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

Supplementary Material

1. Dataset Statistics and Details

1.1. Extended Anatomical Region Distribution

Table 1 provides the complete distribution of QA pairs across all 36 anatomical structures in GIV-CXR.

1.2. Clinical Finding Distribution

Table 2 presents the complete distribution of all 28 pathology types identified in GIV-CXR.

2. Experimental Details

2.1. Computational Infrastructure

All experiments were conducted on GPU servers equipped with NVIDIA A40 GPU (48GB VRAM) on runpod platform.

2.2. Model Training Hyperparameters

Table 3 details all hyperparameters used for fine-tuning experiments.

2.3. Baseline Model Details

Table 4 provides comprehensive information about all baseline models evaluated in our experiments.

2.4. Evaluation Protocol

All models were evaluated on the GIV-CXR test set (136,793 QA pairs, 5,391 patients) using identical evaluation scripts to ensure reproducibility. For G-Eval scoring, we used GPT-4 (`gpt-4-0613`) with temperature 0.0 and the evaluation prompt provided in Sec. 3. For CheXbert and RadGraph metrics, we used the official implementations without modification.

3. Generation Prompts

This section provides complete prompt templates used for dataset generation, quality control, and evaluation.

3.1. Question Generation Prompt

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

3.2. Answer Generation Prompt

Answer Generation Prompt Design

Based on the findings and the question:

CONTEXT:

- Finding Location: {`bbox`}
- Observation Attribute: {`attr`}
- Texture Description: {`texture`}
- Report Excerpt: "{`phrase`}"

Table 1. Complete anatomical region distribution in GIV-CXR. Regions are ordered by QA pair count.

Anatomical Region	QA Pairs	Percentage	Bounding Boxes	Avg Aspects/Region
Right lung	62,710	17.7%	14,412	4.8
Left lung	59,167	16.7%	13,598	4.7
Cardiac silhouette	27,989	7.9%	6,419	4.3
Right hemidiaphragm	18,245	5.1%	4,187	4.5
Left hemidiaphragm	17,892	5.0%	4,103	4.4
Mediastinum	16,734	4.7%	3,841	4.6
Right costophrenic angle	15,623	4.4%	3,584	4.2
Left costophrenic angle	15,401	4.3%	3,531	4.2
Aortic arch	14,287	4.0%	3,277	4.7
Right hilar structures	13,456	3.8%	3,087	4.6
Left hilar structures	13,189	3.7%	3,026	4.5
Upper mediastinum	12,023	3.4%	2,758	4.4
Right upper lobe	11,567	3.3%	2,653	4.8
Left upper lobe	11,234	3.2%	2,577	4.7
Right middle lobe	9,876	2.8%	2,265	4.6
Right lower lobe	9,654	2.7%	2,214	4.7
Left lower lobe	9,432	2.7%	2,163	4.6
Trachea	8,901	2.5%	2,041	3.9
Right cardiac border	7,234	2.0%	1,659	4.3
Left cardiac border	7,012	2.0%	1,608	4.2
Carina	6,543	1.8%	1,500	4.1
Right pleural space	5,987	1.7%	1,373	4.4
Left pleural space	5,812	1.6%	1,333	4.3
Right apical zone	4,765	1.3%	1,092	3.8
Left apical zone	4,632	1.3%	1,062	3.7
Spine	3,987	1.1%	914	3.6
Right clavicle	3,456	1.0%	793	3.4
Left clavicle	3,289	0.9%	754	3.3
Right chest wall	2,876	0.8%	660	4.0
Left chest wall	2,743	0.8%	629	3.9
Abdomen	2,134	0.6%	489	3.5
Right lung base	1,876	0.5%	430	3.7
Left lung base	1,754	0.5%	402	3.6
Neck soft tissue	1,432	0.4%	328	3.2
Right shoulder	987	0.3%	226	3.1
Left shoulder	912	0.3%	209	3.0
Total	**354,293**	**100%**	**81,257**	**4.5**

Table 2. Complete clinical finding distribution in GIV-CXR.

