Adapting Lightweight Multimodal Language Models for Radiological Visual Question Answering A Practical Approach

Chang Sun [a,b]
chang.sun@maastrichtuniversity.nl

Aditya Shourya [b,c]
a.shourya@student.maastrichtuniversity.nl

[a] Institute of Data Science, Faculty of Science and Engineering, Maastricht University, Maastricht, The Netherlands
[b] Department of Advanced Computing Sciences, Faculty of Science and Engineering, Maastricht University, Maastricht, The Netherlands
[c] work done during an internship at [a]

Recent advancements in vision-language models have significantly enhanced the performance of medical visual question answering (MedVQA). However, challenges remain at each stage of development, including dataset curation, fine-tuning, evaluation, and identifying cases where the model might be ill-conditioned for use. In this study, we fine-tune a lightweight, generalist multimodal model for MedVQA, capable of answering both open and closed-ended queries. Our results show that even a smaller model can demonstrate substantial reasoning capabilities. Although it is not comparable to the current state-of-the-art models, it establishes a new benchmark within its size class. We provide a detailed account of the challenges and common pitfalls encountered during the entirety of model development. Additionally, we introduce a ready-to-use diagnostic tool based on saliency analysis, which aids current expert evaluation techniques and helps providing diagnostic indicators for poor generation quality. Our ablation studies highlight the importance of existing robust dataset curation techniques, such as annealing and multi-stage fine-tuning, to improve model performance.

Date : February 21, 2025
URL : https://github.com/adishourya/MedM

Maastricht University

1. Introduction

In recent years, significant advancements have been made in adapting Large Language Models (**LLMs**), which have already proven quite successful in general-purpose tasks, to specialized medical domains. Efforts by individual researchers [1], [2] and major organizations have resulted in notable models, such as Med-PaLM [3], a text-based medical expert capable of answering both consumer and professional medical questions, and LLaVA-Med [4], a Vision-Language Model (**VLM**) that can process open-ended queries involving biomedical images.

In addition to these advancements, there has been significant progress in improving model performance, consistency, and accessibility. This has enabled individual researchers to perform not just inference but also to fine-tune models for cheap. These, along with other advancements, continue to bring attention back to Medical LLMs. However, much of this interest is possibly inflated. As most of the publically available instruction tuned models are trained on datasets that lack sufficient medical content. Although post-training

innovations have enabled these models to perform well on a range of downstream tasks, they still struggle when applied to the specialized and complex demands of medical AI.

One of the primary reasons for this gap is the significant cost associated with generating labeled medical datasets, along with the complexities involved in encoding medical data, and the inherent challenges of recognizing intricate patterns in healthcare contexts which sometimes even leads to disagreement amongst clinicians. As a result, many generalist models fall short when applied to specialized medical tasks, as they lack the domain-specific expertise and alignment required to handle the nuanced nature of medical data.

The advent of multimodal models has partially addressed the latter part of the challenge, allowing the integration of multimodal datasets without constraining labelers to encode medical records into a single modality. While parallel high-quality image-caption pairs are readily available in large volumes for general domains [5], [6], such resources are scarce in the medical domain, further limiting the application of multimodal approaches in healthcare.

To improve upon generalist models,we turn to fine-tuning modern instruct models with a sufficiently narrow focus, offering the potential for more promising results. Besides demanding large, high-quality datasets Medical AI also needs robust safeguards against model biases. Models trained on biased data could lead to harmful decision-making, particularly if deployed for public use. Mitigating Biases remains an active area of research, with many public datasets cautioning against the deployment of models trained solely on their data.

Despite these hurdles, the potential of medical AI continues to inspire progress. While its practical usage in clinical settings remains limited, researchers and practitioners work to push the boundaries of what is achievable, carefully refining the technology and exploring their upper bound with each iteration.

2. Overview

2.1. Related Study

Our methodology builds upon recent advancements in medical visual and text-based question-answering, drawing insights from prior work in dataset curation, fine-tuning strategies, and evaluation frameworks. Several studies have introduced comprehensive pipelines that span data collection, model training, and rigorous assessment, highlighting the evolving capabilities of large language and vision-language models in the medical domain. Below, we summarize key contributions from related works that have informed our approach.

- **MedVInT-T(D,E) [7] :** presents a complete pipeline for medical visual question-answering , starting from data curation to model fine-tuning and evaluation design. Their approach involves finetuning a VLM Model on a self curated dataset [8] which contains multiple-choice style questions to cover variety of radioloical images and short fill in the blank style questions with the expectations to develop ability of also answering open ended queries. The model, fine-tuned on public benchmarks performs on par with existing MedVQA systems. Additionally, they manually verify a sample of test set results to make the models robust from current limitations of popular evaluation framework. A limitation we discuss in more detail in Section 5.
- **Medpalm [3] :** like [7], introduces a comprehensive framework from scratch. They curate a new dataset, HealthSearchQA sampling from existing medical QA datasets to cover both professional and consumer medical queries. They then fine-tune Flan-PaLM [9] on this dataset, achieving a new state-of-the-art model evaluated by both professional and layman on an extensive set of evaluation and alignment axes.
- **LLava-Med [4]** :curates its own dataset from PubMed Central [10] and, unlike previous approaches, prepares few-shot instruction training data using GPT-4 [11] and performs multi stage finetuning with a

cost-effective approach to achieve state-of-the-art performance in medical visual question answering, demonstrating strong multimodal understanding and instruction-following capabilities.

2.2. Key Contributions

With our article we aim at building upon prior work in medical Visual question-answering, providing the following key contributions:

1. **Reassessing Model Scaling Trends**: [3] demonstrated that scaling PaLM [9] models from 8B to 540B led to nearly a twofold accuracy improvement. However, their 8B model performed only slightly better than random perfomance on their benchmarking dataset. But, More recently, [4] achieved state-of-the-art results with just 8B parameters, showing that smaller models can now exhibit strong reasoning capabilities. To examine the longevity of such scaling-based ablation studies, we evaluate performance on an even smaller 2.8B LLM [12] based model with [13].
2. **Comprehensive Framework**: Like previous studies, we develop a complete pipeline covering data curation, fine-tuning, evaluation, and diagnostics. Additionally, we highlight common pitfalls at each stage of model training and assessment.
3. **Diagnostic Tool for Model Evaluation**:For human evaluation stage, we draw inspiration from [14], and present a ready-to-use diagnostic tool with saliency analysis to help assess model capabilities and limitations. We believe that this tool can serve as a supplementary aid for practicing radiologists during evaluation.

3. Model Architecture

3.1. Instruct Models

Large Multimodal models usually get trained in stages. The first step of training is usually referred to as **pre-training** where we train the model on a big corpus like [5], [6], [15], [16] to develop a **Base Model**. The Base Model in itself does not hold much utility to end-users as their sole objective has been to accurately predict the next token . To make the base model more useful to hold conversation we further train them on **Instruct Datasets** which aligns the model to make them suitable to hold conversations. .**Instruct models** act as the backbone for our application. This core functionality enables the model to generate coherent and contextually appropriate responses depending on the nature of the prompt.

Pre-training on large and diverse datasets is essential, as it allows these models to capture and learn diverse signals across a variety of domains, by enabling **transfer learning**. A technique where they are further trained on smaller, domain-specific datasets to refine their knowledge for a particular task. The effectiveness of an instruction-tuned model is often evaluated based on its performance in downstream applications without requiring extensive task-specific training, demonstrating its ability to generalize.

