Rad-Flamingo: A Multimodal Prompt driven Radiology Report Generation Framework with Patient-Centric Explanations

Anonymous ACL submission

Abstract

In modern healthcare, radiology plays a pivotal role in diagnosing and managing diseases. However, the complexity of medical imaging data and the variability in interpretation can lead to inconsistencies and a lack of patient-centered insights in radiology reports. To address these challenges, we propose a novel multimodal prompt-driven report generation framework **Rad-Flamingo**, that integrates diverse data modalities—such as medical images, and clinical notes—to produce comprehensive and context-aware radiology reports. Our framework leverages innovative prompt engineering techniques to guide vision-language models in synthesizing relevant information, ensuring the generated reports are not only accurate but also understandable to individual patients. A key feature of our framework is its ability to provide patient-centric explanations, offering clear and personalized insights into diagnostic findings and their implications. Experimental results demonstrate this framework's effectiveness in enhancing report quality, improving understandability, and could foster better patient-doctor communication. This approach represents a significant step towards more intelligent, transparent, and human-centered medical AI systems.

1 Introduction

Radiology reports form the basis for clinical diagnostics and guide medical experts in treating patients. Despite their significance, creating radiology reports is a labor-intensive and expert-intensive process frequently plagued with human errors and differing details based on the radiologist's level of experience. Given the very low ratio of radiologists to patients, the laborious process of creating full text radiology reports ends up being one of the workflow's largest obstacles (US, China, and India is 1:10,000, 1:14,772, and 1:100,000, respectively) (Arora, 2014). Given the huge number of cases and the shortage of radiology experts, time-efficiently generating reports is a major hurdle worldwide. Towards this goal, there has been a huge attempt from both industry and academia, with the landscape of AI-based report generation witnessing exponential growth in recent times (Messina et al., 2022). This growth is owed to the evolving capabilities of large language models and vision language models (VLMs) in particular and VLMs have showcased exceptional abilities on a variety of tasks, such as image captioning (Hossain et al., 2019), visual question answering (Lu et al., 2023), and visual common sense reasoning (Zellers et al., 2018). VLMs such as (Thawakar et al., 2024; Moor et al., 2023) show promising efficacy in aligning image with text for medical use cases.

1.1 Motivation

VLMs find an excellent application in generation of radiology reports. However, all generative pretrained models are opaque by design. Report generation systems which are able to generate reports with explanations are better placed to build trust and acceptability amongst patients. Such explanations in case of radiology reports could be **patient-centric or expert-centric**. Patient centric explanations are lucid generated texts, that paraphrases medical keywords in the report while explaining the pathophysiology of the condition in easy to understand language. Furthermore, recent research has demonstrated that large language models can also rationalize their own prediction (Wiegreffe et al., 2021) giving the model an ability to give natural language explanations for its own generated responses. Combining the generation capabilities of VLMs and their self-rationalization abilities, we generate coherent radiology reports along-with patient centric explanations [1].

Generating radiology reports using prompting

[1] All our datasets and scripts will be publicly released.

strategies, let alone multimodal prompting is an under-explored domain. Driven by this motivation, we developed a two step multimodal in-context learning strategy to generate radiology reports along with patient-centric explanations. In the first stage we design few-shot prompts following the standard in-context learning template. For this stage we take an open source VLM model Mini-GPT4 (Zhu et al., 2023) fine-tuned on MIMIC-CXR-JPG dataset (Johnson et al., 2019). This stage acts as the synthetic data generator, which annotates each of the image-report instance with a patient-centric explanation. For verifying the explanations we rely only on medical expert evaluations. Following this, we design our multimodal in-context learning strategy on Med-Flamingo (a fine-tuned Flamingo model) (Moor et al., 2023) to generate a structured radiology report along with patient-centric explanations. We evaluate the outcomes by utilizing classical NLG metrics (BLEU, ROUGE, METEOR) as well as medical expert evaluation score. Further since for medical texts semantic similarity has paramount importance compared to lexical similarity we utilized automated semantic scoring metrics.

Our contributions are:

1. An Augmented IUX dataset Demner-Fushman et al. (2015) with each of 3995 image-report instances augmented with a patient-centric explanation. Our augmentation is performed across all 105 disease classes of IUX dataset. We achieve this via a synthetic data generation pipeline, evaluated by medical experts (Section 4.1).

2. A multimodal prompt based VLM framework, **Rad-Flamingo**, for automated structured radiology report generation and patient-centric explanation. Our method improved quantitative and qualitative scores by 2.3% over existing methods (Section 4.2 and 6).

3. A first-of-its-kind multimodal in-context learning technique for self-rationalization by adding explicit medical knowledge to the prompt. To the best of our knowledge, this method incorporates multimodality and patient understandability for prompt based radiology report generation resulting in a 2.4% increment in performance over existing few-shot prompting techniques (Section 6.2).

2 Background and Definitions

Patient-Centric Explanations: Pathophysiology (McCance et al., 2019) is the study of the functional changes that occur in the body as a result of a disease or injury. It focuses on understanding the mechanisms by which diseases disrupt normal physiological processes. In heart failure, for instance, a reduction in cardiac output leads to compensatory mechanisms like fluid retention, which can cause symptoms like edema and shortness of breath. Therefore, such informations serve as a form of medical explanation with the generated report. We extend this idea to patient-centric explanations, where the pathophysiological explanations are provided along-with the medical reports for ease of understanding from the patients' perspective.

Self-Rationalization: Self-rationalization in large language models (LLMs) (Marasovic et al., 2022; Wiegreffe et al., 2021; Camburu et al., 2018) refers to their ability to generate explanations or justifications for their own outputs. This involves creating reasoning pathways that appear coherent, logical, and aligned with the responses they produce, even though these models do not possess true understanding or awareness. LLMs achieve this by leveraging their vast training data to mimic human reasoning patterns, constructing plausible rationales based on context, prior responses, and linguistic structures. However, these explanations do not serve as a pointer to the internal working of the model, they merely act as a justification to the output. In sensitive domains such as healthcare, an explanation, at the very least plays an important role towards building trust.

In-Context Learning: In-context learning refers to the ability of LLMs to perform tasks by understanding and extrapolating from examples provided within a prompt, without requiring explicit fine-tuning of the model. This technique leverages the model's parametric knowledge and allows users to define the task through natural language instructions and a few input-output examples (often called few-shot learning). The model infers the pattern from the context and applies it to new instances during the same interaction. In-context learning demonstrates the flexibility of LLMs to adapt to diverse tasks, making them highly versatile for applications like text generation, question answering, and code synthesis (Dong et al., 2024).

