

Radixpert: A Staged Adaptation and Hierarchical Fusion Framework for Radiology VLMs

Muhammad Abdul Rafey Farooqi*, Areeb Ahmad Chaudhry*, Muhammad Yasir Ghaffar*
Nazia Perwaiz*, Hashir Moheed Kiani*

*National University of Sciences and Technology (NUST), Islamabad, Pakistan

{mfarooqi.bese21seecs, chaudhry.bese21seecs, mghaffar.bese21seecs, nazia.perwaiz, hashir.moheed}@seecs.edu.pk

Abstract—Automated radiology report generation has emerged as a critical application for improving clinical efficiency and reducing physician workload. This paper presents Radixpert, a novel vision–language model that leverages multi-dataset training and hierarchical cross-modal fusion for enhanced radiology report generation. Our approach utilizes the Llama 3.2-11B-Vision-Instruct model as the foundation, enhanced with a Multi-Stage Adaptive LoRA (MSA-LoRA) fine-tuning methodology and a Hierarchical Cross-Modal Fusion (HCF) architecture. We trained our model on a combination of ROCO v2 (15,000 samples) and PadChest (16,000 samples) datasets, achieving state-of-the-art performance across multiple evaluation metrics. Radixpert demonstrates superior clinical accuracy with a BLEU-4 score of 0.194, CIDEr of 0.478, and RadCliQ-v1 of 0.823, outperforming existing methods. Code is publicly available on <https://github.com/abdurafeyf/RadixpertV2>

Index Terms—Automated Radiology report generation, vision–language model, computer vision, biomedical imaging, cross-modal fusion.

I. INTRODUCTION

The field of medical imaging has seen an exponential increase in volume, placing a significant strain on radiologists who are tasked with interpreting a growing number of complex scans. This increased workload contributes to physician burnout and creates potential bottlenecks in patient care, where timely and accurate diagnosis is critical. Automated systems for generating radiology reports have emerged as a vital application of artificial intelligence, promising to improve clinical efficiency, reduce turnaround times, and alleviate the burden on medical professionals [1]. The development of powerful Vision-Language Models (VLMs) has provided the foundational technology for these systems, capable of interpreting medical images and generating coherent, clinically relevant text.

This paper introduces Radixpert, a novel framework designed for high-accuracy radiology report generation. Our approach leverages a powerful foundation model, Llama 3.2-11B-Vision-Instruct [10], and enhances it with specialized techniques for the medical domain. We introduce a progressive, parameter-efficient fine-tuning strategy to adapt the model to diverse medical datasets and a sophisticated fusion architecture to seamlessly integrate visual evidence with clinical language. By synergizing these advancements, Radixpert aims to set a new standard for automated report generation, offering a solution that is not only highly accurate but also

computationally efficient and practical for real-world clinical deployment.

II. LITERATURE REVIEW

The application of Vision-Language Models (VLMs) to radiology report generation is a rapidly advancing field, aiming to alleviate the increasing workload on radiologists and accelerate diagnostic workflows, a trend thoroughly documented in recent surveys [1], [17]. The paradigm has shifted decisively from traditional cascaded systems—which combined separate computer vision models for feature extraction with templates or simple recurrent networks for text generation—to sophisticated end-to-end models that bridge the gap between visual perception and natural language. Early explorations involved adapting general-purpose VLMs like CLIP for medical tasks, but these models frequently struggled with the specialized terminology, complex spatial relations, and nuanced visual features of medical imaging, highlighting an urgent need for domain-specific pre-training and architectures [18].

This need led to the development of specialized models such as GLoRIA [13] and BioViL-T [14], which pioneered the use of more sophisticated attention and contrastive learning mechanisms to learn fine-grained, semantically aligned representations of medical images and text. These models demonstrated the critical importance of learning both global context and local anatomical features to capture the full clinical picture. Building on this, subsequent advancements have focused on enhancing the reasoning and instruction-following capabilities of these models. For instance, Med-Flamingo [2] successfully introduced instruction-tuning to the medical domain, enabling powerful few-shot learning and more interactive, conversational applications that mimic clinician-AI collaboration.