Models besides being trained on parallel image-text interleave data also get supervised on serialized chat data which covers various topics to reflect human preferences on different aspects such as instruction following, truthfulness, honesty and helpfulness. Most of the models that are deployed undergo robust internal evaluations to measure toxicity, profanity, and other potential issues in the generated captions.Its a common practice to further train on datasets such as the FairFace dataset [17] with the intentions that the model does not capture biases originating from Race, Gender or Age.

We select **PaliGemma-mix-448** [13] as our base models across all of our experiments. These models are particularly well-suited for our study, as they have been pretrained on a diverse and well-curated collection of publicly available datasets [18], [19], [20], [21]. This transparency in pretraining data stands in contrast

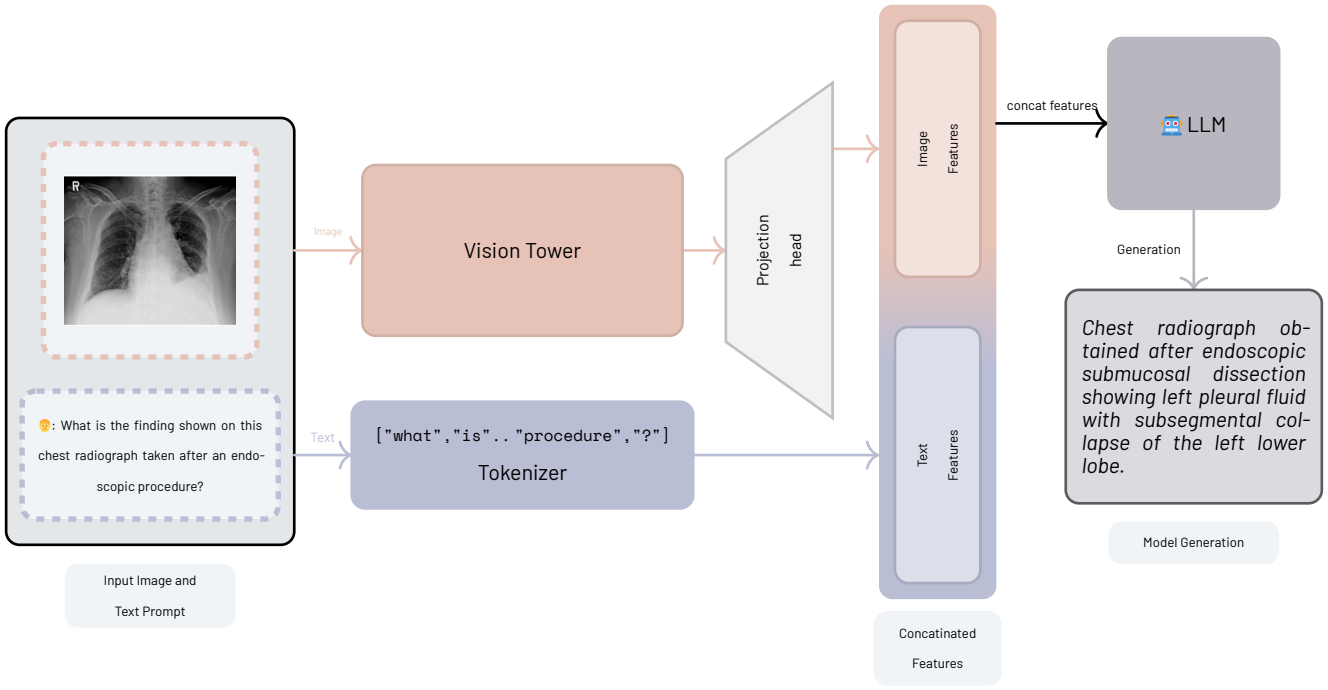

Figure 1: Model Architecture for Generalist VLMs [13]

with other approaches where such details are often undisclosed. This enables a better understanding of the model's zero shot capabilities and limitations.

The model demonstrates to be extremely versatile on transfer across many tasks.By leveraging these models, we aim to obtain the current boundaries of what we can achieve fine tuning lightweight VLM in terms of both accuracy and safety in medical domains.

3.2. Model Architercture

Most lightweight VLMs consist of a Vision Tower, a projector, a connector, and a mini language model, typically following the architecture outlined in Figure 1. For example, Paligemma specifically uses a SigLIP [22] Vision Tower, with roughly 400 million parameters, and the Gemma-2B [12] as its mini language model. Notably, during the multimodal pretraining stage of the model, no weights are frozen in time, allowing all parameters to learn during backpropagation.

We now present further details about the main components of the architecture.

1. **Vision Tower**: The Vision Tower is typically based on a decoder-only transformer architecture [23], [24]. Pretraining starts on a publicly available checkpoint of a shape-optimal Vision Transformer [25],as it is crucial for effectively scaling the model.Selecting a model that was pretrained contrastively on a large scale via the sigmoid loss is typical as the architecture demonstrates state-of-the-art performance when applied to classification tasks, ensuring the model can achieve good results in vision-based applications. At a high level, the Vision Tower takes one or more images as input and applies self-attention across image patches in a non causal manner to produce an output, commonly referred to as **image features**. It is important to note that this output is independent of any text instruction that may accompany the input image.
2. **Projector**: A projection layer is necessary to align the output dimensionality of the vision tower with the token dimension of the language model's vocabulary, which is required for concatenation. While the projection can be implemented using multiple layers, the ablation study by [13] found no significant advantage in employing multiple layers. Consequently,most VLMs use a single-layer projection.

3. **Concatenation**: The text prefix associated with the image is tokenized (here with [26]) and concatenated with the projected features from the vision tower. In practice, the prompt is often padded or truncated to align with the input dimension requirements of the language model.
4. **LLM**: With the concatenated features as input, a decoder-only transformer is employed as the language model. The first token is generated by attending fully to both the image features and the prefix token. Subsequent tokens are then generated autoregressively, relying on previously generated content and the concatinated features.

4. Training Recipe

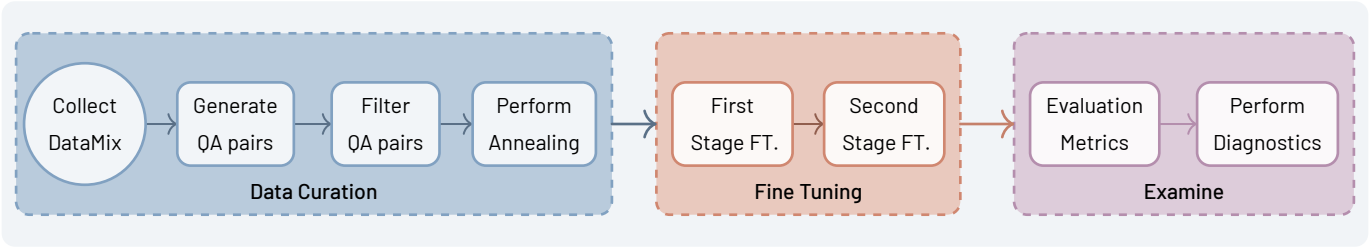

Figure 2: Methodology

Our main methodology that goes into all the stages of the model development can be summarized from Figure 2, We begin by sourcing publicly available radiological datasets and curating the data to ensure it is well-conditioned for training. We then take steps through multiple preprocessing steps to ensure the data is well conditioned for our model to be train on.

Next, we perform a two-stage fine-tuning process to optimize our model performance. The first stage fine-tuning can be seen as making the model learn preliminaries required before move onto the second stage, while the second stage training instruction sets contains examples to improve model's rigour and generalization capability.

After fine-tuning, we move on to evaluation, where we check model's genralizability score using standard natural language metrics and accuracy. We conduct ablation studies to assess the impact of our data curation and fine-tuning choices. Finally, we perform diagnostics with saliency analysis to gain insights over the model's capability and limitation.

