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

Sreevaatsav Bavana[1] Adit Rushil Potta[1] Sai Amrit Patnaik[2] Nidhi Goyal[1]
[1]Mahindra University [2]IIIT Hyderabad
Hyderabad, Telangana, India
bavanasreevaatsav1@gmail.com aditushipot075@gmail.com
saiamritp@gmail.com nidhi.goyal@mahindrauniversity.edu.in
https://github.com/your-project-page

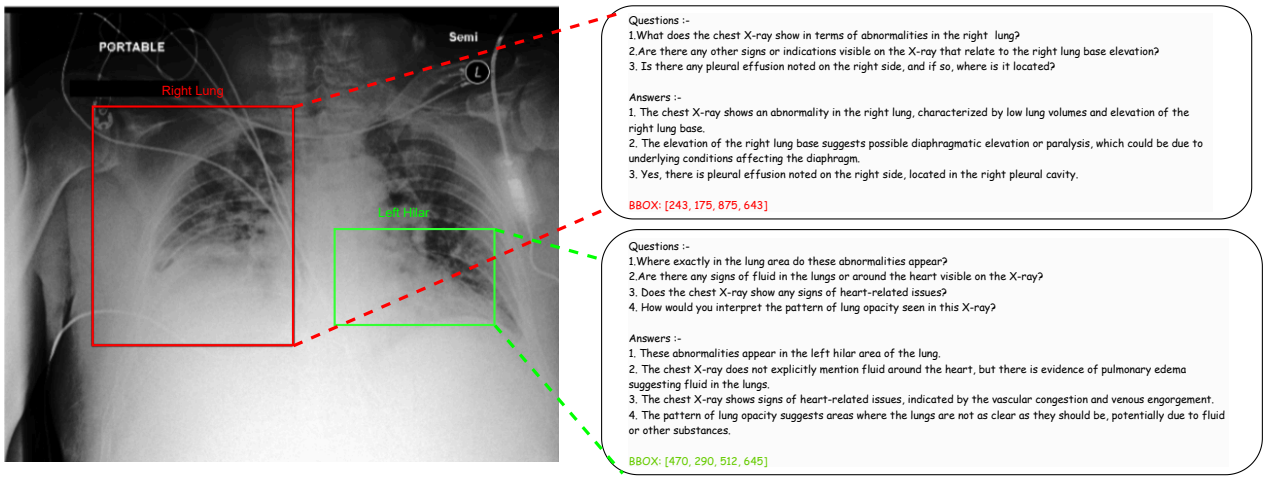


Figure 1. Overview of GIV-CXR: A grounded chest X-ray VQA benchmark providing region-level question-answer pairs with bounding box annotations built on Chest ImaGenome. Each anatomical region receives systematic coverage across five reasoning dimensions with quantified spatial grounding evaluation.

Abstract

Visual question answering in medical imaging requires models to ground predictions in anatomical locations for clinical verification, yet existing benchmarks lack systematic spatial reasoning evaluation infrastructure. While recent datasets include bounding boxes as metadata, no standardized metrics exist to quantify whether models correctly associate predictions with image regions. Additionally, opportunistic question generation yields uneven anatomical coverage, potentially encouraging spurious correlations rather than robust spatial understanding. We introduce GIV-CXR*, a grounded chest X-ray VQA benchmark enabling quantified localization assessment. The benchmark comprises 355,293 question-answer pairs systematically distributed across 36 anatomical structures, with explicit bounding box linkage enabling evaluation via Intersection over Union metrics. We employ structured generation across five reasoning dimensions to ensure comprehensive coverage, implement automated hallucination filtering and radiologist validation for quality control, and provide bias analysis quantifying performance variation across anatomical locations and disease types. Experimental results demonstrate that explicit spatial supervision substantially improves localization while maintaining answer quality, and that our framework generalizes to assess models trained on diverse VQA datasets. This work provides the vision community with standardized infrastructure for developing spatially-aware vision-language models applicable to safety-critical medical domains where localization precision is essential for clinical trust and deployment.*

1. Introduction

Visual question answering in medical imaging requires models to ground predictions in precise anatomical locations. When diagnosing pneumonia, radiologists identify not only the pathology but its exact location, spatial extent, and relationship to surrounding structures. This spatial grounding enables clinical verification, supports differential diagnosis, and builds trust in automated systems [34, 41]. However, current evaluation paradigms for medical VQA focus predominantly on textual accuracy while providing limited infrastructure to assess whether models correctly ground their reasoning in anatomical regions [22, 24].

Recent vision-language models demonstrate strong performance on medical VQA tasks [2, 5, 10, 21, 23, 26, 37]. Yet these models exhibit a systematic limitation: they can generate factually correct answers while attending to anatomically incorrect regions. For instance, a model may correctly identify pleural effusion while localizing attention to the contralateral hemithorax, or describe cardiomegaly while focusing on pulmonary vasculature. Standard VQA evaluation metrics, which assess only textual correctness, cannot detect such localization errors [13].

