LRTA-BioMIC: Lightweight Region-Text Aligned BioMIC-BART for Chest X-ray Report Generation

Anonymous ACL submission

Abstract

The global shortage of radiologists is a major challenge. Radiology is vital for diagnosing and treating diseases, especially in the lungs and heart, using imaging like X-rays. To address this shortage and workload, we introduce ***L**ightweight **R**egion-**T**ext **A**ligned **BioMIC**-BART* (**LRTA-BioMIC**), which generates Chest X-ray reports from X-ray images. *LRTA-BioMIC* is a computationally efficient, Domain Specific, Region Guided Text Aligned language model that integrates tagger information and X-ray embeddings from ViT through cross-attention at every layer of the BioMIC-BART Encoder to generate radiology reports (Findings and Impression). Our model achieves a notable improvement of **9.71%** in BLEU-4 and **0.9%** in ROUGE-L compared to the previous state-of-the-art, *COMG* and *KGVL-BART*, on the *IU-Xray* dataset. *LRTA-BioMIC* also demonstrates competitive performance on the *MIMIC-CXR-JPG* dataset, with a **1.60%** increase in BLEU-4 and a slight **3.53%** decrease in ROUGE-L compared to *RECAP*, the previous state-of-the-art. We will make our codes and resources publicly available.

1 Introduction

VLMs have found huge application in radiology report generation, due to their ability to generate text, coherent to image. However, all vision-language multimodal pipelines are plaqued by improper image-text alignment (Amirloo et al., 2024). Previous literature (Caffagni et al., 2024) shows integrating better image-text alignment can lead to better performance. Additionally, previous VLMs for radiology report generation have used computationally heavy pre-trained vision encoders and language decoders. In lieu of computationally-intensive Vision-Language Models (VLMs), we propose ***L**ightweight **R**egion-**T**ext **A**ligned **BioMIC**-BART* (**LRTA-BioMIC**), which generates Chest X-ray reports from X-ray images. We enabled multimodal processing of chest X-ray images and their corresponding reports on BioBART (Yuan et al., 2022), as it alone lacks image embedding knowledge, by training it on the *MIMIC-CXR-JPG* dataset using the KM-BART architecture (Xing et al., 2021). This resulted in **BioMIC-BART**, which serves as the backbone of *LRTA-BioMIC*, enhancing performance on both *IU-Xray* and *MIMIC-CXR-JPG*. For region-guided feature extraction from MedCLIP (Wang et al., 2022), we used the *Region Selector* from (Tanida et al., 2023) with cross-attention (CA_1). Since MedCLIP is trained on the cosine similarity between chest X-rays and reports, CA_1 enhances contextual chest image embeddings. Its output is then passed to cross-attention (CA_2) at the start of each BioMIC-BART layer, where it serves as keys and values, with the tagger information as the query. This improves textual and regional alignment before processing through BioMIC-BART, a domain-specific encoder-decoder model for chest X-rays.

Our contributions are as follows:

- **LRTA-BioMIC**, a computationally efficient, region-guided, and text-aligned model, achieving **9.71%** and **0.9%** improvements in *BLEU-4* and *ROUGE-L*, respectively, over the previous SoTA. for chest X-ray report generation.
- **BioMIC-BART**, an extension of **BioBART** trained on **MIMIC-CXR-JPG** to process multimodal chest X-ray images and text, serving as the backbone of *LRTA-BioMIC*.

2 Related Work

Early radiology report generation relied on CNN-RNN architectures (Jing et al., 2020, 2017), but recent advancements favor Transformer-based models (Vaswani, 2017). Region-selector Transformers, such as (Tanida et al., 2023) for anatomical