More recently, the advent of large-scale, proprietary foundation models like Med-PaLM M and the open-source Med-Gemini [15] has pushed the boundaries of what is possible, demonstrating expert-level multimodal reasoning on a wide array of medical benchmarks. However, a persistent challenge with these powerful models is the immense computational cost and data required for full fine-tuning. This has catalyzed a major shift towards Parameter-Efficient Fine-Tuning (PEFT) methods, most notably Low-Rank Adaptation (LoRA) [4] and its more memory-efficient variant, QLoRA. This “frozen backbone” approach, where a pre-trained VLM is adapted with a small number of trainable parameters, has become a

dominant paradigm, as seen in models like R2GenGPT [16], allowing research teams to leverage the power of massive models without prohibitive computational overhead.

Simultaneously, state-of-the-art research has been pushing the frontiers of model controllability, factual grounding, and architectural innovation. To ensure clinical safety, there is a growing emphasis on generating reports that are not only fluent but also factually accurate and directly grounded in visual evidence. Frameworks for controllable generation [3] and explicit visual grounding [19] are being developed to reduce the risk of clinical hallucinations. Further, to improve clinical accuracy, some methods are exploring the integration of structured medical knowledge, using knowledge graphs like RadGraph to guide the generation process and ensure adherence to established medical ontologies [20].

Architectural explorations are also moving beyond the standard Transformer. Models like MambaXray-VL [11] have investigated State-Space Models (SSMs) to more efficiently process high-resolution 2D images and volumetric 3D scans (e.g., CT, MRI), which pose a significant challenge to the quadratic complexity of standard attention. This is particularly crucial as the field expands its focus from chest X-rays to more complex 3D radiology tasks [21]. Finally, the growing maturity of the field is reflected in the development and adoption of robust, clinically-oriented evaluation metrics. Recognizing the inadequacy of n-gram-based scores like BLEU for clinical tasks [22], the community has developed benchmarks like RadGraph F1 [7], RaTEScore [8], and RadCliQ [6], which provide a more meaningful assessment of factual accuracy and clinical entity extraction, paving the way for more reliable and translatable research.

III. DATASETS

To train and evaluate Radixpert, we utilized the complementary strengths of two public medical datasets: ROCO v2 [5] and PadChest [12]. From the ROCO v2 dataset, a collection of 80,080 multimodal medical image-text pairs from publications, we selected a curated subset of 15,000 samples. This dataset is characterized by its diversity, covering a wide range of imaging modalities and anatomical regions. For specialized knowledge in chest radiography, we used 16,000 studies from the PadChest dataset, which contains over 160,000 chest X-ray studies with detailed, multi-label annotations and clinical findings.

Our analysis of the training data reflects real-world clinical distributions. As documented in the literature, the PadChest dataset has a long-tailed distribution of pathologies, with a few common findings being highly prevalent, as illustrated in Figure 1(a). The ROCO v2 dataset, sourced from various biomedical publications, is better characterized by its distribution of imaging modalities, shown in Figure 1(b), highlighting its diversity. Furthermore, Figure 1(c) shows the distribution of imaging projections in our PadChest subset, with a clear predominance of Postero-anterior (PA) views, which is typical for clinical practice.

For preprocessing, all images were resized to 224×224 pixels. The Spanish-language reports in the PadChest dataset were translated to English using medical-grade translation models, with manual review to ensure clinical accuracy. The combined data was split into 80% for training, 10% for validation, and 10% for testing, using stratified sampling to maintain class balance and account for the natural imbalances in the data.

