MediVLM: A Vision Language Model for Radiology Report Generation from Medical Images

Debanjan Goswami, Ronast Subedi and Shayok Chakraborty
Department of Computer Science
Florida State University
dgoswami@fsu.edu , rs22ce@fsu.edu, schakraborty2@fsu.edu

Abstract

Generating radiology reports from medical images has garnered sufficient attention in the research community. While existing methods have demonstrated promise, they often tend to generate reports that are factually incomplete and inconsistent, fail to focus on informative regions within an image, and impose strong annotation assumptions, such as bounding box annotations, image level annotations (which can be challenging to obtain) for model training. In this paper, we propose MediVLM, a vision language model (VLM) for radiology report generation from medical images. The proposed model consists of a pre-trained object detector to extract the salient anatomical regions from the images, an image encoder, a text encoder, a module to align the visual and text representations, a cross attention layer to fuse the two representations and finally, a transformer based decoder to generate the final report. MediVLM can generate radiology reports even when no reports are available for training; this is an extremely useful feature, as curating such reports is a labor-intensive task. Further, it computes a severity score (depicting the seriousness of a patient's medical condition) from the generated radiology reports, which can be used to prioritize patients who need immediate medical attention. Our extensive empirical analyses on three benchmark datasets corroborate the promise and potential of our method against competing baselines. Our code is open-sourced in our project webpage at: https://sites.google.com/view/medivlm/home

1 Introduction

Clinical radiology (such as X-rays, MRI scans etc.) is a common type of medical imaging examination and is critical for identifying common diseases such as pneumonia and lung cancer (Johnson et al., 2019; Raoof et al., 2012). Given a radiograph, radiologists manually examine the different anatomical regions and describe both the normal and abnormal findings in a textual report (Goergen et al., 2013). This is a time-consuming and tedious process, given the large volume of radiology images that need to be examined in daily clinical practice; the shortage of trained radiologists in many healthcare systems further aggravates the problem (Rimmer, 2017; Rosenkrantz et al., 2016). To address these challenges, radiology report generation (automatically generating a free-text description for a clinical radiograph) has attracted significant research attention in recent years (Wang et al., 2024). It has the potential to expedite the automation of workflows, alleviate the manual labor of radiologists, and improve the overall quality of healthcare.

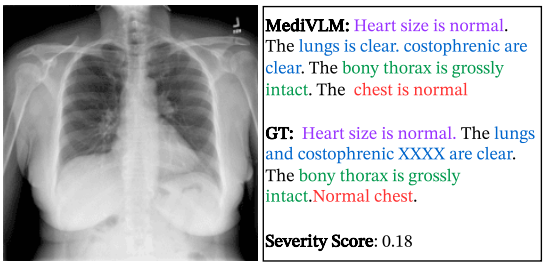


Figure 1: For a given medical image, MediVLM generates a free text radiology report, together with a severity score denoting the seriousness of the patient's medical condition. Best viewed in color.

While existing techniques for automated report generation have depicted encouraging performance, they often tend to generate reports that are incomplete (miss important observations in the images) or inconsistent (contain factually incorrect information) (Miura et al., 2021). Another drawback of current methods is that most of them utilize image-level visual features to generate reports and as a result, fail to focus on specific regions within an image that contain anatomical abnormalities (Xu et al., 2021; Chen et al., 2020; Yeasin et al., 2024). Further, some methods require specialized annotations for report generation. The ICON method (Hou et al., 2024) is based on lesion extraction and

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 10287–10304

requires image-level annotations for model training, which can be challenging to obtain in certain medical settings. The Region-Guided Radiology Report Generation (RGRG) method requires fine-grained annotations, such as bounding boxes, for radiology report generation (Tanida et al., 2023), which are expensive to obtain given the dearth of trained radiologists.

In this paper, we propose *MediVLM*, a vision language model for radiology report generation from medical images to alleviate these challenges. Our proposed method first extracts salient anatomical regions from the given input images using a pre-trained object detector. A visual encoder is then used to extract spatial features from these patches; at the same time, a language model is used as a tokenizer to embed the input radiology reports into latent representations. The image and text embeddings are aligned using contrastive learning and fused together with a trainable cross attention module. Finally, a transformer based language model (decoder), with trainable attention layers, is used to generate the medical report. MediVLM is trained end-to-end using a loss function consisting of a cross entropy loss term and a contrastive loss term.

The proposed MediVLM framework can be trained in an unsupervised manner (when no reports are available for training). We exploit a pseudo-labeling strategy to address the absence of ground truth reports and train the MediVLM using the generated reports, without requiring any architectural modifications. This is an extremely useful feature, as curating radiology reports is an expensive process in terms of time, labor and human expertise. Further, experienced radiologists are busy, which further underscores the challenge of obtaining high quality radiology reports for model training. Our model also generates a severity score S for each generated report, which depicts the criticality of the patient's medical condition. This score can be taken into consideration while scheduling doctors' appointments and can potentially mitigate the long waiting times for patients with serious medical conditions. Figure 1 depicts a sample output of our system for a given medical image.

Our contributions in this paper can be summarized as follows:

- We propose MediVLM, a VLM architecture to generate radiology reports from medical images, that consistently depicts impressive performance against competing baselines.
- While existing methods require images together with corresponding reports, our method can be trained in an unsupervised manner (when no reports are available for training). This is an extremely useful feature as obtaining reports is a labor-intensive, tedious task and experienced radiologists are rare and busy.
- Our MediVLM is also equipped with a capability to compute a severity score corresponding to each generated report, that quantifies the seriousness of the patient's medical condition, which can be used to prioritize patients who need immediate medical attention.
- We conduct extensive empirical studies on three benchmark datasets; our results demonstrate the efficacy of our method over competing baselines.

2 Related Work

Automatic radiology report generation has attracted significant research attention in recent years; please refer to (Wang et al., 2024) for a detailed survey.

Conventional methods: Earlier efforts adopted CNN-RNN architectures to analyze medical images and generate diagnostic reports (Jing et al., 2019, 2018; Li et al., 2018). More recent studies have used the transformer model due to its effectiveness (Vaswani et al., 2017). Chen *et al.* (Chen et al., 2020) proposed a memory-driven transformer (R2Gen) to generate radiology reports, which introduces a memory module and a memory-driven conditional layer normalization module into the transformer decoder architecture. The design of the memory module and layer normalization inspired several subsequent research. Chen *et al.* (Chen et al., 2021) proposed cross-modal memory networks (CMN) to enhance the encoder-decoder framework for radiology report generation, where a shared memory was designed to record the alignment between images and texts so as to facilitate the interaction and generation across modalities. A few strategies have been exploited to better attend to abnormal regions in an image, such as iteratively aligning visual features and disease tags (You et al., 2021), contrasting normal and abnormal images (Liu et al., 2021b) and exploiting medical knowledge graphs (Liu et al., 2021a). A radiologist-minded report generation framework, X-RGen, was proposed by Chen *et al.* (Chen et al., 2024), which mimics the behavior of human radiologists by breaking the entire process

down into four principal phases: initial observation, cross-region analysis, medical interpretation and report generation. Yeasin *et al.* (Yeasin et al., 2024) developed a transformer-based system called Auto-Rad for Lumbar Spinal Stenosis diagnostics and reporting. Hierarchical approaches have also been studied (Johnson et al., 2016; Liu et al., 2019; Nooralahzadeh et al., 2021), in which high-level concepts are first extracted from the image and subsequently decoded into individual sentences. Wang *et al.* (Wang et al., 2022) introduced a multi-head transformer that was applied to patch features from a CNN backbone, with each head assigned to a specific anatomical region, generating sentences exclusively for that region. Along similar lines, Tanida *et al.* (Tanida et al., 2023) proposed an explainable framework called Region-Guided Radiology Report Generation (RGRG) method that detects anatomical regions and generates individual descriptions for each.

