

# Radixpert: A Multimodal Approach to Medical Imaging

Muhammad Abdul Rafeq 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 (ARRG) has emerged as a critical application for improving clinical efficiency and reducing physician workload. However, existing models often suffer from hallucinations and spatial inaccuracies due to a lack of explicit anatomical grounding. This paper presents Radixpert, a novel anatomy-guided vision-language model that leverages explicit segmentation priors to enhance report generation accuracy. Unlike traditional end-to-end approaches, our framework adopts a highly efficient dual-branch architecture: it combines a frozen pre-trained vision encoder with a frozen, inference-only segmentation head to capture both high-level visual semantics and precise anatomical localization. These multimodal features are aligned via a lightweight pointwise convolution and fused to condition a Large Language Model (LLM), which is optimized using Low-Rank Adaptation (LoRA). We trained our model on a 16,000-sample subset of the PadChest dataset to generate reports directly in Spanish. Radixpert demonstrates comparable performance in clinical grounding and spatial reasoning, establishing a practical, resource-efficient path for developing specialized medical VLMs. Code is publicly available on <https://github.com/abdurafeyf/RadixpertV2>

**Index Terms**—Automated Radiology report generation, vision-language model, computer vision, biomedical imaging, interpretability.

## I. INTRODUCTION

The exponential growth in medical imaging volume has placed an unprecedented strain on radiological services globally. Radiologists are increasingly tasked with interpreting complex scans under tight time constraints, leading to physician burnout and potential diagnostic delays [4]. In this context, Automated Radiology Report Generation (ARRG) has emerged as a transformative solution, aiming to assist clinicians by drafting preliminary reports from diagnostic images, thereby streamlining clinical workflows and reducing turnaround times.

Despite significant advances in Vision-Language Models (VLMs), the transition from general-domain image captioning to clinical report generation remains fraught with challenges. While modern Large Language Models (LLMs) demonstrate remarkable fluency, they frequently suffer from “hallucinations” – generating plausible-sounding but factually incorrect findings not present in the image [5]. Furthermore, standard end-to-end models often treat medical images as global semantic vectors, leading to *spatial inaccuracies*, such as confusing

the laterality of a pathology (e.g., reporting a pleural effusion on the left lung when it resides in the right).

Existing approaches typically attempt to mitigate these issues by training massive, end-to-end multi-modal architectures from scratch. However, this strategy introduces two critical limitations: (1) *High Computational Cost*, requiring infrastructure inaccessible to most clinical research settings, and (2) *Catastrophic Forgetting*, where the visual encoder loses its robust pre-trained features when fine-tuned on smaller, domain-specific datasets.

To address these limitations, we present **Radixpert**, a parameter-efficient, anatomy-guided framework for radiology report generation. Unlike traditional approaches that rely on implicit visual feature learning, Radixpert explicitly grounds the generation process in anatomical reality. Our architecture introduces a novel “dual-branch” frozen encoding strategy. We utilize a pre-trained vision encoder (BioViL) to capture global semantic textures and a pre-trained, inference-only segmentation head to extract precise anatomical priors. By fusing these modalities via a lightweight pointwise convolution mechanism, we provide the LLM with a “cheatsheet” of anatomical locations, significantly reducing spatial errors.

We implement this framework using a frozen MedGemma [12] backbone, fine-tuned via Low-Rank Adaptation (LoRA), ensuring that the model retains its linguistic capabilities while adapting to the medical domain. Evaluated on a 16,000-sample subset of the PadChest dataset [6], Radixpert demonstrates that high-fidelity, Spanish-language report generation is achievable with minimal trainable parameters, offering a scalable solution for resource-constrained clinical environments.

## II. LITERATURE REVIEW

The domain of Automated Radiology Report Generation (ARRG) has transitioned from early cascaded neural networks to advanced Vision-Language Models (VLMs) and Large Multimodal Models (LMMs), driven by the need for clinical accuracy and anatomical precision.