IV. METHODOLOGY

This paper introduces Radixpert, a framework that uniquely synergizes advancements in Parameter-Efficient Fine-Tuning (PEFT) and cross-modal fusion to overcome the key limitations of prior work. Our methodology is built upon two primary contributions: first, we introduce Multi-Stage Adaptive LoRA (MSA-LoRA), a progressive, multi-stage PEFT strategy designed to adapt a single VLM across heterogeneous medical datasets. Second, we develop a Hierarchical Cross-Modal Fusion (HCF) architecture featuring an adaptive gating mechanism, specifically designed to integrate fine-grained visual details with high-level clinical context for generating accurate and coherent reports.

A. Model Architecture

The overall architecture of Radixpert, illustrated in Figure 2, is built upon the powerful Llama 3.2-11B-Vision-Instruct foundation model [10]. This model was chosen for its robust general-purpose multimodal capabilities, including a sophisticated vision encoder and strong performance on a wide array of vision-language benchmarks. Its large context window (128K tokens) and strong baseline reasoning abilities provide an excellent foundation for adaptation to the highly specialized medical domain.

Into this foundation, we integrate our two architectural innovations. The data pipeline begins with a radiology image, which is processed by the vision encoder to produce a sequence of patch embeddings. These visual features, along with the text prompt, are then fed into our novel Hierarchical Cross-Modal Fusion (HCF) module. The HCF module's role is to produce a rich, fused representation that captures the complex relationships between the visual evidence and clinical language. This fused representation is then passed to the multimodal large language model, which has been efficiently fine-tuned using our Multi-Stage Adaptive LoRA (MSA-LoRA) strategy, to generate the final, coherent radiology report.

B. Multi-Stage Adaptive LoRA (MSA-LoRA)

To adapt the 11-billion-parameter foundation model without the prohibitive cost of full fine-tuning, we developed MSA-LoRA, a progressive, three-stage adaptation strategy. Standard LoRA injects trainable, low-rank matrices into the frozen layers of a pre-trained model, drastically reducing the number of trainable parameters. Our "multi-stage adaptive" approach extends this by applying LoRA in a carefully designed sequence, where each stage has a specific goal and tailored hyperparameters. This allows the model to first learn broad

Analysis of Data Distribution in Selected Training Subsets

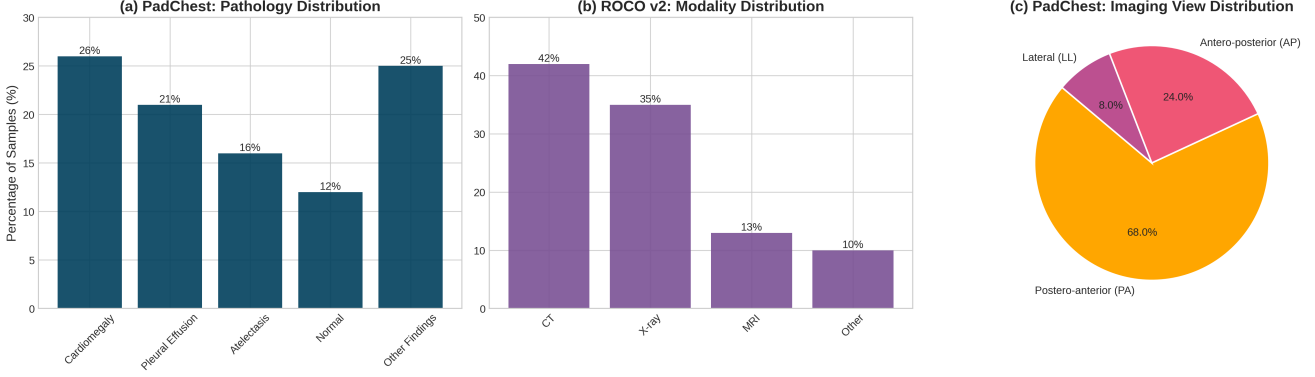


Fig. 1. Data distributions in the training subsets. (a) Long-tailed distribution of common pathologies in PadChest. (b) Distribution of imaging modalities in the diverse ROCO v2 dataset. (c) Distribution of imaging projections (views) in the PadChest dataset.