Other methods: Reinforcement learning (RL) based methods have also been exploited for radiology report generation and improving clinical accuracy (Nishino et al., 2022; Delbrouck et al., 2022). Qin and Song (Qin and Song, 2022) proposed an RL approach over a cross-modal memory (CMM) to better align visual and textual features for radiology report generation. Multiple instance learning has also been used for histopathology report generation by aligning whole slide images and diagnostic reports from local and global granularity (Guo et al., 2024). Data augmentation techniques like mixup have been used in this context to ensure that the representations of the semantically equivalent lesions align with the same attributes, so as to maintain inter-report consistency (Hou et al., 2024). A line of research has focused on exploiting the temporal structure, i.e. prior images and reports (if available) for report generation (Bannur et al., 2023; Hou et al., 2023). Researchers have also studied the problem of report generation in the unpaired setting (where paired image-report data is unavailable for training) by leveraging information in two distinct datasets, one containing reports and the other containing images (Hirsch et al., 2024b,a).

While these methods have depicted encouraging performance, they often tend to generate reports that are inaccurate and incomplete (Miura et al., 2021), fail to focus on the informative anatomical regions in the images (Chen et al., 2020; Yeasin et al., 2024; Xu et al., 2021) and impose strong annotation constraints, such as image-level or bounding box annotations (Tanida et al., 2023; Hou et al., 2024).

Image Captioning: Radiology report generation is largely inspired by research in image captioning (Vinyals et al., 2015; Xu et al., 2015; You et al., 2016). However, while the key ideas from the image captioning domain can be applied to radiology report generation, there are a few important differences: (i) radiology reports are much longer and more diverse than typical image captions, due to the multiple anatomical regions within an image; (ii) generating descriptions of specific but crucial abnormalities is challenging due to the heavy data imbalance towards normal images and normal reports.

3 Method

3.1 Overview

Figure 2 depicts an overview of the MediVLM architecture. It consists of five modules, which are described in detail next. We also discuss the loss function used to train the VLM and how MediVLM computes a severity score for each generated report. One of the features of our framework is that it can generate reports when only images and no reports are available for training. This is achieved by using a frozen ClinicalBERT encoder and a fine-tuned BioT5 decoder to derive coherent pseudo-reports from the informative patches identified by Faster R-CNN, and then using the pseudo-reports as surrogate supervision to fine-tune a GPT-2 decoder. Our framework is modular, allows seamless integration of other image / text encoders and decoder components (depending on the application) and can be trained in an unsupervised manner to generate lucid medical reports.

3.2 Modules

(1) Object detector (frozen): Given a medical image, we first applied a Faster R-CNN object detection model (Ren et al., 2015), pre-trained on the MS-CXR dataset (Boecking et al., 2022) with a ResNet-34 backbone. Faster R-CNN consists of a region proposal network (RPN), which generates object proposal bounding boxes of potential anatomical regions. The image patches corresponding to these bounding boxes were used for further analysis. This allows the model to focus on the important regions of interest within the image for generating the report.

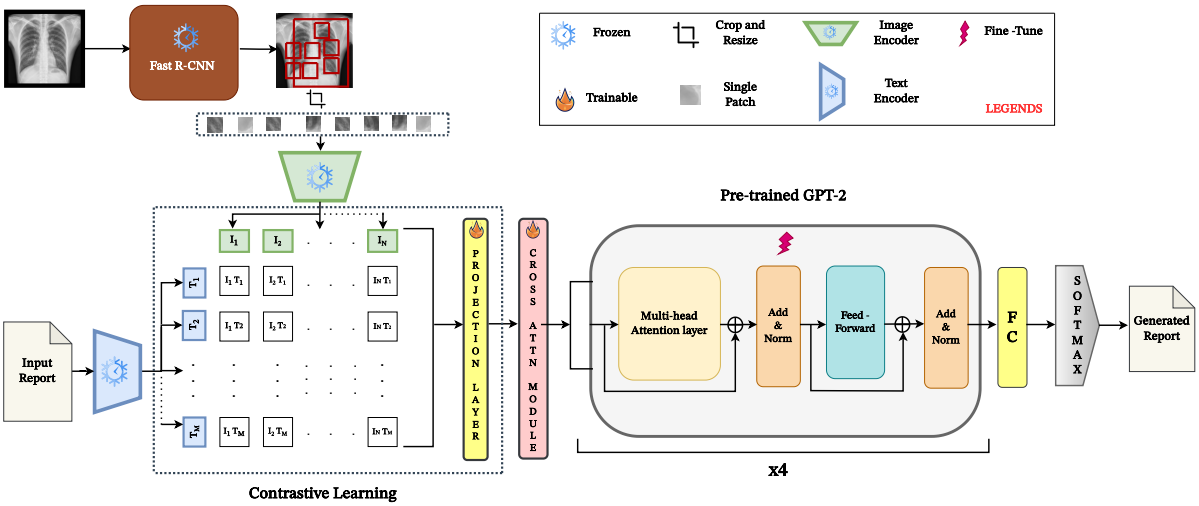


Figure 2: The proposed MediVLM architecture

(2) Image encoder (frozen): We used the p most confident patches (produced by Faster R-CNN), which were cropped and resized into fixed dimensions. A CLIP-ViT model (Radford et al., 2021) was then used as the image encoder to extract spatial features from these patches. While our model extracts Region of Interest (ROI) from the images, it does not discard spatial position information. The bounding box coordinates (x, y, w, h) generated by Faster R-CNN were encoded to capture the spatial location of each cropped and resized patch by normalizing these coordinates and passing them through a Multi-Layer Perceptron (MLP) to produce fixed-size vectors matching CLIP's embedding dimensions. These position embeddings were then concatenated with CLIP's image embeddings to provide the GPT-2 decoder (detailed below) with clinically interpretable spatial priors.

(3) Text encoder (frozen): We used ClinicalBERT (Huang et al., 2020) as the text encoder to tokenize the input radiology reports and embed them into latent representations. ClinicalBERT is a BERT model specifically designed for clinical text processing, and is thus well-suited for our application. The model is pre-trained on a large corpus of clinical notes from the Medical Information Mart for Intensive Care III (MIMIC-III) dataset. Our initial experiments revealed that using ClinicalBERT as the text encoder, rather than the CLIP text encoder, produces much better text embeddings, which improve the overall performance of the model. Our analysis also revealed that the input radiology reports often contain partial / incomplete sentences and incoherent text. To address this, we applied a BioT5 (Zhang et al., 2023b) decoder (fine-tuned with both clinical labels and partial reports from MIMIC-CXR) on the encoded text, to produce a refined embedding that corresponds to a concrete and coherent textual report.

(4) Image-text alignment and fusion (trainable): As mentioned before, we used CLIP-ViT as the image encoder and ClinicalBERT as the text encoder. As these models are pre-trained on different datasets, it is necessary to align the image and text representations before attempting to fuse them. After the image and text were encoded, we sampled N (image-text) pairs (v_i, t_i), $i = 1, \ldots N$ in a training mini-batch. The i^{th} image-text pair was aligned using the contrastive loss function (Radford et al., 2021):

$$\mathcal{L}^i_{contrast} = -\log \frac{\exp(\langle v_i, t_i \rangle / \tau)}{\sum_{j=1}^{N} \exp(\langle v_i, t_j \rangle / \tau)} \quad (1)$$

where $\langle . \rangle$ denotes cosine similarity and τ is a temperature parameter. The goal of this loss term was to learn modified image and text representations, such that positive (image-text) pairs were encoded to similar (closer) representations and negative pairs were encoded to different (farther) representations. This alignment is conducted in the projection layer (Figure 2). Following (Chang and Venkataraman, 2025), the modified visual and text representations (after alignment) were passed on to a cross attention layer to fuse the two into a latent representation:

$$\mathcal{F} = \text{CrossAttention}(V_{aligned}, T_{aligned}) \quad (2)$$

The final fused representation $\mathcal{F}$ was fed into a language model for report generation.

