have led to the development of powerful LMMs (Li et al., 2022b; Kim et al., 2021; Ramesh et al., 2021; Radford et al., 2021; Li et al., 2023; Driess et al., 2023; Yang et al., 2023; Ye et al., 2023; Chen et al., 2023a,b). These LMMs have demonstrated remarkable capabilities, facilitating the development of multimodal applications across various domains (Wang et al., 2024). The medical domain has also leveraged the incredible capabilities of LMMs to develop interpretable diagnostic applications such as the generation of medical reports and the visual medical question answer (Med-VQA) (Lin et al., 2023). Despite their remarkable capabilities, LMMs are not entirely reliable for the development of *clinically trustworthy* multimodal medical applications (Sonicki, 2024) due to two major bottlenecks.

Critical regions in medical images are often not uniformly distributed; they concentrate essential information at specific spatial locations (Tascon-Morales et al., 2024). This fine-grained understanding is vital for report generation (Zou et al., 2024) and interactive question answering (Tascon-Morales et al., 2023), and interest in region-specific analysis is rising (Bai et al., 2025a; Zhang et al., 2024b; Peng et al., 2024; Yung et al., 2024; Tascon-Morales et al., 2024). However, most Med-VQA datasets provide either grounding annotations or question–answer pairs but not both (Table 1) and existing QAs seldom include open-ended, region-focused queries. Given the need for exceptional precision in medicine (Bélisle-Pipon, 2024), incorporating region-specific information improves interpretability, mitigates hallucinations, and enables rigorous model probing. While recent efforts pursue region-aware comprehension (Chen et al., 2023c; Liu et al., 2024b; Zou et al., 2024), a comprehensive resource that couples QAs with

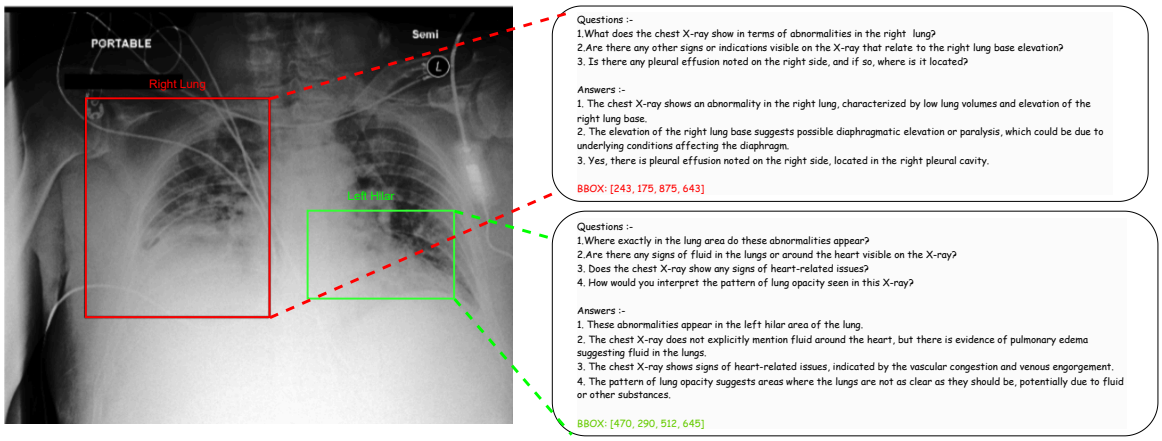


Figure 1: An overview of what GIV-CXR provides to the multimodal interpretability research community. Built upon the Chest ImaGenome dataset, GIV-CXR provides region level deep reasoning question answer pairs along with visual grounding.

explicit spatial grounding at scale remains lacking. Given this level of precision, large-scale supervision relies on expert radiologists to annotate and validate data via iterative review, multi-reader adjudication, standardized protocols, and quality checks. Such fine-grained work is time, coordination, and cost-intensive, demanding specialized expertise and institutional resources. As a result, high-quality datasets are scarce and model development remains bottlenecked. Curating densely grounded datasets therefore imposes a prohibitive annotation burden on already resource-constrained healthcare systems, limiting both the pace and the breadth of progress.

To address these limitations, we move from exhaustive per-example annotation to an *automated, self-evaluating* generation pipeline. The proposed pipeline synthesizes *region-specific* question–answer (QA) pairs mirroring expert radiologists' reasoning. Because LMM-assisted generation is vulnerable to hallucinations, generation is coupled with LLM driven filtering and validation that enforces factuality, semantic consistency, and answer–region coherence. We build on *Chest ImaGenome* (Wu et al., 2021), a large-scale multimodal Chest X-ray dataset (Fig. 1) which contains clinically meaningful findings and anatomy. We manually curate a seed set of critical descriptors and feed them to an instruction-tuned Grok via *radiologist-informed* prompts. These prompts distilled by expert feedback and refined iterative regulate generation of open-ended, clinically specific questions anchored to particular image regions and require correspondingly grounded answers.

To regulate biases and spurious generations, we integrate an evaluation module comprising filtering and a *DeepSeek* (Liu et al., 2024a) based judge to score factuality, semantic alignment, and box–answer coherence, alongside linguistic quality and reasoning depth (Liu et al., 2023a). We additionally compute standard NLG metrics (BLEU, ROUGE-1, ROUGE-L, METEOR, BERTScore) and localization scores (mIoU) to provide complementary, reproducible diagnostics. The generated GIV-CXR dataset using this pipeline comprises 892,364 QA pairs across 191,654 critical regions drawn from 21,680 images. Curation is ongoing, with the goal of scaling to all 242,072 images in the base corpus, thereby creating a comprehensive resource for grounded reasoning in medical imaging. We evaluated seven standard LMMs on a curated test set derived from our proposed dataset. The models exhibited low performance, highlighting their limited capability for region-specific reasoning. we fine-tuned two of these models on a smaller subset of our dataset, carefully balancing resource constraints. Despite being trained on a fraction of the full dataset, the fine-tuned models showed notable improvements across standard answer generation (BLEU, ROUGE-1, ROUGE-L, METEOR, BERTScore) and grounding metrics (mIoU). The generated answers are well-reasoned, and the models successfully aligned their responses with relevant image regions. These findings lay the foundation for further research into the development of interpretable LMMs for medical imaging. We provide a dedicated section discussing the impact of our dataset, potential future directions, and the

types of datasets that could be built upon it. Our contributions can be summarised as follows:

- We introduce a radiologist-informed, LMM-driven self validating pipeline that generates region-specific QAs reducing domain experts.
- We release a large-scale, densely grounded Med-VQA resource with 20,534 images, 354,293 bounding boxes, and 81,257 region-anchored QA pairs, with ongoing curation toward full Chest ImaGenome coverage (Wu et al., 2021).
- Across seven LMMs, we demonstrate low performance and demonstrate that modest fine-tuning on GIV-CXR significantly improves reasoning and grounding, as measured by G-Eval, standard NLG metrics, and mIoU.

2 Related Works

2.1 Medical Visual Question Answering Datasets

In recent years, several datasets have been developed to advance Medical Visual Question Answering (Med-VQA), each addressing specific challenges across clinical domains. VQA-RAD (Antol et al., 2015) is a foundational resource with over 3,000 QA pairs focused on radiology images, particularly Chest X-ray , PathVQA (He et al., 2020) targets fine-grained pathology analysis with 32,000+ QA pairs for histopathological images. RadGenome-Chest CT (Zhang et al., 2024a) offers structured annotations for model training, MIMIC-Diff-VQA (Hu et al., 2023) addresses differential diagnosis reasoning by comparing two X-ray images; and MIMIC-CXR-VQA (Bae et al., 2024), built on MIMIC-CXR (Johnson et al., 2019), introduces diverse question templates tailored for thoracic radiology to aid chest abnormality detection.

2.2 Medical Visual Grounding Datasets

Beyond VQA, several datasets focus on grounding and segmentation tasks, contributing to broader AI applications in medical imaging. EHRXQA (Bae et al., 2023), a dataset that integrates image-based and structured data queries, enhances multi-modal learning in QA and VQA applications. For segmentation and classification tasks, CXLSeg (Nimalsiri et al., 2023) and CheXmask DB (Gaggion et al., 2024) contribute by providing labeled segmentation masks.

2.3 LLM-as-a-Judge Metrics for Medical VQA Evaluation

Evaluation of open-ended generative VQA especially in medicine has shifted towards reference-free, human-aligned metrics. Traditional metrics like BLEU/ROUGE correlate poorly with human judgments on tasks requiring creativity or factual precision. G-Eval (Liu et al., 2023b) instantiates an LLM-as-a-judge (GPT-4) with structured chain-of-thought and form-filling. Evaluators are given explicit criteria, the LLM reasons step-by-step, and then scores outputs for correctness, coherence, context alignment, and fluency.

3 GIV-CXR Curation pipeline

In this section, we expand upon the automated curation pipeline for the proposed GIV-CXR dataset. A visual schema of the entire pipeline is provided in Figure 2. The dataset is built upon the Chest ImaGenome dataset (Wu et al., 2021). The dataset construction process mainly consists of three phases. In the first phase as explained in sec 3.1, we preprocess and generate the question-answer-bbox triplets from the base dataset. the next two phases involves extensive filtering on the curated data and removal of hallucinated sampled, ensuring consistent and high quality data samples. The prompt design strategy is covered in sec 3.2. The final phase comprising post-processing and the details to reach the final stage of GIV-CXR is discussed in the section.

3.1 Phase 1: VQA generation

3.1.1 Extracting critical keywords from Chest ImaGenome

The Chest ImaGenome dataset (Wu et al., 2021) represents each frontal Chest X-ray as a scene graph. We iteratively parse each graph to extract *reason_for_exam*, *objects*, and *attributes* JSON fields sufficient to generate semantically meaningful QA pairs. In the current version of the dataset release, we restrict supervision to single regions and omit inter-regional *relations*, which we plan to incorporate in future versions to enable connectivity reasoning. Texture cues and anatomical findings supply rich visual descriptors, while disease/abnormality phrases derived from reports preserve clinical semantics.

Curation and filtering were conducted under close radiologist supervision as illustrated in Fig 1

Dataset	# Images	Properties	Annotations	Primary Task	Groundable	Region-specific
VQA-RAD (Antol et al., 2015)	3.5 K	Radiology VQA (X-ray, CT, MRI)	QA	VQA	×	×
SLAKE (Liu et al., 2021)	14.0 K	Multi-modality VQA	QA	VQA	×	×
PathVQA (He et al., 2020)	33 K	Histopathology VQA	QA	VQA	×	×
Chest ImaGenome (Wu et al., 2021)	242 K	Region findings	BBox	Report Gen.	✓	✓
ChestX-ray8 (Wang et al., 2017)	112 K	14 diseases, partial boxes	BBox	Classification	✓	×
EHRXQA (Bae et al., 2023)	46 K	Image + structured queries	QA	QA/VQA	✓	×
VinDr-CXR (Nguyen et al., 2022)	18 K	Thoracic findings	BBox	Classification	✓	×
Diff-VQA (Cho et al., 2024)	700 K	Paired-image diff questions	QA	VQA	×	×
CXL-Seg (Li et al., 2022a)	243 K	Segmentation masks	Mask	Segmentation	×	×
CheXmaskDB (Gaggion et al.)	657 K	Anatomical masks (1024^2)	Mask	Segmentation	✓	×
GIV-CXR++ (Ours)	**21.6 K**	Dense grounding + QA	BBox + QA	VQA	✓	✓

Table 1: Comparison of major medical-imaging datasets. ✓ denotes the presence of region-level information suitable for grounding; × denotes its absence. Existing Med-VQA corpora do not jointly provide bounding boxes *and* region-specific QA pairs at scale.

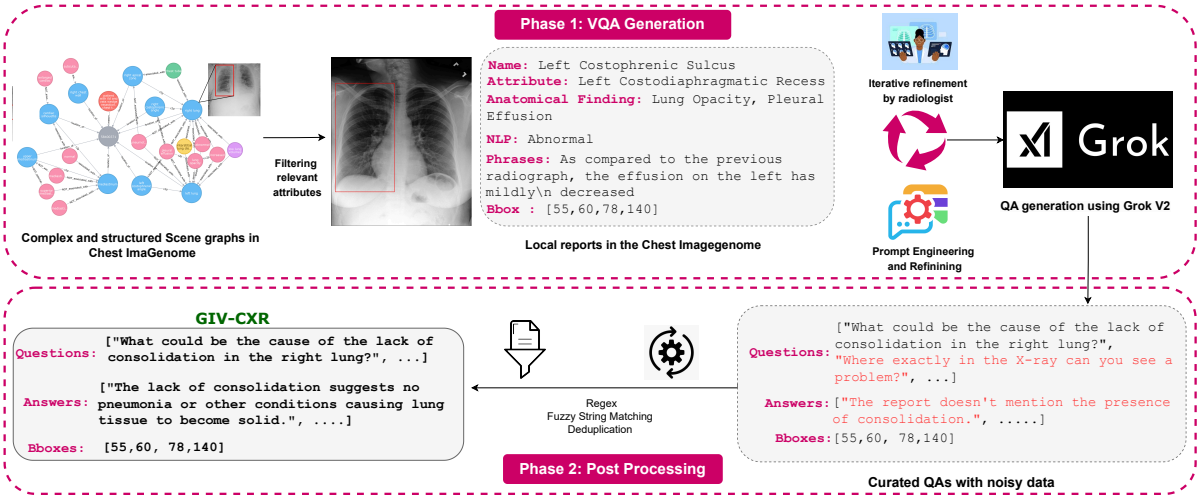


Figure 2: Curation pipeline for GIV-CXR.

with multiple iterations to ensure clinical fidelity. Our pipeline mirrors radiologists' region-first reading, emphasizing fine-grained localization and evidence attribution, thereby improving the dataset's reliability and alignment with real-world radiological practice.

3.1.2 Generating Question-Answer Pairs using XAI's Grok-2

Leveraging bounding box-specific data, we systematically generate clinically relevant question-answer (QA) pairs that align with radiologists' reasoning process in Chest X-ray analysis. To achieve this, we designed structured prompts that incorporate anatomical regions, attributes, texture cues, and descriptive phrases, ensuring that each question is contextually grounded and medically relevant. These prompts guide the LMMs in generating precise and meaningful questions that focus on abnormalities, locations, causes, and textures while explicitly avoiding speculative interpretations or temporal assumptions.

Once the questions are generated, the LLM is strategically prompted again to produce corresponding answers. This process involves conditioning the LLM on the generated questions and contextual information associated with the specific anatomical region. By infusing domain-specific medical knowledge into the response generation process, the model ensures that all answers remain concise, clinically precise, and strictly aligned with the provided data. Each QA pair is then systematically linked to its respective bounding box, image ID, and associated attributes, maintaining traceability, contextual integrity, and relevance for downstream clinical and research applications.

3.2 Prompt design for data generation

We selected XAI's Grok-2 due to its better performance on the MedQA (VALS AI, 2025) benchmark conducted by Vals AI, demonstrating its advanced

medical context comprehension. Key parameters were identified and incorporated into prompt designing to ensure their alignment with the core goals: 1) Enhancing in-depth perceptual reasoning through well-grounded QA pairs. The generated QA pairs are designed with consistent alignment to corresponding anatomical region. This approach enables the model trained on this data to learn structured medical reasoning beyond what is achievable with conventional Visual Question Answering (VQA) datasets, and 2) By incorporating feedback from radiologists, we iteratively refined our prompts to eliminate assumptions about underlying conditions and maintained the dataset's integrity. The prompts used are given in the appendix.

1. **Clinical Relevance:** The prompt asks the model to frame questions that: Identify abnormalities visible in the region. Explore the location, cause, and significance of the findings. Focus on texture information if available. This aligns the questions with the diagnostic process that radiologists follow when interpreting X-rays.

2. **Avoidance of ambiguity:** The question generation prompt explicitly restricts speculative or overly generic questions, such as *"Can you explain what this means?"*; *"Should I be concerned about this finding?"*, and *"Why is this happening?"*. Instead, the questions are designed to be specific, factual information that can be derived directly from the given attributes and phrases. For the answers generation, we hinder the LLM from referencing the reports directly, instead the model is advised strictly to answer each question from a visual point of view, given the observations at that specific zone. We also discourage the model from assuming or giving self-explanatory questions or answers.

3. **Exclusion of temporal comparisons for questions:** Temporal comparisons (e.g., changes from previous exams) are deliberately avoided in the questions. This ensures that the QA pairs focus solely on the current findings, which is crucial for standalone diagnostic insights.

3.3 Phase 2: Data Quality Enhancement Through Question-Answer Pair Filtering

This specifically targeted QA pairs that referenced report text rather than direct visual observations from the X-ray images, as well as questions that were generically specific about spatial location in the entire X-ray, rather than focusing on a specific region.

Our filtering approach utilized regular expression pattern to remove these references. These patterns included phrases like "as per the report", "mentioned in the report,", "according to the report", "Where exactly in the X-ray", and "In which part of the X-ray". Hence, we filtered 15403 QA pairs and $369,696$ reduced to $354,293$ pairs. Notably, while the filtering process affected many entries, it did not eliminate any complete entries from the dataset, as evidenced by the unchanged count of $81,257$ entries.

3.4 Phase 3: Self Evaluating Hallucinated Question-Answers Filtering

The generated question-answer (QA) pairs may contain hallucinations. These include factual inconsistencies, contradictions, or clinically irrelevant reasoning, such as speculative explanations for normal findings. To identify and remove such hallucinated pairs, we employed a Deep-Seek based verification system that evaluates each QA against the corresponding report data.

We provided the model with structured report metadata: anatomical attributes, texture cues, and descriptive phrases associated with the region of interest.

Human Validation of LLM-based Hallucination Filtering

To assess the reliability of the LLM-based hallucination filter, we conducted a small validation study with clinically-certified radiologists. A random sample of 160 QA pairs was drawn from the development set. For each QA pair, we recorded the LLM-judge's prediction of whether the QA was hallucinated as well as its accompanying explanation. Three radiologists independently annotated the same pairs for (i) factual correctness with respect to the source report and (ii) clinical appropriateness. Disagreements were resolved by majority vote.

Out of the 160 pairs, the radiologists judged 9 (5.6%) to be clinically inaccurate and identified 22

(13.8%) instances in which the LLM's assessment was incorrect. Most errors involved unsupported claims or spurious reasoning about normal findings. After incorporating these insights to refine the prompts and re-running the filtering step on the same sample, the LLM's predictions were fully aligned with the radiologists' judgments. This validation confirms that the final LLM-based filter is consistent with expert evaluation and highlights the importance of expert-in-the-loop calibration for medical QA hallucination detection.

The prompt incorporated these insights from radiologists to flag hallucinated QA pairs based on three primary criteria: (i) unsupported factual claims, (ii) contradictions with the report, and (iii) clinically unhelpful reasoning (e.g., justifying normality). QA pairs were assigned a hallucination flag (is hallucination: 1 or 0), a confidence score (0.0–1.0), and an explanation for the classification. Out of 393,425 QA pairs across 78,593 entries, 71,723 (18.2%) were classified as hallucinated. Notably, hallucination rates varied across anatomical regions. The regions with the highest rates included:

Common patterns observed in hallucinations included: 1) **Negation hallucinations:** 9.9% of cases involved inappropriate reasoning about the absence of findings (e.g., "reason for no pneumothorax"). 2) **Normality-based hallucinations:** 58.9% of hallucinated pairs were linked to answers or questions referencing normal findings without supportive diagnostic context. 3) **Left-right confusion:** Present in 1.0% of hallucinated cases.

Table 2: Hallucination Rates by Anatomical Region

Region	Hallucination Rate
Right clavicle	60.0%
Left clavicle	47.8%
Abdomen	29.9%
Right hemidiaphragm	28.6%
Left apical zone	27.9%

Frequently used terms in hallucinated answers included: *normal*, *lung*, *left*, *absence*, and *pleural*, suggesting a pattern of overgeneralization or unsupported negations. Post filtering, the dataset was left with high-quality, clinically relevant QA pairs.

4 Dataset overview

The curated dataset comprises of 20,534 Chest X-ray images from the MIMIC-CXR dataset, resulting in a total of 354,293 question-answer-bounding box pairs after post-processing. More analysis and statistics about dataset are shared in apendix.

Split	# QAs	# Imgs	# BBoxes
Original	354,293	20,534	81,257
Train	150,000	19,194	66,615
Test	7,500	1000	3,916

Table 3: Dataset statistics.

5 Experiments and Results

5.1 Sampling training and test sets from GIV-CXR

The training set was sampled by region-level QA counts Fig 3 and filtered by mean bounding box area, selecting regions up to the aortic arch to balance anatomical diversity with spatial consistency. Train/test splits were then matched for anatomical distribution to support generalization across comparable regions

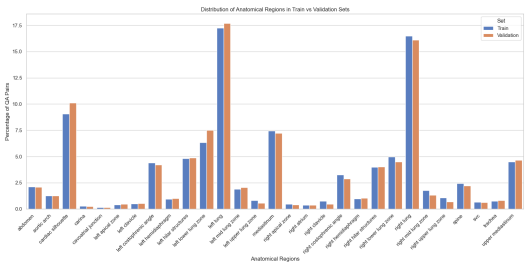


Figure 3: Sampling distribution of train-test splits

5.2 Evaluation Metrics

To evaluate grounded question–answer generation, we adopt **G-Eval**(Liu et al., 2023b) as our primary criterion. Complementing this, and following grounded text-generation practice (Liu et al., 2024b), we report standard NLG metrics: BLEU (Papineni et al., 2002), ROUGE1/L (Ganesan, 2018), BERTScore (Zhang et al., 2019) and localization performance via mIoU for box alignment with ground truth (Rezatofighi et al., 2019).

5.3 Baselines

The models in our evaluation can be categorized into two groups: (1) *Unfine-tuned models*, which are directly used for inference without adaptation, and (2) *Fine-tuned models*, which are finetuned using our dataset to improve both question-answering (QA) and grounding capabilities. We evaluate medical domain-pretrained multimodal models: CheXagent (Chen et al., 2024) and MedGemma-4B

Model	BLEU	ROUGE-L	ROUGE-1	BERTScore	mIoU
CheXagent	10.59	27.05	28.43	28.20	-
MedGemma	14.24	36.84	41.07	55.20	-
GPT-4o-mini	18.65	47.53	50.37	65.35	19.33
LLaMA-3.2-11B†	18.42	51.22	21.18	19.04	-
LLaMA-3.2-11B*	48.87	70.71	73.93	80.05	-
Qwen-2.5VL-7B (Bbox output)†	29.60	42.43	38.44	47.04	11.93
Qwen-2.5VL-7B (Bbox output)*	42.56	66.40	69.78	77.57	68.12
Qwen-2.5VL-7B (Bbox input)†	38.42	50.17	46.30	53.07	-
Qwen-2.5VL-7B (Bbox input)*	28.04	51.22	55.33	66.04	-

Table 4: Model performance on language generation metrics. All values are F1 scores in percentages. * indicates **fine-tuned models**. † indicates **unfine-tuned models**.

Model	GIV-CXR
MedGemma (Sellergren et al., 2025)	3.47
CheXagent (Chen et al., 2024)	3.22
GPT-4o-mini (Achiam et al., 2023)	3.13
Qwen-2.5 VL 7B (Bbox output)†	2.97
Qwen-2.5VL 7B (Bbox output)*	3.98
LLaMa 3.2 11B†	2.18
LLaMA 3.2 11B*	3.86
Qwen-2.5 VL 7B (Bbox input)†	2.92
Qwen-2.5VL 7B (Bbox input)*	3.67

Table 5: G-Eval results of different models on the GIV-CXR dataset. * indicates models **fine-tuned** on task-specific data. † indicates **unfine-tuned models**.

(multi-modal) (Sellergren et al., 2025) for visual question answering, and use GPT-4o-mini (Achiam et al., 2023) separately for grounding. These systems perform strongly on generic multimodal benchmarks but lack supervision targeted at fine-grained, region-level reasoning in medical images.

We fine-tuned LLaMA-3.2-11B (Touvron et al., 2023) on a *50k* QA subset to predict answers from image–question pairs (no grounding). For Qwen-2.5 VL-7B (Bai et al., 2025b), we train two variants on *150k* QA pairs: (i) answer + bounding-box generation; and (ii) the same, with explicit box supervision using <box_start> and <box_end> tokens to reinforce region-specific learning.

5.4 Discussion

Table 4 and 5 highlights the impact of fine-tuning on region-specific reasoning.

Pretrained performance: Despite strong general multimodal capability, domain-pretrained baselines underperform on our benchmark: CheXagent (Chen et al., 2024) attains *3.22* G-Eval (Liu et al., 2023b) and *27.05* ROUGE-L (Ganesan, 2018), while MedGemma-4B (Sellergren et al., 2025) improves to *3.47* G-Eval and *36.84* ROUGE-L, yet still fails to reason effectively at fine-grained, region level. Similarly, the unfine-tuned LLaMA-3.2-11B reaches only *2.18* G-Eval on the GIV-CXR, while Qwen-2.5 VL (bbox output) achieves *2.97*. These results further highlight that pretrained models, despite strong general capability, lack the region-specific grounding needed for precise medical reasoning.

Effect of fine-tuning (QA): Fine-tuning LLaMA-3.2-11B (Touvron et al., 2023) on *50k* curated QA pairs yields substantial gains: *3.83* G-Eval (Liu et al., 2023b) and *70.71* ROUGE-L (Ganesan, 2018) surpassing all unfine-tuned models and indicating that even modest, high-quality supervision improves medical QA.

Impact of bounding-box supervision: For Qwen-2.5-VL-7B (Bai et al., 2025b), adding explicit box supervision with <box_start> / <box_end> further boosts performance: G-Eval rises to *3.67* (Liu et al., 2023b) and ROUGE-L to *51.22* (Ganesan, 2018), showing that structured spatial cues benefit both generation and grounding.

Performance on Visual grounding: The fine-tuned Qwen-2.5-VL-7B outperforms

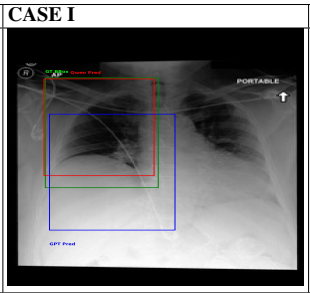
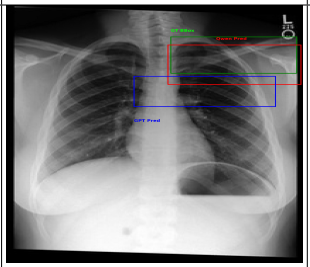
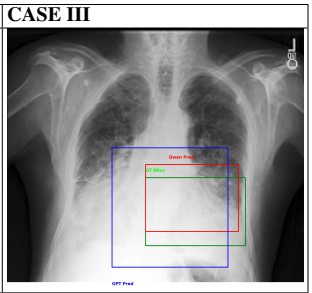
	CASE I	CASE II	CASE III
			
Question	What could be the cause of the atelectasis visible in the right lung?	What might explain the absence of acute osseous abnormalities in the left clavicle area on the X-ray?	What abnormality is visible in the cardiac silhouette area of the Chest X-ray ?
Ground Truth	The atelectasis in the right lung could be due to obstruction of the airways.	The absence of acute osseous abnormalities indicates that there are no recent bone injuries or diseases affecting the left clavicle.	The lung opacity in the left lower lung zone could be due to atelectasis or pulmonary edema.
GPT-4o-mini (Achiam et al., 2023)	The atelectasis in the right lung could be due to several causes, including mucus plugging, foreign body aspiration, pleural effusion, or compression from a mass. Other potential causes include postoperative states, particularly after thoracic surgery, or prolonged bed rest leading to inadequate lung expansion.	The absence of acute osseous abnormalities in the left clavicle area on the X-ray may be explained by factors such as no recent trauma or injury to the area, an absence of underlying pathology like tumors or infections, or a well-healed previous injury. Additionally, the positioning of the clavicle and the quality of the X-ray may influence visibility. Proper alignment and sufficient imaging technique can lead to clear portrayals of normal anatomy without visible abnormalities.	The lung opacity seen in the right lung could be due to several causes, including pneumonia, pulmonary edema, a neoplasm (tumor), or pleural effusion. The specific characteristics (size, shape, borders) of the opacity would further narrow down the differential diagnosis.
Qwen* (Wang et al., 2024)	Atelectasis in the right lung could be due to various reasons including obstruction of the airways, compression from outside the lung, or lack of surfactant.	The absence of acute osseous abnormalities in the left clavicle area indicates that there are no recent bone injuries or diseases affecting this region.	The lung opacity in the left lower lung zone could be due to atelectasis.

Table 6: Questions from GIV-CXR answered by GPT-4o-mini (Achiam et al., 2023) and Qwen* (fine-tuned Qwen-2.5VL (Wang et al., 2024)) on GIV-CXR for grounding. These examples highlight the limitations of LVLMs like GPT-4o-mini (Achiam et al., 2023) in visual grounding with precise answers, which can be effectively improved by fine-tuning with our dataset.

Model	VQA-RAD	MIMIC-CXR
CheXagent (Chen et al., 2024)	2.91	3.02
LLaMA 3.2 11B *	2.98	2.98
MedGemma 4B (Sellergren et al., 2025)	3.49	3.06
Qwen-2.5 VL (Bbox output)*	2.98	2.75

Table 7: G-Eval performance comparison across VQA-RAD and MIMIC-CXR datasets.

GPT-4o-mini (Achiam et al., 2023) on localization with an mIoU of *68.12* (Rezatofighi et al., 2019)., while it's capability improved from 11.93 MIoU (pretrained model) demonstrating strong grounding capacity even with limited training data.

Generalization over other benchmarks: Table 7 reports G-Eval scores for models *trained only on our* GIV-CXR and evaluated on two benchmarks (VQA-RAD and MIMIC-CXR). The fine-tuned LLaMA-3.2-11B attains *2.98/2.98*, matching or exceeding CheXagent on VQA-RAD (*2.91*) and approaching it on MIMIC-CXR (*3.02*); MedGemma-4B remains strongest overall (*3.49/3.06*). Our grounding-oriented Qwen-2.5-VL (bbox) reaches *2.98/2.75*, suggesting a modest trade-off between localization and text generation. Overall, models trained on GIV-CXR generalize well to datasets they never saw during training, achieving performance comparable to specialized systems already tuned on those benchmarks.

6 Limitations

Our dataset inherits biases from the source corpora (MIMIC-CXR / Chest ImaGenome), including potential demographic skews. It is region-imbalanced (lungs $> 55\%$ of QA pairs, with abdomen, spine, and clavicles under-represented) and disease-skewed toward pneumonia (50.2%), which may hinder learning for rarer yet clinically important conditions. The aggressive hallucination filtering—coupled with expert-guided prompt design—removed 18.2% of generated QA pairs, likely eliminating some valid edge cases. Finally, the resource is limited to single-modality Chest X-ray , which may constrain generalizability to

other imaging modalities and multi-modal clinical contexts (e.g., integration with history, laboratory results, or additional imaging).

7 Acknowledgement

We sincerely thank the authors of MIMIC-CXR (Johnson et al., 2019) and ChestImaGenome (Wu et al., 2021) for making their invaluable datasets available. We also acknowledge the broader community of medical imaging research for their contributions and inspiration. We used ChatGPT and Perplexity AI to improve our research workflow and grammatical corrections.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kweon, Jungwoo Oh, Lei Ji, Eric Chang, Tackeun Kim, and 1 others. 2023. Ehrxqa: A multi-modal question answering dataset for electronic health records with chest x-ray images. *Advances in Neural Information Processing Systems*, 36:3867–3880.

Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kweon, Jungwoo Oh, Lei JI, Eric Chang, Tackeun Kim, and 1 others. 2024. Mimic-ext-mimic-cxr-vqa: A complex, diverse, and large-scale visual question answering dataset for chest x-ray images.

Long Bai, Guankun Wang, Mobarakol Islam, Lalithkumar Seenivasan, An Wang, and Hongliang Ren. 2025a. Surgical-vqla++: Adversarial contrastive learning for calibrated robust visual question-localized answering in robotic surgery. *Information Fusion*, 113:102602.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Jean-Christophe Bélisle-Pipon. 2024. Why we need to be careful with llms in medicine. *Frontiers in Medicine*, 11:1495582.

Feilong Chen, Minglun Han, Haozhi Zhao, Qingyang Zhang, Jing Shi, Shuang Xu, and Bo Xu. 2023a. X-llm: Bootstrapping advanced large language models by treating multi-modalities as foreign languages. *arXiv preprint arXiv:2305.04160*.

Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, and 1 others. 2023b. Pali-x: On scaling up a multilingual vision and language model. *arXiv preprint arXiv:2305.18565*.

Zhihao Chen, Yang Zhou, Anh Tran, Junting Zhao, Liang Wan, Gideon Su Kai Ooi, Lionel Tim-Ee Cheng, Choon Hua Thng, Xinxing Xu, Yong Liu, and 1 others. 2023c. Medical phrase grounding with region-phrase context contrastive alignment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 371–381. Springer.

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, and 4 others. 2024. A vision-language foundation model to enhance efficiency of chest x-ray interpretation. *Preprint*, arXiv:2401.12208.

Yeongjae Cho, Taehee Kim, Heejun Shin, Sungzoon Cho, and Dongmyung Shin. 2024. Pretraining vision-language model for difference visual question answering in longitudinal chest x-rays. *arXiv preprint arXiv:2402.08966*.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, and 1 others. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

Nicolas Gaggion, Candelaria Mosquera, Martina Aineseder, Lucas Mansilla, Diego Milone, and Enzo Ferrante. Chexmask database: a large-scale dataset of anatomical segmentation masks for chest x-ray images.

Nicolás Gaggion, Candelaria Mosquera, Lucas Mansilla, Julia Mariel Saidman, Martina Aineseder, Diego H Milone, and Enzo Ferrante. 2024. Chexmask: a large-scale dataset of anatomical segmentation masks for multi-center chest x-ray images. *Scientific Data*, 11(1):511.

Kavita Ganesan. 2018. Rouge 2.0: Updated and improved measures for evaluation of summarization tasks. *arXiv preprint arXiv:1803.01937*.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*.

Xinyue Hu, L Gu, Q An, M Zhang, L Liu, K Kobayashi, T Harada, R Summers, and Y Zhu. 2023. Medical-diff-vqa: a large-scale medical dataset for difference visual question answering on chest x-ray images. *PhysioNet*, 12:13.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR.

Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. 2022a. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022b. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Zhihong Lin, Donghao Zhang, Qingyi Tao, Danli Shi, Gholamreza Haffari, Qi Wu, Mingguang He, and Zongyuan Ge. 2023. Medical visual question answering: A survey. *Artificial Intelligence in Medicine*, 143:102611.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654. IEEE.

Bo Liu, Ke Zou, Liming Zhan, Zexin Lu, Xiaoyu Dong, Yidi Chen, Chengqiang Xie, Jiannong Cao, Xiao-Ming Wu, and Huazhu Fu. 2024b. Gemex: A large-scale, groundable, and explainable medical vqa benchmark for chest x-ray diagnosis. *arXiv preprint arXiv:2411.16778*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: Nlg evaluation using gpt-4 with better human alignment. *Preprint*, arXiv:2303.16634.

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, and 1 others. 2022. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. *Scientific Data*, 9(1):429.

Wimukthi Nimalsiri, Mahela Hennayake, Kasun Rathnayake, Thanuja D Ambegoda, and Dulani Meedeniya. 2023. Cxlseg dataset: Chest x-ray with lung segmentation. In *2023 International Conference On Cyber Management And Engineering (CyMaEn)*, pages 327–331. IEEE.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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.

Peixi Peng, Wanshu Fan, Wenfei Liu, Xing Yang, and Dongsheng Zhou. 2024. Prior-posterior knowledge prompting-and-reasoning for surgical visual question localized-answering. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. 2019. 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.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, and 62 others. 2025. Medgemma technical report. *Preprint*, arXiv:2507.05201.

Zdenko Sonicki. 2024. Large multi-modal models–the present or future of artificial intelligence in medicine? *Croatian Medical Journal*, 65(1):1.

Sergio Tascon-Morales, Pablo Márquez-Neila, and Raphael Sznitman. 2023. Localized questions in medical visual question answering. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 361–370. Springer.

Sergio Tascon-Morales, Pablo Márquez-Neila, and Raphael Sznitman. 2024. Targeted visual prompting for medical visual question answering. In *International Workshop on Applications of Medical AI*, pages 64–73. Springer.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

VALS AI. 2025. Medqa benchmark. Accessed: 2025-02-25.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106.

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, and 1 others. 2021. Chest imagenome dataset for clinical reasoning. *arXiv preprint arXiv:2108.00316*.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, and 1 others. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Ka-Wai Yung, Jayaram Sivaraj, Danail Stoyanov, Stavros Loukogeorgakis, and Evangelos B Mazomenos. 2024. Region-specific retrieval augmentation for longitudinal visual question answering: A mix-and-match paradigm. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 585–594. Springer.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Jiayu Lei, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024a. Radgenome-chest ct: A grounded vision-language dataset for chest ct analysis. *arXiv preprint arXiv:2404.16754*.

Yue Zhang, Wanshu Fan, Peixi Peng, Xin Yang, Dongsheng Zhou, and Xiaopeng Wei. 2024b. Dual modality prompt learning for visual question-grounded answering in robotic surgery. *Visual Computing for Industry, Biomedicine, and Art*, 7(1):9.

Ke Zou, Yang Bai, Zhihao Chen, Yang Zhou, Yidi Chen, Kai Ren, Meng Wang, Xuedong Yuan, Xiaojing Shen, and Huazhu Fu. 2024. Medrg: Medical report grounding with multi-modal large language model. *arXiv preprint arXiv:2404.06798*.

A Appendix

A.1 Dataset Analysis

A.1.1 QA Pair Density across Anatomical Regions

Figure 5 illustrates the distribution of Question-Answer (QA) pairs across different anatomical regions in the curated dataset.

The left and right lungs lead with over 55,000 pairs each, underscoring their importance in diagnosing pulmonary conditions. The cardiac silhouette also shows high density, highlighting its diagnostic relevance. Regions such as the mediastinum, costophrenic angles, and hilar structures are moderately represented. In contrast, areas like the abdomen, spine, clavicles, and trachea have fewer QA pairs (5,000–10,000), suggesting they are less frequently the focus in chest X-rays, while the apical zones and other regions occur the least. This distribution mirrors clinical priorities while also indicating that additional annotations in low-density regions could further enhance the dataset's balance and overall utility.

A.1.2 Findings and diseases mentioned in the dataset

As illustrated in Figure 4 Lung opacity is the most common finding, constituting 26.1% of cases, followed by pleural effusion (14.9%) and pneumothorax (9.1%). Additionally, over 20% of findings fall under "Others," encompassing less common or mixed conditions. These distributions highlight the prevalence of opacities and effusions, aligning with their frequent occurrence in pulmonary

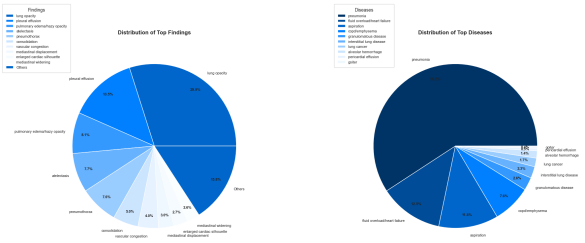


Figure 4: Top findings and diseases in the dataset curated. The left panel illustrates the distribution of the most frequent findings, while the right panel highlights the distribution of the most common diseases mentioned.

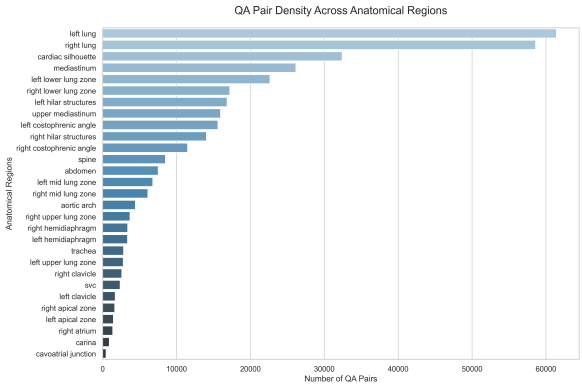


Figure 5: Distribution of curated dataset among regions in an X-ray.

and pleural diseases. Among diseases, pneumonia is the most dominant, accounting for 59.9% of mentions, reflecting its high clinical significance. Fluid overload/heart failure follows at 11.2%, emphasizing the role of cardiac conditions in chest X-ray analysis. Lung cancer and pleural effusion, as primary disease entities, each contribute approximately 1–2% of cases. While the dataset is naturally skewed towards pneumonia, it still captures a diverse range of abnormalities and disease processes relevant to chest imaging.

A.2 Prompts used in the work

Answer Generation Prompt Design

Based on the findings and the question:
CONTEXT: Finding Location: `{bbox}`
Observation Attribute: `{attr}`
Texture Description: `{texture}`
Report Excerpt: `"{phrase}"`
Question: `"{question}"`

Objective: Compose a concise and professional response that clearly explains the significance of the findings. The response should:

- Avoid overly technical terms or speculative language.
- Remain accessible, factual, and aligned with the provided report details.
- Focus solely on the clinical findings relevant to the question.
- Avoid assumptions, opinions, or additional context beyond what is directly supported by the report.

Guidelines:

- DO NOT respond with opinions or personal reasoning.
- Stick strictly to the provided information when answering the question.
- Ensure the response does not reference the report explicitly (e.g., avoid phrases like "As mentioned in the report," "The report states...," or "Not provided in the report").
- Maintain a professional tone, answering as a medical expert interpreting the X-ray findings.
- The report is only for generating answers, but its details should not appear in the response.
- Provide answers in the same order as the corresponding questions.

Expected Output Format (JSON):

```
{
    "answers": [".....",..]
}
```

Question Generation Prompt Design

Given the following chest X-ray findings for a specific zone (i.e., the finding location):
CONTEXT: Finding Location: `{bbox}`
Observation Attribute: `{attr}`
Texture Description: `{texture}`
Report Excerpt: `"{phrase}"`

Draft simple and formal questions that a person might ask to understand the condition and findings about the zone from the given X-ray. Avoid overly technical phrasing and ensure that the questions directly relate to the provided details.
The questions should focus on:

- Identifying any abnormalities visible in the given zone.
- Determining the location of the abnormality.
- Understanding the cause of the abnormality.
- Locating suspicious areas in the X-ray region.
- Identifying potential diseases (if explicitly mentioned in the report).
- Understanding texture information in the region (if present in the report).

Guidelines:

- Frame questions strictly based on the given data.
- Do not mention the presence of the report in the questions.
- Avoid subjective or speculative phrasing such as:
  - "Is my condition...?"
  - "Should I be concerned about...?"

- "Why is this happening?"

- Do not frame generic questions for the entire X-ray; questions must be region-specific.
- Do not assume prior scans or temporal comparisons.

Expected Output Format (JSON):

```
{
    "questions": [".....",..]
}
```

Medical Expert Hallucination Detection Prompt

You are a medical expert evaluating whether questions and answers about the chest X-ray contain hallucinations or not, based on the given report.
GIVEN MEDICAL REPORT INFORMATION:
Region of interest: `{bbox_name}`

Attributes: `{json.dumps(attr_list)}`
Texture cues: `{json.dumps(texture_cues)}`
Report phrases: `{json.dumps(phrases)}`

The report is formatted with pipes (|) to separate different attributes:

- anatomicalfinding|yes|X means the anatomical finding X is present
- anatomicalfinding|no|X means the anatomical finding X is absent
- disease|yes|X means disease X is present
- disease|no|X means disease X is ruled out
- texture|yes|X means texture X is present
- texture|no|X means texture X is absent
- nlp|yes|normal means report describes this area as normal
- nlp|yes|abnormal means report describes this area as abnormal

QUESTION-ANSWER PAIRS TO EVALUATE:
`{json.dumps(qa_pairs, indent=2)}`

For each question-answer pair, determine if the information in answer AND question is supported by the report.
A hallucination is when a question or answer:

1. States something as fact that isn't mentioned in report
2. Contradicts information in report
3. Makes claims about findings that aren't supported in report

There are certain entries which are also not useful for the diagnosis. These are mainly reverse negation of the findings OR reason for normal findings. For example, "What might be the reason for the absence of pneumothorax in the right lung?" these kind of questions or answers are not useful per radiologists as they ask for the reason for normal findings. So, for these kind of entries, classify them as hallucination and respond with the corresponding explanation.

In the cases where, there are multiple facts in a report, and if a question or the corresponding answer is supported by any of the facts, then it is not a hallucination. Which means, not every fact in the report needs to be supported by the question or the answer. On the other hand, if a question or the answer is not supported by any of the facts from the report, then it is a hallucination.
STRICTLY MAINTAIN THE ORDER OF THE QUESTION-ANSWER PAIRS.
Output a JSON array where each element is an object with these fields:

- "is_hallucination": 0 if the question and answer are fully supported by the report, 1 if it contains any hallucination (Also the case of reasonings for normal findings as specified above)
- "explanation": Brief explanation of your decision
- "score": A score between 0 and 1 indicating your confidence level in this decision
  - 1.0: Absolute certainty (clear evidence in report)
  - 0.8-0.9: High confidence (strong indications in report)
  - 0.5-0.7: Moderate confidence (some indications but not explicit)
  - 0.1-0.4: Low confidence (limited information available)

Remember that, based on your confidence of the decision, you can assign a score between 0 and 1 to your decision.
These confidence scores will be essential for human medical expert validation, so please be precise and thorough in your analysis.
Return ONLY the JSON array, nothing else.

Expert Evaluator Assessment Prompt

As an expert evaluator, your task is to assess the accuracy and precision of the model's response compared to the provided ground truth. Your evaluation should consider the relevance, completeness, and correctness of the response.
`{question_part}`
Ground Truth Answer: `"{reference}"`
Model Response: `"{prediction}"`

Please rate the model response on a scale from 1 to 5
Criteria:
Correctness (1-5) — Does the answer factually align with the provided ground truth?
Provide only the numerical score.