LARGE MULTIMODAL MODEL FOR REAL-WORLD RADIOLOGY REPORT GENERATION

Brian Nlong Zhao[1]*, **Xinyang Jiang**[2], **Xufang Luo**[2], **Zilong Wang**[2], **Yifan Yang**[2], **Bo Li**[3], **Javier Alvarez-Valle**[4], **Matthew P. Lungren**[4], **Dongsheng Li**[2], **Lili Qiu**[2]
[1]University of Southern California, [2]Microsoft Research Asia,
[3]Nanyang Technological University, [4]Microsoft Health Futures,
briannlz@usc.edu, libo0013@ntu.edu.sg
{xinyangjiang,xufang.luo,yifanyang,wangzilong}@microsoft.com
{jaalvare,mlungren,dongsheng.li,liliqiu}@microsoft.com

ABSTRACT

While automatic report generation has demonstrated promising results using deep learning-based methods, deploying these algorithms in real-world scenarios remains challenging. Compared to conventional report generation, real-world report generation requires model to follow the instruction from the radiologists and consider contextual information. Thus, this paper focuses on developing a practical report generation method that supports real-world clinical practice. To tackle the challenges posed by the limited availability of clinical data, we propose a GPT-based unified data generation pipeline designed to produce high-quality data. Consequently, we present a new benchmark dataset *MIMIC-R3G*, comprising five representative tasks pertinent to real-world medical report generation. We propose Domain-enhanced Multi-modal Model (*DeMMo*), where an additional medical domain vision encoder is incorporated into the general domain multi-modal LLM to enhance its ability on specific domains. This approach aims to harness the specialized capabilities of the medical domain vision encoder while leveraging the robustness and versatility of the general domain multi-modal LLM. Comprehensive experiments demonstrate that our approach attains competitive performance across all real-world tasks compared to existing interactive report generation frameworks and state-of-the-art encoder-decoder style report generation models.

1 INTRODUCTION

Radiology report generation is one of the straightforward yet essential task in computer-aided diagnosis (CAD) systems. It aims to automatically generate a text description of the patient's radiology images including professional medical diagnosis. Recent works can automatically generate radiology report accurately within seconds, which largely reduces the workload of professional radiologists in clinical routines (Jing et al., 2018; Chen et al., 2020; Liu et al., 2021; Wang et al., 2022a; Huang et al., 2023).

Most previous works treat radiology report generation as a captioning task, where a text decoder generate medical report based on extracted image features (Nicolson et al., 2023). In real clinical practice, however, the scenario and procedure might be more complex than a straightforward captioning task. Specifically, in real-world scenarios, the model is required to follow broader instructions of the radiologists and to consider different types of context information. For example, radiologists usually need to refer to the patient's X-ray images and reports from previous visits in order to write a more comprehensive report that includes progress or changes in the abnormalities. Also in many cases, patients are required to undergo some other medical examinations beside radiology screenings. All these kinds of extra information could affect how radiologists read the radiographs and write the final report for the patient. Therefore, this paper focuses on developing

*Work done during internship at Microsoft Research Asia

a practical report generation method that supports real-world clinical practice containing various interactions and external context information.

Large Language Models (LLMs) pose significant potential in performing real-world report generation tasks, owning to their capabilities in interaction, instruction following, medical domain knowledge and long sequence generation. However, to develop a LLM based real-world report generation system, there are two challenges to overcome. The first challenge is the scarcity of clinical data that comprises different scenarios, which could hinder the development of robust LLMs. Current medical report generation datasets are predominantly obtained from hospital or clinical databases. The information available in these datasets is generally limited to medical images and associated structured reports (Johnson et al., 2019; Demner-Fushman et al., 2016), lacking supplementary information that might influence radiologist's reasoning in formulating a diagnosis. The second challenge involves fine-tuning LLMs, which typically possess billions of parameters to optimize, requiring considerable computational resources. Specifically for medical domain, the training process can be laborious, as it is vulnerable to overfitting, particularly when dealing with a relatively small amount of data.

To address the challenges in data scarcity, we examine the real-world clinical requirements and propose a new benchmark dataset, named *MIMIC-R3G* (Real-world Radiology Report Generation). *MIMIC-R3G* contains five representative tasks pertinent to the medical report generation context: report generation with no context, report revision, template-based report generation, report generation based on patient's previous visits, and report generation incorporating patient's other information including medical records and laboratory tests. Building on these tasks, we introduce a unified automatic data generation pipeline to generate instructions, context, and reports in accordance with the ground truth report and images, using specific system messages and ground truth reports as input to direct ChatGPT (OpenAI, 2022) for generation.

To address the challenge of fine-tuning LLMs for real-world report generation, we propose Domain-enhanced Multimodal Model (*DeMMo*) with pathological guidance, where a medical domain vision encoder is incorporated into Flamingo (Alayrac et al., 2022), a pretrained general domain large multimodal model. *DeMMo* effectively enhances the domain-specific capabilities of the pretrained LLM while retaining its general medical domain knowledge. Under the guidance of pathological information, *DeMMo* is able to zoom in on specific regions of interest in a medical image to focus on detailed characteristics of particular lesions or structures. Comprehensive experiments on the *MIMIC-R3G* benchmark demonstrate that our method achieves promising results on all real-world report generation tasks, compared to existing interactive report generation framework and state-of-the-art encoder-decoder style report generation models.

In summary, the contributions of this paper are as follows:

- We present a new problem setting for real-world report generation that emulates clinical practices by incorporating various clinical interactions and contextual information.
- We propose the first real-world report generation benchmark dataset *MIMIC-R3G*, where a unified framework designed to automatically generate the requisite context data, leveraging the power of LLM.
- We develop *DeMMo* , a large multimodal model with domain-specific capability enhanced via incorporating a general domain Flamingo with an additional medical vision encoder and pathological information for further guidance.

2 Related Works

Report Generation Traditional methods use an encoder-decoder regime, where an encoder is used to extract image features, and a decoder is used to generate text from the features. The combination of CNN encoder and RNN decoder were utilized in earlier works (Jing et al., 2018; Xue et al., 2018; Wang et al., 2018; Hou et al., 2021). With the advent of Transformer architecture, researchers have explored the use of Transformer with specialized memory or attention mechanisms for report generation (Cornia et al., 2020; Chen et al., 2020; 2021; You et al., 2021). To further improve performance, many works incorporated pre-extracted pathology labels and domain-specific knowledge graphs as priors in the generation pipeline (Liu et al., 2021; Wang et al., 2022b; Huang et al., 2023; Li et al., 2023d). Some retrieval-based approaches have also gained prominence in recent years (Endo et al.,

2021; Jeong et al., 2023). These methods predominantly employ contrastive learning techniques to retrieve probable texts from the training set as inference outcome. Building on existing approaches, several studies (Wu et al., 2022; Zhu et al., 2023) have also taken real-world clinical scenarios into account, but primarily focusing on the single task of incorporating reports from previous visits as a generation prior. We expand on this and propose a unified task formulation of real-world report generation.

Large Language Models With the strong ability in natural language processing and generation, Large Language Models (LLMs) have shown significant potentials in performing real-world report generation tasks. State-of-the-art LLMs (Brown et al., 2020; Touvron et al., 2023; Chowdhery et al., 2022) are highly interactive and capable of following instructions for various language tasks (Ouyang et al., 2022), making it poses high potential in dealing with real-world clinical scenarios. Furthermore, the extensive volume of training data equips LLMs with the capacity to internalize domain-specific knowledge and exhibit reasoning capabilities within the medical field. Without fine-tuning on specific medical dataset, ChatGPT (OpenAI, 2022) is tested to pass the US Medical Licensing Exams (USMLE) (Kung et al., 2023) showing its promising ability to reason and process language in the medical domain. Finally, LLMs demonstrate proficiency in generating more extensive and complex text sequences, making them well-suited for medical report generation tasks.

Multimodal LLMs With the remarkable success of LLMs, researchers started to explore the possibility to integrate visual modality into LLMs for various visual-language tasks. Early works such as BLIP-2 (Li et al., 2023c) leveraged a query Transformer to connect visual features to LLM. Flamingo (Alayrac et al., 2022) introduced extra trainable layers within LLM in addition to a Transformer to bridge visual and language modalities. LLaVa (Liu et al., 2023) and MIMIC-IT (Li et al., 2023a) leveraged GPT/ChatGPT to build visual instruction tuning datasets and developed multimodal LLMs as general instruction-follow visual agents. Following their ideas, we construct a real-world report generation dataset by building a unified data generation pipeline leveraging ChatGPT.