(5) Language model (fine-tuned): We used the following layers from the pre-trained GPT-2 model (Radford et al., 2019), as our language decoder.

Multi-Head Self-Attention (MHSA): Transformer models adopt the scaled dot-product attention, where the output is a weighted sum of the values (V), where the weight assigned to each value is determined by the dot-product of the query (Q) with all the keys (K) (Vaswani et al., 2017):

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{n}}\right)V \quad (3)$$

The multi-head mechanism enables the model to run through the scaled dot-product attention multiple times in parallel, and jointly attend to information from different representation subspaces at different positions (Vaswani et al., 2017). The independent attention outputs are concatenated, followed by linear transformation:

$$\text{MHSA}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \quad (4)$$

where each attention head is computed as:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (5)$$

where W_i^Q, W_i^K, W_i^V and W^O are parameter matrices to be learned. Attention dropout is added for better generalization.

Feed-Forward Network (FFN): The feed-forward network (FFN) operates on the output of the multi-head attention and is used to further process the attention outputs and capture more complex transformations. The FFN consists of two fully connected layers with the non-linear GELU activation function applied between them.

Residual Connection and Layer Normalization: Layer normalization stabilizes the inputs to each transformer layer by normalizing the hidden unit activations within the layer, ensuring they have zero mean and unit variance. Residual connections add the input of a layer back to its output, creating a shortcut for the gradient during backpropagation. Residual connections prevent vanishing or exploding gradients by enabling gradients to flow more directly through the network, especially in deep models like GPT-2. Layer normalization and residual connections are applied after the multi-head self-attention and feed-forward network stages to ensure stability and efficient learning.

As shown in Figure 2, these layers constitute one transformer block; we concatenated four such blocks in MediVLM, which were fine-tuned using our training data.

Report generation: After the input tokens pass through all the stacked transformer blocks, we get a refined set of hidden representations, one for each token in the input sequence. The fully connected layer takes this hidden state of each token from the final transformer block and maps it to the size of the vocabulary, producing logits. The logits are then passed through a softmax function, which transforms them into a probability distribution, allowing the model to predict the most likely next token in the sequence, thus generating the final report.

3.3 Loss Function

To train MediVLM, we used the cross entropy loss between the predicted tokens in the generated report and the ground truth tokens:

$$\mathcal{L}_{CE} = -\sum_{k=1}^{K} y_k \log(p_k) \quad (6)$$

where y_k is the true value, and p_k is the predicted probability for token k. We also included a contrastive loss term (as depicted in Equation (1)). The overall loss function is given by:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{CE} + \lambda_2 \mathcal{L}_{contrast} \quad (7)$$

where λ_1 and λ_2 are weights governing the relative importance of the two terms. Note that, the cross entropy loss was used to train the cross attention fusion module, as well as fine-tune the transformer blocks in the GPT-2 decoder. The contrastive loss was used to align the visual and text embeddings (from the respective encoders) within the projection layer, before passing them on to the cross attention fusion module.

3.4 Severity Score Computation

One of the useful features of MediVLM is that it computes a severity score corresponding to each generated report, depicting the seriousness of the patient's medical condition. The severity score of a report d was computed using the term frequency-inverse document frequency (TF-IDF) score (Sammut and Webb, 2011). The TF-IDF score of a term t in a document d is a metric that measures how relevant the term is to the document. We constructed a set $T_s = \{t_1, t_2, \ldots, t_m\}$ containing a collection of terms relevant to severity. T_s might include terms like "*severe*", "*critical*", "*unstable*", "*emergency*", "*shock*" etc. The severity score for a report d was computed as the sum of the TF-IDF scores of the

severity-related terms T_s in d:

$$S(d) = \sum_{t_i \in T_s} \text{TF-IDF}(t_i, d) \quad (8)$$

We normalized the severity scores across all the generated reports to derive a score between 0 and 1. Our empirical results demonstrate that such a simple strategy of computing the severity score can be useful in understanding the criticality of a patient's medical condition.

3.5 Unsupervised Training of MediVLM

We propose a novel unsupervised learning framework for radiology report generation that operates in the absence of annotated reports during training, thereby addressing a critical bottleneck in medical AI, where expert-curated reports are expensive to acquire. Given an image x_i, the Faster R-CNN detector produces semantic label and box pairs (l_i, b_i); the label l_i is passed as an input to the frozen ClinicalBERT model to produce an embedding. As with supervised training, a BioT5 decoder is used to produce concrete clinical sentence embeddings based on the ClinicalBERT encoded output, and refine the clinical text via autoregressive decoding to generate a pseudo-report. Unlike existing methods that rely on fully supervised data, our method learns to produce coherent and diagnostically relevant pseudo-reports from raw semantic labels or *partially observed text*. As before, the image and text embeddings are aligned via contrastive learning and passed as input to a GPT-2 decoder, to generate reports autoregressively. The training targets are the pseudo-reports generated from the pseudo-labeling pipeline, and the decoder is optimized to minimize the token-level cross-entropy loss between the generated sequence and the pseudo-report. To the best of our knowledge, this is the first framework to combine CLIP-style (image - text) alignment, ClinicalBERT-based encoding, and BioT5-driven modeling in a self-supervised medical imaging pipeline. This setup enables our GPT-2 model to learn to generate clinically coherent reports even in the complete absence of annotated training reports.

4 Experiments and Results

Datasets. We used three benchmark datasets to study the performance of MediVLM: (i) **IU X-Ray** (Demner-Fushman et al., 2016), a public radiography dataset collected by Indiana University; (ii) **CASIA-CXR** (Metmer and Yang, 2024), consisting of high-resolution chest radiographs accompanied by narrative reports written in French (we used the Pneumonia category in our experiments); (iii) **MIMIC-CXR** (Johnson et al., 2019), the largest publicly available dataset of chest radiographs with free-text radiology reports. The statistics of these datasets are shown in Section A.2 of the Appendix.

Evaluation Metrics. We evaluated the performance of our model on widely used natural language generation (NLG) metrics such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005) and ROUGE-L (Lin, 2004), which measure the similarity between the generated and reference reports. We also evaluated the performance of our model using metrics that assess the clinical / diagnostic relevance of AI generated medical reports, such as BERTScore(Zhang et al., 2019), RadGraph-F1(Yu et al., 2023) and RaTEScore(Zhao et al., 2024).

Comparison Baselines. We used five recent methods as comparison baselines in our work: (i) R2Gen (Chen et al., 2020); (ii) X-RGen (Chen et al., 2024); (iii) CMN (Chen et al., 2021); (iv) HistGen (Guo et al., 2024); and (v) RL (Qin and Song, 2022). These baselines were selected to cover a wide range of report generation techniques, including memory-driven transformers, methods that mimic the behavior of human radiologists, methods based on multiple instance learning and reinforcement learning. For the MIMIC-CXR dataset, we used several other comparison baselines, as detailed in Table 1.

Implementation Details. Please refer to Section A.1 of the Appendix for our implementation and training parameter details.

