MAIRA at RRG24: A specialised large multimodal model for radiology report generation

Shaury Srivastav[1], Mercy Ranjit[1], Fernando Pérez-García[2], Kenza Bouzid[2], Shruthi Bannur[2], Daniel C. Castro[2], Anton Schwaighofer [2], Harshita Sharma[2], Maximilian Ilse[2], Valentina Salvatelli[2], Sam Bond-Taylor[2], Fabian Falck[2], Anja Thieme[2], Hannah Richardson[2], Matthew P. Lungren[3], Stephanie L. Hyland[2], Javier Alvarez-Valle[2]

[1]Microsoft Research India,
[2]Health Futures, Microsoft Research,
[3]Microsoft Health and Life Sciences

Correspondence: meranjit@microsoft.com

Abstract

This paper discusses the participation of the MSR MAIRA team in the Large-Scale Radiology Report Generation Shared Task Challenge, as part of the BioNLP workshop at ACL 2024. We present a radiology-specific multimodal model designed to generate radiological reports from chest X-Rays (CXRs). Our proposed model combines a CXR-specific image encoder RAD-DINO (Pérez-García et al., 2024) with a Large Language Model (LLM) based on Vicuna-7B, via a multi-layer perceptron (MLP) adapter. Both the adapter and the LLM have been fine-tuned in a single-stage training setup to generate radiology reports. Experimental results indicate that a joint training setup with findings and impression sections improves findings prediction. Additionally, incorporating lateral images alongside frontal images when available further enhances all metrics. More information and resources about MAIRA can be found on the project website: http://aka.ms/maira.

1 Introduction

An impactful application of natural language generation in the medical field involves the creation of support systems that interpret patient X-ray images and produce a draft report detailing the clinical findings in these images. Such systems have the potential to enhance and expedite radiology reporting workflows. In this regard, a Shared Task Challenge for Radiology Report Generation has been organised as part of the ACL 2024 BioNLP workshop[1] (RRG24; Xu et al., 2024). This shared task challenge uses the first large-scale collection of radiology report generation datasets based on MIMIC-CXR (Johnson et al., 2019), CheXpert (Chambon et al., 2024), PadChest (Bustos et al., 2020), BIMCV-COVID19 (Vayá et al., 2020), and Open-i (Nguyen et al., 2022) datasets. This paper covers the participation of the MAIRA (Multimodal AI for Radiology Applications) team at Microsoft Research in this challenge.

We build on the architecture and training approach from our earlier work MAIRA-1 (Hyland et al., 2024). This approach combines a CXR-specific image encoder (RAD-DINO, Pérez-García et al. (2024)) with a pretrained LLM (Vicuna-7B 1.5, Chiang et al. (2023)) via an adapter which is an MLP of 4 layers. Both LLM and adapter are finetuned in a single stage for the task of radiology report generation, while the image encoder is pretrained following the self-supervised DINOv2 approach (Oquab et al., 2024). For this competition, we produce a variant of MAIRA-1 which is trained only on public data, and further extended it to incorporate lateral images. We share the following outcomes:

1. A joint training setup for findings and impression prediction improves the metrics for findings generation.

2. Coupling lateral views with frontal views when available shows improvement in both the clinical and lexical metrics.

3. We show that scaling the model size from Vicuna-7B to 13B further helps to improve all the metrics. We also show that smaller models like Phi-3-mini with 3.8B parameters is on-par with the larger models.

[1] https://stanford-aimi.github.io/RRG24/

Proceedings of the 23rd Workshop on Biomedical Language Processing, pages 597–602

Table 1: Number of studies with a given section across the data sources and splits in the RRG24 challenge dataset.

Split	Section	MIMIC-CXR	CheXpert	Open-i	PadChest	BIMCV-COVID19
Train	Findings	148 374	45 491	3252	101 752	45 525
	Impression	181 166	181 619	3628	–	–
Validation	Findings	3799	1112	85	2641	1202
	Impression	4650	4589	92	–	–

2 Method

2.1 Dataset

The dataset statistics of the RRG24 challenge are available in Table 1. Each study can have multiple frontal and/or lateral images. We processed this dataset into two versions: frontal-only (Table 2), where each record contains only one frontal image, and frontal-with-laterals (Table 3), where each record contains one frontal or a frontal and a lateral image. Hence, each record in the RRG24 dataset may be split into more than one record in these datasets.

