

MMRad-22K: A Structured Multimodal Evidence Dataset for Chest X-ray Report Generation

Yichen Zhao^{1*}, Zelin Peng^{1*}, Fenghe Tang^{2,3}, Piao Yang⁴, Yu Huang¹, Wei Shen^{1†}

¹ MoE Key Lab of Artificial Intelligence, AI Institute,

School of Computer Science, Shanghai Jiao Tong University, Shanghai, China

² School of Biomedical Engineering, Division of Life Sciences and Medicine,

University of Science and Technology of China (USTC), Hefei, Anhui 230026, China

³ Center for Medical Imaging, Robotics, Analytic Computing & Learning (MIRACLE),
Suzhou Institute for Advanced Research, USTC, Suzhou, Jiangsu 215123, China

⁴ Department of Radiology, The First Affiliated Hospital,

Zhejiang University School of Medicine, Hangzhou, Zhejiang, China

{zyc_sjtu, zelin.peng, wei.shen}@sjtu.edu.cn

Abstract

Chest X-ray (CXR) reporting follows a region-based clinical workflow in which radiologists inspect anatomical regions and integrate localized findings into a final report. However, existing resources for CXR report generation provide these supervision signals in fragmented forms. We introduce MMRad-22K, a dataset that organizes regional textual observations, anatomical grounding coordinates, localized image evidence, and report targets into structured multimodal evidence units for CXR report generation. To motivate this formulation, we first compare different evidence formats for report generation and find that structured multimodal evidence is generally more useful than text-only or bounding box-based evidence. We then adapt a unified LVLM backbone using MMRad-22K and show that adaptation with multimodal evidence outperforms both textual-evidence adaptation and end-to-end adaptation on language and clinically oriented metrics. Under the same evaluation protocol, the adapted model also reaches a performance level comparable to several open-source LVLM references. Together, these results support MMRad-22K as a practical structured multimodal resource for training and evaluating CXR report generation aligned with clinical reading workflows. The code is available at <https://github.com/qiuzyc/MMRad/>.

1 Introduction

Medical image analysis is fundamental to computer-assisted diagnosis, enabling clinically relevant information to be extracted from medical images. Chest X-ray (CXR) report generation aims to automatically produce clinically coherent radiology reports from medical images and has become

a central task in medical vision-language modeling (Jing et al., 2018; Li et al., 2018; Miura et al., 2021; Tang et al., 2025; Peng et al., 2026; Li et al., 2025). Existing approaches have evolved from encoder-decoder frameworks such as R2Gen (Chen et al., 2020) to recent large vision-language models (LVLMs) (Wang et al., 2023; Liu et al., 2025c). Despite substantial progress, most current paradigms still formulate report generation as a global image-to-text mapping problem, where a full CXR image is directly transformed into a report through autoregressive decoding (Fig. 1, panel B(i)). Such supervision provides limited guidance on how localized findings are linked to the final report.

This limitation is especially important in chest X-ray interpretation, where reports are formed by integrating findings from multiple regions rather than from a single global summary. In clinical reading, radiologists inspect localized visual patterns, assess their clinical significance, and then synthesize these observations into the final report (North, 2024). As illustrated in Fig. 1, panel A, regional observations are naturally associated with localized visual evidence during this process (*orange box*). Recent medical LVLMs have begun to explore richer forms of intermediate evidence. For example, Med-R2 (Wang et al., 2026a) and MRG-R1 (Wang et al., 2026b) introduce textual observations to improve report generation quality and factual consistency (Fig. 1, panel B(ii)), while models such as CheXagent (Chen et al., 2024), MAIRA-2 (Baner et al., 2024), and RadVLM (Deperrois et al., 2025) incorporate anatomical grounding, regional cues, or broader medical assistance capabilities (Fig. 1, panel B(iii)). However, textual observations provide limited visual grounding, and bounding-box coordinates indicate location without preserving localized image details. At the dataset level,

*Equal contribution.

†Corresponding author.

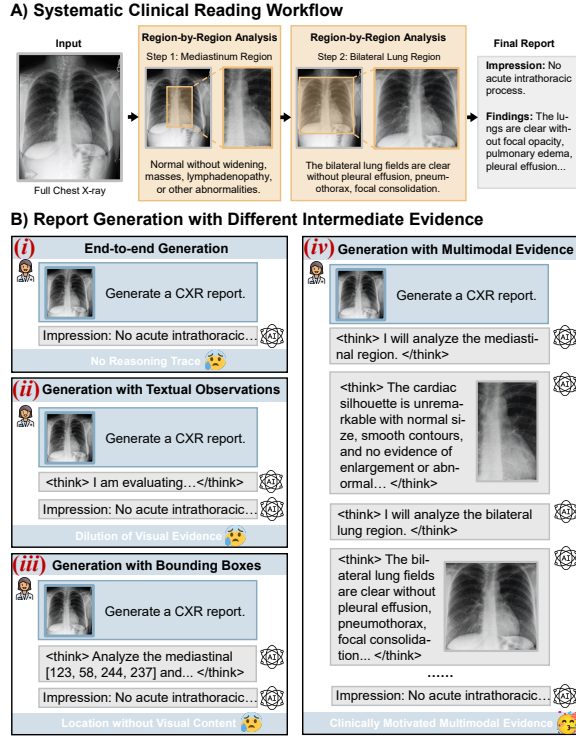


Figure 1: **Illustration of clinical reading and intermediate evidence for CXR report generation.** (A) Radiologists inspect a full chest X-ray through region-level analysis and synthesize these observations into a final report. (B) Different report generation settings use different forms of intermediate evidence: (i) end-to-end generation, (ii) textual observations, (iii) anatomical bounding boxes, and (iv) multimodal evidence.

existing radiology resources provide only fragmented supervision signals, such as report-level targets (Johnson et al., 2019), anatomical grounding coordinates (Liu et al., 2025b), or grounded reasoning traces (Liu et al., 2025a), but rarely organize regional textual observations, localized image evidence, grounding annotations, and report targets into a unified multimodal report-generation-oriented structure. As a result, the utility of such structured multimodal evidence for chest X-ray report generation has not been systematically studied.

To better understand this issue, we first conduct a pilot comparison of different localized evidence formats for chest X-ray report generation, including textual observations, bounding boxes, localized images, and paired multimodal evidence. As shown in Fig. 2, across multiple advanced closed-source LVLMs and evaluation metrics, the multimodal setting generally performs the best among the compared formats. This suggests that structured multimodal evidence is a promising supervision format for report generation, because com-

binning regional textual observations with localized images can provide complementary support. However, existing radiology resources do not explicitly organize it in a reusable study-level form. Driven by these findings, we introduce **MMRad-22K**, a study-level chest X-ray report generation dataset organized around anatomy-guided multimodal evidence. Built upon MIMIC-CXR (Johnson et al., 2019) and GEMeX-ThinkVG (Liu et al., 2025a), MMRad-22K reorganizes fragmented supervision signals into report-oriented evidence units within each study, providing a structured multimodal resource for chest X-ray report generation. Finally, we evaluate MMRad-22K through adaptation experiments on a unified LVLM backbone Anole. The results show that using MMRad-22K improves report generation quality over standard end-to-end generation and textual evidence settings under the same framework. In addition, the adapted model achieves comparable performance relative to representative open-source LVLMs, despite not introducing a specialized model architecture.

Our contributions are summarized as follows. (1) We introduce **MMRad-22K**, a study-level multimodal dataset that establishes a structured evidence-to-report interface for CXR report generation, transforming fragmented regional observations, anatomical grounding coordinates, localized images, and report targets into unified multimodal evidence units. (2) We present a task-oriented construction pipeline that reformulates fragmented localized supervision into report-generation-oriented evidence units, with multi-stage verification and clinician evaluation to improve clinical accuracy and consistency. (3) We provide empirical evidence that structured multimodal supervision is useful for CXR report generation: pilot comparisons suggest multimodal evidence is more effective than text-only or bounding box-based alternatives, and controlled adaptation experiments show that MMRad-22K supports stronger report generation than standard end-to-end and textual-evidence settings; under the same evaluation protocol, the adapted model also reaches a performance level comparable to several open-source LVLM references.

2 Related Work

2.1 End-to-End CXR Report Generation

Early methods mainly followed encoder-decoder image-to-text paradigms, including memory-driven frameworks such as R2Gen (Chen et al., 2020).

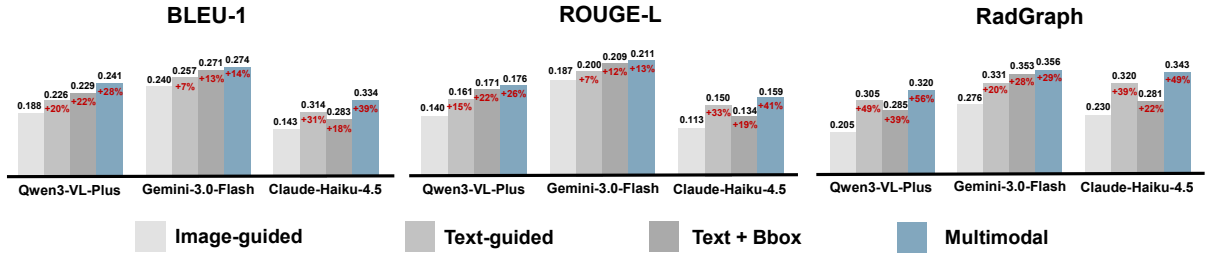


Figure 2: **Performance under different intermediate evidence formats.** Multimodal evidence generally yields the best or comparable performance across multiple LVLMs and evaluation metrics. Percentage improvements are computed relative to the Image-guided baseline within each metric.

