RadTextAid : A Novel Efficient Pipeline Utilizing Lightweight Multi-Modal Models To Develop Chest X-ray Report Writing Assistance For Radiologists

Mahmud Wasif Nafee and **Tasmia Rahman Aanika** and **Dr. Taufiq Hasan**
Bangladesh University of Engineering and Technology (BUET)
Department of BME , ECE Building , West Palashi
Dhaka-1205, Bangladesh

Abstract

Deciphering chest X-rays plays a critical role in diagnosing diseases such as pneumonia, lung cancer, and cardiomegaly. Radiologists often contend with significant workloads, managing large volumes of data under resource constraints, which can lead to exhaustion and burnout. To tackle these challenges, this study explores the use of advanced deep learning techniques, including multi-modal systems like Vision-Language Models (VLMs), to automate the generation of chest X-ray reports, aiming to improve both accuracy and efficiency.

The primary objective is to develop a system capable of identifying diagnostic labels, producing clinically standardized reports, and accommodating various imaging modules. By leveraging multi-label classifiers alongside object detection algorithms, this approach enhances the detection of abnormalities and the overall quality of generated reports, facilitating quicker and more informed therapeutic decisions.

In summary, this research seeks to optimize diagnostic processes, alleviate radiologists' workloads, and elevate patient care by integrating AI technologies into the healthcare system, ultimately fostering greater efficiency and effectiveness.

Introduction

Interpreting chest X-rays is a crucial task in the medical field, as it plays a significant role in diagnosing a wide range of diseases. The ability to accurately read and interpret these images is vital for timely and effective patient care. However, the process is complex and demands a high level of expertise and attention to detail from radiologists. Radiologists frequently dedicate a significant amount of time to carefully analyze every chest X-ray. Conducting a comprehensive investigation is crucial in order to prevent misdiagnosis, which can lead to significant repercussions for the patient's well-being. Nevertheless, this laborious procedure might result in inefficiencies within the healthcare system, especially in environments with a large number of patients or a lack of skilled radiologists.

Automated Radiography Report Generation can be a solution to this problem. However, automated radiological text generation comes with its own challenges. One of the key challenges is being able to detect clinical abnormalities by observing chest X-Rays and then document those clearly in the reports.Chest X-ray reports contain information such as medical history, prior report references, routine physiological observations, and pathological findings. While the first two cannot be inferred from the X-ray, the latter two—routine observations and abnormal findings—can be detected visually, enabling an automated model to extract features and identify such findings. Chest X-ray reports in publicly available datasets often contain mostly routine physiological observations, with limited references to medical history or clinical abnormalities. This abundance of routine information poses a challenge for state-of-the-art Vision Language Models (VLMs) in automated report generation, as these models rely on token loss. When captions are largely similar, the models struggle to learn, hindering their ability to generate differential annotations, such as clinical abnormalities, in the reports. Though the abnormality annotations are important pieces of information according to multiple radiologists we interviewed, the model does not get heavily penalized when predicting the wrong pathological condition or outright ignoring it, simply generating text related to routine observations.

To address the issue of failing to generate pathological findings, one approach is to train the VLM on text related to clinical abnormalities only, rather than the entire report. This will focus the model on generating pathological findings, which require more attention and time. Another approach is using a CNN-based classifier to detect CXR abnormalities, which can then guide the VLM with targeted prompts for better results.

Inspired by the above observations, we propose RadTextAid, a novel efficient pipeline to train lightweight multimodal model with the goal of assisting radiologists in writing reports by generating findings related to clinical abnormalities. The pipeline will consist of a multimodal vision language model known as PaliGemma that will be trained to generate the findings, Llama 3.1 8b to extract only the pathological findings to prepare the training corpus and a CheXNet classifier to generate tags needed for prompt guidance of the vision language model. We summarize contributions as follows.

- We propose a novel efficient pipeline to train vision lan-

guage models for domain specifics task such as X-ray report generation.

- We investigate potential of the latest state-of-the-art large language models to extract clinically critical information from medical reports with few shot prompting
- We investigate potential of Convolutional Neural Network generated outputs to construct prompt to guide the vision language model

Related Works

Automated Chest X-ray Report Generation

Chest radiography holds the position of being the most prevalent imaging examination worldwide, playing a crucial role in the screening, diagnosis, and treatment of numerous life-threatening illnesses. In contrast to other image captioning tasks that prioritize coherence, medical image captioning necessitates a greater emphasis on accuracy in identifying anomalies and extracting information, while still maintaining coherence. This means that the generated report should be easily comprehensible and effectively convey precise medical information.(Srinivasan et al. 2020)

