

Text-Enhanced Medical Visual Question Answering

Chih-Ying Liu^{1*}, Fan Diao^{1*}

¹Stanford University

{yingl029, fdiao}@stanford.edu

Abstract

Medical VQA is a task that requires both computer vision and natural language processing and is critical in clinical decision-making. In this project, we investigated the impact of augmenting additional medical textual knowledge and various designs of fusion modules in the medical VQA system. We conducted our experiments on the VQA-RAD dataset with both binary and multi-class settings. We observed that transformer-based fusion modules perform better than linear fusion and that text enhancement is helpful in binary classification tasks and when the training data is limited. We discussed how imbalanced training labels influence our systems and when text enhancement improves performance. Our best model achieves 81.7% accuracy on closed-ended questions and 66.7% accuracy on all questions.

1. Introduction

Motivation Leveraging multimodal medical data, including visual and text information, to improve critical medical tasks presents an interesting yet challenging opportunity. Medical Visual Question Answering (VQA) is one of the critical medical tasks. VQA [3] is a multidisciplinary problem that combines computer vision and natural language processing. Medical VQA helps to assist in clinical decision-making by answering an image-related question according to the image content [11]. A medical VQA system can potentially answer the physician’s questions and help improve the efficiency of medical professionals [20].

Problem This project aims to gain valuable insights into effectively leveraging multimodal medical data for healthcare. Specifically, we focus on the medical VQA problem and investigate the two topics:

1. **Augmenting with additional textual knowledge:** We aim to investigate whether enhancing the text encoder of the model with additional textual inputs, such as

real-world medical entrance exam questions, can enhance the accuracy of the medical VQA system.

2. **Impact of Multimodal Fusion Modules:** We explore different fusion modules for combining the image features and text features to improve the performance of the medical VQA system.

Input and Output We formulate medical VQA as a classification task. The input to our algorithm is an image and a question. We then use a model consisting of a CLIP [26] base model, a multimodal fusion module and a classifier, to output a predicted answer from a pre-defined set of answer candidates.

Training Training consists of two stages: additional text fine-tuning and MedVQA fine-tuning.

1. **Stage 1: Additional Medical Text Fine-tuning:** We use MedMCQA [24] to fine-tune the text encoder of CLIP. This stage applies to methods with text enhancement and is not needed for methods without text enhancement. The model takes one question and 4 options as inputs and predicts scores for each option.
2. **Stage 2: MedVQA fine-tuning:** We use VQA-RAD [16] to fine-tune the CLIP. This stage applies to all methods, including the baseline method defined in the Section 3. The model takes an image and a question as inputs and predicts scores for each answer candidate. We categorize questions as close-ended if their answer is "yes" or "no"; otherwise, they are open-ended. For open-ended questions, answer candidates are defined by the training set, and our model predicts a score for each candidate.

Experiments We conducted a series of experiments to evaluate the impact of additional medical text fine-tuning on model performance and to compare the performance of different multimodal fusion modules. We built models on top of various CLIP base models (CLIP [26], PubMed-CLIP [8], PMC-CLIP [19]) and experimented under two

*These authors contributed equally to this work

distinct settings: binary classification on close-ended questions only and multi-class classification on all (open-ended and closed-ended) questions.

Overview of the Results The highest accuracy on binary classification tasks, 81.7%, is achieved by the text-enhanced CLIP with transformer decoder fusion. The highest overall accuracy on multi-classification tasks is 66.7%. It's achieved by CLIP with transformer encoder fusion. Text enhancement demonstrates the largest improvement of 3.6% accuracy on binary classification tasks. Transformer-based fusion module shows the highest increase of 4% accuracy on binary classification tasks and over 20% overall accuracy on multi-class classification tasks.

2. Related Work

Medical CLIP Vision-and-language tasks, such as visual question answering and image-text retrieval, require the systems to understand both the visual and text contents. Vision-text contrastive learning like CLIP [26] is trained to match image and text pairs while pulling others apart. The joint training on large-scale image-text pairs generates effective multimodal features that can support both vision-only and vision-language downstream tasks. However, applying CLIP to the medical domain is not a trivial task because publicly available medical data is orders of magnitude lower than general domain data, and medical images contain subtle and fine-grained key point features, such as the difference between "pneumonia" and "consolidation". PubMedCLIP [8] fine-tune CLIP on ROCO [25], a medical image-text dataset containing 80k examples. MedCLIP [30] overcomes data limitation problems by decoupling images and texts for contrastive learning to make full usage of all image-only, text-only, and image-text medical data. Recent work proposed larger medical image-text datasets to further improve the effectiveness of contrastive pre-training. PMC-CLIP [19] proposed and contrastively pre-trained a CLIP-based model on the PMC-OA dataset, which contains 1.6M examples. Furthermore, BiomedCLIP [32] proposed and pre-trained their models on their PMC-15M dataset, containing 15 million examples. Models pre-trained on large-scale and diverse datasets learn better representations for downstream tasks. We are going to experiment with both CLIP that is pre-trained on general image-text pairs and medical CLIP that is fine-tuned or pre-trained on medical image-text pairs in this project.