4.1. Data Curation

Fine-tuning a medical VLM not only requires a large network but also a vast amount of high-quality data to ensure sufficient variety and coverage. A large dataset is particularly beneficial even if some examples contain noisy or fuzzy labels, as big datasets often provide enough reliable signal to counteract such inconsistencies. This is especially critical in the medical domain, where training datasets such as [27] are curated by sampling from open-access databases like [28] and then developes pipelines for automated extraction and filtration to generate annotated labels. While these annotations may sometimes contain fuzzy labels, the overall curation process ensures a diverse and representative dataset.

In this section, we describe our data curation techniques to train a model capable of answering open-ended radiological questions. We carefully work with datasets that already contain pre-processed data, eliminating the need for additional steps like de-identification or upsampling to ensure fairness. Our primary focus is on extracting high-quality signals for the model to learn visual concepts effectively. To achieve this, we generate question-answer pairs, which are more conducive to learning when compared to captioning-based approaches.

4.1.1. Determining the Data Mix

To develop a medical question-answering model, it is essential to carefully determine the proportion of different pathologies in the dataset. A lack of sufficient representation for certain conditions can lead to poor generalizability, limiting the model's ability to perform well across diverse real-world cases.

Moreover, understanding the composition of the training mix is crucial for setting realistic expectations regarding model performance during evaluation. To assess the quality of commonly used data mixtures, we rely on scaling law experiments, discussed in Section 4.2.

Below, we examine several widely used data mixtures that were considered for training our VLM.

- **SLAKE [29]**, contains detailed semantic annotations labeled by practicing physicians, covering a wide range of the human anatomy. This serves as an excellent preliminary radiological knowledge base for fine-tuning a medical VLM such as ours. While models like [13] demonstrate reasonable zero-shot performance in recognizing some body parts most likely due to their exposure to similar examples from open-web datasets during pre-training but this capability remains imperfect in the instruct model.

  To address this limitation, we finetune on SLAKE as the initial stage of our fine-tuning curriculum. Specifically, we save a checkpoint after training the model on SLAKE for five epochs, and this checkpoint serves as the starting point for further training for all of our subsequent models. This approach aligns with the curriculum learning strategy outlined in [30], which highlights the importance of first enabling a model to recognize simpler visual concepts before advancing to more complex structures.*
- **PMC-VQA** dataset [7] is constructed by sampling from [31] and is designed to facilitate medical visual question answering. It includes multiple-choice and fill-in-the-blank-style questions, providing a structured approach to assessing model performance. The dataset is generated and filtered using a large language model like [11], a process we further discus in Section 4.1.2. To evaluate its effectiveness, the authors train **MedVInT-TE** and **MedVInT-TD** on PMC-VQA, achieving performance scores comparable to the current state-of-the-art [4] on benchmark datasets like [32] , [33]
- **Radiology Objects in Context (ROCOv2)** dataset [27] derives from [28] and contains 79,789 radiological image-caption pairs. It has been extensively used in medical AI research like [34] for tasks such as image captioning, multi-class classification, and vision-language model pretraining. With an average caption length of approximately 20 words per example, ROCOv2 offers a rich and structured source of radiological

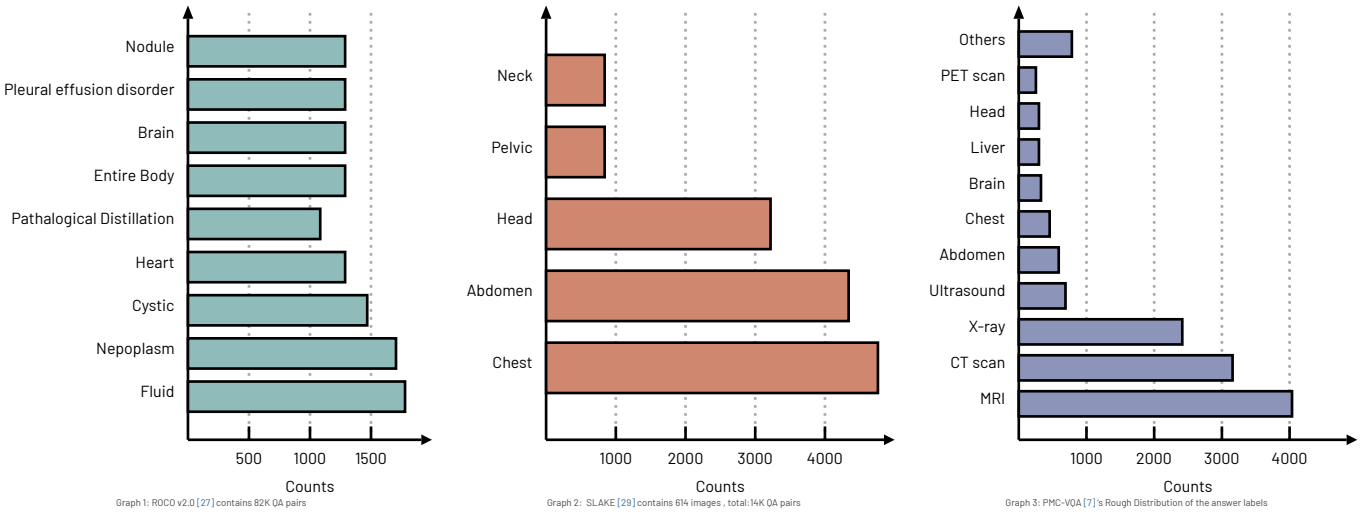

Figure 3: Value Counts of conditions across candidate datamixes

*!! *currently we have only done this with our pmc model.*

image-text data, making it particularly suitable for developing and refining open-ended vision-language models in the medical domain.

- **MedPix 2.0 [35]** is similar to [27], in utilizing a semi-automatic pipeline to extract radiological image-text pairs. It incorporates a manual curation process by sampling from MedPix® [36] to remove noisy labels. This curation effort spans over 12,000 cases, each containing images, diagnoses, and treatment information. Among all of our available datasets, this data mixture offers one of the most comprehensive collections of both textual and visual radiological concepts, making it particularly well-suited for annealing a technique we explore further in Section 4.1.3.

The **PMC-VQA** dataset provides direct image-question-answer pairs, making it suitable for fine-tuning a medical VQA model without the need for any preprocessing. Whereas **ROCOv2** and **MedPix 2.0** contain image-caption pairs, requiring additional processing to be useful for VQA training. We take inspiration from [7] and generate synthetic question-answer pairs from these datasets using a large language model. However, raw synthetic data may contain noise or inconsistencies, necessitating a filtering step to ensure high-quality supervision. Only after this refinement process can the resulting dataset effectively support training a medical VQA model. We discuss this synthesis and filtering process in the next section.†‡

4.1.2. Generating Question Answer Pairs

While training a VLM with a captioning task is an active area of research, modeling medical VLM often focuses on specific diagnostic queries rather than broad observations. This makes VQA particularly relevant for the medical domain, as it prioritizes answering clinically meaningful questions. The concept of training a VLM with VQA as its primary objective stems from the idea that posing insightful questions challenges the model to engage with more abstract visual concepts.

```py
def generate_qapairs(caption):
 # Construct the prompt for ollama for mcq based questions
 prompt = f"""
 Based on the following medical image captions generate appropriate and insightful question for the caption. Treat this caption as the ground truth to generate your question: {caption}
 """
 response = ollama.chat(model='llama3.1',
   messages=[{
     'role': 'user',
     'content': prompt}])

 # Return the generated text from the response
 return response['message']['content'].strip()
```

Listing 1:

We leverage several techniques to generate synthetic question-answer pairs from text alone using large language models, drawing inspiration from [38], [7], and [4]. The generation process requires no self made annotations and can be tailored to generate for both open and close-ended questions.For an example, we use Listing 1 to generate question answer pairs from [35] to train our VLM. We use [39] to generate our question-answer pairs mainly for two reasons: firstly, to generate question-answer pairs at scale at a low cost, making it feasible for small and individual research groups.And secondly, because the pretraining datamix for [39] is known, allowing us to inspect common pitfalls associated with language models that have not been pretrained on extensive medical content, which is the case for most LLMs that are publicly available. We discuss some of such pitfalls in Section 7.