Table 2: RRG24 frontal-only dataset. Each record corresponds to exactly one frontal image.

Section	Train	Test	Val.	Hidden
Findings	359 351	2900	9215	1133
Impression	381 075	3205	9691	1495
Total	740 426	6105	18 906	2628

Table 3: RRG24 frontal-with-laterals dataset. Each record corresponds to a frontal image (F) with or without a lateral image (L).

Section	View	Train	Test	Val.	Hidden
Findings	F	174 977	1193	4420	562
	F+L	196 023	1897	5081	629
Impression	F	226 673	1392	5792	825
	F+L	165 835	2005	4193	734
Total		763 508	6487	19 486	2750

The image encoder is trained using the images from the MIMIC-CXR (Johnson et al., 2019), CheXpert (Chambon et al., 2024), PadChest (Bustos et al., 2020), NIH-CXR (Wang et al., 2017), and BRAX (Reis et al., 2022) datasets. Both frontal and lateral view images were used. The dataset statistics by source are available in Table 4. The images in the validation and test sets of RRG24 challenge were excluded.

Table 4: Training datasets for the image encoder.

Source	View	No. of images
MIMIC-CXR	Frontal, lateral	367 932
CheXpert	Frontal, lateral	218 180
PadChest	Frontal, lateral	156 432
NIH-CXR	Frontal	112 120
BRAX	Frontal, lateral	41 260

2.2 Model architecture

We leverage the MAIRA-1 (Hyland et al., 2024) architecture that consists of a CXR-specific image encoder, an adapter layer and an LLM. The image encoder (Pérez-García et al., 2024) is a ViT-B model (Dosovitskiy et al., 2020). The input image resolution is 518×518. The LLM is Vicuna-7B 1.5 (Chiang et al., 2023). The adapter is an MLP with 4 layers with a hidden size of 1024 for all the layers. The prompt setup we used is available in Table 5.

2.3 Training details

For the image encoder training, we follow RAD-DINO (Pérez-García et al., 2024) and use an image-level objective, a patch-level objective, and a regulariser to encourage uniform span of the features within a batch. We initialised the model with the weights of the pre-trained DINOv2 ViT-B (Oquab et al., 2024) and trained with the chest X-ray images in Table 4 for an additional 60k training steps, with an effective batch size of 640. We used the AdamW optimizer with a base learning rate of 0.001 and a cosine learning rate scheduler with linear warm-up.

For MAIRA-RRG24 training, we keep the image encoder frozen. We just train the adapter and the LLM with a standard auto-regressive language modelling loss. We use a cosine learning rate scheduler with a warm-up of 0.03 and learning rate of

Table 5: Prompt for the different tasks. F: frontal, L: lateral.

Setting	View	Prompt
Findings	F	Given the frontal image `{image_tokens}`, provide a description of the findings in the radiology study.
	F+L	Given the frontal image `{image_tokens}` and the lateral image `{lateral_image_tokens}`, provide a description of the findings in the radiology study.
Impression	F	Given the frontal image `{image_tokens}`, provide a summary impression in the radiology study.
	F+L	Given the frontal image `{image_tokens}` and the lateral image `{lateral_image_tokens}`, provide a summary impression in the radiology study.

Table 6: Experimental settings. F: frontal, L: lateral.

Setting	View	Task
Findings (F)	F	findings prediction
Findings + Impression (F)	F	findings and impression prediction (multi-task)
Findings (F+L)	F+L	findings prediction
Findings + Impression (F+L)	F+L	findings and impression prediction (multi-task)

2×10^{-5}. We train for 3 epochs with a global batch size of 128. During evaluation, as there could be multiple predictions for a study (a study could have more than one frontal/lateral images), we use `GPT-4` with the prompt defined in Table 10 to select the best prediction.

2.4 Evaluation metrics

We use the `vilmedic` package (Delbrouck et al., 2022b) for computing the metrics. ROUGE-L (Lin, 2004), BLEU-4 (Papineni et al., 2002) and BERTScore (Zhang et al., 2019) were used to measure the lexical performance. F1-CheXbert (Smit et al., 2020) and F1-Radgraph (Delbrouck et al., 2022a) were used to measure the clinical performance.