Medical VQA Image-text retrieval, classification, and vision question answering are common downstream tasks used to evaluate the performance of a pre-trained vision-language model. In this project, we focus on medical VQA. The answers in medical VQA could be either close-end,

such as Yes/No, or open-end. Medical VQA can be formulated as a classification or a generative problem.

Most existing methods [8, 19, 32, 5, 13] formulates medical VQA as a classification problem. Given a question and an image, the model is required to select an answer from a predefined set of answer candidates. The advantage of such approaches is that the complexity of the task is reduced by treating medical VQA as a classification task. However, these approaches struggle to accurately predict open-ended questions, since these answers are more varied and have low frequency than close-ended answers. General VQA frameworks include an image encoder, a text encoder, a multimodal fusion module, and a classifier. Due to the small scale of medical VQA dataset, a common approach is to first contrastively pre-train CLIP-based models on large-scale medical image caption datasets, then fine-tune on downstream medical VQA datasets. PubMedCLIP [8] incorporates the vision encoder of PubMedCLIP into two VQA frameworks – MEVF (Mixture of Enhanced Visual Features)[22] and QCR (question-conditioned reasoning) [31], and fine-tune on medical VQA datasets. PMC-CLIP [19] uses a self-attention transformer as a fusion module which takes image features and text features as inputs and predicts the masked tokens jointly with contrastive pre-training. It is applied to medical VQA tasks by computing similarity between output embeddings from the fusion module and text encodings of each answer candidate. BiomedCLIP [32] uses METER (Multimodal End-to-end Transformer) [6] framework, which is a transformer-based co-attention multimodal fusion module that produces cross-modal representations over the image and text features, which are then fed into a classifier.

On the other hand, with the development of transformers and large language models, there has been more work [2, 15, 17, 33, 12, 28] that recently approaches medical VQA as a generative task. These approaches allow image-question features to interact with step-by-step answer predictions, and therefore may enhance long answer generation. However, they often generate many non-existent answers, resulting in low accuracy, which is why they are not yet the mainstream in medical VQA. Q2ATransformer [21] is a classification-based method that integrates advantages of generation-based methods. It uses an answer-querying decoder that performs cross attention between learnable answers embedding and image-text features, where the model can interact with answer information like generation-based approaches. MUMC [17] uses a multimodal encoder to incorporate image and text features, which is then fed into an answer decoder to generate answers. It selects the final answer by comparing the similarity of the generated answer with a set of answer candidates. Recently, there are more works that leverage large language models [23, 27] in medical VQA. PMC-VQA [33] aligns visual information from a

pre-trained vision encoder with a large language model by a transformer-based multimodal decoder. PeFoMed [12] utilizes the pre-trained weights of a general domain LLM and ViT [10], and fine-tune them by only updating the vision projection layer and the low-rank adaptation layer (LoRA) [14]. It achieves the state-of-the-art performance on the VQA-RAD [16] dataset with overall accuracy 81.6%.

3. Method

Figure 1 illustrates our approach. We dissected our model designs into three dimensions: base CLIP model, multimodal fusion module, with or without text enhancement. Training contained two stages: additional text fine-tuning and MedVQA fine-tuning.

3.1. Additional Text Fine-tuning

We fine-tune the text encoder of CLIP on the MedM-CQA dataset. Since each sample contains 1 question and 4 options, we concatenated the question with each option to produce 4 input texts. We added a single linear layer classifier on top of the text encoder to predict scores for the 4 options.

We use cross-entropy loss for multi-class classification:

$$L = - \sum_{i=1}^N \log \left(\frac{e^{s_{y_i}}}{\sum_j e^{s_j}} \right) \quad (1)$$

where N is the number of samples, s_j is the predicted score of the class j , and y_i is the correct class for example i .

3.2. MedVQA Fine-tuning

We experimented with two settings: (1) binary classification tasks, where the model handles only close-ended questions, and (2) multi-class classification tasks, where the model handles both close-ended and open-ended questions.

Given an image $\mathbf{I} \in \mathbb{R}^{H \times W \times C}$, a question encoding $\mathbf{Q} \in \mathbb{R}^{L \times M}$, where L is the sequence length and M is the text embedding dimension:

1. **Binary classification models** output a scalar score $s \in \mathbb{R}$ and use the binary cross-entropy loss:

$$L = - \frac{1}{N} \sum_{i=1}^N [y_i \log(s_i) + (1 - y_i) \log(1 - s_i)] \quad (2)$$

where y_i is the correct class for example i .

2. **Multi-class classification models** output a score vector $\mathbf{s} \in \mathbb{R}^C$, where C is the number of labels, and use multi-class cross-entropy loss:

$$L = - \sum_{i=1}^N \log \left(\frac{e^{s_i}}{\sum_j e^{s_j}} \right) \quad (3)$$

Our system consists of three components: (1) CLIP base model, (2) multimodal fusion module, and (3) classifier.