4.1.3. Annealing and Filtering

Annealing is a data curation techniques aimed at improving model's overall performance by splicing small amounts of high quality data in existing data mix. With the objective that it increases models expectation to capture known to be good signals that might otherwise be difficult to detect in a diverse data containing both good but also some fuzzily labeled examples. Ablation studies by [39] found that Llama 8B model exhibited a

†our experiments with [37] is currently in development
‡PMC-15M [4] is unavailable to the public at the time of writing.

24% improvement on grade-school-level mathematical problem-solving [40] and a 6.4% increase in performance on competition-level mathematical reasoning tasks [41] after annealing. However, these gains were significantly less pronounced in larger models, such as the Llama 3 405B [39], indicating that larger model has higher capabilities of being robust to bad examples in the datamix during pre-training. Given that our VLM has approximately 4B parameters, it serves as an interesting approach to inspect the yielded benifits from annealing. We leverage Medpix v2.0 [35] as our primary dataset for annealing ROCO v2.0. Although Medpix v2.0 has a relatively limited volume, it serves as a highly valuable source of both case studies and literature, making it particularly well-suited for enhancing the model's medical reasoning and domain-specific knowledge acquisition.

A critical component of effective annealing is systematic **filtering**, which ensures that only high-quality and domain-relevant data is incorporated into the dataset. As illustrated in Figure Figure 6, the process begins with a medical corpus, which is filtered based on pathological relevance to maintain specificity. Unlike conventional upsampling strategies that primarily aim to increase the variety of rare cases, our approach prioritizes enriching common pathology cases (see Figure Figure 5). After sampling on selected pathologies, we process the captions using [39] LLama3 8B to generate Literature based and Image-related questions using specific prompt template for each. The underlying idea of including literature based questions in the datamix is that a stronger grasp of medical literature may enhance image-based reasoning capabilities, thereby improving the model's multimodal understanding. The generated questions then undergo subjective human evaluation as the last step before completing our data curation process.

Although our filtration technique ensures that only high-quality samples are added to the data mix, but it does not scale efficiently when dealing with large volumes of annealing data. As the dataset size increases,human labelling to filter data becomes infeasible for small research groups. We discuss some of these limitations further in Section 7.

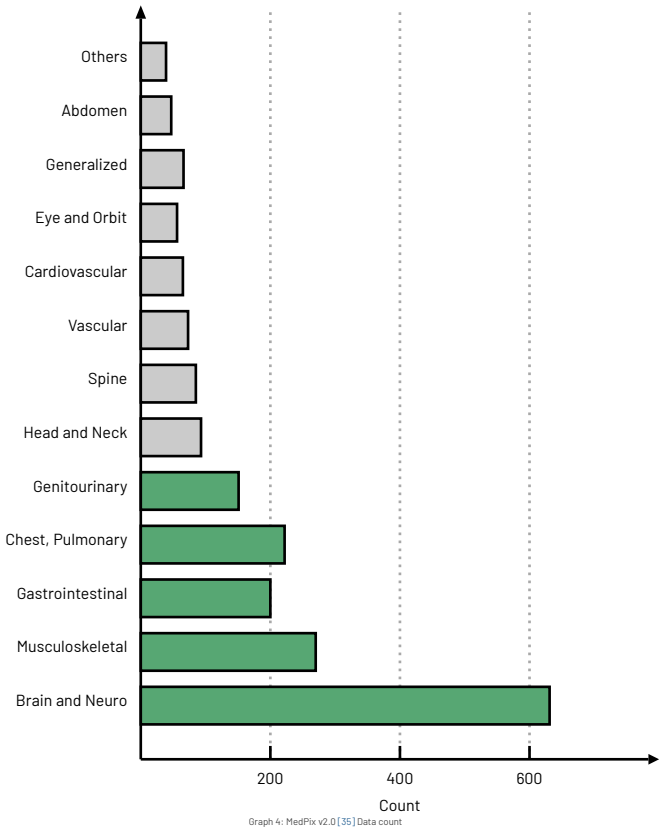

Figure 5: Selecting Pathologies for Annealing (here [35])

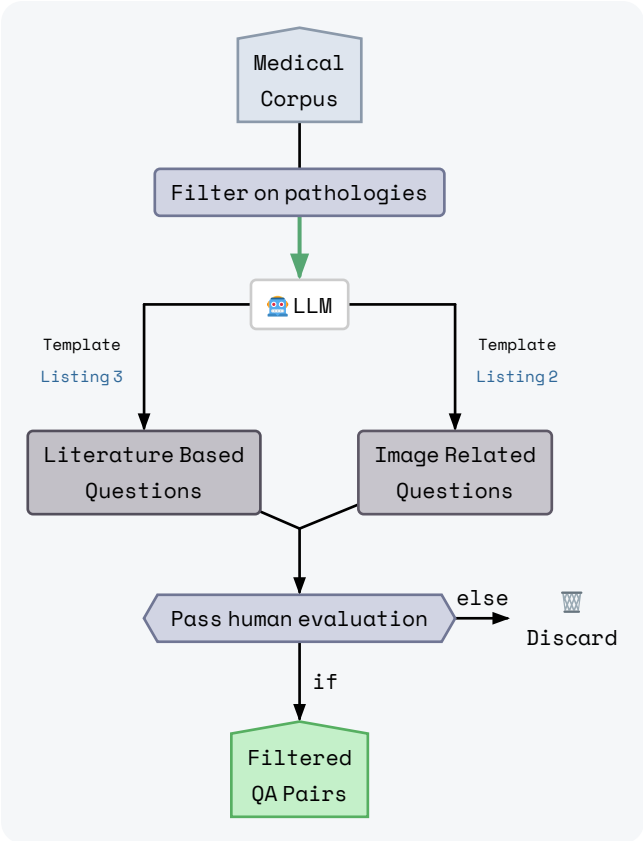

Figure 6: Generation and Filtering for Annealing a given Datamix (here [27])

4.2. Finetuning and Scaling Law

Lora [42] Finetuning Hyperparameters for all of our models: `r:12 , lora_alpha:32 , target_modules = {q,j,v,o,up,down}_proj , task_type : "causal_lm"`