Existing medical VQA benchmarks exhibit two key limitations. First, while several datasets include bounding box annotations [7, 25, 44], these serve primarily as metadata rather than evaluation targets. The field lacks standardized quantitative metrics analogous to Intersection over Union in object detection [36] to measure whether models correctly associate predictions with anatomical locations. Second, question generation in prior work distributes opportunistically based on visible pathology without ensuring systematic coverage [3, 12]. This design may encourage models to exploit spurious correlations between pathologies and image regions rather than learning robust visual features generalizable across diverse anatomical contexts.

We present GIV-CXR, a benchmark enabling quantified spatial reasoning evaluation in medical VQA. The dataset comprises 355,293 question-answer pairs distributed across 36 anatomical structures in 20,534 chest radiographs, with each instance explicitly linked to bounding boxes enabling direct localization assessment via mean IoU metrics. Our key design principle is region-centric generation: we systematically ensure each anatomical region receives balanced coverage across five reasoning types, including abnormality detection, spatial localization, causal reasoning, texture characterization, and differential diagnosis. This structure encourages models to learn generalizable visual representations independent of specific pathological presentations.

We establish rigorous quality control through automated hallucination filtering and independent validation from board-certified radiologists, achieving 82.4% acceptance with substantial inter-rater agreement. We conduct comprehensive bias analysis using bootstrap permutation testing, revealing that location-level performance variation affects models more strongly than disease-level bias. Experimental evaluation demonstrates that explicit spatial supervision substantially improves localization accuracy while maintaining answer generation quality. Our evaluation framework applies to models trained on arbitrary VQA corpora, positioning GIV-CXR as a standardized testbed for spatial reasoning assessment. We release the complete dataset and evaluation code[1]. Our main contributions are:

- We establish standardized localization metrics for medical VQA, demonstrating that spatial supervision improves grounding while maintaining answer quality and enabling systematic assessment of arbitrary models.
- We introduce region-centric generation ensuring balanced coverage across anatomical structures and reasoning types, promoting generalizable visual features over spurious correlations
- We provide rigorous quality assurance through automated filtering and radiologist validation, with comprehensive bias analysis quantifying performance variation.

2. Related Works

2.1. Medical Visual Question Answering

Medical VQA has progressed from small manually annotated datasets to large-scale automated benchmarks. Early efforts like VQA-RAD [21] (315 pairs) and SLAKE [23] (14K pairs with knowledge graphs) established foundational protocols, while PathVQA [17] (33K pairs) targeted histopathology reasoning. Recent work pursues scale through automated generation: MIMIC-CXR-VQA [3] produces 500K pairs via templates, Diff-VQA [12] introduces temporal reasoning across paired images, and GEMeX [25] achieves 1.6M pairs across 151K chest X-rays with bounding box annotations. However, these bounding boxes serve primarily as construction metadata rather than evaluation targets. No existing medical VQA benchmark provides standardized spatial metrics analogous to IoU in object detection [36], preventing systematic assessment of whether models correctly associate textual predictions with anatomical regions.

2.2. Visual Grounding in Medical Imaging

Several efforts target spatial understanding through grounding and segmentation, yet remain disconnected from question answering evaluation. Chest ImaGenome [44] provides 242K X-rays with scene graphs mapping structures to descriptions, while MS-CXR [7] advances phrase grounding through vision-language alignment. Medical contrastive learning work [9] demonstrates improved localization on referring expressions. RadGenome-CT [48] extends to vol-

[1] Available upon publication

Dataset	# Imgs	# QAs	Modality	Annotations	Bbox	Region QA	Grounding Metrics	Quality Validation
Medical VQA Datasets								
VQA-RAD [21]	3.5K	315	Multi	QA	×	×	×	Manual curation
SLAKE [23]	14K	14K	Multi	QA + KG	×	×	×	Expert review
PathVQA [17]	33K	33K	Histo	QA	×	×	×	Pathologist review
MIMIC-CXR-VQA [3]	377K	500K	CXR	QA	×	×	×	Template-based
Diff-VQA [12]	700K	700K	CXR	QA (paired)	×	×	×	Quality control
GEMeX [25]	151K	1.6M	CXR	QA + BBox	✓	✓	×	Radiologist prompts
Grounding & Segmentation Datasets								
Chest ImaGenome [44]	242K	—	CXR	BBox + Scene	✓	×	×	Radiologist review
MS-CXR [7]	1.1K	1.1K	CXR	BBox + QA	✓	✓	Qualitative	Manual annotation
ChestX-ray8 [43]	112K	—	CXR	BBox (partial)	✓	×	×	Weak supervision
CXL-Seg [32]	243K	—	CXR	Mask	×	×	×	Segmentation labels
CheXmask [15]	657K	—	CXR	Mask (1024^2)	✓	×	×	Multi-center validation
GIV-CXR (Ours)	**20.5K**	**355K**	**CXR**	**QA + BBox**	✓	✓	**mIoU**	**82.4% radiologist**

Table 1. Comparison of medical imaging datasets for visual question answering and spatial grounding. GIV-CXR uniquely integrates region-specific question answering with quantified grounding assessment.

umetric 3D grounding, while CXLSeg [32] and CheXmask [15] provide pixel-level segmentation. Despite these resources, medical grounding evaluation remains qualitative, relying on visual inspection rather than quantitative metrics. This contrasts with general-domain grounding benchmarks [20, 46] that routinely report pointing accuracy and IoU scores, establishing reproducible protocols absent in medical imaging.