3 Related Work

Report Generation: Radiology report generation has been receiving a lot of attention lately, and several models have been developed based on the encoder-decoder architecture that was first used for image captioning tasks (Vinyals et al., 2014; Xu et al., 2015; Pan et al., 2020). However, report generation poses additional challenges compared to image captioning, as medical reports are typically longer and coherent with respect to captions. In an encoder decoder setting it becomes very difficult to generate long-form reports coherent with the medical image. Furthermore, bias in medical datasets makes it difficult to generate comprehensive, long-form reports. To address these challenges, researchers have proposed various methods. Wang et al. (2021), introduced an image-text matching branch to facilitate report generation, utilizing report features to augment image characteristics and consequently minimize the impact of data bias. They also employed a hierarchical LSTM structure for the generation of long-form text. Chen et al. (2020a) and Wang et al. (2022b) introduced additional memory modules to store past information, which can be utilized during the decoding process to improve long-text generation performance.

Another type of work aims to mitigate data bias by incorporating external knowledge information, with the most representative approach being the integration of knowledge graphs Li et al. (2019, 2023b); Huang et al. (2023); Liu et al. (2021); Zhang et al. (2020); Kale et al. (2023). Zhang et al. (2020) and Liu et al. (2021) combined preconstructed graphs representing relationships between diseases and organs using graph neural networks, enabling more effective feature learning for abnormalities. Li et al. (2023b) developed a dynamic approach that updates the graph with new knowledge in real-time. Huang et al. (2023) incorporated knowledge from a symptom graph into the decoding stage using an injected knowledge distiller.

These methods are able to generate reports as caption with very high accuracy. However, they do not have the ability of free-form text generation possesed by pretrained VLMs. Therefore, VLMs become very effective for free-form text generation.

Vision Language Models: A significant area of research in natural language processing (NLP) and computer vision is the exploration of vision language model (VLM) learning techniques. This VLM aims to bridge the gap between visual and textual information, enabling machines to understand and generate content that combines both modalities. Recent studies have demonstrated the potential of VLM models in various tasks, such as image captioning (Zhu et al., 2023), visual question answering (Liu et al., 2023b; Maaz et al., 2024), and image generation (Zhang et al., 2023). Developing on these medical VLMs like (Li et al., 2023a) and (Abdin et al., 2024) show impressive performance on medical NLP use cases.

4 Methodology

In the first stage, as per Figure 1, we use a finetuned MiniGPT4 model to synthetically generate patient-centric explanations (which are subsequently human evaluated) for each image report pair. The model is finetuned on MIMIC-CXR-JPG (Johnson et al., 2019) dataset, a large-scale repository of chest X-ray images and corresponding reports in the form of findings and impressions. Finetuning allows the model to re-parameterize its weights to learn to align a chest X-ray to its corresponding report. Given this finetuned model, we design a three-shot prompt template to generate patient-centric explanations for an X-ray image and its corresponding report. Therefore, this stage appends all the existing dataset samples with a patient-centric explanation. The explanations generated are evaluated by medical-experts resulting in creation of a gold-label dataset consisting of image-report-PCE. This created and human evaluated dataset then serves as a standard against which we compare the outcomes of the second stage.

In the second stage, we use this newly augmented dataset to perform in-context learning with a vision-language model that has been pretrained on a medical dataset. This approach allows the model to incorporate the nuances of patient-centric explanations while maintaining its ability to provide clinically accurate and detailed radiological reports.

4.1 Stage I (Synthetic Data Generation)

To fine-tune the MiniGPT4 (Zhu et al., 2023) model we follow the technique in Thawakar et al. (2024). We combine textual information from a medical large language model (LLM) and visual characteristics from a pre-trained medical vision encoder (VLM) given the X-ray. In particular, our large

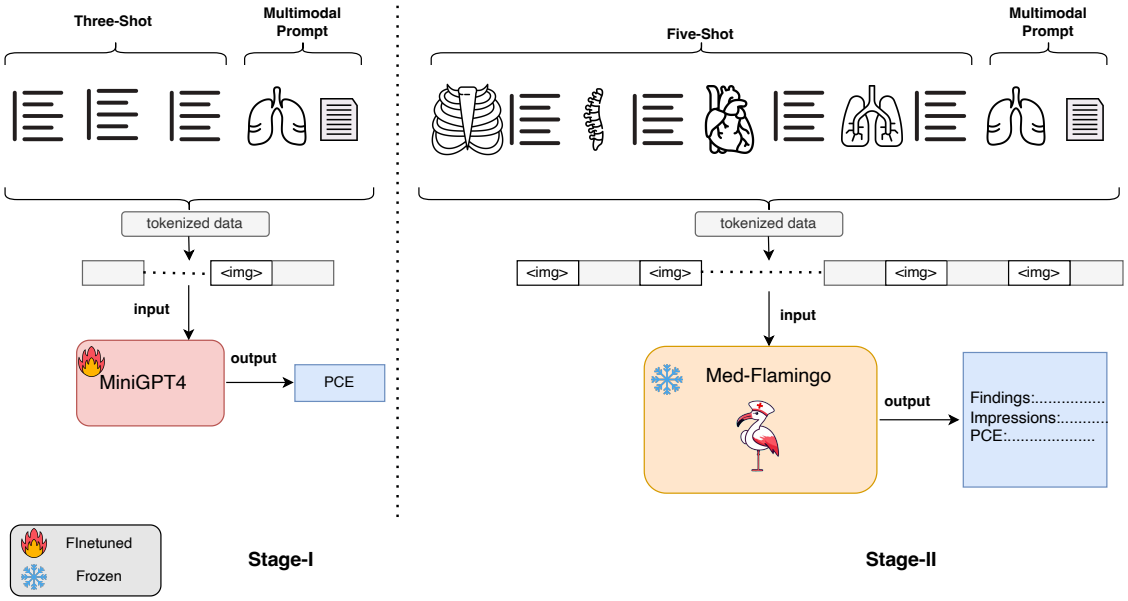


Figure 1: Stage I: Refers to the synthetic data generation stage, which annotate the existing IUX dataset with patient centric explanations. Stage II: Refers to the report generation stage where we design multimodal in-context prompts using the annotated data from stage I. Additionally, the fire symbol represents the finetuned model and ice symbol represent using frozen weights of a model not finetuned by us. PCE refers to the abbreviation of patient-centric explanation.

language model (LLM) is based on the recently developed Vicuna model (Zheng et al., 2024), and we use MedClip (Wang et al., 2022c) as a vision encoder.

Given an X-ray $x \in \mathbb{R}^{H \times W \times C}$, the vision encoder E_{img} encodes the image as $E_{\text{img}}(x)$. Then, the raw embeddings are transformed to an output dimension of 512 using a linear projection head.

$$V_p = f_v(E_{\text{img}}(x)) \quad (1)$$

where E_{img} is the vision encoder, f_v is the projection head. We use a trainable linear transformation layer to close the gap between the embedding space of the language decoder and image-level features , denoted as t. This layer transforms the image-level features, represented by V_p, into corresponding language-decoder embedding tokens, denoted as L_v:

$$L_v = t(V_p) \quad (2)$$

Following this we employ a few-shot prompting strategy to generate patient-centric explanations for a given image-report pair.