More recently, LVLMs have been adapted to this task, as exemplified by SEI (Liu et al., 2024), FedMRG (Metmer and Yang, 2025), R2GenGPT (Wang et al., 2023), and MLRG (Liu et al., 2025c). Despite this progress, end-to-end image-to-report generation remains a standard formulation in chest X-ray report generation.

2.2 Intermediate Evidence for CXR Report Generation

Recent medical LVLMs have begun to incorporate richer supervision signals beyond global image-report pairs. Some approaches improve reasoning and factuality through intermediate textual evidence, rationale-style supervision, or reflection strategies, including LVMed-R2 (Wang et al., 2026a), MRG-R1 (Wang et al., 2026b), and BoxMed-RL (Jing et al., 2026). Other frameworks emphasize grounding and region-aware supervision, such as CheXagent (Chen et al., 2024), MAIRA-2 (Bannur et al., 2024), RadVLM (Deperois et al., 2025), RGRG (Tanida et al., 2023), and ClinCoT (Liu et al., 2026). These studies highlight the value of region-aware reasoning and grounding signals for medical vision-language modeling. However, such supervision is typically embedded in model architectures, training objectives, or multitask systems, rather than explicitly organized as a reusable report-generation-oriented resource that pairs textual observations with localized images.

2.3 Radiology Datasets and Resources

Several public datasets support medical vision-language learning. MIMIC-CXR (Johnson et al., 2019) mainly provides report-level supervision, while other resources introduce complementary signals such as anatomical grounding and reasoning traces (Liu et al., 2025b,a; Wu et al., 2021). Although these datasets provide valuable supervision, their annotations remain fragmented across differ-

ent tasks and resources. MMRad-22K addresses this gap by defining a structured evidence-to-report learning framework, where localized visual evidence, regional clinical observations, and final report generation are organized into a unified interface that enables joint training and separate evaluation of intermediate evidence and final reports.

3 Does Multimodal Evidence Help CXR Report Generation?

Chest X-ray report generation requires models to transform visual observations into clinically coherent textual descriptions. In practice, intermediate supervision may be provided in different forms, such as localized image regions, textual observations, grounding coordinates, or combinations of these signals.

We conduct a pilot oracle-trace analysis of several intermediate evidence formats. For each study, the model always receives the full chest X-ray image, and we additionally provide evidence from the same anatomical region in one of the following formats: (1) **Image-guided**: regional image crop; (2) **Text-guided**: regional textual observation describing that region; (3) **Text + BBox**: the same regional textual observation augmented with anatomical grounding coordinates; and (4) **Multimodal**: both the regional textual observation and corresponding image crop.

Evaluation Setup. We conduct the pilot study on 1,097 CXR studies derived from MIMIC-CXR (Johnson et al., 2019) and GEMeX-ThinkVG (Liu et al., 2025a), where each study contains report targets together with region-level textual and visual evidence used to instantiate different evidence formats. We evaluate three advanced closed-source LVLMs, including Qwen3-VL-Plus (Bai et al., 2025), Gemini-3.0-Flash (Google, 2025), and Claude-Haiku-4.5 (An-

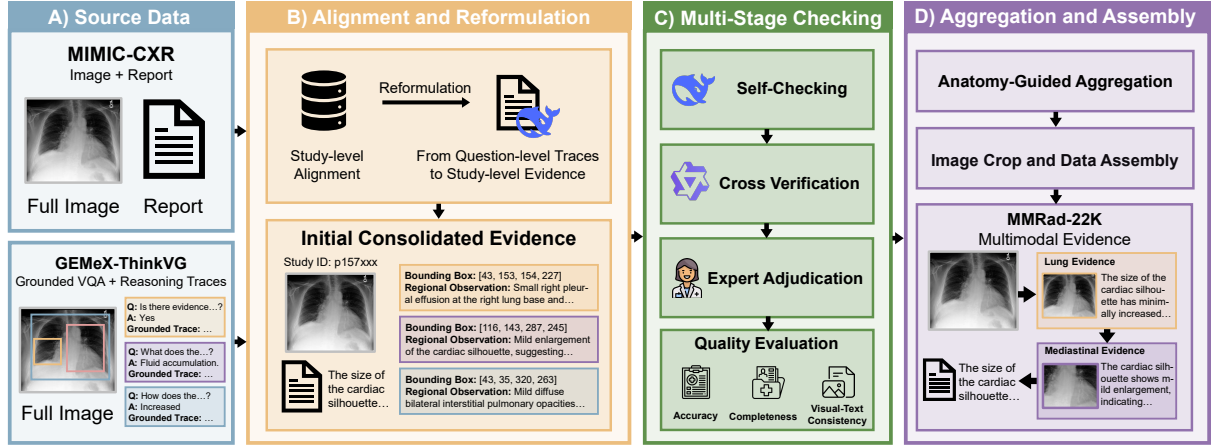


Figure 3: **Construction pipeline of MMRad-22K.** Starting from chest X-ray studies and grounded regional annotations, we reorganize fragmented question-level supervision into study-level report-oriented evidence units. The pipeline includes study-level alignment, evidence reformulation, multi-stage verification, quality assessment, and anatomy-guided aggregation. The final dataset unifies regional textual observations, anatomical coordinates, localized image evidence, and report targets into a structured multimodal resource for chest X-ray report generation.

thropic, 2025), using NLG metrics BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), and the clinical efficacy (CE) metric RadGraph (Delbrouck et al., 2022). Prompt templates and input examples are provided in Appendix A.1.

Multimodal localized evidence yields the most consistent gains. Fig. 2 summarizes the performance of multiple advanced closed-source LVLs under different localized evidence settings. Across models and metrics, the multimodal format generally performs best or remains competitive in our pilot comparison. In contrast, adding anatomical bounding-box coordinates yields smaller and less consistent gains over text-guided formats. These results suggest that localized textual and visual evidence can provide complementary support for CXR report generation.

However, existing radiology datasets provide only fragmented supervision signals, such as reports, grounding annotations, or textual rationales, rather than a unified report-generation-oriented multimodal evidence structure. To bridge this gap, we construct **MMRad-22K**, which organizes textual observations, anatomical grounding coordinates, and localized image evidence within each study as structured multimodal supervision for CXR report generation.

4 MMRad-22K Dataset

4.1 Data Construction

MMRad-22K is designed as a report-generation-oriented multimodal evidence dataset for chest

X-ray reasoning. To construct this supervision, we leverage complementary signals from two study-aligned resources built upon the same radiographs, as shown in Fig. 3, panel A. MIMIC-CXR (Johnson et al., 2019) provides chest X-ray images and report-level supervision, while GEMeX-ThinkVG (Liu et al., 2025a) provides grounded VQA-style reasoning traces and anatomical bounding boxes. Our goal is to transform these fragmented grounded traces into anatomy-structured localized multimodal evidence suitable for study-level report generation.

Study-Level Alignment and Evidence Reformulation. Grounded supervision in GEMeX-ThinkVG (Liu et al., 2025a) is organized as independent question-specific traces, which are not directly suited for study-level report generation. As shown in Fig. 3, panel B, we first align the grounded traces and anatomical bounding boxes from GEMeX-ThinkVG with their corresponding MIMIC-CXR (Johnson et al., 2019) studies, associating localized evidence with the full chest X-ray image and report target. GEMeX-ThinkVG contains 202,384 grounded VQA traces, which are reorganized after alignment into 21,994 study-level samples, with approximately 9 localized evidence traces per study on average. We then use DeepSeek-v3 (DeepSeek-AI et al., 2025) to reformulate these fragmented VQA traces into report-oriented evidence units, converting question-answer supervision into anatomy-associated regional observations for chest X-ray report generation.

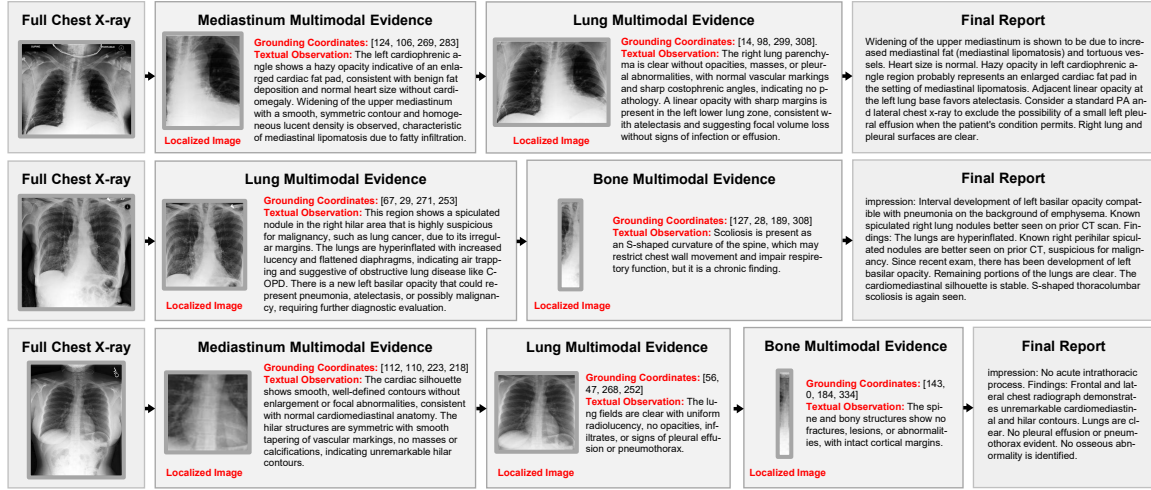


Figure 4: **Examples of MMRad-22K Dataset.** Each sample includes the full CXR image, structured multimodal evidence, and the final report.