Medical LLMs Numerous works have applied LLM within the medical domain through fintuning a general domain LLM. Med-PaLM (Singhal et al., 2022) and Med-PaLM 2 (Singhal et al., 2023) are medical domain-specific language models developed through instruction fine-tuning based on general domain LLMs. Med-PaLM M (Tu et al., 2023) further fine-tunes PaLM-E to the medical domain using multimodal medical data for medical vision-language tasks. LLaVa-Med (Li et al., 2023b) and Med-Flamingo (Moor et al., 2023) similarly fine-tune their general domain base models using domain-specific data to enhance medical question-answering and conversational capabilities. Different from fine-tuning LLMs for medical domain, ChatCAD (Wang et al., 2023) and ChatCAD+ (Zhao et al., 2023) interact with users by connecting medical domain models with ChatGPT via language prompts. We observe that this framework is capable of doing all of our proposed real-world report generation tasks without training using extra task-specific data.

3 R3G: Real-world Radiology Report Generation

In contrast to conventional report generation models, Real-world Radiology Report Generation (R3G) poses two significant differences. Firstly, it necessitates the model to adhere to the user's requests and instructions. Secondly, in addition to the medical image itself, the model must possess the capability to comprehend and utilize external contextual information in order to produce a more precise report. As a results, we propose several representative sub-tasks that resembles these two requirements, all of which are essential features widely applicable in clinical practice. The instances drawn from these representative sub-tasks will be used to train and evaluate our proposed report generation model.

Report generation This sub-task is the conventional report generation task without any additional instructions from radiologist or context information.

Report revision Reports generated models may be sub-optimal in some cases, and and human professionals are still required to review and revise the output reports prior to submission. Therefore, it is desirable for the model to possess the capability of revising the report based on straightforward instructions to further alleviate the workload of the human professional.

Template In real-world scenarios, clinics or hospitals may employ structured report templates. These templates may comprise a list of common abnormalities or regions, and the radiologist is

required to fill in the corresponding findings or absence of abnormalities. In sum, we want the model to be capable of generating report following any form of input template.

Previous Radiology Image and Report as Context In typical clinical practice, patients undergo multiple radiology screenings. It is essential for radiologists to write medical reports that not only focus on the current radiology image but also reference the patient's previous medical images and reports. This approach enables the production of a more informative report that can address the alterations in the disease progression compared to previous visits.

Medical Records and Lab Tests as Context Patient's medical records, including medical condition history, along with medical exams like blood tests and pulmonary function tests, are vital for accurate diagnosis. Medical records and lab tests are all crucial context information for radiologists to write reports, so the model should also posses the ability generate reports based on them.

4 *MIMIC-R3G*: DATASET FOR REAL-WORLD REPORT GENERATION

4.1 TASK FORMULATION

We formulate the proposed real-world report generation tasks under a unified instruction-following paradigm, so we can fully utilize the instruction-following capabilities of a Large Language Model (LLM). Specifically, we format the proposed real-world report generation tasks into a unified single-round instruction-following example: (V_i, I_i, C_i, R'_i), representing the i-th example in the dataset, where V_i denotes a set of medical images; I denotes the instruction from the user; C_i refers to the context information provided to facilitate the report generation; and R'_i refers to the ground truth report associated with the medical images V_i, instruction I_i, and context C_i in the generated dataset. For all the sub-tasks, V_i is directly utilized from the dataset.

4.2 DATA GENERATION

Existing large-scale report generation datasets, such as MIMIC-CXR (Johnson et al., 2019), are not tailored for real-world report generation as they lack user instructions I_i and contextual information C_i paired with corresponding responsive report R'_i . The manual collection of such instructional and contextual data is prohibitively costly and may raise privacy concerns. Hence, we propose to harness the capabilities of ChatGPT and construct a unified pipeline to automatically generate diverse and relevant real-world clinical text data based on existing ground truth reports in conventional datasets.

The primary goal is to either design or generate instructions I_i and context C_i, and also possibly modify the ground truth report R_i from dataset into R'_i according to different sub-tasks. To generate a dataset of a single real-world report generation task, the objection of our pipeline is $\{(V_i, R_i)\}_{i=1}^{N} \mapsto \{(V_i, I_i, C_i, R'_i)\}_{i=1}^{N}$, where N is the number of examples of an existing report generation dataset.

The medical image V_i stays un-changed and directly comes from the original dataset. We devise different task-specific system messages to generate the required I_i, C_i, and R'_i for distinct tasks. Using the ground truth report R_i as input, along with in-context examples (omitted in examples) to guide the output format, the response can be filtered and parsed accordingly into the required data components. Next we will elaborate on how request from each sub-task is organized as an instruction-following example, and how the examples are produced for each sub-task. We show one data generation example, and please refer to supplementary for examples of other tasks.

Report Generation For basic report generation task, the data sample follows (V_i, I_i, C_i, R'_i), where V_i and $R'_i = R_i$ are directly utilized from report generation dataset. I_i is a manually designed instruction, and C_i is kept empty.

Report revision For report revision task, $R'_i = R_i$ come from the report generation dataset, I_i is the instruction of how to revise or correct the report, and C_i is the report that the user wants the model to revise. To generate I_i and C_i for this task, we employ our proposed pipeline to produce a slightly modified report based on the input ground truth report, along with the instructions of how to revise the modified report into the correct ground truth report. We show an example of system message used and ChatGPT response of report revision task here.

Template I_i is a manually designed instruction, *e.g.*, *Fill in the template based on the give medical images*. I_i and R_i' are the empty template and the corresponding filled template. We collect 10 report templates with help of medical professionals, and we leverage our pipeline to generate the structured version of the ground truth report based on a given template.

Previous Visit as Context I_i is manually designed instruction telling the model to generate report based on both the medical images and report from last visit. C_i can be the retrieved previous report of the same patient from the dataset, and $R_i' = R_i$ is the ground truth report. In the case when previous report is unavailable, we can select a random report from the dataset as the context C_i and employ the proposed pipeline to generate a modified report R_i' from both the current report and C_i as a pseudo previous report. The modified report R_i' should have the diagnosis unchanged compared to R_i but with more descriptions on comparisons between two reports. It should be noted that V_i in this task can include medical images of the patient from their previous visit as well.

Medical Records and Lab Tests as Context. Similarly, $R_i' = R_i$ comes from the original dataset, and I_i is a manually designed instructions. C_i here represents the additional medical conditions or medical examination results that the patient may posses. Since the ground-truth medical report exhibits a strong correlation with the external context information, our pipeline generates C_i by extracting medical indications and inferring the plausible medical conditions, medical examinations and exam results based on the ground truth medical report.

Example of Generating Report Revision Data

System Message: You will make some minor diagnosis error when reading a chest x-ray radiograph. You will be given a correct chest x-ray report with a findings section, and you will be asked to rewrite the report with a few diagnostic errors. Then provide the instructions of how to correct your wrong diagnosis.
Input Report: PA and lateral views of the chest provided. The lungs are adequately aerated. There is a focal consolidation at the left lung base adjacent to the lateral hemidiaphragm. There is mild vascular engorgement. There is bilateral apical pleural thickening. The cardio mediastinal silhouette is remarkable for aortic arch calcifications. The heart is top normal in size.

Response: *Report:* PA and lateral views of the chest provided. The lungs are adequately aerated. There is no focal consolidation. There is mild vascular engorgement. There is no bilateral apical pleural thickening. The cardio mediastinal silhouette is remarkable for aortic arch calcifications. The heart is mildly enlarged in size.
Instructions: Add focal consolidation. The patient has bilateral apical pleural thickening.