We follow a standard few-shot prompting strategy with three examples in the prompt. In the prompt we write Explanations as a placeholder for patient-centric explanation. The prompt template goes as follows:

Example 1:

```
Findings:......
Impressions:.......
Explanations:......
```

Example 2:

```
Findings:......
Impressions:.......
Explanations:......
```

Example 3:

```
Findings:......
Impressions:.......
Explanations:......
```

Your Turn:

```
Findings:......
Impressions:.......
<img>Looking at the Xray, findings
and impressions generate an explanation
```

For the synthetic data generation we consider the IUX (Demner-Fushman et al., 2015) dataset, the

generated explanations are appended to each instance of the IUX dataset. For designing the prompt we sample three image-report (findings and impressions) pairs from each of the disease classes. We take assistance of medical experts to append each of the samples with patient-centric explanations. Subsequently, we pass the prompt as per Stage I in Fig 1 for the fine-tuned model to learn in-context. Fine-tuning the model on a large corpus, such as MIMIC-CXR-JPG (Johnson et al., 2019), helps the model to condition on the context provided in the prompt. We provide the full prompt samples in Appendix A. Therefore, the model is able to generate good quality explanations tailoring to our requirement. (the details are in appendix D). An **Augmented Dataset** is now created which consists of Image, report (Findings and Impressions) and patient-centric explanation Fig. (2) that serves as a standard against which we compare the outcomes of the second stage.

4.2 Stage II (Radiology Report Generation)

In this stage we follow the Med-Flamingo model (Moor et al., 2023) which is finetuned on a medical dataset. Med-Flamingo is developed on the Open-Flamingo Awadalla et al. (2023) architecture which possesses the ability of few-shot learning from multimodal inputs. The language modeling in Med-Flamingo is represented in eq 3

$$p(y_\ell \mid x_{1:\ell-1}, y_{1:\ell-1}) = \prod_{\ell=1}^{L} p(y_\ell \mid y_{1:\ell-1}, x_{1:\ell-1}) \tag{3}$$

where y_ℓ refers to the ℓ_{th} language token, $y_{1:\ell-1}$ to the set of prior language tokens, and $x_{1:\ell-1}$ to the set of prior visual tokens. Here the language tokens contain the information of reports and PCEs and the image tokens contain the information of chest X-rays. While fine-tuning, the input is annotated in the form of interleaved image text data, which makes it effective for multimodal few-shot learning. We exploit this interleaved template to design our proposed prompt as per Stage II in Fig 1. The interleaved input prompt-design while fine-tuning enables the model to condition on the multi-modal context. We choose five examples for each disease class from the **Augmented Dataset** compiled in stage I. Pivoting on the idea of interleaved image text data prompt, we set up our framework for multimodal in-context learning for which the prompt template is demonstrated below:

Example 1:

```
<img>Findings:......
Impressions:.......
Explanations:......
```

Example 2:

```
<img>Findings:......
Impressions:.......
Explanations:......
```

Example 3:

```
<img>Findings:......
Impressions:.......
Explanations:......
```

Example 4:

```
<img>Findings:......
Impressions:.......
Explanations:......
```

Example 5:

```
<img>Findings:......
Impressions:.......
Explanations:......
```

Your Turn:

```
<img>Looking at the xray generate
findings and impressions and a explana-
tion
```

Prompt examples are provided in the Appendix B. Med-Flamingo with our proposed multimodal prompt template is referred to as **Rad-Flamingo**.

5 Experiments

5.1 Dataset

In **stage I** we consider the MIMIC-CXR-JPG (Johnson et al., 2019) dataset for fine-tuning. MIMIC-CXR-JPG dataset comprises 473,057 images and 206,563 reports from 63,478 patients. The official splits, i.e. 368,960 for training, 2,991 for validation, and 5,159 for testing are used for fine-tuning our model. Subsequent to this we follow our prompting technique (**Section 4.1**) to generate patient-centric explanations and append it to each instance of the

Metrics	Models						
	R2GEN (Chen et al., 2020b)	**R2GenCMN** (Chen et al., 2021)	**Joint-TraiNet** (Yang et al., 2023)	**M2KT** (Yang et al., 2022)	**Open-Flamingo** (Awadalla et al., 2023)	**XProNet** (Wang et al., 2022a)	**Rad-Flamingo**
BLEU-1	0.355	0.372	0.359	0.366	0.293	0.353	0.323
BLEU-2	0.223	0.233	0.226	0.213	0.195	0.221	0.232
BLEU-3	0.152	0.153	0.155	0.146	0.155	0.150	0.183
BLEU-4	0.103	0.105	0.102	0.104	0.071	0.105	0.081
METEOR	0.141	0.150	0.142	0.152	0.165	0.141	0.170
ROUGE	0.278	0.282	0.278	0.267	0.223	0.281	0.223

Table 1: Lexical similarity performance of Rad-Flamingo compared to baselines using classical metrics (BLEU, METEOR, ROUGE).

IUX dataset (Demner-Fushman et al., 2015).
In **stage II** we use the **Augmented dataset** from the previous step and design our prompts as per Fig 1. The dataset consists of 7,470 chest X-Ray images and 3,955 radiology reports. The number of patients are equal to the number of reports however, each patient corresponds to two xray images i.e. frontal and lateral. Therefore, number of images are twice the number of reports. We append a patient-centric explanation to each of 3955 radiology reports.

5.2 Experimental Setup

In **stage-1** training, the model is fine-tuned to gain alignment between X-ray image features and corresponding reports by training over a large set of image-report pairs. The result obtained from the injected projection layer is considered as a gentle cue for our medically tuned VLM model, guiding it to generate appropriate report based on the finding and impression that match the given X-ray images. For preprocessing we follow Thawakar et al. (2024) where we utilize high quality interactive report summaries of MIMIC-CXR-JPG. The train set contains 213,514 image report pairs for training. During training, the model is trained for 320k total training steps with a batch size of 16 using 3 NVIDIA A100 (80GB) GPUs.
In **stage-II** we utilize predetermined prompts as shown in the previous section (4.2).

For each X-ray image instance we take the corresponding finding, impression and patient centric explanation and put it in the following format:

<image> Findings Impression Explanation|endofchunk|.

Five of these aforementioned multimodal prompt were followed by the query prompt described below:

<image> + You are a helpful medical assistant. You are provided with images, findings, impressions and explanation.Looking at this image generate Findings, Impressions and Explanations.