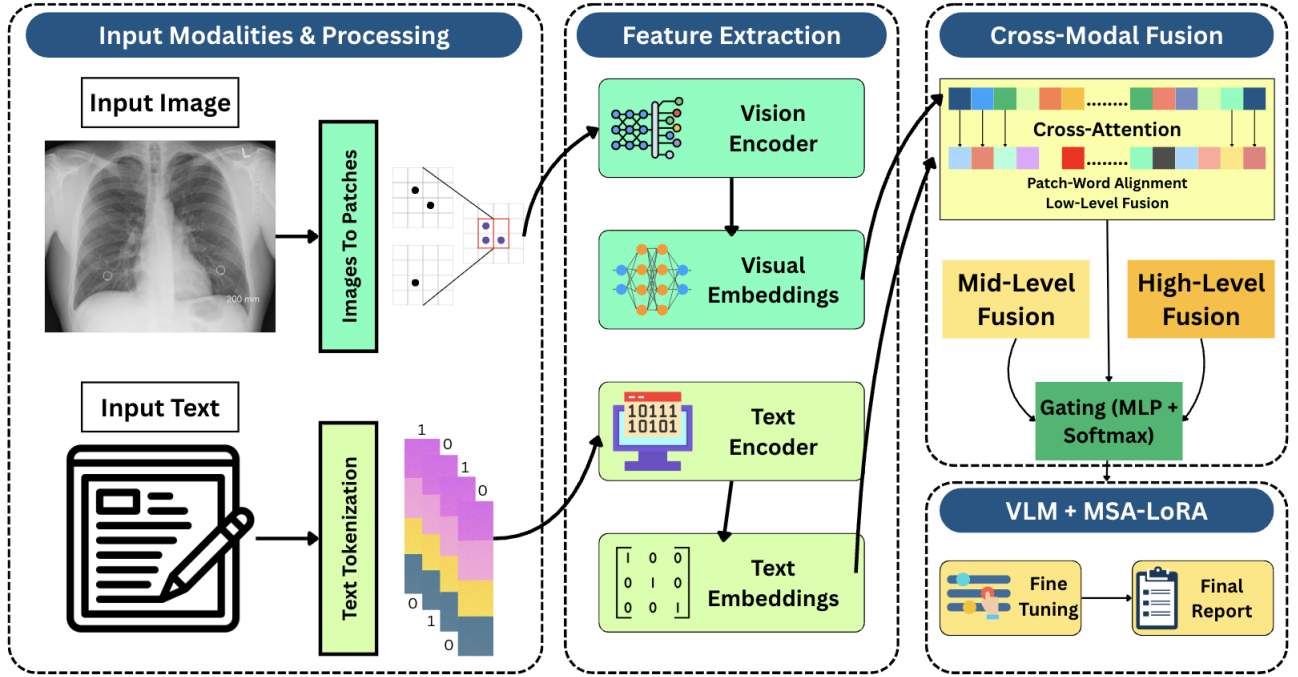


Fig. 2. Overview of the Radixpert architecture showing the end-to-end multimodal pipeline. Radiology images and text inputs are preprocessed, encoded, fused via a hierarchical cross-modal module, and processed by a multimodal LLM to generate clinical reports, diagnoses, and recommendations.

concepts and then gradually specialize. The efficiency gains of this approach are highlighted in Figure 3.