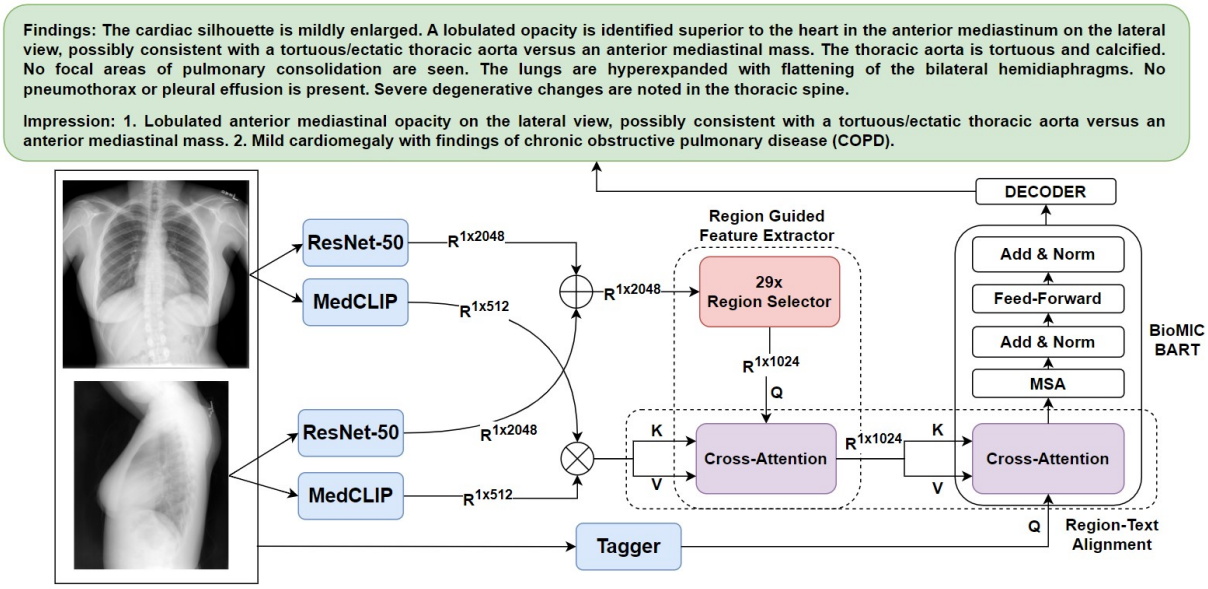

Figure 1: Architecture of **LRTA-BioMIC**. Chest X-ray images (PA & LL) are processed via ResNet-50 and MedCLIP to extract visual features. A 29-region selector refines region-specific embeddings. Textual tags, along with selected regions, aid image-text alignment in BioMIC-BART, which generates the final radiology report.

region detection and (Li et al., 2023) with its *Unify, Align, and then Refine (UAR)* strategy, further improved image-report alignment. Assistive systems (Nicolson et al., 2024), organ-specific masks (Gu et al., 2024), and observation-guided reasoning (Hou et al., 2023b,a) have also enhanced disease identification and report generation. Knowledge graphs, highlighted by (Zhang et al., 2020) and formalized in (Kale et al., 2023), improve multimodal learning, while prompt-based methods like (Jin et al., 2024) enhance rare disease representation. Despite newer Transformer-driven innovations, established models such as CMCA (Song et al., 2022), KnowMat (Yang et al., 2022), and CMM-RL (Qin and Song, 2022) remain robust and effective in Chest X-ray report generation. **LRTA-BioMIC** leverages **BioMIC-BART** for efficient multimodal processing, unlike previous architectures that either relied on computationally expensive VLMs or ineffective fused embeddings. Additionally, it incorporates selected image regions and text alignment, enhancing report quality.

3 Methodology

LRTA-BioMIC is trained by first developing **BioMIC-BART**, an extension of BART designed to process multimodal data, specifically chest X-ray images and medical text. The pretrained **BioMIC-BART** weights serve as the backbone for training our ***Lightweight Region-Text Aligned BioMIC-BART*** (**LRTA-BioMIC**), which incorporates region-level visual features and enhances text-image alignment.

3.1 BioMIC-BART

We build upon *BioBART-Large*, a language model trained on full-text PubMed articles (Yuan et al., 2022). While effective, its performance on Chest X-ray report generation is constrained due to a lack of radiology-specific training. To address this, we augment it with multimodal supervision using image-text pairs from MIMIC-CXR-JPG (Johnson et al., 2019), inspired by methods from (Xing et al., 2021), which effectively model image-text contextual relations. More detail is mentioned in Section 11.

3.2 Region-Guided Feature Extraction

To preprocess Chest X-rays, we extract multi-scale visual embeddings using ResNet-50 (He et al., 2016) and MedCLIP-ResNet50 (Wang et al., 2022). Given a chest X-ray $\mathbf{I}$, we obtain:

$$\begin{aligned} \mathbf{F}_{\text{res}}^{\text{PA}} &= \text{ResNet}(\mathbf{I}_{\text{PA}}) \in \mathbb{R}^{1\times 2048}, \\ \mathbf{F}_{\text{res}}^{\text{LL}} &= \text{ResNet}(\mathbf{I}_{\text{LL}}) \in \mathbb{R}^{1\times 2048}. \end{aligned} \quad (1)$$

$$\begin{aligned} \mathbf{F}_{\text{clip}}^{\text{PA}} &= \text{MedCLIP}(\mathbf{I}_{\text{PA}}) \in \mathbb{R}^{1\times 512}, \\ \mathbf{F}_{\text{clip}}^{\text{LL}} &= \text{MedCLIP}(\mathbf{I}_{\text{LL}}) \in \mathbb{R}^{1\times 512}. \end{aligned} \quad (2)$$

For comprehensive feature fusion, we compute:

$$\mathbf{F}_{\text{res}}^{\text{sum}} = \mathbf{F}_{\text{res}}^{\text{PA}} + \mathbf{F}_{\text{res}}^{\text{LL}} \in \mathbb{R}^{1\times 2048}, \quad (3)$$

Dataset	Model	NLG Metrics						CE Metrics		
		B-1	B-2	B-3	B-4	MTR	R-L	P	R	F_1
MIMIC -CXR	RGRG	0.373	0.249	0.175	0.126	0.168	0.264	0.461	0.475	0.447
	COMG	0.363	0.235	0.167	0.124	0.128	0.290	0.424	0.291	0.345
	PROMPTMRG	0.398	–	–	0.112	0.157	0.268	**0.501**	**0.509**	**0.476**
	ORGAN	0.407	0.256	0.172	0.123	0.162	**0.293**	0.416	0.418	0.385
	RECAP	**0.429**	**0.267**	0.177	0.125	0.168	0.288	0.389	0.443	0.393
	LRTA-BIOMIC	0.418	0.261	**0.179**	**0.127**	**0.171**	0.283	0.496	0.481	0.459
IU X-RAY	RGRG	0.266	–	–	0.063	0.146	0.180	0.183	0.187	0.180
	COMG	**0.536**	0.378	0.275	0.206	0.218	0.383	-	-	-
	PROMPTMRG	0.401	–	–	0.098	0.160	0.281	0.213	0.229	0.211
	ORGAN	0.510	0.346	0.255	0.195	0.205	0.399	-	-	-
	KGVL-BART	0.423	0.256	0.194	0.165	0.500	0.444	-	-	-
	LRTA-BIOMIC	0.527	**0.384**	**0.279**	**0.226**	**0.522**	**0.448**	**0.221**	**0.223**	**0.218**
ABLN	LRTA-BIOMIC$_1$	0.398	0.274	0.213	0.176	0.412	0.374	-	-	-
	LRTA-BIOMIC$_2$	0.483	0.359	0.275	0.211	0.510	0.427	-	-	-
	LRTA-BIOMIC$_3$	0.462	0.339	0.257	0.199	0.498	0.402	-	-	-
	LRTA-BIOMIC	0.527	0.384	0.279	0.226	0.522	0.448	-	-	-