Clinical Consistency Preservation and Verification. During the reformulation process, LLMs may introduce summarization errors or hallucinated content. To improve consistency, we adopt a multi-stage verification pipeline, as shown in Fig. 3, panel C. First, DeepSeek-v3 (DeepSeek-AI et al., 2025) performs a self-check by comparing the reformulated evidence traces from the same study, mainly to remove unsupported content and improve local coherence. Second, we perform study-level consistency checking with Qwen2.5-72B (Qwen et al., 2025), using the MIMIC-CXR reference report as a quality-control signal to identify potential contradictions between the reformulated evidence and the documented findings. The reference report is used only for verification rather than for generating new evidence content. Cases that remain uncertain after this step are further reviewed through clinician adjudication. A coarse summary of the verification pipeline is provided in Appendix A.2.1.

To assess the quality of the reformulated evidence units, we randomly sampled 900 cases (4% of the full dataset) for blinded evaluation by two clinicians with radiology experience. Each case was independently rated on three 5-point Likert scales measuring **Clinical Accuracy**, **Completeness**, and **Visual-Text Consistency**. Across the sampled cases, the evidence units achieved average scores of 4.6/5.0, 4.3/5.0, and 4.9/5.0 for the three dimensions, respectively, with 96.7%, 83.3%, and 95.6% of cases receiving scores of at least 4 from both clinicians. Inter-rater agreement measured by Cohen’s κ was 0.64, 0.50, and 0.56, respectively.

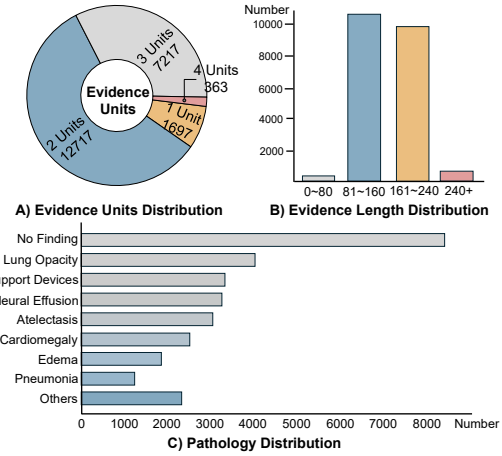


Figure 5: **Statistics of MMRad-22K.** The dataset contains structured regional evidence with diverse clinical findings and varying evidence lengths.

The slightly lower completeness scores mainly reflect that some secondary findings described in the original reports are not explicitly covered by the available evidence units. Given the strong concentration of ratings in the high-score range, the percentage of cases jointly rated at least 4 by both clinicians provides a more direct indication of evidence quality than inter-rater agreement. Overall, the reformulated evidence units show good quality. We further perform stratified analysis across different clinical subgroups. Detailed evaluation rubrics, subgroup analysis, and error analysis are provided in Appendix A.2.2 and A.2.3.

Anatomy-Guided Evidence Organization and Dataset Assembly. The reformulated evidence units are often fine-grained and heterogeneous in structure. To obtain a more consistent and learn-

Dataset	Report Target	Textual Evidence	Anatomical Grounding	Multimodal Evidence
EHRXQA	✗	✗	✗	✗
MIMIC-CXR	✓	✗	✗	✗
GEMeX	✗	✓	✓	✗
Chest ImaGenome	✗	✓	✓	✗
GEMeX-ThinkVG	✗	✓	✓	✗
MMRad-22K	✓	✓	✓	✓

Table 1: **Comparison with related chest X-ray datasets in terms of report targets and intermediate evidence.** MMRad-22K uniquely combines regional textual observations, anatomical grounding coordinates, structured multimodal evidence, and report-level targets within each study.

able study-level representation, we organize the evidence into four coarse anatomy-guided groups: mediastinal, lung, bone, and other regions. Although this grouping is intentionally coarse, it provides a practical abstraction for model learning while remaining broadly consistent with classical chest X-ray interpretation frameworks in radiology textbooks and prior technical work (Collins and Stern, 2012; Wu et al., 2021). Importantly, within each anatomical group, the original regional observations and their associated bounding boxes are retained and organized together rather than re-summarized. Based on the grouped bounding boxes, we further extract localized image crops from the original MIMIC-CXR image and pair them with the corresponding textual observations and coordinates to form multimodal evidence units for each region. Examples of MMRad-22K are shown in Fig. 4, with additional examples provided in Appendix A.3.1.

4.2 Data Statistics

As shown in Fig. 5, MMRad-22K contains 21,994 study-level multimodal evidence samples from 7,975 patients. Most studies contain 2–3 regional evidence groups (Fig. 5, panel A). The evidence text is also relatively rich, with most samples ranging from 81 to 240 words in length (Fig. 5, panel B). The dataset covers a broad range of findings based on CheXpert (Irvin et al., 2019) labels (Fig. 5, panel C), including lung opacity, pleural effusion, atelectasis, cardiomegaly, edema, and pneumonia, in addition to normal studies (No Finding). Additional anatomy-group statistics are provided in the Appendix A.3.

4.2.1 Data Comparison

As shown in Table 1, existing radiology datasets provide only partial supervision for chest X-ray report generation. MIMIC-CXR (Johnson et al., 2019) and EHRXQA (Bae et al., 2023) provide report-level targets and multimodal QA pairs, respectively, but do not include explicit intermediate evidence. Chest ImaGenome (Wu et al., 2021) provides bounding boxes and localized labels, but does not organize them into coherent study-level multimodal evidence for report composition. GEMeX (Liu et al., 2025b) and GEMeX-ThinkVG (Liu et al., 2025a) provide grounded question-level textual supervision, yet their annotations remain organized as VQA-oriented explanations or reasoning traces rather than report-generation-oriented regional evidence.

In contrast, MMRad-22K is not a simple concatenation of existing datasets, but a task-oriented restructuring of fragmented supervision into a unified multimodal resource for report generation. Specifically, it converts question-level grounded traces into study-level, structured evidence units that are aligned with the original report target. Each sample is organized around a single study and explicitly links regional textual observations, anatomical coordinates, localized images, and the final report within the same report-generation context, which is not directly supported by prior datasets.

5 MMRad-22K for CXR Report Generation

Section 3 suggests that localized multimodal evidence benefits CXR report generation. We next evaluate whether MMRad-22K effectively supports LVLM adaptation and provides advantages beyond standard report supervision.

5.1 Training Settings

Dataset Split. We randomly split MMRad-22K at the patient level. Following this protocol, 237 patients are reserved for testing, while the remaining 7738 patients are used for training.

Baseline Model. We adopt Anole-7B (Chern et al., 2024) as the base unified LVLM because of its native multimodal autoregressive generation capability. As a general-domain model without specialized radiology report-generation pretraining, Anole provides a relatively controlled testbed for evaluating the contribution of MMRad-22K supervision, as shown in Fig. 6. Unlike conventional LVLMs de-

Setting	NLG Metrics					CE Metrics				
	B-1↑	B-2↑	B-3↑	MTR↑	R-L↑	RadCliQ↓	RG _e ↑	RG _{er} ↑	RG _{ber} ↑	RaTE↑
Zero-shot	0.096	0.034	0.007	0.104	0.095	2.102	0.082	0.070	0.041	0.464
End-to-end	0.126	0.040	0.008	0.107	0.103	1.821	0.112	0.099	0.067	0.462
Textual	0.175	0.075	0.032	0.144	0.150	1.541	0.149	0.133	0.092	0.493
Grounded	0.192	0.091	0.046	0.166	0.173	1.409	0.180	0.161	0.112	0.528
Generated	0.193	0.090	0.043	0.167	0.172	1.391	0.187	0.173	0.123	0.529

Table 2: **Chest X-ray report generation results under different evidence settings on the Anole backbone.** Statistical tests confirm that the two multimodal evidence settings, *Generated* and *Grounded*, do not differ significantly from each other, while both significantly outperform weaker supervision settings. **Bold** denotes the best result. Abbreviations: B-n: BLEU-n; MTR: METEOR; R-L: ROUGE-L; RG: RadGraph.

signed primarily for visual understanding, Anole enables unified generation over both textual and visual tokens within a single autoregressive framework, making it suitable for modeling the multimodal evidence structures in MMRad-22K and the final radiology report within one generation process. We fine-tune the model using LoRA (Hu et al., 2021) applied to the query, key, and value projections of the attention modules with rank $r = 16$ and scaling factor $\alpha = 32$, leveraging the training framework established in (Chern et al., 2025). Training is conducted for 50,000 steps on two NVIDIA H800 GPUs using a learning rate of 1×10^{-5} and batch size 2.

Evaluation Metrics. We evaluate report generation quality using standard NLG metrics, including BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and ROUGE-L (Lin, 2004), together with clinical efficacy (CE) metrics including RadGraph (Delbrouck et al., 2022), RadCliQ (Yu et al., 2023), and RaTEScore (Zhao et al., 2024). Detailed descriptions of the evaluation metrics are provided in Appendix A.4.