2.3. Automated Generation and Quality Control

LLM-based generation reduces expert annotation costs [3, 25] but introduces hallucinations and factual inconsistencies [13, 40]. Quality approaches vary in rigor: GEMeX employs radiologist-in-the-loop prompt refinement, while template constraints [3] limit errors at the cost of diversity. G-Eval [28] offers scalable evaluation requiring calibration against human judgment. A critical gap in existing work is lack of transparent error analysis. While some datasets report overall acceptance rates, few provide detailed taxonomies of rejection reasons, inter-annotator agreement statistics, or quantitative bias analysis across anatomical regions and pathology types.

2.4. Evaluation Metrics for Vision-Language Tasks

VQA evaluation employs answer-matching metrics (accuracy, BLEU [33], ROUGE [16], BERTScore [47]) and medical-specific measures (CheXbert [39], RadGraph [18]), which assess textual correctness without spatial reasoning signal. Object detection evaluates localization through IoU [36], with general-domain grounding work [20, 27, 46] demonstrating that explicit spatial supervision improves both localization and reasoning. Medical grounding evaluation remains qualitative [7], limiting reproducibility and cross-model comparison. We introduce mIoU-based spatial evaluation for medical VQA, treating localization as a first-class metric alongside answer accuracy and bridging evaluation paradigms from object detection and question answering.

3. Dataset Overview

GIV-CXR comprises 355,293 question-answer pairs across 20,534 MIMIC-CXR chest radiographs [19], with 81,257 bounding boxes spanning 36 anatomical structures. Table 2 summarizes core characteristics. Our generation framework systematically covers five reasoning dimensions per anatomical region: abnormality detection, spatial localization, causal reasoning, texture characterization, and differential diagnosis, yielding 78% of regions with at least four reasoning aspects. Clinical findings follow realistic long-tail distributions (lung opacity 50.2%, pleural effusion 25.3%, atelectasis 16.8%), with 28.6% addressing normal regions to mitigate normality bias. Bounding boxes span diverse spatial scales: 34% small, 42% medium, and 24% large regions, enabling evaluation across varying localization difficulty.

Quality assurance employed automated hallucination filtering removing 18.2% of generated pairs (error taxonomy: 58.9% normality bias, 12.7% terminology errors, 9.9% negation errors), followed by independent radiologist validation on 2,487 pairs achieving 82.4% acceptance (95% CI: [80.9, 83.9]) with 86.4% inter-annotator agreement. Splits comprise training (150K QAs) and test (7,500 QAs) sets with proportional anatomical and pathology distributions. Tables 3 and 4 present baseline performance, demonstrating substantial improvements through fine-tuning. Dataset and

Characteristic	Value
Images / QA pairs / BBoxes	20.5K / 355K / 81K
Anatomical structures	36
QAs per image / bbox	17.3 / 4.37
Question / Answer length (words)	11.2 / 8.4
Detection / Causal / Localization	23.8% / 21.2% / 19.4%
Texture / Differential diagnosis	18.9% / 16.7%
Multi-aspect coverage (mean)	4.5 per region
Abnormal / Normal findings	71.4% / 28.6%
Top finding (lung opacity)	50.2%
BBox size: Small / Medium / Large	34% / 42% / 24%
Radiologist acceptance rate	82.4% [80.9, 83.9]
Hallucination rejection rate	18.2%

Table 2. Dataset statistics for GIV-CXR. Multi-aspect coverage indicates mean reasoning dimensions per region. Bbox sizes: small (< 20K px^2), medium (20K–60K px^2), large (> 60K px^2).

Model	G-Eval
MedGemma-4B [38]	3.47
CheXagent [11]	3.22
GPT-4o-mini [1]	3.13
Qwen-2.5-VL-7B (bbox output)†	2.97
Qwen-2.5-VL-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-2.5-VL-7B (bbox input)*	3.67

Table 3. G-Eval scores on GIV-CXR. * = fine-tuned; † = pre-trained.

Model	CheXbert κ	RadGraph F1
CheXagent	0.25	0.22
MedGemma-4B	0.30	0.39
GPT-4o-mini	0.43	0.47
LLaMA-3.2-11B*	0.62	0.63
Qwen-2.5-VL-7B (bbox output)*	0.58	0.59
Qwen-2.5-VL-7B (bbox input)*	0.44	0.43

Table 4. Clinical metrics on GIV-CXR. CheXbert [39] κ and RadGraph [18] F1 per QA pair. * = fine-tuned.

code will be released under CC-BY-NC 4.0 license upon publication.