CLIP Base Model CLIP [26] consists of (1) an vision encoder that encodes images into embeddings $\mathbf{v} \in \mathbb{R}^D$ where D is the projected dimension and (2) an text encoder that encodes captions into embeddings $\mathbf{t} \in \mathbb{R}^D$, the same dimension as image features.

Given a batch of N text and image pairs, CLIP is pre-trained to maximize the cosine similarity of the embeddings of N real pairs, while minimizing the cosine similarity of the $N^2 - N$ incorrect pairings. The scaled cosine similarity a_{ij} between image i and text j is

$$a_{ij} = (\tilde{v}_i \cdot \tilde{t}_j) e^t \quad (4)$$

where \tilde{v}_i and \tilde{t}_j are normalized visual and text encodings, and t is a learnable temperature. The probability of predicting text j given image i is computed by normalizing across j by softmax,

$$p_{ij}^{v \rightarrow t} = \frac{e^{a_{ij}}}{\sum_j e^{a_{ij}}} \quad (5)$$

and the reversed text-to-image probability is computed by normalizing across i . The pre-training loss is the symmetric cross entropy loss, which includes an image-to-text term,

$$\mathcal{L}^{v \rightarrow t} = - \frac{1}{N} \sum_{i=1}^N - \log \frac{e^{a_{ii}}}{\sum_j e^{a_{ij}}} \quad (6)$$

Similarly, we can compute $\mathcal{L}^{t \rightarrow v}$ and then reach to

$$\mathcal{L} = \frac{\mathcal{L}^{v \rightarrow t} + \mathcal{L}^{t \rightarrow v}}{2} \quad (7)$$

as the final image-text contrastive (ITC) loss.

We experimented with two CLIP base models (1) CLIP [26] and (2) PubMedCLIP [8].

- **CLIP** [1] The visual encoder is a Vision Transformer (ViT) [10], and the text encoder is a Transformer [29]. The model is pre-trained on 400 million image-text pairs in the general domain.
- **PubMedCLIP** [7] PubMedCLIP is derived from CLIP by fine-tuning it on the ROCO [25] dataset, which consists of 80,000 medical image-caption pairs.

To investigate whether adding additional text knowledge improves the performance, we compared the performance of using default text encoder from base models and the text encoder that is additionally fine-tuned with medical question answering mentioned in Section 3.1.

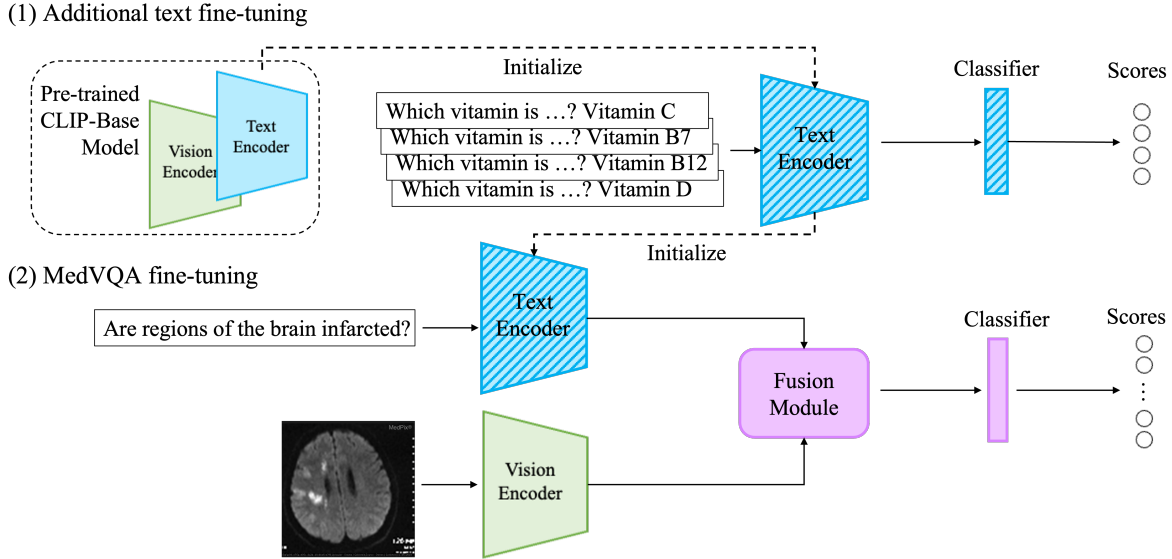


Figure 1. Summary of our method. Our method includes two stage training: additional text fine-tuning on MedMCQA and medical VQA fine-tuning on VQA-RAD. In the second stage, the multimodal fusion module generates a multimodal representation from text and image encodings, which are then fed into a classifier to predict scores.

Multimodal Fusion Module A fusion module incorporates image and text features into combined image-text features, which are then fed into a classifier. We experiment with CLIP and PubMedCLIP using three types of fusion modules: (1) linear, (2) transformer encoder, and (3) transformer decoder.