5.2 Controlled Comparison of Supervision Settings on the Anole Backbone

Starting from the original zero-shot Anole baseline, we compare five settings while keeping the report generation task unchanged, as shown in Table 2. **Zero-shot** directly generates a report without task-specific adaptation. **End-to-end** adapts the model to generate the report directly from the full chest X-ray image. **Textual** first generates regional textual evidence and then produces the final report. **Generated** further generates localized visual evidence in Anole’s native autoregressive generation space together with regional textual evidence before report generation. **Grounded** first generates regional

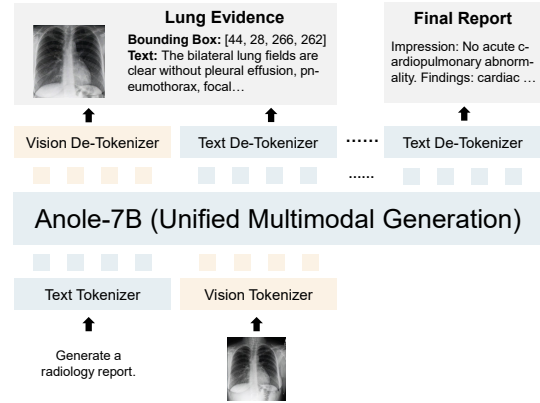


Figure 6: **Unified multimodal framework based on MMRad-22K.** Each training sample is organized as a unified multimodal sequence.

textual evidence and predicts anatomical regions, then maps the predicted regions back to the source chest X-ray and extracts localized image crops as grounded visual evidence before report generation. Notably, the intermediate evidence is generated or predicted by the model itself rather than provided as external prompts. Additional details on evidence construction and the difference between Generated and Grounded are provided in Appendix A.5.

The *Zero-shot* Anole baseline performs poorly across evaluation metrics, indicating that the original general-domain unified model is not directly suited to chest X-ray report generation. Adapting the model in the *End-to-end* setting already yields clear improvements, showing that task-specific adaptation is important in this domain.

Modeling regional textual evidence in the *Textual* setting further improves performance over the end-to-end baseline, suggesting that intermediate textual evidence provides useful structured support beyond direct report generation from the full image alone. Extending this intermediate evidence to

Models	Param.	Domain	NLG Metrics				CE Metrics			
			B-1↑	B-2↑	MTR↑	R-L↑	RadCliQ↓	RG _e ↑	RG _{er} ↑	RaTE↑
<i>Large Vision-Language Models</i>										
Deepseek-VL2	4.5B	G	0.105	0.033	0.090	0.103	1.888	0.084	0.075	0.495
Qwen3-VL	8B	G	0.167	0.054	0.125	0.133	1.357	0.158	0.144	0.530
InternVL3.5	8B	G	0.179	0.059	0.135	0.147	1.316	0.152	0.145	0.532
LLaVA-Med-1.5	7B	M	0.127	0.011	0.076	0.100	1.755	0.072	0.063	0.426
<i>Unified Large Language Models</i>										
Bagel	7B	G	0.111	0.040	0.131	0.113	1.602	0.157	0.135	0.525
UniMedVL	14B	M	0.157	0.054	0.147	0.128	1.504	0.157	0.145	0.519
Anole	7B	G	0.096	0.034	0.104	0.095	2.102	0.082	0.070	0.464
Anole-MMRad	7B	M	0.193	0.090	0.167	0.172	1.391	0.187	0.173	0.529

Table 3: **Contextual comparison of MMRad-22K.** All models are evaluated under the same chest X-ray report generation task, test split, and evaluation protocol, while Anole-MMRad is adapted using MMRad-22K. Anole-MMRad reaches a performance level on par with representative open-source LVLM references, further supporting the effectiveness of MMRad-22K for adapting LVLMs to chest X-ray report generation. **Bold** denotes the best result. Abbreviations: Param.: activated parameters; G: general; M: medical; B-n: BLEU-n; MTR: METEOR; R-L: ROUGE-L; RG: RadGraph.

multimodal evidence yields additional gains. Both multimodal settings *Generated* and *Grounded* outperform the text-only setting, further supporting the value of combining textual evidence with localized visual evidence, indicating that multimodal evidence derived from MMRad-22K provides complementary support for report generation.

Overall, these results suggest that MMRad-22K provides useful structured supervision for chest X-ray report generation. Relative to standard end-to-end adaptation, textual evidence improves report generation quality, and localized visual evidence brings further complementary gains.

5.3 MMRad-22K Enables Effective Adaptation to CXR Report Generation

To provide external context for the effectiveness of MMRad-22K, we additionally report the performance of the adapted Anole model (Anole-MMRad) together with several representative open-source LVLM references, including general-domain models DeepSeek-VL2 (Wu et al., 2024), Qwen3-VL (Bai et al., 2025), InternVL3.5 (Wang et al., 2025), Anole (Chern et al., 2024), and Bagel (Deng et al., 2025), as well as medical-domain models LLaVA-Med-1.5 (Li et al., 2023) and UniMedVL (Ning et al., 2026). Some of these reference models have prior exposure to medical report-generation-related training during their original development.

Table 3 shows that adapting Anole with MMRad-22K yields substantial gains over its original zero-

Models	SSIM ↑	ImgSim ↑	mIoU ↑
<i>Large Vision-Language Models</i>			
Qwen3-VL	N/A	N/A	0.596
<i>Unified Large Language Models</i>			
Bagel	0.305	0.756	0.506
UniMedVL	0.183	0.767	0.488
Anole	N/A	N/A	N/A
Anole-MMRad	0.494	0.885	0.680

Table 4: **Evaluation of intermediate evidence quality.** SSIM and ImgSim measure the similarity between generated local images and predicted-box crops, while mIoU measures the overlap between predicted and reference bounding boxes.

shot performance. Under the same evaluation protocol, Anole-MMRad also reaches a performance level on par with representative open-source LVLM references. These results further support the practical utility of MMRad-22K as a multimodal supervision resource for adapting LVLMs to chest X-ray report generation.

We also assess whether models can effectively utilize the intermediate evidence. Specifically, we evaluate localization accuracy using mIoU between predicted and reference bounding boxes, and evaluate visual evidence quality using SSIM and BioMedCLIP (Zhang et al., 2025) image similarity (ImgSim) between generated local images and the corresponding predicted-box crops. SSIM and ImgSim are only applicable to models capable of

Models	RadCliQ ↓	RGe ↑	RGer ↑	RaTE ↑
<i>Large Vision-Language Models</i>				
Qwen3-VL	1.485	0.138	0.124	0.474
<i>Unified Large Language Models</i>				
Bagel	1.577	0.136	0.126	0.482
Anole	2.139	0.066	0.058	0.449
Anole-MMRad	1.554	0.160	0.143	0.513

Table 5: **Zero-shot transfer evaluation on IU X-Ray.** Models are evaluated without adaptation on the external dataset using CE metrics.

generating localized images.

As shown in Table 4, Anole-MMRad achieves the strongest overall localization and visual evidence quality, demonstrating that MMRad-22K enables models to learn grounded intermediate evidence generation. In contrast, the original Anole does not reliably produce such intermediate evidence.

To evaluate whether the learned multimodal supervision transfers beyond the MIMIC-CXR domain, we further conduct a zero-shot evaluation on the IU X-Ray (Demner-Fushman et al., 2016) dataset. Since reporting styles may vary across datasets, we focus on CE metrics that are less sensitive to surface-level wording differences. As shown in Table 5, Anole-MMRad consistently improves over the original Anole backbone and remains competitive with Qwen3-VL, suggesting that the multimodal evidence supervision learned from MMRad-22K provides transferable benefits beyond the training domain.

5.4 Case Study

Figure 7 illustrates the model’s systematic diagnostic trace on a representative CXR. The model first examines the mediastinal region by generating a localized image and textual observations that rule out cardiac and mediastinal abnormalities. It then transitions to the lung region where it identifies clear lung fields with high spatial accuracy. The final report closely aligns with the reference report. Additional case studies and examples of failure cases are provided in Appendix A.6.

6 Conclusion

In this work, we introduce MMRad-22K, a study-level structured multimodal dataset that unifies regional textual observations, anatomical grounding coordinates, localized images, and report targets into a structured format. Experiments show

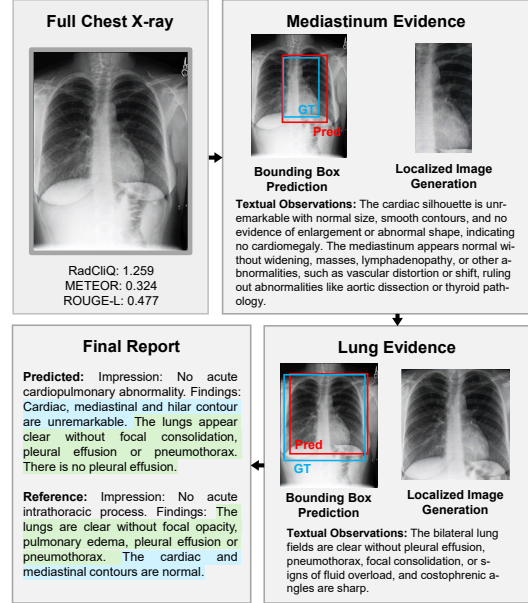


Figure 7: **Qualitative analysis by Anole-MMRad.** Green shading and blue shading represent consistent descriptions for the lung and mediastinal regions.

that this structured supervision is useful for CXR report generation: multimodal evidence outperforms weaker evidence formats in pilot comparisons, and adapting a unified LVLM with MMRad-22K improves over standard end-to-end and textual-evidence settings while reaching a performance level comparable to several open-source LVLM references. Overall, MMRad-22K provides a practical structured multimodal resource for training and evaluating CXR report generation models.

Limitations

Despite the encouraging results provided by MMRad-22K, this study has several limitations that open avenues for future work. First, the dataset is constructed by reorganizing existing report-aligned public resources rather than through new prospective expert annotation collected specifically for study-level report generation. Although we apply multi-stage verification and clinician evaluation, the resulting evidence units may still reflect source reporting conventions and annotation biases. Second, MMRad-22K has imbalanced coverage across anatomical regions and findings, with lung- and mediastinal-related evidence accounting for the majority of samples, which may limit representation of rarer regions or subtle abnormalities. Third, because the dataset is derived from MIMIC-CXR, it remains to be validated whether the observed gains generalize to other institutions, reporting styles,

and real clinical workflows.