Table 1: Experimental Results of our model and baselines on the IU X-RAY dataset and the MIMIC-CXR-JPG dataset. The best results are in **boldface**, and the underlined are the second-best results. We also include Ablation study marked by "ABLN" performed on IU X-RAY dataset. A one-tailed t-test between *LRTA-BioMIC* and *COMG* (best-performing baseline) on the BLEU-4 score yields **p = 0.0138 (< 0.05)**, confirming LRTA-BioMIC's statistically significant improvement for chest X-ray report generation.

$$\mathbf{F}_{\text{clip}}^{\text{concat}} = \text{concat}(\mathbf{F}_{\text{clip}}^{\text{PA}}, \mathbf{F}_{\text{clip}}^{\text{LL}}) \in \mathbb{R}^{1\times 1024}. \quad (4)$$

Additionally, $\mathbf{F}_{\text{region}} \in \mathbb{R}^{1\times 1024}$ (region-level embeddings) are extracted via 29-region selection (Tanida et al., 2023) and transformed using a multi-layer perceptron (MLP). The final visual representation is refined using cross-attention CA_1.

$$\mathbf{F}_{\text{rg}} = \text{softmax}\left(\frac{\mathbf{Q}_{\text{region}}\mathbf{K}_{\text{clip}}^{\top}}{\sqrt{d}}\right)\mathbf{V}_{\text{clip}}, \quad (5)$$

where:

$$\begin{aligned}\mathbf{Q}_{\text{region}} &= \mathbf{F}_{\text{region}},\\ \mathbf{K}_{\text{clip}} &= \mathbf{F}_{\text{clip}}^{\text{concat}},\\ \mathbf{V}_{\text{clip}} &= \mathbf{F}_{\text{clip}}^{\text{concat}}.\end{aligned} \quad (6)$$

Here, the **query** attends to preselected anatomical regions, ensuring that **keys** and **values** represent contextualized visual features. This enriched representation $\mathbf{F}_{\text{rg}}$ encodes spatially guided semantic information for improved report generation.

3.3 Region-Text Alignment via Cross Attention

To align textual features with the region-guided embeddings, we integrate an additional cross-attention (CA_2) into each encoder of BioMIC-BART. Given textual token embeddings $\mathbf{H}_T \in \mathbb{R}^{M\times d}$ from the MeSH or NegBio tagger (Kale et al., 2023; Peng et al., 2018), and region-guided image embeddings $\mathbf{F}_{\text{rg}}$, CA_2 is computed as:

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{H}_T\mathbf{W}_Q(\mathbf{F}_{\text{rg}}\mathbf{W}_K)^{\top}}{\sqrt{d}}\right)\mathbf{F}_{\text{rg}}\mathbf{W}_V, \quad (7)$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d\times d}$ are trainable projection matrices.

This operation enhances textual representations by grounding them in localized visual features, ensuring alignment with relevant anatomical regions. The enriched embeddings are then processed through subsequent layers of BioMIC-BART, including *Multi-Head Self-Attention*, *Layer Normalization*, and *Feed-Forward Networks*, with residual connections ensuring stability. The decoder then generates the final report $\hat{\mathbf{T}}$, selecting the most probable candidate sequence $\mathbf{T}'$ from the distribution:

$$\hat{\mathbf{T}} = \arg\max_{\mathbf{T}'} P(\mathbf{T}'|\mathbf{H}_T, \mathbf{F}_{\text{rg}}; \theta), \quad (8)$$

4 Experiments and Results

We evaluated LRTA-BioMIC with various architectural modifications and benchmarked it against OpenAI's GPT-4o (Achiam et al., 2023), Google's

Gemini (Team et al., 2023) (refer Section 10), and previous models—RGRG (Tanida et al., 2023), COMG (Gu et al., 2024), PromptMRG (Jin et al., 2024), ORGan (Hou et al., 2023b), RECAP (Hou et al., 2023a), and KGVL-BART (Kale et al., 2023)—on both *IU-Xray* and *MIMIC-CXR-JPG*.

LRTA-BioMIC outperformed the prior state-of-the-art, achieving **9.71%** and **0.9%** improvements in *BLEU-4* and *ROUGE-L* on *IU-Xray* (compared to COMG and KGVL-BART). On *MIMIC-CXR-JPG*, it showed a **1.60%** increase in *BLEU-4* but a slight **3.53%** drop in *ROUGE-L* (compared to RECAP and ORGan). Additionally, it achieved a **3.32%** improvement in the Clinical Efficacy (CE) F1-score (CheXbert (Smit et al., 2020)) on *IU-Xray* and performed best on *MIMIC-CXR-JPG*, except for a **3.70%** decrease compared to PromptMRG. Evaluation metrics are detailed in Section 9, while model limitations are discussed in Section 13.

LRTA-BioMIC performed relatively better on *IU-Xray* due to its backbone, *BioMIC-BART*, which is pre-trained on *MIMIC-CXR-JPG*, enriching it with medical terminology and multimodal processing capabilities. When fine-tuned on *IU-Xray*, it leverages prior exposure to a larger dataset, enhancing performance. Inspired by KM-BART (Xing et al., 2021), BioMIC-BART details are in Section 11. Below, we outline our architectural modifications for ablation studies in report generation.

- $LRTA - BioMIC_1$: Removed the *Region Guided Feature Extractor* while retaining all other components.
- $LRTA - BioMIC_2$: Ablated Cross-Attention in the Encoder, using a direct addition of embeddings from the *Region Guided Feature Extractor* and BERT *Tagger* embeddings.
- $LRTA - BioMIC_3$: Removed *BioMIC-BART* and used the original BART from Facebook (Lewis, 2019).
- $LRTA - BioMIC$: Our final report generation architecture as shown in Figure 1.

As shown in Table 1, removing the *Region Guided Feature Extractor* ($LRTA - BioMIC_1$) led to an **22.12%** and **16.52%** decrease in BLEU-4 and ROUGE-L score from our SoTA model, $LRTA - BioMIC$, highlighting the importance of extracting features from 29 specific chest X-ray regions (Tanida et al., 2023). Replacing Cross-Attention with a simple addition of embeddings ($LRTA - BioMIC_2$) reduced the BLEU-4 and ROUGE-L score by **6.64%** and **4.69%**, this underscores the value of effective embedding integration. In $LRTA - BioMIC_3$, replacing *BioMIC-BART* with Facebook's original BART (Lewis, 2019) resulted in a decline of **11.95%** and **10.27%** in BLEU-4 and ROUGE-L, demonstrating the need for domain-specific radiology context along with diverse medical terminology through fine-tuning on PubMed texts. All the other metrics also demonstrated a consistent boost in our final architecture, $LRTA - BioMIC$ (c.f Table 2, Section 7).

4.1 Computational Resources

Experiments were conducted using *A100 GPUs*. *BioMIC-BART* training required *four A100 GPUs* (80GB each) and took approximately *26* hours. *LRTA-BioMIC* fine-tuning on *MIMIC-CXR-JPG* and *IU-Xray* was significantly lightweight, running on a single GPU with just *6GB to 7GB* of memory. Fine-tuning took only *4.5* hours for *MIMIC-CXR-JPG* and *1.5* hours for *IU-Xray*, highlighting its efficiency (c.f. Section 12).

5 Conclusion and Future Work

In place of computationally intensive VLMs, we propose **LRTA-BioMIC**, a computationally efficient, domain-specific, region-guided, and text-aligned language model with ViT, achieving SoTA Chest X-ray report generation. We extend **BioBART**, originally trained on full PubMed texts, by further training it on *MIMIC-CXR-JPG* to enable efficient multimodal processing, naming it **BioMIC-BART**. Our approach improves *BLEU-4* and *ROUGE-L* by **9.71%** and **0.9%** on *IU-Xray*, and by **1.60%** in *BLEU-4* on *MIMIC-CXR-JPG*, with a slight **3.53%** decrease in *ROUGE-L* compared to prior SoTA models. In future work, we will explore transfer learning, augmentation, and in-context learning to improve adaptability to small, long-tail imbalanced datasets and varying clinical settings. Additionally, incorporating reports from other radiology domains, such as CT, MRI, and X-ray of different organs, may enhance the model's understanding of medical language and structural patterns, leading to more accurate and context-aware report generation.

6 Limitations

The IU Chest X-ray and MIMIC-CXR-JPG datasets (c.f Section 8) provide publicly available chest X-ray images paired with radiology reports, though access to MIMIC-CXR-JPG is restricted due to privacy regulations such as HIPAA. Annotating medical reports is costly and requires domain expertise, limiting the availability of large-scale datasets for research. MIMIC-CXR-JPG primarily includes ICU patients, potentially skewing models toward severe disease cases. Another limitation is that our method evaluates chest X-rays in isolation, whereas clinical assessments often compare them with prior scans for a more comprehensive diagnosis. Moreover, MIMIC-CXR-JPG contains descriptions of non-anatomical objects, such as surgical clips, which are not addressed by our approach. Lastly, while our framework is tailored for radiology report generation from chest X-rays, expanding it to other imaging modalities, such as CT or MRI, remains an important future direction.