4.1 Main Results

The results are depicted in Table 1. For the IU X-Ray dataset, MediVLM comprehensively outperforms all the baselines and achieves the highest scores in terms of all the metrics. This shows the efficacy of our method and its ability to outperform recently proposed methods like HistGen and X-RGen. For the CASIA-CXR (Pneumonia) dataset, R2Gen achieves the highest BLEU-4 score and X-RGen achieves the highest BLEU-3 score. MediVLM achieves the best results in the other 4 metrics. These results are particularly encouraging as they corroborate that MediVLM can depict promising results even when the radiology reports are provided in a language other than English. We com-

Dataset	Method	Year	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
IU X-Ray	R2Gen(Chen et al., 2020)	2020	0.451	0.288	0.212	0.167	0.183	0.362
	X-RGen(Chen et al., 2024)	2024	0.454	0.290	0.210	0.161	0.187	0.361
	CMN(Chen et al., 2021)	2021	0.408	0.242	0.160	0.113	0.165	0.313
	HistGen(Guo et al., 2024)	2024	0.427	0.249	0.172	0.128	0.161	0.335
	RL(Qin and Song, 2022)	2022	0.280	0.135	0.064	0.029	0.119	0.220
	Ours		**0.471**	**0.296**	**0.231**	**0.170**	**0.194**	**0.375**
CASIA-CXR (Pneumonia)	R2Gen(Chen et al., 2020)	2020	0.623	0.568	0.530	**0.498**	0.357	0.614
	X-RGen(Chen et al., 2024)	2024	0.642	0.578	**0.531**	0.492	0.352	0.615
	CMN(Chen et al., 2021)	2021	0.565	0.501	0.458	0.424	0.313	0.536
	HistGen(Guo et al., 2024)	2024	0.587	0.525	0.488	0.455	0.315	0.541
	RL(Qin and Song, 2022)	2022	0.182	0.133	0.107	0.088	0.142	0.186
	Ours		**0.673**	**0.582**	0.510	0.476	**0.371**	**0.628**
MIMIC-CXR	R2Gen(Chen et al., 2020)	2020	0.353	0.218	0.145	0.103	0.142	0.277
	CMN (Chen et al., 2021)	2021	0.353	0.218	0.148	0.106	0.142	0.278
	PPKED(Liu et al., 2021a)	2021	0.360	0.224	0.149	0.106	0.149	0.284
	$\mathcal{M}^2$ TR. PROG. (Nooralahzadeh et al., 2021)	2021	**0.378**	0.232	0.154	0.107	0.145	0.272
	AlignTransformer(You et al., 2021)	2021	**0.378**	0.235	0.156	0.112	**0.158**	0.283
	KnowMat(Yang et al., 2022)	2022	0.363	0.228	0.156	0.115	-	0.284
	CMCA(Song et al., 2022)	2022	0.360	0.227	0.156	0.117	0.148	0.287
	RAMT(Zhang et al., 2023a)	2023	0.362	0.229	0.157	0.113	0.153	0.284
	MedCycle(Hirsch et al., 2024a)	2024	0.352	0.194	0.114	0.070	0.132	0.241
	MedRAT(Hirsch et al., 2024b)	2024	0.365	-	-	0.086	0.132	0.251
	Ours		0.377	**0.246**	**0.158**	**0.123**	0.149	**0.293**

Table 1: Performance comparison of MediVLM on the IU X-Ray, CASIA-CXR and MIMIC-CXR datasets. For IU X-Ray and CASIA-CXR, the results of the baselines were obtained using the official codes provided by the authors. The results of the baselines for MIMIC-CXR were cited from (Tanida et al., 2023; Hou et al., 2024; Hirsch et al., 2024a,b). The best results are marked in **bold** and the second best results are underlined.

Dataset	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
IU X-Ray	0.241	0.158	0.124	0.077	0.09	0.188
CASIA-CXR (Pneumonia)	0.378	0.278	0.227	0.294	0.191	0.293
MIMIC-CXR	0.203	0.121	0.096	0.059	0.104	0.168

Table 2: Performance of MediVLM in the **unsupervised setup** (only images, no reports available for training) on the IU X-Ray, CASIA-CXR, and MIMIC-CXR datasets.

pared the performance of our method against several recent baselines for the MIMIC-CXR dataset. MediVLM achieves the highest scores across 4 metrics and the second highest score across 1 metric. These results corroborate the promise and potential of MediVLM to generate high quality radiology reports from medical images, and address the real-world challenge of the shortage of trained radiologists in many healthcare systems.

4.2 Results using Clinical / Diagnostic Relevance Metrics

Table 3 reports the results of MediVLM and several recent baseline methods on the IU X-Ray dataset, using 3 metrics based on clinical / diagnostic relevance: BERTScore, RadGraph-F1 and RaTEScore. MediVLM comprehensively outperforms all the baselines in terms of all the three metrics. X-RGen produces the second best results and MediVLM outperforms X-RGen by 0.093, 0.019 and 0.036 in terms of the three metrics respectively. This corroborates the efficacy of MediVLM to generate reports that capture clinically relevant information.

4.3 Unsupervised Training Results

Table 2 reports the results of MediVLM in the challenging setting where only images and no reports are available during training. As expected, the values are lower than those in the completely supervised setup (Table 1). However, our unsupervised method outperforms the RL baseline (supervised) for the CASIA-CXR dataset in terms of all the metrics. It also outperforms the RL baseline in terms of the BLEU-2, BLEU-3 and BLEU-4 metrics for the IU X-Ray dataset. This shows the efficacy of our pseudo-labeling pipeline (detailed in Section 3.5) to generate pseudo-reports and address the challenge of training MediVLM in the absence of ground truth radiology reports. These results are particularly encouraging, considering the shortage of trained radiologists for the labor-intensive task of report transcription from medical images.

Method	BERTScore	RadGraph-F1	RaTEScore
R2Gen(Chen et al., 2020)	0.487	0.251	0.603
X-RGen(Chen et al., 2024)	<u>0.523</u>	<u>0.387</u>	<u>0.686</u>
CMN(Chen et al., 2021)	0.509	0.281	0.603
HistGen(Guo et al., 2024)	0.466	0.205	0.559
RL(Qin and Song, 2022)	0.331	0.228	0.389
RGRG(Tanida et al., 2023)	0.437	0.223	0.620
VLCI(Chen et al., 2023)	0.455	0.288	0.679
RadFM(Wu et al., 2023)	0.459	0.230	0.627
RaDialog(Pellegrini et al., 2023)	0.444	0.205	0.586
CvT2DistilGPT2(Nicolson et al., 2023)	0.482	0.265	0.620
Ours	**0.616**	**0.406**	**0.722**

Table 3: Performance comparison of MediVLM on the IU X-Ray dataset using clinical / diagnostic relevance metrics (BERTScore, RadGraph-F1 and RaTEScore). The results of the first 5 baselines were obtained using the official codes provided by the authors. The results of the next 5 baselines were cited from https://rexrank.ai/?utm_source=chatgpt.com. Best results are marked in **bold** and the second best results are <u>underlined</u>.

ClinicalBERT	Selective Patching	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
✓	×	0.426	0.222	0.18	0.145	0.138	0.311
×	✓	0.396	0.204	0.191	0.136	0.120	0.307
✓	✓	**0.471**	**0.296**	**0.231**	**0.170**	**0.194**	**0.375**

Table 4: Ablation study results . Best results are marked in **bold**.

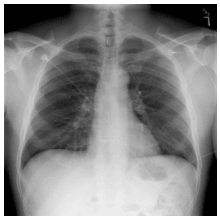
Performance Analysis. Medical reports are often known to contain language complexities like abbreviations, unfinished sentences etc. which can degrade the performance of models trained to generate reports. MediVLM leverages a pre-trained ClinicalBERT text encoder (that is specifically trained for clinical text processing) and a fine-tuned BioT5 decoder that refines clinical text via autoregressive decoding and produces coherent sentence embeddings from partially observed / incomplete text, thereby mitigating the aforementioned issues. Further, MediVLM uses a pre-trained Faster R-CNN model to extract the salient patches from an input medical image that enable the model to focus on the clinically relevant portions, and avoid irrelevant / redundant details. These factors result in the improved performance of MediVLM in terms of both natural language generation metrics, as well as clinical / diagnostic relevance metrics, as evident from Tables 1 and 3 respectively.