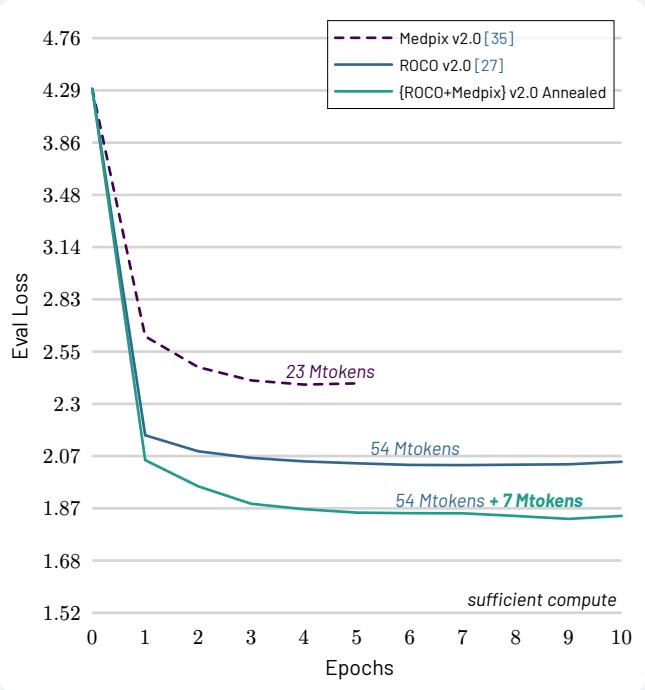

Figure 7: Evaluation loss for answering Open Ended Questions

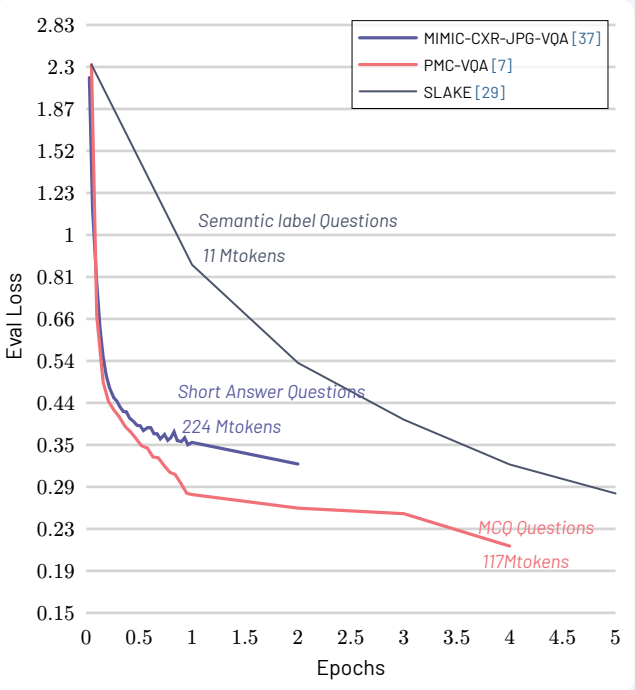

Figure 8: Evaluation loss for answering Closed Based Questions

*Mtokens: Million Tokens

Finetuning: in our approach we carry this out in two stages. As discussed very briefly in Section 4.1.1, we begin by fine-tuning using the SLAKE dataset [29] before training them on ([27] , [36] , [28] , [37]) separately as distinct instruction sets. We employ LoRA [42], a low-rank adapting method, to perform parameter efficient finetuning [43] of our models specifically by targeting the attention heads of both the vision tower and the large language model.This allows us to significantly lower computational and storage costs.This deviates from traditional approaches were only the last few layers or all of the model's parameter were trained to adapt the base model to a new domain. Figures Figure 7 and Figure 8 records ours evaluation loss across epochs of training. We discuss about their evaluation in Section 5.

For answering open ended questions we see evaluation loss decrease with increasing tokens in Figure 7. But the pattern breaks for answering close ended questions refer Figure 8. We discuss the effect of task dependence on finetuning below.

Scaling Law for Finetuning: scaling properties for LLM fine-tuning are highly task and data-dependent [44]. This means that the optimal fine-tuning method and the scaling behavior can vary depending on the characteristics of the fine-tuning dataset.Specifically we can have scaling exponents differ depending on the question-answer templates of our datamix. Since [27] and [36] contains open ended questions with caption length averaging around 20.The task dependence is not aparent.And mostly can be fitted by training on more but similar datamix with higher volume (As D_f is only volume dependent in Equation 1).

$$\tilde{L}(X, D_f) = A * \frac{1}{X^{\alpha}} * \frac{1}{{D_f}^{\beta}} + E \tag{1}$$

The task dependence becomes easier to analyze in dataset mixes for close-ended questions, especially when different templates are used for question-answer pairs,like in Figure 8. Typically, a higher-volume dataset is expected to yield greater improvements in fewer epochs. However, this trend does not hold when comparing [37] to [8]. Despite having less data, [8] shows greater learning gains in fewer epochs, primarily because it

consists mostly of multiple-choice questions, which have a lower expected loss compared to the template used in [37]. This observation suggests that different scaling laws are needed for such datasets.

Unfortunately, here we do not have enough data to fit a line to predict our loss on open ended questions when trained on higher volume in Figure 7. But we can see an improvement in generalizability of annealing; since its more prudent to perform annealing than to do scaling law experiments on small datasets like [36] .

5. Evaluation and Results

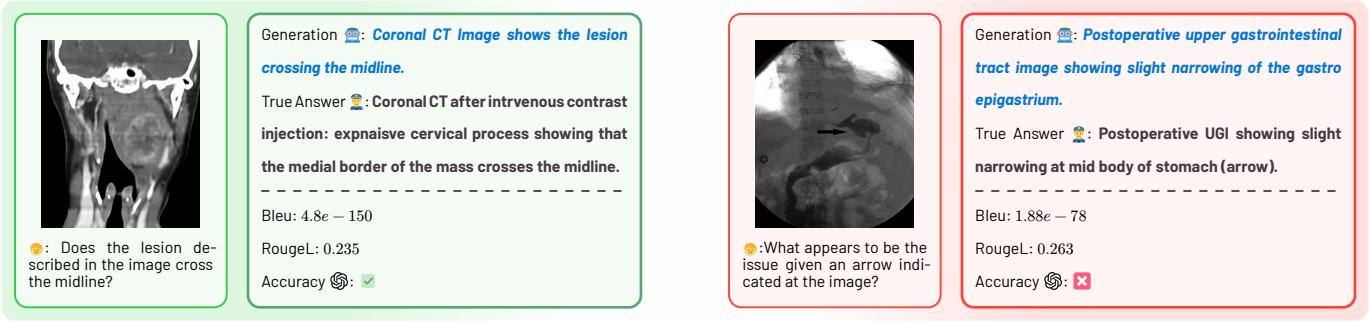

Figure 9: Generation and Evaluation on test samples from [27]

5.1. limitations of standard evaluation metrics

Before presenting our evaluation strategy, it is essential to highlight the shortcomings of standard NLP metrics and conventional accuracy calculations in fairly assessing generative models.

- **Multiple-Choice Accuracy Bias**: Many accuracy calculations for VLMs rely on multiple-choice questions, making scoring straightforward. However, this approach is inherently flawed, as expected accuracy is directly influenced by the number of answer choices. A possible solution is to increase the number of options, but this is often infeasible, especially for complex medical queries. While human annotation can enhance option quality, automated scoring remains an over-estimate of the true accuracy.

  Multiple-choice evaluation methods also suffers from model stochasticity. Since generative models can produce different outputs for the same prompt across multiple runs, accuracy scores may vary significantly on the same test set.

- **Limitations of Standard NLP Metrics**: Metrics like BLEU [45] and ROUGE [46], while widely used in NLP, are poorly suited for evaluating open-ended medical responses. These metrics primarily measure n-gram overlap, which fails to capture factuality or reasoning correctness. Since, open ended questions inherently can be answered in multiple valid ways, a low score does not necessarily indicate a poor response, but more importantly a high score does not ensure clinical reliability.

- Despite their limitations, we present these metrics for completeness while emphasizing the need for more robust evaluation strategies in Table 1 and Table 2.

5.2. Our Evaluation Methodology

- **Closed-Set Question Evaluation**: For multiple-choice questions, such as those in [8], we report accuracy as the primary metric.