Ethical Considerations

MMRad-22K is intended solely as a research resource for developing and evaluating chest X-ray report generation systems. It is derived from existing public radiology resources and should not be used as a clinical decision-making tool or as a substitute for expert interpretation. As with other medical vision-language resources, potential risks include propagation of source reporting biases, incomplete coverage of rare findings, and over-reliance on automatically generated outputs in safety-critical settings. We therefore emphasize that models trained on MMRad-22K are for research use only, and their clinical reliability and generalization to new institutions remain to be established.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant Nos. 62322604 and 62576207. Large language models were used to support evaluation in Sections 3 and 5, dataset construction in Section 4, and limited grammar checking during manuscript preparation. All technical content, experimental claims, and final wording were reviewed and verified by the authors.

References

- Anthropic. 2025. [Introducing claude haiku 4.5](#).
- Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kweon, Jungwoo Oh, Lei Ji, Eric I-Chao Chang, Tackeun Kim, and Edward Choi. 2023. [EHRXQA: A multi-modal question answering dataset for electronic health records with chest x-ray images](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. [Qwen3-vl technical report](#). *Preprint*, arXiv:2511.21631.
- Satanjeev Banerjee and Alon Lavie. 2005. [METEOR: An automatic metric for MT evaluation with improved correlation with human judgments](#). In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, Felix Meissen, Mercy Ranjit, Shaury Srivastav, Julia Gong, Noel C. F. Codella, Fabian Falck, Ozan Oktay, Matthew P. Lungren, Maria Teodora Wetscherek, and 2 others. 2024. [Maira-2: Grounded radiology report generation](#). *Preprint*, arXiv:2406.04449.
- Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. [Generating radiology reports via memory-driven transformer](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1439–1449, Online. Association for Computational Linguistics.
- 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, Jena Jitsev, and 4 others. 2024. [A vision-language foundation model to enhance efficiency of chest x-ray interpretation](#). *Preprint*, arXiv:2401.12208.
- Ethan Chern, Zhulin Hu, Steffi Chern, Siqi Kou, Jiadi Su, Yan Ma, Zhijie Deng, and Pengfei Liu. 2025. [Thinking with generated images](#). *Preprint*, arXiv:2505.22525.
- Ethan Chern, Jiadi Su, Yan Ma, and Pengfei Liu. 2024. [Anole: An open, autoregressive, native large multi-modal models for interleaved image-text generation](#). *Preprint*, arXiv:2407.06135.
- Jannette Collins and Eric J Stern. 2012. *Chest radiology: the essentials*. Lippincott Williams & Wilkins.
- DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. [Deepseek-v3 technical report](#). *Preprint*, arXiv:2412.19437.
- Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. 2022. [Improving the factual correctness of radiology report generation with semantic rewards](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4348–4360, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez,

- Sameer K. Antani, George R. Thoma, and Clement J. McDonald. 2016. [Preparing a collection of radiology examinations for distribution and retrieval](#). *J. Am. Medical Informatics Assoc.*, 23(2):304–310.
- Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. 2025. [Emerging properties in unified multimodal pretraining](#). *Preprint*, arXiv:2505.14683.
- Nicolas Deperrois, Hidetoshi Matsuo, Samuel Ruipérez-Campillo, Moritz Vandenhirtz, Sonia Laguna, Alain Rysler, Koji Fujimoto, Mizuho Nishio, Thomas M. Sutter, Julia E. Vogt, Jonas Kluckert, Thomas Frauenfelder, Christian Blüthgen, Farhad Nooralahzadeh, and Michael Krauthammer. 2025. [Radvlm: A multi-task conversational vision-language model for radiology](#). *Preprint*, arXiv:2502.03333.
- Google. 2025. [Gemini 3 flash: frontier intelligence built for speed](#).
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. [Lora: Low-rank adaptation of large language models](#). *Preprint*, arXiv:2106.09685.
- Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silvana Ciurea-Illcus, Christopher Chute, Henrik Marklund, Behzad Haghighi, Robyn L. Ball, Katie S. Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. 2019. [Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison](#). In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 590–597. AAAI Press.
- Baoyu Jing, Pengtao Xie, and Eric Xing. 2018. [On the automatic generation of medical imaging reports](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2577–2586, Melbourne, Australia. Association for Computational Linguistics.
- Peiyuan Jing, Kinhei Lee, Zhenxuan Zhang, Huichi Zhou, Zhengqing Yuan, Zhifan Gao, Lei Zhu, Giorgos Papanastasiou, Yingying Fang, and Guang Yang. 2026. [Reason like a radiologist: Chain-of-thought and reinforcement learning for verifiable report generation](#). *Medical Image Analysis*, 109:103910.
- Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. 2019. [Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs](#). *Preprint*, arXiv:1901.07042.
- Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. [Llava-med: Training a large language-and-vision assistant for biomedicine in one day](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Yifan Li, Fenghe Tang, Yingtai Li, and Shaohua Kevin Zhou. 2025. [Medreason-r1: Learning to reason for ct diagnosis with reinforcement learning and local zoom](#). *Preprint*, arXiv:2510.19626.
- Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. 2018. [Hybrid retrieval-generation reinforced agent for medical image report generation](#). In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 1537–1547.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Bo Liu, Xiangyu Zhao, Along He, Yidi Chen, Huazhu Fu, and Xiao-Ming Wu. 2025a. [Gemex-rmcot: An enhanced med-vqa dataset for region-aware multimodal chain-of-thought reasoning](#). In *Proceedings of the 33rd ACM International Conference on Multimedia, MM 2025, Dublin, Ireland, October 27-31, 2025*, pages 13213–13220. ACM.
- Bo Liu, Ke Zou, Li-Ming Zhan, Zexin Lu, Xiaoyu Dong, Yidi Chen, Chengqiang Xie, Jiannong Cao, Xiao-Ming Wu, and Huazhu Fu. 2025b. [Gemex: A large-scale, groundable, and explainable medical VQA benchmark for chest x-ray diagnosis](#). In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pages 21310–21320. IEEE.
- Kang Liu, Zhuoqi Ma, Xiaolu Kang, Yunan Li, Kun Xie, Zhicheng Jiao, and Qiguang Miao. 2025c. [Enhanced contrastive learning with multi-view longitudinal data for chest x-ray report generation](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 10348–10359. Computer Vision Foundation / IEEE.
- Kang Liu, Zhuoqi Ma, Xiaolu Kang, Zhusi Zhong, Zhicheng Jiao, Grayson Baird, Harrison X. Bai, and Qiguang Miao. 2024. [Structural entities extraction and patient indications incorporation for chest x-ray report generation](#). In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2024 - 27th International Conference, Marrakesh, Morocco, October 6-10, 2024, Proceedings, Part III*, volume 15003 of *Lecture Notes in Computer Science*, pages 433–443. Springer.

- Xiwei Liu, Yulong Li, Xinlin Zhuang, Xuhui Li, Jianxu Chen, Haolin Yang, Imran Razzak, and Yutong Xie. 2026. [Clincot: Clinical-aware visual chain-of-thought for medical vision language models](#). *Preprint*, arXiv:2603.01124.
- Hichem Metmer and Xiaoshan Yang. 2025. [Fedmrg: federated medical report generation via text-aware learning rate adjustment and multi-level prototype collaboration](#). *Multim. Syst.*, 31(2):170.
- Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. 2021. [Improving factual completeness and consistency of image-to-text radiology report generation](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5288–5304, Online. Association for Computational Linguistics.
- Junzhi Ning, Wei Li, Cheng Tang, Jiashi Lin, Chenglong Ma, Chaoyang Zhang, Jiayao Liu, Ying Chen, Shujian Gao, Yuandong Pu, Huihui Xu, Chenhui Gou, Ziyang Huang, Yi Xin, Qi Qin, Diping Song, Bin Fu, Guang Yang, Yuanfeng Ji, and 8 others. 2026. [Unimedvl: Unifying medical multimodal understanding and generation through observation-knowledge-analysis](#). *Preprint*, arXiv:2510.15710.
- Zach North. 2024. How to interpret chest radiographs (x-rays): a systematic approach.
- 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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Zelin Peng, Yichen Zhao, Yu Huang, Piao Yang, Feilong Tang, Zhengqin Xu, Xiaokang Yang, and Wei Shen. 2026. [Nearl: Interacted query adaptation with orthogonal regularization for medical vision-language understanding](#). *Preprint*, arXiv:2508.04101.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Fenghe Tang, Chengqi Dong, Wenxin Ma, Zikang Xu, Heqin Zhu, Zihang Jiang, Rongsheng Wang, Yuhao Wang, Chenxu Wu, and Shaohua Kevin Zhou. 2025. [U-bench: A comprehensive understanding of u-net through 100-variant benchmarking](#). *Preprint*, arXiv:2510.07041.
- Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. 2023. [Interactive and explainable region-guided radiology report generation](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 7433–7442. IEEE.
- Hao Wang, Shuchang Ye, Jinghao Lin, Usman Naseem, and Jinman Kim. 2026a. [Med-r2: Perception and reflection-driven complex reasoning for medical report generation](#). *Preprint*, arXiv:2504.02885.
- Pengyu Wang, Shuchang Ye, Usman Naseem, and Jinman Kim. 2026b. [Mrg-r1: Reinforcement learning for clinically aligned medical report generation](#). *Preprint*, arXiv:2512.16145.
- Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, and 56 others. 2025. [Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency](#). *Preprint*, arXiv:2508.18265.
- Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. 2023. [R2gengpt: Radiology report generation with frozen llms](#). *Preprint*, arXiv:2309.09812.
- 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, Leo A. Celi, and Mehdi Moradi. 2021. [Chest imagenome dataset for clinical reasoning](#). *Preprint*, arXiv:2108.00316.
- Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, and 8 others. 2024. [Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding](#). *Preprint*, arXiv:2412.10302.
- Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y. Ng, Curtis P. Langlotz, Vasanth Kumar Venugopal, and Pranav Rajpurkar. 2023. [Evaluating progress in automatic chest x-ray radiology report generation](#). *Patterns*, 4(9):100802.
- Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, and 5 others. 2025. [Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs](#). *Preprint*, arXiv:2303.00915.
- Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. [RaTEScore: A metric for radiology report generation](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15004–15019, Miami, Florida, USA. Association for Computational Linguistics.

