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.

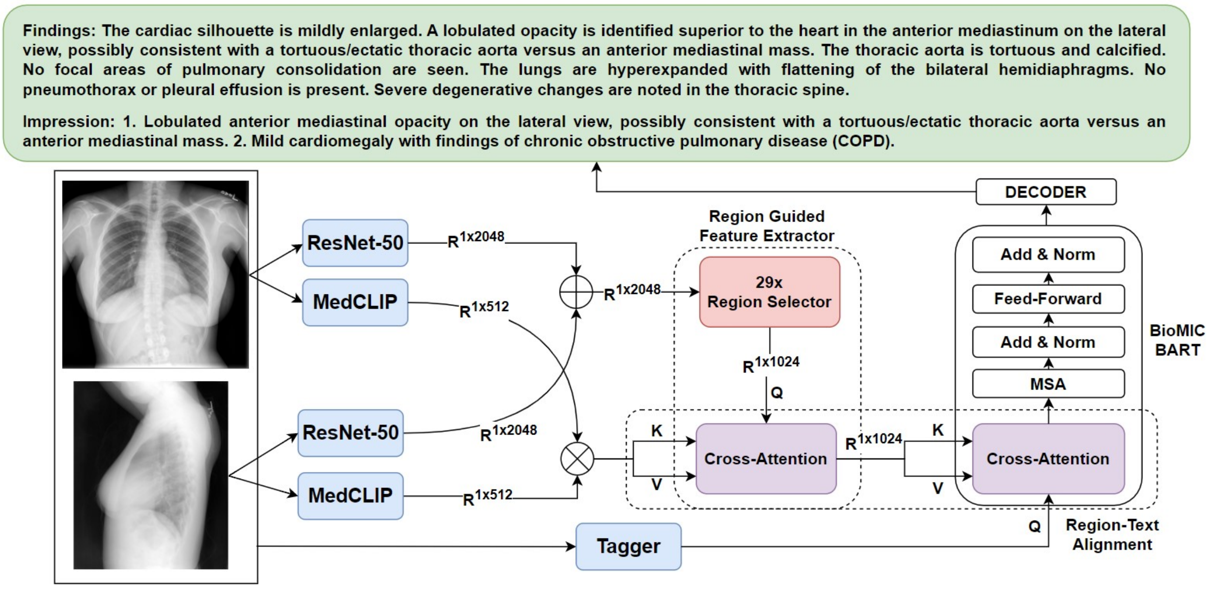
1 Introduction

Vision-Language Models (VLMs) are widely used in radiology report generation due to their ability to generate coherent text from images. However, existing pipelines often suffer from poor image-text alignment (Amirloo et al., 2024), which affects generation quality. Prior work (Caffagni et al., 2024) has shown that improving alignment enhances performance. Moreover, many VLMs rely on heavy pre-trained encoders and decoders, limiting their practicality. To address these limitations, we propose ***L**ightweight **R**egion-**T**ext **A**ligned **BioMIC**-BART* (**LRTA-BioMIC**)—a model that efficiently generates chest X-ray reports. We extend BioBART (Yuan et al., 2022), which lacks image embedding capability, by training it on the *MIMIC-CXR-JPG* dataset using KM-BART's dual-stream training (Xing et al., 2021). The resulting **BioMIC-BART** forms the backbone of LRTA-BioMIC (cf. Table 2), enhancing performance on *IU-Xray* and *MIMIC-CXR-JPG*. Our model uses region-guided features from MedCLIP (Wang et al., 2022), refined via the Region Selector from (Tanida et al., 2023) with cross-attention (CA_1), to enhance contextual visual embeddings. These are further aligned with textual tags through a second cross-attention (CA_2) in BioMIC-BART layers, improving region-text coherence during encoding.

Earlier methods in this domain ranged from CNN-RNN pipelines (Jing et al., 2020, 2017) to Transformer-based models (Vaswani, 2017). Region-aware methods (Tanida et al., 2023; Li et al., 2023) improved alignment, while organ-specific masks (Gu et al., 2024) and observation-guided reasoning (Hou et al., 2023b,a) boosted disease detection. Knowledge graphs (Zhang et al., 2020; Kale et al., 2023) and prompt-based techniques (Jin et al., 2024) further enriched text generation. Despite these advances, existing models like CMCA (Song et al., 2022), KnowMat (Yang et al., 2022), and CMM-RL (Qin and Song, 2022) remain resource-intensive or alignment-limited. LRTA-BioMIC overcomes these drawbacks by combining efficient multimodal processing with improved alignment via region selection and text grounding.

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*.



Findings: The cardiac silhouette is mildly enlarged. A lobulated opacity is identified superior to the heart in the anterior mediastinum on the lateral view, possibly consistent with a tortuous/ectatic thoracic aorta versus an anterior mediastinal mass. The thoracic aorta is tortuous and calcified. No focal areas of pulmonary consolidation are seen. The lungs are hyperexpanded with flattening of the bilateral hemidiaphragms. No pneumothorax or pleural effusion is present. Severe degenerative changes are noted in the thoracic spine.

Impression: 1. Lobulated anterior mediastinal opacity on the lateral view, possibly consistent with a tortuous/ectatic thoracic aorta versus an anterior mediastinal mass. 2. Mild cardiomegaly with findings of chronic obstructive pulmonary disease (COPD).

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.

2 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 ***L**ightweight **R**egion-**T**ext **A**ligned **BioMIC**-BART* (**LRTA-BioMIC**), which incorporates region-level visual features and enhances text-image alignment.

2.1 BioMIC-BART

We build upon *BioBART-Large*, a 442M-parameter 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 limited 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 from (Xing et al., 2021), which effectively model image-text contextual relations. Details in Section 11.