$I_i = \langle$Text in Instructions section$\rangle$
$C_i = \langle$Text in Report section$\rangle$

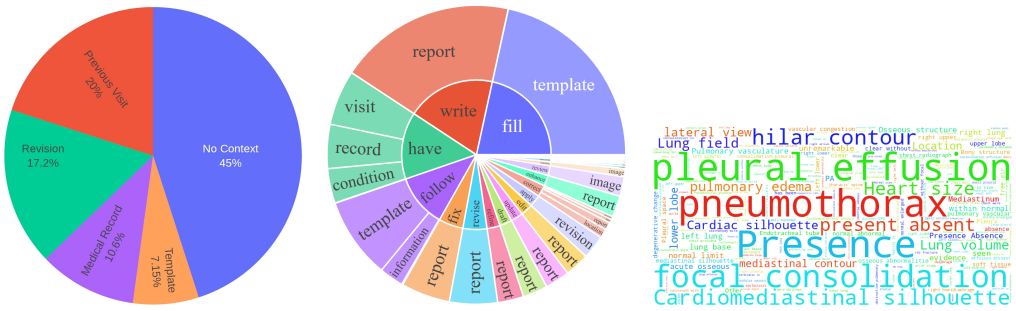
Data Quality Control Since the ground truth report R_i' is either identical to original report R_i or rewritten by ChatGPT while preserving the medical diagnosis intact, our pipeline is able to produce accurate data with very few factual errors. Furthermore, the generated data has undergone both automatic and manual data quality exam and control processes. Specifically, after generating the data, ChatGPT will be prompted again with the generated data as input and is required to check correctness. For report revision task, ChatGPT checks the correctness of revision instructions. For template task, it checks whether the generated ground truth follow the diagnosis of original ground truth and the format of the given template. For medical records and tests task, it checks whether the generated context is diagnostically consistent with the ground truth report. If there exists any incorrectness or inconsistency in generated data, our pipeline will try to regenerate or skip to the next sample. Unsatisfactory generated reports are further filtered by comparing the labels of generated context and ground truth report. We use CheXpert labeler (Irvin et al., 2019), an automatic tool to extract labels of common observation from radiology reports, to extract and compare the labels of C_i and R_i' to ensure that no information leakage is presented in the generated context. For manual examination, a medical professional validates the clinical correctness of a subset of the generated data.

4.3 Dataset Statistics and Analysis

Using our data generation pipeline, we generate a novel dataset based on a large report generation dataset MIMIC-CXR (Johnson et al., 2019), named *MIMIC-R3G*. Since MIMIC-CXR already contains patients' previous reports, we directly use the report from dataset as ground truth and retrieved previous report as context without generation. The statistics of original MIMIC-CXR data and generated data for all tasks is shown in table 1.

	No Context	Revision	Template	Previous Report	Medical Record	Total
#Train	140,781	53,649	22,348	62,576	33,265	312,619
#Test	2,020	2,020	1,703	1,253	1,969	8,965

Table 1: Statistics of *MIMIC-R3G*



(a) Percentage of tasks (b) Frequent verb-noun pairs. (c) Word cloud of context.

Figure 1: Visualizations of *MIMIC-R3G* statistics. (a) shows the general distribution of data of different tasks. (b) shows the frequent verb-noun pairs appeared in instructions of our dataset, where inner circle is the root verb, and outer ring represents the paired nouns.

No Context

Instruction: Act as a radiologist and write a diagnostic radiology report for the patient based on their chest radiographs
Report: The patient is status post recent median sternotomy and coronary artery bypass surgery. Cardiomediastinal contours are stable in appearance compared to previous post-operative radiographs. Moderate left pleural effusion has slightly increased in size since the previous radiograph, and is associated with adjacent left basilar atelectasis. On the other hand, a small right pleural effusion has decreased in size and nearly resolved.

Report Revision

Instruction: Update the medical report of the given chest x-ray images with these changes: replace "significantly increased" with "slightly increased"
Context: Report: The patient ... Moderate left pleural effusion has *significantly increased* in size since the previous radiograph, and is associated with adjacent left basilar atelectasis ... still present.
Report: (same as no context)

Medical Record as Context

Instruction: Based on the chest x-ray images and patient's medical details, draft a detailed diagnostic medical report
Context: Medical conditions of the patient: Echocardiogram shows decreased cardiac output or decreased ejection fraction, pulmonary function tests show decreased lung capacity or decreased oxygen saturation
Report: (same as no context)

Template

Instructions: Please act as a radiologist and write a diagnostic radiology report for the patient based on their chest radiographs, the format should follow the template
Context: Template:
Cardiomediastinal contours: [stable/unstable]
...
bony structures: [normal/abnormal]
Report:
Cardiomediastinal contours: stable.
...
Bony structures: not mentioned.

Previous Visit as Context

Instruction: Please write a diagnostic radiology report for the patient based on their chest radiographs considering the report from last visit
Context: Medical report from last visit: Following removal of left-sided chest tube, there is a probable residual tiny left apical pneumothorax. Otherwise, no short interval change in the appearance of the chest since the recent study performed earlier the same date.
Report: (same as no context)

Table 2: Examples of *MIMIC-R3G* generated using the same report on different tasks

A subset of the generated dataset has undergone validation by a medical professional. Specifically, approximately twenty samples from each sub-task were randomly selected, and a doctor evaluated the quality of each sample to ensure that the generated content are suitable with its corresponding ground-truth report. In summary, our medical professionals have determined that no significant factual discrepancies exist between the content generated by GPT and the ground-truth reports across all samples, and no information in the generated content is present in the ground truth report. This validation affirms the capability of our pipeline to generate clinically correct content and filter out generated contents with any factual errors or data leakage.

5 METHODOLOGY

For real-world report generation, our objective is to train a model that given the image-text input $x = (V, I, C)$ generate output text $y = R'$, therefore the generation process can be formalized as $p_\theta(y \mid x)$ where θ represents the model parameters to be optimized. Our model is built upon Flamingo (Alayrac et al., 2022) due to its training efficiency and good performance. To further enhance the domain specific capability of the general domain flamingo model, *DeMMo* inserts an additional domain specific medical encoder to the perceiver resampler of flamingo. A parameter efficient prompt tuning method is adopted to fine-tune the model for medical domain while preserving its generalization ability.

5.1 FLAMINGO AS MODEL BASIS

Flamingo is a family of vision language model that is capable of generating language conditioned on interleaved text and image sequences. By connecting visual encoder and LLM with a perceiver resampler, Flamingo models the likelihood of text output y conditioned on interleaved image and text input as a next-token prediction task: $p_\theta(y \mid x) = \prod_{l=1}^{L} p(y_l \mid y_{<l}, x_{<l})$, where $y_{<l}$ and $x_{<l}$ are the sets of text and image tokens preceding y_l, the l-th input text token.

The general domain visual encoder of Flamingo exhibit greater diversity and generalization ability, but cannot fully capture the detailed visual feature in medical domain. Consequently, a domain specific encoder is required to capture the nuances and specific characteristics of medical images. In this paper, we employ BioViL (Boecking et al., 2022) as our medical vision encoder. To capitalize on the robust generalizability and expedite convergence, the original pretrained general domain visual encoder in Flamingo is still preserved in conjunction with the newly introduced medical encoder.

5.2 MODEL ARCHITECTURE AND FINE-TUNING APPROACH

We propose Domain enhanced Mulimodal Model (*DeMMo*), which incorporates a domain specific visual encoder into the general domain Flamingo.

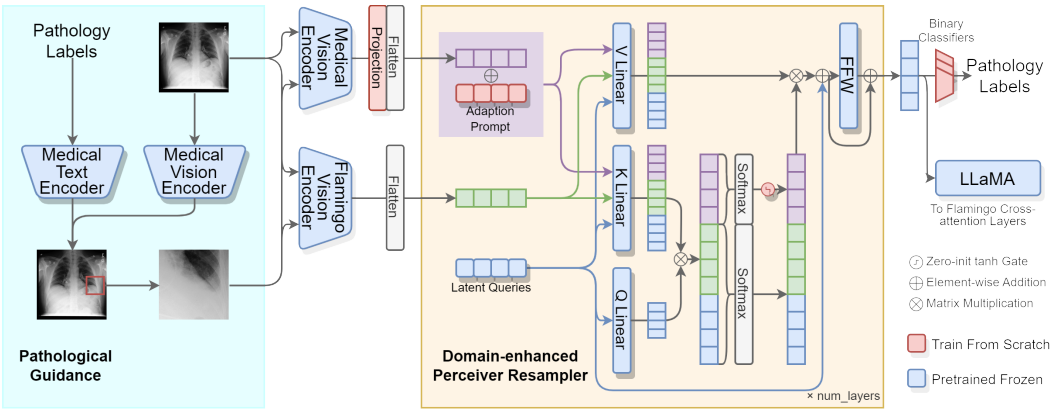


Figure 2: Architecture of *DeMMo*

As shown in figure 2, given a set of images V_i that contains k images, the original Flamingo vision encoder outputs $n \times n$ grid features $\boldsymbol{X}_f \in \mathbb{R}^{k \times n \times n \times d_f}$, and the medical vision encoder outputs an $m \times m$ grid features $\boldsymbol{X}_m \in \mathbb{R}^{k \times m \times m \times d_m}$, where d_f and d_m are feature dimensions of Flamingo vision encoder and medical vision encoder, respectively. After applying a projection $\boldsymbol{W} \in \mathbb{R}^{d_m \times d_f}$ to $\boldsymbol{X}_m$ followed by flattening both grid features, we get $\boldsymbol{X}_f \in \mathbb{R}^{kn^2 \times d_f}$ and $\boldsymbol{X}_m \in \mathbb{R}^{km^2 \times d_f}$. We adopt the idea of LLaMA-Adapter (Zhang et al., 2023) to insert a learnable adaption prompt $\boldsymbol{P}_l \in \mathbb{R}^{m^2 \times d_f}$ into the perceiver resampler independently for each layer l. Each flattened feature from medical vision encoder is then added element-wise to $\boldsymbol{P}_l$ to form the medical visual feature prepared for attention. Similar to vanilla Flamingo, a predefined number of latent queries are cross-attended to the concatenation of queries and visual features. Formally, denote t as the number of