7 Ethical Considerations

The authors of both the *IU X-ray* (Demner-Fushman et al., 2016) and the *MIMIC-CXR-JPG* (Johnson et al., 2019) dataset have implemented techniques for de-identifying patient information. Both datasets ensure that data is anonymized, which protects patient identity and adheres to ethical standards in healthcare research. This comprehensive de-identification process allows our model to operate without disclosing any sensitive information regarding individual patients. *BioMIC-BART* is trained over BART. While Pre-trained Language Models (PLMs) like BART are advantageous for various natural language processing tasks, they can introduce biases present in their training corpora (Gallegos et al., 2023; Navigli et al., 2023). Despite efforts to mitigate bias, it is challenging to completely eliminate biased or discriminatory content in the model's representations.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Elmira Amirloo, Jean-Philippe Fauconnier, Christoph Roesmann, Christian Kerl, Rinu Boney, Yusu Qian, Zirui Wang, Afshin Dehghan, Yinfei Yang, Zhe Gan, and Peter Grasch. 2024. Understanding alignment in multimodal llms: A comprehensive study. *ArXiv*, abs/2407.02477.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. 2022. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pages 1–21. Springer.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. 2024. The revolution of multimodal large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13590–13618, Bangkok, Thailand. Association for Computational Linguistics.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

Gabriel Forgues, Joelle Pineau, Jean-Marie Larchevêque, and Réal Tremblay. 2014. Bootstrapping dialog systems with word embeddings. In *Nips, modern machine learning and natural language processing workshop*, volume 2, page 168.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2023. Bias and fairness in large language models: A survey. *arXiv preprint arXiv:2309.00770*.

Tiancheng Gu, Dongnan Liu, Zhiyuan Li, and Weidong Cai. 2024. Complex organ mask guided radiology report generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7995–8004.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Wenjun Hou, Yi Cheng, Kaishuai Xu, Wenjie Li, and Jiang Liu. 2023a. Recap: Towards precise radiology report generation via dynamic disease progression reasoning. *arXiv preprint arXiv:2310.13864*.

Wenjun Hou, Kaishuai Xu, Yi Cheng, Wenjie Li, and Jiang Liu. 2023b. Organ: observation-guided radiology report generation via tree reasoning. *arXiv preprint arXiv:2306.06466*.

Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. 2024. Promptmrg: Diagnosis-driven prompts for medical report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2607–2615.

Baoyu Jing, Zeya Wang, and Eric Xing. 2020. Show, describe and conclude: On exploiting the structure information of chest x-ray reports. *arXiv preprint arXiv:2004.12274*.

Baoyu Jing, Pengtao Xie, and Eric Xing. 2017. On the automatic generation of medical imaging reports. *arXiv preprint arXiv:1711.08195*.

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. 2019. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.

Kaveri Kale, Pushpak Bhattacharyya, Milind Gune, Aditya Shetty, and Rustom Lawyer. 2023. Kgvl-bart: Knowledge graph augmented visual language bart for radiology report generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3401–3411.

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

Thomas K Landauer and Susan T Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

M Lewis. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Yaowei Li, Bang Yang, Xuxin Cheng, Zhihong Zhu, Hongxiang Li, and Yuexian Zou. 2023. Unify, align and refine: Multi-level semantic alignment for radiology report generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2863–2874.

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

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: origins, inventory, and discussion. *ACM Journal of Data and Information Quality*, 15(2):1–21.

Aaron Nicolson, Jinghui Liu, Jason Dowling, Anthony Nguyen, and Bevan Koopman. 2024. e-health csiro at rrg24: Entropy-augmented self-critical sequence training for radiology report generation. *arXiv preprint arXiv:2408.03500*.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24.

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.

Yifan Peng, Xiaosong Wang, Le Lu, Mohammadhadi Bagheri, Ronald Summers, and Zhiyong Lu. 2018. Negbio: a high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Summits on Translational Science Proceedings*, 2018:188.

Han Qin and Yan Song. 2022. Reinforced cross-modal alignment for radiology report generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 448–458.

Vasile Rus and Mihai Lintean. 2012. An optimal assessment of natural language student input using word-to-word similarity metrics. In *Intelligent Tutoring Systems: 11th International Conference, ITS 2012, Chania, Crete, Greece, June 14-18, 2012. Proceedings 11*, pages 675–676. Springer.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren. 2020. Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert. *arXiv preprint arXiv:2004.09167*.

Xiao Song, Xiaodan Zhang, Junzhong Ji, Ying Liu, and Pengxu Wei. 2022. Cross-modal contrastive attention model for medical report generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2388–2397.

Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. 2023. Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7433–7442.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*.

Yiran Xing, Zai Shi, Zhao Meng, Gerhard Lakemeyer, Yunpu Ma, and Roger Wattenhofer. 2021. Km-bart: Knowledge enhanced multimodal bart for visual commonsense generation. *arXiv preprint arXiv:2101.00419*.

Shuxin Yang, Xian Wu, Shen Ge, S Kevin Zhou, and Li Xiao. 2022. Knowledge matters: Chest radiology report generation with general and specific knowledge. *Medical image analysis*, 80:102510.

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022. Biobart: Pretraining and evaluation of a biomedical generative language model. *arXiv preprint arXiv:2204.03905*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Yuille, and Daguang Xu. 2020. When radiology report generation meets knowledge graph. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12910–12917.

Appendix

8 Dataset

The **MIMIC-CXR-JPG** dataset (Johnson et al., 2019) and the **IU-Xray** dataset (Demner-Fushman et al., 2016) are among the most reliable and widely used benchmarks in radiology report generation research. These datasets have been extensively utilized in previous works due to their high-quality imaging studies and corresponding radiology reports. **MIMIC-CXR-JPG** includes 227,835 imaging studies from 65,379 patients treated at Beth Israel Deaconess Medical Center's Emergency Department from 2011 to 2016, providing a total of 377,110 chest X-ray images along with free-text de-identified radiology reports. **IU-Xray**, while comparatively smaller in size with 7,470 chest X-ray images and 3,825 patient reports, offers certain advantages. Unlike MIMIC-CXR-JPG, which contains unstructured free-text reports, IU-Xray follows a structured template consisting of two key sections: *Findings*, which provides a detailed description of the radiograph, and *Impression*, which serves as a summary or inference of the report. Additionally, IU-Xray is balanced in terms of normal and abnormal reports, making it a valuable dataset for evaluating model performance across different case distributions. Given their significance in the field, we have utilized both datasets in our research to ensure robustness and comparability with existing methods.

9 Evaluation Metrics

In our evaluation process, we employed several metrics, including BLEU (Papineni et al., 2002), CIDEr (Vedantam et al., 2015), METEOR (Banerjee and Lavie, 2005), BERTScore (Zhang et al., 2019), ROUGE-L (Lin, 2004), and Embedding-Based Metrics (Rus and Lintean, 2012; Landauer and Dumais, 1997; Forgues et al., 2014). *BLEU* effectively measures translation quality by comparing n-grams from the generated outputs with reference translations. *CIDEr* emphasizes capturing the consensus between human judgments and model predictions by quantifying n-gram overlaps. *METEOR* improves robustness to lexical variations by considering both precision and recall through stemming and synonyms. *BERTScore* utilizes contextual embeddings to evaluate fluency and coherence by assessing semantic similarities between generated texts and references. *ROUGE-L* specifically evaluates summarization quality by measuring the longest common subsequence (LCS) between generated summaries and reference summaries. Embedding-Based Metrics assess semantic similarities between generated and reference outputs.

While NLG metrics are widely used and reliable for report generation evaluation, they do not capture all clinically relevant aspects of the gener-

Model	B-1	B-2	B-3	B-4	Cider	MTR	Dist-2	BertScore	Rouge-L	E-avg
GPT-4o	0.183	0.070	0.032	0.002	-	0.287	0.349	0.628	0.187	0.934
Gemini	0.176	0.072	0.027	0.001	-	0.204	**0.383**	0.582	0.173	0.916
$LRTA - BioMIC_1$	0.398	0.274	0.213	0.176	0.888	0.412	0.317	0.812	0.374	0.946
$LRTA - BioMIC_2$	0.483	0.359	0.275	0.211	0.974	0.510	0.339	**0.902**	0.427	0.962
$LRTA - BioMIC_3$	0.462	0.339	0.257	0.199	0.934	0.498	0.324	0.871	0.402	0.963
$LRTA - BioMIC$	**0.527**	**0.384**	**0.279**	**0.226**	**1.013**	**0.522**	0.347	0.898	**0.448**	**0.969**

Table 2: Performance comparison of $LRTA - BioMIC$ against multiple Ablation architecture (c.f section 4), GPT-4o and Gemini across multiple evaluation metrics. $LRTA - BioMIC$ achieves the highest scores in most metrics, outperforming state-of-the-art vision-language models. B-i represents BLEU scores with i-gram overlap, ROUGE-L denotes the longest common subsequence measure, MTR refers to the METEOR score, Dist-2 indicates distinct bigram diversity, and E-avg represents the average embedding-based metric.

ated reports. To address this limitation, we adopt *CheXbert* (Smit et al., 2020) to label the generated reports and compare them with the disease labels from the reference reports. Due to space constraints and the fact that previous works have omitted several NLG metrics, we provide a detailed breakdown of all NLG metric results in the appendix to facilitate further comparison of ablation studies, mentioned in the Table 2.

10 Comparision with GPT-4o and Gemini

We evaluated our model with various architectural modifications and benchmarked it against OpenAI's GPT-4o (Achiam et al., 2023) and Google's Gemini (Team et al., 2023). The prompt provided was: *"The bot is given a chest X-ray image and must generate a report consisting of Findings and Impression. Findings provide a detailed description of the radiograph, while Impression serves as a summary or inference of the report."*

The results are presented in Table 2. We observed an improvement of **139.57%**, **158.96%** in ROUGE-L and **42.99%**, **54.30%** in BERTScore when comparing LRTA-BioMIC to GPT-4o and Gemini. Although the BLEU score is significantly lower for GPT-4o and Gemini, their BERTScore remains decent. Notably, Gemini achieved an **10.37%** higher Distinct-2 score than LRTA-BioMIC; however, a better Distinct-2 score does not necessarily indicate superior performance. In medical report generation, excessive diversity can lead to incoherence, inconsistency, and potential loss of medical accuracy, as reports often require standardized phrasing and necessary repetitions. In the future, we would like to see more studies exploring few-shot learning and in-context learning with additional experiments.

11 BioMIC-BART

Figure 2 illustrates the architecture of our **BioMIC-BART**, which is built upon BioBART (Yuan et al., 2022), a model trained on full PubMed texts. While BioBART is rich in general medical contexts, it lacks specialized, refined knowledge of chest X-rays and their associated conditions. To address this limitation, we draw inspiration from (Xing et al., 2021), which extended the BART model to process multimodal data comprising images and text. Their dataset includes Conceptual Captions (Sharma et al., 2018), SBU (Ordonez et al., 2011), COCO (Lin et al., 2014), and Visual Genome (Krishna et al., 2017). We give the details of our visual feature extractor, What are the tokens we use and the encoder decoder.