2.2 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:

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

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

For comprehensive feature fusion, we compute:

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

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

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

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

where:

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

Here, the **query** attends to preselected anatomical regions, ensuring that **keys** and **values** represent

Dataset	Model	NLG Metrics						CE Metrics		
		B-1	B-2	B-3	B-4	MTR	R-L	P	R	F_1
MIMIC -CXR -JPG	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	0.156	0.161	0.149
	LRTA-BioMIC$_2$	0.483	0.359	0.275	0.211	0.510	0.427	0.204	0.207	0.202
	LRTA-BioMIC$_3$	0.462	0.339	0.257	0.199	0.498	0.402	0.197	0.203	0.195
	LRTA-BioMIC$_4$	0.464	0.345	0.265	0.203	0.516	0.414	0.202	0.203	0.199
	LRTA-BioMIC	0.527	0.384	0.279	0.226	0.522	0.448	0.221	0.223	0.218

Table 1: Experimental Results of our model and baselines on the IU-Xray 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-Xray dataset.A comprehensive one-tailed t-test analysis between *LRTA-BioMIC* and all five baselines across three key metrics—BLEU-4, ROUGE-L, and Clinical Efficacy—on both datasets was conducted (30 tests total) to validate the model's effectiveness (see Section 14).

Aspect	**BioMIC-BART**	**LRTA-BioMIC**
Base Model	BioBART	BioMIC-BART
Training Cost	High	Low to Moderate
Inference	Efficient	Efficient
RGFE	Not present	Present
RTA	Not present	Present
Purpose	Serves as backbone weight for LRTA-BioMIC	Designed for generation of report
Objective	Tuned to process multi-modal chest X-ray image and text	Generate reports via lightweight region-text aligned model
Report Suitability	Not suitable alone to generate report	Suitable alone to generate report
Visualization	cf. Figure 2	cf. Figure 1

Table 2: Distinction between BioMIC-BART and LRTA-BioMIC. RGFE: Region-Guided Feature Extraction; RTA: Region-Text Aligner.

contextualized visual features. This enriched representation $\mathbf{F}_{\text{rg}}$ encodes spatially guided semantic information for improved report generation.

2.3 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) for IU-Xray, and from (Alfarghaly et al., 2021)[1], for MIMIC-CXR-JPG. , and region-guided image embeddings $\mathbf{F}_{\text{rg}}$, CA_2 is computed as:

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

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 generates the final report by selecting the most likely sequence, using contextual features from textual embeddings $\mathbf{H}_T$ and region-guided visuals $\mathbf{F}_{\text{rg}}$, already integrated via cross-attention (CA_2).

[1] Tags were generated by applying the pre-trained models directly to the images. https://www.kaggle.com/datasets/tasmiarahmanaanika/automated-radiology-105-tags

3 Experiments and Results

We evaluated LRTA-BioMIC with architectural variations and benchmarked it against GPT-4o (Achiam et al., 2023), Gemini (Team et al., 2023) (see Section 9), prior 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 *IU-Xray* and *MIMIC-CXR-JPG*, and the CheXpert dataset (Section 10). LRTA-BioMIC outperformed previous SoTA with **+9.71% BLEU-4** and **+0.9% ROUGE-L** on *IU-Xray* (vs. COMG, KGVL-BART). On *MIMIC-CXR-JPG*, it achieved **+1.60% BLEU-4** but **-3.53% ROUGE-L** (vs. RECAP, ORGan). For Clinical Efficacy F1 (CheXbert (Smit et al., 2020)), it improved by **+3.32%** on *IU-Xray* and was best on *MIMIC-CXR-JPG* except for a **-3.70%** drop vs. PromptMRG. Evaluation metrics and error analysis are in Sections 8 and 13, respectively.

LRTA-BioMIC Superior Performance on the IU-Xray Dataset. While training and inferencing on IU X-ray, the model benefits from joint exposure to the MIMIC-CXR-JPG dataset from the backbone BioMIC BART, which is over 30 times larger and provides more diverse sample of radiology reports. This auxiliary supervision helps learn generalized parameters that transfer well to the smaller IU dataset. However, when evaluating on MIMIC, there is no such external source to support training, making it a more difficult generalization challenge. Replacing *BioMIC-BART* with BART, lacks radiology-specific pretraining and access to MIMIC during IU training, results in a notable drop of **11.95%** in BLEU-4 and **10.27%** in ROUGE-L. Since no available dataset is fully unbiased or clinically exhaustive (Song et al., 2024), leveraging complementary datasets can enhance robustness. While one may consider augmenting IU data to improve MIMIC performance, the small size of IU limits its effectiveness in adapting model parameters for the much larger MIMIC test set. Below, we present our ablation studies.

- $LRTA - BioMIC_1$: Removed the *Region Guided Feature Extractor*.
- $LRTA\text{-}BioMIC_2$: Ablated Region-Text Aligner.
- $LRTA - BioMIC_3$: Replaced *BioMIC-BART* with the original BART (Lewis, 2019).
- $LRTA - BioMIC_4$: Replaced *BioMIC-BART* with the BioBART.(Yuan et al., 2022).
- $LRTA - BioMIC$: Our final report generation architecture as shown in Figure 1.

As shown in Table 1, removing the *Region Guided Feature Extractor* ($LRTA\text{-}BioMIC_1$) caused a sharp **22.12%** and **16.52%** drop in BLEU-4 and ROUGE-L, confirming the importance of extracting features from 29 chest X-ray regions (Tanida et al., 2023). Replacing cross-attention with simple embedding addition ($LRTA\text{-}BioMIC_2$) reduced BLEU-4 and ROUGE-L by **6.64%** and **4.69%**, showing the need for effective fusion. Using vanilla BART ($LRTA\text{-}BioMIC_3$) led to a **11.95%** and **10.27%** drop, while swapping in BioBART ($LRTA\text{-}BioMIC_4$) gave only minor gains of **0.2%** and **2.9%**. This highlights the need for radiology-specific tuning beyond generic biomedical pretraining. Other metrics also favor $LRTA$-BioMIC (c.f. Table 3, Section 6).

3.1 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, 15).

4 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, a full-fledged systematic study of various data configuration strategies including transfer learning and dataset augmentation could be helpful to improve performance and generalization.

5 Limitations

The IU Chest X-ray and MIMIC-CXR-JPG datasets (cf. Section 7) 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.