- **Linear** A linear fusion module simply concatenates the two 512-dim image and text features from the projection layers.
- **Transformer Encoder** We concatenate the image features $\mathbf{v} \in \mathbb{R}^{L_v \times M}$ and text features $\mathbf{t} \in \mathbb{R}^{L_t \times M}$ together as $\mathbf{f} \in \mathbb{R}^{(L_v+L_t) \times M}$, where M is the hidden state feature dimension. Then, feed the concatenated features \mathbf{f} into a two-layer transformer encoder [29]. Each transformer encoder layer contains a multi-head self-attention module to compute the relation between each pair of features.

From the transformer outputs, we select the feature at the [EOS] token [9] to obtain the 512-dim image-text feature, which is then fed into the classifier.

- **Transformer Decoder** Instead of concatenating image features and text features together, here we use a two-layer transformer decoder [29] to learn the cross-attention between them. Each transformer decoder layer consists of a multi-head self-attention module followed by a multi-head cross-attention module. The text features are first fed into the self-attention module to compute the relation between each pair of text features. Then the cross-attention module takes (1) the

self-attention outputs as the query and (2) the image features as the key and the value, to compute the relation between text features and image features.

Similarly, we take the feature at the [EOS] token to feed into the classifier.

We implemented the transformer encoder and decoder based on transformer code from CS231N course assignment 3[4]. For the encoder, we removed the cross-attention layer. For the decoder, we use the cross-attention layer without masking.

PMC-CLIP In addition to combining pre-trained CLIP and PubMedCLIP models with fusion modules, we also adapted PMC-CLIP [18], which features a fusion module jointly pre-trained with visual and text encoders on large-scale image-text pairs.

The PMC-CLIP model includes a ResNet-based visual encoder, a BERT-based text encoder, and a self-attention Transformer-based fusion module. The fusion module takes concatenated text features $\mathbf{t} \in \mathbb{R}^{L_t \times M}$ and image features $\mathbf{v} \in \mathbb{R}^M$ as inputs and applies self-attention, where M is the projected dimension and L_t is the maximum sequence length. The text features are linearly transformed from the last hidden states.

The pre-training loss comprises an image-text contrastive loss, defined by Equation 7, and a masked language modeling (MLM) loss. The input text tokens are randomly masked with a probability of 15%, and the outputs of the fusion module are fed into an MLM projection layer to predict the masked tokens.

We take the last hidden states from the fusion module and apply average pooling over the time sequence to obtain the 768-dim image-text feature, which are then fed into the classifier.

Classifier We used a two-layer fully connected network as a classifier, which takes image-text features as inputs and outputs either (1) a scalar for binary classification tasks or (2) a C -dim score vector for multi-class classification tasks, where C is the number of labels.

3.3. Baseline

We considered the model that uses CLIP pre-trained on general domain with linear fusion as our baseline.

4. Dataset

VQA-RAD For the medical VQA task, we utilize the VQA-RAD dataset [16] as shown in Figure 2. This dataset contains 315 images and 3,515 corresponding questions, with each image linked to multiple questions. Questions are categorized as close-ended if the answer is “yes” or “no”, and otherwise as open-ended. The training set includes 458 answer candidates, where basic string processing maps semantically identical answers to the same label. We conduct experiments under two settings: using only close-ended questions and using all questions. The dataset is split into training, validation, and testing sets, containing 2,681, 383, and 408 question-answer pairs respectively. For close-ended questions, the training, validation, and testing sets contain 1498, 215, and 266 question-answer pairs respectively. Images are reshaped to 224×224 and normalized.

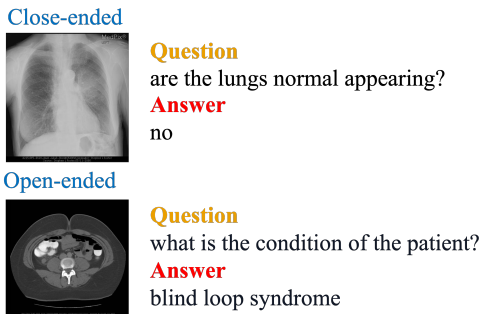


Figure 2. Examples of the VQA-RAD dataset.

MedMCQA We use the MedMCQA dataset [24], to fine-tune the text encoder of CLIP and enhance its medical knowledge. MedMCQA is a medical Multiple-Choice Question Answering (MCQA) dataset. Each sample includes a question text, 4 options, and the correct option (i.e. 1, 2, 3, 4). The original dataset contains 182822, 4183, and

6150 samples in the training, validation, and testing set respectively. The base CLIP models are pre-trained with a maximum token length of 77, thus, we filtered out any samples where concatenated input texts exceed the maximum token length. After filtering, there are 177605, 4088, and 6119 samples in the training, validation, and testing set respectively. Below is an example:

Question: Characteristic X Ray finding in ASD is:
Option A: Enlarged left ventricle
Option B: Enlarged left atria
Option C: Pulmonary pletheora
Option D: PAH
Answer: Option C

5. Experiments and Results

5.1. Additional Text Fine-tuning

Experiment Setting For fine-tuning on the MedMCQA dataset, we used AdamW as the optimizer with a learning rate of 5×10^{-6} , batch size 16, and trained for 3 epochs. We experimented with three base model initializations: CLIP, PubMedCLIP and PMC-CLIP.