11.1 Visual Feature Extractor

Following previous work on Vision Transformers, we use MedCLIP (Wang et al., 2022), pretrained on the MIMIC-CXR-JPG chest X-ray image and report pair dataset, to extract visual embeddings. These embeddings are then fed into the Transformer-based cross-modal encoder. We include both the Posteroanterior (PA) and Lateral (LL - Lateral View) images, if available, to provide BioMIC-BART with contextual information from multiple perspectives. The PA view is the standard frontal chest X-ray, while the LL view offers a side perspective, helping to better assess the depth and localization of abnormalities. Using both views enhances the model's understanding of anatomical structures and improves diagnostic accuracy.

11.2 Token Embeddings

We utilize *CXR-BERT-general* (Boecking et al., 2022), a domain-specific language model tailored on chest X-ray (CXR) reports. It is pretrained

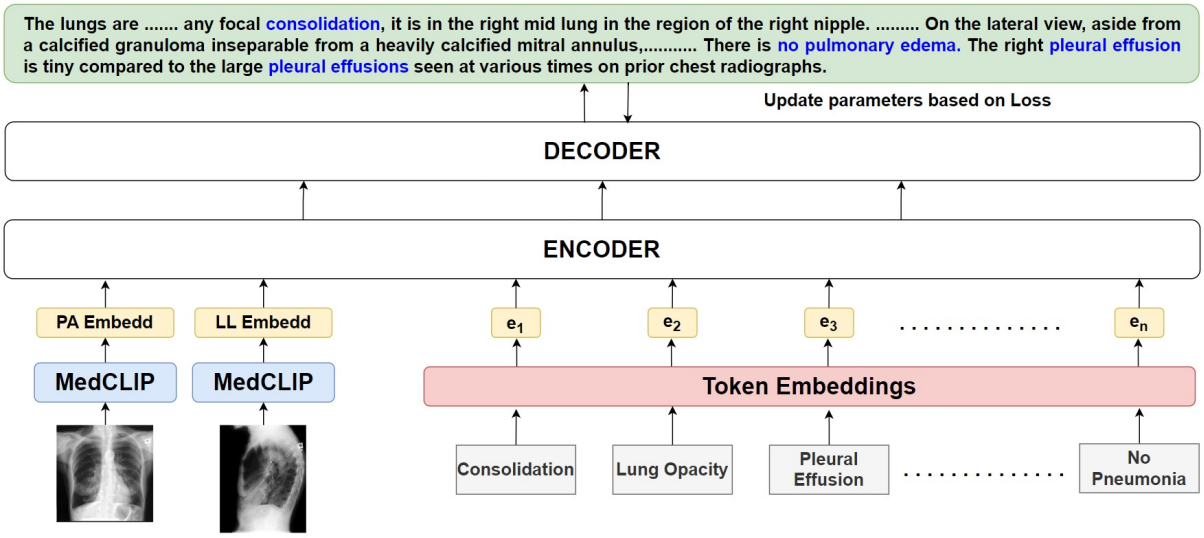

Figure 2

from a randomly initialized BERT model using Masked Language Modeling (MLM) on PubMed abstracts and clinical notes from the publicly available MIMIC-III and MIMIC-CXR-JPG datasets. This model extracts token embeddings, where the tokens are expert-annotated medical tags inherent to the dataset. Combined with X-ray image embeddings from MedCLIP, these token representations enhance the model's ability to capture a richer contextual understanding of multimodal radiology Chest X-ray data.

11.3 Encoder-Decoder

The model architecture consists of 12 encoder-decoder layers designed to effectively process and integrate multimodal data. The encoder receives both image embeddings, extracted from Posteroanterior (PA) and Lateral (LL) chest X-ray views using MedCLIP, and token embeddings, derived from chest X-ray tags using CXR-BERT-general. The entire model is trained on the official MIMIC-CXR-JPG train split. The model parameters are updated based on the loss calculated during training, which measures the discrepancy between the predicted and actual diagnostic outcomes. This loss is backpropagated through the network, adjusting the weights of both the encoder and decoder to minimize error and improve the model's performance.

Although a simple model like this alone cannot produce meaningful radiology reports on unseen data, transferring the contextual multimodal understanding of BioMIC-BART to our architecture, LRTA-BioMIC, as illustrated in Figure 1, enhances performance compared to using BART alone (Lewis, 2019) (refer to Section 4).

12 Parameter and Computational Resources

We divide our experiments into two parts. The first set of experiments involves *BioMIC-BART*, while the second focuses on *LRTA-BioMIC* using both the *MIMIC-CXR-JPG* and *IU-Xray* datasets. Experiments with *BioMIC-BART* are computationally expensive, whereas all experiments with *LRTA-BioMIC* our final architecture, a domain-specific region-guided language model combined with ViT are computationally efficient. The common parameters across both experiments include the *GELU* activation function and the *Adam optimizer*. Additionally, a weight decay of *0.001* was applied for regularization. All experiments were conducted using one or more *A100 GPUs*.

12.1 BioMIC-BART

We conducted a grid search to determine the optimal hyperparameters. Among the tested learning rates (*3e-4*, *3e-5*, and *3e-6*), we found *3e-5* to yield the best performance. Similarly, we evaluated batch sizes of *48* and *64*, with a batch size of *48* performing better over *20* epochs. The training split followed the official *MIMIC-CXR-JPG* partition, which we further subdivided into a *90-10* split: *90%* of the training data was used for pretraining *BioMIC-BART*, while the remaining *10%* was allocated for fine-tuning the *LRTA-BioMIC* architecture with a random seed of *42*. The experiments were conducted on *four A100 GPUs*, each with

80GB of memory. Each training run took approximately *26* hours to complete.