5.1 Bias within Training Data

From a machine learning perspective, dataset bias can affect even a "perfect" model because such bias originates from the data itself, not the model's architecture or training process. If a dataset is biased—such as being skewed or containing spurious correlations—then the model trained on it will inherently reflect these biases in its predictions, regardless of its accuracy or sophistication (ari, 2023; Bourgin and Peterson, 2024; Haider, 2024). We focused on co-occurrences of the X-ray diagnosis and the long-tail issue of the dataset for our analysis, as mentioned in (Song et al., 2024). Instead of identifying critical disease features, models may infer the attributes of one disease solely by relying on the presence of others, which confuses the recognition of visual realities and the generation of accurate reports. In Figures 4 and 5, we observe that the imbalance in the data can lead to biased predictions and poor performance on minority classes. IU is more skewed, while MIMIC is less so, but both depict long-tail issues arising from dataset bias. For co-occurrences of the X-ray diagnosis, refer to Figures 6 and 7; models may infer the attributes of one disease solely based on others. For example, "Support Devices" co-occurred with "Pleural Effusion" in **11.18%** of cases (in MIMIC-CXR-JPG) and also showed high co-occurrence with most diagnoses. This is likely because the data is collected from ICU patients, leading to biased sampling. If we now test the model on a dataset not collected from ICU patients, the model may still assign undue importance to "Support Devices," which is not clinically relevant, resulting in biased predictions. Mitigating bias is not within the scope of this work, but causal approaches appear promising for future research, as shown in (Jones and colleagues, 2024; Jones et al., 2023; Song et al., 2024).

5.2 Real-World Clinical Deployment

As discussed in Section 5.1, the lack of balanced datasets and the presence of spurious correlations pose significant challenges to deploying such models in real-world clinical settings. While existing models, including ours, are not yet ready for direct clinical use, they hold strong potential as **assistive tools for radiologists**. Specifically, they can aid in generating provisional reports and automating routine tasks. For instance, even an experienced clinician typically requires 5–10 minutes to interpret and compose a radiology report (Hou et al., 2023b), whereas our model can generate a preliminary report in less than a second per instance. This efficiency makes it well-suited for handling straightforward or repetitive cases, thereby streamlining the diagnostic workflow. However, the primary barrier to achieving reliable end-to-end automated reporting lies in the underlying data biases. Addressing this requires the development and curation of more representative, bias-free datasets—a goal that is essential yet notably difficult to accomplish in practice.

6 Ethical Considerations

The authors of both the *IU-Xray* (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

2023. Understanding bias in machine learning models. *Arize Blog*.

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*.

Omar Alfarghaly, Rana Khaled, Abeer Elkorany, Maha Helal, and Aly Fahmy. 2021. Automated radiology report generation using conditioned transformers. *Informatics in Medicine Unlocked*, 24:100557.

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.

NN Bourgin and NN Peterson. 2024. Modelling dataset bias in machine-learned theories of economic decision-making. *PMC*.

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.

Zhang Chen, Yixin Zhang, Yixuan Zhang, Derek Lin, Jialin Song, and Yizhou Yu. 2020. Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2003.10152*.

Zhenzhou Chen, Zhiqiang Tian, Jun Zhu, Chang Li, and Shu Du. 2022. C-cam: Causal cam for weakly supervised semantic segmentation on medical image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11676–11685.

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.

Mark Endo, Rayan Krishnan, Viswesh Krishna, Andrew Y. Ng, and Pranav Rajpurkar. 2021. Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In *Proceedings of Machine Learning for Health*, volume 158 of *Proceedings of Machine Learning Research*, pages 209–219. PMLR.

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.

Chowdhury Mohammad Rakin Haider. 2024. *Identifying Induced Bias in Machine Learning*. Tech report, CERIAS, Purdue University.

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*.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Christopher Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jason Seekins, David A Mong, Safwan Halabi, Jason K Sandberg, Robyn Jones, David B Larson, Curtis P Langlotz, Bhavik N Patel, Matthew P Lungren, and Andrew Y Ng. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *arXiv preprint arXiv:1901.07031*.

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*.

Author Jones and colleagues. 2024. A causal perspective on dataset bias in machine learning for medical imaging. *Nature Machine Intelligence*.

Charles Jones, Daniel C. Castro, Fabio De Sousa Ribeiro, Ozan Oktay, Melissa McCradden, and Ben Glocker. 2023. No fair lunch: A causal perspective on dataset bias in machine learning for medical imaging. *arXiv preprint arXiv:2307.16526*.

Jeremy Jones and Liz Silverstone. 2024. Chest radiograph. *Radiopaedia.org*.

Kaveri Kale, Pushpak Bhattacharyya, Milind Gune, Aditya Shetty, and Rustom Lawyer. 2023. Kgvlbart: 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.

Yuta Miura, Kazunari Hara, and Yuta Hayashi. 2021. Improving radiology report generation with memory-driven transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14344–14353.

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.

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, Jiafan Liu, Yan Liu, Yun Li, Wenbin Lei, and Ruxin Wang. 2024. Rethinking radiology report generation via causal inspired counterfactual augmentation. In *Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 1–10.

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. Kmbart: 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

7 Dataset

The **MIMIC-CXR-JPG** (Johnson et al., 2019) and **IU-Xray** (Demner-Fushman et al., 2016) datasets are widely used benchmarks in radiology report generation. MIMIC-CXR-JPG comprises 377,110 chest X-rays from 227,835 studies across 65,379 patients (2011–2016), paired with free-text, de-identified reports. IU-Xray, though smaller with 7,470 images and 3,825 reports, provides structured reports with distinct *Findings* and *Impression* sections, and a balanced distribution of normal and abnormal cases. We use both datasets to ensure robustness and comparability with prior work.

8 Evaluation Metrics

We evaluate using 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* and *CIDEr* assess n-gram overlaps; *METEOR* accounts for synonyms and recall; *BERTScore* measures contextual semantic similarity; *ROUGE-L* evaluates summarization via longest common subsequence; and Embedding-Based Metrics compute semantic similarity.

Since these NLG metrics may miss clinical accuracy, we use *CheXbert* (Smit et al., 2020) to extract disease labels from generated reports and compare them to references. Due to space constraints and prior works omitting some metrics, detailed NLG results are reported in the appendix (Table 3) for comprehensive ablation analysis.

9 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 3. 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

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
$LRTA - BioMIC_4$	0.464	0.345	0.265	0.203	0.966	0.516	0.301	0.885	0.414	0.958
$LRTA - BioMIC$	**0.527**	**0.384**	**0.279**	**0.226**	**1.013**	**0.522**	0.347	0.898	**0.448**	**0.969**

Table 3: Performance comparison of $LRTA - BioMIC$ against multiple Ablation architecture (c.f section 3), 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.

