# Enhancing Radiology Report Generation: The Impact of Locally Grounded Vision and Language Training

Sergio Sanchez S, Muhammad Awais , Yi-Zhe Song, Josef Kittler

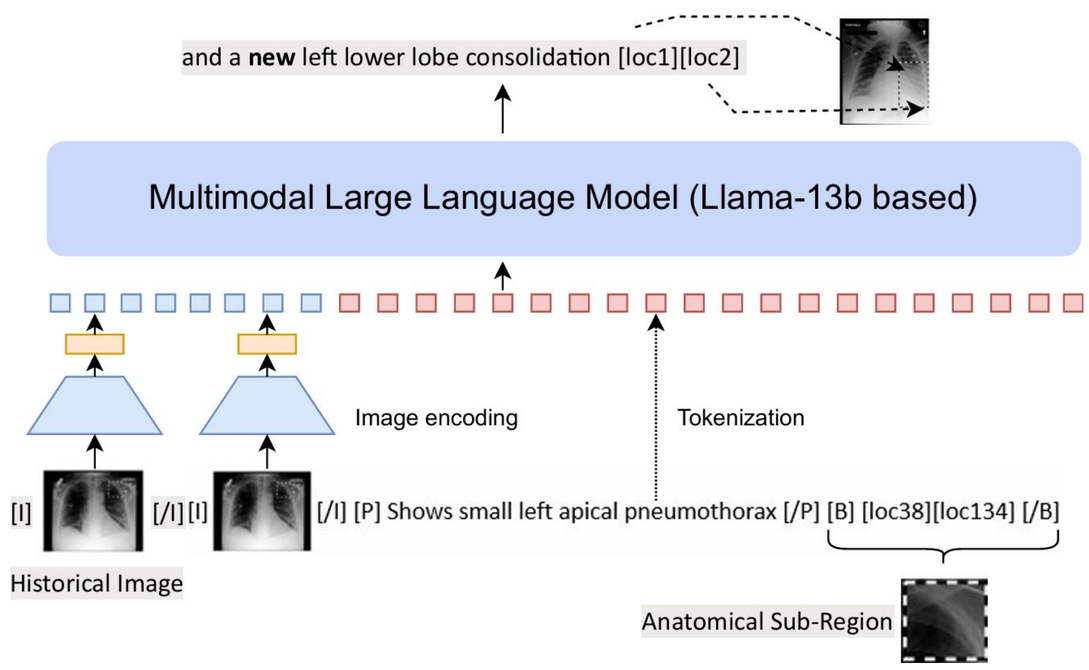
People-Centred AI
UNIVERSITY OF SURREY

BMVC 2024

## Abstract

We present a model for automated radiology report generation that enhances spatial understanding by introducing learnable location tokens. This model links x-ray images with report texts and annotated region anchors, marking key anatomical areas. A grounding prompt-tuning strategy directly connects report text to specific x-ray regions, improving accuracy and relevance in generated reports.

## Architecture




## Methodology

**Model Overview:**
The model integrates both current and historical X-ray images with annotated reports, using special tokens to identify the location of findings in images.

**Grounded Input Representations:**
The model uses bounding boxes on X-rays to link text descriptions to specific anatomical areas. Continuous bounding box coordinates are converted into discrete location tokens for precise spatial encoding.

**Architecture Components:**

- **Image Encoder:** A 3D Vision Transformer processes each image, representing both 2D and 3D X-ray inputs.
- **Perceiver Aggregation:** Aggregates visual embeddings using a perceiver module, normalizing embeddings for integration.
- **Multi-Modal Fusion:** Combines visual and text embeddings into a sequence, processed by a large language model to predict report text.
- **Fine-tuning Approach:**Uses LoRA (Low-Rank Adaptation) for efficient fine-tuning, adapting the large language model to radiology-specific data while preserving its general medical knowledge.

## Results

Table 1: Comparison of our proposed model with foundation model baselines, using results reported by RadFM on the same test set as RadFM for the tasks of RRG and Medical Visual Question Answering.

| Methods | Params | Textual metrics | | Semantic metrics | |
|---|---|---|---|---|---|
| | | Rouge | Bleu | UMLSp | UMLSr |
| *Radiology Report Generation* | | | | | |
| MedFlamingo [15] | 9B | 8.39 | 8.78 | 2.65 | 1.04 |
| MedVint [26] | 7B | 1.73 | 4.72 | 9.93 | 1.45 |
| RadFM [24] | 14B | 12.81 | 18.22 | 22.49 | 12.07 |
| Ours | 14B | **20.03** | **25.14** | **32.63** | **22.15** |
| *Improvement* | | +7.22 | +6.92 | +10.14 | +10.08 |
| *Medical Visual Question Answering* | | | | | |
| MedFlamingo [15] | 9B | 35.97 | 38.64 | 18.7 | 14.52 |
| RadFM [24] | 14B | 52.24 | 52.74 | 62.12 | 42.82 |
| Ours | 14B | **62.55** | **63.8** | **71.27** | **51.79** |
| *Improvement* | | +10.31 | +11.06 | +9.15 | +8.97 |

Table 2: Comparison of the proposed model with foundation model baselines on the same test set used in RadFM for the tasks Modality Recognition and Medical Disease Diagnosis.

| Methods | Params | Accuracy | F1 |
|---|---|---|---|
| *Modality Recognition* | | | |
| MedFlamingo [15] | 9B | 32.87 | - |
| MedVint [26] | 7B | 84.25 | - |
| RadFM [24] | 14B | **92.95** | - |
| Ours | 14B | **92.95** | - |
| *Disease Diagnosis* | | | |
| MedFlamingo [15] | 9B | 50.13 | 66.13 |
| MedVint [26] | 7B | 49.36 | 66.99 |
| RadFM [24] | 14B | **80.62** | **80.10** |
| Ours | 14B | 80.5 | 80 |

## Conclusion

1. Grounded Multimodal Large Language Models (MLLMs) show potential for enhancing radiology report generation (RRG).
2. Addressed the challenge of aligning text with specific image regions, leading to more interpretable and clinically useful reports.
3. This approach produces high-quality, grounded reports with versatility across different MLLMs.

RESEARCH POSTER PRESENTATION DESIGN © 2015
www.PosterPresentations.com