12.2 LRTA-BioMIC

We conducted a grid search to determine the optimal hyperparameters. Among the tested learning rates (*3e-4* and *3e-5*), we found *3e-5* to yield the best performance. Similarly, we evaluated batch sizes of *4*, *8*, and *12* over *20* epochs and found that a batch size of *4* performed the best. For fine-tuning on the *MIMIC-CXR-JPG* dataset, we randomly selected *10%* of the official training split using a seed of *42*, while the official test split was used for evaluation. For fine-tuning on the *IU-Xray* dataset, since there is no official train-validation-test distribution, we partitioned the data into *80%*, *10%*, and *10%* splits, respectively, using a random seed of *42*. Fine-tuning on *MIMIC-CXR-JPG* required *7GB* of GPU memory, whereas *IU-Xray* required *6GB*, both with a batch size of *4*. The additional *1GB* of GPU memory for *MIMIC-CXR-JPG* was due to its larger training set. Training on *MIMIC-CXR-JPG* took approximately *4.5* hours, while training on *IU-Xray* was significantly faster, requiring only about *1.5* hours.

13 Error Analysis

We conducted an analysis to identify weaknesses in LRTA-BioMIC. We identified two key weaknesses: **Numerical Discrepancies (Weakness-A).** In Table 3, we observe that the gold report mentions an **8mm nodule**, whereas the generated report states a **1cm nodule**. Although the difference is small, in a sensitive domain like healthcare, even minor inaccuracies can be critical. Similarly, in the second gold report under the same limitation, our model, LRTA-BioMIC, correctly identified the spatiality by mentioning **healed left rib fractures** but failed to specify the **9th rib fracture**, which was explicitly mentioned in the gold report. **Incomplete Transfer of Findings to Impression (Weakness-B):** In Table 3, at first glance, it may appear that our model underperforms due to missing details in the impression section. However, upon closer inspection, we observe that LRTA-BioMIC correctly generates the relevant observations and includes them in the findings section but omits them in the impression. This behavior is influenced by the inherent bias in the *MIMIC-CXR-JPG* and *IU-Xray* datasets, where findings are typically detailed, while impressions tend to be concise. Due to the brevity of the impression section, the model sometimes hallucinates by omitting crucial conclusions that should be included in the impression while only mentioning them in the findings.

Weaknesses	GOLD REPORT	GENERATED REPORT
Weakness-A	Findings: The cardiac and mediastinal contours are within normal limits. The lungs are well-inflated and clear. There is an **8mm nodule in the left lower lobe**, XXXX calcified granuloma. There is no pneumothorax or effusion. Bony structures of the thorax are intact with minimal early degenerative change. Impression: 1. No evidence of acute cardiopulmonary process. 2. 8mm nodule in the left lung base, XXXX calcified granuloma.	Findings: Heart size and pulmonary vascularity appear within normal limits. The lungs are free of focal airspace disease. No pleural effusion or pneumothorax is seen. In the **left lower lobe, there has been development of a 1 cm diameter nodule**. This appears somewhat XXXX from the prior study. Mild degenerative changes are present in the XXXX. Impression: 1 cm of left lung base nodule. XXXX since the previous study.
	Findings: Stable cardiomediastinal silhouette. Pulmonary vascularity is within normal limits. Hyperlucent apices. Negative for focal airspace disease or consolidation. Negative for pneumothorax or pleural effusion. **Healed remote left 9th rib fracture.** Right shoulder: Negative for fracture or dislocation. Impression: Chest. Right shoulder. 1. No acute cardiopulmonary abnormality. 2. Negative for right shoulder fracture or dislocation.	Findings: The lungs are clear bilaterally. Specifically, no evidence of focal consolidation, pneumothorax, or pleural effusion. Cardiomediastinal silhouette is unremarkable. Stable apical hyperlucency. Visualized osseous structures of the thorax are without acute abnormality. **Healed left rib fractures are noted.** Impression: No acute cardiopulmonary abnormality.
Weakness-B	Findings: The heart is normal in size. The mediastinum is stable. Left-sided chest XXXX is again visualized with tip at cavoatrial junction. There is no pneumothorax. Numerous bilateral pulmonary nodules have increased in size and number XXXX compared to prior study. The dominant nodule/mass in the left midlung is also mildly increased. There is no pleural effusion. **Impression: Interval increase in size and number of innumerable bilateral pulmonary nodules consistent with worsening metastatic disease.**	Findings: Heart size within normal limits, stable mediastinal and hilar contours. Left-sided subclavian central venous catheter. No focal alveolar consolidation, no definite pleural effusion seen. **There are multiple bilateral pulmonary nodules. Prominent left midlung nodule may be secondary to overlapping structures.** No typical findings of pulmonary edema. Impression: No acute findings
	Findings: There is a calcified granuloma in the left upper lobe. Lungs otherwise are believed to be clear. The heart is normal. There are calcified left hilar and mediastinal lymph XXXX. The skeletal structures show some senescent changes. **Impression: Old granulomatous disease. No acute pulmonary disease.**	Findings: **There is a calcified granuloma in the left upper lobe.** There are calcified left hilar and mediastinal lymph XXXX, consistent with prior granulomatous disease. The lungs are otherwise clear without evidence of acute infiltrate or effusion. Impression: No acute cardiopulmonary abnormalities.

Table 3: Table highlighting model weaknesses by presenting two examples per weakness, comparing the gold reports with the generated reports.