Metric The primary metric is accuracy. Because the test set labels are not public, we report the accuracy on the validation set. We consider this case to be fine because the goal of this fine-tuning stage is to let the model learn additional medical knowledge but the final goal of this project is the performance of medical VQA discussed in Section 5.2.

Result The results are reported in Table 1. Our results are comparable to those of fine-tuned BERT-base, which achieved 35% accuracy, as reported by [24].

Base Model	MedMCQA (%)
CLIP	36.6
PubMedCLIP	37.0
PMC-CLIP	37.5

Table 1. Accuracy scores on MedMCQA validation set.

5.2. MedVQA Fine-tuning

Experiment Setting For fine-tuning on the VQA-RAD dataset, we used AdamW as the optimizer with a learning rate of 2×10^{-6} , which linearly decays, batch size 16, and trained for 20 epochs for binary tasks and 30 epochs for multi-class tasks. We store the model with the highest validation accuracy during training and reports scores on the test set.

Metric For experiments using only close-ended questions, we report the close-ended testing accuracy. For experiments using all questions, we report the overall, close-ended, and open-ended testing accuracy.

5.2.1 Binary Classification on Close-ended Questions

The results of the fine-tuning on close-ended questions of the VQA-RAD dataset are reported in Table 2. The baseline method, CLIP with linear fusion achieves 76.5% accuracy. The text-enhanced CLIP with transformer decoder fusion achieves the highest accuracy, 81.7%. When using CLIP as the base model, text enhancement consistently increases the accuracy of this task regardless of the fusion module used. Text enhancement also improves accuracy on PubMedCLIP combined with linear fusion. However, text enhancement doesn't show improvement on PubMedCLIP combined with transformer-based fusion modules or on PMC-CLIP.

Across all settings, transformer-based fusion modules consistently outperform the linear fusion modules. The largest improvements are a 4% accuracy increase from 76.5% (using CLIP with linear fusion) to 80.5% (using CLIP with transformer decoder), and also a 4% accuracy increase from 77.7% (using text-enhanced CLIP with linear fusion) to 81.7% (using text-enhanced with transformer decoder).

5.2.2 Multi-class Classification on All Questions

The results of fine-tuning on all questions of the VQA-RAD dataset are reported in Table 3. All transformer-based fusion methods outperform the baseline, CLIP with linear fusion. The highest overall accuracy, 66.7%, is achieved by CLIP with transformer encoder fusion. Classification on open-ended questions is particularly challenging due to the low frequency of many answers in the training dataset. While linear fusion fails completely on open-ended questions, transformer-based fusion methods achieve non-zero scores, with the highest open-ended accuracy of 48.4% achieved by the transformer encoder combined with PubMedCLIP. Comparing transformer-based fusion modules with and without pre-training, PMC-CLIP, which has a fusion module jointly pre-trained, performs worse than the from-scratch transformer-based fusion modules combined with CLIP-based models. This suggests that fine-tuning transformer-based fusion modules is sufficient to learn the relationship between image and text features. The suboptimal performance of PMC-CLIP may be due to the difference in capability between the visual encoders of CLIP/PubMedCLIP (ViT) and PMC-CLIP (ResNet). Additionally, text enhancement does not improve performance on this multi-class task, as shown in the binary task.

6. Discussion

6.1. Predictions over Different Types of Questions

In this section, we will analyze the predictions of four multi-class models on closed- and open-ended questions across various question types. The VQA-RAD dataset includes ten question types as shown in Figure 5. Figure 3 presents the predictions for both closed/open-ended questions and different question types. (a) For closed-ended questions, the CLIP + Linear model performs the worst on the "modality" question type, with an accuracy of 53.8%. For open-ended questions, PMC-CLIP and CLIP + Transformer Encoder/Decoder can predict correct answers for some questions, while CLIP tends to predict "yes" or "no," resulting in zero accuracy. (b) PMC-CLIP is weaker on the "abnormal" question type compared to CLIP + Transformer Encoder/Decoder. As shown in Figure 3(c) (d), common error occurs on "modality" and "abnormal" closed-ended questions, where all four models fail with probabilities of 64.3% and 55.6%, respectively.

An interesting observation in Figure 3 (d) is that when presented with a chest x-ray image and asked "what kind of image is this?", the CLIP + Transformer Encoder/Decoder provided the reasonable answer "chest x-ray," while the ground truth was "x-ray." This indicates that multiple reasonable labels can exist for a single question.

6.2. Reduce Overfitting by More Data

For fine-tuning on VQA-RAD, we initially used a data source containing 821, 119, and 251 close-ended question-answer pairs in the training, validation, and testing set. Table 4 shows that for the binary classification on close-ended questions, test accuracies are significantly lower than validation accuracies, with a gap ranging from 8.5% to 17.3%. The models have overfitted to the training and validation set. The likely reason is that the data size is relatively small.

To reduce the overfitting, we discovered a larger VQA-RAD data source containing 1498, 215, and 266 question-answer pairs in the training, validation, and testing set. Training and validation sets have 82.5% more data than the initial dataset. By fine-tuning with more data, table 5 shows that the gaps are smaller for the same task and setting. The table indicates that the models are less overfitted and generalize better.