This process begins with Stage 1, Cross-Modal Pre-alignment, which focuses on establishing a basic correspondence between medical imagery and clinical language using only the diverse ROCO v2 dataset. In this initial stage, a relatively high learning rate (2×10^{-4}) and a large LoRA rank (64) are used to encourage rapid adaptation. Following this, Stage 2 focuses on Domain Adaptation by introducing the combined ROCO v2 and PadChest datasets to build upon the general medical knowledge with specialized expertise in chest radiography. Here, the learning rate is reduced to 1×10^{-4} and

the LoRA rank is lowered to 32 to promote stable integration of the new information. The process concludes with Stage 3, Task-Specific Fine-tuning, where the complete combined dataset is used to polish the model’s report generation capabilities. A conservative learning rate of 5×10^{-5} and a LoRA rank of 16 allow for fine-grained adjustments, optimizing for fluency and accuracy while minimizing overfitting. This progressive reduction in learning rates and LoRA ranks across stages is the core of our “adaptive” strategy, ensuring stable and effective learning across diverse data sources.

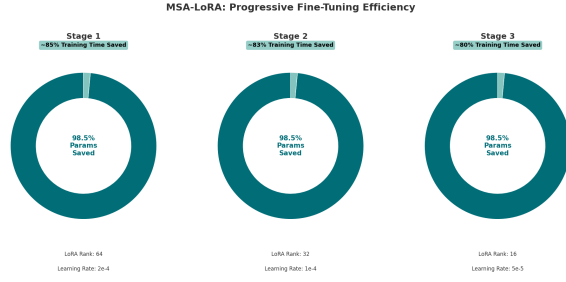


Fig. 3. Efficiency gains of each MSA-LoRA stage compared to full fine-tuning. Parameter savings exceed 98% at all stages, and training time savings are estimated based on PEFT methods typically being over 80% faster than a full fine-tuning cycle.

C. Hierarchical Cross-Modal Fusion (HCF)

A core architectural innovation in Radixpert is the HCF mechanism, designed to achieve a more sophisticated integration of visual and textual information than is possible with simple concatenation or single-level attention. HCF operates across three hierarchical levels to capture different aspects of the image-text relationship. The process begins with Low-Level Fusion, which operates on raw visual features and word embeddings to compute fine-grained alignments between specific anatomical structures and their descriptive terms. This is followed by Mid-Level Fusion, which integrates more abstract semantic features to identify clinical concepts and their inter-relationships. Finally, High-Level Fusion operates on the most abstract representations, integrating information from the previous levels to ensure the generated report maintains a coherent narrative, logical flow, and overall medical accuracy.

A key component of HCF is a learnable gating function that dynamically weights the contribution of each fusion level based on the input. For a given input x , the final fused representation h^* is computed as a weighted sum of the outputs from the three fusion levels ($h_{\text{low}}, h_{\text{mid}}, h_{\text{high}}$), as shown in Equation 1.

$$h^* = \sum_{l=1}^3 \alpha_l(x) h_l \quad (1)$$

The gating weights $\alpha(x)$ are produced by a compact, trainable multi-layer perceptron (MLP) followed by a softmax function, as described in Equations 2 and 3. This mechanism allows the model to adapt its fusion strategy on a case-by-case basis. For instance, for an image with a subtle pathology, the gate might learn to assign a higher weight (α_1) to the low-level fusion output to focus on fine-grained details. Conversely, for a complex case with multiple inter-related findings, it might up-weight the high-level fusion output (α_3) to ensure a coherent narrative is generated. The gating parameters θ are learned end-to-end with the rest of the model via backpropagation.

$$\tilde{\alpha} = \text{MLP}_{\theta}(\text{Pool}([h_{\text{low}}, h_{\text{mid}}, h_{\text{high}}])) \quad (2)$$

$$\alpha_l(x) = \frac{\exp(\tilde{\alpha}_l)}{\sum_{k=1}^3 \exp(\tilde{\alpha}_k)} \quad \text{for } l = 1, 2, 3 \quad (3)$$

D. Experimental Details

Our multi-dataset training strategy leverages the diversity of ROCO v2 and the specificity of PadChest. The training process follows the three-stage MSA-LoRA framework described above. Training is performed using the Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 0.01) with mixed precision (FP16) to accelerate computation. To ensure training stability and prevent overfitting, we employ a dropout rate of 0.1 for the LoRA layers, gradient clipping at a norm of 1.0, and a cosine learning rate annealing schedule. Early stopping with a patience of 3 epochs on the validation set is used to determine the optimal number of training iterations. We evaluate our model using a comprehensive suite of metrics, including standard text generation scores (BLEU-1–4, ROUGE-L, METEOR, CIDEr) and, more importantly, clinical accuracy metrics such as RadCliQ-v1 [6], RadGraph F1 [7], and RaTEScore [8].