4. Dataset Curation

4.1. Gold Set Creation through Expert-Guided Prompt Refinement

Our generation pipeline employs a validated gold set establishing quality standards for subsequent data creation. We developed initial prompts covering five reasoning dimensions (detection, localization, causal reasoning, texture analysis, differential diagnosis) across 36 anatomical structures. Three board-certified radiologists iteratively refined these prompts to align with authentic clinical questioning patterns, transforming research-oriented phrasing into natural formulations reflecting diagnostic practice. Radiologists demonstrated contextual adaptability across clinical settings: teaching rounds emphasize pathophysiological mechanisms, patient consultations use accessible language, and emergency triage focuses on critical decisions. This ensures models learn robust visual representations applicable across diverse deployment contexts.

Using refined prompts, we generated 2,500 gold set question-answer pairs with spatial grounding via Grok-2 [45], stratified across anatomical regions and pathology distributions. Each entry underwent independent validation by three radiologists assessing medical accuracy, spatial grounding appropriateness, question naturalness, and clinical relevance. The gold set achieved 84.2% acceptance, with rejected cases providing insights into generation failures. This validated set serves three functions: establishing quality thresholds for full-scale generation, informing automated hallucination detection rules, and guiding prompt calibration. This design ensures the complete dataset inherits expert-validated grounding and natural question-answering patterns.

4.2. Responsible Large-Scale Generation

Full generation employed Grok-2 with gold-calibrated prompts across 20,534 images, implementing safeguards aligned with medical AI ethics [34, 41]. Medical safety filtering removed emergency care questions and treatment recommendations beyond diagnostic scope. Privacy protection detected patient identifiers including names, dates, and institutional details [30]. Bias monitoring tracked distribution across anatomical regions, pathology types, and demographic considerations to prevent systematic under-representation [8]. Fairness auditing verified consistent question difficulty across pathology prevalence levels, avoiding scenarios where rare conditions receive specialist-level questions while common findings receive basic queries. Generation parameters constrained over-representation of visually obvious abnormalities, ensuring subtle findings receive adequate coverage for comprehensive visual feature learning.

4.3. Automated Hallucination Detection

Automated quality control filtered 18.2% of generated output (79,124 from 434,417 pairs) before human validation. Our detection framework employed five rule-based classifiers calibrated against gold set rejection patterns. Normality-based hallucination detection (58.9% of rejec-

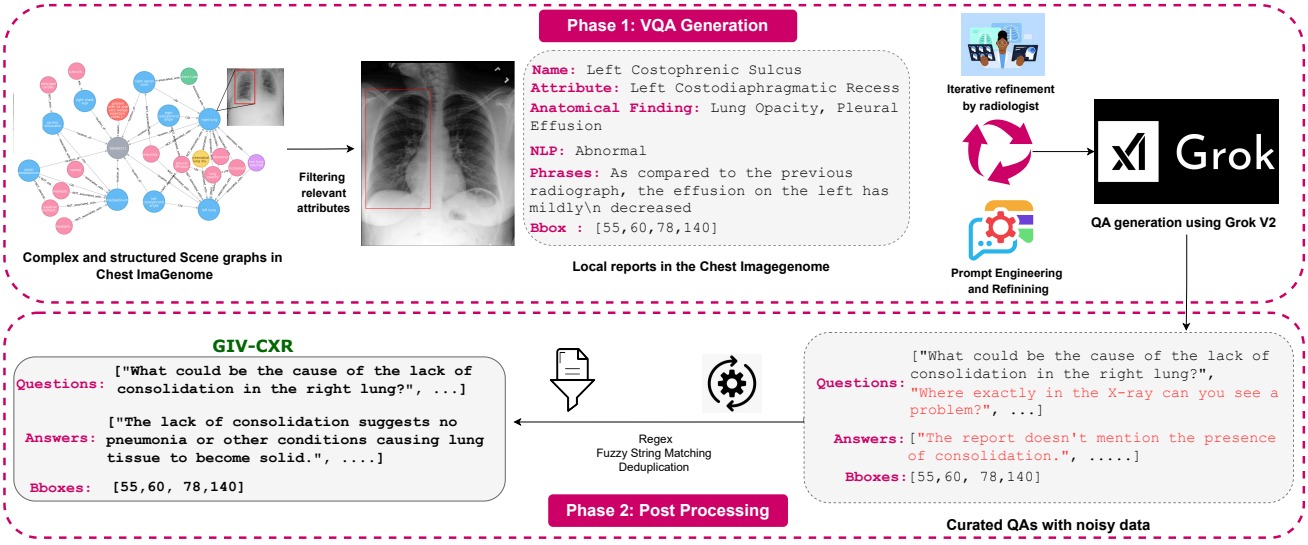


Figure 2. Curation pipeline for GIV-CXR.

tions) identified questions claiming abnormalities in normal regions by cross-referencing Chest ImaGenome labels [44] with generated content using clinical ontology matching [6]. Medical terminology validation (12.7%) flagged anatomically impossible descriptions through UMLS [6]. Negation error detection (9.9%) identified incorrect negative descriptors via dependency parsing [31]. Spatial confusion detection (1.0%) caught orientation mistakes and implausible locations. Logical inconsistency detection (8.0%) identified contradictions using semantic similarity [35]. The remaining 9.5% of rejections involved multi-category violations not attributable to a single classifier. Detection thresholds matched gold set rejection rates, ensuring alignment with expert standards. Questions with multiple error flags were removed entirely, prioritizing reliability over scale.