A Appendix

A.1 More Information about Pilot Study

A.1.1 Prompt Templates

We use a shared report-generation objective across all settings, with common output requirements and the same report format. Because the additional evidence differs in modality, the prompt wording is minimally adapted to describe the corresponding input format (*red mark*). Thus, the comparison keeps the task objective and reporting constraints fixed while varying the form of additional region-level evidence.

Image-guided

You are a Senior Radiologist Consultant.

Based on the provided full CXR image and **localized image** of a specific anatomical region, generate a concise medical report.

[Format Requirements]

1. Include ONLY two sections: Impression and Findings.
2. Impression Section: Provide 1-2 sentences summarizing the key conclusions.
3. Findings Section: List 3-5 specific observations in brief, clear statements.
4. Keep the entire report around 100 words.

[Example of Desired Format]

Impression: Slight improvement in moderate pulmonary edema. No evidence of pneumonia. Findings: Lung volumes remain low however with slightly improved aeration compared to yesterday's examination. Moderate pulmonary edema is slightly improved. Bibasilar linear and plate like atelectasis is unchanged. A right PICC is unchanged in position. There is no pleural effusion or pneumothorax. Residual ethiodol from prior TACE projects over the right upper quadrant.

<FULL IMAGE INPUT>

<REGION IMAGE INPUT> The localized image shows the [REGION NAME] region.

Text-guided

You are a Senior Radiologist Consultant.

Based on the provided full CXR image and **localized textual** observation of a specific anatomical region, generate a concise medical report.

[Format Requirements]

1. Include ONLY two sections: Impression and Findings.
2. Impression Section: Provide 1-2 sentences summarizing the key conclusions.
3. Findings Section: List 3-5 specific observations in brief, clear statements.
4. Keep the entire report around 100 words.

[Example of Desired Format]

Impression: Slight improvement in moderate pulmonary edema. No evidence of pneumonia. Findings: Lung volumes remain low however with slightly improved aeration compared to yesterday's examination. Moderate pulmonary edema is slightly improved. Bibasilar linear and plate like atelectasis is unchanged. A right PICC is unchanged in position. There is no pleural effusion or pneumothorax. Residual ethiodol from prior TACE projects over the right upper quadrant.

<FULL IMAGE INPUT>

The localized textual observation shows the [REGION NAME] region, indicating [REGION TEXTUAL OBSERVATION INPUT].

Text + bbox

You are a Senior Radiologist Consultant.

Based on the provided full CXR image and **localized textual observation** of a specific anatomical region and its **grounding coordinates**, generate a concise medical report.

[Format Requirements]

1. Include ONLY two sections: Impression and Findings.
2. Impression Section: Provide 1-2 sentences summarizing the key conclusions.
3. Findings Section: List 3-5 specific observations in brief, clear statements.
4. Keep the entire report around 100 words.

[Example of Desired Format]

Impression: Slight improvement in moderate pulmonary edema. No evidence of pneumonia. Findings: Lung volumes remain low however with slightly improved aeration compared to yesterday's examination. Moderate pulmonary edema is slightly improved. Bibasilar linear and plate like atelectasis is unchanged. A right PICC is unchanged in position. There is no pleural effusion or pneumothorax. Residual ethiodol from prior TACE projects over the right upper quadrant.

<FULL IMAGE INPUT>

The relative bounding box [REGION BOUNDING BOX INPUT] shows the [REGION NAME] region , indicating [REGION TEXTUAL OBSERVATION INPUT].

A.1.2 Input Example

We provide one example to illustrate how the same study case is instantiated under different evidence

Multimodal

You are a Senior Radiologist Consultant.

Based on the provided full CXR image and **localized image** of a specific anatomical region and its **textual observation**, generate a concise medical report.

[Format Requirements]

1. Include ONLY two sections: Impression and Findings.
2. Impression Section: Provide 1-2 sentences summarizing the key conclusions.
3. Findings Section: List 3-5 specific observations in brief, clear statements.
4. Keep the entire report around 100 words.

[Example of Desired Format]

Impression: Slight improvement in moderate pulmonary edema. No evidence of pneumonia. Findings: Lung volumes remain low however with slightly improved aeration compared to yesterday's examination. Moderate pulmonary edema is slightly improved. Bibasilar linear and plate like atelectasis is unchanged. A right PICC is unchanged in position. There is no pleural effusion or pneumothorax. Residual ethiodol from prior TACE projects over the right upper quadrant.

<FULL IMAGE INPUT>

<REGION IMAGE INPUT> The localized textual observation shows the [REGION NAME] region, indicating [REGION TEXTUAL OBSERVATION INPUT].

formats. The example includes the full chest X-ray image together with its region-level evidence, showing how image-guided, text-guided, text + bbox, and multimodal inputs differ while sharing the same report-generation target.

Full CXR Image

Format	Full CXR Image	Localized Image	Bounding Box	Textual Observation
Image-guided	✓	✓		
Text-guided	✓			✓
Text + Bbox	✓		✓	✓
Multimodal	✓	✓		✓

Localized Image

Bounding Box: [0.369, 0.375, 0.744, 0.711]

Textual Observation: The cardiac silhouette shows moderate cardiomegaly with no interval change, indicating chronic volume overload likely due to heart failure. The hilar regions demonstrate pulmonary vascular congestion with engorged vessels, supporting left heart failure.

A.2 More Information about Dataset Construction

A.2.1 Summary of the Verification Pipeline

Approximately 14% of reformulated evidence units were revised during the self-check stage. A coarse error breakdown suggests that about 9% involved omission of source-supported content, around 4% involved hallucinated or unsupported statements, and roughly 1% involved mismatches between bounding boxes and the corresponding textual observations. In the cross-check stage, Qwen2.5-72B was used as a verifier to compare the reformulated evidence against the corresponding study-level report. About 2% of cases were flagged at this stage for potential report-level inconsistencies or uncertain anatomical correspondence and were sent to clinician adjudication.

A.2.2 Clinician Evaluation Rubrics

Clinical Accuracy evaluates whether the regional observations are medically correct and consistent with the reference study findings, as shown in Table 6. **Completeness** measures whether the reformulated evidence units cover the clinically important findings documented in the study, as shown in Table 7. **Visual-Text Consistency** evaluates

whether each regional text description is supported by, and spatially aligned with, its associated localized image crop, as shown in Table 8.

A.2.3 Stratified Clinician Evaluation and Error Analysis

To further assess the reliability of MMRad-22K evidence units, we stratify the 900 clinician-reviewed cases by clinical subgroups, including normal studies and cases with lung, cardiac, pleural, or other abnormalities, as shown in Table 9. These categories are non-exclusive. Clinical accuracy and visual-text consistency remain consistently high across subgroups. The main variation is observed in completeness.

Error analysis in Table 10 further shows that most lower-rated cases are caused by incomplete regional evidence, mainly due to omitted minor secondary findings. Region-text mismatch, unsupported observations, and ambiguous localization are relatively uncommon, while major abnormalities are generally preserved.

Score	Clinical Accuracy
5	Fully clinically accurate and well aligned with the reference report.
4	Clinically accurate with only negligible imprecision.
3	Generally correct but contains minor inaccuracies or ambiguous descriptions.
2	Multiple clinically inaccurate or misleading observations.
1	Major medical errors or hallucinated findings inconsistent with the reference report.

Table 6: Rubric for clinical accuracy evaluation.

Score	Completeness
5	Comprehensive coverage of clinically relevant findings described in the reference report.
4	Largely complete with only minor omissions.
3	Covers major findings but misses some secondary observations.
2	Several clinically relevant findings are omitted.
1	Most clinically important findings are missing.

Table 7: Rubric for completeness evaluation.

A.3 More Information about the MMRad-22K Dataset

A.3.1 Additional Dataset Examples

Fig. 8 presents additional examples of MMRad-22K. Each sample includes the full CXR image, one to four anatomy-guided multimodal evidence

Score	Visual-Text Consistency
5	Strong visual-text correspondence with clear and well-localized visual support.
4	Good visual-text alignment with only minor mismatch or ambiguity.
3	Partial alignment between the localized image evidence and the textual observation.
2	Weak or unclear correspondence between the image evidence and the text.
1	The localized image evidence is inconsistent with the corresponding textual observation.

Table 8: Rubric for visual-text consistency evaluation.

units, and the corresponding final report. These examples illustrate the study-level organization of the dataset, where localized image crops, textual observations, and grounding coordinates are grouped into clinically relevant regions and linked to the report target. They also show the variability of MMRad-22K across studies, ranging from simpler cases with a single evidence unit to more complex cases involving multiple anatomical groups such as lung, mediastinal, bone, and other regions.

A.3.2 Anatomy-group Statistics

We further report the distribution of fine-grained anatomical regions in MMRad-22K. As shown in Figure 9, the dataset exhibits a clearly long-tailed regional distribution, with *lung* and *mediastinal* related regions accounting for the largest proportion of evidence, broadly consistent with the dominant focus of CXR interpretation. At the same time, less frequent but clinically relevant regions from the *bone* and *others* groups are also represented, including spine, abdomen, and support-device-related regions. This suggests that the coarse anatomy-guided grouping used in MMRad-22K still preserves substantial fine-grained regional diversity.