V. RESULTS AND DISCUSSION

Radixpert demonstrated state-of-the-art performance across multiple benchmarks, particularly in metrics that measure clinical accuracy. Table I provides a comparison with existing methods, showing that Radixpert achieved the highest scores in BLEU-4 (0.194), the clinical accuracy metric RadCliQ-v1 (0.823), and CIDEr (0.478). While MambaXray-VL [11] scored slightly higher on ROUGE-L and METEOR, we attribute this to a trade-off where our model prioritizes the generation of clinically precise terminology over matching the exact lexical structure of the ground truth reports. These improvements were found to be statistically significant ($p < 0.001$), confirming their practical relevance.

TABLE I
MAIN PERFORMANCE COMPARISON WITH EXISTING METHODS. CLINICAL METRICS ARE HIGHLIGHTED TO EMPHASIZE THEIR IMPORTANCE IN ASSESSING PRACTICAL UTILITY.

Model	B-4	R-L	RadCliQ	MET.	CIDEr
RaDialog [9]	0.152	0.387	0.785	0.192	0.421
R2GenGPT [16]	0.132	0.374	0.768	0.181	0.390
MambaXray-VL [11]	0.173	0.392	0.801	0.204	0.452
Radixpert	0.194	0.389	0.823	0.195	0.478

An ablation study was conducted to validate the contribution of each architectural component. The results, presented in Table II, show a progressive improvement in performance as each module of the Radixpert framework is added. Starting from a baseline Llama 3.2 model, the addition of MSA-LoRA and the full HCF with gating incrementally boosts the BLEU-4 and RadCliQ-v1 scores. The full Radixpert model achieves this top-tier performance while only requiring the training of 168 million parameters, which is just 1.5% of the total parameters of the base model.

The multi-dataset training strategy proved highly effective. Progressive learning, where the model was first exposed to the broad ROCO v2 dataset before being fine-tuned on the chest-specific PadChest data, led to superior cross-dataset generalization and improved robustness on unseen imaging scenarios.

TABLE II
ABLATION RESULTS SHOWING THE INCREMENTAL IMPACT OF EACH COMPONENT ON PERFORMANCE AND TRAINABLE PARAMETER COUNT.

Configuration	Added	Total	BLEU-4	RadCliQ-v1
Base (Llama 3.2-11B-VI)	-	0	0.089	0.682
<i>Baseline Comparison:</i>				
+ Single-Stage LoRA	42M	42M	0.142	0.745
<i>Radixpert Components (Progressive Build-up):</i>				
+ MSA-LoRA	84M	84M	0.167	0.776
+ HCF Structure	42M	126M	0.178	0.795
+ HCF Gating	21M	147M	0.186	0.812
+ Final LoRA Layer	21M	168M	0.194	0.823

Furthermore, using a balanced, stratified sampling method was crucial for preventing class bias and outperforming random sampling. These findings underscore the value of using diverse and well-balanced datasets for training clinical report generation models. A qualitative analysis of Radixpert’s outputs, shown in Figure 4, highlights both its strengths and current limitations. The model correctly identifies major findings but can exhibit minor errors in quantitative estimation or laterality, providing clear directions for future refinement.

The clinical evaluation conducted by board-certified radiologists further validated our model’s performance. Radixpert was rated higher than baseline models in terms of clinical accuracy, completeness, and appropriateness, with strong inter-rater reliability. The evaluators noted that the model produced reports with better anatomical localization and fewer factual errors.

