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

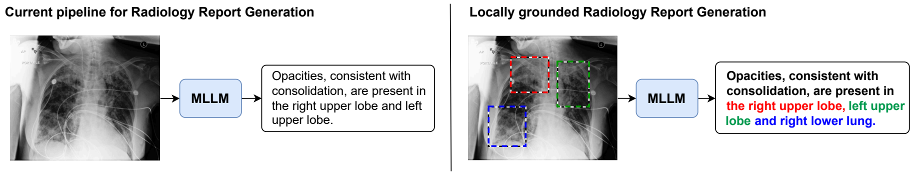
Sergio Sanchez Santiesteban[1]
s.sanchezsantiesteban@surrey.ac.uk
Muhammad Awais[1, 2]
m.a.rana@surrey.ac.u
Yi-Zhe Song[1, 2]
y.song@surrey.ac.uk
Josef Kittler[1]
j.kittler@surrey.ac.uk

[1] Centre for Vision, Speech and Signal Processing (CVSSP)
University of Surrey
Surrey GU2 7XH, UK

[2] Surrey Institute for People-Centred AI (PAI)
University of Surrey
Surrey GU2 7XH, UK

Abstract

In the medical domain, the integration of multimodal data—specifically radiology images paired with corresponding reports—presents a valuable opportunity for enhanced diagnostics. Recently, there has been growing interest in using Multimodal Large Language Models (MLLMs) for this purpose, due to their proficiency in learning effectively from the limited examples typical in specialized fields like radiology. Traditionally, radiologists generate reports by scrutinizing specific regions of an x-ray for changes, which are then systematically described with references to anatomical structures in the report's text. Existing methodologies, however, often process the radiographic image as a whole, which requires the fine-grained alignment to be learnt during the training phase through predominantly global optimization objectives. During pretraining this approach overlooks the subtleties of local image-to-text correspondences which results in automatically generated reports that are deficient in critical grounding elements, subsequently impeding the explanation of model predictions. In this paper, we introduce a novel dataset of interleaved radiology images with locally aligned phrase grounding annotations provided by radiologists. Drawing on grounding techniques employed in general-domain MLLMs, our methodology introduces learnable location tokens to enhance understanding of spatial relationships for model. We structure our training samples as sequences that encompass entire x-ray images, corresponding report texts, and region anchors. The region anchors are defined as sequences composed of the aforementioned location tokens to denote specific anatomical areas of interest. Combined with a grounding prompt-tuning strategy, this dataset fosters a direct connection between the radiology report's text and specific regions of the x-ray image. Our evaluation, conducted on large-scale public datasets, demonstrates that our proposed approach significantly refines the capabilities of existing MLLMs for radiology report generation.


Current pipeline for Radiology Report Generation
MLLM
Locally grounded Radiology Report Generation
MLLM

Opacities, consistent with consolidation, are present in the right upper lobe and left upper lobe.

Opacities, consistent with consolidation, are present in the right upper lobe, left upper lobe and right lower lung.

Figure 1: Comparison of Radiology Report Generation Pipelines: The left panel depicts the global alignment method, generating reports without localized detail in the X-ray image, while the right panel showcases the locally grounded approach, highlighting corresponding areas for specific findings in the report, enhancing detail and localization.

1 Introduction

Radiology Report Generation (RRG) is a critical task in the medical field, involving the creation of written reports based on diagnostic radiographic images, particularly X-ray images. These reports are essential for documenting patient care and assisting clinicians in understanding medical images to make informed decisions about the patient treatment. In recent years, pretrained Multimodal Large Language Models (MLLMs) have gained attention in the medical community [7, 14, 24]. Pretrained MLLMs are powerful neural networks known for their ability to learn from limited examples, a common challenge in specialized fields like radiology. These models are designed to handle both images and text prompts during training, offering an opportunity to improve RRG by leveraging both visual content and textual descriptions. However, existing MLLM-based approaches often struggle to accurately capture the nuances of RRG. Radiologists meticulously examine specific regions within an X-ray image and reference these regions in their textual descriptions [4]. This precision in referencing image regions enhances the interpretability of the generated reports. The majority of current RRG approaches do not exploit the training of locally grounded descriptions (Figure 1).

To address these challenges, we propose a new approach to enhance RRG using grounded MLLMs. We introduce a novel fine-grained interleaved dataset by combining existing large-scale X-ray datasets, including well-known sources like MIMIC-CXR [4], with detailed region-text annotations. This dataset explicitly captures the crucial relationship between localized image regions and textual phrases, enabling our model to generate precise radiology reports, akin to radiologists' practices. We incorporate location tokens into our methodology, improving the model's ability to connect specific regions within X-ray images with their corresponding textual descriptions. This feature aligns the model more closely with radiologists' workflows, establishing a direct link between the radiology report's text and specific regions of interest within the X-ray image.

Recognizing the importance of temporal context in interpreting radiographic images, we leverage previous X-ray images, alongside the current one, providing the model with valuable historical information. This temporal context enriches the model's capacity to generate coherent and context-aware radiology reports, a crucial aspect of clinical decision-making. Our proposed method not only significantly enhances the performance of state-of-the-art MLLMs for RRG but also results in the production of higher quality reports. These improved reports incorporate essential grounding information, ultimately enhancing the interpretability of the model-generated reports and their clinical utility in patient care.

Moreover, our approach is designed to be generic and extendable, making it applicable to a wide range of Multimodal Large Language Models for RRG. This generality underscores the potential impact of our approach in advancing the field of medical image analysis and report generation. In this paper, we provide a comprehensive exposition of our methodology, detailing our dataset creation process, training strategies, and evaluation results. Through rigorous experiments, we demonstrate the significant advancements achieved in refining RRG with grounded MLLMs, offering a promising avenue for improved healthcare decision-making. In summary, our main contributions are:

1. We introduce a unique fine-grained interleaved dataset, created by combining large-scale X-ray datasets with region-text annotations, providing a valuable resource for training and evaluation.
2. Our method incorporates location tokens to MLLMs for radiology, facilitating the model's ability to associate specific image regions with corresponding textual phrases, mirroring the practice of radiologists.
3. We leverage temporal context through the inclusion of previous X-ray images alongside the current one. This enhances the model's capacity to generate coherent and context-aware radiology reports by reducing hallucinations.
4. We demonstrate that our proposed approach significantly improves the performance of state-of-the-art MLLMs for RRG, resulting in higher quality reports that incorporate essential grounding information.

2 Related work

Foundation models. In recent times, there has been a notable proliferation of generative Large Language Models (LLMs), exemplified by commercial models like GPT-4 [17], and PaLM-2 [20], as well as open-source alternatives like Llama2 [22]. This surge in interest has extended to the domain of multimodal foundation models, where significant progress has been made in various applications involving natural images. Notable examples in this context include BLIP-2 [8], Flamingo [1], LlaVa [13], and Vila [10].

In the medical domain, the development of relevant LLMs and Very Large Language Models (VLLMs) has also garnered attention. Models like LLava-Med [7], Medflamingo [14], and RadFM [24] have been tailored to address specific medical applications, including RRG. Despite these advancements, a persistent challenge in these models lies in their susceptibility to hallucinations and the introduction of errors in the generated radiology reports.

Interleaved datasets. In contrast to the natural scenery domain, which has abundant resources like MMC4 [27], Visual Genome [5], and LION-5B [21], the medical domain lacks extensive multimodal datasets. The most widely used multimodal medical dataset is MIMIC-CXR [4], which contains only chest X-ray images with captions, totaling 224,000 samples. PMC-OA [11] provides a dataset of 1.6 million image-caption pairs, but many 3D medical scans are presented as 2D slices. Medical Visual Question Answering (VQA) datasets, like VQA-RAD [6], SLAKE [12], and PMC-VQA [25], also exist but are limited to 2D images. Med-Flamingo [15] offers a dataset called MTB with approximately 800,000 image-text pairs, but it is not open-source. RadFM [24] combines various existing medical datasets, including MIMIC-CXR [4] and PMC-OA [11], to create MedMD, which contains 16 million

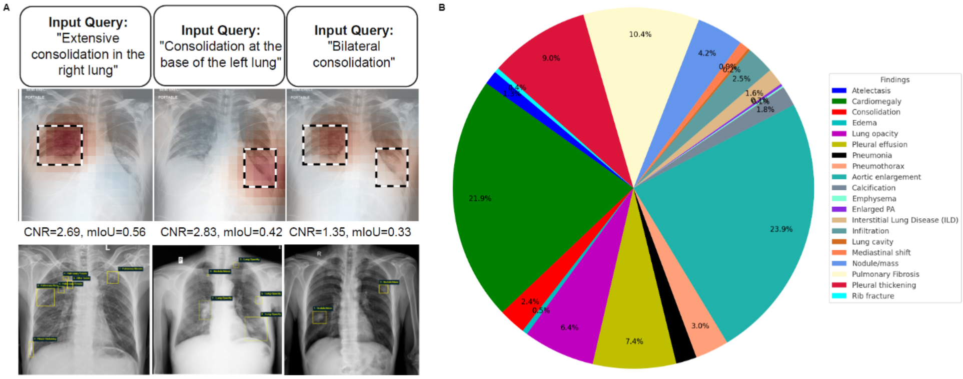


Figure 2: **A**: Top row illustrates MS-CXR dataset examples, featuring heatmaps of latent vector similarities over dashed, radiologist-provided ground-truth annotations; bottom row displays VinDr-CXR images with radiologist-marked local labels, figures from [4, 16]. **B**: Shows the distribution of annotation pairs by clinical findings in the small dataset of grounded image-text pairs.

2D image-text pairs, including 15.5 million 2D images and 500,000 3D scans with captions or diagnosis labels.

These datasets primarily feature complete radiology images paired with radiology report text, using a global pairing approach. This approach poses challenges for models in learning detailed alignments between text references and specific image regions. Additionally, it hampers the generation of visually grounded reports, which our research aims to address.

3 Dataset curation

We introduce a Small Dataset of Grounded Image-Text Pairs, which is created based on image-text pairs from MS-CXR [4] VinDr-CXR [16]. The MS-CXR dataset offers 1,153 image-sentence pairs, each including a bounding box and a radiology text description, verified by two board-certified radiologists, and equally distributed across eight cardiopulmonary conditions. The VinDr-CXR dataset comprises 18,000 images annotated by 17 experienced radiologists, featuring 22 local and 6 global labels identifying suspected radiological abnormalities and diseases. See Figure 2 for examples.

We construct a pipeline to extract and link phrases referring to abnormalities in the caption to their corresponding image regions. The pipeline mainly consists of three steps: 1) generating abnormality-bounding-box pairs, 2) producing referring-expression-bounding-box pairs and 3) adding historical images from MIMIC-CXR when they are available. We describe these steps in detail below:

Step-1: Generating abnormality-bounding-box pairs: Given an image-text pair, we turn phrase annotations into sentences resembling how findings are reported in usual radiology reports. We use a large language model to automatically do this for all instances that require it.

Step-2: Producing referring-expression-bounding-box pairs: In order to endow the model with the ability to ground abnormality descriptions, we transform bounding box coordinates into a referencing system using special tokens.

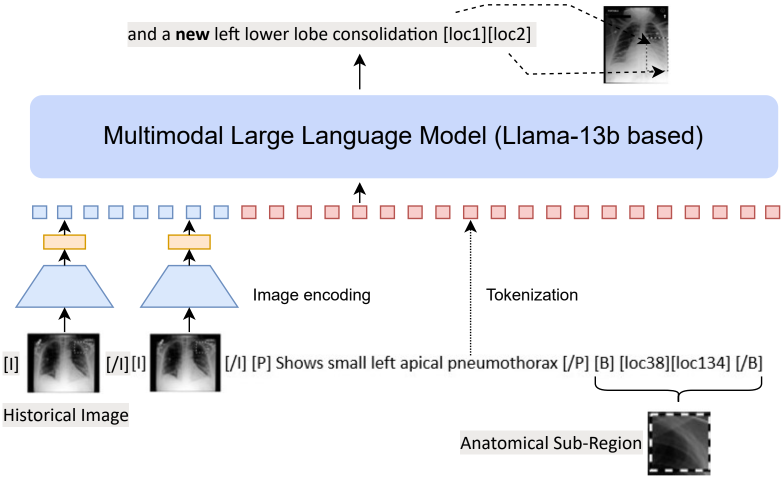

and a **new** left lower lobe consolidation [loc1][loc2]
Multimodal Large Language Model (Llama-13b based)
Image encoding
Tokenization
[I]
[/I] [I]
[/I] [P] Shows small left apical pneumothorax [/P] [B] [loc38][loc134] [/B]
Historical Image
Anatomical Sub-Region

Figure 3: The Architecture of the Proposed Multimodal Large Language Model for Radiology Report Generation. Our model, integrates both current and historical X-ray images with annotated ground truth text reports. These reports include special tokens that identify the location of relevant findings within the images. During training, the model learns to associate these tokens with the corresponding anatomical sub-regions, allowing for the generation of detailed reports that include localized descriptions of findings.

Step-3: Adding historical images: We add historical images or images from previous studies to complement the information necessary for accurately aligning reports which rely on comparisons. Not all images have previous ones available.

The final dataset contains 20000 image-sentence pairs. See Figure 2 for data samples and a detailed distribution of annotation pairs across different clinical findings. For additional dataset details see *Supplementary material*.

4 Methodology

In this section, we present our proposed method for visually-grounded RRG using MLLMs. Sec. 4.1 explains the process to prepare the data to train a visually-grounded MLLM, Sec. 4.2 introduces the overall architecture of the model. Sec. 4.3 presents the training details.

4.1 A Grounded MLLM for RRG

Our proposed model incorporates grounding and reference abilities through a procedure inspired by Kosmos-2 [19]. The model is designed to handle input from user-defined bounding boxes on images, providing visual feedback in the form of bounding boxes and linking the textual outputs to the visual elements. To enhance the model's ability to ground and refer to specific visual content, we incorporate pairs of grounded radiology images and text reports into the training set. For elements like descriptive phrases associated with specific bounding boxes in these pairs, we convert the bounding box coordinates into a series of location tokens. These tokens are subsequently combined with textual tokens to achieve a coherent encoding process.

Grounded Input Representations: In the process involving a grounded image-text pair, we start by transforming the continuous coordinates of the bounding boxes associated with a text span into a sequence of discrete location tokens [19]. For an image of width W and height H, we subdivide each dimension into P equal segments, resulting in $P \times P$ bins. Each bin encompasses a pixel area defined by $\left(\frac{W}{P}\right) \times \left(\frac{H}{P}\right)$. We assign a location token to each bin to denote its coordinates, using the center pixel of each bin as the reference point for bounding box calculations. Consequently, we generate $P \times P$ location tokens, which are incorporated into the text vocabulary for integrated image-text modeling.

Bounding boxes are specified by their top-left (x_1, y_1) and bottom-right (x_2, y_2) coordinates. These coordinates are discretized into location tokens. To denote a bounding box, we string together the top-left and bottom-right location tokens with boundary markers, forming the sequence: <box><loc$_1$><loc$_2$></box>. When a text span corresponds to multiple bounding boxes, we interpose the location tokens of these boxes with a delimiter token <delim>. This structure informs the model about the connection between the image regions within the bounding boxes and the associated text span.

4.2 Architecture Details

Our model follows the architecture of RadFM [24], which consists of three core components, **i)** Image encoder, **ii)** Perceiver aggregator and **iii)** LLM. Different to RadFM, we train on grounded image-text pairs. We allow spatial positions within both inputs and outputs, enabling visual prompts as inputs and grounded objects in the model outputs. Notably, the original RadFM model or any other SOTA model cannot perform object grounding or accept region inputs. We describe each component in the architecture as follows:

Image encoder. From a single instance in our dataset, represented as $\mathcal{X} = \{\mathcal{T}, \mathcal{V}\}$ where $\mathcal{V} = \{v_1, v_2, \dots, v_N\}$, each image v_i is initially processed independently using a visual encoder, denoted Φ_{vis}. In line with the approach used in RadFM, we employ a 3D Vision Transformer (ViT) to accommodate both 2D and 3D image inputs. For processing 2D images, we introduce an additional dimension representing depth by duplicating the image layers, thus representing each scanned image as $v_i \in \mathbb{R}^{H \times W \times D_i \times C}$. Here, C stands for the number of channels, while H, W, and D_i refer to the image's height, width, and depth, respectively.

Aggregation with Perceiver. After visual encoding, we adopt a perceiver [3] module Φ_{per} to aggregate visual representation. Specifically, Φ_{per} follows the classical perceiver architecture with a fix number of learnable queries as the latent array input, and the visual embedding v_i is treated as the byte array input, so that the final output embeddings will be normalized into the same length with the pre-defined learnable query sequence. The aggregation procedure can be formulated as $u_i = \Phi_{\text{per}}(v_i) \in \mathbb{R}^{P \times d}$, where u_i refers to the aggregated visual embedding, P is the pre-defined sequence length number of the learnable queries and d is the feature dimension.

Multi-modal fusion. We combine visual embeddings and text embeddings, which are generated from tokenization, to integrate visual and linguistic data. In this approach, special placeholder tokens designated for images are directly replaced by the corresponding visual embeddings. This mixed sequence is then input into a decoder-only large language model (Φ_{llm}). Within this model, the self-attention layers of the transformer architecture are effectively utilized as multi-modal fusion mechanisms. The process can be expressed as follows: $p = \Phi_{\text{llm}}(\text{concat}(t_1, u_1, t_2, u_2, t_3, \dots))$, where t_i and u_i are the text and visual embeddings respectively, and p represents the probability distribution for predicting subsequent tokens.

We utilize a 12-layer 3D Vision Transformer (ViT) with 768 feature dimensions for the visual encoder. Additionally, a 6-layer transformer decoder, known as the perceiver, is employed, featuring a learnable latent array of dimensions 32×5120. This configuration ensures that all images are embedded into a feature space of 32×5120.

When inserting them into the text embedding, we will add two extra special tokens <image>, </image> at the beginning and ending respectively to distinguish them from common text tokens. We initialize our large language model with the MedLLaMA-13B model introduced by PMC-LLaMA [23], which has further fine-tuned the LLaMA-13B [22] model on the medical corpus. Our final model has **14B** parameters. A Low-Rank Adaptation (LoRA) [2] based strategy is used for fine-tuning the LLM. While training, instead of fine-tuning all of the weights, we finetune two smaller matrices in LoRA that approximate the original larger matrix. After that, the fine-tuned adaptor is fed into the pretrained model and utilised for inference. The LoRA adaptation ensures faster training and avoids forgetting original knowledge embedded in the LLM trained and fine-tuned on generic natural language instructions.

4.3 Model training

Image preprocessing. We apply similar pre-processing steps as in RadFM [24]: resize the images to 512×512. In our dataset, each training sample consists of two elements, *i.e.*, $\mathcal{X} = \{\mathcal{T}, \mathcal{V}\}$, where $\mathcal{T}$ refers to the language part of the report, with special placeholder tokens for images, *e.g.*, "<image-1> <image-2> *A layering left-sided pleural effusion is moderate in size and new since the prior study*". $\mathcal{V}$ refer to the visual parts containing a set of 2D image scans, *i.e.*, $\mathcal{V} = \{v_1, v_2, \ldots, v_N\}$, $v_i \in \mathbb{R}^{H \times W \times C}$ or $v_i \in \mathbb{R}^{H \times W \times D \times C}$, H, W, D, C are height, width, depth, channel respectively, corresponding to the "<image-*i*>" token in $\mathcal{T}$. $\mathcal{T}$ and $\mathcal{V}$ serve as interleaved prompts of language and image input to the model. The primary aim is to predict the likelihood of the text tokens in $\mathcal{T}$, based on the xray images, which is expressed as:

$$p(\mathcal{T}|\mathcal{V}) = \prod p(\mathcal{T}_l|\mathcal{V}_{<l}, \mathcal{T}_{<l}), \tag{1}$$