4.4. Expert Validation and Bias Analysis

Three radiologists independently validated 2,487 stratified random samples (0.70% of retained dataset) across anatomical regions, pathology types, and complexity levels, achieving 82.43% acceptance (95% CI: [80.94%, 83.92%]) with 86.4% inter-annotator agreement (Fleiss $\kappa = 0.78$). Rejection causes included subtle medical inaccuracies (9.2%), ambiguous grounding (5.1%), and complex phrasing (3.3%). Comparison with gold set acceptance (84.2%) confirmed generation consistency. Bootstrap permutation testing (10,000 iterations) revealed location-dependent performance variation in 50% of models (Cohen's $d = 0.27$), with lower accuracy for apical and peripheral regions versus central structures, while disease-level performance remained balanced ($d = 0.08$). These findings inform training augmentation strategies for under-represented spatial regions, critical for developing vision models with robust generalization across anatomical contexts.

Table 5. 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 observed hallucination patterns included normality overgeneralization and unsupported negations, with common terms: *normal*, *lung*, *left*, *absence*, *pleural*. Post-filtering yielded high-quality, clinically relevant pairs enabling robust visual grounding evaluation.

5. Experiments and Results

5.1. Experimental Setup

Evaluation metrics. We employ G-Eval [29] as our primary metric for answer quality, complemented by standard NLG metrics (BLEU [33], ROUGE-1/L [16], BERTScore [47]). For spatial reasoning evaluation, we adopt mean IoU [36] following object detection conventions, enabling quantified assessment of localization accuracy. Clinical validity is measured through CheXbert [39] (Cohen's κ) and RadGraph [18] (entity F1) applied per question-answer pair.

Baselines. We evaluate pretrained vision-language models: CheXagent [11], MedGemma-4B [38], and GPT-4o-

Split	# QAs	# Imgs	# BBoxes
Total	355,293	20,534	81,257
Train	150,000	19,194	66,615
Test	7,500	1,000	3,916

Table 6. Dataset splits. Train and test sets maintain proportional anatomical and pathology distributions.

Model	G-Eval	ROUGE-L	BERTScore	mIoU
Pretrained (Zero-shot)				
CheXagent	3.22	27.05	85.12	—
MedGemma-4B	3.47	36.84	87.21	—
LLaMA-3.2-11B	2.18	19.43	82.34	—
Qwen-2.5-VL-7B	2.97	24.68	84.57	11.93
GPT-4o-mini	3.13	—	—	39.04
Fine-tuned on GIV-CXR				
LLaMA-3.2-11B (50K)	3.86	70.71	92.48	—
Qwen-BBox-Output (150K)	3.98	51.22	88.63	**68.12**
Qwen-BBox-Input (150K)	3.67	62.36	94.26	—

Table 7. Performance on GIV-CXR test set. Fine-tuning yields substantial improvements in both text generation and spatial grounding. Qwen-BBox-Output achieves 68.12 mIoU versus GPT-4o-mini's 39.04, demonstrating effective learning of region-specific visual features. Dashes indicate metrics not applicable to a given model configuration.

Model	CheXbert (κ)	RadGraph (F1)
CheXagent	0.25	0.22
MedGemma-4B	0.30	0.39
LLaMA-3.2-11B	0.28	0.26
Qwen-2.5-VL-7B	0.35	0.33
LLaMA-3.2-11B (fine-tuned)	**0.62**	**0.63**
Qwen-BBox-Output	0.58	0.59
Qwen-BBox-Input	0.44	0.43

Table 8. Clinical validation metrics. Fine-tuned models achieve substantial clinical agreement, with rankings consistent across evaluation dimensions.

Model	VQA-RAD	MIMIC-CXR
CheXagent	2.91	3.02
MedGemma-4B	3.49	3.06
LLaMA-3.2-11B (GIV-CXR)	2.98	2.98
Qwen-BBox-Output (GIV-CXR)	2.98	2.75

Table 9. Cross-dataset generalization. Models fine-tuned only on GIV-CXR achieve competitive performance on external benchmarks without exposure during training.

mini [1]. These models demonstrate strong performance on general multimodal benchmarks but lack explicit spatial grounding supervision.

Fine-tuned models. We fine-tune LLaMA-3.2-11B [42] on 50K question-answer pairs for text generation without spatial supervision. For Qwen-2.5-VL-7B [4], we train two variants on 150K pairs: Qwen-BBox-Output predicting answers with bounding boxes, and Qwen-BBox-Input using explicit box tokens to reinforce region-specific visual feature learning.