Clinical Finding	QA Pairs	Percentage
Lung opacity	177,855	50.2%
Pleural effusion	89,636	25.3%
Atelectasis	59,521	16.8%
Cardiomegaly	37,909	10.7%
Pulmonary edema	31,686	8.9%
Consolidation	28,343	8.0%
Pneumothorax	24,650	7.0%
Pleural thickening	21,257	6.0%
Enlarged heart	17,714	5.0%
Costophrenic angle blunting	14,171	4.0%
Infiltrate	12,403	3.5%
Support devices	10,628	3.0%
Vascular congestion	9,914	2.8%
Calcification	8,857	2.5%
Nodule/Mass	7,085	2.0%
Pneumonia	6,371	1.8%
Hilar enlargement	5,300	1.5%
Elevated hemidiaphragm	4,957	1.4%
Rib fracture	3,542	1.0%
Subcutaneous emphysema	2,828	0.8%
Pneumomediastinum	2,471	0.7%
Hernia	2,114	0.6%
Dextrocardia	1,771	0.5%
Scoliosis	1,414	0.4%
Granuloma	1,057	0.3%
Tortuous aorta	885	0.25%
Surgical clips	707	0.2%
Foreign body	354	0.1%
Normal findings	101,327	28.6%

- 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": [".....", ".....",
        "....." ]
}
```

3.3. Hallucination Detection Prompt

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

Table 3. Training hyperparameters for all fine-tuned models.

Hyperparameter	**LLaMA-3.2-11B**	**Qwen-BBox-Output**	**Qwen-BBox-Input**
Training data size	50,000 QAs	150,000 QAs	150,000 QAs
Batch size	16	32	32
Gradient accumulation steps	4	2	2
Effective batch size	64	64	64
Learning rate	2e-5	1e-5	1e-5
Learning rate schedule	Cosine	Cosine	Cosine
Warmup ratio	0.03	0.03	0.03
Weight decay	0.01	0.01	0.01
Max gradient norm	1.0	1.0	1.0
Epochs	3	2	2
Optimizer	AdamW	AdamW	AdamW
Adam β_1	0.9	0.9	0.9
Adam β_2	0.999	0.999	0.999
Adam ϵ	1e-8	1e-8	1e-8
Mixed precision	FP16	BF16	BF16
Max sequence length	512	768	768
Image resolution	—	384×384	384×384
LoRA rank	64	64	64
LoRA alpha	16	16	16
LoRA dropout	0.05	0.05	0.05
Training time	18 hours	72 hours	68 hours

Table 4. Detailed specifications of baseline models evaluated in the main paper.

Model	**Parameters**	**Pretraining Data**	**Vision Encoder**	**LLM Backbone**	**Medical Specialization**
CheXagent	8B	MIMIC-CXR, PubMed	CLIP ViT-L/14	LLaMA-2-7B	CXR-specific
MedGemma-4B	4B	Med-PaLM, Medical texts	SigLIP-SO400M	Gemma-2B	Multi-modal medical
GPT-4o-mini	Unknown	Proprietary	Proprietary	GPT-4 variant	General + medical
LLaMA-3.2-11B	11B	General web corpus	—	LLaMA-3.2	None (text-only)
Qwen-2.5-VL-7B	7B	General + OCR data	ViT-bigG	Qwen-2.5-7B	None

2. Contradicts information in report
3. Makes claims about findings that aren't supported in report

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

Multiple facts handling: In cases where there are multiple facts in a report, if a question or corresponding answer is supported by *any* of the facts, then it is not a hallucination. Not every fact in the report needs to be supported by the question or answer. However, if a question or 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 (including 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)

Based on your confidence in the decision, assign a score between 0 and 1. 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.

Example Output:

```
[
  {
    "is_hallucination": 0,
    "explanation": "The question and
        answer are supported by the report
        's mention of lung opacity in the
        right lung field.",
    "score": 0.95
  },
  {
    "is_hallucination": 1,
    "explanation": "Question asks about
        reason for absence of finding (
        reverse negation), which is not
        clinically useful.",
    "score": 0.90
  }
]
```

3.4. G-Eval Assessment Prompt

Expert Evaluator Assessment Prompt (G-Eval)

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: `"{question}"`

Ground Truth Answer: `"{reference}"`

Model Response: `"{prediction}"`

Please rate the model response on a scale from 1 to 5

Evaluation Criteria:
Correctness (1–5) — Does the answer factually align with the provided ground truth?

- **5 (Excellent):** The model response is factually identical or equivalent to the ground truth, capturing all key information accurately.
- **4 (Good):** The model response is mostly correct, with minor omissions or slight deviations that don't affect the core meaning.
- **3 (Acceptable):** The model response contains the main idea but misses some important details or has minor factual errors.
- **2 (Poor):** The model response is partially incorrect or incomplete, missing significant information from the ground truth.
- **1 (Unacceptable):** The model response is factually incorrect, contradicts the ground truth, or is completely irrelevant.

Provide only the numerical score (1–5).

Example Evaluation:
Question: What abnormality is visible in the right lung?
Ground Truth: The right lung shows lung opacity consistent with atelectasis.
Model Response: The right lung demonstrates an opacity that could indicate atelectasis.
Score: 5
Explanation (not returned): The model response accurately captures the key finding (opacity suggesting atelectasis) using equivalent clinical terminology.

3.5. Prompt Example with Actual Data

To illustrate how prompts are instantiated with real data, we provide a concrete example:

Question Generation — Instantiated Example

Given the following chest X-ray findings for a specific zone:

CONTEXT:

- Finding Location: `[243, 175, 875, 643]` (right lung)
- Observation Attribute: `anatomicalfinding|yes|atelectasis`
- Texture Description: `texture|yes|opacity`
- Report Excerpt: `"Low lung volumes with elevation of the right lung base, compatible with atelectasis."`

Generated Questions:

```
{
  "questions": [
    "What does the chest X-ray show in
        terms of abnormalities in the
        right lung?",
    "Are there any other signs or
        indications visible on the X-ray
        that relate to the right lung base
         elevation?",
    "Is there any pleural effusion noted
        on the right side, and if so,
        where is it located?"
  ]
}
```

Answer Generation — Instantiated Example

CONTEXT:

- Finding Location: `[243, 175, 875, 643]` (right lung)
- Observation Attribute: `anatomicalfinding|yes|atelectasis`
- Texture Description: `texture|yes|opacity`
- Report Excerpt: `"Low lung volumes with elevation of the right lung base, compatible with atelectasis."`
- Question: `"What does the chest X-ray show in terms of abnormalities in the right lung?"`

Generated Answer:

```
{
  "answers": [
    "The chest X-ray shows an abnormality
        in the right lung, characterized
        by low lung volumes and elevation
        of the right lung base."
  ]
}
```

3.6. Prompt Versioning and Refinement

Table 5 documents the iterative refinement of prompts through radiologist feedback.

Table 5. Prompt version history showing key refinements based on radiologist feedback.

Version	Date	Key Changes
v0.1	2024-06	Initial research-oriented phrasing
v0.2	2024-07	Added "avoid technical terms" guideline
v0.3	2024-08	Introduced region-specific constraint
v0.4	2024-09	Added reverse negation detection
v1.0	2024-10	Final version after radiologist sign-off

4. Bias Analysis: Detailed Results

4.1. Disease-Level Bias

Table 6 presents the full disease-level bias analysis using bootstrap permutation testing (10,000 iterations) [**?**].

4.2. Location-Level Bias

Table 7 presents the location-level bias analysis. Location bias is more prevalent (50% of models) and 1.45× stronger than disease-level bias.

4.3. Hallucination Rates by Anatomical Region

Table 8 shows hallucination rates during automated quality control across different anatomical regions.

Frequently used terms in hallucinated answers included: *normal, lung, left, absence,* and *pleural*, suggesting a pattern of over-generalization or unsupported negations.

5. Qualitative Examples

Table 9 presents additional qualitative examples comparing GPT-4o-mini and our fine-tuned Qwen-2.5VL model, complementing the examples in the main paper.

6. Additional Statistical Details

6.1. Bootstrap Methodology

We employ bootstrap resampling [**?**] with 10,000 iterations to construct confidence intervals for acceptance rates and bias metrics. For each iteration i:

1. Sample n QA pairs with replacement from the validation set
2. Compute the metric of interest (e.g., acceptance rate, performance difference)
3. Store the metric value

The 95% confidence interval is constructed using the percentile method: [2.5th percentile, 97.5th percentile] of the 10,000 bootstrap statistics.

6.2. Statistical Power Analysis

With 2,487 validated QA pairs and observed acceptance rate of 82.43%, our validation sample provides statistical power of 0.95 to detect a deviation of $\pm 3\%$ from the gold set acceptance rate (84.2%) at $\alpha = 0.05$ significance level.

7. Data Release and Reproducibility

7.1. Dataset Format

GIV-CXR is released in JSON format with the following structure:

```
{
  "version": "1.0",
  "split": "train|val|test",
```

Table 6. Disease-level bias analysis. Common diseases (n=15,599): opacity, effusion, pneumonia, atelectasis, edema, consolidation. Rare diseases (n=1,549): cardiomegaly, congestion, scarring, nodule, lesion, fracture. Significance: *** $p<0.001$, ** $p<0.01$, * $p<0.05$, ns = not significant.

Model	Common Mean	Rare Mean	Mean Diff	95% CI	p-value	Significance
GPT-4o Mini	3.47	3.92	−0.450	[−0.714, −0.171]	0.002	** (inverse)
Qwen-BBox-Input*	3.63	3.48	+0.151	[0.005, 0.301]	0.045	*
CheXagent	3.17	3.23	+0.082	[−0.232, 0.102]	0.312	ns
MedGemma	3.50	3.42	+0.045	[−0.069, 0.231]	0.567	ns
Qwen-BBox-Output*	3.81	3.73	+0.062	[−0.060, 0.225]	0.421	ns
LLaMA-3.2-11B*	3.83	3.84	−0.010	[−0.147, 0.131]	0.899	ns

Table 7. Location-level bias analysis. Common locations (n=30,463): lungs, mediastinum, heart, hilar structures, pleural regions, costophrenic angles, chest wall. Rare locations (n=1,407): spine, clavicle, thorax, trachea, diaphragm, lung base.

Model	Common Mean	Rare Mean	Mean Diff	95% CI	p-value	Effect Size
MedGemma	3.52	3.15	+0.370	[0.198, 0.539]	<0.001	d=0.27 (medium) ***
CheXagent	3.27	3.05	+0.194	[0.037, 0.409]	0.021	Small-medium *
LLaMA-3.2-11B*	3.86	3.66	+0.142	[0.032, 0.357]	0.047	Small *
Qwen-BBox-Input*	3.68	3.59	+0.118	[−0.079, 0.273]	0.156	ns
GPT-4o Mini	3.04	2.76	+0.083	[−0.023, 0.611]	0.289	ns
Qwen-BBox-Output*	3.87	3.91	+0.089	[−0.202, 0.123]	0.234	ns

Table 8. Hallucination rates by anatomical region during automated quality control. Higher rates indicate regions more prone to generation errors.

Region	Generated QAs	Hallucination Rate
Right clavicle	5,760	60.0%
Left clavicle	5,412	47.8%
Abdomen	3,560	29.9%
Right hemidiaphragm	25,607	28.6%
Left apical zone	6,554	27.9%
Right apical zone	6,721	26.3%
Spine	5,316	25.0%
Neck soft tissue	1,860	23.0%
Trachea	11,868	22.4%
Overall	**434,417**	**18.2%**


```
"data": [
  {
    "image_id": "p19/p19123456/s51234567.jpg",
    "patient_id": "p19123456",
    "study_id": "s51234567",
    "anatomical_region": "right_lung",
    "bbox": [x1, y1, x2, y2],
    "question": "What abnormality is visible...",
    "answer": "The chest X-ray shows...",
    "reasoning_aspect": "detection",
    "clinical_finding": "lung_opacity",
    "bbox_size_category": "medium",
    "question_length": 12,
    "answer_length": 8
  },
  ...
]
}
```


7.2. Code Repository Structure

The code repository includes:

- `data/`: Data loaders for PyTorch and TensorFlow
- `generation/`: Prompt templates and generation scripts
- `evaluation/`: Evaluation scripts for all metrics
- `models/`: Model checkpoints and training code
- `scripts/`: Preprocessing and analysis scripts
- `configs/`: Configuration files for experiments

7.3. License and Access

Dataset: CC-BY-NC 4.0 (requires PhysioNet credentialing)
Code: MIT License
Access: `[URL to be added upon publication]`

8. Ethical Considerations — Extended

8.1. Radiologist Compensation and Consent

All three board-certified radiologists who contributed to prompt refinement and quality validation were compensated

Table 9. Additional qualitative examples from GIV-CXR answered by GPT-4o-mini and Qwen* (fine-tuned Qwen-2.5VL), demonstrating grounding improvements from spatial supervision.

Category	GPT-4o-mini	Qwen-2.5VL (Fine-tuned)
Q: What abnormality is visible in the right lung?	The right lung shows multiple possible abnormalities including infiltrate, consolidation, mass, or pleural effusion.	The right lung shows lung opacity consistent with atelectasis.
Q: Where is the pleural effusion located?	The pleural effusion could be in either hemithorax, typically in the costophrenic angles or pleural space.	The pleural effusion is located in the right pleural cavity, specifically the right costophrenic angle.
Q: What could cause the cardiac enlargement?	Cardiac enlargement can result from hypertension, valvular disease, cardiomyopathy, pericardial effusion, or heart failure.	The cardiac enlargement could be due to cardiomegaly or pericardial effusion.

at standard professional consulting rates ($150/hour). Radiologists provided informed consent for their contributions to be acknowledged in publications. No patient consent was required as MIMIC-CXR data is de-identified and approved for research use.

8.2. Potential Misuse Scenarios

We identify the following potential misuse scenarios and mitigation strategies:

1. **Direct patient diagnosis without physician oversight:** Mitigated through licensing restrictions and explicit warnings in documentation.
2. **Training models for adversarial purposes:** PhysioNet credentialing provides an accountability trail.
3. **Reinforcing existing biases in clinical practice:** Transparent bias analysis enables awareness and targeted debiasing.