A.4 More Information about Evaluation Metrics

- **BLEU** (Papineni et al., 2002) measures n-gram precision between the generated report and the reference report.
- **METEOR** (Banerjee and Lavie, 2005) combines precision, recall, and a fragmentation penalty, while additionally considering word order and synonym matching.
- **ROUGE** (Lin, 2004) measures n-gram recall with respect to the reference report.

Clinical Subgroup	Occurrence Rate	Clinical Accuracy	Completeness	Visual-Text Consistency
Normal	28.9%	4.83	4.58	4.96
Lung Abnormality	44.4%	4.51	4.08	4.83
Cardiac Abnormality	45.6%	4.57	4.18	4.85
Pleural Abnormality	26.7%	4.54	4.17	4.85
Others	22.2%	4.28	3.92	4.62

Table 9: **Stratified clinician evaluation of MMRad-22K evidence units.** Clinical subgroups are non-exclusive. Scores are averaged over the clinician-reviewed cases using three 5-point Likert scales.

Primary Error Type	Ratio
Incomplete Regional Evidence	76.8%
Minor Wording or Detail Issue	18.9%
Over-generalized or Unsupported Observation	2.8%
Mismatch or Ambiguous Crop	1.5%

Table 10: **Error analysis of clinician-reviewed evidence units.** Most residual issues are related to incomplete regional evidence, mainly due to omitted minor secondary findings.

- **RadGraph F1** (Delbrouck et al., 2022) extracts radiology entities and relations and computes an F1 score over their overlap, mainly for chest X-ray reports.
- **RaTEScore** (Zhao et al., 2024) evaluates radiology reports at the entity level by matching clinically important entities with type-aware embedding similarity, making it robust to synonyms and sensitive to negation.
- **RadCliQ** (Yu et al., 2023) combines BLEU, BERTScore, CheXbert vector similarity, and RadGraph F1 into a composite score that is optimized to better align with radiologist judgment.

A.5 More Information About Evidence Construction and Inference Settings

Each anatomy-guided evidence unit in MMRad-22K contains three aligned components: a regional textual observation, an anatomical bounding box, and a localized image crop. By training on these paired multimodal evidence structures together with the final report target, the adapted Anole model acquires both localization-related prediction ability and visual evidence generation ability within a unified autoregressive framework.

This design gives rise to two practical ways of incorporating localized visual evidence at inference time. In the **Generated** setting, we follow Anole’s native multimodal generation paradigm

and use autoregressively generated visual evidence together with regional textual observations. In the **Grounded** setting, the model first predicts anatomical regions, which are then mapped back to the source chest X-ray to extract localized image crops as grounded visual evidence, which is aligned with the crop-based localized evidence setting examined in Section 3.

A.6 More Cases Generated by Anole-MMRad

In addition to the main case study, we provide more qualitative examples to illustrate how the structured multimodal evidence is used across different studies. These examples are intended to complement the quantitative results by showing the intermediate regional evidence, the predicted bounding boxes, the generated localized images, and the final reports within the same study-level generation process. They also help clarify that MMRad-22K does not only provide final report targets, but organizes intermediate evidence in a form that can be inspected together with the generated report.

A.6.1 Representative Cases

Figure 10 presents representative examples of report generation with structured multimodal evidence. The examples cover both normal and abnormal studies, illustrating that the structured regional evidence can support clinically coherent study-level reports across different levels of difficulty. In relatively normal cases, the model produces findings that closely match the reference reports. In abnormal cases, it is able to combine multiple localized evidence units to capture major findings such as pulmonary edema, pleural effusion, or bibasal opacities, demonstrating the benefit of anatomy-guided multimodal evidence for report generation.

A.6.2 Failure Cases

Figure 11 presents representative failure cases. In both examples, the model retains a broadly reason-

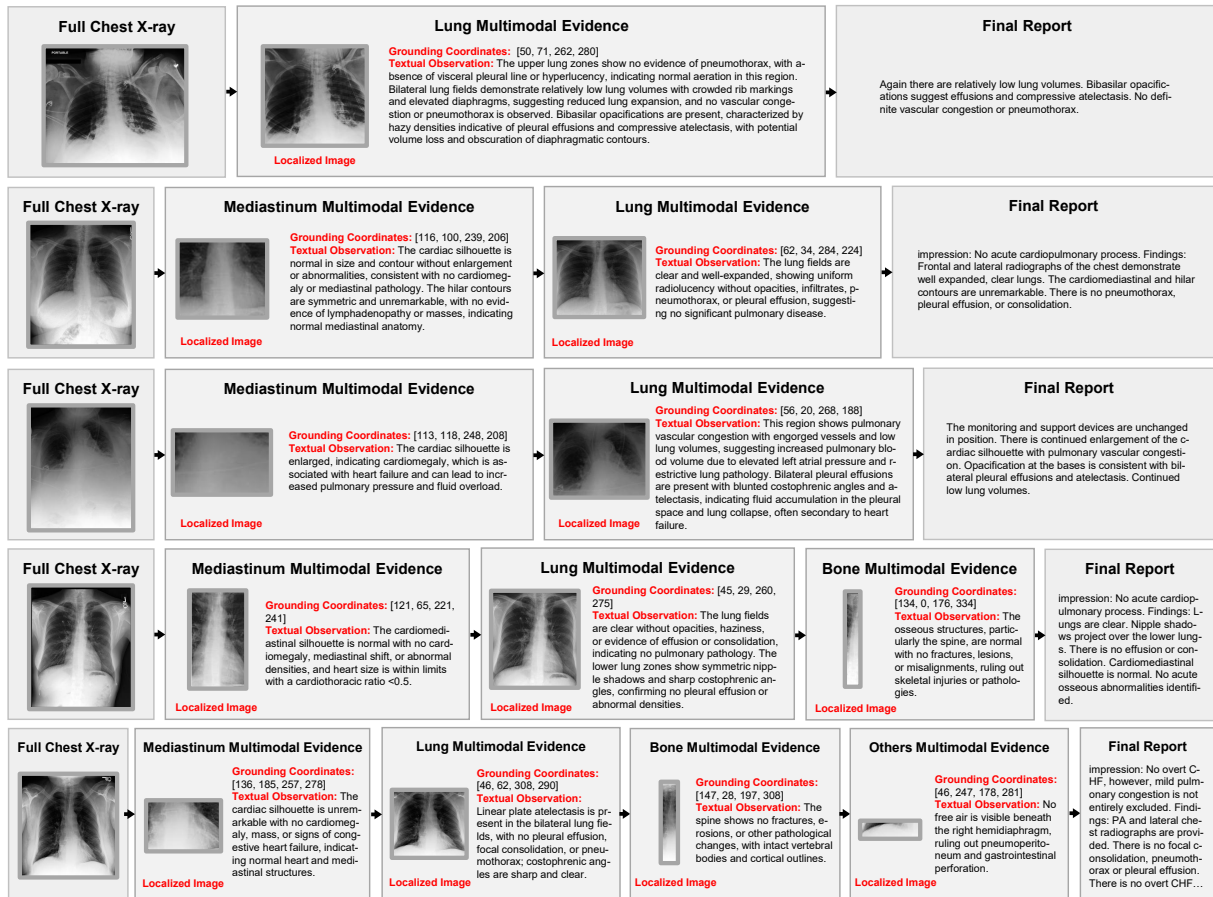


Figure 8: **More Examples of MMRad-22K Dataset.** Each sample includes the full CXR image, structured multimodal evidence, and the final report.

able study-level impression but does not fully capture all findings described in the reference report. The first case omits low lung volumes, whereas the second misses a more localized left lower lobe linear opacity/atelectatic change. These cases suggest that, even with anatomy-guided multimodal evidence, some supported findings may be under-emphasized or omitted when they are integrated into the final study-level report.

A.7 Ethics, Data Use, and Disclosure Information

A.7.1 Potential Risks

MMRad-22K is constructed for research on chest X-ray report generation and multimodal evidence modeling. Potential risks include propagation of source reporting biases, under-representation of rare findings or patient subgroups, and inappropriate over-trust in automatically generated reports. In addition, because the dataset reorganizes report-aligned supervision into study-level evidence, generated outputs may appear clinically plausible even when they omit important details. For these rea-

sons, MMRad-22K and models trained on it should be used only in research settings and not for direct clinical decision-making.

A.7.2 Data Sources and Usage Terms

MMRad-22K is derived from MIMIC-CXR (Johnson et al., 2019) and GEMeX-ThinkVG (Liu et al., 2025a), both of which are existing research resources with their own access conditions and usage requirements. Our use of these resources is limited to research purposes and follows their intended access setting. MMRad-22K is released only as a derived research artifact consistent with the original access restrictions, and it is not intended to circumvent the access controls, licensing conditions, or redistribution boundaries of the underlying data sources.