Data splits. Training (150K QAs) and test (7,500 QAs) sets maintain proportional anatomical and pathology distributions.

5.2. VLMs Performance

Table 7 presents results on GIV-CXR test set. Pretrained models underperform on region-specific reasoning despite general capabilities. CheXagent achieves 3.22 G-Eval and 27.05 ROUGE-L, while MedGemma-4B reaches 3.47 G-Eval and 36.84 ROUGE-L. Unfine-tuned LLaMA-3.2-11B (2.18 G-Eval) and Qwen-2.5-VL (2.97 G-Eval) confirm insufficient spatial grounding in pretrained representations.

Fine-tuning dramatically improves performance. LLaMA-3.2-11B achieves 3.86 G-Eval and 70.71 ROUGE-L (40-50 point improvement), demonstrating that modest high-quality supervision enables strong medical reasoning. Critically, Qwen-BBox-Output reaches 68.12 mIoU, outperforming GPT-4o-mini (39.04 mIoU) by 74% and improving 5.7× over pretrained baseline (11.93 mIoU). This validates that explicit spatial supervision improves localization while maintaining competitive text generation (3.98 G-Eval), confirming spatial grounding and answer quality as complementary learning objectives rather than competing trade-offs.

Clinical validation (Table 8) shows fine-tuned LLaMA-3.2-11B achieving highest agreement (κ=0.62, RadGraph F1=0.63), with Qwen-BBox-Output following closely (κ=0.58, F1=0.59). Model rankings correlate across all metrics, confirming alignment between text quality and clinical validity.

Cross-dataset evaluation (Table 9) demonstrates generalization. Models trained exclusively on GIV-CXR achieve competitive performance on VQA-RAD and MIMIC-CXR without additional tuning. Fine-tuned LLaMA-3.2-11B reaches 2.98 G-Eval on both benchmarks, matching CheXagent (2.91/3.02). This confirms our region-centric design promotes generalizable visual representations rather than dataset-specific pattern recognition.

5.3. Bias Analysis

We quantify bias using bootstrap permutation testing [14] (10,000 iterations) with median-split stratification. Table 10 presents results across disease and location dimensions.

Model	Disease-Level		Location-Level	
	Δ	p	Δ	p
CheXagent	+0.082	0.312	+0.194	**0.021**
MedGemma-4B	+0.045	0.567	+0.370	**<0.001**
GPT-4o-mini	−0.450	**0.002**	+0.083	0.289
LLaMA-3.2-11B	−0.010	0.899	+0.142	**0.047**
Qwen-BBox-Output	+0.062	0.421	+0.089	0.234
Qwen-BBox-Input	+0.151	**0.045**	+0.118	0.156

Table 10. Bias analysis. Δ = performance difference (common minus rare); bold indicates significance ($p < 0.05$). Location-level bias (50% of models) exceeds disease-level bias (33%), with Qwen-BBox-Output showing no significant bias at either level.

Disease-level bias affects only 2/6 models (33%). GPT-4o-mini exhibits inverse bias favoring rare diseases ($\Delta = -0.450$, $p = 0.002$), while Qwen-BBox-Input favors common diseases ($\Delta = +0.151$, $p = 0.045$). Fine-tuned LLaMA shows no significant bias ($\Delta = -0.010$, $p = 0.899$).

Location-level bias is more prevalent, affecting 3/6 models (50%) with stronger effects. MedGemma-4B shows strongest bias ($\Delta = +0.370$, $p < 0.001$, Cohen's $d = 0.27$), followed by CheXagent and LLaMA with smaller effects. Location-level bias is 1.25× stronger than disease-level bias, indicating spatial distribution imbalance poses greater fairness challenges than pathology prevalence for vision models learning anatomical representations.

Critically, Qwen-BBox-Output exhibits no significant bias at either level ($p > 0.2$), suggesting explicit spatial supervision serves as an implicit debiasing mechanism by grounding predictions in visual evidence rather than statistical priors. This finding offers important architectural insights for developing fair medical vision systems through evidence-based reasoning.

5.4. Discussion

Our experiments reveal three key findings for medical vision research. First, region-specific supervision is essential for spatial reasoning. Pretrained models underperform on fine-grained localization despite strong general capabilities, with 40-50 point ROUGE-L improvements after targeted fine-tuning demonstrating the importance of anatomical grounding in learned visual representations. Second, explicit spatial supervision enhances both localization and generation. Qwen-BBox-Output achieves 68.12 mIoU (74% improvement over GPT-4o-mini) while maintaining competitive answer quality, validating spatial grounding as a learnable visual capability complementary to text generation. Third, region-centric design promotes generalizable visual features. Models trained exclusively on GIV-CXR transfer effectively to external benchmarks without additional tuning, confirming our systematic anatomical coverage captures robust spatial reasoning patterns rather than dataset-specific correlations.