Recent developments in computational machine translation, as noted in Ref (Sirshar et al. 2022), have shown that by utilizing a robust sequence model, significant progress can be achieved in obtaining state-of-the-art results. This is accomplished by explicitly optimizing the probability of successful translation in an end-to-end manner, where the input sequence is provided for both training and inference purposes. In this architecture, the decoder is implemented using a long short-term memory (LSTM) network, which is further enhanced by incorporating attention mechanism. The attention mechanism operates based on the same principle of utilizing the source information for the target language conversion. The Contrastive Attention (CA) approach is put out by Liu et al.(acl) to efficiently capture and characterise aberrant regions. The CA model distils the contrastive information by comparing the current input image with normal images rather than concentrating just on it. Kaur et al.(Kaur and Mittal 2022) introduce a deep neural network called RadioBERT that utilizes contextual word representations to generate valuable radiological reports from CXR images. This network incorporates distilBERT for contextual word representation and applies sentiment analysis to rearrange the generated sentences, ensuring that abnormal descriptions are placed at the beginning of the report. Srinivasan et al. (Srinivasan et al. 2021) introduce a deep neural network as a means to accomplish this task. The proposed network utilises a set of Chest X-Ray images to predict the medical tags and provide a comprehensible radiology report.

To overcome limits of typical RNNs, use of Transformer is now being initiated. CNX-B2(Alqahtani et al. 2024) is a Convolutional Neural Network (CNN) combined with a Transformer approach to generate medical reports. The work of (Alfarghaly et al. 2021) involves fine-tuning a pre-trained ChexNet model to predict specific identifiers from the image, generating weighted semantic features from the pre-trained embeddings of the predicted tags, and producing complete medical reports by conditioning a pre-trained GPT-2 model on the visual and semantic features. A study (Tsaniya, Fatichah, and Suciati 2024) developed a model for generating medical reports by utilizing the transformer approach and implementing image enhancement techniques. In their paper (Mondal et al. 2023), the authors introduce EfficienTransNet, an automatic chest X-ray report generation approach based on CNN-Transformers. EfficienTransNet incorporates clinical history or indications to enhance the report generation process and align with radiologists' workflow, which is mostly overlooked in recent research.

In order to enhance the creation of CXR reports, (Nicolson, Dowling, and Koopman 2023) propose the exploration of warm starting the encoder and decoder using up-to-date computer vision and natural language processing checkpoints, such as the Vision Transformer (ViT) and PubMedBERT. TrMRG (Mohsan et al. 2023), also known as the Transformer Medical report generator, is a comprehensive model that utilizes the Transformer architecture to generate reports. This model incorporates pre-trained computer vision (CV) and language models, making it a powerful tool for report generation. The multimodal model proposed in (Veras Magalhães et al. 2024) utilizes the Swin Transformer as the image encoder. This enables the model to perform hierarchical mapping and enhance its perception without significantly increasing the number of parameters or computational costs. Additionally, the model incorporates GPT-2 as the textual decoder. (Wang et al. 2023) introduces ME-Transformer that incorporates a transformer framework and includes "expert tokens" into both the transformer encoder and decoder, representing many experts.

Method

Framework Overview

As illustrated in Figure 1 , we start the framework with a Chest X-ray image as input, which flows through a multi-label classifier based on CheXNet-121. This classifier scans the available image for relevant features and produces 105 diagnostic tags that show the detected conditions or abnormalities. These tags provide semantic information about the X-ray findings and are visualized in the system diagram as individual elements, offering clarity and interpretability of the multi-label classification results.

Next, the initial Chest X-ray image along with the produced tags, which are text outputs, are then fed into a multimodal Vision-Language Model (VLM) like Florence 2 or PaliGemma. The VLM, fuses the visual characteristics from the image itself along with the semantic information obtained by the transforming tags, allowing for an encyclopedic analysis of each datapoint. The system uses this data to create a detailed diagnostic report with observations of inferred conditions and abnormalities along with extra clinical insights. The proposed framework utilizes multi-modal analysis aimed at improving the accuracy and efficiency of the diagnostic process, so that it can bring positive contribution when reporting Chest X-ray images.