6.3. Problem of Skewed Training Labels

In the multi-class classification task, although the CLIP + Linear Fusion model can get 75.5% on close-ended questions, surprisingly we saw a 0% accuracy on open-ended questions. We analyzed the problem, and we think the skewed training labels make the training more challenging. The training set has 3064 samples and 458 labels. "Yes" appears 829 times. "No" appears 884 times. However,

Model	Method		VQA-RAD (%) Closed
	Fusion	Text++	
CLIP	Linear	✗	76.5
		✓	77.7
	TransEnc	✗	77.3
		✓	80.9
	TransDec	✗	80.5
		✓	81.7
PubMed CLIP	Linear	✗	75.3
		✓	76.9
	TransEnc	✗	78.8
		✓	78.8
	TransDec	✗	80.5
		✓	78.1
PMC -CLIP	Pre-Trans	✗	80.5
		✓	76.5

Table 2. Accuracy scores on VQA-RAD closed-ended-questions testing set. TransEnc = Transformer Encoder, TransDec = Transformer Decoder, Pre-Trans = Pre-trained Transformer, Text++ = Text-Enhancement.

Model	Method		VQA-RAD (%)		
	Fusion	Text++	Overall	Closed	Open
CLIP	Linear	✗	45.8	75.5	0.0
		✓	44.9	72.9	0.0
	TransEnc	✗	66.7	78.5	47.8
		✓	64.5	78.9	41.4
	TransDec	✗	61.3	75.7	37.6
		✓	62.8	77.7	40.1
PubMed CLIP	Linear	✗	44.6	72.5	0.0
		✓	44.4	72.1	0.0
	TransEnc	✗	65.4	72.3	48.4
		✓	65.2	77.7	45.9
	TransDec	✗	65.7	80.1	42.7
		✓	63.7	78.9	40.1
PMC -CLIP	Pre-Trans	✗	58.1	78.1	26.1
		✓	55.6	76.9	21.7

Table 3. Accuracy scores on VQA-RAD all-questions testing set. TransEnc = Transformer Encoder, TransDec = Transformer Decoder, Pre-Trans = Pre-trained Transformer, Text++ = Text-Enhancement.

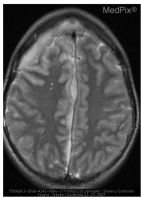



	(a)	(b)	(c)	(d)
				
Question	Can fluids be highlighted with this modality?	Where is the ascending colon?	Describe the lung abnormalities?	What kind of image is this?
Answer	yes	posterior to the appendix	pulmonary nodules	x-ray
CLIP + Linear	no ✗	no	✗ hemidiaphragm ✗	yes ✗
CLIP + TransEnc	yes ✓	posterior to the appendix	✓ bilateral ✗	chest x-ray ✗
CLIP + TransDec	yes ✓	posterior to the appendix	✓ yes ✗	chest x-ray ✗
PMC-CLIP	yes ✓	axial	✗ choroid plexus ✗	no ✗

Figure 3. Examples from the VQA-RAD dataset. The question types of four questions are respectively (a) modality, (b) position, (c) abnormality, and (d) modality.

Model	Method		VQA-RAD Binary (%)	
	Fusion	Text++	Test Acc	Val Acc
CLIP	Linear	✗	63.4	80.7
CLIP	Linear	✓	68.9	79.0
PubMedCLIP	Linear	✗	61.8	74.0
PubMedCLIP	Linear	✓	69.7	78.2

Table 4. Initial source: test accuracies vs validation accuracies. Text++ = Text-Enhancement

Model	Method		VQA-RAD Binary (%)	
	Fusion	Text++	Test Acc	Val Acc
CLIP	Linear	✗	76.5	76.6
CLIP	Linear	✓	77.7	82.6
PubMedCLIP	Linear	✗	75.3	78.3
PubMedCLIP	Linear	✓	76.9	76.7

Table 5. New source with more data: test accuracies vs validation accuracies. Text++ = Text-Enhancement

164 labels appear only once, and 120 labels appear only twice. The fine-tuned model tends to predict close-ended

(yes/no) answers rather than open-ended answers. Figure 4 (a) shows the distribution of answers against question types.

Out of 157 open-ended questions, the CLIP + Linear Fusion model only predicts 4 open-ended answers. The model predicts close-ended answers for all close-ended questions.

As a comparison, the CLIP + Transformer Encoder model achieves 47.8% accuracy on open-ended questions. The confusion matrix in figure 4 (b) shows that this model predicts open-ended answers for most of the open-ended questions (143 out of 157). It indicates the transformer encoder fusion module, by learning the relation between the image and the question, helps the model understand deeper about the open-ended questions and that mitigates the impact of skewed labels to some extent.

To further overcome the problem of skewed training labels, a future work (see Section 7) is to explore using weighted loss and assign larger weights to low-frequency labels.