Bias analysis reveals location-level performance variation (50% of models affected) exceeds disease-level bias (33%), with stronger effect sizes. This asymmetry suggests spatial distribution imbalance in medical imaging datasets warrants greater attention in fairness research than pathology prevalence alone. The finding that bounding box supervision eliminates both bias types in Qwen-BBox-Output indicates architectural approaches encouraging visual evidence grounding may inherently promote fairness in learned representations.

6. Conclusion

We present GIV-CXR, enabling quantified spatial reasoning evaluation through 354,293 question-answer pairs across 36 anatomical structures with mIoU-based localization metrics. Our region-centric generation promotes generalizable visual features over spurious correlations, achieving 5.7× localization improvement through spatial supervision while maintaining answer quality. Models generalize effectively to external benchmarks, providing the vision community with infrastructure for spatially-grounded vision-language understanding in safety-critical applications.

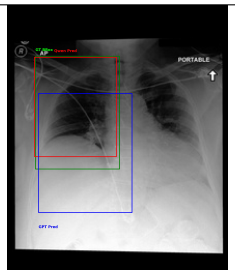
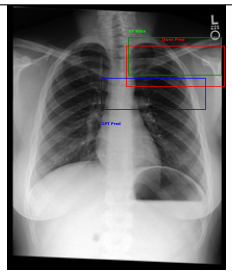
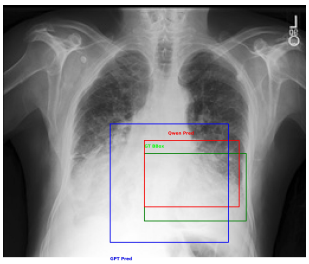
7. Limitations

GIV-CXR focuses on frontal chest radiographs from MIMIC-CXR, limiting diversity in imaging views and patient demographics for visual representation learning. Region-level bounding boxes provide coarser localization than pixel-precise segmentation. Automated filtering rejected 18.2% of validated samples, indicating quality limitations. Location-level bias affects 50% of models, suggesting learned visual features may not generalize uniformly across anatomical regions.

8. Ethical Considerations

GIV-CXR uses IRB-approved, de-identified MIMIC-CXR data requiring PhysioNet credentialing. Automated filters removed patient identifiers and emergency content. This benchmark supports research in vision-language understanding; clinical deployment requires independent validation and regulatory approval. Comprehensive bias analysis enables fairness-aware model development. Released under CC-BY-NC 4.0 license.

Table 11. Qualitative comparison showing GPT-4o-mini and fine-tuned Qwen-2.5VL responses. Fine-tuning on GIV-CXR improves both spatial grounding precision and answer specificity.

	CASE I	**CASE II**	**CASE III**
Image			
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?	What abnormality is visible in the cardiac silhouette area?
Ground Truth	The atelectasis could be due to obstruction of the airways.	The absence indicates no recent bone injuries or diseases affecting the left clavicle.	The lung opacity in the left lower lung zone could be atelectasis or pulmonary edema.
GPT-4o-mini	The atelectasis could be due to several causes, including mucus plugging, foreign body aspiration, pleural effusion, or compression from a mass.	The absence may be explained by no recent trauma or injury, absence of underlying pathology, or a well-healed previous injury.	The lung opacity seen in the right lung could be due to several causes, including pneumonia, pulmonary edema, a neoplasm, or pleural effusion.
Qwen (fine-tuned)	Atelectasis could be due to obstruction of the airways, compression from outside the lung, or lack of surfactant.	The absence indicates no recent bone injuries or diseases affecting this region.	The lung opacity in the left lower lung zone could be due to atelectasis.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 4, 6

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2

[3] Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kweon, Jungwoo Oh, Lei JI, Eric Chang, Tackeun Kim, et al. Mimic-ext-mimic-cxr-vqa: A complex, diverse, and large-scale visual question answering dataset for chest x-ray images, 2024. 2, 3

[4] 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, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. 6

[5] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2502.13923*, 2025. 2

[6] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl_1):D267–D270, 2004. 5

[7] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision-language processing. In *European Conference on Computer Vision*, pages 1–21. Springer, 2022. 2, 3

[8] Irene Y Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. Can ai help reduce disparities in general medical and mental health care? *AMA Journal of Ethics*, 21(2):167–179, 2019. 4

[9] 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, et al. Medical phrase grounding with region-phrase context contrastive alignment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 371–381. Springer, 2023. 2

[10] Zhihong Chen, Maya Varma, Justin Xu, Magdalini Paschali, Dave Van Veen, Andrew Johnston, Alaa Youssef, Louis Blankemeier, Christian Bluethgen, Stephan Altmayer, et al. A vision-language foundation model to enhance efficiency of

chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*, 2024. 2

[11] 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. 4, 5

[12] Yeongjae Cho, Taehee Kim, Heejun Shin, Sungzoon Cho, and Dongmyung Shin. Pretraining vision-language model for difference visual question answering in longitudinal chest x-rays. *arXiv preprint arXiv:2402.08966*, 2024. 2, 3

[13] Arash Dehghan, Siddharth Ramachandran, Nassir Navab, Shadi Albarqouni, and Paul A Keane. The potential and pitfalls of large language models in medical diagnosis. *The Lancet Digital Health*, 6(5):e318–e321, 2024. 2, 3