6 Result and Analysis

Our evaluation emphasizes the performance of the Flamingo family of models (Moor et al., 2023) (Awadalla et al., 2023), as these models provide the essential few-shot learning capabilities needed for our prompt-based report generating framework. One possible comparison of **Rad-Flamingo** could be done with other vision-language models, such as Med-Phi (Abdin et al., 2024) and Med-LLaVA (Li et al., 2023a). However, these models do not have the ability to accept multimodal prompt and hence were deliberately excluded as baselines from our analysis. Our results analyse the effectiveness of our multimodal prompt in generating reports with patient-centric explanation. Tables 1 and 2 compare the scores over the generated report and patient-centric explanations. Additionally, we present zero shot experiments on open-source VLMs in Appendix D.4.2

6.1 Lexical Metrics

In this section, we evaluate the quality of generated reports by **Rad-Flamingo** and compare them against baselines using classical lexical similarity metrics such as BLEU (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007), and ROUGE (Lin, 2004) as shown in Table 1. These metrics provide a convenient means of measuring word overlap and syntactic similarity between generated and reference texts. **Rad-Flamingo** performs similar to the baselines on lexical similarity metrics. However, these metrics find less application in medical domain. This arises due to their inability to account for the deeper semantic relevance and contextual accuracy required in specialized content, such as medical data. For example, the sentences *"There is focal consolidation"* and *"There is no*

Metrics	Rad-Flamingo	Rad-Flamingo w/oI	Open-Flamingo	Open-Flamingo w/oI
BertScore	**0.875**	0.855	0.863	0.834
BioClinicalBertScore	**0.895**	0.879	0.885	0.854
RadGraphF1	**0.285**	0.273	0.279	0.269

Table 2: Performance comparison of Rad-Flamingo and Open-Flamingo models on clinical evaluation metrics using proposed multimodal few-shot prompting framework. The table includes ablation studies highlighting the impact of removing image modalities (w/oI) from the few-shot prompts. We do a metric wise significance testing in Appendix D.2

focal consolidation" are lexically very similar yet semantically very dissimilar. Therefore, semantic similarity plays a greater role in evaluating generated medical texts.

Our few-shot prompting technique show comparable performance in some of the lexical metrics. While these metrics offer a preliminary measure of performance, they do not fully reflect the real utility of generated medical texts. This analysis underscores the need for more domain-specific evaluation frameworks that can assess not only linguistic fluency and coherence but also the contextual alignment of generated texts in medical domain.

6.2 Semantic Metrics

We choose semantic metrics for clinical evaluation like BioClinicalBERTScore [2] (Lee et al., 2019), BERTScore (Zhang et al., 2019) and RadGraphF1 (Jain et al., 2021). In table 2 column Rad-Flamingo represents the setting where we prompt the Med-Flamingo model with proposed multimodal few-shot prompt. The Rad-Flamingo w/oI column reflects a configuration where images are excluded from the few-shot prompt examples, while all other components remain identical to Rad-Flamingo. A similar ablation strategy is applied to Open-Flamingo and Open-Flamingo w/oI for consistency.

Both the BERTScore and ClinicalBERTScore for Rad-Flamingo show a 1.4% increase compared to Open-Flamingo. This shows our proposed multimodal prompt template effectively generates report with better performance than existing models. Similar increase is found in case of RadGraphF1 scores. This result signifies the benefit of our proposed **multimodal prompt** template of Rad-Flamingo, over Open-Flamingo. To show the utility of multimodality in our prompt template, we remove the images from the examples and pass it to the Rad-Flamingo and Open-Flamingo models. Rad-Flamingo w/oI and Open-Flamingo w/oI represents those settings. We see the scores drop significantly by 2.4% percent indicating the utility of the multimodal prompt in integrating different data-modalities and helps the model to generate task-specific outputs. This approach effectively addresses challenges in both unimodal and multimodal data modes. So domain specific metrics are crucial to understand the utility of the multimodal prompt strategy developed by us. Therefore, we observe from the metrics that the semantic similarity scores help us analyze the performance better for task-specific output. Overall, the best performance is given by Rad-Flamingo as the underlying Med-Flamingo model is finetuned on medical data. However, comparing the scores with Open-Flamingo exhibits the effectiveness and utility of our proposed **multimodal prompt** framework. We provide further experiments on the role of patient-centric explanations as part of the prompt Appendix D.

6.3 Qualitative Evaluation

Owing to the subjective nature and the semantic complexity which medical data possesses, evaluation by medical expert becomes very important to have a rigorous examination of a proposed system. We consulted four expert-medical professionals to evaluate our generated reports and corresponding patient-centric explanations. We perform an extensive expert evaluation of the generated outputs. The evaluation criteria is divided into two criterions namely, Understandability and Medical Comprehensiveness. Whereas **Understandability** is Patient Centric, **Medical Comprehensiveness** measures the output based on its completeness from a medical experts perspective. Following this we created five levels of grading: 1 (very poor), 2 (poor),

[2] BioClinicalBERT is taken from huggingface. Underlying model is BioBERT trained on MIMIC III dataset. https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT

Models	Rad-Flamingo	
	Understandability	Medical Comprehensiveness
Cardiomegaly	3.44 ± 0.67	3.25 ± 0.43
Pulmonary Atelectasis	3.33 ± 1.36	3.4 ± 0.5
Nodules	3.21 ± 1.05	$3.01 \pm .70$
Opacity	2.06 ± 0.54	2.5 ± 0.54
Calcified Granuloma	3.75 ± 0.82	3.13 ± 0.41
Pulmonary Fibrosis	3.0 ± 0.63	2.8 ± 0.58
Consolidation	3.2 ± 0.39	3.1 ± 0.56
Pneumothorax	3.6 ± 0.8	3.63 ± 0.6
Granuloma	3.4 ± 0.95	3.1 ± 0.7
Bronchiestasis	3.25 ± 0.44	3.1 ± 0.46

Table 3: The table presents the mean and standard deviation of scores provided by four medical professionals for each of the chosen disease class, highlighting the effectiveness of the proposed prompting method after stage II.

3 (good), 4 (very good), 5 (excellent) for both criterion. Subsequently, for each disease class we get four scores and the table shows a mean and standard deviation over these four scores for each criterion.

The expert evaluation also shows our proposed prompting method, gives promising performance. An output sample of our system is given in Appendix B. In Table 3 we express the results of expert evaluation on the aforementioned two criterion: patient-centric understandability (Understandability) and expert-centric completeness (Medical Comprehensiveness). The expert-evaluated scores indicate that understandability is in the upper half of the range of possible scores. Further Medical Comprehensiveness is also above average which indicates that the explanations generated are correct. However, the medical expert professionals depth of knowledge might not be achieved giving further scope in the future to explore this line of work. Furthermore, the ability to generated such explanations demonstrates that our multimodal prompting strategy effectively generates explanations that are both understandable to patients and simplify complex medical terminology.