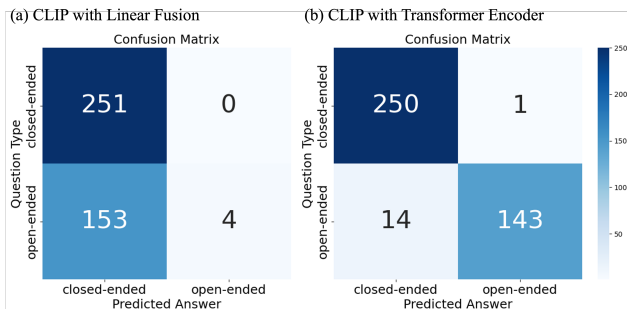


Figure 4. Answer distribution against close/open-ended questions of (a) CLIP with linear fusion and (b) CLIP with Transformer encoder.

6.4. When Does Text-Enhancement Help

In this section, we will discuss the two possible factors that determine whether text enhancement improves performance. We will focus exclusively on closed-ended questions, as experiments on all questions are affected by imbalanced-label issues, as discussed in Section 6.3.

Embedded Knowledge from Medical Image Caption Datasets Table 2 shows that text enhancement improves the performance of general-domain CLIP with any fusion modules. However, text enhancement does not improve performance of PMC-CLIP and PubMedCLIP combined with Transformer-based fusion. We infer that this is because PubMedCLIP and PMC-CLIP are fine-tuned or pre-trained on medical image caption datasets. Therefore, further fine-tuning on the MedMCQA dataset may not only fail to learn extra medical knowledge but may also result in the loss of general medical knowledge.

Data Size We observed a 5.6% to 7.9% performance gain when fine-tuning on the smaller VQA-RAD data source, as

shown in Table 4. However, the performance gain decreases to about 1% when fine-tuning on a larger data source, as shown in Table 5. We infer that additional medical text fine-tuning is more beneficial when the medical VQA dataset is small, as the text module cannot learn sufficiently general medical knowledge from a very small medical VQA dataset.

7. Conclusion and Future Work

This project investigated the impact of augmenting additional medical textual knowledge and various designs of fusion modules on the medical VQA system. We experimented with systems incorporating or excluding text enhancement, using either linear or transformer-based fusion modules, and employing CLIP-based models pre-trained on general or medical image-caption pairs. We observed that transformer-based fusion modules outperform linear fusion, especially on open-ended questions where the system requires more informative fused image-text features. Text enhancement is beneficial in binary classification tasks involving only closed-ended questions and when the training dataset size is small. For binary classification tasks, we achieved the highest accuracy of 81.7% with CLIP + Transformer decoder + text enhancement. For multi-class classification tasks, CLIP + Transformer encoder achieved the highest overall accuracy of 66.7%.

As discussed in Section 6.3, we inferred that the low performance on open-ended questions is due to the imbalanced and long-tailed label distribution. There are a total of 458 labels, with many labels appearing only once or twice. We also observed that there may be multiple reasonable answer labels for some questions. To address label imbalance problems, we can design a weighted loss function that gives more weight to low-frequency labels or develop a sampling strategy based on label frequency. Another direction is to shift to generative-based methods using large language models, which may be better at generating varied and long answers for open-ended questions. We leave these directions for future work.

8. Appendix

Figure 5 VQA-RAD Question Type Distribution

9. Contribution and Acknowledgment

- This project is not shared with projects from other classes.
- Chih-Ying Liu implemented the fine-tuning script on the VQA-RAD data, linear fusion module, and adapting PMC-CLIP.
- Fan Diao implemented the fine-tuning script on the MedMCQA data, transformer encoder/decoder fusion

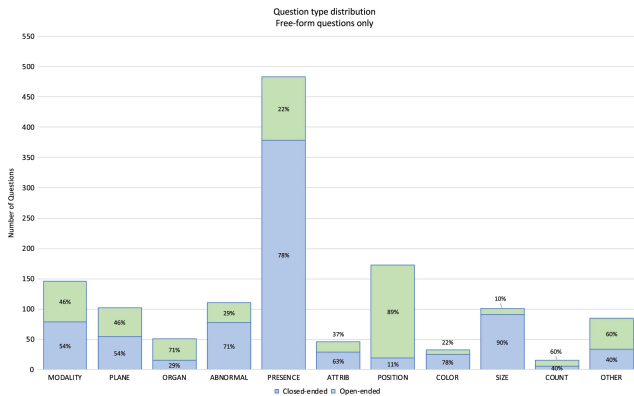


Figure 5. Question type distribution of the VQA-RAD dataset. [16]

modules, preprocessing of the new source VQA-RAD data, and fine-tune text modules.

- Both Chih-Ying Liu and Fan Diao fine-tune models on the VQA-RAD data, and write the paper together.
- We adapt PMC-CLIP code from the official repo [18], transformer implementation from CS 231N course assignment 3[4], CLIP pre-trained weights from [1] and PubMedCLIP pre-trained weights from [7].