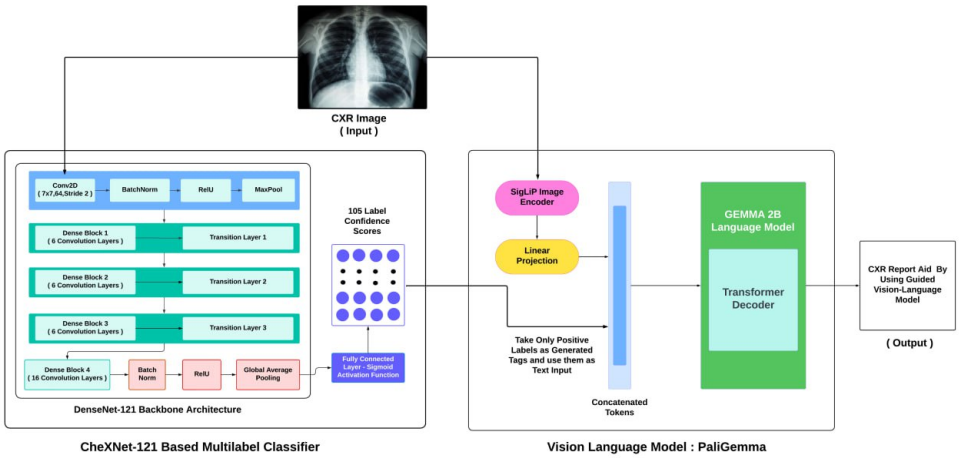


Figure 1: Our proposed RadTextAid Model Architecture.

Hence, we can break this framework into two important parts:

1. The multi-label classifier based on CheXNet-121 and 105 Tags
2. The multi-modal Vision-Language Model (VLM), i.e. Florence 2 or PaliGemma

The multi-label classifier based on CheXNet-121 and 105 Tags

The chest x-ray image is first passed through a CNN model to produce the tags' predictions. Our base model is a Chexnet (Rajpurkar et al. 2017), which is a Densenet121 model (Huang, Liu, and Weinberger 2016) pre-trained on ChestX-ray14 dataset to detect and localize 14 types of diseases or anomalies from the images. However, 14 tags were determined not to be sufficient for our task of conditioning report generation. So, we chose a fine-tuned model to classify the manual tags from the IU-Xray dataset (Demner-Fushman et al. 2015) by removing the final layer and adding a new final layer containing 105 nodes for the most occurring manual tags from the dataset.

The positive tags here indicate the current physiological conditions present in the chest X-ray. By passing only these as prompts to the next stage, we can further provide an attention system for the model for extracting features and generating the final text report.

The model typically uses a multi-label classification loss function, such as the binary cross-entropy (BCE) loss, to handle multiple diagnostic labels. So, the average loss per sample within a batch is calculated by :

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M}\left[y_{ij}\log(\hat{y}_{ij}) + (1-y_{ij})\log(1-\hat{y}_{ij})\right]$$

Where, N denotes the number of samples (batch size), M denotes the number of labels (e.g., $M = 105$ for 105 diagnostic tags), $y_{ij} \in \{0, 1\}$ is the ground truth binary label for the j-th class of the i-th sample, $\hat{y}_{ij} \in [0, 1]$ is the predicted probability for the j-th class of the i-th sample, and $\log$ is the natural logarithm.

The idea was inspired from the visual feature extraction part of CDGPT2 (Alfarghaly et al. 2021) , a conditioned transformer model to generate radiology reports.

The multi-modal Vision-Language Model (VLM), i.e. Florence 2 or PaliGemma

The input image and generated tags (the positive tags only that were found from the multi-label classifier) are then passed through a multi-modal VLM, like - pretrained Florence-2 or PaliGemma. Here, the tags act as text prompts.

Florence-2

Florence-2 is a novel Vision Foundation Model (Xiao et al. 2023) for various vision and vision-language tasks with a unified sequence-to-sequence architecture. It uses **DaViT (Dual Attention Vision Transformer)** (Ding et al. 2022) as its vision encoder to process images into token embeddings $V \in \mathbb{R}^{N_v \times D_v}$, where N_v represents the number of visual tokens and D_v their dimensionality. An extended version of the tokenizer combines these embeddings with task-related text prompts $T_{\text{prompt}} \in \mathbb{R}^{N_t \times D}$.

The model uses a multi-modal transformer-based encoder-decoder , with input $X = [V', T_{\text{prompt}}]$, where V' is the dimensionally aligned projection of V. It uses a standard cross-entropy language modeling objective for training:

$$\mathcal{L} = -\sum_{i=1}^{|y|}\log P_\theta(y_i \mid y_{<i}, x),$$

where, y represents the target sequence, and x combines visual and textual inputs.

Trained on FLD-5B, a large-scale dataset with 126M images and more than 5 billion annotations (text, region-text pairs, text-phrase-region triplets), this empowers the model to learn multiple levels of spatial and semantic granularity (PLN) [33]. It attains SOTA performances, such as 143.3 CIDEr in COCO Captioning and 55.5 mIoU in ADE20K segmentation, exceeding the results of larger models while remaining lightweight with relatively fewer parameters (e.g., Florence-2-L with 0.77 billion parameters).