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 necessary phrasing and repetitions.

10 Evaluation on CheXpert

Model	**Precision**	**Recall**	**F1**
R2Gen	–	–	0.191
M^2 Trans	–	–	0.326
CXR-RePaiR-Select	–	–	0.352
RGRG (reproduced)	0.381	0.397	0.389
LRTA-BioMIC (Ours)	0.424	0.429	**0.426**

Table 4: Performance comparison on the CheXpert dataset.

CheXpert is a large-scale chest X-ray dataset that differs from IU-Xray and MIMIC-CXR-JPG in that it provides structured labels indicating the presence, absence, or uncertainty of 14 clinical observations, rather than free-text radiology reports. All prior and recent models discussed in Table 1 have only evaluated on IU and MIMIC, as these are considered the most reliable and widely adopted datasets. To assess generalizability beyond these conventional benchmarks, we evaluated our model LRTA-BioMIC alongside four baselines. Results for M^2 Trans (Chen et al., 2020), R2Gen (Miura et al., 2021), and CXR-RePaiR-Select are taken from (Endo et al., 2021). We acknowledge the authors of RGRG (Tanida et al., 2023) for their open-source repository, which enabled accurate reproduction. Using openly available CheXpert data (Irvin et al., 2019), we randomly sampled 40K studies and split them into 80% train, 10% validation, and 10% test sets, using final label verdicts as gold labels. Since CheXpert does not contain free-text reports, evaluation using natural language generation (NLG) metrics is not feasible; instead, we report F1 scores based on label classification. Our model achieved an F1 score of 0.426, reflecting a **+9.5%** improvement over the best-performing baseline RGRG (cf. Table 4). We encourage future work to explore datasets beyond IU-Xray and MIMIC-CXR-JPG to ensure broader robustness and generalization.

11 BioMIC-BART

Figure 2 illustrates the architecture of our **BioMIC-BART**, which is built upon BioBART (Yuan et al., 2022), a language model pretrained on full-text biomedical literature from PubMed. Compared to the original BART model, BioBART incorporates domain-specific biomedical terminology and contextual reasoning, making it more suitable for medical applications. While BioBART captures rich biomedical knowledge, it is primarily trained on textual data and lacks grounding in visual representations or domain-specific imaging patterns observed in chest X-rays. Such grounding is essential for radiology tasks where language often tightly correlates with anatomical regions. To bridge this gap, we draw inspiration from (Xing et al., 2021), which extended the BART architecture to handle multimodal inputs comprising both images and text. We adopt this architecture—originally trained on general vision-language datasets such as Conceptual Captions (Sharma et al., 2018), SBU (Ordonez et al., 2011), COCO (Lin et al., 2014), and Visual Genome (Krishna et al., 2017)—but re-train it on

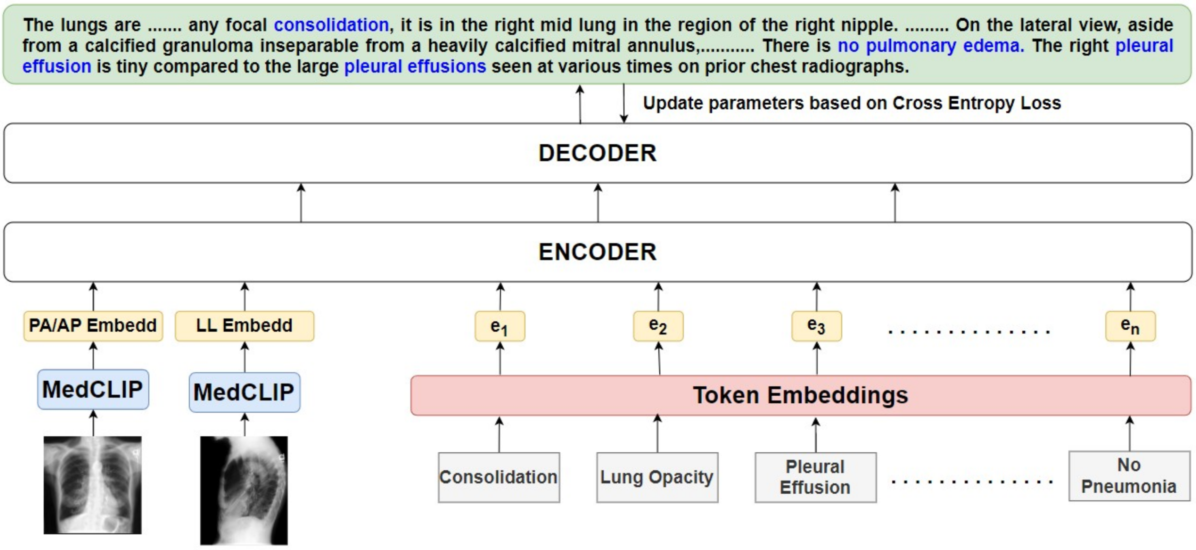

The lungs are .... any focal consolidation, it is in the right mid lung in the region of the right nipple. .... On the lateral view, aside from a calcified granuloma inseparable from a heavily calcified mitral annulus,.......... There is no pulmonary edema. The right pleural effusion is tiny compared to the large pleural effusions seen at various times on prior chest radiographs.
Update parameters based on Cross Entropy Loss
DECODER
ENCODER
PA/AP Embedd
LL Embedd
e_1
e_2
e_3
e_n
MedCLIP
MedCLIP
Token Embeddings
Consolidation
Lung Opacity
Pleural Effusion
No Pneumonia

Figure 2: Illustration of the **BioMIC-BART** architecture, an extension of BioBART designed for multimodal processing. It integrates chest X-ray image embeddings and radiology tag embeddings into a unified encoder-decoder framework to enhance multimodal radiological understanding.

chest X-ray images and associated reports from the MIMIC-CXR dataset. (Xing et al., 2021) provides an in-depth analysis of the role of each component and demonstrates their importance toward achieving multimodal processing. In our work, we reuse the same architecture but apply it specifically to the domain of Chest X-ray Report Generation.