A.7.3 Intended Use and Distribution Boundary

The intended use of MMRad-22K is academic research on chest X-ray report generation, multimodal evidence organization, and related evalu-

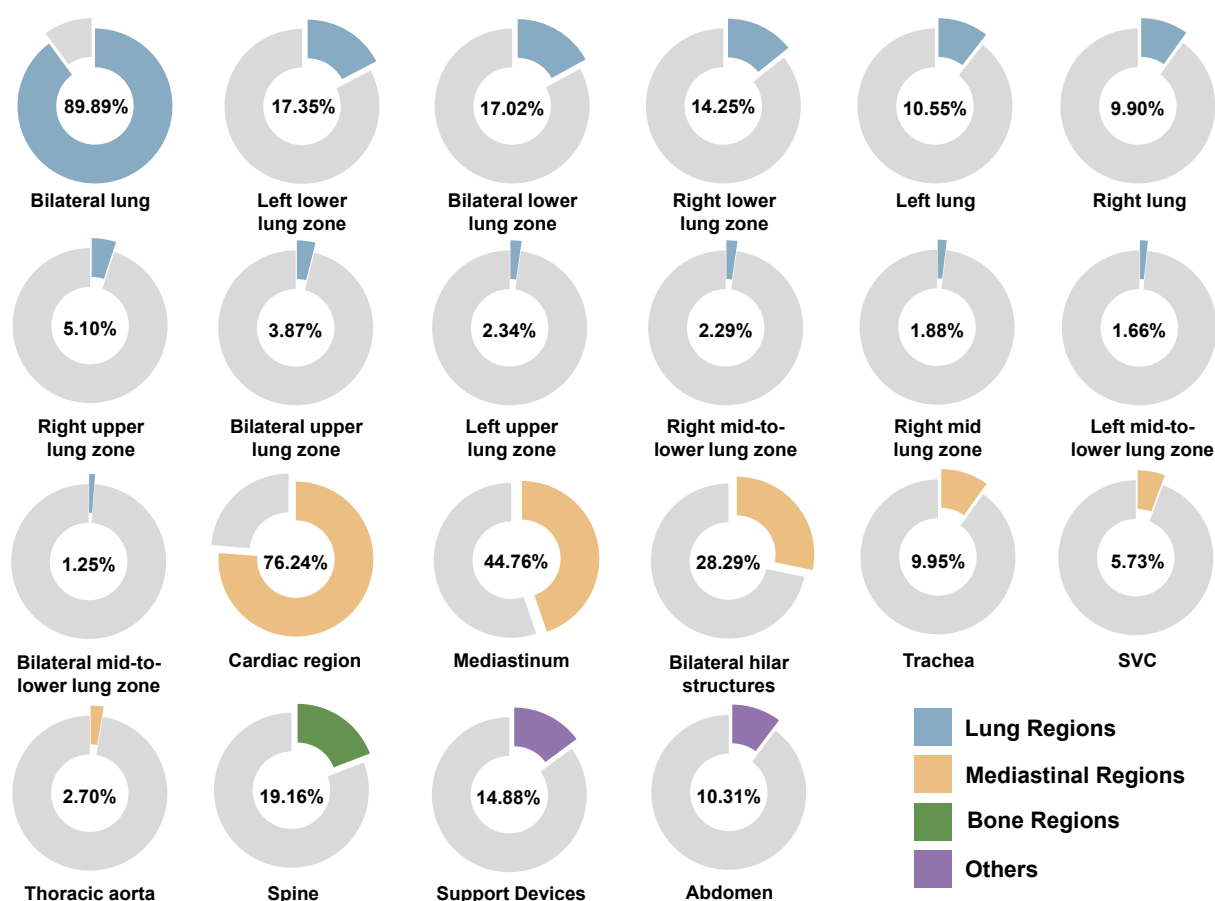


Figure 9: **Fine-grained anatomical region distribution in MMRad-22K.** For readability, only regions with frequency greater than 1% are shown. The distribution is long-tailed, with lung- and mediastinal-related regions predominating, while bone, abdominal, and support-device regions remain represented.

ation. It is not intended for clinical deployment, patient care, or autonomous medical reporting. Any released derivative artifact should remain compatible with the original data-access conditions, and the dataset should not be used outside research environments where these conditions do not apply.

A.7.4 Use of AI Assistants

Large language models were used to support evaluation in Sections 3 and 5, and during dataset construction to reformulate question-level grounded traces into report-generation-oriented evidence units and to support intermediate verification, as described in Section 4. These model-assisted outputs were not accepted without control. AI assistants were also used for limited grammar checking during manuscript preparation, with all technical content, experimental claims, and final wording reviewed and revised by the authors.

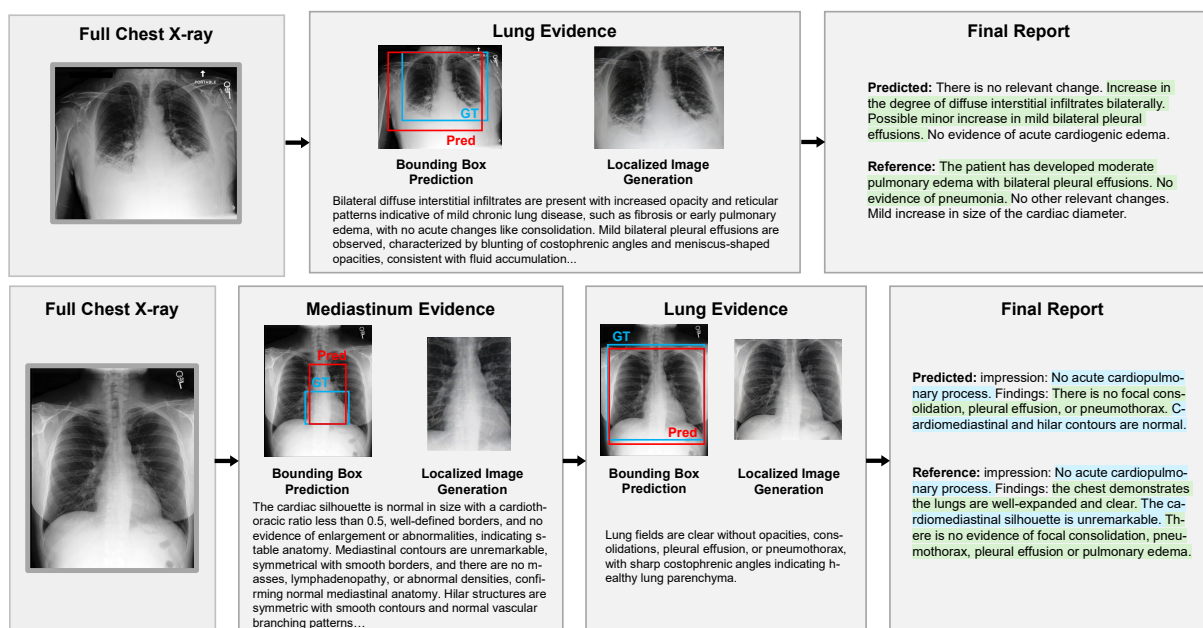


Figure 10: **Representative cases of report generation with structured multimodal evidence.** Each example shows the full chest X-ray image, regional evidence units, the predicted report, and the reference report. The cases cover both relatively normal and abnormal studies, illustrating that structured multimodal evidence can support clinically coherent study-level report generation. Green shading and blue shading represent consistent descriptions for the lung and mediastinal region.

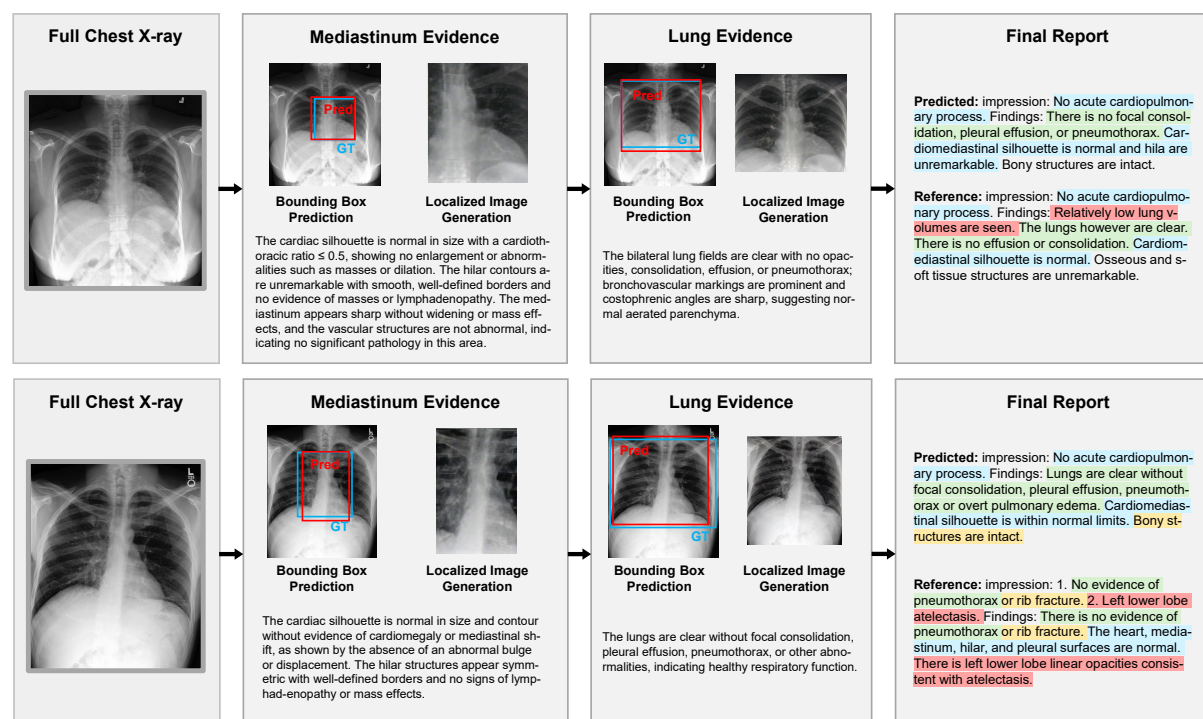


Figure 11: **Failure cases of report generation with structured multimodal evidence.** Each example shows the full chest X-ray image, regional evidence units, the predicted report, and the reference report. Although the overall study-level impression remains broadly reasonable, these cases show that some supported findings may still be under-emphasized or omitted in the final report. Green shading, blue shading, and orange shading represent consistent descriptions for the lung, mediastinal, and bone region. Red shading represents errors.