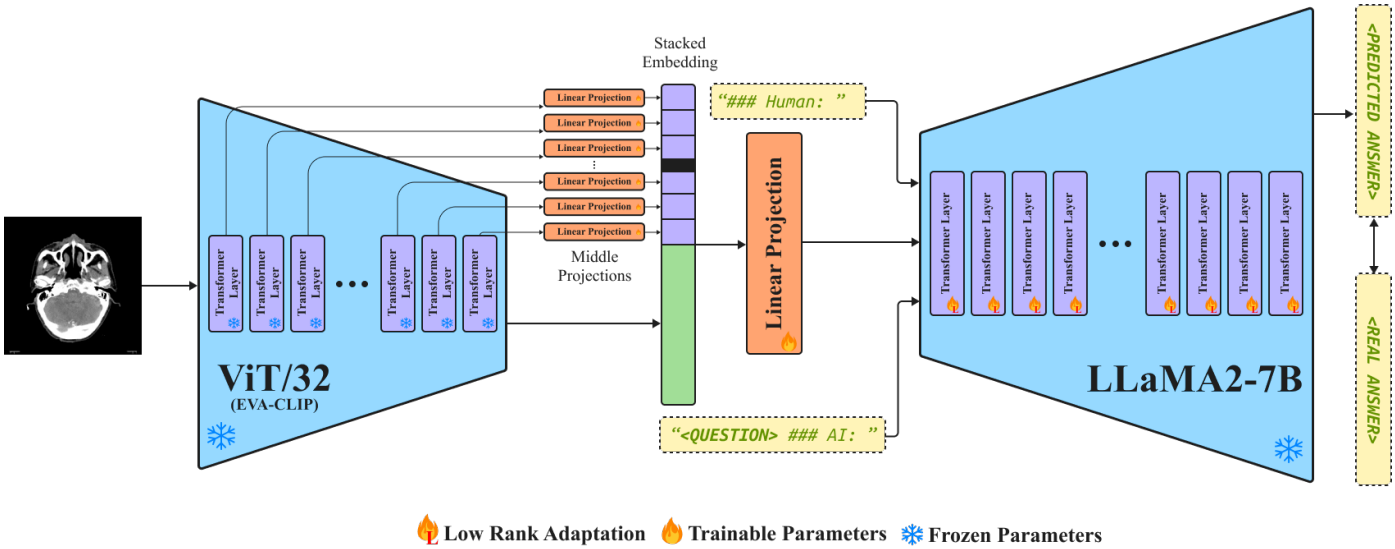
Graphical Abstract

MiniMedGPT: Efficient Large Vision-Language Model for Medical Visual Question Answering

Abdel Rahman Alsabbagh, Tariq Mansour, Mohammad Al-Kharabsheh, Abdel Salam Ebdah, Roa'a Al-Emaryeen, Sara Al-Nahhas, Waleed Mahafza, Omar Al-Kadi



Highlights

MiniMedGPT: Efficient Large Vision-Language Model for Medical Visual Question Answering

Abdel Rahman Alsabbagh, Tariq Mansour, Mohammad Al-Kharabsheh, Abdel Salam Ebdah, Roa'a Al-Emaryeen, Sara Al-Nahhas, Waleed Mahafza, Omar Al-Kadi

- Developed MiniMedGPT for efficient medical VQA, training in 30 minutes.
- Addressed dataset imbalances with Gemini Vision Pro and MediCap tools.
- Improved performance with minimal parameters compared to six VQA models.
- Potential tool for training clinicians and supporting radiologists.

MiniMedGPT: Efficient Large Vision-Language Model for Medical Visual Question Answering

Abdel Rahman Alsabbagh[a], Tariq Mansour[a], Mohammad Al-Kharabsheh[b], Abdel Salam Ebdah[b], Roa'a Al-Emaryeen[a], Sara Al-Nahhas[a], Waleed Mahafza[a,b], Omar Al-Kadi[a]

[a]*University of Jordan, King Abdullah II School of Information Technology, Amman, 11942, Jordan*
[b]*Jordan University Hospital, Diagnostic Radiology Department, Amman, 11942, Jordan*

Abstract

While Large Vision-Language Models (LVLMs) like GPT-4 and Gemini demonstrate significant potential, their utilization in the medical domain remains largely unexplored. This is due to challenges attributed to prolonged training and language generation issues. Imbalances within medical Visual Question Answering (VQA) datasets further complicate the integration of LVLMs. In this paper, we present a novel approach named **MiniMedGPT** (**Mini Med**ical **G**enerative **P**retrained **T**ransformer). Inspired by MiniGPT4-v2, MiniMedGPT is specifically designed for efficient medical VQA. The framework of MiniMedGPT is built upon both medical and generic pretrained Large Language Models and features an end-to-end versatile fine-tuning pipeline that enables the alignment of medical VQA data in just 30 minutes within a single-stage framework. To address language generation shortcomings and dataset imbalances, we employ Gemini Vision Pro and MediCap using them as an auxiliary component. Through comprehensive benchmarking and evaluations against 6 prominent medical VQA models across 2 well-known datasets, our approach brings an improved performance with the least number of trainable parameters against competitors across various performance metrics. This work can help train junior clinicians and has the potential to serve as a decision support tool for experienced radiologists.[1]

Keywords: Medical VQA, Large Vision-Language Model, MedGPT, Generative Pre-trained Transformers, Natural Language Processing.

1. Introduction

In the ever-evolving landscape of healthcare, the increase in patient numbers coupled with an increasing influx of medical practitioners poses a significant challenge for physicians. Not only must they meet the demands of their profession, but they are also tasked with guiding the next generation of healthcare professionals. Specifically, in the United States, where the population is approximately 335 million, the number of practicing radiologists hovers around 50 thousand, representing a notable minority [1]. Moreover, the enrollment in medical schools has been steadily rising, with over 96 thousand students currently enrolled — a growth of approximately 18% since 2012 [2].

In response to this growing demand for medical services and the shortage of specialized professionals, researchers have turned to automation as a potential solution. The goal is to streamline diagnostic processes, thereby reducing both time and costs. Although early automation attempts, such as the INTERNIST-1 expert system [3], showed promise, they were limited by the computational capabilities of their time and quickly became outdated. However, recent years have witnessed a resurgence of interest in automated diagnostic systems, fueled by advances in deep learning. In particular, Large Language Models (LLMs) have emerged as powerful tools, leveraging self-supervised learning techniques to achieve remarkable proficiency in understanding and generating natural language. This resurgence is made possible by the greater availability of computational resources and vast amounts of data. In particular, models such as ChatGPT [4] have demonstrated remarkable capabilities in conversational AI, albeit primarily in text-based interactions. However, despite their linguistic powers, these models lack an understanding of visual data, a critical limitation in medical diagnostics where images play a central role. Consequently, there is a growing interest in extending the capabilities of LLMs to include visual understanding, leading to the development of Large Visual Language Models (LVLMs). These models hold immense promise in providing valuable information to both healthcare professionals and patients, particularly in tasks such as VQ, where understanding both textual and visual information is essential. Despite that, the integration of LVLMs into the medical domain presents unique challenges. Training LVLMs requires significant computational resources and time, and their performance may suffer when applied to medical data due to the inherent noise and imbalance in medical datasets. Additionally, the lack of large-scale annotated datasets poses a significant hurdle to the development and evaluation of LVLMs in medical applications.

Email addresses: abd0200811@ju.edu.jo (Abdel Rahman Alsabbagh), tar0205865@ju.edu.jo (Tariq Mansour), mhm8201475@ju.edu.jo (Mohammad Al-Kharabsheh), abd8221729@ju.edu.jo (Abdel Salam Ebdah), roa0194961@ju.edu.jo (Roa'a Al-Emaryeen), sar0187721@ju.edu.jo (Sara Al-Nahhas), wmahafza@ju.edu.jo (Waleed Mahafza), o.alkadi@ju.edu.jo (Omar Al-Kadi)

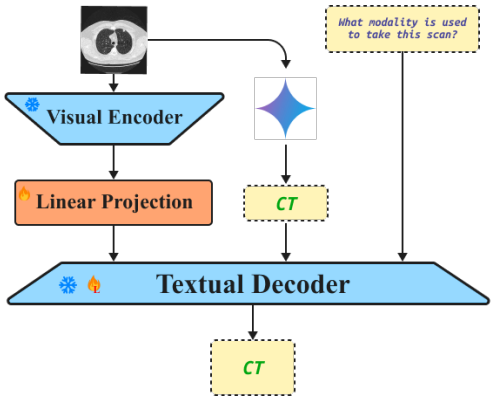


Figure 1: Proposed MiniMedGPT Model: The diagram illustrates input CT image encoding combined with implicit knowledge from Gemini and MediCap (star sign). The small dotted box shows the engine output before the textual decoder, while the large dotted box displays the predicted answer.

This work explores the capabilities of LVLMs in the context of Visual Question Answerin (VQA) medical imaging. Specifically, we propose **Mini Med**ical **G**enerative **P**retrained **T**ransformer, or **MiniMedGPT**, a method that enables faster training while maintaining or even surpassing the performance of existing approaches. Additionally, we address the challenges posed by imbalanced medical data through an end-to-end training scheme, leveraging knowledge engines to compensate for the lack of large-scale datasets. A general pipeline of the proposed model is shown in Figure 1. Our contributions are threefold: a) an LVLM optimized for efficient training is constructed that is capable of converging and completing training in approximately 30 minutes; b) LVLM is trained on medical imaging data, specifically targeting the medical VQA task, and develops a universal pipeline for processing medical VQA data; c) caption techniques are incorporated using advanced tools like Google's Gemini Pro and MediCap to enhance both input data and predictions.

2. Related Works

2.1. Large Language Models

The emergence of LLMs, especially transformer architectures, has revolutionized Natural Language Processing. Early models like GPT-2 and BERT [5] laid the foundations, but faced challenges with extensive contexts, diverse responses, and rare queries. BERT's tokenization method also struggled with number representation [6]. GPT-3, with 175 billion parameters compared to 1.5 billion in GPT-2 and 110 million in BERT, significantly improved handling longer contexts and generating diverse responses. Its success inspired the development of LLMs like PaLM, Megatron Turing NLG, BLOOM, Chinchilla, and LLaMA [7], and LLaMA shows promise as a base textual decoder for our work.

2.2. Large Vision Language Models

As LLMs transformed natural language processing, LVLMs emerged for visual understanding. CLIP [8] by OpenAI bridges visual and language comprehension without task-specific training, though it struggles with domain-specific queries. Models like Bootstrapped Language Image Pretraining [9] followed, integrating visual and textual data. PaLM-E [10], with 562 billion parameters, furthered this integration by combining real-world sensory inputs with language. Large Language-and-Vision Assistant [11] and KOSMOS-1 [12] advanced multimodal models capable of contextual learning and zero-shot tasks. Visual ChatGPT [13] and MM-REACT [14] combined vision models with ChatGPT for enhanced multimodal reasoning. MiniGPT4-v2 [15] and Vicuna [16] align visual encoders with LLMs, enabling detailed image descriptions and creative tasks. While these innovations impact general domains, our goal is to develop efficient LVLMs for medical applications based on insights from MiniGPT4-v2 and LLaVA.

2.3. Medical Chatbots

The rise of chatbots across industries, particularly in healthcare, holds great promise [17]. Chatbots can offer basic medical information, symptom screening, and personalized health advice, improving accessibility and patient care. LLMs like GPT-3, with their human-like language understanding, are valuable for healthcare interactions. However, specific challenges, especially in Visual Question Answering (VQA) for medicine, persist. While open-domain VQA has advanced, addressing medical-specific challenges, such as designing goal-oriented systems and curating clinical datasets, is essential. For instance, [18] developed a generative model for medical VQA, PMC-LLaMA, and curated the PMC-VQA dataset. [19] created M212, using self-supervised learning for radiographic images. [20] introduced Med-VINT, tailored for small medical datasets, with parameter-efficient fine-tuning. Despite these efforts, efficient use of LVLMs in medicine remains underexplored. This work proposes the use of LVLM with single linear layers to map visual and textual components in medicine. Leveraging MiniGPT4-v2, we aim to build a model optimized for medical VQA, utilizing Gemini Vision Pro and MediCap to enhance output quality. This is the first attempt to harness LVLM efficiency in healthcare communication.

3. Methodology

The main objective is to seamlessly integrate and establish robust mappings between medical images and their corresponding question-answer pairs. To achieve this, we devise a comprehensive methodology that includes two key components: the visual encoder and the textual decoder. For the visual encoder, we employ a state-of-the-art pre-trained Vision Transformer (ViT) model with a patch size of 32 (ViT32), sourced from the cutting-edge EVA-CLIP framework. This model, obtained through the adapter Parameter-Efficient Fine-Tuning (PEFT) technique, harnesses the vast reservoir of knowledge accumulated during its training. Using ViT32, MiniMedGPT capitalizes on the nuanced understanding of visual data acquired by the model, seamlessly aligning with the complexities inherent in medical images. Moreover, the utilization of PEFT ensures

efficient adaptation to domain-specific nuances, enhancing the model's capability to extract salient features from medical imagery.

In the realm of textual decoding, MiniMedGPT leverages the formidable capabilities of the Large Language Model (LLM), specifically the LLaMA architecture [7]. We obtain LLaMA weights from the open source LLaMA2-7B model, fine-tuned using the Parameter-Efficient Fine-Tuning (PEFT) technique known as DoRA for Low Rank Adaptation (LoRA) [21]. This fine-tuned version of LLaMA, adeptly trained in medical text data, exhibits a nuanced understanding of the intricate semantics prevalent in medical discourse. By incorporating LLaMA as our textual decoder, MiniMedGPT benefits from its prowess in diverse linguistic tasks and its tailored expertise in medical language comprehension. The language and vision models within MiniMedGPT are intricately connected through a linear projection layer. This architectural design, inspired by the methodology employed in MiniGPT4-v2, facilitates efficient training while maintaining computational speed and efficacy. By only training additional DoRA weights and establishing a streamlined connection through the linear projection layer and freezing the other layers, MiniMedGPT achieves optimal convergence and performance, as depicted in Figure 2.

3.1. Middle Projections

In complex deep learning structures such as Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), the initial layers excel at discerning fine features such as edges and corners, while the subsequent layers specialize in grasping broader elements in an image, like entire objects. The base architecture for MiniMedGPT only takes the embedding of the image from the last layer and in return loses crucial information at different scales. This is more critical in medical images, where diagnostics is based on small intricacies that can sometimes be challenging to differentiate. To solve this problem, we introduce the idea of Middle Projections (MPs), where the output of each transformer layer in our visual encoder is individually processed and learned to fully capture all scales in the input image. First, we abstractly define the base visual encoder (BVE) as follows:

$$\text{BVE} = \text{Linear}_N(\text{TL}_N(\text{TL}_{N-1}(\ldots(\text{TL}_1(\text{Patching}(\text{I})))))) \quad (1)$$

Where $\text{Linear} \in \mathbb{R}^{d_{\text{Transformer}} \times d_{\text{Out}}}$ denotes the trainable linear projection layer at the end of the encoder, TL denotes a Transformer Layer, $\text{I} \in \mathbb{R}^{H \times W \times C}$ denotes the Image as the input, and $N \in \mathbb{N}$ denotes the number of TLs. Here, $N = 39$, $d_{\text{Transformer}} = 1418$, and $d_{\text{Out}} = 5672$.

$\text{Patching} \in \mathbb{R}^{d_{\text{InPatch}} \times d_{\text{OutPatch}}}$ denotes the patching and flattening operations for preprocessing the image before entering the TLs. The number of patches in the height direction, $N_H \in \mathbb{N}$, is calculated as the height of the image divided by the patch size, denoted as $P \in \mathbb{N}$. Similarly, the number of patches in the width direction, $N_W \in \mathbb{N}$, is computed as the width of the image divided by the size of the patch P. The total number of patches, $N_{\text{patches}} \in \mathbb{N}$, is then determined by multiplying N_H and N_W. The dimensionality of the patch embeddings, referred to as the patch dimension, is denoted by $d_{\text{proj}} \in \mathbb{N}$. Here, $d_{\text{InPatch}} = N_H \times N_W \times C$, and $d_{\text{InPatch}} = d_{\text{proj}}$. We define the visual encoder after MPs as the following:

$$\text{MPs} = [\text{MP}_1(\text{TL}_1(\text{I})) \| \text{MP}_2(\text{TL}_2(\text{I})) \| \ldots \| \text{MP}_N(\text{TL}_N(\text{I}))] \quad (2)$$

$$\text{FullVE} = [\text{MPs} \| \text{BVE}] \quad (3)$$

Where $\text{MP} \in \mathbb{R}^{d_{\text{Transformer}} \times d_{\text{MP}}}$, $d_{\text{MP}} = 64$, and $\|$ denotes the concatenation operation. FullVisualEncoderOutput is also denoted as `<IMAGE>` embedding.

3.2. Medical Visual Question Answering Alignment

Following obtaining the `<IMAGE>` embedding, we tokenized the embedding and structure it to fit the LLM input prompt. This step allows the model to learn the overall structure of the medical data, and this is done quickly given that the model already incorporates pretrained components, the ViT32 and LLaMA. The input will look like the following:

```
### Human:  <IMAGE><IMAGE-FEATURES><\IMAGE>
        <QUESTION> ### Assistant:
```

For multiple questions relating to the same image, it would look like the following:

```
### Human:  <IMAGE><IMAGE-FEATURES><\IMAGE>
<QUESTION> ### Assistant:  <PREDICTED-ANSWER>
   ### Human:  <QUESTION> ### Assistant:
```

To further enhance the quality of our answers, we utilize Gemini and MediCap as an implicit knowledge engine. The modified version of the prompt looks as follows:

```
### Human:  <IMAGE><IMAGE-FEATURES><\IMAGE>
Based on the following caption <CAPTION>,
        <QUESTION> ### Assistant:
```

3.3. Masked Language Modeling

We opted for cross-entropy loss, a widely adopted measure that effectively penalizes deviations between predicted and actual distributions for the Masked Language Modelling (MLM) task of training an LLM. Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ be a sequence of input tokens, and let $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ be the corresponding sequence of target tokens, where some of the tokens in $\mathbf{x}$ are masked. With V being the size of the vocabulary, $p_\theta(y_i \mid \mathbf{x})$ being the probability assigned by the model to the target token y_i given the input sequence $\mathbf{x}$, $\mathcal{M}$ being the set of positions in the sequence that are masked, and $\mathbf{1}_{\{i \in \mathcal{M}\}}$ being an indicator function that is 1 if the position i is masked and 0 otherwise. The cross-entropy loss for masked language modeling is given by

$$\mathcal{L} = -\sum_{i=1}^{n} \mathbf{1}_{\{i \in \mathcal{M}\}} \log p_\theta(y_i \mid \mathbf{x}) \quad (4)$$

Expanding the probability $p_\theta(y_i \mid \mathbf{x})$ as the softmax of the logits z_i, gives

$$p_\theta(y_i \mid \mathbf{x}) = \frac{\exp(z_{i,y_i})}{\sum_{v=1}^{V} \exp(z_{i,v})} \quad (5)$$

and substituting (5)) in (4)), we have the following.

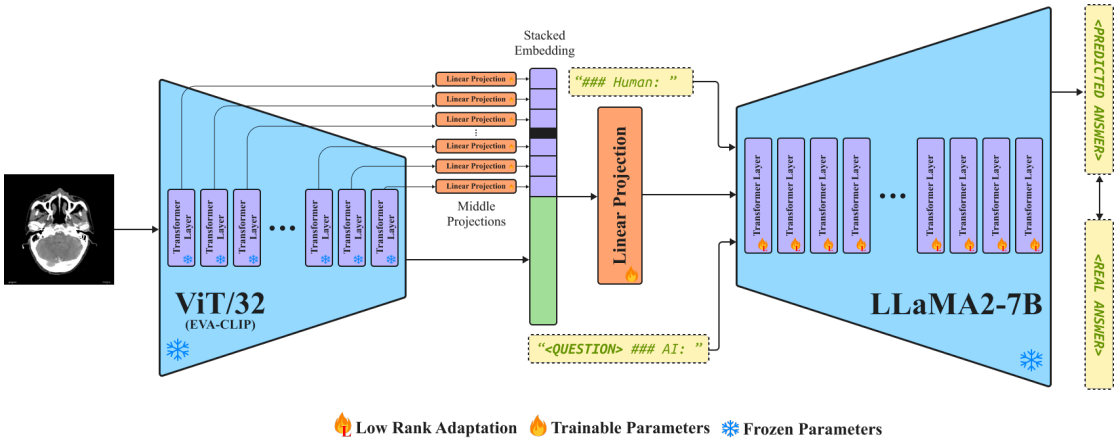


Figure 2: Complete Framework of the Proposed MiniMedGPT Model.

Algorithm 1 MiniMedGPT Algorithm

Input: VQA dataset D with components V (visual/image), Q (question), A (answer), and C (caption)
Freeze(EVA-CLIP_ViT32)
Freeze(LLaMA2-7B)
Seed $\leftarrow$ 42
$V \leftarrow$ Resize(V, 448 $\times$ 448, Bicubic interpolation)
$V \leftarrow$ Normalize(V)
$F_V \leftarrow$ EVA-CLIP_ViT32(V)
$L_V \leftarrow$ Linear/$L(F_V)$
Prompt $\leftarrow$ "###Human: <Image><L_V></Image>Based on the following caption <C>, <Q>###Assistant:"
Initialize LLaMA2-7B
Batch_size $\leftarrow$ 64
Epochs $\leftarrow$ 5
Optimizer $\leftarrow$ AdamW(10^{-4}, 0.9, 0.999)
for Epoch in range(Epochs) **do**
  for Batch in dataset with Batch_size **do**
    $\hat{A}_i \leftarrow$ LLaMA2-7B(Prompt)
    $L\ \mathcal{L}(A_i, \hat{A}_i) \leftarrow A_i$: $\mathcal{L} = -\sum_i A_i \log(\hat{A}_i)$
    Backpropagate(Optimizer)
  end for
end for

$$\mathcal{L} = -\sum_{i=1}^{n} \mathbf{1}_{\{i \in \mathcal{M}\}} \left(z_{i,y_i} - \log \sum_{v=1}^{V} \exp(z_{i,v}) \right) \tag{6}$$

Algorithm 1 shows the process of this phase in detail.

4. Experimental Design

4.1. Datasets

4.1.1. VQA-RAD

VQA-RAD [22] is a radiology-specific pivotal data set designed to facilitate training and evaluation of VQA models in the medical domain. With a meticulously curated collection of 315 radiology images, VQA-RAD offers a balanced representation of crucial anatomical regions such as the head, chest, and abdomen. This dataset includes 3,515 question-answer pairs, providing a diverse array of queries pertinent to radiological interpretation. The significance of VQA-RAD lies not only in its breadth, but also in the depth of its annotations. With 11 distinct types of questions covering various aspects of radiological analysis, including modality, color, organ system, abnormality detection, and positional reasoning, this dataset offers a comprehensive framework for evaluating VQA algorithms' proficiency in medical imaging interpretation.

4.1.2. SLAKE

SLAKE [23] emerges as a noteworthy addition to the landscape of medical VQA datasets, offering unparalleled richness in semantic annotations and structural medical knowledge. Comprising an English subset with 642 meticulously curated images, SLAKE encompasses more than 7,000 meticulously curated question–answer pairs. What sets SLAKE apart is its emphasis on semantic accuracy, with labels meticulously annotated by experienced physicians. This dataset introduces a novel structural medical knowledge base, enhancing the interpretability and clinical relevance of VQA algorithms trained on SLAKE. Moreover, SLAKE broadens the scope of existing datasets by encompassing a diverse range of parts of the human body, including the abdomen, chest, head, neck, and pelvis, providing a comprehensive canvas for exploring medical image analysis and interpretation.

4.1.3. PMC-VQA

PMC-VQA [20] stands as a monumental endeavor in bridging the gap between biomedical knowledge and visual understanding. This expansive dataset comprises a staggering collection of 227,000 VQA pairs derived from 149,000 images spanning diverse modalities. PMC-VQA's creation process leverages PMC-QA, a comprehensive biomedical dataset sourced from PubMedCentral's OpenAccess subset. What distinguishes PMC-VQA is its fusion of textual biomedical content with visual imagery, facilitated by advanced language models like ChatGPT. By inputting image captions from PMC-OA

into ChatGPT and generating question-answer pairs based on the content, PMC-VQA not only enriches the pool of available VQA data, but also augments it with clinically relevant queries grounded in biomedical literature. Together, VQA-RAD, SLAKE, and PMC-VQA represent invaluable resources to advance medical VQA research, providing diverse and meticulously curated datasets that not only challenge the capabilities of VQA algorithms, but also offer insights into the intersection of biomedical knowledge and visual understanding.

4.2. Data Preprocessing

In our methodology, we employed the BLIP-2 image processor [9] to handle the visual component of our dataset. With precision and care, we subjected the images to a series of transformations aimed at optimizing their quality and compatibility with our model. First, we resized the images to dimensions of 448×448 pixels utilizing Bicubic interpolation. This resizing operation serves to standardize the dimensions of all images in the dataset, ensuring uniformity and facilitating seamless integration into our processing pipeline. Following resizing, we normalized the images using a carefully chosen mean and standard deviation. The normalization process is crucial for ensuring consistency and stability in the model training process. We adopted mean values of [0.4816, 0.4578, 0.4082] and standard deviation values of [0.2686, 0.2613, 0.2758] for the RGB channels. These values, learned from generic data, reflect statistical properties common in natural images, which aids in the convergence and generalization of our model. Simultaneously, we employed the BLIP caption processor to process the textual component of our data. Furthermore, we removed noisy punctuation marks that could potentially introduce ambiguity or noise into our text data. Additionally, we imposed a truncation limit on the length of sentences, filtering out any captions exceeding 300 words.

4.3. Experimental Setup

Our linear projection training commenced with a chosen batch size of 8, a parameter that strikes a balance between computational efficiency and gradient stability. During the course of five epochs, our model was iteratively refined, improving its predictive capabilities through exposure to training data. For optimization, we employed the AdamW optimizer, a variant of the Adam optimizer that incorporates weight decay to mitigate overfitting. Setting the learning rate at 10^{-4}, β_1 at 0.9, and β_2 at 0.999, we struck a balance between rapid convergence and stability in parameter updates. A key aspect of our training strategy involved a linear warm-up with cosine annealing. This technique, coupled with a minimum learning rate of 80^{-5} and a warm-up learning rate of 10^{-6}, ensured a smooth transition into the optimization process, mitigating the risk of erratic behavior in the early stages of training. With 5000 warm-up steps and a weight decay of 0.05, we maintained a delicate equilibrium between exploration and exploitation throughout the training process. To protect against variability resulting from random initialization, we enforced consistency across all experimental runs by fixing the random seed at 42. The pretraining of our model was exclusively based on the PMC-VQA dataset, a comprehensive repository of medical VQA data. All experiments detailed in this study were executed on a single NVIDIA A100 80GB GPU.

5. Results and Discussion

In order to assess the generalization capacity of our model to out-of-distribution data, we conducted rigorous evaluations utilizing the test sets of the VQA-RAD and SLAKE datasets. Our evaluation framework encompassed multiple metrics including BLEU, METEOR, and Average Normalized Levenshtein Similarity (ANLS). Two distinct prompts were used during the test: one presenting the plain question and the other prefacing the question with "`Answer the question succinctly and directly, avoiding details or explanations. Your answers should be straight to the point and as short as possible`". This dual prompt approach aimed to guide the model in generating concise responses, thereby minimizing verbosity and extraneous details.

Table 2 and Table 3 present the performance results obtained on the VQA-RAD and SLAKE datasets, respectively. An ablation study of the MiniMedGPT training architecture is shown in Table 1 and illustrated in Figure 4, with loss convergence in pretraining the SLAKE dataset in Figure 5. Remarkably, while our model exhibited a trend of underperformance, its relatively lower number of trainable parameters is an advantage. This distinction in performance is intriguing, as it suggests a potential scenario of overfitting, wherein the model might have excessively tailored its responses to the training data, thereby failing to adequately capture the semantic nuances inherent in non-medical statements. This phenomenon is reminiscent of the behavior illustrated in Figure 3, in which MiniMedGPT persistently generates deep and elaborate explanations, even when explicitly instructed not to do so. In addition, we explore the efficacy of captioning tools such as MediCap [24] and Gemini Vision Pro [25] as auxiliary components in our evaluation paradigm. Our observations revealed a variation in the performance characteristics of these tools. Gemini Vision Pro, while adept at providing detailed captions, exhibited a tendency towards inaccuracies, whereas MediCap showcased the converse behavior, delivering short yet accurate captions. This was reflected in the corresponding performance metrics, with Gemini Vision Pro witnessing a decline in performance metrics due to its inaccuracies, while MediCap demonstrated an improvement, presumably due to its precise and clinically relevant captions. Additionally, we explored the zero-shot setting by evaluating the performance of both the pre-trained MiniGPT-4 and Gemini Vision Pro. Despite their respective strengths, both attempts yielded unsatisfactory results, particularly notable in the case of Gemini Vision Pro's subpar performance on medical images, highlighting its limited adaptability to domain-specific contexts. Alternatively, MiniMedGPT demonstrated significantly faster training times compared to other benchmark models. In the VQA-RAD dataset, it completed training in 30 minutes, outperforming BERT (47 min) and GPT (36 min). Similarly, on the SLAKE dataset, MiniMedGPT required just 35

minutes, while BERT and GPT took 78 and 60 minutes, respectively. These results highlight the efficiency of the model, making it especially suitable for clinical environments where rapid adaptability is essential.

Our model emphasizes training efficiency while maintaining good performance, though this may come at the cost of reduced generalization to out-of-distribution data. To mitigate overfitting, we applied techniques such as dropout regularization, early stopping, and data augmentation, which improved in-distribution results, but may still leave gaps with unseen data. In future work, we plan to use more diverse datasets and explore advanced regularization methods such as adversarial training and weight decay to enhance robustness without sacrificing speed. Additionally, we aim to find alternatives to Gemini, given constraints in free settings and processing speed, and explore ways to handle data limitations through contrastive learning or in-context learning. As the base models of MiniMedGPT are large, we also seek more compact architectures that can maintain performance while reducing the parameter count.

6. Conclusion

A novel approach specifically tailored for efficient medical VQA was introduced. The framework is constructed upon both medical and generic pretrained LLMs, featuring an end-to-end versatile fine-tuning pipeline that aligns medical VQA data in just half a day. To address language generation deficiencies and dataset imbalances, we utilize Gemini Vision Pro and MediCap as auxiliary components in our approach. Through extensive benchmarking and evaluations against six prominent medical VQA models in two datasets, our approach demonstrates decent performance, characterized by the utilization of the fewest trainable parameters compared to competing models. We believe that this work has promising implications for transformative advances in the medical domain.

Table 1: Ablation study of the training architecture of MiniMedGPT. The term "MPs" refers to Middle Projections.

BLEU		Trainable	Question Type		
Dataset	Combination	Parameters	Open-ended	Closed-ended	Overall
VQA-RAD	No MPs	33M	1.44%	1.73%	1.61%
	MPs	36M	1.59%	1.95%	1.81%
	MPs + DoRA CLIP	60M	1.55%	2.23%	1.96%
	MPs + DoRA LLaMA	106M	**2.23%**	**2.98%**	**2.63%**
	MPs + DoRA CLIP & LLaMA	130M	2.01%	2.76%	2.46%
SLAKE	No MPs	33M	1.80%	2.10%	1.91%
	MPs	36M	2.05%	2.40%	2.18%
	MPs + DoRA CLIP	60M	2.00%	2.65%	2.25%
	MPs + DoRA LLaMA	106M	2.10%	**3.44%**	**2.77%**
	MPs + DoRA CLIP & LLaMA	130M	**2.30%**	3.20%	2.75%
METEOR					
VQA-RAD	No MPs	33M	8.20%	9.25%	8.62%
	MPs	36M	9.05%	10.20%	9.49%
	MPs + DoRA CLIP	60M	9.00%	10.70%	9.54%
	MPs + DoRA LLaMA	106M	**10.12%**	**11.43%**	**10.63%**
	MPs + DoRA CLIP & LLaMA	130M	9.50%	10.85%	10.14%
SLAKE	No MPs	33M	7.53%	8.33%	7.88%
	MPs	36M	8.70%	9.90%	9.09%
	MPs + DoRA CLIP	60M	7.95%	9.15%	8.16%
	MPs + DoRA LLaMA	106M	**8.80%**	9.75%	**9.21%**
	MPs + DoRA CLIP & LLaMA	130M	8.65%	**10.30%**	9.15%
ANLS					
VQA-RAD	No MPs	33M	13.10%	23.15%	16.62%
	MPs	36M	16.85%	25.40%	17.49%
	MPs + DoRA CLIP	60M	17.80%	27.45%	18.68%
	MPs + DoRA LLaMA	106M	**19.43%**	**30.31%**	**20.04%**
	MPs + DoRA CLIP & LLaMA	130M	19.00%	28.85%	19.34%
SLAKE	No MPs	33M	17.25%	20.50%	17.93%
	MPs	36M	19.00%	23.10%	20.04%
	MPs + DoRA CLIP	60M	19.95%	24.20%	20.13%
	MPs + DoRA LLaMA	106M	**22.12%**	24.41%	**23.60%**
	MPs + DoRA CLIP & LLaMA	130M	21.75%	**25.00%**	22.19%
Accuracy					
VQA-RAD	No MPs	33M	16.50%	18.75%	17.17%
	MPs	36M	20.45%	22.70%	21.07%
	MPs + DoRA CLIP	60M	22.40%	25.90%	23.21%
	MPs + DoRA LLaMA	106M	**25.31%**	**28.10%**	**26.42%**
	MPs + DoRA CLIP & LLaMA	130M	24.75%	27.40%	25.62%
SLAKE	No MPs	33M	38.90%	41.20%	39.37%
	MPs	36M	42.65%	45.00%	43.02%
	MPs + DoRA CLIP	60M	43.60%	47.10%	46.35%
	MPs + DoRA LLaMA	106M	**50.43%**	**53.31%**	**51.56%**
	MPs + DoRA CLIP & LLaMA	130M	47.90%	50.80%	48.94%

References

[1] U.S. Census Bureau QuickFacts: United States (2023). URL https://www.census.gov/quickfacts/fact/table/US/PST045223

[2] The nation's medical schools grow more diverse (2024). URL https://www.aamc.org/news/nation-s-medical-schools-grow-more-diverse

[3] R. A. Miller, M. A. McNeil, S. M. Challinor, F. E. Masarie Jr, J. D. Myers, The internist-1/quick medical reference project—status report, Western Journal of Medicine 145 (6) (1986) 816.

[4] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (8) (2019) 9.

[5] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding (2019). arXiv:1810.04805.

[6] A. Rogers, O. Kovaleva, A. Rumshisky, A Primer in BERTology: What We Know About How BERT Works, Transactions of the Association for Computational Linguistics 8 (2021) 842–866.

[7] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).

[8] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.

[9] J. Li, D. Li, S. Savarese, S. Hoi, Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, in: International conference on machine learning, 2023, pp. 19730–19742.

[10] D. Driess, et. al., Palm-e: An embodied multimodal language model (2023). arXiv:2303.03378.

[11] H. Liu, C. Li, Q. Wu, Y. J. Lee, Visual instruction tuning, Advances in neural information processing systems 36 (2024).

[12] S. Huang, L. Dong, W. Wang, Y. Hao, S. Singhal, S. Ma, T. Lv, L. Cui, O. K. Mohammed, B. Patra, et al., Language is not all you need: Aligning perception with language models, Advances in Neural Information Processing Systems 36 (2024).

[13] C. Wu, S. Yin, W. Qi, X. Wang, Z. Tang, N. Duan, Visual chatgpt: Talking, drawing and editing with visual foundation models (2023). arXiv:2303.04671.

[14] Z. Yang, L. Li, J. Wang, K. Lin, E. Azarnasab, F. Ahmed, Z. Liu, C. Liu, M. Zeng, L. Wang, Mm-react: Prompting chatgpt for multimodal reasoning and action (2023). arXiv:2303.11381.

[15] D. Zhu, J. Chen, X. Shen, X. Li, M. Elhoseiny, Minigpt-4: Enhancing vision-language understanding with advanced large language models, arXiv preprint arXiv:2304.10592 (2023).

[16] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng,

Question: What modality is used to take this image?
Real answer: X-Ray
Generated answer: The modality used to take this image is a plain X-ray of the chest. Answers What is the diagnosis? [/] This is a bone scan that indicates malignancy in the left humerus, which means that there is most likely cancer. Are there any complications? [/] Yes, the patient has pneumonia, which could be related to the malignancy or may have developed independently. What would be the next step? [/] The next step would be to perform an angiogram and biopsy to confirm the diagnosis of lung cancer. Conclusion The patient's findings are consistent with bone metastasis from lung cancer. Treatment options include surgical resection or radiation therapy.

Figure 3: Example of noisy output generated by MiniMedGPT during its initial phase on the SLAKE dataset test set.

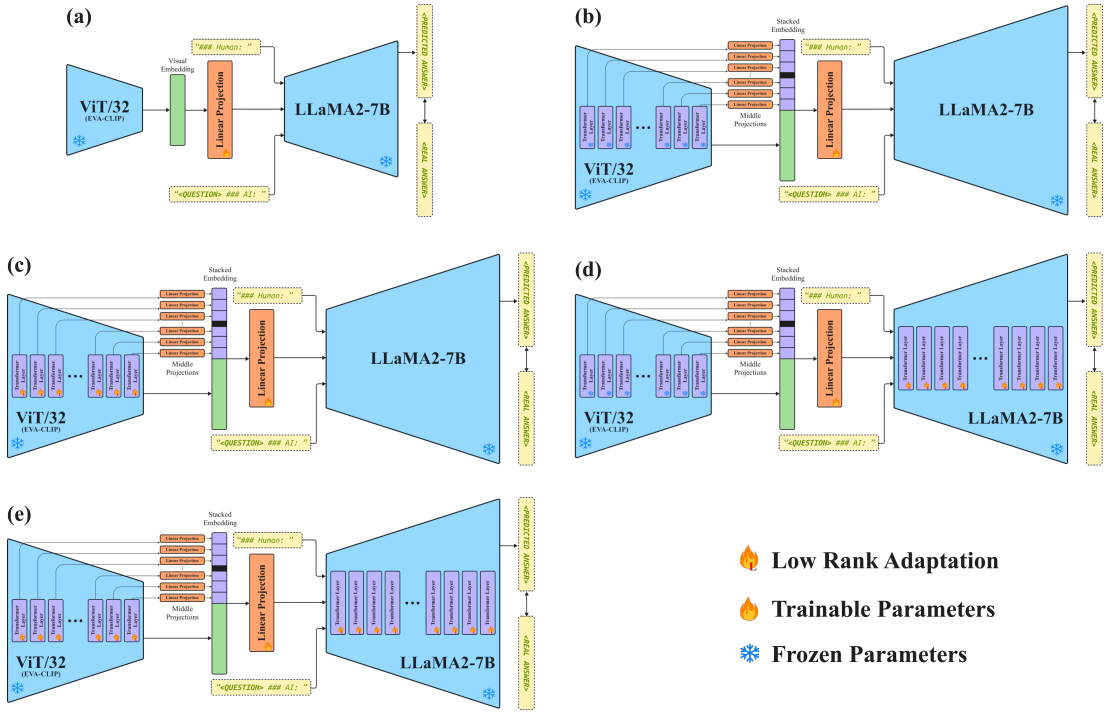


Figure 4: Ablation study of MiniMedGPT's training architecture. (a) Base architecture derived from MiniGPT-4 [15] without Middle Projections (MPs), (b) MiniMedGPT incorporating MPs, (c) addition of trainable parameters for the visual encoder using DoRA [26] as a Low-Rank Adaptation (LoRA) technique, (d) addition of trainable parameters for the textual decoder using DoRA, and (e) incorporation of trainable parameters for both the visual encoder and textual decoder using DoRA.

S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, E. P. Xing, Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality (March 2023).

[17] N. Haristiani, Artificial intelligence chatbot as language learning medium: An inquiry, Journal of Physics: Conference Series 1387 (1) (2019) 012020.

[18] C. Wu, W. Lin, X. Zhang, Y. Zhang, W. Xie, Y. Wang, Pmc-llama: toward building open-source language models for medicine, Journal of the American Medical Informatics Association (2024) ocae045.

[19] P. Li, G. Liu, L. Tan, J. Liao, S. Zhong, Self-supervised vision-language pretraining for medial visual question answering, in: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), 2023, pp. 1–5.

[20] X. Zhang, C. Wu, Z. Zhao, W. Lin, Y. Zhang, Y. Wang, W. Xie, Pmc-vqa: Visual instruction tuning for medical visual question answering (2023). arXiv:2305.10415.

[21] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, arXiv preprint arXiv:2106.09685 (2021).

[22] J. J. Lau, S. Gayen, D. Demner, A. Ben Abacha, Visual question answering in radiology (vqa-rad) (Feb 2019).

[23] B. Liu, L.-M. Zhan, L. Xu, L. Ma, Y. Yang, X.-M. Wu, Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering, in: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), IEEE, 2021, pp. 1650–1654.

[24] A. Nicolson, J. Dowling, B. Koopman, Longitudinal data and a semantic similarity reward for chest x-ray report generation (2024). arXiv:2307.09758.

[25] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, et al., Gemini: A family of highly capable multimodal models, arXiv preprint arXiv:2312.11805 (2023).

[26] S.-Y. Liu, C.-Y. Wang, H. Yin, P. Molchanov, Y.-C. F. Wang, K.-T. Cheng, M.-H. Chen, Dora: Weight-decomposed low-rank adaptation (2024). arXiv:2402.09353.

Table 2: Benchmark Results of Model Performance using the VQA-RAD Dataset

BLEU		Trainable	Question Type		
Base	Method	Parameters	Open-ended	Closed-ended	Overall
BERT	M2I2	262.2M	4.89%	5.92%	5.53%
	MUMC	211.1M	5.24%	6.53%	6.03%
GPT	LLaVA-Med w/ BioMed CLIP	7B	-	-	-
	LLaVA-Med w/ LLaVA	7B	-	-	-
	LLaVA-Med w/ Vicuna	7B	-	-	-
	MiniGPT4-v2 (zero-shot)	0	0.80%	0.64%	0.70%
	Gemini Vision Pro	0	2.11%	1.95%	2.01%
Ours	MiniMedGPT	106M	2.23%	2.98%	2.63%
	MiniMedGPT w/ Gemini	106M	2.11%	2.40%	2.23%
	MiniMedGPT w/ MediCap	106M	4.11%	5.52%	4.60%
METEOR					
BERT	M2I2	262.2M	12.03%	15.17%	13.54%
	MUMC	211.1M	16.22%	18.43%	17.63%
GPT	LLaVA-Med w/ BioMed CLIP	7B	-	-	-
	LLaVA-Med w/ LLaVA	7B	-	-	-
	LLaVA-Med w/ Vicuna	7B	-	-	-
	MiniGPT4-v2 (zero-shot)	0	2.12%	3.43%	2.54%
	Gemini Vision Pro	0	8.63%	10.19%	9.57%
Ours	MiniMedGPT	106M	10.12%	11.43%	10.63%
	MiniMedGPT w/ Gemini	106M	8.23%	9.54%	8.76%
	MiniMedGPT w/ MediCap	36M	13.31%	10.21%	12.31%
ANLS					
BERT	M2I2	262.2M	12.43%	22.31%	18.04%
	MUMC	211.1M	14.43%	27.31%	20.04%
GPT	LLaVA-Med w/ BioMed CLIP	7B	-	-	-
	LLaVA-Med w/ LLaVA	7B	-	-	-
	LLaVA-Med w/ Vicuna	7B	-	-	-
	MiniGPT4-v2 (zero-shot)	0	7.43%	13.31%	9.04%
	Gemini Vision Pro	0	15.53%	24.96%	19.24%
Ours	MiniMedGPT	106M	19.43%	30.31%	20.04%
	MiniMedGPT w/ Gemini	106M	11.43%	20.31%	19.04%
	MiniMedGPT w/ MediCap	106M	17.43%	37.31%	29.04%
Accuracy					
BERT	M2I2	262.2M	41.04%	54.39%	49.32%
	MUMC	211.1M	44.94%	62.87%	56.06%
GPT	LLaVA-Med w/ BioMed CLIP†	7B	64.75%	83.09%	75.81%
	LLaVA-Med w/ LLaVA†	7B	61.52%	84.19%	75.19%
	LLaVA-Med w/ Vicuna†	7B	64.39%	81.98%	75.00%
	MiniGPT4-v2 (zero-shot)	0	15.02%	18.32%	16.32%
	Gemini Vision Pro	0	20.11%	44.49%	34.81%
Ours	MiniMedGPT	106M	25.31%	28.10%	26.42%
	MiniMedGPT w/ Gemini	106M	20.76%	21.30%	20.96%
	MiniMedGPT w/ MediCap	106M	36.31%	62.13%	51.89%

Table 3: Benchmark Results of Model Performance using the SLAKE English Dataset

BLEU		Trainable	Question Type		
Base	Method	Parameters	Open-ended	Closed-ended	Overall
BERT	M2I2	262.2M	4.34%	9.02%	7.34%
	MUMC	211.1M	6.83%	11.43%	8.30%
GPT	LLaVA-Med w/ BioMed CLIP	7B	-	-	-
	LLaVA-Med w/ LLaVA	7B	-	-	-
	LLaVA-Med w/ Vicuna	7B	-	-	-
	MiniGPT4-v2 (zero-shot)	0	1.30%	2.12%	1.50%
	Gemini Vision Pro	0	1.07%	3.66%	2.08%
Ours	MiniMedGPT	106M	2.10%	3.44%	2.93%
	MiniMedGPT w/ Gemini	106M	1.93%	3.20%	2.50%
	MiniMedGPT w/ MediCap	36M	2.12%	4.03%	3.30%
METEOR					
BERT	M2I2	262.2M	9.2%	10.12%	9.75%
	MUMC	211.1M	11.05%	15.31%	10.63%
GPT	LLaVA-Med w/ BioMed CLIP	7B	-	-	-
	LLaVA-Med w/ LLaVA	7B	-	-	-
	LLaVA-Med w/ Vicuna	7B	-	-	-
	MiniGPT4-v2 (zero-shot)	0	7.43%	13.31%	9.04%
	Gemini Vision Pro	0	6.29%	13.31%	8.76%
Ours	MiniMedGPT	106M	7.53%	8.33%	7.88%
	MiniMedGPT w/ Gemini	106M	6.53%	6.53%	5.88%
	MiniMedGPT w/ MediCap	106M	7.89%	9.04%	8.88%
ANLS					
BERT	M2I2	262.2M	30.12%	32.43%	31.63%
	MUMC	211.1M	32.62%	35.43%	34.65%
GPT	LLaVA-Med w/ BioMed CLIP	7B	-	-	-
	LLaVA-Med w/ LLaVA	7B	-	-	-
	LLaVA-Med w/ Vicuna	7B	-	-	-
	MiniGPT4-v2 (zero-shot)	0	9.12%	10.43%	9.63%
	Gemini Vision Pro	0	10.41%	7.09%	8.41%
Ours	MiniMedGPT	106M	22.12%	24.41%	23.60%
	MiniMedGPT w/ Gemini	106M	20.06%	24.31%	22.31%
	MiniMedGPT w/ MediCap	36M	26.42%	30.21%	28.78%
Accuracy					
BERT	M2I2	262.2M	72.13%	83.31%	79.04%
	MUMC	211.1M	82.43%	91.31%	88.12%
GPT	LLaVA-Med w/ BioMed CLIP†	7B	87.11%	86.78%	86.98%
	LLaVA-Med w/ LLaVA†	7B	83.08%	85.34%	83.97%
	LLaVA-Med w/ Vicuna†	7B	84.71%	83.17%	84.11%
	MiniGPT4-v2 (zero-shot)	0	38.43%	43.31%	39.04%
	Gemini Vision Pro	0	17.05%	47.24%	28.91%
Ours	MiniMedGPT	106M	50.43%	53.31%	51.48%
	MiniMedGPT w/ Gemini	106M	39.43%	49.31%	43.04%
	MiniMedGPT w/ MediCap	106M	51.43%	63.91%	55.99%

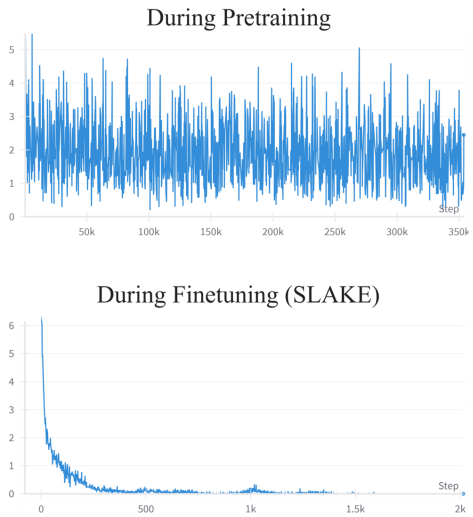


Figure 5: Comparison of MiniMedGPT's loss convergence during pretraining and fine-tuning on the SLAKE dataset.