Automated Radiology Report Generation using Vision-Language Models

Yuting Fang
z5518340

Lingxiao Jiang
z5562632

Jiabei Wu
z5597215

Zhiwei Guo
z5550962

Peidong Liang
z5603156

***Abstract*—The objective of this study is to generate accurate and clinically meaningful diagnostic reports from chest X-ray images using vision-language models. The development process involved the creation of six model variants, with the foundation being the BERT-style Medical Vision Language Learner (MedViLL), which has undergone pre-training on MIMIC-CXR-JPG. In the proposed framework, the original computer vision is substituted with three distinct vision models, namely CheXNet, ResNet101, and Vision Transformer (ViT), while the decoder from Clinical T5 is adopted for the generation of diagnostic text. We named the part including BERT-encoder and Clinical T5 decoder as RadScribe. Furthermore, an investigation is conducted into a label-enhanced approach, whereby clinical labels are extracted from the ground truth reports, which are utilized to train a classifier. Then, we integrate the label we get from the classifier with the vectorized images by cross-attention. The experimental results demonstrate that the models with a label-enhanced approach outperform other models, and the ViT-Label-RadScribe configuration outperforms all other variants and achieves better performance than the baseline MedViLL model.**

I. Introduction

A. Background

In recent decades, the field of radiological imaging has undergone revolutionary changes, making the accurate and efficient interpretation of medical images crucial in modern diagnostics [1]. The generation of radiology reports, to automatically produce free-text descriptions for clinical radiographs (e.g. chest X-rays), has emerged as a compelling research direction in the domains of artificial intelligence and clinical medicine [2]. It can considerably expedite the automation of workflows and enhance the quality and standardization of healthcare [2].

Recent advancements in vision-language models (VLMs) have significantly accelerated progress in this area. Foundation models have driven revolutionary advancements in complex multimodal applications, in which VLMs focus on integrating visual and language modalities [3]. By leveraging paired datasets during pre-training, models such as CLIP align images with text, making them highly effective for tasks requiring multimodal reasoning [3].

B. Motivation

Based on the above background, considerable scope exists for enhancement in the implementation of AI in this field. Recent breakthroughs in multimodal foundation models have demonstrated that AI systems trained on a vast quantity of unlabeled data can be adapted and achieve state-of-the-art accuracy in a wide range of downstream specialized tasks, including biomedical problems. Numerous issues require further investigation and resolution.

C. Contribution

In our project, we try to leverage a BERT-based model with the weights learned from the article 'Multi-Modal Understanding and Generation for Medical Images and Text via Vision-Language Pre-Training.'Based on the structure, we change the picture embedding and Decoder of the original model of the article, and use 10k instances from the MIMIC-CXR-JPG to adapt the new part of the model. Based on our results, it is evident that the model's performance has been enhanced to a certain extent.

II. Relative Works

Traditional radiology report generation methods primarily adopt CNN-based encoder-decoder frameworks, which often suffer from limited receptive fields and weak global context modeling, resulting in fragmented and clinically inconsistent reports [4] [5]. Recent studies have demonstrated the potential of vision-language transformer architectures in bridging visual semantics and medical narrative generation. For example, METransformer introduces learnable expert tokens to capture distinct semantic regions of chest X-rays, improving the alignment between image features and textual descriptions [6]. Similarly, GIT-CXR proposes an end-to-end generative vision-language model tailored for chest X-ray interpretation [4], while ChestX-Transcribe explores multi-task learning to enhance factual correctness and clinical coherence in report generation [5].

However, most existing approaches still face challenges in handling multi-view chest X-rays, generating view-specific content, and generalizing across varying input formats. Addressing these gaps, UniXGen introduces a unified transformer-based model that supports both report and image generation through bidirectional learning, incorporating special tokens for view control and leveraging multi-view information to improve diagnostic accuracy [7].