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) view and the Lateral (LL) view, if available, using the Anteroposterior (AP) view only when the PA view is unavailable, to provide BioMIC-BART with contextual information from multiple perspectives. The PA/AP view is the standard 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 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 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 multimodal radiology Chest X-ray data.

11.3 Encoder-Decoder

The model architecture comprises 12 encoder-decoder layers designed to effectively process and integrate multimodal data. The encoder receives two types of embeddings: image embeddings, extracted from Posteroanterior (PA) or Anteriorposterior (AP) and Lateral (LL) chest X-ray views using MedCLIP, and token embeddings, derived from chest X-ray tags using CXR-BERT-general.

To accommodate the variability in available image views within the dataset, we used the Anteroposterior (AP) view only when the PA view was unavailable. If neither PA nor AP views were present, a zero matrix was passed in place of the image embedding. Similarly, if the Lateral view was missing, a zero matrix was also used. This decision was guided by the view distribution within the MIMIC-CXR-JPG dataset (Table 5), which is heavily skewed toward AP images, a result of data collection in ICU settings where patients are typically bedridden and AP imaging is more feasible. However, despite their prevalence, AP views are diagnostically inferior to PA views, which radiologists prefer whenever feasible due to their superior image quality (Jones and Silverstone, 2024). Therefore, when both views are available, our method prioritizes PA over AP to ensure higher reliability.

Moreover, in real-world clinical scenarios, where patients are ambulatory, PA views are far more commonly acquired than AP. Thus, designing a model that prioritizes PA view not only aligns with clinical best practices but also better generalizes to non-ICU environments where PA imaging predominates.

View Combination	Number of Patients
PA + AP + Lateral	183
PA + Lateral	85,077
AP + Lateral	19,978
PA + AP	7,424
Only PA	483
Only AP	112,289
Only Lateral	2,401
Total Studies	227,835
Total Image Count	377,110

Table 5: Distribution of Chest X-ray View Combinations in MIMIC-CXR-JPG

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 3).

12 Parameter and Computational Resources

We categorize our experiments into two groups: (1) *BioMIC-BART*, a computationally intensive large-scale language model, and (2) *LRTA-BioMIC*, our final lightweight model combining ViT with a region-guided language decoder. Both setups used the *GELU* activation function, the *Adam* optimizer with a weight decay of *0.001*, run on *A100 GPUs*.

BioMIC-BART. A grid search over learning rates (*3e-4*, *3e-5*, *3e-6*) and batch sizes (*48*, *64*) identified *3e-5* and *48* as optimal. Training was performed on 4×A100 GPUs (80GB each) for *20* epochs using 90% of the *MIMIC-CXR-JPG* train split. The remaining 10% was reserved for fine-tuning LRTA-BioMIC. Each full training run took approximately *26 hours* to complete end-to-end, including checkpointing, logging, and intermediate evaluations.

LRTA-BioMIC. Grid search selected *3e-5* learning rate and *batch size 4* over *20* epochs. For MIMIC-CXR-JPG, 10% of the official training split was used; for IU-Xray, an 80-10-10 custom split was created using a fixed random seed. GPU memory usage was *7GB* (MIMIC) and *6GB* (IU-Xray). Training duration was *4.5 hours* (MIMIC) and *1.5 hours* (IU-Xray), underscoring the model's efficiency, fast turnaround, and suitability for lightweight deployments.

13 Error Analysis

We conducted an analysis to identify weaknesses in LRTA-BioMIC. We identified two key weaknesses: **Numerical Discrepancies (Weakness-A).** In Table 6, 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 6, 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 6: Table highlighting model weaknesses by presenting two examples per weakness, comparing the gold reports with the generated reports.

14 Statistical Significance Analysis

- **IU-Xray Dataset:**
  - **Bleu-4:** The one-tailed t-test yields $p = 0.0138$ (< 0.05) when compared to COMG (the best-performing baseline), indicating a statistically significant improvement.
  - **Rouge-L:** The p-value is < 0.05 for 4 out of 5 baseline models (Table 7). For COMG, while the difference is not marginally significant, LRTA-BioMIC still achieves a higher score.
  - **CE Metric:** LRTA-BioMIC achieves a p-value < 0.05, indicating statistical significance when compared to RGRG. While it is not marginally significant on PromptMRG, it still outperforms in terms of F1-score.

- **MIMIC-CXR Dataset:**
  - **Bleu-4:** For 3 out of 5 baseline models (Table 1), the p-value is < 0.05, demonstrating significance. For the remaining two models, the difference is not statistically significant.
  - **Rouge-L:** The one-tailed t-test shows that none of the baseline models are statistically significant compared to LRTA-BioMIC.
  - **CE Metric:** LRTA-BioMIC is statistically significant ($p < 0.05$) in 3 out of 5 baseline models. For the remaining two models, the difference is not statistically significant.

- **None** of the baseline models were significantly stronger than **LRTA-BioMIC** on any metric.

- A total of 30 paired one tailed t-tests were conducted (3 metrics $\times$ 5 baselines $\times$ 2 datasets):
  - **LRTA-BioMIC** showed significantly stronger performance ($p < 0.05$) in **21** instances.
  - In the remaining 9 cases:
    - There were only **4** instances where a baseline outperformed LRTA-BioMIC.
    - **None** of these 4 instances were statistically significant.

- On the **MIMIC-CXR-JPG** dataset:
  - Out of 9 evaluated metrics, **LRTA-BioMIC** ranked as either the best-performing or second-best-performing model in **8** of them.

- On the **IU-Xray** dataset:
  - Out of 9 metrics, **LRTA-BioMIC** achieved the best performance in **8** and was the second-best in only one metric (**BLEU-1**).

The detailed significance tests validate the effectiveness of our proposed model (cf. Table 7). Our model achieves state-of-the-art performance on the **IU-Xray** dataset and delivers competitive results on the large-scale **MIMIC-CXR-JPG** dataset when compared with existing baseline models. Notably, this is accomplished while adhering to the primary objective of our work—developing a model that is comparatively lightweight and efficient relative to existing approaches.

15 How Lightweight is LRTA-BioMIC

"Lightweight" is used here as a comparative term relative to existing architectures. The efficiency claims apply exclusively to **LRTA-BioMIC**. While **BioMIC-BART**—a robust backbone of LRTA-BioMIC enriched with chest X-ray data and multimodal learning capabilities—initially required approximately 26 hours of pre-training on multiple A100 GPUs (refer to Section 12), once trained, it can be efficiently loaded from Hugging Face or a local device, eliminating the need for extensive retraining. All subsequent training and inference on new data are performed solely on LRTA-BioMIC.

Our objective is to achieve state-of-the-art performance or results comparable to existing approaches, which typically rely on heavy computational resources. In contrast, our method offers an efficient alternative with competitive performance. Table 8 presents a comparative analysis of the time and GPU resources required to train each model on a new dataset.

Compared to existing architectures, LRTA-BioMIC demonstrates remarkable efficiency in GPU memory consumption. With a requirement of only 6–7 GB, it achieves approximately

Metric	IU-Xray					MIMIC-CXR-JPG				
	RGRG	COMG	PromptMRG	ORGan	KGVL-BART	RGRG	COMG	PromptMRG	ORGan	RECAP
Bleu-4	✓	✓	✓	✓	✓	×	✓	✓	✓	×
Rouge-L	✓	✓	✓	✓	×	✓	×	✓	×	×
CE	✓	–	×	–	–	×	✓	×	✓	✓

Table 7: ✓ indicates that **LRTA-BioMIC** has a **p-value** < 0.05, signifying a statistically significant improvement over the corresponding baseline model (mentioned in the column) on the specified dataset (also mentioned in the column) for the metric listed in the corresponding row. **However, × does not imply that LRTA-BioMIC performs worse than the baseline; it only indicates that the p-value was not** < 0.05. Refer Section 14 for deeper analysis.

Model	**GPU (GB)**	**Tr. Hours**
RGRG	48	45
COMG	147–179	5
PromptMRG	24	24
ORGan	24	–
RECAP	24	–
KGVL-BART	60–80	4-5
LRTA-BioMIC	6–7	2–5

Table 8: Comparison of GPU memory usage and training time required for different models when adapting to a new dataset.

85% reduction in GPU memory compared to RGRG (48 GB), over **95% reduction** compared to COMG (147–179 GB), around **75% less** than PromptMRG, ORGan, and RECAP (each using 24 GB), and over **85% less** than KGVL-BART (60–80 GB). These significant reductions validate our claim that **LRTA-BioMIC is a lightweight architecture**, offering a computationally efficient alternative to existing models, without compromising performance—and in many cases, even achieving superior results.

16 Generalizability to Other Modalities: MRI and CT

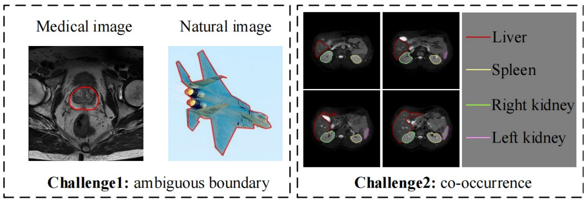


Figure 3: Challenges in radiology imaging such as MRI: (1) ambiguous boundaries between foreground and background, and (2) severe anatomical co-occurrence of organs. Figure adapted from (Chen et al., 2022).

Scalability to MRI and CT. Although our current work focuses on chest X-rays, the proposed **LRTA-BioMIC** framework is generalizable to other radiology domains such as MRI and CT. As illustrated in Figure 3, both modalities face two key challenges: unclear anatomical boundaries and frequent co-occurrence of organs (Chen et al., 2022).

Our approach addresses these via:

- **Region-Guided Feature Extraction**: Helps resolve both challenges by enabling region-specific representations (e.g., isolating left/right kidney) and learning sharper boundaries through training.
- **Region-Text Aligner**: Helps resolve co-occurrence by aligning disease tags (e.g., chronic kidney disease) to precise anatomical regions.

While the methodology applies to MRI and CT, experimenting with these modalities is out of scope for the current work, as it would require new datasets, extensive retraining of backbone models (e.g., MedCLIP, ResNet, BioBART), and curated tag information. Nevertheless, we will publicly release our model to support future research in these directions.

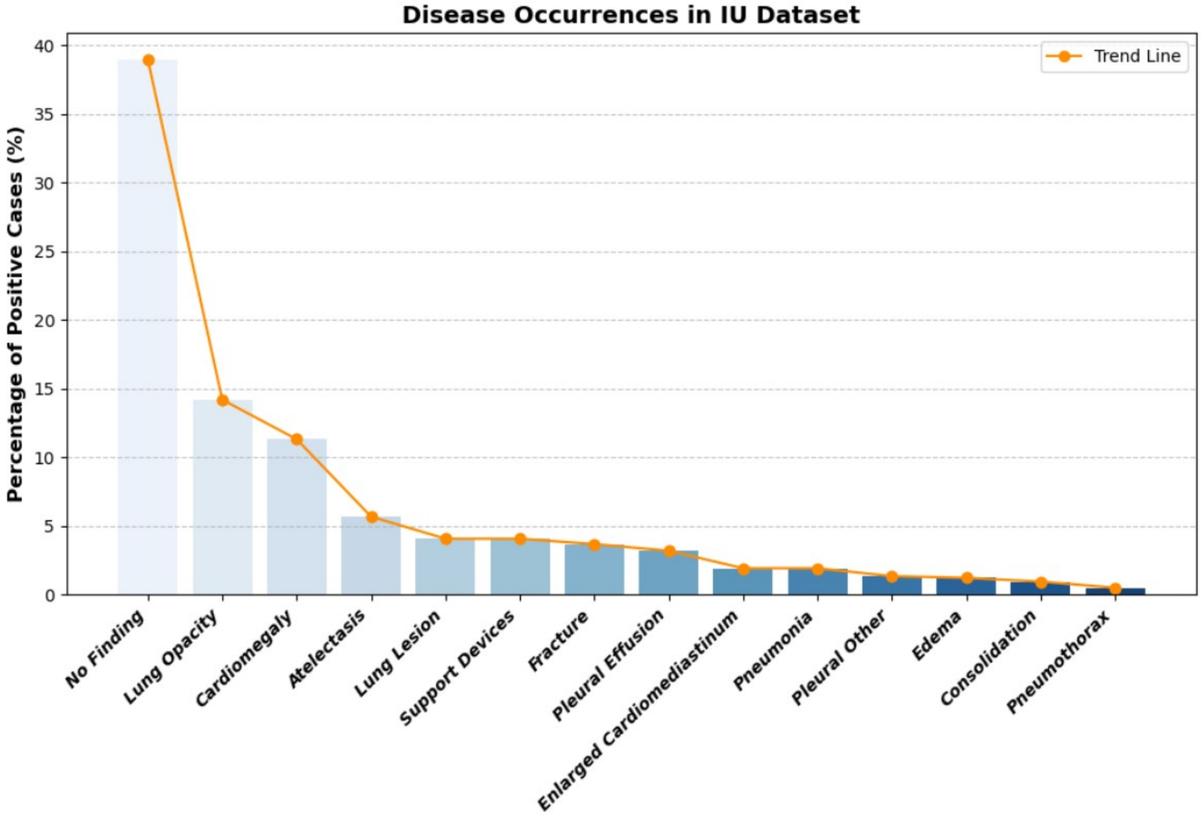


Figure 4: Distribution of diagnostic labels in the IU Chest X-ray dataset. A clear imbalance is visible, indicating a long-tail distribution where a small number of labels dominate the dataset.

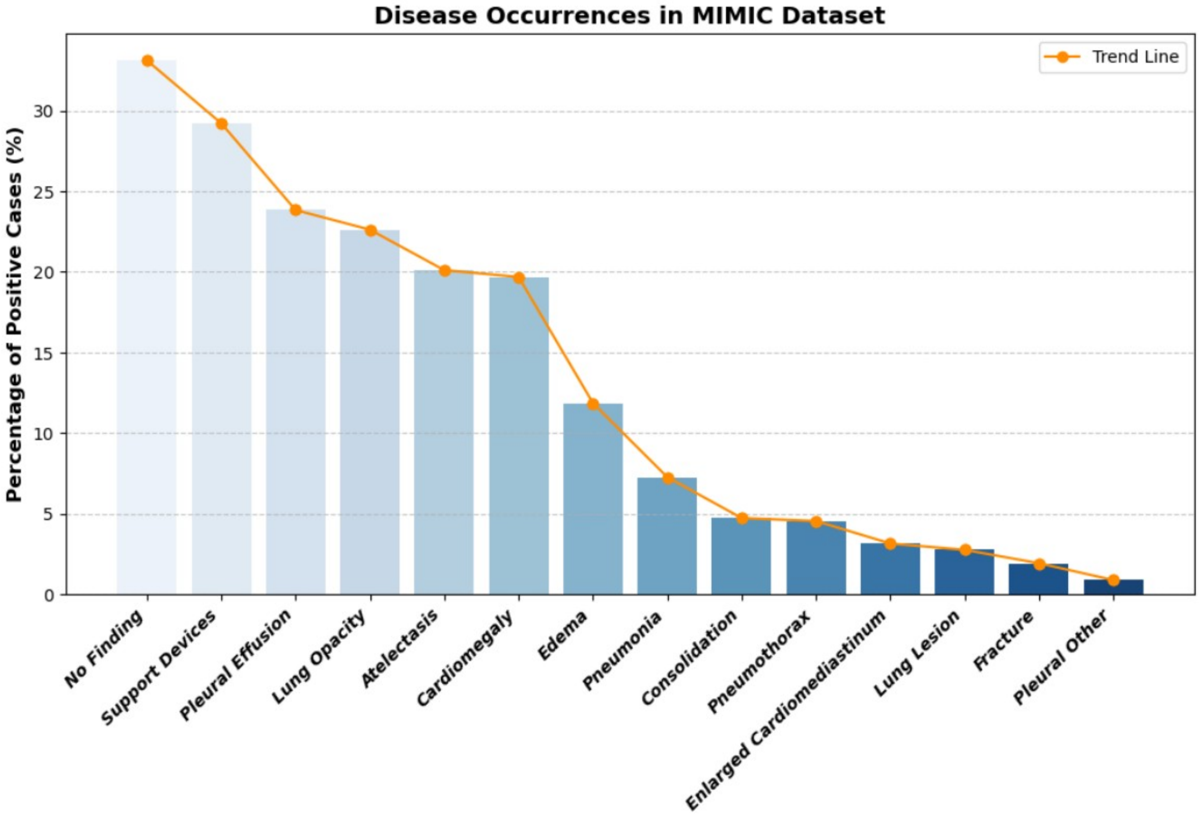


Figure 5: Label distribution in the MIMIC-CXR-JPG dataset. While less skewed than IU, the dataset still exhibits long-tail characteristics, reflecting the imbalance in class representation.

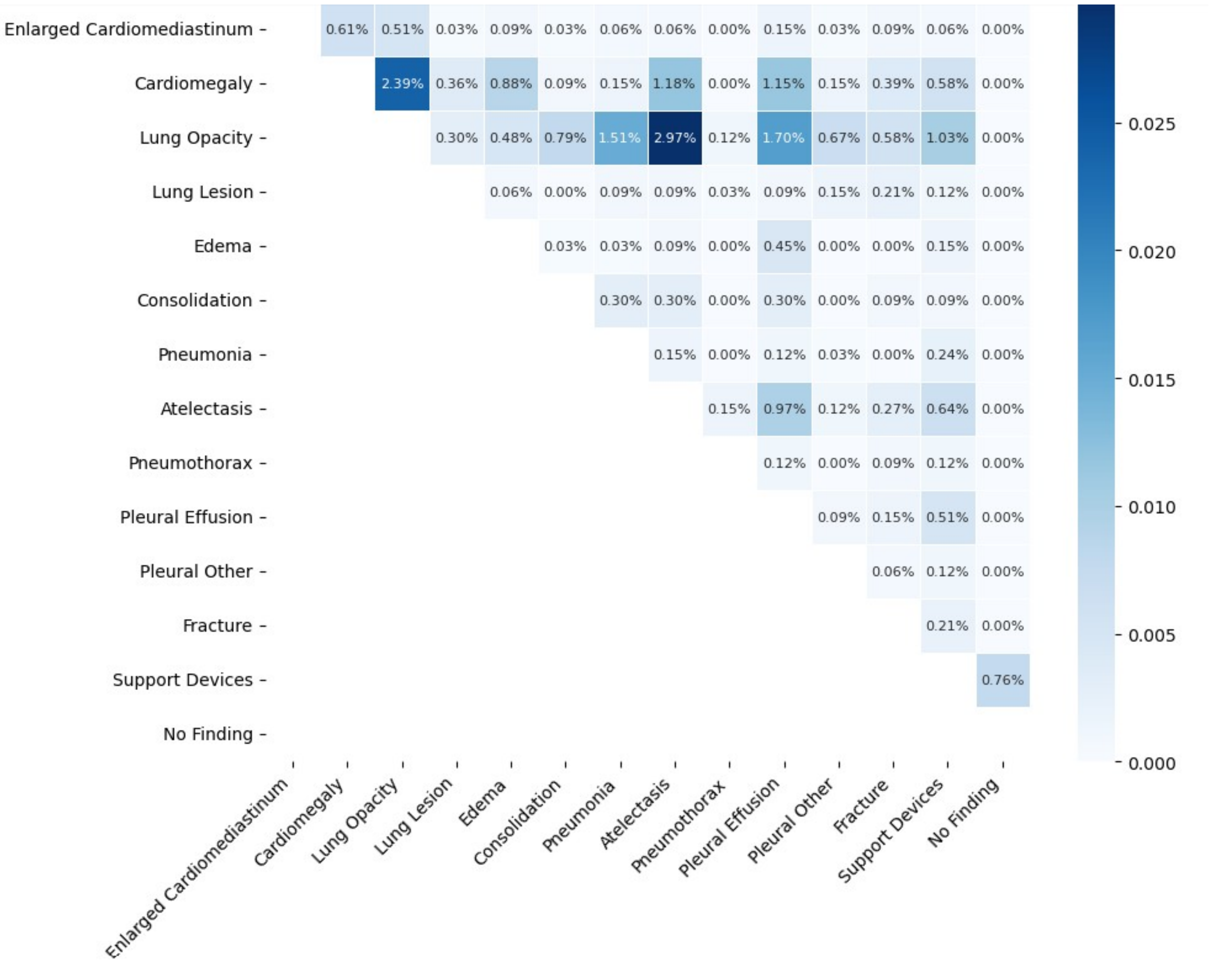


Figure 6: Co-occurrence matrix of diagnostic labels in the IU Chest X-ray dataset. Significant overlaps between certain conditions highlight potential spurious correlations that may bias model training.

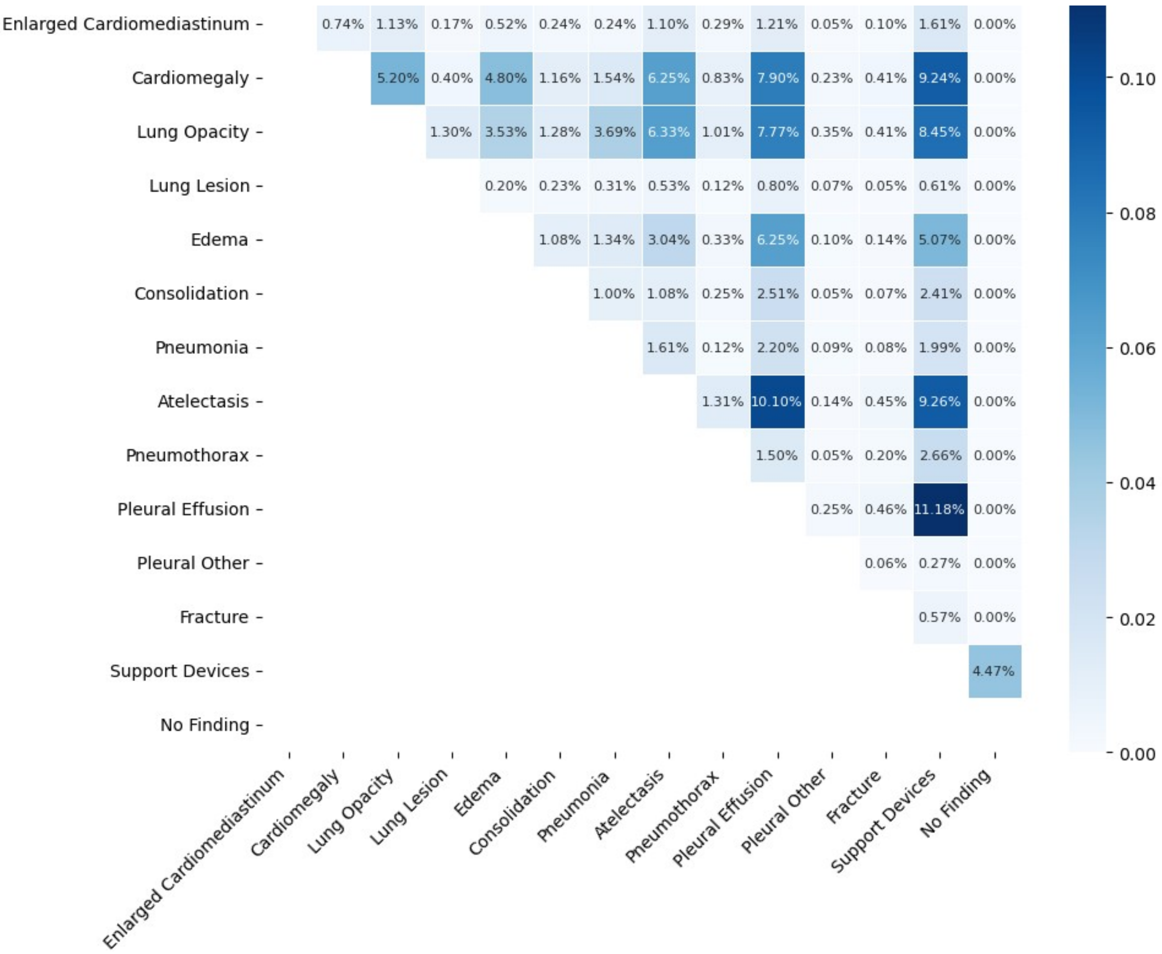


Figure 7: Co-occurrence matrix of diagnostic labels in the MIMIC-CXR-JPG dataset. Frequent co-occurrences, particularly involving support devices, indicate label dependencies that could affect model generalization.