References

- [1] O. AI. Model openai/clip-vit-base-patch32. <https://huggingface.co/openai/clip-vit-base-patch32>. Accessed: 2024-06-03.
- [2] R. Ambati and C. R. Dudyala. A sequence-to-sequence model approach for imageclef 2018 medical domain visual question answering. In *2018 15th IEEE India Council International Conference (INDICON)*, pages 1–6. IEEE, 2018.
- [3] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.
- [4] C. Course. Assignment 3 - cs231n: Convolutional neural networks for visual recognition. <https://cs231n.github.io/assignments2024/assignment3/>, 2024. Accessed: 2024-06-03.
- [5] T. Do, B. X. Nguyen, E. Tjiputra, M. Tran, Q. D. Tran, and A. Nguyen. Multiple meta-model quantifying for medical visual question answering. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, pages 64–74. Springer, 2021.
- [6] Z.-Y. Dou, Y. Xu, Z. Gan, J. Wang, S. Wang, L. Wang, C. Zhu, P. Zhang, L. Yuan, N. Peng, et al. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18166–18176, 2022.
- [7] S. Eslami, G. de Melo, and C. Meinel. Model flaviagammarino/pubmed-clip-vit-base-patch32. <https://huggingface.co/flaviagammarino/pubmed-clip-vit-base-patch32>. Accessed: 2024-06-03.
- [8] 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? *CoRR*, abs/2112.13906, 2021.
- [9] H. Face. Clip model implementation in transformers library. https://github.com/huggingface/transformers/blob/main/src/transformers/models/clip/modeling_clip.py, 2024. Accessed: 2024-06-03.
- [10] Y. Fang, W. Wang, B. Xie, Q. Sun, L. Wu, X. Wang, T. Huang, X. Wang, and Y. Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369, 2023.
- [11] S. A. Hasan, Y. Ling, O. Farri, J. Liu, H. Müller, and M. P. Lungren. Overview of imageclef 2018 medical domain visual question answering task. In *CLEF (Working Notes)*, 2018.
- [12] J. He, P. Li, G. Liu, Z. Zhao, and S. Zhong. Pefomed: Parameter efficient fine-tuning on multimodal large language models for medical visual question answering. *arXiv preprint arXiv:2401.02797*, 2024.
- [13] X. He, Y. Zhang, L. Mou, E. Xing, and P. Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.
- [14] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.
- [15] Y. Khare, V. Bagal, M. Mathew, A. Devi, U. D. Priyakumar, and C. Jawahar. Mmbert: Multimodal bert pretraining for improved medical vqa. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1033–1036. IEEE, 2021.
- [16] 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*, 5(1):1–10, 2018.
- [17] P. Li, G. Liu, J. He, Z. Zhao, and S. Zhong. Masked vision and language pre-training with unimodal and multimodal contrastive losses for medical visual question answering. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 374–383. Springer, 2023.
- [18] W. Lin, Z. Zhao, X. Zhang, C. Wu, Y. Zhang, Y. Wang, and W. Xie. Pmc-clip official repo. <https://github.com/WeixiongLin/PMC-CLIP>. Accessed: 2024-06-03.
- [19] W. Lin, Z. Zhao, X. Zhang, C. Wu, Y. Zhang, Y. Wang, and W. Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 525–536. Springer, 2023.
- [20] Z. Lin, D. Zhang, Q. Tao, D. Shi, G. Haffari, Q. Wu, M. He, and Z. Ge. Medical visual question answering: A survey. *Artificial Intelligence in Medicine*, page 102611, 2023.

- [21] Y. Liu, Z. Wang, D. Xu, and L. Zhou. Q2atransformer: Improving medical vqa via an answer querying decoder. In *International Conference on Information Processing in Medical Imaging*, pages 445–456. Springer, 2023.
- [22] B. D. Nguyen, T.-T. Do, B. X. Nguyen, T. Do, E. Tjiputra, and Q. D. Tran. Overcoming data limitation in medical visual question answering. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*, pages 522–530. Springer, 2019.
- [23] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.
- [24] A. Pal, L. K. Umapathi, and M. Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In G. Flores, G. H. Chen, T. Pollard, J. C. Ho, and T. Naumann, editors, *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR, 07–08 Apr 2022.
- [25] O. Pelka, S. Koitka, J. Rückert, F. Nensa, and C. Friedrich. Radiology objects in context (roco): A multimodal image dataset. In *CVII-STENT/LABELS@MICCAI*, 2018.
- [26] 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*, pages 8748–8763. PMLR, 2021.
- [27] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- [28] T. Tu, S. Azizi, D. Driess, M. Schaeckermann, M. Amin, P.-C. Chang, A. Carroll, C. Lau, R. Tanno, I. Ktena, et al. Towards generalist biomedical ai. *NEJM AI*, 1(3):AIoa2300138, 2024.
- [29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [30] Z. Wang, Z. Wu, D. Agarwal, and J. Sun. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*, 2022.
- [31] L.-M. Zhan, B. Liu, L. Fan, J. Chen, and X.-M. Wu. Medical visual question answering via conditional reasoning. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2345–2354, 2020.
- [32] S. Zhang, Y. Xu, N. Usuyama, H. Xu, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.
- [33] X. Zhang, C. Wu, Z. Zhao, W. Lin, Y. Zhang, Y. Wang, and W. Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023.