[14] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994. 6

[15] Nicolás Gaggion, Candelaria Mosquera, Lucas Mansilla, Julia Mariel Saidman, Martina Aineseder, Diego H Milone, and Enzo Ferrante. Chexmask: a large-scale dataset of anatomical segmentation masks for multi-center chest x-ray images. *Scientific Data*, 11(1):511, 2024. 3

[16] Kavita Ganesan. Rouge 2.0: Updated and improved measures for evaluation of summarization tasks. *arXiv preprint arXiv:1803.01937*, 2018. 3, 5

[17] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020. 2, 3

[18] Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duber, Tan Pham, Thomas Watber, Matthew P Lungren, Andrew Y Ng, Curtis P Langlotz, and Pranav Rajpurkar. Radgraph: Extracting clinical entities and relations from radiology reports. In *Proceedings of the 1st Workshop on NLP for Medical Conversations*, 2021. 3, 4, 5

[19] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019. 3

[20] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, 2014. 3

[21] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. In *Scientific data*, pages 1–10, 2018. 2, 3

[22] Zhihong Lin, Donghao Zhang, Qingyi Tao, Danli Shi, Gholamreza Haffari, Qi Wu, Mingguang He, and Zongyuan Ge. Medical visual question answering: A survey. *Artificial Intelligence in Medicine*, 143:102611, 2023. 2

[23] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 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, 2021. 2, 3

[24] Bo Liu, Li-Ming Zhan, Li Xu, Xiao-Ming Wu, Lin Ma, Yan Yang, and Ya Zhang. Medical visual question answering: A survey. *Artificial Intelligence in Medicine*, 143:102611, 2024. 2

[25] Bo Liu, Ke Zou, Liming Zhan, Zexin Lu, Xiaoyu Dong, Yidi Chen, Chengqiang Xie, Jiannong Cao, Xiao-Ming Wu, and Huazhu Fu. Gemex: A large-scale, groundable, and explainable medical vqa benchmark for chest x-ray diagnosis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. 2, 3

[26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[27] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2024. 3

[28] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*, 2023. 3

[29] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment, 2023. 5

[30] Zengjian Liu, Buzhou Tang, Xiaolong Wang, and Qingcai Chen. De-identification of clinical notes via recurrent neural network and conditional random field. *Journal of Biomedical Informatics*, 58:S34–S42, 2015. 4

[31] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, 2014. 5

[32] Wimukthi Nimalsiri, Mahela Hennayake, Kasun Rathnayake, Thanuja D Ambegoda, and Dulani Meedeniya. Cxlseg dataset: Chest x-ray with lung segmentation. In *2023 International Conference On Cyber Management And Engineering (CyMaEn)*, pages 327–331. IEEE, 2023. 3

[33] 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. 3, 5

[34] Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J Topol. Ai in health and medicine. *Nature Medicine*, 28(1): 31–38, 2022. 2, 4

[35] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference*

on Natural Language Processing (EMNLP-IJCNLP), pages 3982–3992, 2019. 5

[36] 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. 2, 3, 5

[37] 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*, 2024. 2

[38] 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, Lu Yang, Kejia Chen, Per Bjornsson, Shashir Reddy, Ryan Brush, Kenneth Philbrick, Mercy Asiedu, Ines Mezerreg, Howard Hu, Howard Yang, Richa Tiwari, Sunny Jansen, Preeti Singh, Yun Liu, Shekoofeh Azizi, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Riviere, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Elena Buchatskaya, Jean-Baptiste Alayrac, Dmitry Lepikhin, Vlad Feinberg, Sebastian Borgeaud, Alek Andreev, Cassidy Hardin, Robert Dadashi, Léonard Hussenot, Armand Joulin, Olivier Bachem, Yossi Matias, Katherine Chou, Avinatan Hassidim, Kavi Goel, Clement Farabet, Joelle Barral, Tris Warkentin, Jonathon Shlens, David Fleet, Victor Cotruta, Omar Sanseviero, Gus Martins, Phoebe Kirk, Anand Rao, Shravya Shetty, David F. Steiner, Can Kirmizibayrak, Rory Pilgrim, Daniel Golden, and Lin Yang. Medgemma technical report, 2025. 4, 5

[39] 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. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1500–1519, 2020. 3, 4, 5

[40] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature Medicine*, 29(8):1930–1940, 2023. 3

[41] Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1): 44–56, 2019. 2, 4

[42] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 6

[43] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. 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, 2017. 3

[44] 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. 2, 3, 5

[45] xAI. Grok-2: Technical report. *xAI Technical Reports*, 2024. 4

[46] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer, 2016. 3

[47] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019. 3, 5

[48] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Jiayu Lei, Ya Zhang, Yanfeng Wang, and Weidi Xie. Radgenome-chest ct: A grounded vision-language dataset for chest ct analysis. *arXiv preprint arXiv:2404.16754*, 2024. 2