§not yet done!

- To address the limitations of standard accuracy calculations, we draw inspiration from [7] and perform inference five times[§]. If the model generates different answer labels in at least three out of five runs, we dock a point to account for inconsistency.

- **Open-Ended Question Evaluation**: For free-form responses in datasets such as [27] and [36], we leverage llm based evaluation. Specifically, we use a predefined template[¶] to systematically assess model generations via [11].

5.3. Ablation studies across annealing and pre-stage finetuning

- We observed a performance improvement with annealing, even when using a small volume of annealing data, as shown in Table 1. While larger annealing datasets may further enhance robustness, our results indicate that even modest annealing can yield benefits for cheap.

Metrics	Medpix v2.0	ROCO v2.0	ROCO v2.0 + Medpix v2.0
rouge-s	0.311 ± 0.255	0.325 ± 0.132	0.334 ± 0.122
rouge-m	0.167 ± 0.082	0.179 ± 0.124	0.181 ± 0.124
rouge-l	0.308 ± 0.125	0.278 ± 0.120	0.304 ± 0.180
Bleu	0.055 ± 0.111	0.059 ± 0.090	0.077 ± 0.065
Accuracy	34/200 **(82/200)**	63/200 **(113/200)**	71/200 **(113/200)**

Table 1: results with no pre-stage finetuning ± 1 standard deviation (**LLava-Med** [4] as baseline)[‖]

-

Metrics	SLAKE	Medpix v2.0	ROCO v2.0	ROCO v2.0 + Medpix v2.0	PMC-VQA
rouge-s	-	? ± ?	? ± ?	? ± ?	-
rouge-m	-	? ± ?	? ± ?	? ± ?	-
rouge-l	-	? ± ?	? ± ?	? ± ?	-
Bleu	-	? ± ?	? ± ?	? ± ?	-
Accuracy	68/100 **(87/100)**	? **(82/200)**	? **(113/200)**	? **(113/200)**	0.31046

Table 2: results with pre-stage ± 1 standard deviation (**LLava-Med** [4] as baseline)[**]

6. Inspecting Saliency Maps

In our early model experiments, we observed generations indicative of what is commonly referred to as model hallucinations [47] when trained on an insufficient volume of image question answer pairs. This suggests that the LLMs can learn spurious signals from image features or text input IDs for a given pathology when scaling is inadequate. As noted by [3], most natural language metrics are poorly suited for evaluating medical question answering, and accuracy alone fails to capture the intermediate biases that the models learn.We believe that a subjective inspection of Saliency maps should be an axis of evaluation before a model is publically made available like the works in [48].

To mitigate this, one approach is to keep text input IDs constant or independent of the image, ensuring that the model primarily learns from image features. However, this approach has not led to good performance for visual question answering (VQA) on open-ended questions in general domains [49]. This suggests that a larger

[¶]add prompt
[‖]we will transpose the results later
[**]we will transpose the results later

volume of images or more sampled questions with adequate noise in the text input IDs is required for effective fine-tuning.But both of the approaches are constrained by high cost of labeling.

We believe more efforts can be made on detecting the source of bad signals by analyzing saliency maps. Although saliency is not the same as explainability [50], experts can often identify diagnostic indicators, as saliency in self-attention is fundamentally tied to the weights of queries and keys learnt by the layers of the model [48].

In our experiments, we follow from the works of [14], [51] and conduct saliency analysis on the intermediate layers of the LLM as our diagnostic approach. As prior to the concatenation layer, there is no interaction between the text input IDs and the image features. In many fine-tuning schemes for VLMs in general domains, the weights of the vision tower are typically frozen, and only the weights of the LLM are trained. This approach ensures that the image features, up to the projection layer, remain unchanged, with the expectation that the attention heads of the LLM will learn to filter and select only the relevant signals to guide the generation process. We visualize and analyze the interactions between the text input IDs, image features, and response tokens that occurs exclusively within the attention heads of the LLM.

6.1. Raw Attentions

We can use raw attention to perform saliency analysis by examining the interactions between queries and keys, this interaction can be seen as attempting to search for affinity or relevance of an interesting token with the rest. This relation is important as a collection of such relations ultimately guides the model's generation. In earlier layers, attention can be noisy compared to convolutional-based models, where saliency appears more contiguous. This is primarily because capturing relationships between features requires $O(n)$ operations for convolutional sequence-to-sequence models, while self-attention mechanisms can achieve this in constant time $O(1)$, making them more efficient and scalable [52]. Like convolutional models, self-attention-based models learn higher levels of abstraction as they propagate through the layers. So it becomes interesting to search for such relations in the last few layers , In Figure 10 we inspect 8th attention head of the penultimate layer. Inspecting these relationships is easy as it only involves collecting queries from target tokens and visualizing their attention over the keys of other tokens, either within the same or across different modalities. In Figure 10, we inspect a single head of attention by collecting all response tokens as the target query and plotting the average saliency over the image (higher saliency highlighted in red). Additionally, we can select specific image patches like the brightest spot in Figure 10 as query and plot the saliency on the keys of the response tokens, as shown in Figure 11. (higher saliency highlighted in blue, top 7 tokens underlined).

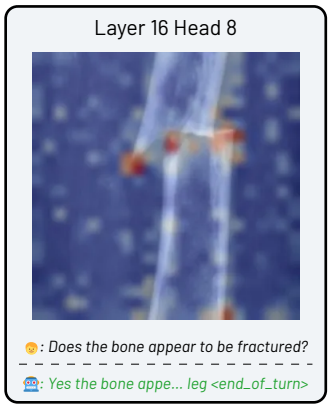

Figure 10: Interested Region

: Yes the Bone appears to be broken as there is a broken bone sticking out of the side of the leg <end_of_turn>

Figure 11: Saliency on response tokens

$$\underset{\text{Scaled Dot Product Attention}}{A} = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \text{ or } \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

query from selected response tokens; query from image patch; keys from image patches; keys from response

6.2. Attention Rollout

In Transformer-based models, raw attention weights do not always provide meaningful insights into token importance as we saw before. As information propagates through multiple layers, embeddings become increasingly mixed. This is because self-attention does not inherently preserve token identity across layers; rather, it continuously blends representations from multiple input tokens. As a result, by the time we reach higher layers, individual token contributions become obscure, and raw attention weights fail to capture the original token relationship [51] . Moreover, raw attention saliency maps often appear noisy and less interpretable compared to methods like [53], [54], which provide more structured visualizations of important regions.

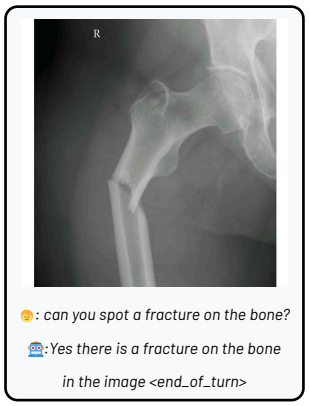

Figure 13: Test Input Example from [27]

To address this issue, we look into recent methods like [51], [55]. Attention Rollout tries to recursively aggregate attention across layers while also accommodating for skip connections. Our interest in using [51] as a diagnostic tool mostly comes as an inspiration from [48], where medical experts leveraged saliency-based explainability methods like Attention Rollout to study chest X-rays and CT scans for disease classification in COVID-19 and pneumonia cases. Furthermore, their analysis also identified cases where a well scaled model struggled in correctly classifying the disease. This further confirms our belief that saliency inspection from a practicing radiologist should be considered an essential axis of evaluation for vision-language models, particularly in medical question answering.††

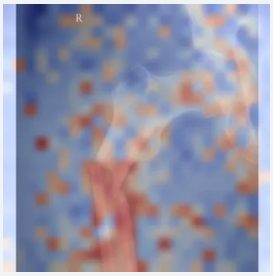

Figure 14: Rollout Saliency

7. Discussion and Limitation

Despite the improvements introduced in our methodology, several limitations remain, highlighting areas for future research and refinement.