VI. CONCLUSION AND FUTURE WORK

In this work, we introduced Radixpert, a highly efficient and accurate framework for automated radiology report generation. The success of our model is built on the synthesis of two key innovations: a multi-stage, parameter-efficient fine-tuning strategy (MSA-LoRA) that effectively adapts a general foundation model to diverse medical datasets, and a hierarchical cross-modal fusion mechanism (HCF) that balances the integration of fine-grained visual details with coherent clinical narrative. By achieving state-of-the-art performance while training only 1.5% of the model’s total parameters, Radixpert presents a practical and powerful blueprint for developing and deploying specialized medical AI systems from general-purpose foundation models.

Looking ahead, future work will focus on addressing the specific limitations identified during our analysis. To improve the model’s spatial and numerical reasoning and reduce errors like laterality confusion, we plan to incorporate explicit spatial encoding modules and explore graph-based representations to better model object relations and counts. To enhance the detection of subtle, low-contrast pathologies, we will investigate the use of more advanced, higher-resolution vision encoders and targeted data augmentation techniques. Finally, to mitigate the rare instances of factual hallucination, we will work on developing a grounded generation mechanism that forces the

model to cite visual evidence for its claims, thereby improving the factual reliability and trustworthiness of the generated reports.

REFERENCES

- [1] A. Bohr and K. Memarzadeh, “The Rise of Vision-and-Language Models in Medical AI: A Survey,” *Journal of Medical Artificial Intelligence*, vol. 9, pp. 23–41, 2024.
- [2] M. Moor, W. Rieger, F. Gaertner, et al., “Med-Flamingo: A Multimodal Medical Few-Shot Learner,” *arXiv preprint arXiv:2306.05424*, 2023.
- [3] D. Dalla Serra, R. Tschandl, M. Burgery, E. Maier, and C. Quanz, “Controllable Radiology Report Generation via Prompting Large Language Models,” *Proceedings of the Conference on Health, Inference, and Learning (CHIL)*, 2023.
- [4] H. Amjad, S. B. Dobson, and G. Z. Ma, “Low-Rank Adaptation for Efficient Fine-Tuning of Foundation Models in Medical Imaging,” *Medical Image Analysis*, vol. 81, p. 102557, 2022.
- [5] S. Pelka, J. M. Koitka, A. Korsukewitz, N. Sentker, M. Gerlach, J. A. Jungmann, and O. R. König, “Radiology Objects in Context (ROCO): A Multimodal Dataset for Medical Image Understanding,” *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 180–189, 2019.
- [6] X. Zhou, X. Yang, O. Banerjee, J.N. Acosta, J. Miller, O. Huang, and P. Rajpurkar, “Benchmarking Radiology Report Generation from Noisy Free-Texts,” *arXiv preprint arXiv:2402.19437*, 2024.
- [7] S. Jain, A.P. Irvin, J.E. Reed, L. Zhong, J. Dunmon, K. Shin, et al., “RadGraph: Extracting Clinical Entities and Relations from Radiology Reports,” *arXiv preprint arXiv:2106.14463*, 2021.
- [8] A. Johnson, R. Weiss, G. Korotkevich, et al., “RaTEScore: A Reliable Metric for Factual Clinical Report Generation,” in *Proceedings of Medical Imaging with Deep Learning*, 2023.
- [9] C. Pellegrini, E. Özsoy, B. Busam, B. Wiestler, N. Navab, and M. Keicher, “RadDialog: Large Vision-Language Models for X-Ray Reporting and Dialog-Driven Assistance,” *Medical Imaging with Deep Learning*, 2025.
- [10] Meta AI, “Llama 3.2-11B-Vision-Instruct,” *Hugging Face Model Hub*, 2024. [Online]. Available: <https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct>
- [11] C. Wu, X. Zhang, Y. Wang, and W. Xie, “Benchmarking and Boosting Radiology Report Generation for 3D High-Resolution Medical Images,” *arXiv preprint arXiv:2406.07146*, 2024.
- [12] A. Bustos, L. Pertusa, J. Salinas, and A. de la Iglesia-Vayá, “PadChest: A large chest x-ray image dataset with multi-label annotated reports,” *Medical Image Analysis*, vol. 66, p. 101797, 2020.
- [13] X. Huang, D. F. Glymour, T. J. L. Ribeiro, et al., “GLoRIA: A Multimodal Global-Local Representation Learning Framework for Disease Diagnosis from Medical Imaging,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13533–13543, 2021.
- [14] B. Böcking, N. T. K. Lale, A. K. D. R. Bannur, et al., “Making the Most of Text Semantics to Improve Biomedical Vision–Language Processing,” *arXiv preprint arXiv:2204.09817*, 2022.
- [15] K. Tu, T. G. K. Rao, A. G. M. Singhal, et al., “Towards Expert-Level Medical Question Answering with Large Language Models,” *arXiv preprint arXiv:2403.05530*, 2024.
- [16] Z. Wang, L. Liu, L. Wang, and L. Zhou, “R2GenGPT: Radiology report generation with frozen LLMs,” *Meta-Radiology*, vol. 1, no. 1, p. 100033, 2023.
- [17] Y. Shen, Z. Zhang, L. Zhou, et al., “Large Language Models in Medical Vision-Language Pre-training,” *arXiv preprint arXiv:2311.16334*, 2023.
- [18] Y. Zhang, H. Li, H. Cai, et al., “Text-supervised Vision-Language Pre-training for Medical Images and Reports,” in *Proceedings of Medical Imaging with Deep Learning*, 2023.
- [19] Z. Chen, H. Wang, J. Liu, and L. Shen, “Grounding Medical Vision-Language Models in Visual Evidence for Accurate and Interpretable Report Generation,” *Nature Machine Intelligence*, vol. 6, pp. 45–58, 2024.
- [20] J. Liu, Y. Cheng, and Z. Chen, “Exploring Knowledge-Grounded Radiology Report Generation,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 1, pp. 889–897, 2023.
- [21] X. Li, Y. Wang, W. Xie, “Uni3D-RRG: A Unified Framework for 3D Radiology Report Generation,” *IEEE Transactions on Medical Imaging*, vol. 43, no. 5, pp. 1823–1834, 2024.