latent queries. At layer l, $\boldsymbol{Q}_l \in \mathbb{R}^{t \times d_f}$ is the latent queries, and $\boldsymbol{V}_l = \boldsymbol{K}_l \in \mathbb{R}^{\left(km^2+kn^2+t\right) \times d_f}$ is the concatenation of medical visual features, original visual features from Flamingo vision encoder, and the latent queries. Then, the similarity scores are computed as $S_l = \left(\boldsymbol{Q}_l \boldsymbol{W}_l^Q\right)\left(\boldsymbol{K}_l \boldsymbol{W}_l^K\right)^\top / \sqrt{d_h} \in \mathbb{R}^{t \times \left(km^2+kn^2+t\right)}$, where $\boldsymbol{W}_l^Q, \boldsymbol{W}_l^K \in \mathbb{R}^{d_f \times d_h}$ are query and key projections respectively at layer l, and d_h represents the hidden feature dimension.

Since the linear projection and the adaption prompt is randomly initialized, training with newly introduced medical visual features might introduce instability on top of the pretrained weights. Therefore, we follow (Zhang et al., 2023) to independently apply softmax on the similarity score corresponding to the medical visual features. A zero-initialized $\tanh$ gate is utilized on the corresponding output attention score, ensuring a gradual increase of influence due to medical visual features.

Pathological Guidance To provide further medical domain-specific guidance, we leverage the phrase grounding ability of the medical vision encoder and incorporate pathology label guidance in our training and inference pipeline. Specifically, given a chest medical image and a pathology phrase, BioViL (Boecking et al., 2022) is able to output a heatmap on the image associated with the phrase. In our training phase, we apply the CheXpert labels extracted from the ground truth report to find maxima on the heatmap and crop a zoomed in region of interest for each image. We proceed by concatenating the zoomed in regions of interest with the original images as the input fed into the perceiver resampler. Additionally, to enable this guidance during inference when ground truth labels are not available, we augment the perceiver resampler with binary classifiers for each pathology category, which imposes additional constraints to ensure that the latent query output of the perceiver contains pathology categorization information. During inference, the medical image is initially passed through the perceiver resampler only to obtain the corresponding pathology label. Subsequently, this label is utilized to extract zoomed-in region of interest from the original image. Finally, the extracted region along with the full medical images undergoes a full forward pass through the entire pipeline.

The rest are same as vanilla Flamingo model, where the attended queries pass through another feed-forward network before next layer, and the last perceiver layer output is inserted into Flamingo Cross-attention layers. We only tune the medical vision encoder projection, adaption prompts, zero-initialized gates, and the binary disease classifiers, along with the Flamingo cross-attention layer in LLaMA.

6 EXPERIMENTS

6.1 DATASET AND METRICS

Following (Chen et al., 2020; 2021; Wang et al., 2022a; Nicolson et al., 2023), we use the frontal view images in *MIMIC-R3G* and focus on generating the findings. This results in 140,781 reports and 156,969 images in training set and 2020 reports and 2274 images in test set. We adopt natural language generation (NLG) metrics that measures text similarity between generated and ground truth report, including BLEU (B@n) (Papineni et al., 2002), METEOR (M) (Banerjee & Lavie, 2005), and ROUGE-L (R-L) (Lin, 2004). Following previous works, we also utilize CheXpert, an automatic labeler tool to extract observation labels from chest X-ray reports, to evaluate clinical efficacy (CE) in terms of micro-averaged label precision (P), recall (R), and F1-score (F1).

6.2 BASELINES

ChatCAD+ (Zhao et al., 2023) is an interactive report generation framework that connects medical image disease classifier and report generator with ChatGPT and online knowledge databases.

ChatCAD+ with replaced report generator ChatCAD+ utilize an R2Gen model trained on MIMIC-CXR dataset as the report generator backbone. We replace this report generator by our proposed model trained solely on MIMIC-CXR data as well. As this baseline employs the same report generator as *DeMMo*, it offers a more equitable comparison to illustrate the effectiveness of training an end-to-end multi-modal LLM using task-specific data generation, rather than linking it to an LLM via text prompts.

CvT-212DistilGPT2 We adapted CvT-212DistilGPT2 (Nicolson et al., 2023), a conventional report generation model to consider context and instructions. We then trained it on *MIMIC-R3G* to evaluate its real-world report generation performance.

Task	Method	B@1	B@2	B@3	B@4	M	R-L	P	R	F1
No Context	ChatCAD+	0.307	0.160	0.088	0.052	0.266	0.189	0.335	**0.613**	0.433
	ChatCAD+*	0.346	0.187	0.106	0.063	0.268	0.200	0.376	0.564	0.451
	Enc-Dec	0.338	0.205	0.133	0.095	0.277	0.236	0.350	0.190	0.246
	Ours	**0.375**	**0.227**	**0.146**	**0.103**	**0.296**	**0.242**	**0.500**	0.461	**0.480**
Revision	ChatCAD+	0.639	0.571	0.521	0.479	0.719	0.655	0.860	**0.866**	0.863
	ChatCAD+*	0.652	0.583	0.531	0.488	0.719	0.659	0.869	0.853	0.861
	Enc-Dec	0.389	0.241	0.159	0.114	0.304	0.253	0.506	0.368	0.426
	Ours	**0.784**	**0.686**	**0.641**	**0.630**	**0.740**	**0.726**	**0.894**	0.837	**0.865**
Template	ChatCAD+	0.506	0.433	0.381	0.340	0.443	0.409	0.553	**0.572**	0.562
	ChatCAD+*	0.496	0.427	0.377	0.338	0.443	0.407	0.556	0.571	0.564
	Enc-Dec	0.165	0.086	0.047	0.029	0.15	0.155	0.536	0.266	0.356
	Ours	**0.534**	**0.461**	**0.409**	**0.367**	**0.533**	**0.483**	**0.684**	0.564	**0.618**
Previous Report	ChatCAD+	0.310	0.168	0.100	0.063	0.290	0.199	0.511	0.523	0.516
	ChatCAD+*	0.308	0.168	0.099	0.063	0.292	0.200	0.509	**0.526**	**0.517**
	Enc-Dec	0.377	**0.237**	**0.158**	**0.115**	**0.301**	**0.255**	0.437	0.281	0.342
	Ours	**0.383**	0.231	0.147	0.098	0.287	0.242	**0.511**	0.493	0.502
Medical Record	ChatCAD+	0.179	0.090	0.050	0.031	0.227	0.123	0.456	**0.588**	0.513
	ChatCAD+*	0.180	0.092	0.051	0.031	0.238	0.130	0.468	0.563	0.511
	Enc-Dec	0.376	0.234	0.155	0.112	0.296	0.251	0.463	0.305	0.367
	Ours	**0.377**	**0.254**	**0.183**	**0.142**	**0.335**	**0.292**	**0.580**	0.468	**0.518**

Table 3: Comparison of our model with baselines on the test sets of our real-world report generation dataset. ChatCAD+* refers to the ChatCAD+ model using our model trained on no-context data (i.e. original MIMIC-CXR data) as report generator. Enc-Dec refers to the CvT-212DistilGPT2 encoder-decoder model. B@n, M, R-L represent the NLG metrics BLEU, METEOR, and ROUGE-L respectively. P, R, F1 represent the CE metrics CheXpert precision, recall, and F1-score respectively.

6.3 Main Result

The models are trained on *MIMIC-R3G* for 8 epochs with 4 batch size in all experiments. We use an ADAMW optimizer with $\beta_1 = 0.9, \beta_2 = 0.999$ and weight decay of 0.01. The learning rate is set to 1e-4 with a 1000-step warm-up and a cosine decay schedule. More detailed implementation of the experiments are elaborated in supplementary.

As shown in Table 3, we compare *DeMMo* with aforementioned baselines on each of the *MIMIC-R3G* sub-tasks, respectively. Note that the test splits of each sub-task are not identical and hence the performance is not comparable across difference tasks. In terms of most NLG metrics, *DeMMo* outperforms the baselines by a large margin on all tasks. In terms of CE metrics, *DeMMo* achieves the highest F1 score on 4 sub-tasks. While generally achieve low performance in terms of NLG metrics, methods using ChatCAD+ achieves the highest CE metrics on 'Previous Report' task. This is because ChatCAD+ employs an extra pretrained image disease classifier trained with additional annotations that gives more accurate prior information to the ChatGPT.

6.4 Conventional Report Generation

To show the efficacy of our model architecture design, we also evaluated the performance of *DeMMo* on conventional report generation task. Specifically, we train *DeMMo* using the original MIMIC-CXR dataset to compare with other conventional report generation models under the same setting. For a fair comparison, only generation methods that do not use extra medical dataset, knowledge graphs, or disease label or image classifier are compared. The performance of the comparison methods are directly cited from papers. As shown in Table 4, our methods significantly outperform existing conventional report generation methods in terms of CE metrics and a comparable performance in terms of NLG metrics.

Inference Examples of Generation

Ground Truth Report: (s55575670)
the patient has a history of chronic interstitial lung disease with waxing and waning pulmonary edema and infection. today it is largely unchanged with diffuse infiltrative and interstitial opacities stable since ___. bilateral pleural effusion is essentially the same. cardiomediastinal silhouette is stable and demonstrates mild cardiomegaly. there is no pneumothorax. enteric tube is seen once again, entering the stomach and then out of field of view. right-sided picc terminates within the mid svc. an endotracheal tube terminates no less than 6 cm from the carina.

Instruction:
Act as a radiologist and please write a chest x-ray report for the patient based on the chest x-ray images:

ChatCAD+
upon reviewing the patient's chest x-ray, it has been observed that overlying ekg leads are present, and lung volumes are low, limiting assessment. there is diffuse pulmonary edema with probable small bilateral pleural effusions, and it is difficult to exclude a superimposed pneumonia. the heart size cannot be assessed due to the presence of these conditions, and the mediastinal contour appears grossly unremarkable. the bony structures are intact. while there is no sign of cardiomegaly, further evaluation is necessary to investigate potential consolidation and atelectasis, and to monitor the patient's pulmonary edema and pleural effusion.

CvT212DistilGPT2
as compared to the previous radiograph there is no relevant change. the monitoring and support devices are constant. the bilateral parenchymal opacities are constant in extent and severity. unchanged appearance of the cardiac silhouette.

***DeMMo* (Ours):**
in comparison with the study of ___, the monitoring and support devices remain in place. continued enlargement of the cardiac silhouette with pulmonary edema. bibasilar opacifications persist, consistent with a combination of pleural effusion and atelectasis. The patient has no pneumothorax. in the appropriate clinical setting, supervening pneumonia would have to be considered.

Method	B@1	B@2	B@3	B@4	M	P	R	F1
R2Gen (Chen et al., 2020)	0.353	0.218	0.145	0.103	0.142	0.333	0.273	0.276
CMN (Chen et al., 2021)	0.353	0.218	0.148	0.106	0.142	0.334	0.275	0.278
XPRONET (Wang et al., 2022a)	0.344	0.215	0.146	0.105	0.138	-	-	-
CvT-212DistilGPT2 (Nicolson et al., 2023)	**0.395**	**0.249**	**0.172**	**0.127**	0.155	0.365	0.418	0.390
DeMMo (Ours)	0.384	0.231	0.154	0.113	**0.296**	**0.500**	**0.443**	**0.470**

Table 4: Comparison of *DeMMo* with conventional report generation methods. The highest and the second highest performance are highlighted in bold and underline respectively.

6.5 Ablation Study

We conduct ablation experiments to compare the performance of three other model designs. Table 5 reports the performance comparison. Under the same setting mentioned in section 6.1, we train and compare the performance of three other model designs. *DeMMo* outperforms the vanilla Flamingo without using medical vision encoders, showing the importance of adopting a medical vision encoder to enhance the domain-specific ability. The second baseline does not preserve the original Flamingo visual encoder like *DeMMo*, instead it directly replaces it with a medical vision encoder. The comparison results verify that preserving the original visual encoder can retain its general domain knowledge and hence help the performance. The third baseline trains the architecture design with both original Flamingo vision encoder and the medical vision encoder, but without any pathological guidance. Compared to this baseline, *DeMMo* achieves generally higher performance, which highlights the efficacy of the design of pathological guidance in enhancing the model's capabilities for medical domain-specific tasks.

7 Conclusions

In this paper, we propose a highly interactive real-world radiology report generation problem setting (R3G). R3G requires models to be highly interactive, to follow instructions and consider various context information. A new benchmark dataset for the real-world report generation is built with a unified data generation pipeline. A novel Domain-enhanced Multi-Modal (*DeMMo*) model is proposed to enhance the medical domain specific ability of conventional LLM. Experiments demonstrate that *DeMMo* attains competitive performance across all real-world tasks.

Task	Metrics	w/o Medical Encoder	w/o General Encoder	w/o pathological guidance	*DeMMo*
No Context	BLEU@1	0.365	**0.376**	0.373	0.375
	Precision	0.438	0.487	0.491	**0.500**
	Recall	0.411	0.453	0.451	**0.461**
	F1 Score	0.424	0.469	0.470	**0.480**
Revision	BLEU@1	0.737	0.747	0.777	**0.784**
	Precision	0.884	0.894	0.845	**0.894**
	Recall	0.811	0.818	0.817	**0.837**
	F1 Score	0.847	0.854	0.831	**0.865**
Template	BLEU@1	0.470	0.429	0.529	**0.534**
	Precision	0.572	0.659	0.683	**0.684**
	Recall	0.440	0.489	0.530	**0.564**
	F1 Score	0.497	0.561	0.597	**0.618**
Previous Report	BLEU@1	0.356	0.357	0.370	**0.383**
	Precision	0.438	0.500	0.503	**0.511**
	Recall	0.366	0.421	0.436	**0.493**
	F1 Score	0.399	0.457	0.467	**0.502**
Medical Record	BLEU@1	0.374	**0.382**	0.381	0.377
	Precision	0.560	0.573	0.580	**0.580**
	Recall	0.464	0.437	0.446	**0.468**
	F1 score	0.508	0.496	0.504	**0.518**

Table 5: Ablation studies on the performance comparison of different components in *DeMMo*, including medical encoder, general Flamingo encoder, and pathological guidance.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pp. 1–21. Springer, 2022.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, November 2020.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 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)*, pp. 5904–5914, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.459. URL https://aclanthology.org/2021.acl-long.459.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10578–10587, 2020.

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

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

Benjamin Hou, Georgios Kaissis, Ronald M Summers, and Bernhard Kainz. Ratchet: Medical transformer for chest x-ray diagnosis and reporting. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VII 24*, pp. 293–303. Springer, 2021.

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

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Thirty-Third AAAI Conference on Artificial Intelligence*, 2019.

Jaehwan Jeong, Katherine Tian, Andrew Li, Sina Hartung, Fardad Behzadi, Juan Calle, David Osayande, Michael Pohlen, Subathra Adithan, and Pranav Rajpurkar. Multimodal image-text matching improves retrieval-based chest x-ray report generation. *arXiv preprint arXiv:2303.17579*, 2023.

Baoyu Jing, Pengtao Xie, and Eric Xing. On the automatic generation of medical imaging reports. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2577–2586, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1240. URL https://aclanthology.org/P18-1240.

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. *PLoS digital health*, 2(2):e0000198, 2023.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. 2023a.

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023c.

Mingjie Li, Rui Liu, Fuyu Wang, Xiaojun Chang, and Xiaodan Liang. Auxiliary signal-guided knowledge encoder-decoder for medical report generation. *World Wide Web*, 26(1):253–270, 2023d.

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

Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. Exploring and distilling posterior and prior knowledge for radiology report generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13753–13762, 2021.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Cyril Zakka, Yash Dalmia, Eduardo Pontes Reis, Pranav Rajpurkar, and Jure Leskovec. Med-flamingo: a multimodal medical few-shot learner. *arXiv preprint arXiv:2307.15189*, 2023.

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

OpenAI. Introducing chatgpt, 2022. URL https://openai.com/blog/chatgpt/.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*, 2022.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023.

Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. Interactive and explainable region-guided radiology report generation. In *CVPR*, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *arXiv preprint arXiv:2307.14334*, 2023.

Jun Wang, Abhir Bhalerao, and Yulan He. Cross-modal prototype driven network for radiology report generation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pp. 563–579. Springer, 2022a.

Sheng Wang, Zihao Zhao, Xi Ouyang, Qian Wang, and Dinggang Shen. Chatcad: Interactive computer-aided diagnosis on medical image using large language models. *arXiv preprint arXiv:2302.07257*, 2023.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M Summers. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9049–9058, 2018.

Zhanyu Wang, Mingkang Tang, Lei Wang, Xiu Li, and Luping Zhou. A medical semantic-assisted transformer for radiographic report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 655–664. Springer, 2022b.

Xian Wu, Shuxin Yang, Zhaopeng Qiu, Shen Ge, Yangtian Yan, Xingwang Wu, Yefeng Zheng, S Kevin Zhou, and Li Xiao. Deltanet: Conditional medical report generation for covid-19 diagnosis. *arXiv preprint arXiv:2211.13229*, 2022.

Yuan Xue, Tao Xu, L. Rodney Long, Zhiyun Xue, Sameer Antani, George R. Thoma, and Xiaolei Huang. Multimodal recurrent model with attention for automated radiology report generation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*, pp. 457–466, Berlin, Heidelberg, 2018. Springer-Verlag. ISBN 978-3-030-00927-4. doi: 10.1007/978-3-030-00928-1_52. URL https://doi.org/10.1007/978-3-030-00928-1_52.

Di You, Fenglin Liu, Shen Ge, Xiaoxia Xie, Jing Zhang, and Xian Wu. Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III*, pp. 72–82, Berlin, Heidelberg, 2021. Springer-Verlag. ISBN 978-3-030-87198-7. doi: 10.1007/978-3-030-87199-4_7. URL https://doi.org/10.1007/978-3-030-87199-4_7.

Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023.

Zihao Zhao, Sheng Wang, Jinchen Gu, Yitao Zhu, Lanzhuju Mei, Zixu Zhuang, Zhiming Cui, Qian Wang, and Dinggang Shen. Chatcad+: Towards a universal and reliable interactive cad using llms. *arXiv preprint arXiv:2305.15964*, 2023.

Qingqing Zhu, Tejas Sudharshan Mathai, Pritam Mukherjee, Yifan Peng, Ronald M Summers, and Zhiyong Lu. Utilizing longitudinal chest x-rays and reports to pre-fill radiology reports. *arXiv preprint arXiv:2306.08749*, 2023.

A Details of Model Architecture

As shown in the architecture figure in main paper, our approach introduce the medical visual features into the existing pretrained attention in the perceiver resampler of Flamingo using adaption prompts at each layer. After obtaining the similarity scores, to ensure that no instability will be introduced when initializing the model with medical feature introduced, we follow (Zhang et al., 2023) to apply softmax independent on two splits of the similarity score matrix, one on the scores corresponding to the Flamingo visual features and latent queries, and the other one on the scores corresponding to the newly introduced medical visual features. Specifically, S_l could be separated into: $S_l = \left[S_l^m; S_l^f; S_l^q\right]$ where $S_l^m \in \mathbb{R}^{t \times km^2}, S_l^f \in \mathbb{R}^{t \times kn^2}, S_l^q \in \mathbb{R}^{t \times t}$ represent similarity scores of the queries with respect to medical features, Flamingo vision encoder features, and the latent queries, respectively. We then apply a $\tanh$ gate controlled by a zero-initialized trainable parameter g_l. The resulting attention score at layer l is:

$$\text{Attn}_l = \left[\tanh\left(g_l\right) \cdot \text{Softmax}\left(S_l^m\right); \text{Softmax}\left(\left[S_l^f; S_l^q\right]\right)\right]$$

In this way, when the model is initialized, medical visual features will have zero effect, and the forward process is equivalent to the forward process of a pretrained vanilla Flamingo. As the training advances, the gate parameter g_l will be updated to gradually introduce the influence from medical visual features.

B *MIMIC-R3G* Generation Examples

Example of Generating Template Data

System Message: *Template:*
REPORT: [Imaging Protocol]
COMPARISON: [None/Compare with former image]
FINDINGS:
Lines/tubes:
Lungs:
Pleura:
Heart and mediastinum:
Bones:
Support Devices:
You are an assistant who helps format radiology reports using structured templates. You will be given a free-text radiology report, and you need to rewrite the report using this template by fill in the blanks. Leave any section blank if the report doesn't mention related information.
Input Report: FINAL REPORT EXAMINATION: CHEST (PORTABLE AP) INDICATION: ___ year old woman with pleurx for malignant effusion // ?appropriate position, response. please perform by 7 AM COMPARISON: ___ IMPRESSION: As compared to the previous image, no relevant change is seen. The extent of the left pleural effusion and the position of the left pleural drain are constant. Constant appearance of the mild opacity at the right lung bases. No new parenchymal changes. Unchanged size of the cardiac silhouette.

Response:
REPORT: CHEST (PORTABLE AP)
COMPARISON: Compare with former image
FINDINGS:
Lines/tubes:
Lungs: Constant appearance of the mild opacity at the right lung bases. No new parenchymal changes.
Pleura: The extent of the left pleural effusion and the position of the left pleural drain are constant.
Heart and mediastinum: Unchanged size of the cardiac silhouette.
Bones:
Support Devices:

$I_i = \langle$Manually designed instructions$\rangle$
$C_i = \langle$Manually designed template in system message$\rangle$
$R_i' = \langle$Text in response$\rangle$

Example of Generating Medical Record Data

System Message: You are a radiology assistant who help extract information from a chest x-ray radiology report. Output the corresponding section related to indication or medical history, if there are no such sections, output none. You need to infer the patient's possible medical conditions or history based on the indication and/or impression. Only list 2 to 4 of the possible abnormal medical history or conditions that are related to the indication, findings or impression of the report. Be short and concise and only output the list.

Input Report: FINAL REPORT INDICATION: ___-year-old female with fall, weakness. Please assess for pneumonia. TECHNIQUE: AP frontal and lateral chest radiographs were obtained. COMPARISON: None available. FINDINGS: Lung volumes are low. No focal consolidation, pleural effusion or pneumothorax is seen. There is no pulmonary edema. The cardiac silhouette is accentuated by low lung volumes and AP projection. IMPRESSION: Low lung volumes.

Response:
Indication: ___-year-old female with fall, weakness. Please assess for pneumonia.
Medical History:
1. Chronic obstructive pulmonary disease (COPD)
2. Asthma
3. Muscular dystrophy
4. Myasthenia gravis
Medical Tests:
1. Pulmonary Function Test: Reduced total lung capacity, reduced vital capacity, reduced forced vital capacity.
2. Complete Blood Count (CBC): Normal range.
3. Arterial Blood Gas Analysis: Normal pH, normal PaCO2, normal PaO2, normal HCO3.

$I_i = \langle$Manually designed instructions$\rangle$
$C_i = \langle$Text in response$\rangle$
$R_i' = \langle$Ground truth report from dataset$\rangle$

C More on experiments

C.1 Datasets and Implementation Details

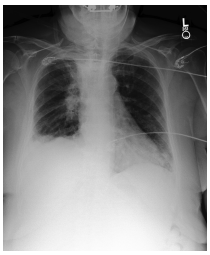
The proposed generated benchmark datasets are built upon the ground-truth report in MIMIC-CXR, which is the largest widely used report generation dataset. It consists of chest X-ray radiographs and reports of 227,835 studies from 64,588 patients, with a total of 227,835 reports and 377,110 x-ray images. The official training and test splits of MIMIC-CXR includes 386,960 images and 222,758 reports in training set and 5159 images and 3269 reports in test set.

We adopt OpenFlamingo (Awadalla et al., 2023), which is an opensource implementation of the Flamingo architecture. We use BioViL (Boecking et al., 2022) as our medical vision encoder. The BioViL medical vision encoder outputs a 15×15 grid of features with feature dimension 2048, which is then flattened and projected into 225 1024-dimensional vectors, which is same as the feature dimension of original CLIP ViT-L/14 encoder in Flamingo. The length of adaption prompt in perceiver sampler is same as the number of visual features from medical vision encoder output, which is 225. We maintain other model design parameters, *e.g.*, hidden dimension and number of attention heads, consistent with the OpenFlamingo implementation. For each data sample, we randomly sample two frontal view chest x-ray images associated with the study, or add a dummy zero-valued image if there is only one available frontal view image. We train the model on *MIMIC-R3G* data for 8 epochs with 4 batch size in all experiments. We use an ADAMW optimizer with $\beta_1 = 0.9, \beta_2 = 0.999$ and weight decay of 0.01 and set the learning rate 1e-4 with a 1000-step warm-up and a cosine decay schedule. Beam search with beam size of 4 is used for report generation.

C.2 Use Cases

In this section we present inference results of out model on all proposed report generation tasks. Input text prompt to the model is the concatenation of context and instruction in arbitrary order.

Inference Examples of No Context Generation

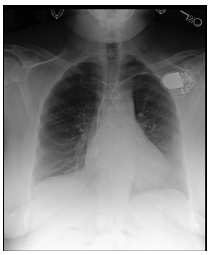


Ground Truth Report: (s50043351)
there is a right pleural effusion which is unchanged since prior exam. again seen is a right hilar opacity consistent with fibrosis, better assessed on recent ct. a subtle left lower lobe opacity is seen, which may represent atelectasis, but pneumonia cannot be excluded. the lungs are otherwise clear. the cardiomediastinal silhouette is unchanged from prior exam. visualized osseus structures are unremarkable.

Instruction:
Act as a radiologist and write a diagnostic radiology report for the patient based on their chest radiographs.

***DeMMo* Output**
in comparison with study of ____, there is continued enlargement of the cardiac silhouette with evidence for elevated pulmonary venous pressure. opacification along the right lateral chest wall suggests increasing right pleural effusion with compressive atelectasis at the base. mild atelectatic changes are seen at the left base.

Inference Examples of No Context Generation

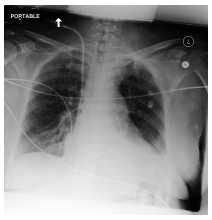


Ground Truth Report: (s56104633)
frontal and lateral views of the chest were obtained. patient is status post median sternotomy and cardiac valve replacement. dual lead left-sided pacemaker is seen with leads extending to the expected position of the right atrium and right ventricle. there may be minimal basilar atelectasis. no focal consolidation is seen. there is no pleural effusion or pneumothorax. the cardiac and mediastinal silhouettes are stable and unremarkable.

Instruction:
Based on the given chest X-ray images, please write a diagnostic radiology report for the patient.

***DeMMo* Output**
pa and lateral views of the chest provided. left chest wall pacer device is again seen with pacer leads extending to the region of the right atrium and right ventricle. midline sternotomy wires and mediastinal clips are again noted. the lungs are clear without focal consolidation, effusion or pneumothorax. the cardiomediastinal silhouette is normal. imaged osseous structures are intact. no free air below the right hemidiaphragm is seen.

Inference Examples of Generation with Previous Visit as Context



Ground Truth Report: (s59751598)
swan-ganz catheter has been advanced beyond the right hilum, and should be withdrawn for standard positioning, as discussed by telephone with dr. ___ at 9:45 a.m. on ___. new airspace opacity distal to the catheter tip could potentially represent pulmonary hemorrhage, but other etiologies such as atelectasis or aspiration are also possible. improving atelectasis in left lower lobe and persistent small left pleural effusion. incidental calcified granulomas within the left upper lobe.

Context:
The patient has a previous visit with the report: endotracheal tube and nasogastric tube remain in standard position. swan-ganz catheter has been slightly withdrawn, with tip terminating in the central right hilar region. left sided catheter has been removed, with no visible pneumothorax. cardiac silhouette remains mildly enlarged, but previously reported mild edema has nearly resolved. bibasilar retrocardiac atelectasis is present with some improvement on the left, and small left pleural effusion is unchanged.

Instruction:
Considering the patient's previous visit, please write a chest x-ray report for the patient based on the chest x-ray images:

***DeMMo* Output:**
in comparison with the study of ___, the endotracheal and nasogastric tubes have been removed. swan-ganz catheter remains in place. continued enlargement of the cardiac silhouette with indistinct pulmonary vessels suggesting elevated pulmonary venous pressure. opacification at the left base most likely reflects atelectasis and effusion, though in the appropriate clinical setting, supervening pneumonia would have to be considered. no gross evidence of pneumoperitoneum, though this is not an upright view.

Inference Examples of Generation with Previous Visit as Context

Ground Truth Report: (s58402174)
ap portable semi upright view of the chest. lung volumes are low limiting assessment. there is increased bibasilar atelectasis and bronchovascular crowding. overall cardiomediastinal silhouette is unchanged. the right upper extremity access picc line appears in unchanged position extending to the level of the cavoatrial junction. mild congestion is difficult to exclude in the correct clinical setting. no overt signs of edema.

Context:
Medical report from the last visit: rounded bilateral mid lung opacities are again seen, grossly unchanged and likely reflect consolidative infectious process given history of septic emboli. there is unchanged bibasilar opacification, which is likely atelectasis with left greater than right effusions. cardiac silhouette is markedly enlarged, similar to the most recent prior. left picc terminates in the cavoatrial junction. median sternotomy wires are intact.

Instruction:
Please write a diagnosis chest x-ray report by investigating the given chest x-ray images, referencing the patient's previous report:

***DeMMo* Output:**
portable semi-upright radiograph of the chest demonstrates low lung volumes which results in bronchovascular crowding. bibasilar atelectasis is unchanged. the cardiomediastinal and hilar contours are unchanged. there is no pneumothorax, pleural effusion, or consolidation. a left-sided picc line ends at the cavoatrial junction. median sternotomy wires are intact.

Inference Examples of Generation with Patient's Medical Record as Context

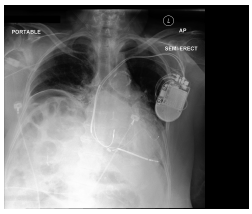
Ground Truth Report: (s54225810)
there is hilar congestion and diffuse bilateral ground glass opacities, most predominant at the bases, slightly improved from prior exam, and most consistent with pulmonary edema. an underlying pneumonia cannot be fully excluded. there are trace bilateral pleural effusions. there is no pneumothorax. the cardiac silhouette is moderately enlarged and unchanged from the prior exam. the mediastinal contours are normal.

Context:
The patient has the following medical history:
Medical History: 1. Multifocal Pneumonia 2. Pulmonary Edema 3. Heart Disease 4. Pleural Effusion Medical Tests: 1. Complete Blood Count (CBC): Elevated white blood cell count 2. Blood Culture: Possible bacterial growth indicating an infection 3. Sputum Culture: Presence of pathogenic bacteria indicating pneumonia 4. B-type Natriuretic Peptide (BNP) Test: Elevated BNP levels indicating heart failure

Instruction:
Examine the given chest x-ray images and patient's medical record and write a medical report detailing the findings.

***DeMMo* Output:**
as compared to the previous radiograph, the patient has developed mild-to-moderate pulmonary edema. in addition, bilateral pleural effusions have newly occurred. the heart continues to be slightly enlarged. no evidence of pneumonia. no pneumothorax.

Inference Examples of Generation with Patient's Medical Record as Context



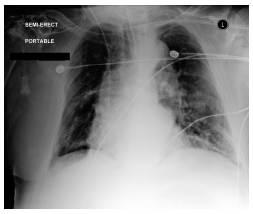
Ground Truth Report: (s58585557)
portable semi-upright radiograph of the chest demonstrates low lung volumes with resultant bronchovascular crowding. bibasilar consolidations may represent atelectasis or pneumonia in the appropriate clinical setting. the cardiomediastinal and hilar contours are unchanged. there is a new lucency beneath the right hemidiaphragm concerning for intra-abdominal free air. right-sided picc line and to the mid svc. unchanged position of the aicd. no pneumothorax.

Context:
The patient has following information: acute respiratory distress syndrome (ards), chronic obstructive pulmonary disease (copd)

Instruction:
Review the attached chest x-ray images and relevant patient information to write a detailed medical report:

***DeMMo* Output:**
portable semi-upright radiograph of the chest demonstrates low lung volumes with resulting bronchovascular crowding. there is an area of increased opacification at the right lung base, which may represent atelectasis, however aspiration or pneumonia could be considered in the appropriate clinical setting. there is a probable small right-sided pleural effusion. the cardiomediastinal and hilar contours are unchanged. a right-sided internal jugular central venous line ends at the cavoatrial junction. a dual-chamber pacemaker is present, with leads in unchanged position. there is no pneumothorax. free air is seen under the right hemidiaphragm, consistent with peritoneal dialysis.

Inference Examples of Generation with Template as Context



Ground Truth Report: (s52030252)
lungs:
- parenchyma: clear
- nodules/masses: absent
cardiomediastinal contour:
- heart size: enlarged
- mediastinal shape: normal
pleura:
- pleural effusion: absent
- pleural thickening: absent
bones:
- ribs/spine/clavicles/scapulae: normal
diaphragm and abdomen:
- diaphragm: normal contour
- abdominal component: not visible
soft tissues and other observations:
- soft tissue: normal
- foreign bodies: absent
- additional findings: atelectatic changes are seen at both lung bases, right more than left. there are improved lung volumes, notably on the left, potentially reflecting increased ventilatory pressure. no new parenchymal opacities. unchanged moderate cardiomegaly with tortuosity of the thoracic aorta.

Context:
Template:
Lungs:
- Parenchyma: [Clear — Consolidation — Interstitial markings — Other: please specify]
- Nodules/Masses: [Absent — Present]
If present, complete the following:
- Number of Nodules/Masses: [#]
- Size of the largest Nodule/Mass: [# mm — # cm]
- Location: [Right Upper Lobe — Right Middle Lobe — Right Lower Lobe — Left Upper Lobe — Left Lower Lobe; specify segment if known]
- Characteristics: [Well-defined — Spiculated — Cavitary — Calcified — Other: [specify]]
- Density: [Solid — Ground glass — Mixed — Other: [specify]]
Cardiomediastinal Contour:
- Heart Size: [Normal — Enlarged]
- Mediastinal Shape: [Normal — Widened — Other: please specify]
Pleura:
- Pleural Effusion: [Absent — Present]
If present, specify side and approximate volume if possible
- Pleural Thickening: [Absent — Present]
If present, describe extent and location
Bones:
- Ribs/Spine/Clavicles/Scapulae: [Normal — Fracture(s) — Lesion(s) — Other: [specify]]
Diaphragm and Abdomen:
- Diaphragm: [Normal contour — Elevated hemidiaphragm — Other: please specify]
- Abdominal Component: [Not visible — Gas under diaphragm — Other: please specify]
Soft Tissues and Other Observations:
- Soft Tissue: [Normal — Abnormality noted: please specify]
- Foreign Bodies: [Absent — Present: please specify location and appearance]
- Additional Findings: [None — Specify: please specify]
Support Devices: [None — pacemaker — PICC — tube — catheter — other]

Instruction:
Please fill this chest x-ray diagnostic report template based on the give chest x-ray radiographs.

***DeMMo* Output:**
lungs:
- parenchyma: interstitial markings
- nodules/masses: absent
cardiomediastinal contour:
- heart size: enlarged
- mediastinal shape: normal
pleura:
- pleural effusion: absent
- pleural thickening: absent
bones:
- ribs/spine/clavicles/scapulae: normal
diaphragm and abdomen:
- diaphragm: normal contour
- abdominal component: not visible
soft tissues and other observations:
- soft tissue: abnormality noted: patchy opacities in the lung bases
- foreign bodies: absent
- additional findings: low lung volumes
support devices: tracheostomy tube, right internal jugular central venous catheter

Inference Examples of Generation with Template as Context

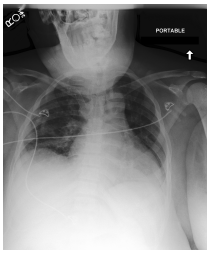
Ground Truth Report: (s55646831)
Bronchovascular markings: Right upper lobe scarring/chronic fibrosis in the right greater than left upper lobes.
Rest of the visualised lung fields: There are low lung volumes. New opacity in the left mid-to-lower lung.
Bilateral hilum:
Cardiac silhouette: The cardiac and mediastinal silhouettes are stable.
Costophrenic angles: No pleural effusion is seen.
Visualised bones & soft tissues:
Support Devices:

Context:
Template:
Bronchovascular markings:
Rest of the visualised lung fields:
Bilateral hilum:
Cardiac silhouette:
Costophrenic angles:
Visualised bones & soft tissues:
Support Devices:

Instruction:
Write a diagnostic radiology report for the patient based on their chest radiographs following the given template:

***DeMMo* Output:**
bronchovascular markings: diffuse airspace consolidation in the right lung
rest of the visualized lung fields: low volume, no evidence of cardiomegaly or pleural effusion
bilateral hilum: no large effusion or pneumothorax
cardiac silhouette: appears normal
costophrenic angles: clear
visualized bones & soft tissues: intact
support devices: none

Inference Examples of Report Revision



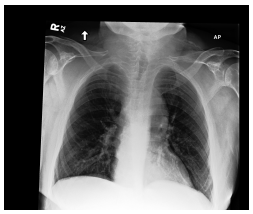
Ground Truth Report: (s57996680)
low lung volumes are present. this accentuates the size of the cardiac silhouette which is likely mildly enlarged. mediastinal and hilar contours are likely within normal limits. a right brachiocephalic venous stent is re- demonstrated. there is crowding of the bronchovascular structures with probable mild pulmonary vascular congestion. no pleural effusion or pneumothorax is identified.

Context:
Incorrect report: Normal lung volumes are present. The cardiac silhouette is of regular size. Mediastinal and hilar contours are likely mildly enlarged. No evidence of a right brachiocephalic venous stent. There is no crowding of the bronchovascular structures and no signs of pulmonary vascular congestion. Pleural effusion and pneumothorax are identified.

Instruction:
1. The lung volumes are described as normal in the incorrect report while they are low in the correct report. This needs to be fixed.
2. The incorrect report incorrectly states that there is no evidence of a right brachiocephalic venous stent, while the correct report states that it is re-demonstrated.
3. The incorrect report mistakenly identifies pleural effusion and pneumothorax, which are not present as per the correct report.
4. The incorrect report states that there is no crowding of the bronchovascular structures and no signs of pulmonary vascular congestion, which is contrary to the correct report. This needs correction.

***DeMMo* Output:**
low lung volumes are present. this accentuates the size of the cardiac silhouette which is enlarged. mediastinal and hilar contours are likely within normal limits. a right brachiocephalic venous stent is re- demonstrated. there is crowding of the bronchovascular structures with mild pulmonary vascular congestion. no pleural effusion or pneumothorax is identified.

Inference Examples of Report Revision



Ground Truth Report: (s56541072)
heart size is normal. the aorta is tortuous. unchanged widening of the mediastinum attributable to mediastinal lipomatosis is re- demonstrated. hilar contours are unremarkable. pulmonary vasculature is not engorged. lungs are clear. no pleural effusion, focal consolidation or pneumothorax is demonstrated. there are no acute osseous abnormalities.

Context:
Report: Heart size is enlarged. The aorta is not tortuous. No evidence of mediastinal lipomatosis. Hilar contours are abnormal. Pulmonary vasculature is engorged. Lungs have a small focal consolidation in the right lower lobe. Pleural effusion is present. There are multiple acute osseous abnormalities.

Instruction:
1. Revise the findings regarding heart size, aorta, mediastinum, hilar contours, pulmonary vasculature and lungs condition. The incorrect report inaccurately states these as abnormal when they are actually normal.
2. Correct the diagnosis of the presence of a small focal consolidation in the right lower lobe, pleural effusion, and multiple acute osseous abnormalities. These conditions are not present according to the correct report.
3. Include the detail about the mediastinal lipomatosis, as it is important information that was omitted in the incorrect report.

***DeMMo* Output:**
heart size is normal. the aorta is mildly tortuous. widening of the right paratracheal stripe is unchanged, and could suggest underlying lymphadenopathy or mediastinal fat or lipomatosis. hilar contours are normal. pulmonary vasculature is not engorged. lungs are clear without focal consolidation. no pleural effusion or pneumothorax is present. there are no acute osseous abnormalities. heart size is normal. the aorta is tortuous. unchanged widening of the mediastinum attributable to mediastinal lipomatosis is re- demonstrated. hilar contours are unremarkable. pulmonary vasculature is not engorged. lungs are clear without focal consolidation. no pleural effusion or pneumothorax is present. there are no acute osseous abnormalities.

D Limitations and Future Works

As introduced in previous sections, our method is a pure generation method without encompassing extra generation priors such as labels from a classifier. In contrast, methods such as Zhao et al. (2023) and You et al. (2021) utilize an image classifier to extract disease labels prior to generation, which ensures diagnostic correctness. Tanida et al. (2023) leverages object detector and use the extracted abnormal regions to guide generation, which also shows promising result. This presents a limitation, as our model's diagnostic accuracy may not be as reliable as methods employing guidance from high accuracy classifiers. Therefore, future works may focus on fusing the model with extra generation prior or guidance to further improve clinical efficacy.

We also observe that our *DeMMo* approach can be generalized to other domains as well using other domain-specific vision encoders. A potential future direction could entail utilizing a CT scan encoder for CT report generation, or developing a universal medical vision encoder for a more unified medical report generation tasks.