3 Experiments

We perform experiments in four settings as defined in Table 6: single-task training for findings generation, joint(multitask) training for findings and impression generation, and combinations of both with or without lateral images alongside frontal images. We use dataset versions in Table 2 and Table 3 when we train for frontal only setup and frontal and lateral setup respectively. We call the model trained with the multitask setting for findings and impression prediction using the frontal and lateral images as MAIRA-RRG24. We also trained MAIRA-RRG24 with a smaller LLM, `Phi-3-mini-4k-instruct` (Abdin et al., 2024) with 3.8B parameters. We also performed an additional scaling experiment for the findings generation task with laterals (third setting in Table 6) using the Vicuna 7B and 13B versions.

3.1 Results

The results of the experiments are available in Table 7. We find that a joint training setup involving findings and impression prediction tasks shows a slight improvement in the findings prediction metrics compared to training for findings prediction alone. Additionally, training with lateral images in addition to frontal images further improves all the metrics. The best experimental setup, which is the Findings+Impression (F+L) setting involves joint training for findings and impression prediction tasks, along with the inclusion of lateral images when available. The results of our best setting on the hidden test set are presented in Table 9. We also trained our best setting, with a smaller model `Phi-3-mini-4k-instruct` and got better or competitive results in all the metrics. The results of the model scaling experiment in Table 8 demonstrate that a larger model size helps to improve the metrics.

Table 7: MAIRA-RRG24 – Experimental results for findings generation task on the public-test set.

Setting	BLEU-4	ROUGE-L	BERTScore	F1-CheXbert	F1-RadGraph
Findings (F)	10.61	26.70	54.54	52.64	24.57
Findings+Impression (F)	10.88	26.86	54.50	55.55	24.68
Findings (F+L)	11.20	26.59	54.53	56.95	24.84
Findings+Impression (F+L)	12.26	28.00	55.76	59.71	26.33
Findings+Impression (F+L) (Phi-3-mini-4k-instruct)	14.84	29.17	58.91	55.87	27.07

Table 8: MAIRA-RRG24 – Model scaling experiment. Public test results for Findings (F+L) setting.

LLM	BLEU-4	ROUGE-L	BERTScore	F1-CheXbert	F1-RadGraph
vicuna-7b-v1.5	11.20	26.59	54.53	56.95	24.84
vicuna-13b-v1.5	12.17	27.86	55.62	59.66	26.21

Table 9: MAIRA-RRG24 – Hidden test set results for the Findings+Impression (F+L) setting.

Task	BLEU-4	ROUGE-L	BERTScore	F1-CheXbert	F1-RadGraph
Findings Generation	11.24	26.58	54.22	57.87	25.48
Impression Prediction	11.66	28.48	51.62	53.27	25.26

Table 10: GPT-4 prompt for selecting the best report for a study when there are multiple records.

```
You are an AI assistant who helps to select the best radiology report from multiple reports written for the same patient. User will send you a list of reports. You will select the best report based on the below criteria.
1. It has the best complete list of findings that contains the findings from other reports as well.
2. Do not contain hallucinations like comparison to a previous report and other noisy details.
3. The writing style matches closely with that of a radiologist.
Return just the number of the index of the list corresponding to the best report. The index starts with 0.
```

4 Limitations

MAIRA-RRG24 does not have access to the prior studies and hence it may generate 'hallucinated' references to prior studies (Bannur et al., 2023).

5 Conclusion

We have presented MAIRA-RRG24, a radiology-adapted large multimodal model based on the MAIRA-1 architecture (Hyland et al., 2024) with a RAD-DINO-like (Pérez-García et al., 2024) image encoder, trained exclusively with the data available for the RRG24 challenge (Xu et al., 2024). It exhibits competitive performance in both lexical and clinical metrics. It benefits from a domain-specific image encoder, a joint training setup for findings and impression prediction leveraging lateral images when available.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Qin Cai, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Yen-Chun Chen, Yi-Ling Chen, Parul Chopra, Xiyang Dai, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Victor Fragoso, Dan Iter, Mei Gao, Min Gao, Jianfeng Gao, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Ce Liu, Mengchen Liu, Weishung Liu, Eric Lin, Zeqi Lin, Chong Luo, Piyush Madan, Matt Mazzola, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael San-

tacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Xin Wang, Lijuan Wang, Chunyu Wang, Yu Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Haiping Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Sonali Yadav, Fan Yang, Jianwei Yang, Ziyi Yang, Yifan Yang, Donghan Yu, Lu Yuan, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Pérez-García, Maximilian Ilse, Daniel C. Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, Anton Schwaighofer, Maria Wetscherek, Matthew P. Lungren, Aditya Nori, Javier Alvarez-Valle, and Ozan Oktay. 2023. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15016–15027.