6.4 Ablation study on patient-centric explanation

We analyze the impact of removing patient-centric explanations (PCEs) from our multimodal few-shot prompting framework by providing only findings and impressions as few-shot examples as shown in Appendix D.4.1. In this setting, the model fails to generate patient-centric explanations, highlighting the necessity of explicitly incorporating PCEs. As per Stage II (Section 4.2), when PCEs are omitted from the prompt template, the prior language tokens do not contain any information about them, leading to the next predicted tokens also lacking PCEs. This demonstrates that without explicit patient-centric guidance, the model is unable to produce explanations tailored for patient understanding as shown in Figure 4 despite being specifically instructed to do so. This ablation study reinforces the importance of our synthetic annotation stage (Section 4.1), which systematically introduces PCEs into the prompting process. We observe that the presence of PCEs directly influences the generation of explanations that simplify medical terminology. The ablation study further confirms that patient-centricity in explanations does not emerge naturally from findings and impressions alone, necessitating an explicit prompting strategy. In summary, this study highlights the crucial role of PCEs in shaping the generated explanations, confirming that our synthetic annotation stage provides significant context. Thus, PCEs serve as a key component in our framework, reinforcing the effectiveness of our multimodal few-shot prompting strategy. Along with this we also present a detailed experiment on understanding the readability of our generated explanations as presented in Appendix D.3

7 Conclusion

Rad-Flamingo introduces a radiology report generation framework that integrates multimodal data with prompt-driven methodologies and patient-centric explanations, enhancing accuracy and understandability. By leveraging vision-language models (VLMs), it automates routine reporting tasks, allowing radiologists to focus on complex cases and save valuable time. A key feature is the patient-centric approach, ensuring that reports are both medically accurate and understandable to non-expert audiences. Additionally, by simplifying complex medical terms, Rad-Flamingo makes radiology reports more accessible, bridging the gap between clinical findings and patient understanding. The results highlight significant potential for improving workflow efficiency and diagnostic support in radiology. However, future work should focus on refining the alignment between vision and language components in VLMs to generate more coherent, reports with improved explanations.

Limitations

In this section we discuss the main limitations of our proposed framework. A notable limitation in our study is the absence of a number of VLMs which possess the same few-shot learning capability as the Flamingo family of models. This restricts us from evaluating the generalizability of our approach. While our method shows promise, validating its performance against a diverse set of few-shot models would provide deeper insights into its strengths and weaknesses. The inclusion of these models would also allow us to better understand how our approach fares in broader scenarios and under varying conditions, such as domain shifts or noisy inputs.

Class imbalance in machine learning occurs when certain classes dominate the training data, causing the model to be biased toward these over-represented classes and perform poorly on minority classes. This is particularly problematic in applications like medical diagnosis, where minority classes are crucial, and can be addressed using techniques like re-sampling, loss adjustment, or robust algorithms.

Another constraint in our evaluation is the lack of a direct comparison with ChatGPT, a widely recognized benchmark in conversational AI. The prompt template we use would require high computational and financial cost to perform a rigorous analysis. These constraints underscore the need for collaborative efforts and accessible research resources to enable comprehensive benchmarking.

Ethical Considerations

The Rad-Flamingo framework enables multimodal, prompt-driven radiology report generation with patient-centric explanations, adhering to strict ethical standards. All medical data is anonymized, and our data augmentation process ensures no risk of identity leakage. Designed to support, not replace, clinicians, it enhances diagnostic accuracy and promotes transparent patient-provider communication. We mitigate bias through diverse training data representing various demographics and medical conditions. Patient explanations are clear, respectful, and free from misleading content. Human oversight ensures outputs align with clinical standards and ethical guidelines, maintaining patient safety, data security, and fairness in medical AI applications.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Richa Arora. 2014. The training and practice of radiology in india: current trends. *Quant. Imaging Med. Surg.*, 4(6):449–450.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *Preprint*, arXiv:2308.01390.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 2021. Cross-modal memory networks for radiology report generation. In *Proceedings of the 59th Annual*

Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 5904–5914, Online. Association for Computational Linguistics.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020a. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1439–1449, Online. Association for Computational Linguistics.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020b. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1439–1449, Online. Association for Computational Linguistics.

Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza M. Rodriguez, Sameer Kiran Antani, George R. Thoma, and Clement J. McDonald. 2015. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association : JAMIA*, 23 2:304–10.

Etienne Denoual and Y. Lepage. 2004. Bleu in characters: Towards automatic mt evaluation in languages without word delimiters. In *International Joint Conference on Natural Language Processing*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi

Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

MD. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2019. A comprehensive survey of deep learning for image captioning. *ACM Comput. Surv.*, 51(6).

Zhongzhen Huang, Xiaofan Zhang, and Shaoting Zhang. 2023. Kiut: Knowledge-injected u-transformer for radiology report generation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19809–19818.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Truong, D. Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curt P. Langlotz, and Pranav Rajpurkar. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. *ArXiv*, abs/2106.14463.

Alistair E. W. 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. *Preprint*, arXiv:1901.07042.

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

Alon Lavie and Abhaya Agarwal. 2007. Meteor: an automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, page 228–231, USA. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Christy Y. Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. 2019. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Preprint*, arXiv:2306.00890.

Mingjie Li, Bingqian Lin, Zicong Chen, Haokun Lin, Xiaodan Liang, and Xiaojun Chang. 2023b. Dynamic graph enhanced contrastive learning for chest x-ray report generation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3334–3343.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. 2021. Exploring and distilling posterior and prior knowledge for radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13753–13762.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *ArXiv*, abs/2304.08485.

Yunyi Liu, Zhanyu Wang, Dong Xu, and Luping Zhou. 2023b. Q2atransformer: Improving medical vqa via an answer querying decoder. *Preprint*, arXiv:2304.01611.

Siyu Lu, Yueming Ding, Mingzhe Liu, Zhengtong Yin, Lirong Yin, and Wenfeng Zheng. 2023. Multiscale feature extraction and fusion of image and text in vqa. *International Journal of Computational Intelligence Systems*, 16(1):54.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. 2024. Video-ChatGPT: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12585–12602, Bangkok, Thailand. Association for Computational Linguistics.

Ana Marasovic, Iz Beltagy, Doug Downey, and Matthew Peters. 2022. Few-shot self-rationalization with natural language prompts. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 410–424, Seattle, United States. Association for Computational Linguistics.

K.L. McCance, S.E. Huether, V.L. Brashers, and N.S. Rote. 2019. *Pathophysiology: The Biologic Basis for Disease in Adults and Children*. Elsevier.

Pablo Messina, Pablo Pino, Denis Parra, Alvaro Soto, Cecilia Besa, Sergio Uribe, Marcelo Andía, Cristian Tejos, Claudia Prieto, and Daniel Capurro. 2022. A survey on deep learning and explainability for automatic report generation from medical images. *ACM Comput. Surv.*, 54(10s).

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. 2023. Med-flamingo: a multimodal medical few-shot learner. In *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pages 353–367. PMLR.

Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. 2020. X-linear attention networks for image captioning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10968–10977.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Amanda Ross and Victor L. Willson. 2017. *One-Way Anova*, pages 21–24. SensePublishers, Rotterdam.

A. Jackson Stenner. 2023. *Measuring Reading Comprehension with the Lexile Framework*, pages 63–88. Springer Nature Singapore, Singapore.

Omkar Chakradhar Thawakar, Abdelrahman M. Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Khan. 2024. XrayGPT: Chest radiographs summarization using large medical vision-language models. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 440–448, Bangkok, Thailand. Association for Computational Linguistics.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and D. Erhan. 2014. Show and tell: A neural image caption generator. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164.

Jun Wang, Abhir Bhalerao, and Yulan He. 2022a. Cross-modal prototype driven network for radiology report generation. In *Computer Vision – ECCV 2022*, pages 563–579, Cham. Springer Nature Switzerland.

Zhanyu Wang, Mingkang Tang, Lei Wang, Xiu Li, and Luping Zhou. 2022b. A medical semantic-assisted transformer fornbsp;radiographic report generation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part III*, page 655–664, Berlin, Heidelberg. Springer-Verlag.

Zhanyu Wang, Luping Zhou, Lei Wang, and Xiu Li. 2021. A self-boosting framework for automated radiographic report generation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2433–2442.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022c. MedCLIP: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3876–3887, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sarah Wiegreffe, Ana Marasović, and Noah A. Smith. 2021. Measuring association between labels and free-text rationales. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10266–10284, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057, Lille, France. PMLR.

Shuxin Yang, Xian Wu, Shen Ge, S. Kevin Zhou, and Li Xiao. 2022. Radiology report generation with a learned knowledge base and multi-modal alignment. *Preprint*, arXiv:2112.15011.

Yan Yang, Jun Yu, Jian Zhang, Weidong Han, Hanliang Jiang, and Qing-An Huang. 2023. Joint embedding of deep visual and semantic features for medical image report generation. *IEEE Transactions on Multimedia*, 25:167–178.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2018. From recognition to cognition: Visual commonsense reasoning. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6713–6724.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3813–3824.

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

Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Loddon Yuille, and Daguang Xu. 2020. When radiology report generation meets knowledge graph. In *AAAI Conference on Artificial Intelligence*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *Preprint*, arXiv:2304.10592.

A Appendix

A.1 Prompt sample from Stage I

Example 1: Findings:Borderline cardiomegaly. Midline sternotomy. Enlarged pulmonary arteries. Clear lungs. Inferior.
Impression: No acute pulmonary findings.
Explanation: The borderline cardiomegaly suggests slight enlargement of the heart, which could be due to chronic pressure overload (possibly from hypertension) or cardiomyopathy. The midline sternotomy indicates a previous cardiac surgery, which could relate to coronary artery disease or valve replacements, considering the context of other findings. Enlarged pulmonary arteries might suggest pulmonary hypertension, which could be secondary to left heart disease or chronic lung conditions, though the lungs are reported as clear. The absence of acute pulmonary findings despite these chronic alterations indicates that there are no new or worsening issues in the pulmonary system at this time.

Example 2: Findings: The cardiac contours are normal. basilar atelectasis. The lungs are clear. Thoracic spondylosis. Lower cervical arthritis.
Impression: Basilar atelectasis. No confluent lobar consolidation or pleural effusion.
Explanation: The chest X-ray shows basilar atelectasis, a partial collapse or incomplete expansion of the lower lung segments, likely due to shallow breathing or reduced lung volume. The lungs are otherwise clear, without any signs of confluent lobar consolidation or pleural effusion, suggesting no infection or fluid accumulation. The cardiac contours are normal, ruling out cardiomegaly or abnormal heart positioning. Degenerative changes, including thoracic spondylosis and lower cervical arthritis, are noted, indicating age-related or chronic wear-and-tear changes in the spine. The overall impression is of minor pulmonary findings with no acute cardiopulmonary disease.

Example3: Findings: The cardiomediastinal silhouette is within normal limits for size and contour. The lungs are normally inflated without evidence of focal airspace disease, pleural effusion, or pneumothorax. Stable calcified granuloma within the right upper lung. No acute bone abnormality.
Impression: No acute cardiopulmonary process.
Explanation: The chest X-ray reveals a normal cardiomediastinal silhouette, indicating no enlargement or abnormality in heart or mediastinal contours. The lungs are well inflated, with no signs of focal airspace disease such as infection, fluid, or collapse. A stable, calcified granuloma in the right upper lung suggests a past, healed infection (e.g., tuberculosis or histoplasmosis) with no current active disease. The absence of pleural effusion, pneumothorax, or acute bone abnormalities supports the lack of any acute thoracic issues. Overall, there is no evidence of an acute cardiopulmonary process, and the findings are consistent with a stable, chronic condition.

Your Turn: Findings: Heart size within normal limits. No focal alveolar consolidation, no definite pleural effusion seen. No typical findings of pulmonary edema. No pneumothorax.
Impression: No acute cardiopulmonary findings.
<img>Looking at the Xray, findings and impressions generate a patient- centric explanation

B Appendix

B.1 Prompt sample from Stage II

The images in figure 1 are placeholders to exhibit the multimodal prompt design. While experiments we took examples per disease class mentioned in the results section.

Example 1: <image> Findings: Heart size mildly enlarged, stable mediastinal and hilar contours. No focal alveolar consolidation, no definite pleural effusion seen. No typical findings of pulmonary edema. No pneumothorax.
Impressions: Mild cardiomegaly, no acute pulmonary findings.
Explanation: It is clear the patient is in an enlarged heart condition (mild cardiomegaly) and has been for some time. The pulmonary system is clear and no evidence of acute lung changes. The overall pathophysiology suggests that the patient is in a chronic, stable state with ongoing changes.The findings are normal and do not suggest any acute or severe events.<|endofchunk|>

Example 2: <image> Findings: The heart is top normal in size. The mediastinum is stable. The lungs are clear.
Impressions: Borderline cardiomegaly without acute disease.
Explanation: It is a case of borderline cardiomegaly, but it is difficult to correlate the finding of enlarged heart with pathophysiology. Borderline cardiomegaly is a radiological finding. It could mean that the heart size is in between normal and big (normal is 55 mm) but in this case it comes to 57 mm.<|endofchunk|>

Example 3: <image> Findings: Stable borderline cardiomegaly, stable mediastinal and hilar contours. No alveolar consolidation, no findings of pleural effusion or pulmonary edema. No pneumothorax.
Impressions: No acute cardiopulmonary findings.
Explanation: There are no acute cardiopulmonary findings, no pulmonary edema, no focal alveolar consolidation, no definite pleural effusion, no large pleural effusions, no pneumothorax, no left apical or basal consolidation, no pulmonary vascular congestion, and no pulmonary infarction; however, bilateral patchy pulmonary opacities and multifocal scattered bibasilar patchy opacities are noted.<|endofchunk|>

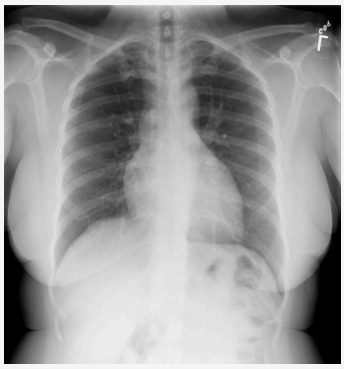
Example 4: <image> Findings: Persistent cardiomegaly. No abnormal airspace consolidation. Resolved interstitial edema. No pneumothorax or pleural effusion.
Impressions: Stable cardiomegaly without acute abnormality.
Explanation: No pneumothorax, no effusion, no infiltrate, no pulmonary congestion, no pleural erythema all point towards non-infectious etiology. Mild cardiomegaly without acute abnormality is also non-specific and without clinical significance. The unremarkable spine suggests degenerative changes and nothing else.<|endofchunk|>

Example 5: <image> Findings: The outside is normal except for slight cardiomegaly.
Impressions: Heart size upper limits normal. Lungs are clear. No evidence of active tuberculosis. No change from prior exam.
Explanation: Slight cardiomegaly. Clear lungs indicate no pulmonary congestion or active disease.<|endofchunk|>

Your Turn: <image> You are a helpful medical assistant. You are provided with images, findings, impressions and explanation.Looking at this image generate Findings, Impressions and Explanations

C Appendix

C.1 Augmented IUX dataset instance



Findings: Heart size mildly enlarged, stable mediastinal and hilar contours. No focal alveolar consolidation, no definite pleural effusion seen. No typical findings of pulmonary edema. No pneumothorax.

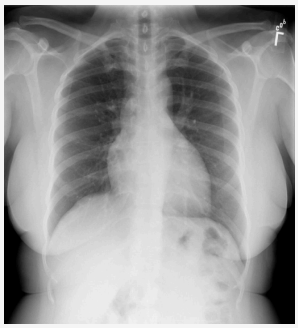
Impressions: Mild cardiomegaly, no acute pulmonary findings

Explanation: It is clear the patient is in an enlarged heart condition (mild cardiomegaly) and has been for some time. The pulmonary system is clear and no evidence of acute lung changes. The overall pathophysiology suggests that the patient is in a chronic, stable state with ongoing changes.The findings are normal and do not suggest any acute or severe events.

Figure 2: Augmented dataset instance showcasing input modalities (e.g., medical images, clinical text) and corresponding annotated outputs, illustrating the report (findings and impression) and patient-centric explanation

C.2 Radiology Report with patient-centric explanation generated by Rad-Flamingo

Image



Generated

Findings: Stable cardiomegaly. No abnormal airspace consolidation. Resolved interstitial edema. No pneumothorax or pleural effusion.

Impressions: Stable cardiomegaly without acute abnormality.

Explanation: No pneumothorax, no effusion, no infiltrate, no pulmonary congestion, no pleural erythema all point towards non-infectious etiology. Mild cardiomegaly without acute abnormality is also non-specific and without clinical significance. The unremarkable spine suggests degenerative

Ground Truth

Findings: Heart size mildly enlarged, stable mediastinal and hilar contours. No focal alveolar consolidation, no definite pleural effusion seen. No typical findings of pulmonary edema. No pneumothorax.

Impressions: Mild cardiomegaly, no acute pulmonary findings

Explanation: It is clear the patient is in an enlarged heart condition (mild cardiomegaly) and has been for some time. The pulmonary system is clear and no evidence of acute lung changes. The overall pathophysiology suggests that the patient is in a chronic, stable state with ongoing changes.The findings are normal and do not suggest any acute or severe events.

Figure 3: Example of output given by Rad-Flamingo. Image and ground truth are from the proposed augmented dataset.

D Appendix

D.1 Medical Expert Evaluation for Stage I outputs

Models	Finetuned MiniGPT-4	
	Understandability	Medical Comprehensiveness
Cardiomegaly	3.56 ± 0.76	3.43 ± 0.52
Pulmonary Atelectasis	3.31 ± 1.26	3.41 ± 0.51
Nodules	3.22 ± 1.46	$3.09 \pm .71$
Opacity	2.07 ± 0.57	2.5 ± 0.54
Calcified Granuloma	3.78 ± 0.82	3.23 ± 0.41
Pulmonary Fibrosis	3.0 ± 0.68	2.7 ± 0.78
Consolidation	3.22 ± 0.69	3.1 ± 0.66
Pneumothorax	3.61 ± 0.81	3.63 ± 0.67
Granuloma	3.44 ± 0.85	3.12 ± 0.71
Bronchiestasis	3.25 ± 0.54	3.11 ± 0.56

Table 4: The table presents the mean and standard deviation of scores provided by four medical professionals for each of the chosen disease class. Highlighting the effectiveness of the proposed finetuning+prompting method in stage I for synthetic annotation with patient-centric explanations. Follows the same trend as Table 3

D.2 Significance testing for Semantic Metrics

Metrics	F-statistic	p-value
BioClinicalBertScore	30.00	0.0001
BertScore	30.01	0.0001
RadGraphF1	30.00	0.0001

Table 5: Statistical significance analysis using one-way ANOVA for BERTScore, BioClinicalBERTScore, and RadGraphF1 scores across four evaluation settings: Rad-Flamingo, Rad-Flamingo w/oI, Open-Flamingo, and Open-Flamingo w/oI. The results indicate significant differences in scores, as determined by F-statistics and p-values ($p < 0.05$).

Extending our analysis in the results section, we further provide significance testing for the BERTScore, BioClinicalBERTScore, and RadGraphF1 scores of Rad-Flamingo, Rad-Flamingo w/oI, Open-Flamingo, and Open-Flamingo w/oI.

Null Hypothesis (H_0): There is no significant difference between the <score-name>. **Alternative Hypothesis** (H_1): There is significant difference between the <score-name>. As each of the output from the models are mean of generated reports over the chosen disease classes, we take them as the group mean for the one-way ANOVA test (Ross and Willson, 2017). Therefore, we consider the four evaluation setting as four groups of data, We get $F\text{-}statistic = 30.00$ and $p\text{-}value \approx 0.0001$ respectively. Consequently, $F\text{-}statistic > F_{critical}$ and $p\text{-}value < 0.05$, satisfying these conditions we can reject the Null Hypothesis thereby establishing the values are significantly different. Similarly, we get $F\text{-}statistic = 30.01$ and $p\text{-}value \approx 0.0001$ respectively. As the BioClinicalBERTScores are similar to the BERTScore we get similar $F\text{-}statistic$ and $p\text{-}value$. Consequently, $F\text{-}statistic > F_{critical}$ and

$p\text{-}value < 0.05$, satisfying these conditions we can reject the Null Hypothesis thereby establishing the values are significantly different. Lastly, we get $F\text{-}statistic = 30.00$ and $p\text{-}value \approx 0.0001$ respectively. Consequently, $F\text{-}statistic > F_{critical}$ and $p\text{-}value < 0.05$, satisfying these conditions we can reject the Null Hypothesis thereby establishing the values are significantly different.

D.3 Readability measure

We perform an additional evaluation to increase experimental validity of our proposed multimodal few-shot prompting strategy. To evaluate the human understandability of the generated explanations we evaluate them with reading measure technique like Lexile Reading Measure (Stenner, 2023). A Lexile measure is a standardized score that assesses both the reading ability of individuals and the complexity of written texts, represented on a scale typically ranging from below 200L to above 1600L. This measure helps educators, parents, and students identify reading materials that align with a reader's current ability level, ensuring an appropriate level of challenge to support comprehension and skill development. We also evaluate on CharBLEU metric (Denoual and Lepage, 2004) since in medical text spelling plays a crucial role.

Models	Rad-Flamingo	
	Generated	**Ground Truth**
Lexile Measure	69.28	63.6
CharBLEU	0.298	0.283

Table 6: The table highlights the readability and spelling accuracy of the generated explanations, demonstrating their alignment with patient comprehension needs and medical domain standards.

Table 6 represents two columns where the ground truth corresponds to the synthetically annotated instances in stage-I and generated corresponds to the output explanations generated by our proposed prompting technique in stage-II. The scores show a 8.9% increase in the readability of the generated explanations. The score provided is an average over all the ten selected diseases as per Table 3. Averaging across all values indicates an overall increase in readability; however, for certain disease classes, no improvement is observed. The readability scores confirm that the generated explanations become more comprehensible. Notably, explanations from Stage II exhibit enhanced readability compared to those from Stage I, demonstrating the effectiveness of our proposed prompt design in improving clarity.

D.4 Further Experiments

D.4.1 Ablation study on patient-centric explanation as part of the prompt template

We ablate our prompt in the following manner:

Example 1: <image> Findings: Heart size mildly enlarged, stable mediastinal and hilar contours. No focal alveolar consolidation, no definite pleural effusion seen. No typical findings of pulmonary edema. No pneumothorax.
Impressions: Mild cardiomegaly, no acute pulmonary findings. <|endofchunk|>

Example 2: <image> Findings: The heart is top normal in size. The mediastinum is stable. The lungs are clear.
Impressions: Borderline cardiomegaly without acute disease.<|endofchunk|>

Example 3: <image> Findings: Stable borderline cardiomegaly, stable mediastinal and hilar contours. No alveolar consolidation, no findings of pleural effusion or pulmonary edema. No pneumothorax.
Impressions: No acute cardiopulmonary findings.<|endofchunk|>

Example 4: <image> Findings: Persistent cardiomegaly. No abnormal airspace consolidation. Resolved interstitial edema. No pneumothorax or pleural effusion.
Impressions: Stable cardiomegaly without acute abnormality.<|endofchunk|>

Example 5: <image> Findings: The outside is normal except for slight cardiomegaly.
Impressions: Heart size upper limits normal. Lungs are clear. No evidence of active tuberculosis. No change from prior exam.<|endofchunk|>

Your Turn: <image> You are a helpful medical assistant. You are provided with images, findings, impressions and explanation.Looking at this image generate Findings, Impressions and Explanations

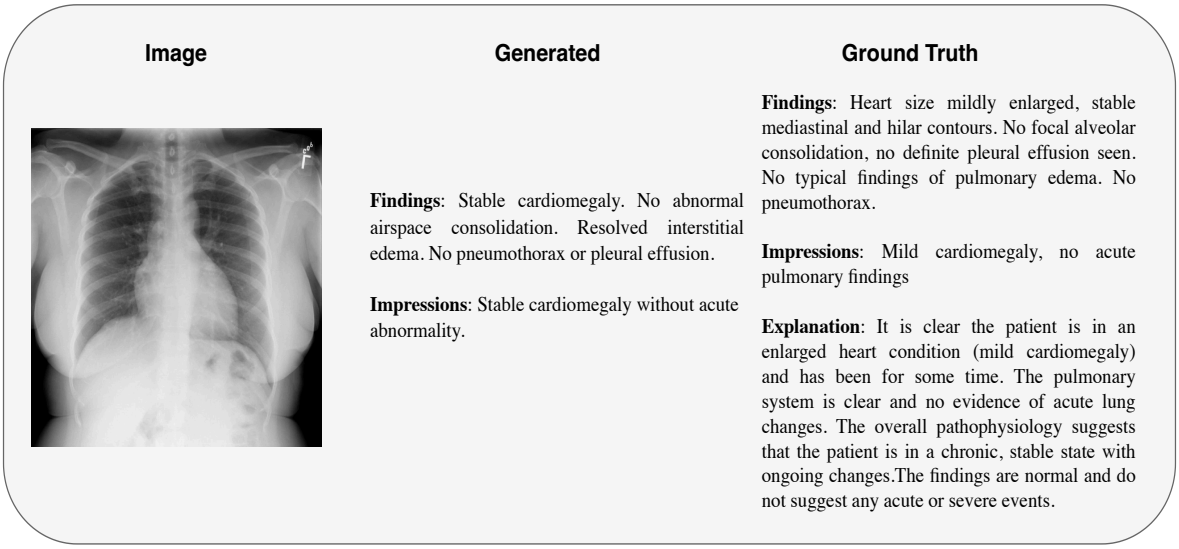


Figure 4: Example of output given by Rad-Flamingo after ablating patient-centric explanation

D.4.2 Zero-Shot experiments on open-source VLMs

Table 7 presents the zero-shot evaluation of two open-source vision-language models (VLMs), Llava Liu et al. (2023a) and Llama3.2-vision (Grattafiori et al., 2024). Unlike the Flamingo family, which supports few-shot learning, these models lack such capabilities, necessitating zero-shot experiments where images and instructions are provided to generate findings, impressions, and patient-centric explanations. The results show a significant performance decline, highlighting their limitations in medical report generation without few-shot adaptation. This reinforces the effectiveness of our multimodal few-shot prompting strategy in improving diagnostic accuracy, interpretability, and bias reduction. Additionally, the results validate the importance of our two-stage framework, which first generates findings and impressions before integrating patient-centric explanations, ensuring more structured and reliable outputs. These findings emphasize the necessity of few-shot prompting in AI-driven diagnostic radiology and demonstrate the advantages of a structured generation pipeline for maintaining accuracy and contextual relevance in

medical imaging applications.

Metrics	Llava (Zero-Shot)	Llama 3.2-Vision (Zero-Shot)
BertScore	0.70	0.55
BioClinicalBertScore	0.81	0.57
RadGraphF1	0.225	0.172

Table 7: Table 7:Zero-shot evaluation results for open-source vision-language models (VLMs), Llava Liu et al. (2023a) and Llama3.2-vision (Grattafiori et al., 2024). The significant performance drop highlights the limitations of these models in generating high-quality medical reports without few-shot adaptation, reinforcing the effectiveness of our multimodal few-shot prompting strategy and the necessity of a two-stage framework for structured report generation.