4.4 Ablation Studies

We conducted ablation studies on the IU X-Ray dataset to study the effects of two components in our MediVLM architecture: (i) extracting the image patches containing the salient anatomical regions using the Fast R-CNN model; and (ii) using the ClinicalBERT as the text encoder to tokenize the input radiology reports. The results are reported in Table 4. We note that, if we pass the entire image to the visual encoder (instead of passing only the salient anatomical regions), the performance degrades in terms of all the metrics. This shows that extracting the salient image patches for the model to focus on, results in a performance improvement. We also note that using the CLIP text encoder (instead of ClinicalBERT) to tokenize the reports, results in a performance drop in terms of all the metrics. This shows the utility of ClinicalBERT, which is specially trained for clinical text processing, as the text encoder of MediVLM. A general purpose text encoder like CLIP may not be able to aptly capture the information specific to medical reports. The best results are obtained when both the components are included, as in our MediVLM architecture.

5 Visual Illustrations: Unsupervised Training of MediVLM

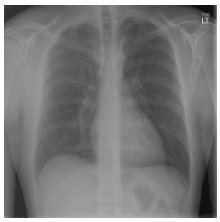
Figure 3 depicts a couple of low severity examples on the IU X-Ray dataset, when MediVLM is trained in an unsupervised manner (only images and no reports are available for training). For the first image, the GT report mentions that *the heart size is normal, the lungs are clear with no signs of consolidation, the hilar and mediastinal contours are normal, and there are no acute abnormalities*. All these findings are appropriately captured in the MediVLM report. Further, since there are no abnor-



MediVLM(Unsupervised): The heart is of normal size with no abnormalities. The lungs are clear with no signs of consolidation or fluid accumulation. The structures around the lungs, including the mediastinum and hilar regions are normal. Overall, there are no acute or concerning findings.

GT: Heart size is normal. The lungs are clear. There are no focal air space consolidations. No pleural effusions or pneumothoraces. The hilar and mediastinal contours are normal. Normal pulmonary vascularity. No acute abnormality.

Severity Score: 0.22



MediVLM(Unsupervised): The heart is normal size and the contours of the mediastinal area are unremarkable. The lungs appear clear with no visible signs of infection, fluid accumulation. The rib structures are intact with no evidence of fractures or dislocations. Overall, the examination reveals no acute issues with the heart or lungs.

GT: Heart size and cardiomediastinal contours are normal. Lungs are clear without focal airspace opacity, pleural effusion, or pneumothorax. No displaced rib fracture.Negative for acute cardiopulmonary findings.

Severity Score: 0.23

Figure 3: Low severity examples on the IU X-Ray dataset. Report generated using **unsupervised training of MediVLM** (only images, no reports available for training). Matching texts in the ground truth report and the generated output are highlighted with the same color. Best viewed in color.

malities, this has been designated as a low severity case (score = 0.22). The same observations are evident for the second image, where MediVLM correctly captures the important findings and designates it to be a low severity case, as there are no significant abnormalities.

These examples demonstrate that MediVLM can appropriately capture the important findings in a medical image even when no radiology reports are available for training. It can also compute a severity score which appropriately reflects the seriousness of a patient's medical condition, even when it is trained in an unsupervised manner. These results are extremely important from a practical standpoint, given the shortage of trained radiologists to conduct the labor-intensive task of report transcription.

We include an analysis of the values of different parameters (Section A.3) and more visual illustrations demonstrating the performance of MediVLM and the severity scores (Section A.4) in the Appendix.

6 Conclusion

In this paper, we proposed MediVLM, a vision language model for radiology report generation from medical images. Apart from report generation, MediVLM also furnishes a severity score for each generated report, depicting the seriousness of the patient's medical condition, which can be used to prioritize patients who need immediate medical attention. Further, it can also generate reports when only images and no reports are available for training; this is an extremely useful feature, as curating such reports is a labor-intensive task. Our extensive empirical results on three benchmark datasets (including a dataset where the reports are provided in French) corroborated the promise and potential of our framework against competing baselines.

7 Limitations

Although our framework has depicted promising performance, it still has some limitations. The textual reports and severity scores generated by MediVLM were not validated by medical professionals for their correctness (this is an integral part of our ongoing research). A qualitative user study with domain experts will further validate the usefulness of MediVLM for clinical applications. Further, our framework generates radiology reports from Chest X-ray images; future investigations should extend its applicability to other types of medical images.

8 Acknowledgment

This research was supported in part by the National Science Foundation under Grant Number: IIS-2143424 (NSF CAREER Award).

References

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

Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, and 1 others. 2023. Learning

to exploit temporal structure for biomedical vision-language processing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15016–15027.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay. 2022. Making the most of text semantics to improve biomedical vision–language processing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1033–1047. Association for Computational Linguistics.

Tzu-Tao Chang and Shivaram Venkataraman. 2025. LV-XAttn: Distributed cross-attention for long visual inputs in multimodal large language models. In *arXiv:2502.02406v2*.

Qi Chen, Yutong Xie, Biao Wu, Xiaomin Chen, James Ang, Minh-Son To, Xiaojun Chang, and Qi Wu. 2024. Act like a radiologist: Radiology report generation across anatomical regions. In *Asian Conference on Computer Vision (ACCV)*, pages 1–17.

Weixing Chen, Yang Liu, Ce Wang, Jiarui Zhu, Guanbin Li, and Liang Lin. 2023. Cross-modal causal intervention for medical report generation. *Preprint*, arXiv:2303.09117.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 2021. Cross-modal memory networks for radiology report generation. In *Association for Computational Linguistics and International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 5904–5914, Online.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. Generating radiology reports via memory-driven transformer. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis P Langlotz. 2022. Improving the factual correctness of radiology report generation with semantic rewards. In *Empirical Methods in Natural Language Processing (EMNLP)*.

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

Stacy K Goergen, Felicity J Pool, Tari J Turner, Jane E Grimm, Mark N Appleyard, Carmel Crock, Michael C Fahey, Michael F Fay, Nicholas J Ferris, Susan M Liew, and 1 others. 2013. Evidence-based guideline for the written radiology report: Methods, recommendations and implementation challenges. *Journal of medical imaging and radiation oncology*, 57(1):1–7.

Zhengrui Guo, Jiabo Ma, Yingxue Xu, Yihui Wang, Liansheng Wang, and Hao Chen. 2024. Histgen: Histopathology report generation via local-global feature encoding and cross-modal context interaction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 189–199. Springer.

Elad Hirsch, Gefen Dawidowicz, and Ayellet Tal. 2024a. Medcycle: unpaired medical report generation via cycle-consistency. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

Elad Hirsch, Gefen Dawidowicz, and Ayellet Tal. 2024b. MedRat: Unpaired medical report generation via auxiliary tasks. In *European Conference on Computer Vision (ECCV)*, pages 18–35. Springer.

Wenjun Hou, Yi Cheng, Kaishuai Xu, Yan Hu, Wenjie Li, and Jiang Liu. 2024. ICON: Improving inter-report consistency in radiology report generation via lesion-aware mixup augmentation. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Wenjun Hou, Yi Cheng, Kaishuai Xu, Wenjie Li, and Jiang Liu. 2023. RECAP: Towards precise radiology report generation via dynamic disease progression reasoning. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2020. Clinicalbert: Modeling clinical notes and predicting hospital readmission. In *Conference on Health, Inference, and Learning (CHIL) Workshop*.

Baoyu Jing, Zeya Wang, and Eric Xing. 2019. Show, describe and conclude: On exploiting the structure information of chest X-ray reports. In *Association for Computational Linguistics (ACL)*.

Baoyu Jing, Pengtao Xie, and Eric Xing. 2018. On the automatic generation of medical imaging reports. In *Association for Computational Linguistics (ACL)*.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.

Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. DenseCap: Fully convolutional localization networks for dense captioning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4565–4574.

Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. 2018. Hybrid retrieval-generation reinforced agent for medical image report generation. *Advances in Neural Information Processing Systems (NeurIPS)*, 31.

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

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

Fenglin Liu, Changchang Yin, Xian Wu, Shen Ge, Yuexian Zou, Ping Zhang, and Xu Sun. 2021b. Contrastive attention for automatic chest X-ray report generation. In *Association for Computational Linguistics (ACL)*.

Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. 2019. Clinically accurate chest x-ray report generation. In *Machine Learning for Healthcare Conference*, pages 249–269. PMLR.

Hichem Metmer and Xiaoshan Yang. 2024. An open chest x-ray dataset with benchmarks for automatic radiology report generation in French. *Neurocomputing*, 609:128478.

Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. 2021. Improving factual completeness and consistency of image-to-text radiology report generation. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 5288 – 5304.

Aaron Nicolson, Jason Dowling, and Bevan Koopman. 2023. Improving chest X-ray report generation by leveraging warm starting. *Artificial Intelligence in Medicine*, 144:102633.

Toru Nishino, Yasuhide Miura, Tomoki Taniguchi, Tomoko Ohkuma, Yuki Suzuki, Shoji Kido, and Noriyuki Tomiyama. 2022. Factual accuracy is not enough: Planning consistent description order for radiology report generation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 7123–7138.

Farhad Nooralahzadeh, Nicolas Perez Gonzalez, Thomas Frauenfelder, Koji Fujimoto, and Michael Krauthammer. 2021. Progressive transformer-based generation of radiology reports. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Association for Computational Linguistics (ACL)*.

Chantal Pellegrini, Ege Özsoy, Benjamin Busam, Nassir Navab, and Matthias Keicher. 2023. Radialog: A large vision-language model for radiology report generation and conversational assistance. *arXiv preprint arXiv:2311.18681*.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. In *Technical Report, OpenAI*.

Suhail Raoof, David Feigin, Arthur Sung, Sabiha Raoof, Lavanya Irugulpati, and Edward C Rosenow III. 2012. Interpretation of plain chest roentgenogram. *Chest*, 141(2):545–558.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 28.

Abi Rimmer. 2017. Radiologist shortage leaves patient care at risk, warns royal college. *BMJ: British Medical Journal (Online)*, 359.

Andrew B Rosenkrantz, Danny R Hughes, and Richard Duszak Jr. 2016. The US radiologist workforce: an analysis of temporal and geographic variation by using large national datasets. *Radiology*, 279(1):175–184.

C. Sammut and G. Webb. 2011. *Encyclopedia of Machine Learning*. Springer.

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

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30.

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

Lin Wang, Munan Ning, Donghuan Lu, Dong Wei, Yefeng Zheng, and Jie Chen. 2022. An inclusive task-aware framework for radiology report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 568–577. Springer.

Xinyi Wang, Grazziela Figueredo, Ruizhe Li, Wei Emma Zhang, Weitong Chen, and Xin Chen. 2024. A survey of deep learning-based radiology report generation using multimodal data. *arXiv preprint arXiv:2405.12833*.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *arXiv preprint arXiv:2308.02463*.

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 *International Conference on Machine Learning (ICML)*, pages 2048–2057. PMLR.

Wenting Xu, Chang Qi, Zhenghua Xu, and Thomas Lukasiewicz. 2021. Reinforced medical report generation with x-linear attention and repetition penalty. In *AAAI Conference on Artificial Intelligence*.

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

Mohammed Yeasin, Kazi Ashraf Moinuddin, Felix Havugimana, Lijia Wang, and Paul Park. 2024. AutoRad: End-to-end report generation from lumber spine mri using vision-language model. *Journal of Clinical Medicine*, (23).

Di You, Fenglin Liu, Shen Ge, Xiaoxia Xie, Jing Zhang, and Xian Wu. 2021. AlignTransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 72–82. Springer.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4651–4659.

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y Ng, and 1 others. 2023. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9).

Ke Zhang, Hanliang Jiang, Jian Zhang, Qingming Huang, Jianping Fan, Jun Yu, and Weidong Han. 2023a. Semi-supervised medical report generation via graph-guided hybrid feature consistency. *IEEE Transactions on Multimedia*, 26:904–915.

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

Yuxuan Zhang, Shangqing Wu, Yicheng Liu, Kangmin Huang, Kai Wang, Xin Gao Wang, Yuxuan Zhang, Xiaoyong Li, and Dong Yu. 2023b. BioT5: Enriching cross-modal integration in biology with chemical knowledge and natural language associations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Ratescore: A metric for radiology report generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15004–15019.

A Appendix

In this Appendix, we provide the following:

- The implementation and training parameter details (Section A.1)
- The split (train / val / test) used to train our models for each dataset (Section A.2)
- Parameter value analysis (Section A.3)
- Visual illustrations depicting the performance of MediVLM (Section A.4)

A.1 Implementation Details

Each input image was resized to 224×224 pixels. Patches of size 28×28 were then extracted using the Fast R-CNN object detector. We used the top 8 patches with the highest confidence scores, which were passed to the CLIP-ViT-L/14 model for visual feature extraction. The ClinicalBERT text encoder was used to tokenize the input report. The BioT5 decoder was used in conjunction with ClinicalBERT and was fine-tuned to refine clinical text to produce coherent sentence embeddings. The GPT-2 decoder model was fine-tuned with learning rate of $2.e^{-5}$ and a batch size of 32. The fine-tuning converges between 30 and 50 epochs. We used the *AdamW* optimizer to fine tune the GPT-2 decoder. The same parameters were used to train the cross attention layer. The values of λ_1 and λ_2 in Equation (7) were taken as 1 and 0.7 respectively. The value of the temperature parameter τ in Equation (1) was taken as 0.07. We used Python's Natural Language Toolkit (NLTK) to capture tokens from the input radiology reports, and NLTK's *SentimentAnalyzer* to identify the positive and negative tokens. The negative tokens were used to construct the set T_s for severity computation. Our code will be made publicly available upon acceptance of our paper.

A.2 Dataset Splits

The number of samples in the training, validation and test sets for the IU X-Ray, CASIA-CXR and MIMIC-CXR datasets are reported in Table 5.

A.3 Parameter Value Analysis

In this section, we study the effects of different parameters on the performance of MediVLM.

Number of patches. We conducted an experiment to study the effect of the number of patches selected by the Faster R-CNN model, which are passed onto the image encoder. The results on the IU X-Ray dataset for $6, 8, 12$ and 16 patches are reported in Table 6. We note that using 8 patches produces the best results across most of the metrics; we therefore used this value in our experiments.

Number of sentences. We also conducted an experiment to study the effect of the number of sentences (maximum number of allowable tokens) in the report generated by MediVLM. The results on the IU X-Ray dataset with $60, 77$ and 80 maximum allowable tokens (which correspond to 3, 4 and 5 sentences respectively) are reported in Table 7. We used 77 max tokens (4 sentences) in our experiments, as it produces the best results across most of the metrics.

Weight parameters. We further conducted an experiment to study the effects of the weight parameters λ_1 and λ_2 in the training loss function (Equation (7)). The results on the IU X-ray dataset with different combinations of the parameter values are presented in Table 8. We note that $\lambda_1 = 1$ and $\lambda_2 = 0.7$ produces the best results across most of the metrics. We therefore uses these values in our experiments.

A.4 Visual Illustrations

A.4.1 MediVLM vs. Baselines

Figure 4 provides comparative visual illustrations of the reports generated by all the methods for two given medical images from the IU X-Ray dataset. For the first image, the GT report mentions four major findings: (i) *the lungs are clear (with no effusion or pneumothorax);* (ii) *the heart size is normal;* (iii) *the bony thorax is unremarkable; and* (iv) *cardiopulmonary abnormalities are absent*. The baseline methods capture some of these findings, but fail to capture all four. For instance, RL only captures information about the clear lungs. HistGen mentions about the normal heart and no pleural effusion. R2Gen and X-RGen capture information about normal heart, clear lungs and no effusions. CMN captures information about the normal heart and lungs, with no pleural effusions and no bony findings. MediVLM is the only method that correctly captures all these four findings (evident from the matching colors in the text).

The same observation is evident for the second image, where MediVLM aptly captures all the important findings mentioned in the GT report (*normal heart size, prominent right paratracheal soft tissue density, rounded mass with correct measurement in the right middle lobe, absence of pleural effusions, intact bony thorax, right mid lung mass with mild right paratracheal soft tissue*), and even recommends a further imaging with a CT scan of the chest. The baseline methods each capture only a subset of the findings, but not all of them. These results further corroborate the efficacy of MediVLM over the competing baselines and account for its superior performance in Table 1.

A.4.2 MediVLM: Severity Scores

Visual illustrations of a few low severity cases on the IU X-Ray dataset are show in Figure 5. For the first image, for instance, the GT report mentions that *the cardiac and mediastinal contours are within normal limits, the lungs are clear, the bony structures are intact and there are no acute findings overall*. Each of these findings is captured correctly in the MediVLM report. This has been designated as a low severity case (score = 0.18) as there are no acute abnormalities. The same pattern is evident in all the other images, where MediVLM captures the primary findings mentioned in the GT report. Further, these reports are mostly normal with few serious concerns and are thus designated as low severity cases.

Visual illustrations of a few high severity cases on the IU X-Ray dataset are shown in Figure 6. For the first image, for instance, the GT report mentions that *the heart is mildly enlarged, the lung volumes are low, the bony structures are within normal limits and there is no free air under the diaphragm. A mild amount of abnormality is visible in the transverse colon. Overall, there are no acute cardiopulmonary findings.* All of these findings are aptly captured in the report generated by MediVLM. This has been designated as a moderately high severity case (score = 0.69) due to the mildly enlarged heart and the presence of an abnormality in the transverse colon. A similar pattern is evident for all the examples

Splits	Datasets					
	IU X-Ray		MIMIC-CXR		CASIA-CXR*	
	Image	Report	Image	Report	Image	Report
Train	5.2K	2.8K	369.0K	222.8K	1.8k	1.8k
Val	0.7K	0.4K	3.0K	1.8K	0.1k	0.1k
Test	1.5K	0.8K	5.2K	3.3K	0.1k	0.1k

Table 5: Details of the IU X-RAY, MIMIC-CXR, and CASIA-CXR*(Pneumonia only) datasets

No. of patches	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
6	0.437	0.269	0.228	0.167	0.165	0.348
8	0.471	**0.296**	**0.231**	0.170	**0.194**	**0.375**
12	**0.477**	0.279	0.227	**0.199**	0.181	0.350
16	0.473	0.283	0.226	0.197	0.179	0.353

Table 6: Study of the effect of the number of patches used for visual encoding on the IU X-Ray dataset. Best results are marked in **bold**.

No. of sentences	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
3 (60 tokens)	0.463	0.257	0.201	0.169	0.165	0.351
4 (77 tokens)	**0.471**	**0.296**	**0.231**	0.170	**0.194**	**0.375**
5 (80 tokens)	**0.471**	0.249	0.211	**0.192**	0.170	0.346

Table 7: Study of the effect of the maximum number of allowable tokens (number of sentences) in the generated report on the IU X-Ray dataset. Best results are marked in **bold**.

	λ_1 (CE)	λ_2 (Contrastive)	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
	1	0.5	0.438	0.267	0.21	0.158	0.189	0.367
	1	1	0.463	0.292	0.225	**0.172**	0.19	0.373
	1	2	0.389	0.247	0.191	0.135	0.154	0.316
	0.5	1	0.378	0.241	0.187	0.136	0.158	0.321
Ours	1	0.7	**0.471**	**0.296**	**0.231**	0.170	**0.194**	**0.375**

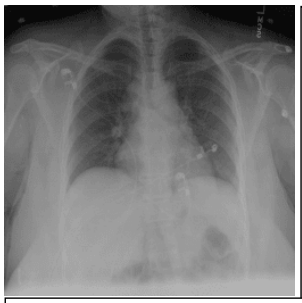
Table 8: Study of the effects of the weight parameters λ_1 and λ_2 in the training loss function (Equation (7)) on the IU X-Ray dataset. Best results are marked in **bold**.

where MediVLM captures the important findings mentioned in the GT report. Due to the presence of notable abnormalities, all of these are designated as high severity cases (with high severity scores).

These examples further demonstrate that MediVLM can capture the important medical information from images and generate high quality reports summarizing the findings. It can also compute a meaningful severity score which appropriately determines the severity of a patient's medical condition, which can be used to identify patients that need immediate medical attention.

A.4.3 MediVLM: Unsupervised Training

A couple of high severity examples on the IU X-Ray dataset are depicted in Figure 7, when MediVLM is trained in an unsupervised manner (only images and no reports are available for training). For the first image, the GT report mentions that *the lung volumes are low, the heart size is borderline enlarged; otherwise, no typical findings of pulmonary edema are noted.* These are all correctly captured in the report generated by MediVLM. This has been designated as a moderately high severity case (score = 0.46) due to the enlarged



R2Gen: Heart size normal. Lungs are clear. No pneumonia effusions edema pneumothorax adenopathy nodules or masses.

X-RGen: The heart is normal in size. The lungs are clear. There is no pleural effusion or pneumothorax .

RL: The lungs are clear. No pleural effusion or pneumothorax.

HistGen: The cardiomediastinal silhouette is normal in size and contour and contour. There is no pneumothorax pleural effusion or focal air space opacity.

CMN: The cardiomediastinal silhouette and pulmonary vasculature are within normal limits in size. The lungs are clear of focal airspace disease pneumothorax or pleural effusion. There are no acute bony findings.

MediVLM: The lungs appear normal without any signs of infection, fluid buildup, or collapsed areas. The heart is of normal size, indicating no visible enlargement or abnormalities. The bones of the chest do not show any significant issues. Overall, there are no signs of any acute heart or lung problems.

GT: Lungs are clear without focal consolidation, effusion, or pneumothorax. Normal heart size. Bony thorax grossly unremarkable. Negative for acute cardiopulmonary abnormality.

R2Gen: The heart is normal in size. The mediastinum is stable. the lungs are clear.

X-RGen: The heart is normal in size. The lungs are clear. There is no pleural effusion or pneumothorax.

RL: The heart is normal in size. The lungs have some problems. There is no pleural effusion or pneumothorax .

HistGen: Lateral views of the chest were obtained. The heart is normal in size and contour. There is no acute bony abnormality.

CMN: The heart is normal in size. The mediastinum is unremarkable. The lungs are clear.

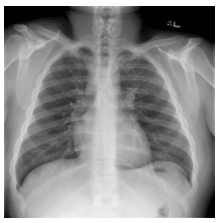
MediVLM: The heart is normal in size, but there is a prominent right paratracheal soft tissue density. A rounded mass measuring approximately 4.6 cm x 3.7 cm is seen in the right middle lobe, with mild surrounding airspace disease or atelectasis. No pleural effusions are noted, and the bony thorax appears intact. The right mid-lung mass, along with the paratracheal soft tissue, may indicate associated lymphadenopathy, and further imaging with a CT scan of the chest is recommended.

GT: The heart is normal in size. Prominent right paratracheal soft tissue density. Rounded mass in the right middle lobe measures approximately 4.6 cm x 3.7 cm. There is mild surrounding airspace disease and/or atelectasis. No pleural effusions noted. The visualized bony thorax appears grossly intact. Right mid lung mass with mild right paratracheal soft tissue may be secondary to associated lymphadenopathy. Further imaging with CT of the chest is recommended.

Figure 4: Visual illustration of MediVLM vs. the baselines on the IU X-Ray dataset. Best viewed in color.

heart size and low lung volumes. The same observations are evident for the second image, where MediVLM correctly captures the important findings and designates it to be a high severity case (score = 0.61).

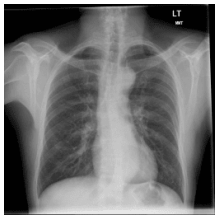
These examples further demonstrate that MediVLM can appropriately capture the important findings in a medical image even when no radiology reports are available for training, corroborating its usefulness in real-world applications, where there is a shortage of trained radiologists to conduct the labor-intensive task of report transcription.



MediVLM: The cardiac and mediastinal contours are normal. The lungs are clear with no signs of infection or abnormalities. The bony structures are intact. Overall, there are no acute findings.

GT: Cardiac and mediastinal contours are within normal limits. The lungs are clear. Bony structures are intact.No acute findings.

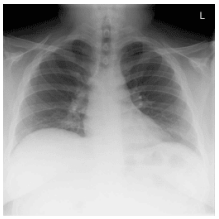
Severity Score: 0.18



MediVLM: The heart is of normal size, but the aorta appears tortuous and slightly dilated. The lungs are clear with no signs of infection, fluid buildup, or air leakage. The bones are intact with no fractures or abnormalities. Overall, there are no acute heart or lung issues detected.

GT: The heart is normal in size. The aorta is tortuous and ectatic. The lungs are clear without focal airspace opacity, pleural effusion, or pneumothorax. The osseous structures are intact.No acute cardiopulmonary findings.

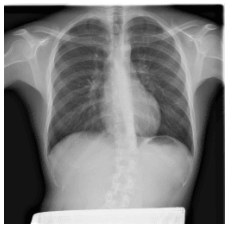
Severity Score: 0.33



MediVLM: The heart size is normal, and the lungs appear clear, although lung volumes are low. There is no evidence of pneumothorax or large pleural effusion. Overall, there are no acute heart or lung abnormalities detected.

GT: Heart size is normal. Lungs are clear. Low lung volumes. There is no pneumothorax or large pleural effusion.No acute cardiopulmonary findings.

Severity Score: 0.17

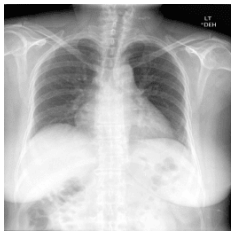


MediVLM: The heart size and pulmonary vascularity are normal, and the lungs show no signs of focal airspace disease. There is no pleural effusion or pneumothorax. Levoscoliosis of the thoracolumbar spine is noted, but there is no evidence of active disease in the chest.

GT: Heart size and pulmonary vascularity appear within normal limits. The lungs are free of focal airspace disease. No pleural effusion or pneumothorax is seen. Levoscoliosis of the thoracolumbar spine is present. No evidence of active disease.

Severity Score: 0.22

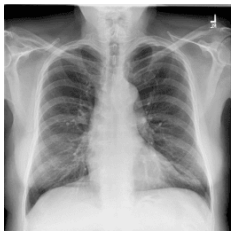
Figure 5: Low severity examples on the IU X-Ray dataset.Best viewed in color.



MediVLM: The heart appears mildly enlarged, and lung volumes are reduced. There are no signs of infection, lung collapse, or significant fluid buildup around the lungs. The bony structures and the area under the diaphragm show no abnormalities, and there is a mild amount of XXXX noted in the transverse colon. Overall, there are no acute issues with the heart or lungs.

GT: The heart is mildly enlarged. Lung volumes are low. There is no focal consolidation, pneumothorax, or large pleural effusion. Bony structures are within normal limits. There is no free air under the diaphragm. There is a mild amount of XXXX seen in the transverse colon. No acute cardiopulmonary findings.

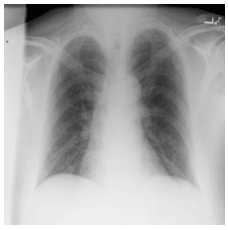
Severity Score: 0.69



MediVLM: The heart size and pulmonary vascularity are normal, and the lungs are clear with no signs of focal airspace disease. There is no pleural effusion or pneumothorax present. Degenerative changes are observed in the spine, and a stent is noted in the abdomen. The thoracic aorta is tortuous, and calcified granulomas are seen, but there is no evidence of active disease.

GT: The heart size and pulmonary vascularity appear within normal limits. The lungs are free of focal airspace disease. No pleural effusion or pneumothorax is seen. Degenerative changes are present in the spine. Stent is noted in the abdomen. The thoracic aorta is tortuous. Calcified granuloma are noted.1. No evidence of active disease.

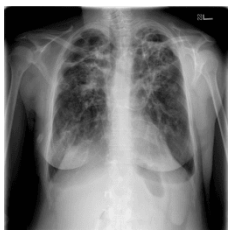
Severity Score: 0.80



MediVLM: The frontal view shows a kyphotic and rotated posture, with low lung volumes and bronchovascular crowding. There are no clear signs of infection, lung collapse, or fluid buildup around the lungs. Due to technical factors, the heart size appears borderline enlarged. There are no definite acute findings in the lungs or heart.

GT: Frontal view kyphotic and rotated, low lung volumes with bronchovascular crowding. Otherwise, no definite airspace consolidation or pleural effusion. Accounting for technical factors heart size XXXX borderline enlarged.Low lung volumes, otherwise, no definite acute findings.

Severity Score: 0.81

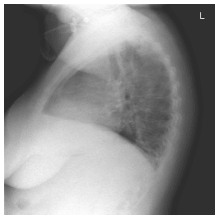


MediVLM: A left-sided medication injection device is in place, with its tip positioned at the cavoatrial junction. The trachea is midline, and the heart is of normal size. There is extensive bronchiectasis, cystic changes, and lung scarring, consistent with the patient's history of cystic fibrosis, without significant change from prior imaging. No signs of new infection, fluid buildup, large pneumothorax, or acute bony abnormalities are present.

GT: Left-sided medication injection XXXX has its tip projecting at the cavoatrial junction. The trachea is midline. Extensive bilateral bronchiectasis, cystic changes, and scarring represents sequela from the patient's cystic fibrosis. No evidence of focal pulmonary infiltrate or pleural effusion. No large pneumothorax has developed in the interim. The overlying bony structures reveal no acute abnormalities. The heart size is normal.1. Extensive pulmonary bronchiectasis and scarring from cystic fibrosis, not significantly XXXX from prior. 2. Left-sided medication injection XXXX has its tip projecting over the cavoatrial junction.

Severity Score: 0.74

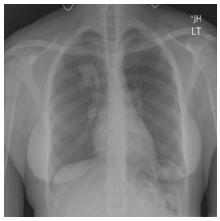
Figure 6: High severity examples on the IU X-Ray dataset.Best viewed in color.



MediVLM(Unsupervised): The lung volumes appear slightly reduced, affecting the overall visualization of structures. No significant fluid accumulation or other concerning findings are noted. The heart size appears slightly prominent, though this may be related to imaging technique. A tracheostomy tube is positioned in a stable location within the airway.

GT: Low lung volumes with bronchovascular crowding. Otherwise, no focal alveolar consolidation, no definite pleural effusion seen. No typical findings of pulmonary edema. Heart size borderline enlarged for technique, tracheostomy tube tip approximately 4.7 cm above the carina.

Severity Score: 0.46



MediVLM(Unsupervised): There is a region of structural change in the right upper lung with the rest of the lungs clear. No unusual fluid accumulation or air pockets are seen. The heart and central structures maintain a normal appearance. Overall, no other significant abnormalities are detected in the evaluation.

GT: Right upper lobe airspace disease consistent with pneumonia given patient's history. The lungs are otherwise clear. No pleural effusions or pneumothoraces. Heart and mediastinum of normal size and contour.Right upper lobe pneumonia.

Severity Score: 0.61

Figure 7: High severity examples on the IU X-Ray dataset. Report generated using **unsupervised training of MediVLM** (only images, no reports available for training). Best viewed in color.