- **Question-Answer Pair Generation**: LLMs such as LLaMA [39], struggle to generate high-quality open-ended queries, particularly when producing lengthy answers. This limitation extends to our automated filtering process Figure 5, which inherits the same issues. To mitigate this, we resorted to a manual filtering approach, ensuring higher-quality question-answer pairs which comes at the cost of increased time for data curation.
- **Ablation Studies on Vision Towers**: Existing ablation studies largely focus on scaling the LLM while sometimes overlooking critical aspects needed of the vision tower for radiological visual question answering. Although increasing the input resolution improves local feature extraction, it might still remain insufficient. We found our VLM struggle with images, such as chest X-rays, where fine-grained details are crucial. Notably, [56] found that a CNN-based ResNet model outperformed self-attention-based vision towers in scenarios requiring strong local feature context. Future work could explore depth-wise convolutions in Vision Transformers [57] to enhance feature extraction efficiency, particularly for small dataset training.

8. Conclusion

- While LLava-Med achieves strong results with an 8B model, our model falls short of state-of-the-art performance. However, despite using less than 1/4th of the total parameters, we achieve significantly better-than-

†† our experiments with employing saliency inspection as a diagnostic toll in visual question answering is still ongoing.

random performance. This validates our first key contribution, demonstrating that smaller models, when trained effectively, can still exhibit strong reasoning capabilities in radiological visual question answering.

- Our diagnostic tool successfully identifies typical group of examples that frequently lead to poor generations, such as chest images.

9. Data Availability

1. Medpix [35] for annealing data : 🤗 https://huggingface.co/datasets/adishourya/MEDPIX-ShortQA
2. Synthetically generated question answer pairs for ROCO V2.0 [27] can be found here:
   - Train split 🤗 https://huggingface.co/datasets/adishourya/ROCO-QA-Train
   - Valid and Test split 🤗 https://huggingface.co/datasets/adishourya/ROCO-QA

10. Code Availability

- All the code that went into finetuning our models and their model card can be found in our https://github.com/adishourya/MedM main repository.
- our fork of [14] can be found here: https://github.com/adishourya/lvlm-interpret

Bibliography

[1] V. C. Markus Zhang, "BabyDoctor." 2023.

[2] T. van Sonsbeek, M. M. Derakhshani, I. Najdenkoska, C. G. M. Snoek, and M. Worring, "Open-Ended Medical Visual Question Answering Through Prefix Tuning of Language Models." [Online]. Available: https://arxiv.org/abs/2303.05977

[3] K. Singhal, S. Azizi, T. Tu, and others, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 03 August 2023, pp. 172–180, 2023, doi: 10.1038/s41586-023-06291-2.

[4] C. Li *et al.*, "LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day." [Online]. Available: https://arxiv.org/abs/2306.00890

[5] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, "Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts," in *CVPR*, 2021.

[6] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning," in *Proceedings of ACL*, 2018.

[7] X. Zhang *et al.*, "PMC-VQA: Visual Instruction Tuning for Medical Visual Question Answering." [Online]. Available: https://arxiv.org/abs/2305.10415

[8] X. Zhang and others, "PMC-VQA Dataset." [Online]. Available: https://huggingface.co/datasets/xmcmic/PMC-VQA

[9] H. W. Chung *et al.*, "Scaling Instruction-Finetuned Language Models." [Online]. Available: https://arxiv.org/abs/2210.11416

[10] National Library of Medicine, "PubMed Central (PMC)." [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/

[11] OpenAI, "ChatGPT: A Large Language Model." 2024.

[12] G. Team *et al.*, "Gemma 2: Improving Open Language Models at a Practical Size." [Online]. Available: https://arxiv.org/abs/2408.00118

[13] L. Beyer* *et al.*, "PaliGemma: A versatile 3B VLM for transfer," *arXiv preprint arXiv:2407.07726*, 2024.

[14] G. B. M. Stan *et al.*, "LVLM-Interpret: An Interpretability Tool for Large Vision-Language Models." [Online]. Available: https://arxiv.org/abs/2404.03118

[15] T.-Y. Hsu, C. L. Giles, and T.-H. K. Huang, "SciCap: Generating Captions for Scientific Figures," *CoRR*, 2021, [Online]. Available: https://arxiv.org/abs/2110.11624

[16] B. Wang, G. Li, X. Zhou, Z. Chen, T. Grossman, and Y. Li, "Screen2Words: Automatic Mobile UI Summarization with Multimodal Learning," *CoRR*, 2021, [Online]. Available: https://arxiv.org/abs/2108.03353

[17] K. Karkkainen and J. Joo, "FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1548–1558.

[18] S. Changpinyo, D. Kukliansy, I. Szpektor, X. Chen, N. Ding, and R. Soricut, "All You May Need for VQA are Image Captions," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, Eds., Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 1947–1963. doi: 10.18653/v1/2022.naacl-main.142.

[19] A. Piergiovanni, W. Kuo, and A. Angelova, "Pre-training image-language transformers for open-vocabulary tasks." [Online]. Available: https://arxiv.org/abs/2209.04372

[20] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, I. Gurevych and Y. Miyao, Eds., Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 2556–2565. doi: 10.18653/v1/P18-1238.

[21] K. Srinivasan, K. Raman, J. Chen, M. Bendersky, and M. Najork, "WIT: Wikipedia-based Image Text Dataset for Multimodal Multilingual Machine Learning," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, in SIGIR '21. ACM, Jul. 2021, pp. 2443–2449. doi: 10.1145/3404835.3463257.

[22] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid Loss for Language Image Pre-Training." [Online]. Available: https://arxiv.org/abs/2303.15343

[23] A. Vaswani *et al.*, "Attention Is All You Need." [Online]. Available: https://arxiv.org/abs/1706.03762

[24] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." [Online]. Available: https://arxiv.org/abs/2010.11929

[25] I. Alabdulmohsin, X. Zhai, A. Kolesnikov, and L. Beyer, "Getting ViT in Shape: Scaling Laws for Compute-Optimal Model Design." [Online]. Available: https://arxiv.org/abs/2305.13035

[26] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing." [Online]. Available: https://arxiv.org/abs/1808.06226

[27] O. Pelka, S. Koitka, J. Rückert, F. Nensa, and C. M. Friedrich, "Radiology Objects in Context (ROCO): A Multimodal Image Dataset," 2018. [Online]. Available: https://labels.tue-image.nl/wp-content/uploads/2018/09/AM-04.pdf

[28] National Library of Medicine, "PMC Open Access Subset." 2003.

[29] B. Liu, L.-M. Zhan, L. Xu, L. Ma, Y. Yang, and X.-M. Wu, "SLAKE: A Semantically-Labeled Knowledge-Enhanced Dataset for Medical Visual Question Answering." [Online]. Available: https://arxiv.org/abs/2102.09542

[30] T. Srinivasan, X. Ren, and J. Thomason, "Curriculum Learning for Data-Efficient Vision-Language Alignment." [Online]. Available: https://arxiv.org/abs/2207.14525

[31] W. Lin and others, "PMC-CLIP: Contrastive Language-Image Pre-training Using Biomedical Documents," *Medical Image Computing and Computer Assisted Intervention*. Springer, pp. 525–536, 2023.