Here, $\mathcal{T}_l$ denotes the l-th token of $\mathcal{T}$, while $\mathcal{V}_{<l}$ and $\mathcal{T}_{<l}$ represent the images and language text prior to the l-th token. These probabilities are parameterized using our generative model (Φ_{Ours}). The final training objective is defined by the negative log-likelihood of the correct next token in the sequence:

$$\mathcal{L}_{\text{reg}} = -\sum w_l \log \Phi_{\text{Ours}}(\mathcal{T}_l|\mathcal{V}_{<l}, \mathcal{T}_{<l}), \tag{2}$$

where w_l is the weighting per token, designed to highlight significant tokens or disregard special tokens.

5 Experiments

For the evaluation of RRG, we use the MIMIC-CXR v2 [4] chest X-ray dataset, which contains longitudinal imaging studies with corresponding radiological reports. Medical Visual Question Answering, Modality Recognition and Disease Diagnosis are reported on the same datasets as in RadFM. For fair comparison, we use the same split as in RadFM [24]. We only use frontal view scans. We use the following evaluation metrics: BLEU [18], ROUGE [9], UMLS Precision [24] and UMLS Recall [24].

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

5.1 Results and Discussion

Our proposed model represents a notable step forward in the field of RRG, as indicated by its encouraging performance improvement detailed in Table 1. We show a +7.22 increase in the Rouge metric, +6.92 in Bleu, and significant gains of +10.14 and +10.08 in UMLSp and UMLSr respectively. These metrics collectively indicate that our model not only excels in capturing the key points that are essential for high-quality reports but also demonstrates a good ability to generate text that is coherent, contextually relevant, and accurate.

The enhancements in Rouge and Bleu scores highlight the model's refined capability to produce reports that are both informative and linguistically sound, closely mirroring the expert-written narratives in structure and content. Furthermore, the substantial improvements in UMLSp and UMLSr metrics emphasize the model's strengthened proficiency in medical semantics.

We extended our evaluation to include the performance of our method on the tasks of Modality Recognition and Medical Disease Diagnosis, as presented in Table 2. Although our method and training were primarily focused on RRG and medical visual question answering, we examined these additional tasks to assess potential performance degradation from the base model. The results demonstrate that our specialized training did not lead to any significant performance loss, confirming the model's capacity to maintain its foundational diagnostic performance, while being significantly better for specific tasks.

Table 3: Ablation study of different variations of our prosed model for RRG on the MIMIC-CXR dataset. **Historic imgs:** indicates whether we train with previous x-rays images (when available) or not; **Grounding:** whether we train with grounding tokens or not and **Impression:** whether we predict this section of the report or not.

Model Variations			MIMIC-CXR			
Historic imgs	**Grounding**	**Impression**	**R**	**B**	**UMLSp**	**UMLSr**
Baseline						
RadFM [24]			12.81	18.22	22.49	12.07
Ours						
✓	-	-	13.55	19.58	23.72	13.41
-	✓	-	15.03	20.84	25.00	14.79
-	-	✓	13.09	19.02	23.14	13.92
✓	✓	-	16.12	21.35	25.80	16.08
✓	✓	✓	**17.00**	**22.14**	**26.31**	**17.24**

How does the inclusion of historical images impact model performance? Incorporating historical X-ray images from previous patient studies into our model significantly enhances the generated radiology reports' quality. Our work represents the first instance of such an approach among state-of-the-art models in this domain. This innovation allows for a contextual analysis that enriches the report, enabling the model to identify and articulate temporal changes, such as *"the pleural effusion is larger than before"* or detecting new findings like *"there is new cardiomegaly."* The effectiveness of this novel integration is clearly supported by the data in Table 3 of our ablation study, demonstrating improvements of 0.74, 1.36, 1.23, and 1.34 in the Rouge, Bleu, UMLSp, and UMLSr metrics, respectively. These results indicate that including historical images enhances the model's ability to provide accurate, context-aware, and clinically relevant radiological assessments, marking a significant advancement in the automated generation of radiology reports.

What impact do grounding tokens have on enhancing RRG? Introducing grounding tokens during training has significantly enhanced our model's ability to correlate textual phrases with specific image regions. This is an important innovation in state-of-the-art large language models for RRG. This technique bolsters the quality of the generated reports, particularly in the *"Findings"* section, by enabling precise alignment between textual descriptions and the corresponding areas in the X-ray images, using spatial references such as left, right, top, bottom, and specific region names. The benefits of this approach are evident in the improvements reported in Table 3 of our ablation study, showing enhancements of 2.22, 2.62, 2.51, and 2.72 in the Rouge, Bleu, UMLSp, and UMLSr metrics respectively. These results confirm that grounding textual phrases to image regions significantly improves the model's accuracy and the clinical relevance of the generated radiology reports, compared to the baseline model that did not utilize grounding tokens.

What are the benefits of exclusively generating the "Findings" section? Our decision to have the model generate only the *"Findings"* sections of radiology reports, excluding the *"Impressions"* sections, is justified by the requirement of accuracy and reliability in automated reporting. The *"Impressions"* section often relies on external patient-specific information and additional studies that our model does not access, increasing the risk of generating incorrect or speculative content. This focused approach is validated by the enhancements observed in our ablation study, detailed in Table 3. We noted performance improvements with increases of 0.28, 0.80, 0.65, and 1.85 in the Rouge, Bleu, UMLSp, and UMLSr metrics, respectively, by omitting the generation of the *"Impressions"* section. This strategy ensures the model's outputs remain grounded in the visual evidence provided by the X-ray images, mitigating the risk of information hallucination and boosting the credibility of the generated reports.

6 Conclusion

Our study has demonstrated the potential of grounded Multimodal Large Language Models (MLLMs) in improving RRG. We've addressed the challenge of aligning text with image regions, leading to more interpretable and clinically useful reports. By introducing a new dataset, location tokens, and leveraging temporal context, we've advanced RRG. This results in higher-quality reports with essential grounding information, benefiting patient care and clinical decisions. Our approach is versatile and can be applied to various MLLMs for RRG, potentially impacting medical image analysis and report generation.

Future work: Future research could focus on refining training methods, using additional multimodal datasets, and exploring techniques to further enhance report interpretability.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. *ArXiv*, abs/2204.14198, 2022.

[2] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[3] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021.

[4] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

[5] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

[6] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018.

[7] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*, 2023.

[8] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ArXiv*, abs/2301.12597, 2023.

[9] Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W04-1013.

[10] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. *arXiv preprint arXiv:2312.07533*, 2023.

[11] Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents. *arXiv preprint arXiv:2303.07240*, 2023.

[12] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654. IEEE, 2021.

[13] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.

[14] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR, 2023.

[15] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Cyril Zakka, Yash Dalmia, Eduardo Pontes Reis, Pranav Rajpurkar, and Jure Leskovec. Med-flamingo: A multimodal medical few-shot learner. July 2023. URL https://arxiv.org/abs/2307.15189. arXiv:2307.15189.

[16] Ha Q Nguyen, Khanh Lam, Linh T Le, Hieu H Pham, Dat Q Tran, Dung B Nguyen, et al. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. *Scientific Data*, 9(1):429, 2022.

[17] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023. URL https://api.semanticscholar.org/CorpusID:257532815.

[18] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://www.aclweb.org/anthology/P02-1040.

[19] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

[20] Orhan Firat et al. Rohan Anil, Andrew M. Dai. Palm 2 technical report. *ArXiv*, abs/2305.10403, 2023. URL https://api.semanticscholar.org/CorpusID:258740735.

[21] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[22] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[23] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-llama: Further finetuning llama on medical papers. *arXiv preprint arXiv:2304.14454*, 2023.

[24] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology. *arXiv preprint arXiv:2308.02463*, 2023.

[25] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *ArXiv*, abs/2305.10415, 2023.

[26] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023.

[27] Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal C4: An open, billion-scale corpus of images interleaved with text. *arXiv preprint arXiv:2304.06939*, 2023.