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. 2020. PadChest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797.

Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P. Langlotz. 2024. Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats. *Preprint*, arXiv:2405.19538.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality.

Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. 2022a. Improving the factual correctness of radiology report generation with semantic rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4348–4360.

Jean-benoit Delbrouck, Khaled Saab, Maya Varma, Sabri Eyuboglu, Pierre Chambon, Jared Dunnmon, Juan Zambrano, Akshay Chaudhari, and Curtis Langlotz. 2022b. ViLMedic: a framework for research at the intersection of vision and language in medical AI. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 23–34.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Stephanie L. Hyland, Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Mercy Ranjit, Anton Schwaighofer, Fernando Pérez-García, Valentina Salvatelli, Shaury Srivastav, Anja Thieme, Noel Codella, Matthew P. Lungren, Maria Teodora Wetscherek, Ozan Oktay, and Javier Alvarez-Valle. 2024. MAIRA-1: A specialised large multimodal model for radiology report generation. *Preprint*, arXiv:2311.13668.

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

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Ha Q. Nguyen, Khanh Lam, Linh T. Le, Hieu H. Pham, Dat Q. Tran, Dung B. Nguyen, Dung D. Le, Chi M. Pham, Hang T. T. Tong, Diep H. Dinh, Cuong D. Do, Luu T. Doan, Cuong N. Nguyen, Binh T. Nguyen, Que V. Nguyen, Au D. Hoang, Hien N. Phan, Anh T. Nguyen, Phuong H. Ho, Dat T. Ngo, Nghia T. Nguyen, Nhan T. Nguyen, Minh Dao, and Van Vu. 2022. VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations. *Scientific Data*, 9(1):429.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2024. DINOv2: Learning robust visual features without supervision. *Preprint*, arXiv:2304.07193.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Fernando Pérez-García, Harshita Sharma, Sam Bond-Taylor, Kenza Bouzid, Valentina Salvatelli, Maximilian Ilse, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Matthew P Lungren, et al. 2024. RAD-DINO: Exploring scalable medical image encoders beyond text supervision. *Preprint*, arXiv:2401.10815.

Eduardo P. Reis, Joselisa P. Q. de Paiva, Maria C. B. da Silva, Guilherme A. S. Ribeiro, Victor F. Paiva, Lucas Bulgarelli, Henrique M. H. Lee, Paulo V. Santos, Vanessa M. Brito, Lucas T. W. Amaral, Gabriel L.

Beraldo, Jorge N. Haidar Filho, Gustavo B. S. Teles, Gilberto Szarf, Tom Pollard, Alistair E. W. Johnson, Leo A. Celi, and Edson Amaro. 2022. BRAX, Brazilian labeled chest X-ray dataset. *Scientific Data*, 9(1):487.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew Lungren. 2020. Combining automatic labelers and expert annotations for accurate radiology report labeling using bert. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1500–1519.

Maria De La Iglesia Vayá, Jose Manuel Saborit, Joaquim Angel Montell, Antonio Pertusa, Aurelia Bustos, Miguel Cazorla, Joaquin Galant, Xavier Barber, Domingo Orozco-Beltrán, Francisco García-García, et al. 2020. BIMCV COVID-19+: a large annotated dataset of RX and CT images from COVID-19 patients. *Preprint*, arXiv:2006.01174.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. 2017. ChestX-Ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Justin Xu, Zhihong Chen, Andrew Johnston, Louis Blankemeier, Maya Varma, Jason Hom, William J. Collins, Ankit Modi, Robert Lloyd, Benjamin Hopkins, Curtis Langlotz, and Jean-Benoit Delbrouck. 2024. Overview of the first shared task on clinical text generation: Rrg24 and "discharge me!". In *The 23rd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Bangkok, Thailand. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.