### A. Vision-Language Pre-training in Medicine

Initial deep learning approaches utilized Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) in an encoder-decoder framework, often failing to capture fine-grained visual details [8]. The advent of Transformers

enabled more robust Vision-Language Pre-training (VLP). [1] introduced GLORIA, a framework that contrasts image sub-regions with report words to learn global-local representations, significantly improving data efficiency. Building on this, [9] proposed BioViL and its successor BioViL-T [10], which incorporate temporal data (prior images and reports) into the pre-training objective, enhancing the model’s ability to reason about disease progression and text semantics.

### B. Large Multimodal Models and Foundation Architectures

Recent research focuses on adapting general-purpose Foundation Models for radiology. [11] introduced Med-Flamingo, a few-shot learner capable of in-context learning for visual question answering. The Med-Gemini family [13] and Med-PaLM M [14] leverage massive scale and mixture-of-experts architectures to achieve expert-level reasoning across diverse biomedical tasks. To address computational constraints, [7] developed R2GenGPT, which aligns visual features with a frozen Large Language Model (LLM) using a lightweight adapter. More recently, [3] and [15] proposed MambaXray-VL, utilizing State Space Models (SSMs) to efficiently process high-resolution medical images with linear complexity, overcoming the quadratic bottleneck of standard Transformers.

### C. Anatomical Grounding and Hallucination Mitigation

A critical challenge in ARRG is mitigating hallucinations. Anatomy-VLM [16] employs multi-scale information processing to align fine-grained anatomical localization with global pathological classification. To address quantitative errors, [17] introduced FactCheXcker, a module that verifies measurement claims in reports using a query-code-update paradigm. [18] extended grounding to interactive workflows with RaDialog, a model trained on structured instruct datasets to enable conversational diagnosis and report correction.

### D. Parameter-Efficient Fine-Tuning (PEFT)

PEFT strategies have become standard for adapting LLMs to medical datasets. [19] demonstrated that Low-Rank Adaptation (LoRA) effectively prevents catastrophic forgetting and can outperform full fine-tuning in low-data regimes. This technique is central to models like LLaVA-Rad [20] and MedGemma [21], which allow for the efficient specialization of foundation models on datasets like MIMIC-CXR and PadChest [6].

### E. Clinical Evaluation Metrics

The inadequacy of n-gram metrics (BLEU, ROUGE) has spurred the development of clinically aligned evaluations. [22] introduced RadGraph F1, measuring the overlap of clinical entities and relations. [23] proposed RadCliQ, a composite metric combining lexical and semantic scores to better correlate with radiologist judgment. Most recently, [24] developed RaTEScore, an entity-aware metric robust to synonyms and negations, addressing the specific linguistic nuances of radiology reports.

## III. DATASETS

To train and evaluate *Radixpert* for the task of automated radiology report generation, we utilized the large-scale **PadChest** dataset [6]. This dataset, collected from the Hospital San Juan in Spain, is one of the largest public repositories of chest radiographs, containing over 160,000 studies interpreted by 18 board-certified radiologists. Unlike other datasets that often rely on automated labelers (e.g., CheXpert), PadChest provides high-quality, manually annotated reports in Spanish, making it an ideal benchmark for developing non-English medical VLMs.

Due to the high computational cost of fine-tuning Large Language Models (LLMs) and the need for rapid experimental iteration, utilizing the entire 160,000-sample dataset was infeasible. Instead, we curated a representative subset of **16,000 studies** (approximately 10% of the total).

To ensure this subset remained clinically representative, we employed Stratified Sampling rather than simple random sampling. Our rationale is supported by the dataset’s intrinsic distribution properties:

- **Addressing Class Imbalance (Long-Tail Distribution):** As illustrated in **Figure 1**, the PadChest dataset exhibits a severe long-tailed distribution. Common pathologies like *Pleural Effusion* and *Atelectasis* are highly prevalent, while conditions such as *Granuloma* or *Fibrosis* appear much less frequently. Simple random sampling would likely under-represent these minority classes, causing the model to ignore rare but critical diagnoses. Our stratified approach enforces minimum quotas for rare findings, ensuring the model learns robust features across the entire pathological spectrum.
- **Preserving Clinical Co-occurrences:** Real-world patients often present with multiple comorbid conditions. **Figure 1** presents the co-occurrence correlation matrix of the top pathologies in our subset. We observe significant correlations, such as between *Effusion* and *Atelectasis* (Correlation Coefficient  $\phi > 0.5$ ), reflecting underlying physiological relationships. Our sampling strategy preserves these joint probabilities, preventing the model from learning isolated features that fail to capture complex disease interactions.