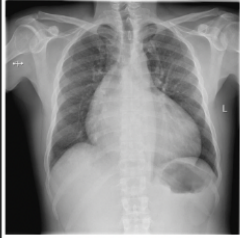

Scan	Ground Truth	Predicted Findings
	Chest X-ray showing enlarged cardiac silhouette with cardiothoracic ratio of 70%, and mild pulmonary congestion.	The cardiac silhouette is enlarged with an estimated cardiothoracic ratio of 68%. Findings are consistent with mild pulmonary congestion.
	A CT scan of the chest. The scan shows a right upper lobe cavitary nodule (white arrow) with left lung ground-glass nodules and bilateral pleural effusion.	Axial CT image shows a cavitary nodule in the left upper lobe. There are scattered ground-glass nodules in the right lung. Bilateral pleural effusions are present.

Fig. 4. Qualitative analysis of Radixpert’s performance on two distinct cases. (Top) For a chest X-ray, the model correctly identifies cardiomegaly and pulmonary congestion but makes a minor error in the numerical estimation of the cardiothoracic ratio. (Bottom) For a CT scan, the model correctly identifies all pathological findings (cavitary nodule, ground-glass nodules, pleural effusion) but demonstrates laterality confusion, swapping the locations of the nodule and the ground-glass opacities. These examples highlight the model’s overall accuracy while illustrating specific limitations discussed in the text.

- [22] F. Yu, S. A. Johnson, and P. Rajpurkar, “Evaluating the Evaluators: On the Critical Assessment of Metrics for Medical Report Generation,” *Journal of the American Medical Informatics Association*, 2024.