In addition to the aforementioned challenges in evaluating the clinical accuracy of AI report generation models, considerable scope exists for enhancement. Recent breakthroughs in multimodal foundation models have demonstrated that AI systems trained on a vast quantity of unlabeled data can be adapted and achieve state-of-the-art accuracy in a wide range of downstream specialized tasks, including biomedical problems [8]. Nonetheless, regulatory and ethical considerations are ongoing challenges, especially when it comes to patient safety and model interpretability [9].

III. Methodology

A. Picture Embedding

The notion that image vectorization, simplification, and edge tracing should be approached as complementary tasks has a long history in both computer graphics and computer vision [10]. The process of picture embedding involves the conversion of a raster image into a vector representation, thereby distilling the image's essential information into a simplified vector. In our model, we mainly use the CheXNet and Vit as our picture vectorization models.

- **ResNet101** is a deep convolutional neural network with 101 layers that uses residual connections to solve the problem of vanishing gradients in deep networks [11]. These shortcut connections allow the model to train very deep architectures more effectively [11]. In our models, we use the weights of ResNet101, which were pre-trained on the Chest X-ray.
- **CheXNet121** is a vision model based on the DenseNet121 architecture, which is pre-trained on large-scale chest X-ray datasets (such as Chest X-ray14) [12]. The model is capable of providing multilabel prediction of disease signs in images, including pneumonia, pneumothorax, cardiomegaly and e.g [13]. Further, the basic model of CheXNEt is based on Convolutional Neural Network(CNN), which is commonly used in computer vision [14].
- **Vision Transformer (ViT)** is an image classification model, which is also pre-trained on the large-scale ImageNet-21k dataset and outperforms in the medical-image analysis area [15]. Compared with traditional CNN, ViT adopts a novel image processing approach, segmenting the input image into multiple patch sequences and modeling the global dependencies by the self-attention mechanism [16].

The finally output of the two models is the feature vectors of the input images.

B. BERT-Based Encoder

Our encoder is a unified BERT-style Transformer that performs early fusion of images and text within a single self-attention stack. An image is converted into a sequence of visual tokens by a visual backbone and embedded with modality and positional information; text is tokenized and embedded in the standard BERT manner, with segment cues indicating modality. The model forms one joint token stream—beginning with a classification token, followed by image tokens, a separator, and text tokens—so that all tokens attend to one another and exchange information across modalities. Optionally, a lightweight cross-attention step conditions visual tokens on the textual stream to tighten alignment when supervision is available [18]. Passing this unified sequence through the Transformer layers yields contextualized multimodal representations that preserve spatial structure and linguistic order, providing a strong foundation for downstream understanding and for conditioning an autoregressive report decoder [19] [20].

C. Decoder

Clinical T5 is a domain-adapted version of Google's Text-to-Text Transfer Transformer (T5), which has been specifically fine-tuned on large-scale clinical datasets, including electronic health records and radiology reports. This specialised pre-training has been shown to result in superior performance by Clinical T5 in tasks such as clinical question answering, medical report generation and information extraction. The text produced by Clinical T5 is characterised by increased accuracy, professionalism and alignment with medical standards when compared with the general T5 [21].

IV. Experiments

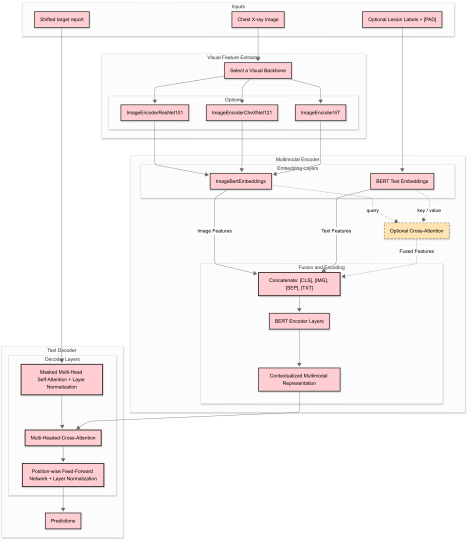


Fig. 1: Framework

A. Framework Diagram

The Figure 1 In short, we have six different models:

- CheXNet-RadScribe
- CheXNet-Label-RadScribe
- ResNet101-RadScribe
- ResNet101-label-RadScribe
- ViT-RadScribe
- Vit-Label-RadScribe

And the larger Diagram can refer to the Methodology part of `COMP9444Project.ipynb`..

B. Experiment Setting

1) Hardware Specifications: All experiments were conducted on a custom-built desktop machine equipped with the hardware detailed in Table I.

TABLE I: Hardware Specifications

Component	Specification
GPU	NVIDIA GeForce RTX 4070 Ti
CUDA	12.9

2) Implementation Details: All models were implemented using PyTorch. For the text generation part, we employed the **ClinicalT5** model as the decoder across all experiments. We trained six main variants of our model on MIMIC-CXR-JPG, combining the three encoders with two different training strategies: one with and one without label enhancement. The `-label-` versions incorporate additional cross-attention layers in the encoder to fuse textual label information with visual features. Further, A total of 10,000 instances were extracted from the original dataset to serve as the training set, and subsequently, 500 new instances were extracted from the original dataset to function as the validation set.

Key hyperparameters are detailed in Table II. We used a batch size of 2 with 8 gradient accumulation steps, resulting in an effective batch size of 16. All training was performed using Automatic Mixed Precision (AMP) to accelerate computation.

TABLE II: Key Hyperparameters

Parameter	Value
Image Size	512x512
Batch Size	2
Gradient Accumulation	8
Effective Batch Size	16
Optimizer	AdamW
Mixed Precision	Enabled (FP16)

3) Training Procedure: Our fine-tuning process is conducted in two sequential stages, with learning rates detailed in Table III.

TABLE III: Learning Rate Configurations for Fine-Tuning

Stage	Component	Learning Rate
Stage 1: Decoder Fine-tuning	Encoder	Frozen
	Decoder	2×10^{-5}
Stage 2: End-to-End Training	Encoder	5×10^{-5} (min 1×10^{-5})
	Decoder	1×10^{-4} (min 2×10^{-5})

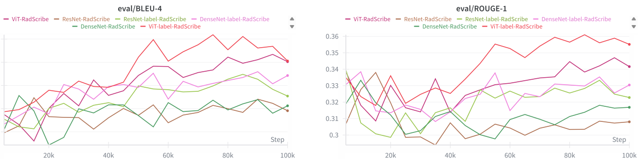
a) Stage 1: Decoder and Cross-Attention Fine-tuning: In the first stage, we perform a focused fine-tuning for 3 epochs. During this phase, the entire visual encoder is kept frozen. For the label-enhanced models, both the ClinicalT5 decoder and the newly added cross-attention layers are trained. For models without label enhancement, only the decoder is trained. This allows the language-generation components to adapt to the task without altering the pre-trained visual features.

b) Stage 2: Full End-to-End Training: Following the initial decoder fine-tuning, the entire model is unfrozen. We then proceed with end-to-end training for up to 20 epochs. In this stage, all parameters, including those of the visual encoder and the decoder, are updated. This allows the visual encoder to fine-tune its features for the specific medical imaging domain. This stage utilizes a cosine learning rate decay schedule with a 10% warmup phase and includes early stopping to prevent overfitting.

4) Evaluation Metrics: To evaluate model performance, we conducted extensive experiments using eight automatic metrics: BLEU-4, ROUGE-L, METEOR, CIDEr, Accuracy, Precision, Recall, and F1-score, which jointly assess both linguistic quality and clinical relevance.

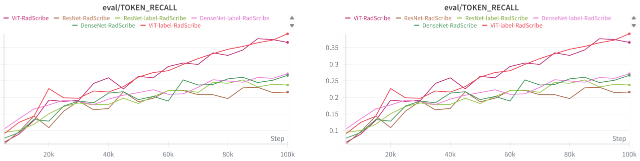
V. Result

Following the training of the six models, the subsequent comparison was made between them and MEdViLL (Baseline).



((a)) Text BLEU-4 ((b)) Text ROUGE-L

Fig. 2: Text-level evaluation metrics (BLEU-4 and ROUGE-L) over training.



((a)) Token F1 ((b)) Token Recall

Fig. 3: Token-level evaluation metrics (F1 and Recall) over training.

Figures 2 and 3 show the performance trends for the selected models across key text-level and token-level metrics during training.

For the text-level metrics BLEU-4 (Figure (a)) and ROUGE-L (Figure (b)), ViT-label-RadScribe consistently achieves the highest or near-highest scores, with the gap over other models widening after approximately 40k steps—indicating stronger fluency and structural coherence. ViT-RadScribe remains competitive but trails its label-supervised counterpart.

For the token-level metrics F1 (Figure (a)) and Recall (Figure (b)), ViT-label-RadScribe exhibits the most stable and sustained gains, finishing with the best overall performance, suggesting superior identification of clinically relevant terms while maintaining accuracy. ViT-RadScribe follows closely but improves at a slightly slower rate.

Tables IV and V present the results of models trained without label supervision. As shown in Table IV, ViT-RadScribe achieves the highest BLEU-4 (0.0882) and CIDEr (0.1288) scores among the unlabeled models, while also performing strongly in METEOR and ROUGE-L. Table V indicates that,

TABLE IV: Unlabeled Text Evaluation

	BLEU-4	CIDEr	METEOR	ROUGE-L
DenseNet-RadScribe	0.070592	0.109908	0.247913	0.199864
ResNet-RadScribe	0.064833	0.100100	0.238848	0.196596
ViT-RadScribe	0.088191	0.128825	0.264228	0.215673
MedViLL (baseline)	0.066000	–	–	–

TABLE V: Unlabeled Token Evaluation

	Accuracy	F1	Precision	Recall
DenseNet-RadScribe	0.996482	0.672507	0.669586	0.675454
ResNet-RadScribe	0.998047	0.631757	0.626588	0.637012
ViT-RadScribe	0.997296	0.773979	0.766252	0.781863
MedViLL (baseline)	0.841000	0.621000	0.698000	0.559000

at the token level, ViT-RadScribe achieves the best F1 (0.7739) and Recall (0.7819), suggesting strong capability in retrieving clinically relevant terms, although DenseNet and ResNet variants also show competitive Accuracy. It is evident that the model ViT-Scribe outperforms the baseline(MedViLL).

TABLE VI: Labeled Text Evaluation

	BLEU-4	CIDEr	METEOR	ROUGE-L
DenseNet-label-RadScribe	0.083249	0.085944	0.261153	0.213036
ResNet-label-RadScribe	0.077390	0.085778	0.255216	0.211238
ViT-label-RadScribe	0.095590	0.109285	0.280670	0.231057
MedViLL (baseline)	0.066000	–	–	–

TABLE VII: Laebled Token Evaluation

	Accuracy	F1	Precision	Recall
DenseNet-label-RadScribe	0.999907	0.737084	0.778203	0.700092
ResNet-label-RadScribe	0.998212	0.765382	0.682939	0.862399
ViT-label-RadScribe	0.999853	0.823560	0.861414	0.788892
MedViLL (baseline)	0.841000	0.621000	0.698000	0.559000

Tables VI and VII summarize the results for models trained with label supervision. Table VI shows that ViT-label-RadScribe achieves the highest scores in BLEU-4 (0.0956), METEOR (0.2807), and ROUGE-L (0.2311). Table VII further confirms its token-level advantage, leading in F1 (0.8236) and Precision (0.8614), while also performing strongly in Recall (0.7889). It can also be detected that our labeled models are better than the MedViLL in nearly every metric.

A comparison between the labeled and unlabeled settings reveals that incorporating label supervision leads to superior performance across most metrics, demonstrating its effectiveness in enhancing both text quality and clinical term accuracy. In particular, within the ViT family, both ViT-label-RadScribe and ViT-RadScribe are the best models in their respective settings. However, the labeled variant consistently outperforms its unlabeled counterpart in key metrics such as BLEU-4, METEOR, F1, and Precision, highlighting the additional benefit of label guidance.

Finally, compared to the baseline MedViLL, all proposed models show substantial improvements, with ViT-label-RadScribe achieving the best performance over the metrics.

VI. Conclusion

We can conclude that we have achieved our goal. The ViT-label-RadScribe model outperforms MedViLL(baseline) and also demonstrates significant potential in automating radiology report generation. Further, we find the ViT-based model always better than others. The Vision Transformer (ViT) architecture excels at capturing global image context [22], proving more effective than CNN-based models for comprehensive diagnostic tasks [23].

For further analysis, we summarize the following strengths, limitations, and some recommended future works.

A. Key Strengths

- We use many Pre-trained Models as our initial states, which gives a high-quality initialization, accelerating convergence and boosting overall performance. It also saves a lot of computing power.
- Also, the label-enhancement strategy in our model provides strong semantic guidance [24] and acts as a powerful regularizer, significantly improving both clinical accuracy and text quality [25].

B. Key Weaknesses

- Interpretability Deficit: Like many deep learning models, our solution operates as a 'black box' [26], lacking transparent reasoning for its diagnostic conclusions, which can hinder clinical trust.
- Dataset & Language Constraints: Trained on a limited 10k-sample subset of MIMIC-CXR and confined to English-only reports, restricting generalizability and global applicability.
- Single-View Input: Relying solely on a single posteroanterior (PA) view provides incomplete diagnostic information compared to clinical practice, which often utilizes multiple views (e.g., lateral).

C. Recommendations for Future Work

- Data Expansion & Balancing: Leverage the full MIMIC-CXR dataset and apply balancing techniques (resampling, cost-sensitive learning) to improve detection of rare conditions.
- Multi-Modal Integration: Develop models that process multi-view images (PA and lateral) and incorporate other modalities (e.g., patient history) for a holistic diagnosis [27].
- Cross-Lingual Development: Explore transfer learning with large multilingual models to adapt report generation for non-English clinical environments [28].
- Enhancing Interpretability: Implement visualization techniques (e.g., Class Activation Maps, Grad-CAM) to highlight image regions influencing model outputs, increasing transparency and clinical trust.

For more details, please refer to the `COMP9444Project.ipynb`.

References

[1] Chen, H., Zhao, W., Li, Y., Zhong, T., Wang, Y., Shang, Y., ... Zhang, T. (2024). 3d-ct-gpt: Generating 3d radiology reports through integration of large vision-language models. arXiv preprint arXiv:2409.19330.

[2] Chen, Z., Song, Y., Chang, T. H., Wan, X. (2020). Generating radiology reports via memory-driven transformer. arXiv preprint arXiv:2010.16056.

[3] Ryu, J. S., Kang, H., Chu, Y., Yang, S. (2025). Vision-language foundation models for medical imaging: a review of current practices and innovations. Biomedical Engineering Letters, 1-22.

[4] Chen, Y. J., Shen, W. H., Chung, H. W., Chiu, C. H., Juan, D. C., Ho, T. Y., ... Ho, T. Y. (2022). Representative image feature extraction via contrastive learning pretraining for chest x-ray report generation. arXiv preprint arXiv:2209.01604.

[5] Singh, P., Singh, S. (2025). ChestX-Transcribe: a multimodal transformer for automated radiology report generation from chest x-rays. Frontiers in Digital Health, 7, 1535168.

[6] Wang, Z., et al. (2023). METransformer: Radiology Report Generation by Transformer with Multiple Learnable Expert Tokens. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). arXiv:2304.02211.

[7] Kim, T., Kim, J., Lee, S., Choi, E. (2023). UniXGen: A Unified Vision-Language Model for Multi-View Chest X-ray Generation and Report Generation. *arXiv preprint arXiv:2302.12172*.v

[8] Tanno, R., Barrett, D.G.T., Sellergren, A. et al. Collaboration between clinicians and vision–language models in radiology report generation. Nat Med 31, 599–608 (2025). https://doi.org/10.1038/s41591-024-03302-1

[9] Foote, H. P., Hong, C., Anwar, M., Borentain, M., Bugin, K., Dreyer, N., ... Lindsell, C. J. (2025). Embracing Generative Artificial Intelligence in Clinical Research and Beyond: Opportunities, Challenges, and Solutions. JACC: Advances, 4(3), 101593.

[10] Olsen, S., Gooch, B. (2011, August). Image simplification and vectorization. In Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Non-Photorealistic Animation and Rendering (pp. 65-74).

[11] Zhang, Q. (2022). A novel ResNet101 model based on dense dilated convolution for image classification. SN Applied Sciences, 4(1), 9.

[12] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., ... Ng, A. Y. (2017). Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225.

[13] Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., ... Natarajan, V. (2023). Large language models encode clinical knowledge. Nature, 620(7972), 172-180.

[14] Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., ... Farhan, L. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. Journal of big Data, 8(1), 53.

[15] Yin, H., Vahdat, A., Alvarez, J. M., Mallya, A., Kautz, J., Molchanov, P. (2022). A-vit: Adaptive tokens for efficient vision transformer. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10809-10818).

[16] Wang, Y., Deng, Y., Zheng, Y., Chattopadhyay, P., Wang, L. (2025). Vision transformers for image classification: A comparative survey. Technologies, 13(1), 32.

[17] Wang, Y., Li, Y., Ma, Y., Li, X., Li, Z., He, X., ... Zhang, S. (2023). METransformer: Radiology report generation by transformer with multiple learnable expert tokens. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 10166–10175).

[18] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[19] Moon, J., Shin, S., Lee, H., Lee, S., Lee, S., Park, S., ... Kim, H. (2022). Multi-modal understanding and generation for medical images and text via vision-language pre-training. IEEE Transactions on Medical Imaging, 41(11), 2927–2940.

[20] Moon, J., Shin, S., Lee, H., Lee, S., Lee, S., Park, S., ... Kim, H. (2022). Multi-modal understanding and generation for medical images and text via vision-language pre-training. IEEE Transactions on Medical Imaging, 41(11), 2927–2940.

[21] Lehman, E., Hernandez, E., Mahajan, D., Wulff, J., Smith, M. J., Ziegler, Z., ... Szolovits, P. (2023). ClinicalT5: Large Language Models Built Using MIMIC Clinical Text. In Proceedings of Machine Learning for Healthcare Conference.

[22] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, 'An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale," *arXiv preprint arXiv:2010.11929*, 2021.

[23] M. Shamshad, 'Transformers in medical imaging: A survey," *Medical Image Analysis*, vol. 88, Art. no. 102548, 2023.

[24] F. Zhang, M. Zhao, and H. Li, 'Contrastive Learning of Medical Visual Representations from Paired Images and Text," in *Proc. Machine Learning for Healthcare*, 2022.

[25] S. Ruder, 'An Overview of Multi-Task Learning in Deep Neural Networks," *arXiv preprint arXiv:1706.05098*, 2017.

[26] M. Ghassemi, K. Oakden-Rayner, and A. Beam, 'The false hope of current approaches to explainable artificial intelligence in health care," *The Lancet Digital Health*, vol. 3, no. 11, pp. e745–e750, 2021.

[27] M. Rubin, A. Kotval, and T. Mungall, 'Large Scale Automated Reading of Frontal and Lateral Chest X-Rays using Dual Convolutional Neural Networks," *arXiv preprint arXiv:1801.07703*, 2018.

[28] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. NAACL*, 2019, pp. 4171–4186.