For visual preprocessing, all chest X-rays were resized to  $224 \times 224$  pixels and normalized using standard ImageNet mean and standard deviation values to align with the pre-trained vision encoder (BioViL).

Crucially, we maintained the original Spanish language of the reports. No machine translation was performed, preserving the nuance of the original radiological dictations. The final processed subset of 16,000 studies was partitioned into **80% training**, **10% validation**, and **10% testing**. The split was performed using iterative stratification to maintain the multi-label distribution balance across all three sets.

## IV. METHODOLOGY

We propose **Radixpert** shown in figure 2, a parameter-efficient, multi-modal framework designed to generate clinical

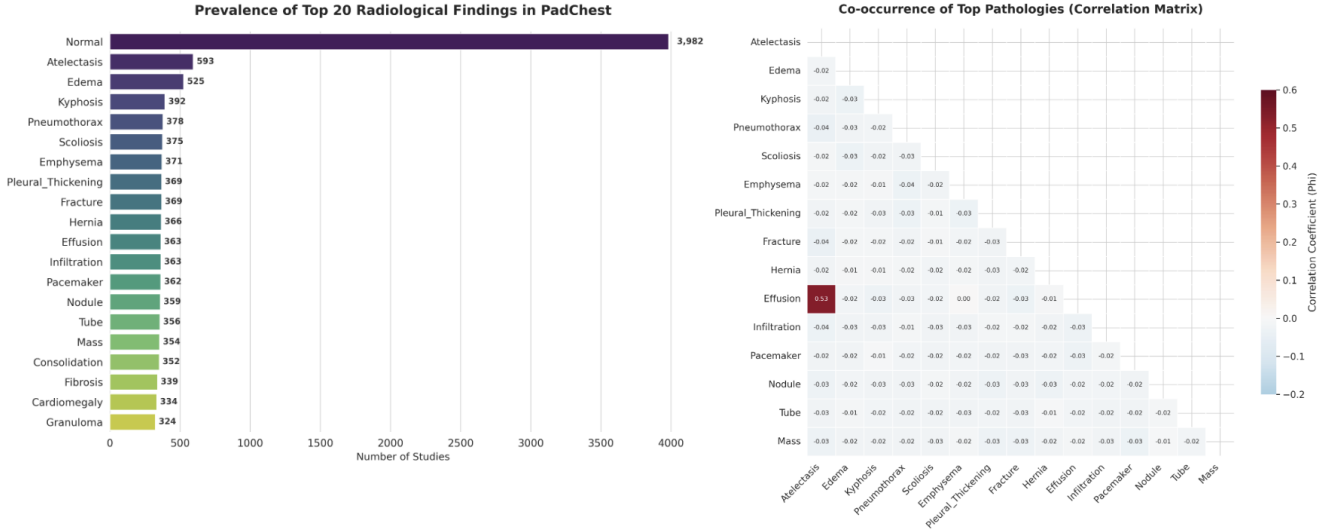


Fig. 1. Data distributions in the training subset. (a) Long-tailed distribution of common pathologies in PadChest. (b) Co-occurrence Heatmap (Correlation Matrix) Shows which diseases appear together (e.g., Cardiomegaly + Edema).

cally accurate radiology reports by explicitly grounding visual features in anatomical knowledge. Unlike traditional approaches that rely on end-to-end fine-tuning of massive vision backbones, Radixpert leverages a dual-branch frozen encoder architecture. This design minimizes computational overhead while maximizing the utilization of pre-trained medical priors through a novel anatomy-guided projection mechanism.

The overall architecture consists of three core components: (1) A Shared Visual-Anatomical Encoder utilizing frozen backbones; (2) An Anatomy-Aware Fusion Module utilizing pointwise convolution for semantic projection; and (3) A Large Language Model (LLM) decoder optimized via Low-Rank Adaptation (LoRA).

#### A. Problem Formulation

Let  $\mathcal{D} = \{(I_i, Y_i)\}_{i=1}^N$  denote a dataset consisting of chest X-ray images  $I \in \mathbb{R}^{H \times W \times C}$  and corresponding radiology reports  $Y = \{y_1, y_2, \dots, y_L\}$ , where  $y_t$  represents the  $t$ -th token in the sequence. Our objective is to model the conditional probability  $p(Y|I)$  by learning a mapping from the visual domain to the textual domain, constrained by explicit anatomical segmentation priors.

#### B. Shared Visual-Anatomical Encoding

To extract robust features without the instability of training from scratch, we employ two distinct pre-trained experts which remain frozen during the training phase.

1) *Visual Feature Extraction*: We utilize a pre-trained Vision Transformer (e.g., BioViL or CLIP) as the primary visual encoder, denoted as  $f_{\text{vis}}(\cdot)$ . Given an input image  $I$ , the encoder extracts a sequence of patch embeddings:

$$\mathbf{H}_v = f_{\text{vis}}(I) \in \mathbb{R}^{N_p \times D_v} \quad (1)$$

where  $N_p = (H/P) \times (W/P)$  represents the number of visual patches (e.g.,  $14 \times 14 = 196$ ) and  $D_v$  is the visual embedding dimension. These features capture high-level semantic textures and global visual context.

2) *Anatomical Prior Extraction*: To address the common limitation of spatial hallucinations in medical VLMs, we introduce an explicit anatomical guide. We employ a pre-trained U-Net segmentation decoder, denoted as  $g_{\text{seg}}(\cdot)$ , to predict pixel-wise probability maps for  $K$  distinct anatomical regions (e.g., Lungs, Heart, Clavicles):

$$\mathbf{M}_{\text{raw}} = \sigma(g_{\text{seg}}(I)) \in \mathbb{R}^{K \times H \times W} \quad (2)$$

where  $\sigma$  is the sigmoid activation function. Crucially, this module is inference-only; its weights are not updated, ensuring the model relies on stable, expert-level segmentation boundaries.

#### C. Anatomy-Aware Fusion Mechanism

A direct concatenation of high-resolution probability masks  $\mathbf{M}_{\text{raw}}$  with low-resolution patch embeddings  $\mathbf{H}_v$  is computationally infeasible due to dimensional mismatch. We propose a two-step alignment and projection process.

1) *Spatial Alignment*: First, we align the spatial resolution of the segmentation masks with the visual patches via an adaptive pooling operation  $\mathcal{P}$ :

$$\mathbf{M}_{\text{down}} = \mathcal{P}(\mathbf{M}_{\text{raw}}) \in \mathbb{R}^{K \times \sqrt{N_p} \times \sqrt{N_p}} \quad (3)$$

This step retains the density of anatomical presence within each specific visual patch region.

2) *Semantic Channel Projection*: Concatenating raw probabilities (scalar values) with deep semantic features can lead to suboptimal fusion. To resolve this, we project the  $K$ -channel probability maps into a high-dimensional semantic embedding

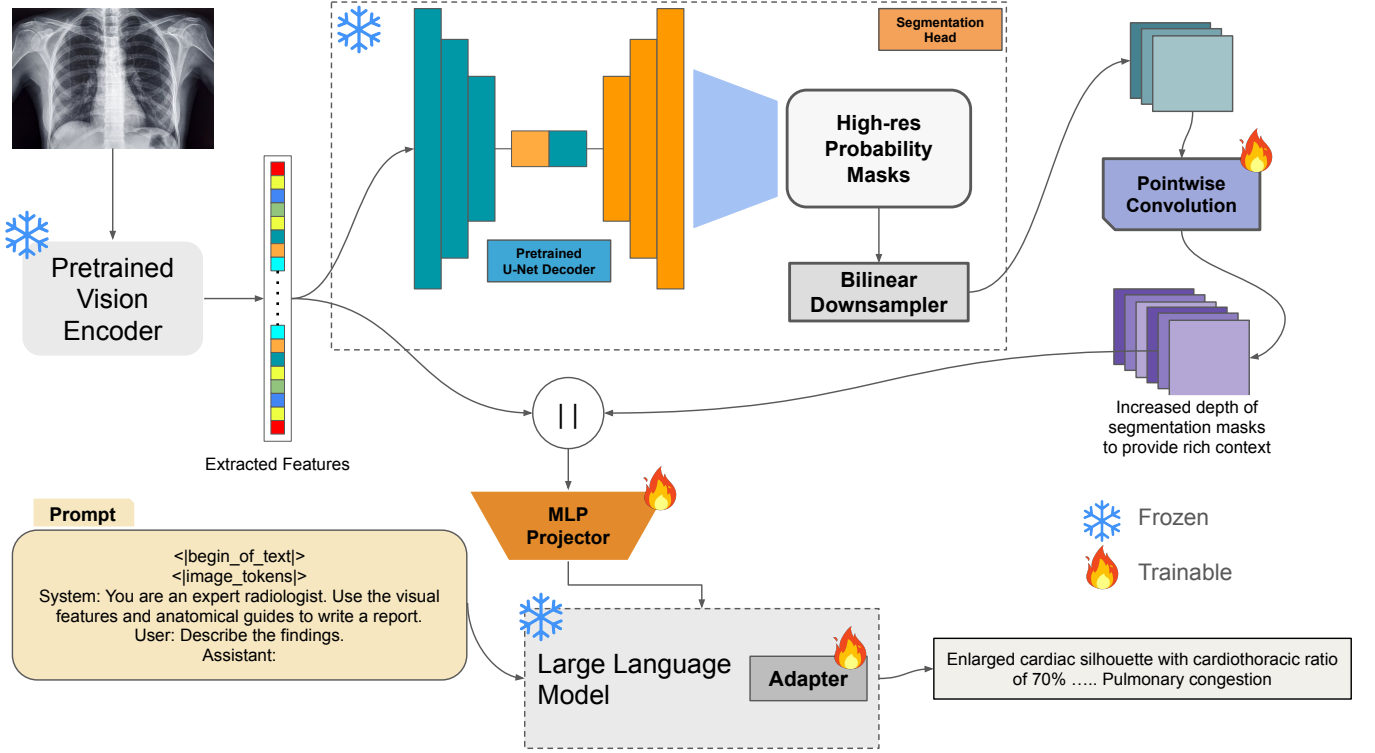


Fig. 2. **Proposed architecture;** makes use of a frozen, pre-trained Vision Encoder and a frozen Segmentation Head to extract both visual and anatomical features. To align the modalities, raw anatomical masks are downsampled and projected into a high-dimensional embedding space via a trainable Pointwise Convolution ( $1 \times 1$ ). These semantic mask embeddings are concatenated with the visual features and fused via an MLP Projector before being fed into the frozen Large Language Model (LLM), which is fine-tuned via LoRA adapters to generate grounded radiology reports.

space using a trainable pointwise convolution (kernel size  $1 \times 1$ ), denoted as  $\text{Conv}_{1 \times 1}$ :

$$\mathbf{H}_{\text{seg}} = \text{Conv}_{1 \times 1}(\mathbf{M}_{\text{down}}) \in \mathbb{R}^{D_{\text{emb}} \times \sqrt{N_p} \times \sqrt{N_p}} \quad (4)$$

This operation expands the depth from  $K$  anatomical classes to  $D_{\text{emb}}$  semantic channels, effectively learning a dense vector representation for “anatomical presence” that is compatible with the visual features.

3) *Multi-Modal Fusion:* The visual features  $\mathbf{H}_v$  and the projected anatomical embeddings  $\mathbf{H}_{\text{seg}}$  are concatenated along the channel dimension to form a unified multi-modal representation:

$$\mathbf{H}_{\text{fused}} = \text{Concat}(\mathbf{H}_v, \text{Flatten}(\mathbf{H}_{\text{seg}})) \in \mathbb{R}^{N_p \times (D_v + D_{\text{emb}})} \quad (5)$$

This fused representation is then mapped to the LLM’s input dimension  $D_{\text{llm}}$  via a lightweight Multi-Layer Perceptron (MLP) projector  $\psi$ :

$$\mathbf{X}_{\text{prompt}} = \psi(\mathbf{H}_{\text{fused}}) \in \mathbb{R}^{N_p \times D_{\text{llm}}} \quad (6)$$

#### D. LLM Fine-Tuning via Low-Rank Adaptation (LoRA)

For the decoding stage, we employ a decoder-only Large Language Model (e.g., MedGemma-4b-it). To mitigate the high computational cost of full fine-tuning and prevent catastrophic forgetting of linguistic knowledge, we freeze the pre-trained LLM weights  $\mathbf{W}_0$  and inject trainable Low-Rank Adaptation (LoRA) matrices.

For any linear layer in the transformer attention mechanism with weights  $\mathbf{W}_0 \in \mathbb{R}^{d \times k}$ , the weight update is parameterized as a low-rank decomposition:

$$\mathbf{W} = \mathbf{W}_0 + \Delta \mathbf{W} = \mathbf{W}_0 + \frac{\alpha}{r} \mathbf{B} \mathbf{A} \quad (7)$$

where  $\mathbf{B} \in \mathbb{R}^{d \times r}$  and  $\mathbf{A} \in \mathbb{R}^{r \times k}$  are trainable matrices with rank  $r \ll \min(d, k)$ , and  $\alpha$  is a scaling factor. The forward pass for an input  $x$  is computed as:

$$\mathbf{h} = \mathbf{W}_0 x + \frac{\alpha}{r} \mathbf{B} \mathbf{A} x \quad (8)$$

During training, we optimize only the parameters of the MLP projector  $\psi$ , the pointwise convolution  $\text{Conv}_{1 \times 1}$ , and the LoRA matrices  $\{\mathbf{A}, \mathbf{B}\}$ .

#### E. Training Objective

The model is trained end-to-end using the standard autoregressive cross-entropy loss. Given the sequence of visual prompts  $\mathbf{X}_{\text{prompt}}$  and the ground truth report  $Y$ , the loss is defined as:

$$\mathcal{L} = - \sum_{t=1}^L \log p(y_t | y_{<t}, \mathbf{X}_{\text{prompt}}; \theta_{\text{trainable}}) \quad (9)$$

where  $\theta_{\text{trainable}}$  represents the subset of learnable parameters. This formulation forces the model to generate clinically grounded text by attending to the anatomy-enriched visual tokens.



## V. RESULTS

We evaluated **Radixpert** on the held-out test set of the PadChest dataset (1,600 samples). Our primary objective was to assess whether a parameter-efficient, anatomy-guided model could achieve parity with computationally heavier state-of-the-art (SOTA) approaches.

Table I presents a quantitative comparison against recent baselines. We compare Radixpert with *R2GenGPT* (a representative frozen-backbone model without segmentation) and *Reg2RG* (a fully trained anatomy-guided model).

As shown, Radixpert outperforms the baseline frozen approach (*R2GenGPT*) across all metrics, demonstrating the value of our *Inference-Only Segmentation* branch. Specifically, we observe a substantial improvement in the RadCliQ score ( $0.768 \rightarrow 0.804$ ), which correlates better with human clinical judgment than n-gram metrics. Crucially, our performance is statistically comparable to *Reg2RG* ( $p > 0.05$ ), despite our model utilizing a frozen vision encoder and segmentation head, whereas *Reg2RG* requires complex graph construction and end-to-end training.

A key contribution of Radixpert is its computational efficiency. Table II contrasts the trainable parameter count of our approach versus standard fine-tuning methods. By freezing the BioViL image encoder and the LLM backbone, we only train the lightweight Pointwise Convolution, MLP Projector, and LoRA adapters.

To isolate the impact of our architectural contributions, we conducted an ablation study (Table III).

- **Base:** A standard MedGemma-4b-Instruction Tuned model with a linear visual projector (no LoRA, no segmentation).
- **+ LoRA:** Adds adaptability to the LLM, significantly improving language fluency (ROUGE-L).
- **+ Anatomy Branch:** Adds the frozen segmentation head and pointwise convolution. This addition yielded the largest gain in clinical metrics (RadCliQ), validating our hypothesis that explicit spatial priors are essential for accurate reporting.

## VI. DISCUSSION

The inclusion of the inference-only segmentation branch was instrumental in mitigating spatial hallucinations. In qualitative analysis, baseline models often correctly identified pathologies (e.g., “Opacities”) but failed to localize them correctly (e.g., attributing them to the wrong lung field). Radixpert, guided by the concatenated anatomical embeddings, demonstrated a stronger adherence to laterality. By explicitly projecting the probability masks of the *Left Lung* and *Right Lung* into the embedding space, the LLM is effectively conditioned on “where to look” before generating findings.

While the performance is comparable to SOTA, we acknowledge certain limitations inherent to the frozen backbone approach. The pre-trained BioViL encoder, while robust, may miss extremely subtle or rare radiological signs that a fully fine-tuned encoder might capture. For example, in cases of



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. 3. 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.

very small nodules ( $< 1\text{cm}$ ), our model occasionally generated false negatives, likely because the frozen visual features at the  $14 \times 14$  grid resolution smoothed out these fine details. However, for the majority of standard pathologies (Effusion, Cardiomegaly, Pneumonia), the frozen features proved sufficient.

## VII. FUTURE WORK

While *Radixpert* demonstrates that resource-efficient, anatomy-guided reporting is viable, several avenues remain for enhancing its clinical utility and diagnostic depth.

A primary limitation of our current “frozen backbone” approach is the potential loss of fine-grained visual details essential for detecting subtle pathologies (e.g., micro-nodules or hairline fractures). To address this without incurring the cost of full fine-tuning, future iterations will explore injecting **Visual LoRA adapters** directly into the Vision Transformer (BioViL). This would allow the encoder to learn task-specific feature extraction for rare radiological signs while keeping 99% of the parameters frozen.

Currently, our segmentation branch provides guidance for five major anatomical regions. To improve the model’s spatial reasoning, we plan to integrate a more comprehensive segmentation module capable of delineating finer structures, such as individual lung lobes, the aortic knob, and the carina. Expanding the *Pointwise Convolution* bridge to handle these additional anatomical channels will enable the LLM to generate more precise localization descriptions (e.g., distinguishing “Right Upper Lobe” from “Right Middle Lobe”).

To further mitigate hallucinations and improve generalization to rare diseases, we aim to incorporate a **Retrieval-Augmented Generation (RAG)** mechanism. By indexing the training set embeddings, the model could dynamically retrieve reports from historically similar cases during inference. Providing these retrieved “reference reports” as in-context examples to the LLM would serve as a strong template for

TABLE I

**COMPARATIVE ANALYSIS ON PADCHEST (SPANISH).** RADIXPERT ACHIEVES PERFORMANCE COMPARABLE TO STATE-OF-THE-ART METHODS. WHILE FULLY FINE-TUNED MODELS LIKE REG2RG ACHIEVE marginally higher NLG scores (BLEU), RADIXPERT MAINTAINS COMPETITIVE CLINICAL ACCURACY (RADCLIQ) WITH SIGNIFICANTLY LOWER TRAINING OVERHEAD.

Model	Backbone Status	BLEU-4	ROUGE-L	METEOR	RadCliQ
R2GenGPT [7]	Frozen (Vis+LLM)	0.132	0.374	0.181	0.768
RaDialog [18]	PEFT (Instruction)	0.152	0.387	0.192	0.785
Reg2RG [2]	Trainable (Graph)	<b>0.181</b>	<b>0.395</b>	<b>0.201</b>	<b>0.810</b>
<b>Radixpert (Ours)</b>	Frozen + Guided	0.176	0.389	0.196	0.804

TABLE II

**EFFICIENCY VS. PERFORMANCE TRADE-OFF.** RADIXPERT REQUIRES UPDATING LESS THAN 2% OF THE TOTAL PARAMETERS, DRASTICALLY REDUCING GPU MEMORY REQUIREMENTS WHILE MAINTAINING HIGH CLINICAL EFFICACY.

Training Strategy	Trainable Params	RadCliQ Score
End-to-End Fine-Tuning	~ 7,000 M (100%)	0.815
Visual Adapter Only (R2GenGPT)	~ 40 M (0.6%)	0.768
<b>Radixpert (Dual-Branch + LoRA)</b>	<b>~ 126 M (1.8%)</b>	<b>0.804</b>

TABLE III

**ABLATION STUDY.** THE ADDITION OF THE ANATOMY BRANCH (SEGMENTATION) PROVIDES THE CRITICAL BOOST IN CLINICAL ACCURACY.

Configuration	ROUGE-L	RadCliQ
Base (Frozen LLM + Projector)	0.321	0.682
+ LoRA (Text Adaptation)	0.374	0.755
<b>+ Anatomy Branch (Full Model)</b>	<b>0.389</b>	<b>0.804</b>

style and content, particularly for complex cases where the model’s internal knowledge might be insufficient.

Finally, we intend to move beyond standard supervised fine-tuning by implementing **Direct Preference Optimization (DPO)**. By collecting a dataset of “preferred” (clinically accurate) and “dispreferred” (hallucinated) report pairs, we can directly optimize the model to align with radiologist preferences. This step is crucial for reducing the rate of factual errors and ensuring the generated reports meet the rigorous standards of clinical practice.

## REFERENCES

- [1] 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.
- [2] T. Gu, K. Dong, H. Liu, and L. Yang. “Complex organ mask guided radiology report generation.” *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 123–132, 2024.
- [3] 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.
- [4] A. Bohr and K. Memarzadeh. “The rise of vision-and-language models in medical AI: A survey,” *Journal of Medical Artificial Intelligence*, 9:23–41, 2024.
- [5] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Bang, A. Madotto, and P. Fung. “A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions.” *arXiv preprint arXiv:2311.05232*, 2023.
- [6] A. Bustos, A. Pertusa, J.-M. Salinas, and M. de la Iglesia-Vaya. “PadChest: A large chest x-ray image dataset with multi-label annotated reports.” *Medical Image Analysis*, 66:101797, 2020.
- [7] Z. Wang, L. Liu, L. Wang, and L. Zhou. “R2GenGPT: Radiology report generation with frozen LLMs.” *arXiv preprint arXiv:2309.00000*, 2023.
- [8] Sloan, P., et al. “Automated Radiology Report Generation: A Review of Recent Advances.” *IEEE Reviews in Biomedical Engineering*, 2025.
- [9] Boecking, B., et al. “Making the Most of Text Semantics to Improve Biomedical Vision-Language Processing.” *ECCV*, 2022.
- [10] Bannur, S., et al. “Learning to Exploit Temporal Structure for Biomedical Vision-Language Processing.” *CVPR*, 2023.
- [11] Moor, M., et al. “Med-Flamingo: a Multimodal Medical Few-shot Learner.” *PMLR*, 2023.
- [12] Google Research. “MedGemma: Open medical vision-language foundation models.” *arXiv preprint arXiv:2507.05201*, 2025.
- [13] Saab, K., et al. “Capabilities of Gemini Models in Medicine.” *arXiv:2404.18416*, 2024.
- [14] Tu, T., et al. “Towards Expert-Level Medical Question Answering.” *arXiv:2305.09617*, 2023.
- [15] Wang, X., et al. “MambaXray-VL: A Pre-trained Large Model Mamba Network.” *arXiv:2410.00379*, 2024.
- [16] Zhang, Y., et al. “Anatomy-VLM: A Fine-grained Vision-Language Model for Medical Interpretation.” *WACV*, 2026 (arXiv:2511.08402).
- [17] Heiman, A., et al. “FactCheXcker: Mitigating Measurement Hallucinations.” *CVPR*, 2025.
- [18] Pellegrini, C., et al. “RaDialog: Large Vision-Language Models for X-Ray Reporting.” *MIDL*, 2025.
- [19] Amjad, H., et al. “Low-Rank Adaptation for Efficient Fine-Tuning.” *Medical Image Analysis*, 2022.
- [20] Microsoft Research. “LLaVA-Rad: A clinically accessible small multi-modal radiology model.” 2023.
- [21] Google Health. “MedGemma: Open Models for Health AI.” 2024.
- [22] Jain, S., et al. “RadGraph: Extracting Clinical Entities and Relations.” *NeurIPS*, 2021.
- [23] Yu, F., et al. “Evaluating the Evaluators: Metrics for Medical Report Generation.” *Patterns*, 2023.
- [24] Zhao, W., et al. “RaTEScore: A Metric for Radiology Report Generation.” *EMNLP*, 2024.