Although Florence-2 is a remarkably versatile model with state-of-the-art performance, training of Florence-2 requires significant computational resources and is heavily dependent on large datasets, thereby limiting its accessibility in low-resource settings.

PaliGemma

PaliGemma is a unified VLM (Beyer et al. 2024) for general-use multi-modal functionalities based on **SigLIP ViT-So400m image encoder** (Zhai et al. 2023) and the **Gemma-2B decoder-only language model** (Team et al. 2024) with fewer than 3 billion parameters.

The SigLIP encoder uses a contrastive pretraining method with sigmoid loss to produce image embeddings, achieving SOTA clip-level visual representation quality. The textual inputs to the Gemma-2B model are processed using a SentencePiece tokenizer ; and then for autoregressive decoding, image tokens and text prompts can be combined into a single sequence. A linear projection aligns the SigLIP output with Gemma-2B's input space, facilitating seamless integration.

PaliGemma adopts a multi-stage pretraining strategy. To begin with, all unimodal components are pretrained separately. Next comes multimodal pretraining, where all the layers in the whole model (except occasionally for the image encoder) are trained on a large variety of vision-language tasks so that it knows how to deal with spatial and relational information correctly.

The model is then further trained on higher-resolution images and domain-focused data in final stages, which leads to better accuracy. The model performs well on all tasks, including captioning, visual question answering (VQA), segmentation and even domain-specific ones like remote sensing and video QA.

So, mathematically (for PaliGemma) input sequence looks like this:

$$\text{tokens} = [\text{image tokens} \ldots, \text{BOS}, \text{prefix tokens} \ldots, \text{SEP}, \text{suffix tokens} \ldots, \text{EOS}, \text{PAD} \ldots]$$

PaliGemma, though comparatively smaller than other models, offers state-of-the-art performance on the various benchmarks making it efficient and flexible too. Nevertheless, the requirement for massive amounts of pretraining data and processing power makes the model hard to use for smaller scale deployments.

Training Corpus Report Pre-Processing

In order to ensure the accuracy and relevance of the generated diagnostic reports, the pre-processing of the training corpus reports focused on eliminating usage of routine or repetitive words as usually contained in Chest X-ray reports. Such words are common in normal cases and can cause skewed results where the Vision-Language Model (VLM) generates a lot of general terms in order to avoid loss in training.

To address this , the dataset is adjusted by utilizing a filtering process with the use of the pre-trained **Llama 3.1-8B model** (Sam and Vavekanand 2024) (Shown in Figure 2).

Experiments

Datasets

MIMIC-CXR Database v2.0.0

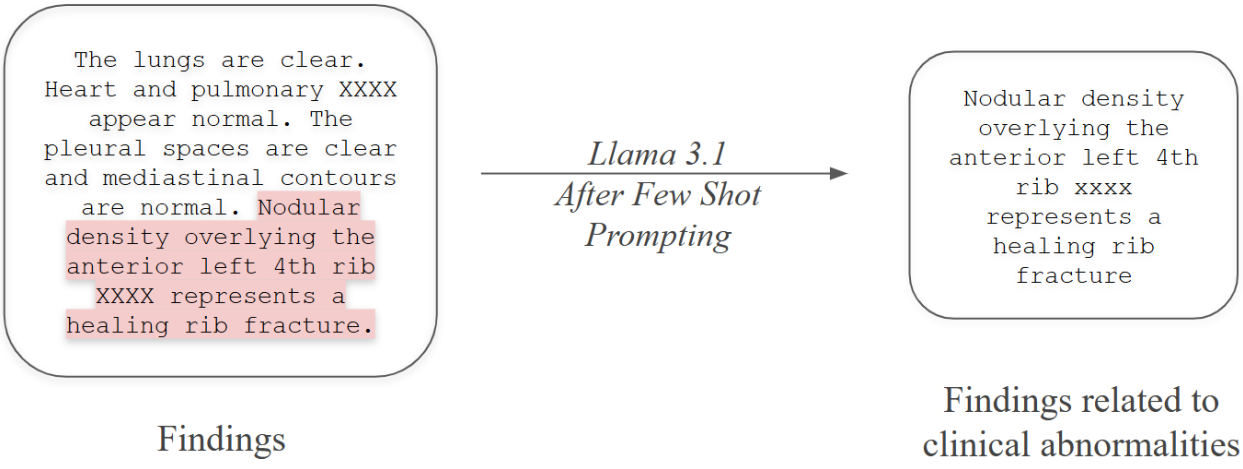


Figure 2: Pre-Processing the Reports before Training.

The publically available, large-scale MIMIC-CXR Database v2.0.0 is intended to facilitate medical imaging research, especially in the area of chest radiography.The dataset contains 377,110 JPG format images and structured labels derived from the 227,827 free-text radiology reports associated with these images. Images in the collection are available in JPG format.(Johnson et al. 2024)

Indiana University Chest X-rays and Reports (OpenI)

Accessible via OpenI, the Indiana University Chest X-rays dataset is an extensive collection of chest radiographs with their corresponding diagnostic reports intended to facilitate medical imaging research and education.The dataset contains 7,470 pairs of images and reports that covers a broad spectrum of both common and unusual thoracic disorders. (Demner-Fushman et al. 2015)

Evaluation Metrics

To evaluate the performance of our finetuned VLM, we have decided to rely on two metrics: BERTScore and F1-cheXbert

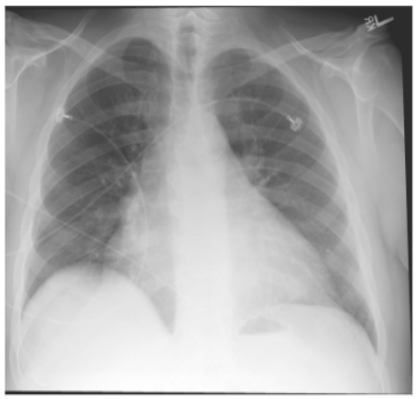
We opt for BERTScore because it is an automated assessment measure for text production. Similar to conventional metrics, BERTScore calculates a similarity score for every token in the candidate sentence in relation to each token in the reference phrase. Instead of relying on exact matches, the system calculates token similarity by utilising contextual embeddings. (Zhang et al. 2020)

The F1-cheXbert (Smit et al. 2020) score utilizes Chexbert transformer to output selected lables on both original and generated reports and then calculates the F1 score between these two sets of labels. To be consistent with previous works, the score is calculated over 5 observations: atelectasis, cardiomegaly, consolidation, edema and pleural effusion.

Implementation details

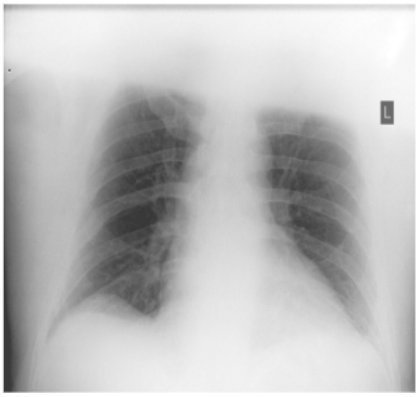
The multilabel classification model is implemented using TensorFlow. With 32 photos per batch and the Adam optimizer, the model was trained end-to-end using mini-batch gradient descent. Binary cross-entropy loss (eq.) was the loss that needed to be examined in this model. It was decided to leave all the model parameters up for fine-tuning.

On the other hand , the Vision-Language Models (VLMs) are implemented using PyTorch. Both models are trained and tested on Intel(R) Xeon(R) CPUs and L4 GPUs provided by Google Colab notebooks. For the multilabel classification task, binary cross-entropy loss is used, while the loss function provided by the VLM packages are used for training the Vision-Language Models. The Adam optimizer with an initial learning rate of $5*10^{-5}$ and a batch size of 16 for multilabel classifiers and 4 for the Vision Language Models is employed for training or finetuning. The VLMs were finetuned for 10 epochs.



Reference findings	there is aortic xxxx vascular calcification
Generated findings	there is aortic vascular calcification

(a) Report Aid Example 1



Reference findings	low lung volumes, mild bibasilar atelectasis
Generated findings	low lung volumes mild bibasilar atelectasis and pleural thickening is noted

(b) Report Aid Example 2

Reference findings	nodular densities consistent with chronic granulomatous disease
Generated findings	evidence of prior granulomatous disease

(c) Report Aid Example 3

Figure 3: Some Report Examples generated from our Proposed Model.

Results

Qualitative analysis of abnormal findings extraction: A Llama 3.1 8b instruct model was used to extract sentences or phrases in the findings that indicated clinical abnormalities. Zero shot prompting the Llama model with the instruction "Extract the sentences or phases that seem to indicate clinical abnormalities" didn't generate fruitful results. That is why a radiologist's help was taken to create a few examples of separating the clinical abnormal findings from the rest of the report. With 12 examples, few shot prompting the Llama model with the same instruction showed much better results as shown in fig.2

Vision Language model comparison and necessity of prompt guidance: There were two possible state-of-the-art lightweight Vision Language Model options with multimodal inputs to use in our pipeline: Paligemma (Beyer et al. 2024) by Google and Florence-2 (Xiao et al. 2023) by Microsoft. While comparing the models, we also tested their performances with a generic prompt("Write a Chest X-ray report") and a tag specific prompt (such as "Write a Chest X-ray report mentioning cardiomegaly", "Write a Chest X-ray report mentioning pleural effusion, consolidation" etc.) From Table-1, it is obvious that Paligemma far outperforms Florence-2 in all the BERTScore metrics and the F1-cheXbert metric. In cases of both VLMs, it is clear that tag-specific prompts enhance the performance of the model. For Florence-2, tag-specific prompts increase the BERTScore (at least 4%) and increases F1-cheXbert score from 0.5516 to 0.700, that is, it generates almost 14.84% more accurate reports. For Paligemma, both types of prompts result in very similar BERTScore metrics, but the F1-cheXbert score shows that the model can generate 2.5% more accurate reports with specific prompt guidance. With the help of this comparative analysis, we can determine that Paligemma with tag-specific prompt guidance should be utilized in our pipeline.

Comparison with captioning baselines: We compared our final methods to three types of models: (1) Convolutional Neural Network encoders with Recurrent Neural Network decoders. (2) Convolutional Neural Network encoders with Transformer decoders. (3) Vision Transformer encoders with LLM decoders. For type-1, we look at attention-based CNN-LSTM architecture mentioned in ref (Sirshar et al. 2022) (acl). For comparison against our proposed methods, we follow the architecture of (Sirshar et al. 2022). Type-2 entails CNN as encoders and transformers as decoders. We saw such work being done in ref (Alqahtani et al. 2024),(Alfarghaly et al. 2021),(Tsaniya, Fatichah, and Suciati 2024) ,(Mondal et al. 2023). For a comparative evaluation, we used the architecture CNX-B2 as described in ref (Alqahtani et al. 2024). Transformers were used end-to-end for both encoding and decoding in type-3 captioning baselines as seen in ref (Wang et al. 2023),(Nicolson, Dowling, and Koopman 2023),(Veras Magalhães et al. 2024). We explored the architecture CvT2DistilGPT2 described in ref (Nicolson, Dowling, and Koopman 2023) for our quantitative analysis.

Table-2 shows the results for both MIMIC-CXR and IU X-Ray. From the table, it is evident that the proposed method achieves a considerable improvement over the baselines. When compared with the best existing method, CvT2DistilGPT2, RadTextAid shows an absolute improvement of 4.8% in the NLG metric (BERTScore) and generates 3.16% more accurate reports according to the F1-cheXbert score. Similar trends are observed in the results for MIMIC-CXR. However, the overall performance of all models drops, which is likely due to the significantly higher average findings length in the MIMIC-CXR dataset compared to the IU X-Ray dataset.

Qualitative results: We present a few qualitative examples of RadTextAid to demonstrate its superiority. The model, trained using our pipeline, has successfully detected and described all the clinical abnormalities as seen in the original report. In the second example, however, the model mistakenly detects and describes a clinical abnormality that is not mentioned in the reference report. This error may have occurred due to the image quality. A more robust preprocessing of the image could help resolve this issue.

Table 1: Comparative Analysis of the two Vision-Language Models.

Used Dataset : OpenI

VLM	BERTScore Precision	BERTScore Recall	BERTScore F1	F1-cheXbert
Florence-2 (with generic prompts)	0.2467	0.2721	0.2507	0.5516
Florence-2 (with tag specific prompts)	0.2753	0.2874	0.2608	0.7000
Paligemma (with generic prompts)	0.3181	0.3334	0.3173	0.7666
Paligemma (with tag specific prompts)	0.3273	0.3322	0.3156	0.7916

Conclusion

The study highlights the effectiveness of a novel complex Vision-Language Model (VLM) framework for developing automated diagnostic aids for chest X-ray diagnosis. The starting point of the model is a CheXNet-121 based multi-

Table 2: Comparative Analysis against Literature Captioning Baselines.

Used Dataset : OpenI

Pipeline	BERTScore Precision	BERTScore Recall	BERTScore F1	F1-cheXbert
Attention-based CNN-LSTM	0.2174	0.2018	0.2123	0.5451
CNX-B2	0.2194	0.1608	0.2268	0.6484
CvT2--Distil--GPT2	0.3089	0.2996	0.3009	0.7600
Rad--TextAid (Proposed)	0.3273	0.3322	0.3156	0.7916

Used Dataset : MIMIC-CXR

Pipeline	BERTScore Precision	BERTScore Recall	BERTScore F1	F1-cheXbert
Attention-based CNN-LSTM	0.1871	0.1926	0.1852	0.5347
CNX-B2	0.1886	0.1606	0.2265	0.5333
CvT2--Distil--GPT2	0.2939	0.2927	0.2902	0.6813
Rad--TextAid (Proposed)	0.2813	0.3034	0.2975	0.6956

label classifier which identifies 105 diagnostic tags from a chest X-ray image associated with different physiological conditions and pathologies. These tags , supplemented with the original image , are then sent through a multi-modal VLM like Florence 2 or PaliGemma, which combines image information with semantic information to create coherent, relevant reports. This integration enhances the interpretability of the results while ensuring clinical relevance.

A key innovation lies in the pre-processing of the training corpus , using the Llama 3.1 pre-trained model, which eliminates all overworked and overused vernacular from the text, which means that the reports generated will be based on only the abnormalities that are of diagnostic import. This helps in reducing the duplication, improving the clinical value, and serves as a very strong base for automatic report creation.

The results validate the strength of the proposed framework, with the presence of quantitative measures such as BERTScore and F1-cheXbert confirming the model's capacity to produce relevant and accurate descriptive diagnosis. Although other architectures were included in the study for comparative analysis, our framework surpasses them all on integrating multi-modal learning approaches and achieving better outcomes, proving its efficacy to optimize radiological processes and enhance patient care.

References

aclanthology.org. https://aclanthology.org/2021.findings-acl.23.pdf. [Accessed 24-06-2024].

Alfarghaly, O.; Khaled, R.; Elkorany, A.; Helal, M.; and Fahmy, A. 2021. Automated radiology report generation using conditioned transformers. *Informatics in Medicine Unlocked* 24:100557.

Alqahtani, F. F.; Mohsan, M. M.; Alshamrani, K.; Zeb, J.; Alhamami, S.; and Alqarni, D. 2024. Cnx-b2: A novel cnn-transformer approach for chest x-ray medical report generation. *IEEE Access* 12:26626–26635.

Beyer, L.; Steiner, A.; Pinto, A. S.; Kolesnikov, A.; Wang, X.; Salz, D.; Neumann, M.; Alabdulmohsin, I.; Tschannen, M.; Bugliarello, E.; Unterthiner, T.; Keysers, D.; Koppula, S.; Liu, F.; Grycner, A.; Gritsenko, A.; Houlsby, N.; Kumar, M.; Rong, K.; Eisenschlos, J.; Kabra, R.; Bauer, M.; Bošnjak, M.; Chen, X.; Minderer, M.; Voigtlaender, P.; Bica, I.; Balazevic, I.; Puigcerver, J.; Papalampidi, P.; Henaff, O.; Xiong, X.; Soricut, R.; Harmsen, J.; and Zhai, X. 2024. Paligemma: A versatile 3b vlm for transfer.

Demner-Fushman, D.; Kohli, M.; Rosenman, M.; Shooshan, S.; Rodriguez, L.; Antani, S.; Thoma, G.; and Mcdonald, C. 2015. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association : JAMIA* 23.

Ding, M.; Xiao, B.; Codella, N.; Luo, P.; Wang, J.; and Yuan, L. 2022. Davit: Dual attention vision transformers.

Huang, G.; Liu, Z.; and Weinberger, K. Q. 2016. Densely connected convolutional networks. *CoRR* abs/1608.06993.

Johnson, A.; Lungren, M.; Peng, Y.; Lu, Z.; Mark, R.; Berkowitz, S.; and Horng, S. 2024. MIMIC-CXR-JPG - chest radiographs with structured labels.

Kaur, N., and Mittal, A. 2022. Radiobert: A deep learning-based system for medical report generation from chest x-ray images using contextual embeddings. *Journal of Biomedical Informatics* 135:104220.

Mohsan, M. M.; Akram, M. U.; Rasool, G.; Alghamdi, N. S.; Baqai, M. A. A.; and Abbas, M. 2023. Vision trans-

former and language model based radiology report generation. *IEEE Access* 11:1814–1824.

Mondal, C.; Pham, D.-S.; Gupta, A.; Ghosh, S.; Tan, T.; and Gedeon, T. 2023. EfficienTransNet: An automated chest x-ray report generation paradigm. In *Proceedings of the 1st International Workshop on Multimodal and Responsible Affective Computing*. New York, NY, USA: ACM.

Nicolson, A.; Dowling, J.; and Koopman, B. 2023. Improving chest x-ray report generation by leveraging warm starting. *Artificial Intelligence in Medicine* 144:102633.

Rajpurkar, P.; Irvin, J.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D. Y.; Bagul, A.; Langlotz, C. P.; Shpanskaya, K. S.; Lungren, M. P.; and Ng, A. Y. 2017. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *CoRR* abs/1711.05225.

Sam, K., and Vavekanand, R. 2024. Llama 3.1: An in-depth analysis of the next generation large language model.

Sirshar, M.; Paracha, M. F. K.; Akram, M. U.; Alghamdi, N. S.; Zaidi, S. Z. Y.; and Fatima, T. 2022. Attention based automated radiology report generation using CNN and LSTM. *PLoS One* 17(1):e0262209.

Smit, A.; Jain, S.; Rajpurkar, P.; Pareek, A.; Ng, A. Y.; and Lungren, M. P. 2020. Chexbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using bert.

Srinivasan, P.; Thapar, D.; Bhavsar, A.; and Nigam, A. 2020. Hierarchical x-ray report generation via pathology tags and multi head attention. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*.

Srinivasan, P.; Thapar, D.; Bhavsar, A.; and Nigam, A. 2021. Hierarchical x-ray report generation via pathology tags and multi head attention. In *Computer Vision – ACCV 2020*, Lecture notes in computer science. Cham: Springer International Publishing. 600–616.

Team, G.; Mesnard, T.; Hardin, C.; Dadashi, R.; Bhupatiraju, S.; Pathak, S.; Sifre, L.; Rivière, M.; Kale, M. S.; Love, J.; Tafti, P.; Hussenot, L.; Sessa, P. G.; Chowdhery, A.; Roberts, A.; Barua, A.; Botev, A.; Castro-Ros, A.; Slone, A.; Héliou, A.; Tacchetti, A.; Bulanova, A.; Paterson, A.; Tsai, B.; Shahriari, B.; Lan, C. L.; Choquette-Choo, C. A.; Crepy, C.; Cer, D.; Ippolito, D.; Reid, D.; Buchatskaya, E.; Ni, E.; Noland, E.; Yan, G.; Tucker, G.; Muraru, G.-C.; Rozhdestvenskiy, G.; Michalewski, H.; Tenney, I.; Grishchenko, I.; Austin, J.; Keeling, J.; Labanowski, J.; Lespiau, J.-B.; Stanway, J.; Brennan, J.; Chen, J.; Ferret, J.; Chiu, J.; Mao-Jones, J.; Lee, K.; Yu, K.; Millican, K.; Sjoesund, L. L.; Lee, L.; Dixon, L.; Reid, M.; Mikuła, M.; Wirth, M.; Sharman, M.; Chinaev, N.; Thain, N.; Bachem, O.; Chang, O.; Wahltinez, O.; Bailey, P.; Michel, P.; Yotov, P.; Chaabouni, R.; Comanescu, R.; Jana, R.; Anil, R.; McIlroy, R.; Liu, R.; Mullins, R.; Smith, S. L.; Borgeaud, S.; Girgin, S.; Douglas, S.; Pandya, S.; Shakeri, S.; De, S.; Klimenko, T.; Hennigan, T.; Feinberg, V.; Stokowiec, W.; hui Chen, Y.; Ahmed, Z.; Gong, Z.; Warkentin, T.; Peran, L.; Giang, M.; Farabet, C.; Vinyals, O.; Dean, J.; Kavukcuoglu, K.; Hassabis, D.; Ghahramani, Z.; Eck, D.; Barral, J.; Pereira, F.; Collins, E.; Joulin, A.; Fiedel, N.; Senter, E.; Andreev, A.; and Kenealy, K. 2024. Gemma: Open models based on gemini research and technology.

Tsaniya, H.; Fatichah, C.; and Suciati, N. 2024. Automatic radiology report generator using transformer with contrast-based image enhancement. *IEEE Access* 12:25429–25442.

Veras Magalhães, G.; L de S Santos, R.; H S Vogado, L.; Cardoso de Paiva, A.; and de Alcântara Dos Santos Neto, P. 2024. XRaySwinGen: Automatic medical reporting for x-ray exams with multimodal model. *Heliyon* 10(7):e27516.

Wang, Z.; Liu, L.; Wang, L.; and Zhou, L. 2023. METransformer: Radiology report generation by transformer with multiple learnable expert tokens.

Xiao, B.; Wu, H.; Xu, W.; Dai, X.; Hu, H.; Lu, Y.; Zeng, M.; Liu, C.; and Yuan, L. 2023. Florence-2: Advancing a unified representation for a variety of vision tasks. *arXiv preprint arXiv:2311.06242*.

Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid loss for language image pre-training.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2020. Bertscore: Evaluating text generation with bert.