[32] J. J. Lau, S. Gayen, A. Ben Abacha, and D. Demner-Fushman, "A dataset of clinically generated visual questions and answers about radiology images," *Scientific data*, vol. 5, no. 1, pp. 1–10, 2018.

[33] A. Ben Abacha, S. A. Hasan, V. V. Datla, D. Demner-Fushman, and H. Müller, "VQA-Med: Overview of the Medical Visual Question Answering Task at ImageCLEF 2019," in *Proceedings of the Conference and Labs of the Evaluation Forum (CLEF) 2019 Working Notes*, Sep. 2019, pp. 9–12.

[34] "Overview of the ImageCLEF 2023: Multimedia Retrieval in Medical, Social Media and Internet Applications," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 14th International Conference of the CLEF Association, CLEF 2023, Thessaloniki, Greece, September 18-21, 2023, Proceedings*, in Lecture Notes in Computer Science, vol. 14163. Springer, 2023, pp. 370–396. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-42448-9_25

[35] I. Siragusa, S. Contino, M. L. Ciura, R. Alicata, and R. Pirrone, "MedPix 2.0: A Comprehensive Multimodal Biomedical Dataset for Advanced AI Applications." [Online]. Available: https://arxiv.org/abs/2407.02994

[36] National Library of Medicine, "MedPix: Free Online Medical Image Database." 2024.

[37] A. E. W. Johnson *et al.*, "MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs." [Online]. Available: https://arxiv.org/abs/1901.07042

[38] D. Zhu, J. Chen, K. Haydarov, X. Shen, W. Zhang, and M. Elhoseiny, "ChatGPT Asks, BLIP-2 Answers: Automatic Questioning Towards Enriched Visual Descriptions." [Online]. Available: https://arxiv.org/abs/2303.06594

[39] H. Touvron *et al.*, "LLaMA: Open and Efficient Foundation Language Models." [Online]. Available: https://arxiv.org/abs/2302.13971

[40] K. Cobbe *et al.*, "Training Verifiers to Solve Math Word Problems," *arXiv preprint arXiv:2110.14168*, 2021.

[41] D. Hendrycks *et al.*, "Measuring Mathematical Problem Solving With the MATH Dataset," *NeurIPS*, 2021.

[42] E. J. Hu *et al.*, "LoRA: Low-Rank Adaptation of Large Language Models." [Online]. Available: https://arxiv.org/abs/2106.09685

[43] L. Xu, H. Xie, S.-Z. J. Qin, X. Tao, and F. L. Wang, "Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment." [Online]. Available: https://arxiv.org/abs/2312.12148

[44] B. Zhang, Z. Liu, C. Cherry, and O. Firat, "When Scaling Meets LLM Finetuning: The Effect of Data, Model and Finetuning Method," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. [Online]. Available: https://openreview.net/forum?id=5HCnKDeTws

[45] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 2002, pp. 311–318. doi: 10.3115/1073083.1073135.

[46] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Text Summarization Branches Out*, Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://www.aclweb.org/anthology/W04-1013

[47] Z. Xu, S. Jain, and M. Kankanhalli, "Hallucination is Inevitable: An Innate Limitation of Large Language Models." [Online]. Available: https://arxiv.org/abs/2401.11817

[48] A. K. Mondal, A. Bhattacharjee, P. Singla, and A. P. Prathosh, "xViTCOS: Explainable Vision Transformer Based COVID-19 Screening Using Radiography," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 10, no. , pp. 1–10, 2022, doi: 10.1109/JTEHM.2021.3134096.

[49] C. Kolling, J. Wehrmann, and R. C. Barros, "Component Analysis for Visual Question Answering Architectures." [Online]. Available: https://arxiv.org/abs/2002.05104

[50] A. Bertrand, A. Pearce, and N. Thain, "Searching for Unintended Biases with Saliency," *PAIR Explorables*, 2022.

[51] H. Chefer, S. Gur, and L. Wolf, "Transformer Interpretability Beyond Attention Visualization," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2021, pp. 782–791. doi: 10.1109/CVPR46437.2021.00084.

[52] S. S. J. S. K. A. Sasha Rush Austin Huang and S. Biderman., "The Annotated Transformer." 2018.

[53] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, Oct. 2019, doi: 10.1007/s11263-019-01228-7.

[54] J. Gildenblat and contributors, "PyTorch library for CAM methods." GitHub, 2021.

[55] J. Gildenblat, "Explainability for Vision Transformers." 2021.

[56] S. Eslami, G. de Melo, and C. Meinel, "Does CLIP Benefit Visual Question Answering in the Medical Domain as Much as it Does in the General Domain?." [Online]. Available: https://arxiv.org/abs/2112.13906

[57] T. Zhang, W. Xu, B. Luo, and G. Wang, "Depth-Wise Convolutions in Vision Transformers for Efficient Training on Small Datasets." [Online]. Available: https://arxiv.org/abs/2407.19394

11. Appendix

- Finetuning Traning Arguments

Training Args	Value
learning_rate	$1e-5$
lr_schedule	constant
label_smoothing	0.0
weight_decay	0.0
fp16	True
gradient accumulation	16
batch_size	6

- All prompts are taken verbatim from [3]
  - for text based base llms from [4]

```py
prompt =f """
You are provided with a text description (Figure Caption) of a figure image from a biomedical
research paper. In some cases, you may have additional text (Figure Context) that mentions
the image. Unfortunately, you don't have access to the actual image.
Below are requirements for generating the questions and answers in the conversation:
- Avoid quoting or referring to specific facts, terms, abbreviations, dates, numbers, or names, as these may reveal the conversation is based on the text information, rather than the image itself. Focus on the visual aspects of the image that can be inferred without the text information.
- Do not use phrases like "mentioned", "caption", "context" in the conversation. Instead, refer to the information as being "in the image."
- Ensure that questions are diverse and cover a range of visual aspects of the image.
- The conversation should include at least 2-3 turns of questions and answers about the visual aspects of the image.
- Answer responsibly, avoiding overconfidence, and do not provide medical advice or diagnostic information. Encourage the user to consult a healthcare professional for advice.
"""
```

Listing 2:

- to generate caption pairs from [4]

```py
# for brief descriptions
brief_prompt = np.random.choice([
 "Describe the image concisely.",
 "Provide a brief description of the given image.",
 "Offer a succinct explanation of the picture presented.",
 "Summarize the visual content of the image.",
 "Give a short and clear explanation of the subsequent image.",
 "Share a concise interpretation of the image provided.",
 "Present a compact description of the photo's key features.",
 "Relay a brief, clear account of the picture shown.",
 "Render a clear and concise summary of the photo.",
 "Write a terse but informative summary of the picture.",
 "Create a compact narrative representing the image presented."
])
# for detailed prompts
detailed_prompts = np.random.choice([
"Describe the following image in detail",
"Provide a detailed description of the given image",
"Give an elaborate explanation of the image you see",
"Share a comprehensive rundown of the presented image",
"Offer a thorough analysis of the image",
"Explain the various aspects of the image before you",
"Clarify the contents of the displayed image with great detail",
"Characterize the image using a well-detailed description",
"Break down the elements of the image in a detailed manner",
"Walk through the important details of the image",
"Portray the image with a rich, descriptive narrative",
"Narrate the contents of the image with precision",
"Analyze the image in a comprehensive and detailed manner",
"Illustrate the image through a descriptive explanation",
"Examine the image closely and share its details",
"Write an exhaustive depiction of the given image"
])
```

Listing 3: