MISS: A GENERATIVE PRETRAINING AND FINETUNING APPROACH FOR MED-VQA

Jiawei Chen[1,2,3] *Dingkang Yang*[1,2,3] *Yue Jiang*[1,2,3] *Yuxuan Lei*[1,2,3] *Lihua Zhang*[1,2,3,4,*]

[1]Academy for Engineering and Technology, Fudan University
[2]Engineering Research Center of AI and Robotics, Shanghai, China
[3]Cognition and Intelligent Technology Laboratory (CIT Lab)
[4]Jilin Provincial Key Laboratory of Intelligence Science and Engineering, Changchun, China

ABSTRACT

Medical visual question answering (VQA) is a challenging multimodal task, where Vision-Language Pre-training (VLP) models can effectively improve the generalization performance. However, most methods in the medical field treat VQA as an answer classification task which is difficult to transfer to practical application scenarios. Additionally, due to the privacy of medical images and the expensive annotation process, large-scale medical image-text pairs datasets for pretraining are severely lacking. In this paper, we propose a large-scale **M**ult**I**-task **S**elf-**S**upervised learning based framework (MISS) for medical VQA tasks. Unlike existing methods, we treat medical VQA as a generative task. We unify the text encoder and multimodal encoder and align image-text features through multi-task learning. Furthermore, we propose a Transfer-and-Caption method that extends the feature space of single-modal image datasets using large language models (LLMs), enabling those traditional medical vision-field task data to be applied to VLP. Experiments show that our method achieves excellent results with fewer multimodal datasets and demonstrates the advantages of generative VQA models. The code and model weights will be released upon the paper's acceptance.

***Index Terms*—** Medical visual question answering, vision-language pre-training, multi-modal learning

1. INTRODUCTION

Medical Visual Question Answering (Med-VQA) is a multi-modal task based on vision and language, aiming to provide corresponding answers to given images and questions. Compared with theVQA of natural images, Med-VQA requires a deeper and more accurate understanding of medical images. At the same time, due to the privacy of medical images and the high cost of high-quality text annotation, large-scale datasets for training Med-VQA are extremely scarce. Therefore, currently, Med-VQA is still a highly challenging task.

Thanks to the development of Convolutional Neural Networks (CNN) and Natural Language Processing (NLP) techniques, some works [16] [15] [17] have attempted to use CNN and Recurrent Neural Networks (RNN) to extract image and text features respectively for VQA tasks. With the emergence of transformers [7], image features and text features can be more easily embedded into the feature space with the same dimension, and VLP models [5] [6] have emerged continuously and have been proven to be effective solutions for downstream multi-modal tasks. While effective, these VLP models suffer from several limitations, such as domain gaps when applied to medical fields.

Currently, some multimodal models specialized in the medical domain have been proposed, such as M3AE [13], MRM [10], and CMITM [11], which unify masked autoencoders (MAE) [8] and masked language modeling (MLM) pre-training to learn joint representations of images and texts; MUMC [12] utilizes masked image modeling(MIM) by sending masked images to the image encoder as data augmentation; PMC-CLIP builds a new large medical dataset and trains it on a CLIP [6] model which pretrained on natural images. However, the above Med-VQA models still have two key problems:

a. They treat Med-VQA as an answer classification task by selecting the most likely answer from a candidate pool as the output. Such models cannot adapt to diverse questions and be transferred to practical application scenarios, as there are no candidate answers in practical applications.

b. They utilize image-text pairs crawled from the article centres for pre-training, which contain a lot of noise, and high-quality open-source medical images for other tasks, such as classification and segmentation, have been ignored.

In this paper, we propose a new pre-training and fine-tuning paradigm for medical image-text tasks, named MISS, and apply it to the Med-VQA task. Unlike previous dual-tower multi-modal models, we treat Med-VQA as an answer-generating task, making our method directly applied to real-world scenarios and generating responses that more closely match human expression. We innovatively unify the text encoder and multi-modal encoder, building a **J**oint **T**ext-

*Corresponding author.

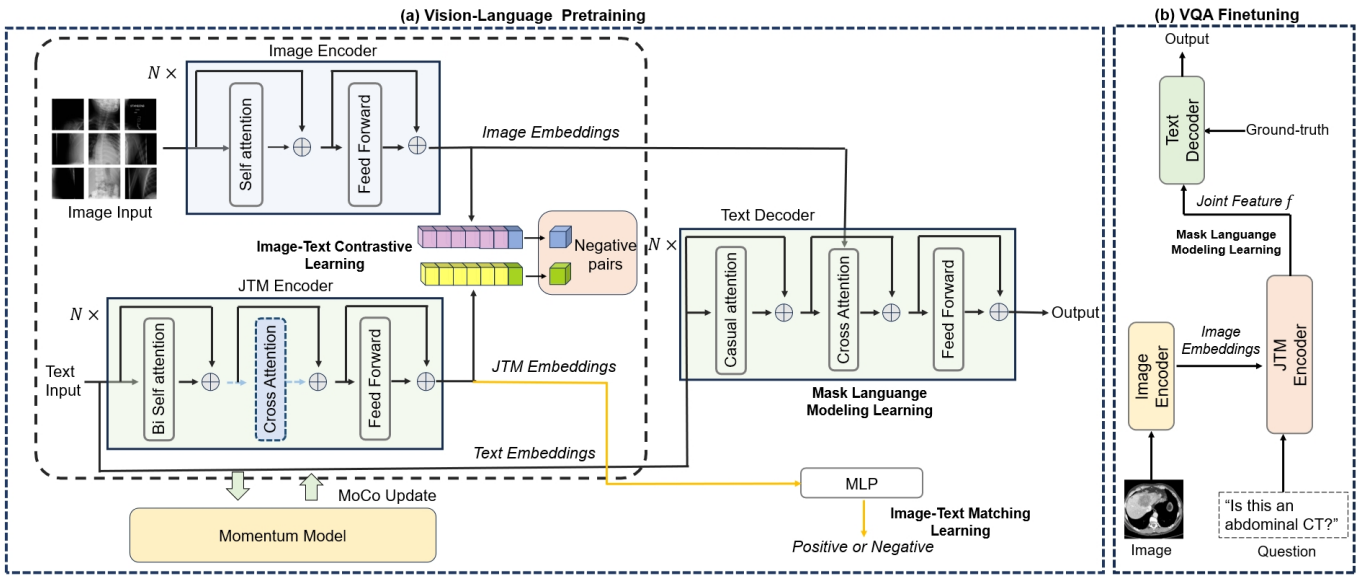

Fig. 1. Pretraining (a) and Finetuning (b) of our proposed method. We propose a pretraining and finetuning framework Miss for Med-VQA tasks which is composed of an image encoder, a JTM encoder, and a text decoder. ITC, ITM, and MLM Learning are used for pretraining. In the finetuning stage, the joint feature interacts with tokenized answers for MLM Learning.

Multimodal (JTM) encoder and enabling it to learn joint feature representations using a multi-task learning approach.

To align multi-modal features using unimodal medical images, we propose a novel method called **Trans**fer and **Cap**tion (TransCap). This method utilizes unlabeled unimodal datasets to construct image-text pairs, making it the first work in the medical field that combines large language models (LLMs) with unimodal image data to construct multi-modal datasets for visual language pretraining and finetuning. We believe that with this pioneering approach, researchers in this field no longer need to be plagued by the lack of relevant high-quality image-text pairs for pretraining.

Our main contributions are as follows:

• We propose a JTM encoder that escapes extracting text and multimodal features in different stages and enhances the efficiency of joint feature representation extraction.

• We present Transcap, a pioneering method for constructing multimodal medical data based on text-free labeled images and LLMs, which will greatly inspire the construction of pretraining data in medical multimodal fields.

• We introduce a new pretraining and fine-tuning framework named MISS. Not considering the Universal Large Vision-Language models (parameters more than 1B), it is the first pure generative VQA model in the medical field.

2. RELATED WORK

2.1. Medical Visual Question Answering

The task of Med-VQA is to provide answers based on professional questions posed by the inquirer regarding medical images. In terms of training paradigms, early works [16] [15] [17] mostly employed RNNs and CNNs to respectively extract textual and visual features. However, these models often suffer from poor generalization. Thanks to the application of transformers, large-scale pretraining has begun to migrate from the textual domain to the multimodal domain. Pretraining VLP models [5] [6] using image-text pairs and finetuning them for downstream tasks has become the preferred approach for most multimodal tasks.

In terms of content output, previous works in Med-VQA have followed the VQA paradigm in the natural image domain, treating VQA as a classification task [16] [15] [17] [18] [19]. Specifically, fully connected layers and softmax layers are installed at the output end of the model to calculate cross-entropy loss for all candidate answers. Recently, some works [12] have also employed text-based decoders, which calculate masked language model (MLM) loss for all candidate answers and select the answer with the smallest loss as the model's output. This approach is referred to as answer ranking. Although these methods achieve good accuracy on some benchmarks, they still treat visual question answering as a simple multi-classification task. When transferred to practical tasks without candidate answer pools, these Med-VQA methods cannot be effective.

In this paper, we propose a pretraining-finetuning paradigm called MISS for Med-VQA tasks. To our knowledge, this is the first work in the medical field that fully treats visual question answering as a text-generation task.

2.2. Visual-Language Pretraining Dataset

Currently, some works train Med-VQA models with the pretraining-finetuning paradigm. However, the medical field faces a shortage of image-text pairs for pretraining. ROCO [20] collects a large-scale unimodal and multimodal medical dataset from PubMed Central articles and constructs an image-text pairs dataset containing multiple types of medical images by expert radiologists. MediCaT [22] extracts images and corresponding captions from 131k openly available biomedical papers to construct a dataset containing more than 217k medical images with corresponding captions. However, these methods are similar to those used in the natural image domain to construct multimodal datasets by extracting news images and titles, and the collected images and captions contain a lot of noise, such as citations, labels, and other irrelevant information. Other high-quality open-source medical images for tasks such as classification and segmentation have been ignored because annotating these images also requires high costs and professional knowledge. We propose an automatic method for generating captions for unimodal images. This pioneering method attempts to utilize LLMs to construct a multimodal medical image-text dataset. The image data is clean, and the captions conform to human expression habits.

3. METHOD

3.1. Overview

We adopt the pretraining and fine-tuning paradigm for training medical VQA models. In the pretraining stage, we first use image-text pairs to enable the model to learn multi-modal feature representation. In the fine-tuning stage, we use image-question pairs to train the model, enabling it to be applied to Med-VQA tasks ultimately. In the following, we will first introduce our model structure, followed by our pretraining method. Finally, we will present the TransCap method and the implementation details of fine-tuning.

3.2. Model Articture

Our model adopts the encoder-decoder structure in its entirety. Unlike most dual-tower image-text VLP models in the past, our model's encoder is divided into only two parts - the image encoder and the JTM encoder. For the image encoder, we borrow from the settings adopted by most recent works, which utilize the visual transformer as the image feature extractor. For an image input I, it's firstly reshaped into flattened 2D patches and then encoded into a embedding sequence $\{x_{<cls>}, x_1, ..., x_n\}$. After that, a 12-layer transformer encoder will extract its high-dimensional features.

The JTM encoder is based on Bert and replaces the text and multi-modal encoders used in recent works. It performs representation learning of text and multi-modal features simultaneously. As shown in Figure 1(a), each JTM encoder is composed of 12 transformer-based layers, with each layer containing a bidirectional self-attention layer, a cross-attention layer, and a feed-forward layer. For each text input T, it's first pre-processed by the tokenizer into a token sequence. Then, we feed it into the JTM encoder for multi-layer representation learning, where it interacts with the image features through cross-attention. Specifically, we define the text representation as $\{w_{<cls>}, w_1, ..., w_n\}$, and the image embeddings are defined as $\{v_{<cls>}, v_1, ..., v_n\}$. Both of these representations fuse and compute multimodal representations through $CrossAttention(Q, K, V) = SoftMax(\frac{QK^T}{\sqrt{d}} + B)V$, where text representation $\{w_{<cls>}, w_1, ..., w_n\}$ generates query vectors Q, and the image representation $\{v_{<cls>}, v_1, ..., v_n\}$ generates key K and value vectors V.

The decoder part of the model includes a text decoder, which aims to decode the multi-modal feature representation obtained by the JTM encoder into an output text representation. The backbone of the text decoder is similar to that of the JTM encoder but replaces the bidirectional self-attention layer in the JTM's per-layer with a causal-attention layer. The text input passes through the causal-attention layer to calculate the text feature representation and then undergoes feature interaction with the multi-modal features through the cross-attention layer. The final features obtained are decoded by the tokenizer to obtain the text output.

3.3. Pretraining

The VLP aims to align the multimodal features while trying to make the image encoder understand the feature distribution of images in high-dimensional space and comprehend the deep semantic information of medical images. Inspired by METER, we choose Image-Text Contrastive Learning (ITC), Image-Text Matching (ITM) and Mask Language Modeling (MLM) tasks for multi-modal pretraining.

To enable the JTM encoder to learn the joint representation of text-multimodal features without being disturbed by the flow of features from another modality during the process of learning one representation, we adopt the method of BLIP [2] and deform the layer structure of the JTM encoder in different pretraining tasks. Specifically, at the beginning of model pretraining, the distance between $\{v_{<cls>}, v_1, ..., v_n\}$ extracted by the image encoder and $\{w_{<cls>}, w_1, ..., w_n\}$ of the JTM encoder in high-dimensional space is too far. At this time, it's difficult for the two features to interact and perform ITM and MLM training. Therefore, at the beginning of training, the JTM encoder will discard the cross-attention layer and extract word embeddings so that narrowing the distance between $\{v_{<cls>}, v_1, ..., v_n\}$ and $\{w_{<cls>}, w_1, ..., w_n\}$ in high-dimensional space through the ITC task. The ITC, ITM, and LM losses are calculated as delineated below.

Image-Text Contrastive Learning aims to learn unimodal representations before fusion [1]. ITC loss measures the distance of two embeddings in the feature space by a ma-

trix similarity measure $\mathbf{S} = A^T B$. Inspired by MoCo, two momentum encoders are created and they respectively have the same architecture as the text encoder and the JTM encoder. Two queues are constructed to store the most recent M image-text representations. The image and text features extracted by the image encoder and JTM encoder are denoted as $e_V(v_{cls})$ and $e_J(t_{cls})$, and those extracted by momentum encoders are denoted as $e'_V(v'_{cls})$ and $e'_J(t'_{cls})$. So we can calculate similarity $\mathbf{S}(I,T) = e_V(v_{cls})^T e'_J(t'_{cls})$ and $\mathbf{S}(T,I) = e_J(t_{cls})^T e'_V(v'_{cls})$, the softmax-normalized similarity between each image-text is calculated as follows:

$$p_m^{I2T} = \frac{exp(\mathbf{S}(I,T_m)/\tau}{\Sigma_{m=1}^{M} exp(\mathbf{S}(I,T_m)/\tau}, p_m^{T2I} = \frac{exp(\mathbf{S}(T,I_m)/\tau}{\Sigma_{m=1}^{M} exp(\mathbf{S}(T,I_m)/\tau} \tag{1}$$

where τ is the temperature parameter. Similarly, we use the above method to calculate the similarity of the embeddings and their ground truth as $y^{I2T}(I)$ and $y^{T2I}(T)$, the cross-entropy H between p and y is calculated as follows:

$$\mathcal{L} = \frac{1}{2}\mathbb{E}_{(I,T)\sim D}[H(y^{I2T}(I), p^{I2T}(I)) + H(y^{T2I}(T), p^{T2I}(T)], \tag{2}$$

which is defined as ITM loss $\mathcal{L}_{itc}$.

Image-text Matching Learning is a binary classification task, which measures visual-semantic similarity between images and texts to match and associate them. Following the setting in ALBEF [1], a linear layer after the JTM encoder is used to predict whether an image-text pair is matched or unmatched given their multimodal feature. The ground-truth label L and the probability of the matched image-text pair P_{IT} are used to calculate the ITM loss:

$$\mathcal{L}_{ITM} = \mathbb{E}_{(I,T)\sim D} H(L, P_{IT}). \tag{3}$$

Mask Language Modeling Learning trains the text decoder by randomly masking some tokens in the word vectors. Unlike most VLP models that adopt text decoders that only receive multi-modal features, our decoder simultaneously accepts input from the original tokenized text and the JTM encoder. As shown in Figure 1(a), after the word vectors $\{w_{<decod>}, w_1, ..., w_n\}$ undergo casual attention to extract word embeddings, they serve as query vectors Q and interact with image embeddings which generate key K and value vectors V to calculate cross-attention. The MLM loss $\mathcal{L}_{MLM}$ is calculated similarly to that adopted in Bert, the details of which will be covered in the Appendix.

3.4. Tranfer and Caption

The purpose of TransCap is to overcome the current challenges in medical image-text datasets, which often contain a large amount of noise since their images and captions are mostly extracted from open-source papers. In the medical field, some unimodal tasks, such as medical image classification and lesion segmentation, often have high-quality open-source data. TransCap aims to utilize these unimodal datasets to construct high-quality multimodal image-text pairs.

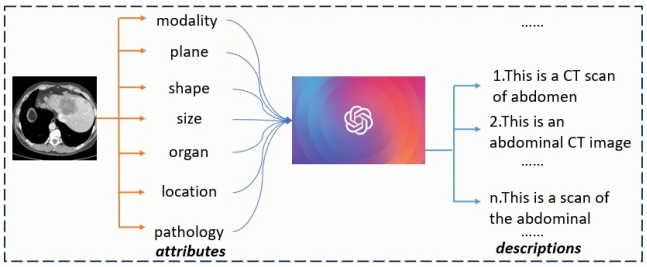

Fig. 2. Transfer and Caption unimodal images. We construct image descriptions based on image attributes and ChatGPT.

TranCap defines six attributes for medical images: modality, plane, shape, size, organ, location, and pathology. As shown in Figure 2, for each input image I, TranScap constructs a corresponding dictionary $dict_I\{\}$ with keys representing the six image attributes. For unimodal medical image datasets, TranScap uses ChatGPT to generate attribute content based on dataset information and task labels. Then, it uses this attribute content as a prompt to input the LLM and requests it to generate multiple ways of expressing the attribute's corresponding textual description. The attribute textual description serves as the value corresponding to the attribute key in $dict_I\{\}$. During pretraining, TranScap constructs captions by randomly sampling the various attribute contents of the input image dictionary. In this way, we can make use of the previously overlooked large amount of high-quality open-source unimodal image data and obtain captions that are more in line with human expression habits through LLMs. More details will be described in the appendix.

3.5. VQA Finetuning

The Med-VQA task requires generating answers to given questions and images. Previous Med-VQA tasks have been treated as classification or rank tasks, making the practical application value of models severely limited. To our knowledge, our model is the first fully generative model in the medical field (both training and testing stages are generative tasks).

As shown in Figure 1(b), during the finetuning stage, the image input I undergoes image encoding to extract image embeddings $\{i_{<cls>}, i_1, ..., i_n\}$. The question input Q is encoded by the JTM encoder to obtain question embeddings $\{q_{<encod>}, q_1, ..., q_n\}$ and then interacted with the $\{i_{<cls>}, i_1, ..., i_n\}$ in the cross-attention layer to obtain joint feature representations $f \in \mathbb{R}^{n+1}$. The tokenized answer is then sent to the casual attention layer to obtain answer embeddings $\{a_{<decod>}, a_1, ..., a_n\}$, which serve as query vectors in the cross-attention layer. These then interact with the joint feature representations, and the final output is used to calculate LM loss $\mathcal{L}_{LM}$ like Bert with the ground truth.

Table 1. Comparsion with other works which have different methods, pretraining images, training paradigm and types of task.

Methods	Pretrain # images	Training paradigm	Type of task	VQA-RAD			SLALKE		
				CLOSED	OPENED	OVERALL	CLOSED	OPENED	OVERALL
Basic Vision-Language Models (parameters <1B)									
MEVF [16]	11,779	Meta Learning	classification	75.1	43.9	-	-	-	-
MMQ [15]	-	Supervised learning	classification	75.8	53.7	67	-	-	-
VQAMIX [17]	-	Supervised learning	classification	79.6	56.6	70.4	-	-	-
AMAM [23]	-	Supervised learning	classification	80.3	63.8	73.3	-	-	-
PUBMEDCLIP-MEVF [19]	-	Pretraing-finetuning	classification	78.1	48.6	66.5	76.2	79.9	77.6
CPRD [18]	22,995	Mean Teacher	classification	80.4	61.1	72.7	83.4	81.2	82.1
MTL [24]	87,952	Pretraing-finetuning	classification	79.8	69.8	75.8	86.1	80.2	82.5
M3AE [13]	298,000	Pretraing-finetuning	classification	83.4	67.2	77	87.8	80.3	83.2
MUMC [12]	387,000	Pretraing-finetuning	ranking	84.2	71.5	79.2	-	-	84.9
ours (base model)	38,800	Pretraing-finetuning	**generating**	80.35	**71.81**	76.05	82.91	**81.47**	82
Large Vision-Language Model (parameters >1B)									
LLaVA (7B) [26]	-	Pretraing-finetuning	ranking	65.07	50.00	-	63.22	78.18	-
LLaVA-Med (7B) [25]	1M	Pretraing-finetuning	ranking	84.19	61.52	-	85.34	83.08	-
LLaVA-Med (13B) [25]	1M	Pretraing-finetuning	ranking	81.98	64.39	-	83.17	84.71	-

4. EXPERIMENT

In this section, we compare MISS with a series of previous medical VQA models. Since most methods treat VQA as a simple classification task or rank task, while our model treats VQA as a generation task, and most methods use different training paradigms and pretraining data scales, a direct result comparison may be unfair for our method. We also conduct extensive ablation studies on our method. Next, we will introduce our dataset, comparative experiments, and ablation studies. The implementation details and baseline of the experiments will be introduced in the appendix.

4.1. Dataset

We consider two Med-VQA benchmarks: the VQA-RAD dataset [14] and the Slake dataset [21]. VQA-RAD contains 315 radiology images and 3,515 QA pairs annotated by clinicians, which are evenly distributed over the head, abdomen, and chest. SLAKE is a semantically-labeled knowledge-enhanced dataset for Med-VQA, which consists of 642 radiology images and 14,028 QA pairs created by experienced physicians. Details of both datasets will be introduced in the Appendix. VQA-RAD doesn't provide a test set, and we extract 1,797 QA pairs to train the model, and the rest 451 pairs are used to test. For Slake, 14,028 QA pairs are divided into 70% training, 15% validation, and 15% testing subsets.

4.2. Comparison with Other Methods

We conduct a comparative evaluation of our method against the existing approaches on VQA-RAD and Slake. Compared to past research, our model is the only one that treats Med-VQA as a generative task, while others have approached VQA as answer classification or ranking tasks. Since generated evaluation metrics such as BELU cannot intuitively represent the model's effectiveness in Med-VQA, we have relied solely on answer accuracy as our primary evaluation metric. For closed-ended questions with only "yes" or "no" answers, we utilize automated evaluation methods. For open-ended questions, we follow the approach outlined in [14] and conduct manual evaluations comparing the generated responses with ground-truth answers. We take the model without TransCap as our base model.

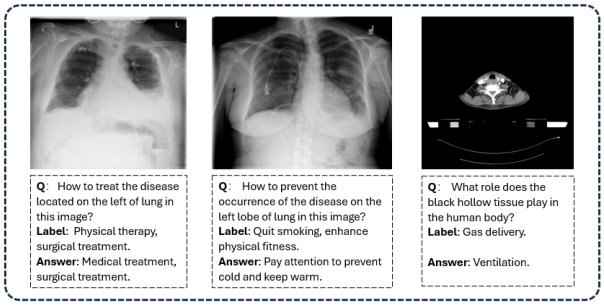

Fig. 3. Answers of our method and the ground truth.

Table 1 demonstrates our comparison with existing methods on Slake and VQA-RAD. Even if the current Large Vision-Language Model achieves better results, the extremely low pre-training cost and parameter amount of the base Vision-Language Model cannot be ignored. Apart from the large-scale vision-language model LLaVA (7B or 13B), our model has the smallest pre-trained image scale compared with methods in the pre-trained fine-tuning paradigm, using only 38,800 images. Nevertheless, our base model achieves the best accuracy in open-ended questions for Slake, which reaches 81.47%, surpassing all methods employing answer classification and ranking tasks. For VQA-RAD, although the test sets selected by each method may vary, the results show that our model still achieves good performance.

Figure 3 showcases a comparison between the responses generated by our generative model and the ground-truth answers for select questions. In some open-ended questions, Our model generates responses that differ from the ground truth but lead to the same destination, this diversity highlight-

Table 2. Ablation studies on different components of our method, "w/o" means the without.

Methods	Pretrain # images	SLALKE CLOSED	SLALKE OPENED	SLALKE OVERALL
ours (w/o pretrain)	0	50.99	3.82	19.6
ours (w/o TranScap)	38,800	82.91	81.47	82
ours (w/o JTM)	38,800	82.82	79.11	80.36
ours (JTM+TranScap)	50,000	83.94	**81.87**	82.47
ours (JTM+TranScap)	70,000	83.94	81.44	82.38
ours (JTM+TranScap)	90,000	**84.51**	81.19	**83**

ing the advantages of a generative Med-VQA model.

4.3. Abslation Studies

To demonstrate the effectiveness of different components of our method, we perform an ablation study on Slake, with the results shown in Table 2. There are several observations drawn from the results. The model without pretraining achieves only 50.99% accuracy on closed-ended questions, indicating that it can understand the task type through VQA fine-tuning, but cannot fully understand the semantics of medical images. When MISS did not utilize the JTM encoder and conventional multi-modal models were used to set up the text encoder and multi-modal encoder, our global accuracy rate was 1.64% lower than that of the base model, indicating that the JTM encoder can extract joint features more efficiently.

When our model uses both the JTM encoder and the TransCap method, we compare the impact of TransCap on our model by increasing the amount of pretraining data. As shown in the table, when using the TransCap method, with only an increase of less than 12k pretraining images, our open-ended accuracy and closed-ended accuracy increased by 1.03 and 0.5, respectively, demonstrating the positive effect of TransCap on VQA performance. Since most of the captions generated by TranCap are judgmental statements, the proportion of judgmental captions in the pretraining data continues to increase, resulting in a slight decrease in accuracy on open-ended questions; at the same time, it also leads to a certain increase in overall accuracy, ultimately reaching 83%.

5. CONCLUSION

In this paper, we propose a pretraining and finetuning framework for medical VQA tasks. We treat medical VQA as a generative task and propose a Joint Text-Multimodal encoder and align image-text features through multi-task learning. Furthermore, we propose a Transfer-and-Caption method that extends the feature space of single-modal image datasets using LLMs, enabling those traditional medical vision-field task data to be applied to VLP. We demonstrate excellent results with fewer multimodal datasets and the advantages of generative VQA models through experiments. We hope that our method will encourage the development of Med-VQA in both data and model aspects.

6. REFERENCES

[1] J. Li, R. Selvaraju, Gotmare *et al.*, "Align before fuse: Vision and language representation learning with momentum distillation," *NIPS*, vol. 34, pp. 9694–9705, 2021.

[2] J. Li, D. Li, Xiong *et al.*, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *ICCV*, 2022, pp. 12 888–12 900.

[3] Y.-C. Chen, L. Li, Yu *et al.*, "Uniter: Universal image-text representation learning," in *European conference on computer vision*. Springer, 2020, pp. 104–120.

[4] J. Devlin, M.-W. Chang, Lee *et al.*, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[5] L. H. Li, M. Yatskar, Yin *et al.*, "Visualbert: A simple and performant baseline for vision and language," *arXiv preprint arXiv:1908.03557*, 2019.

[6] A. Radford, J. W. Kim, Hallacy *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[7] A. Vaswani, N. Shazeer, Parmar *et al.*, "Attention is all you need," *NIPS*, vol. 30, 2017.

[8] K. He, X. Chen, Xie *et al.*, "Masked autoencoders are scalable vision learners," in *CVPR*, 2022, pp. 16 000–16 009.

[9] K. He, H. Fan, Wu *et al.*, "Momentum contrast for unsupervised visual representation learning," 2020, pp. 9729–9738.

[10] S. Zhang, Y. Xu, Usuyama *et al.*, "Large-scale domain-specific pretraining for biomedical vision-language processing," *arXiv preprint arXiv:2303.00915*, 2023.

[11] C. Chen, A. Zhong, Wu *et al.*, "Contrastive masked image-text modeling for medical visual representation learning," in *MICCAI*. Springer, 2023, pp. 493–503.

[12] P. Li, G. Liu, He *et al.*, "Masked vision and language pre-training with unimodal and multimodal contrastive losses for medical visual question answering," in *MICCAI*. Springer, 2023, pp. 374–383.

[13] Z. Chen, Y. Du, Hu *et al.*, "Multi-modal masked autoencoders for medical vision-and-language pre-training," in *MICCAI*. Springer, 2022, pp. 679–689.

[14] J. J. Lau, Gayen *et al.*, "A dataset of clinically generated visual questions and answers about radiology images," *Scientific data*, vol. 5, no. 1, pp. 1–10, 2018.

[15] T. Do, B. X. Nguyen *et al.*, "Multiple meta-model quantifying for medical visual question answering," in *MICCAI*, Cham, 2021, pp. 64–74.

[16] B. D. Nguyen, T.-T. Do, B. X. Nguyen *et al.*, "Overcoming data limitation in medical visual question answering," in *MICCAI*, Cham, 2019, pp. 522–530.

[17] H. Gong, G. Chen, Mao *et al.*, "Vqamix: Conditional triplet mixup for medical visual question answering," *IEEE Transactions on Medical Imaging*, vol. 41, no. 11, pp. 3332–3343, 2022.

[18] B. Liu, L.-M. Zhan, and X.-M. Wu, "Contrastive pretraining and representation distillation for medical visual question answering based on radiology images," in *MICCAI 2021*. Cham: Springer International Publishing, 2021, pp. 210–220.

[19] 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*, vol. abs/2112.13906, 2021. [Online]. Available: https://arxiv.org/abs/2112.13906

[20] O. Pelka, S. Koitka, Rückert *et al.*, "Radiology objects in context (roco): a multimodal image dataset," in *LABELS 2018, MICCAI 2018*, 2018, pp. 180–189.

[21] B. Liu, L.-M. Zhan, Xu *et al.*, "Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering," in *2021 ISBI*, 2021, pp. 1650–1654.

[22] S. M. Sanjay Subramanian, Lucy Lu Wang *et al.*, "MedICaT: A Dataset of Medical Images, Captions, and Textual References," in *Findings of EMNLP*, 2020.

[23] H. Pan, S. He, K. Zhang *et al.*, "Amam: An attention-based multimodal alignment model for medical visual question answering," *KBS*, vol. 255, p. 109763, 2022.

[24] F. Cong, S. Xu *et al.*, "Caption-aware medical vqa via semantic focusing and progressive cross-modality comprehension," in *ACM MM*, 2022, pp. 3569–3577.

[25] C. Li, C. Wong, Zhang *et al.*, "Llava-med: Training a large language-and-vision assistant for biomedicine in one day," *arXiv preprint arXiv:2306.00890*, 2023.

[26] H. Liu, C. Li, Wu *et al.*, "Visual instruction tuning," *arXiv preprint arXiv:2304.08485*, 2023.

A. APPENDIX

A.1. Overview

In the appendix, we will supplement the content that cannot be described in detail due to space limitations in the main text. We will introduce several contents in turn: first, we will introduce the basic principle of Mask Language Modeling(MLM) loss and its calculation method in the pretraining and finetuning process of this article; then, we will use practical examples to specifically explain how the TransCap method proposed in this article constructs multi-modal image-text pairs based on LLM and unimodal image data; after that, we will supplement the implementation details and baselines of our experiments.

A.2. Details of MLM loss

In the MLM task of Bert, 15% of the tokens are randomly selected, 80% of them are replaced with a special token [MASK], 10% are randomly replaced with other words, and the other 10% are left unchanged. The model does not know whether the position is [MASK] or an original token or a random word, so it forces the model to predict the masked token according to the context. In our proposed paradigm, while pretraining the model, the tokenized caption embeddings $\{w_{<decod>}, w_1, ..., w_n\}$ and the image embeddings $\{i_{<cls>}, i_1, ..., i_n\}$ fuse in the cross-attention layer and while finetuning it's joint feature f who fuses with question embeddings $\{q_{<encod>}, q_1, ..., q_n\}$. Both of them then are sent to the feed-forward layer and calculate MLM loss $\mathcal{L}_{MLM}$.

MLM will minimize a cross-entropy loss between predicted results and ground-truth results:

$$\mathcal{L}_{MLM} = \mathbb{E}_{(I,\hat{T})\sim D} H(y^{msk}, p^{msk}(I, \hat{T}) \tag{4}$$

where $\hat{T}$ presents a masked token, $p^{msk}(I, \hat{T})$ presents the predicted probability for the masked token, and y^{msk} is a true distribution of vocabulary.

A.3. Details of TransCap

Transfer and Caption is a method based on large language models(LLMs) to extend the feature space of unimodal image data, which has never been explored in the medical field before. Specifically, medical images have a large amount of high-quality open-source unimodal data in other unimodal tasks, such as image classification and lesion segmentation. The labels of these data contain the attributes of medical images. We define the attributes of medical images as follows: modality, plane, shape, size, organ, location, and pathology. By obtaining the attribute content of images, we input it into LLMs and obtain different ways of describing all attribute content. By freely combining multiple attribute descriptions, we can obtain diverse text descriptions of a certain image. Below, we introduce the process of constructing image-text pairs using the TransCap method with a specific example.

The RSNA-PDC is a chest radiograph (CXR) dataset with a training set of 26,684 CXRs in three categories: Normal, No Lung Opacity/Not Normal, Lung Opacity. Based on the CXR attributes, the labels of each CXR image can be set to the following types: Modality (indicates data type), Class (can be: Normal, No Lung Opacity/Not Normal, Lung Opacity), Nums (indicates the number of lung opacities), Location (indicates the location of lung opacity). For example, one piece of data in the dataset is:

$dict$\{"Img_id": "000db696-cf54-4385-b10b-6b16fbb3f985", "Modality": "CXR", "Class": "Lung Opacity", "Nums": 2, "Location": "the upper left, the upper right"\}

The attribute labels of each data will be input into LLM as a prompt, and it is required to generate a caption to describe the attributes of CXR. Additionally, to realize data enhancement, LLM will be utilized to rephrase, obtaining more captions with the same meaning, but more in line with human expression habits.

Ultimately, one-to-one caption sets for each CXR image are constructed, which contain multiple captions with different expressions. As mentioned in this paper, during pre-training, the captions in the caption set will be randomly sampled to construct an image-text pair, such as {"Img_id": "000db696-cf54-4385-b10b-6b16fbb3f985", "Caption": "This chest X-ray reveals abnormalities in the lung, with the presence of 2 lung opacities located at the upper left, the upper right."}

An image from an unimodal dataset is transformed into an image-text pair, and data enhancement is realized with the LLM to obtain a caption set which is more in line with human expression habits. Figure 4 compares the largest medical multimodal dataset, MedICat, collected from article centres, with image-text pairs generated by TransCap. It can be seen that the image-text pairs generated by the TransCap method contain less noise and are more humanized in terms of both image and caption. We believe that this method of combining LLMs to generate pretrained multimodal medical data will greatly encourage the development of this field.

A.4. Details of Experiments

Baseline: In this paper, we present an image-text pretraining framework tailored for generative tasks and propose a joint text-multimodal encoder to simultaneously extract features from both images and text through multi-task pretraining. To better compare the advantages of our approach with others, we have constructed a baseline for our study based on ALBEF. Specifically, in terms of model architecture, the baseline model comprises a vision transformer (ViT) base as the backbone for the image encoder, consisting of 12 transformer

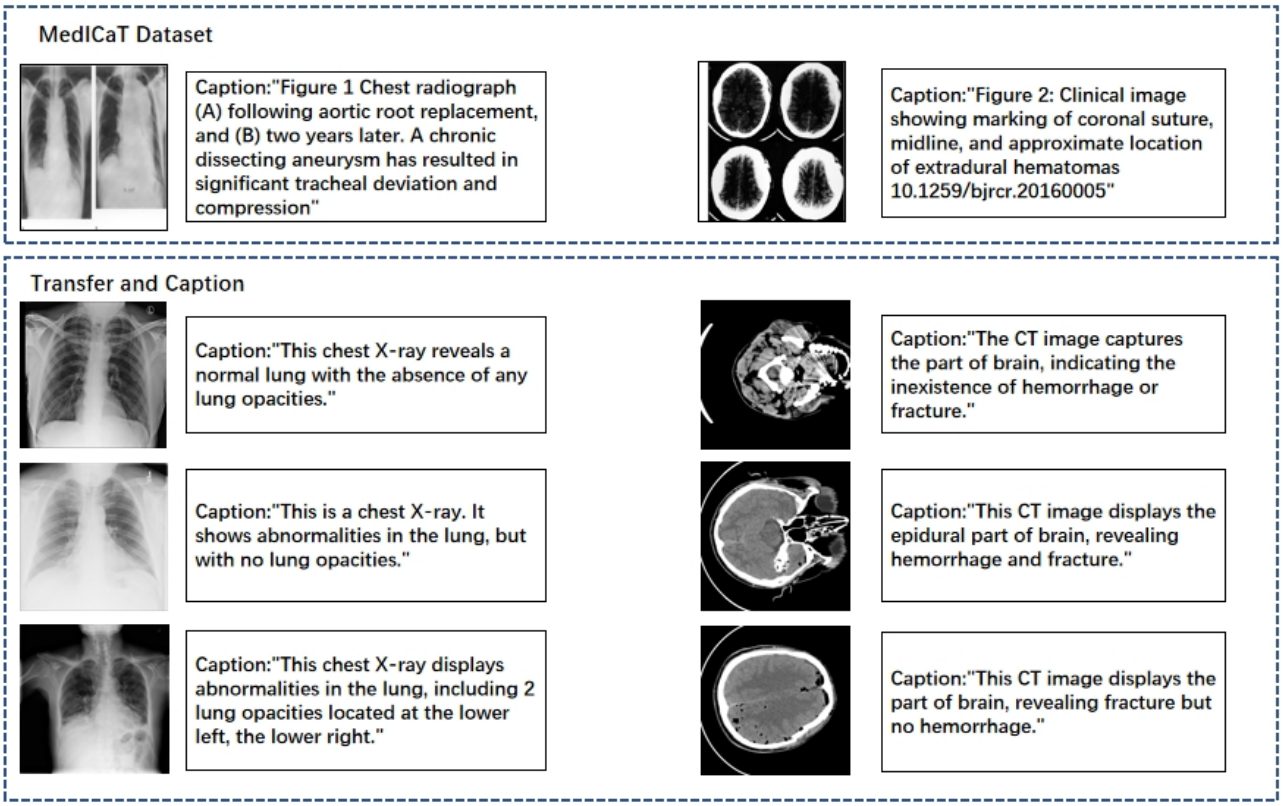

Fig. 4. Comparison of data from MedICat Dataset and image-text pair data generated through TransCap. Image-text pairs generated by the TransCap method contain less noise and are more humanized in terms of both image and caption.

layers; a BERT-based text encoder and multimodal encoder, wherein the first six layers of the BERT model serve as the text encoder identical to the original BERT encoder, while the latter six layers incorporate cross-attention between the self-attention and feed-forward layers to function as the multimodal encoder; and a BERT-based text decoder, with each layer identical in structure to the BERT encoder.

In terms of the pretraining of the baseline model, it still follows ALBEF and sets up three pretraining tasks: Image-Text Contrastive Learning, Image-text Matching Learning, and Mask Language Modeling Learning. For fine-tuning the VQA task, we still use Mask Language Modeling Learning as the fine-tuning task. Since ALBEF still treats VQA as a RANK task in testing, we modified the output end of the model to enable it to directly output text. Finally, we constructed our baseline model based on the above methods, which is the method mentioned in the main text that does not utilize the JTM encoder.

Pretraining Details: To better compare the superiority of our method and the impact of TransCap on our model, we choose to train a basic model that does not utilize the high-quality image-text pairs generated by TransCap but rather employs the MedICaT dataset used by most works in the literature. However, to demonstrate the model's performance, we do not use the entire MedICaT dataset for training as in MAE and MUMC. Instead, we randomly select 38,800 radiological images from MedICaT to pretrain our model, a training scale that is only one-eighth to one-tenth of the training data used by other methods. At the same time, in the ablation experiments, to demonstrate the effectiveness of the TransCap method, we gradually incorporate the image-text pairs generated by TransCap into the basic model training, increasing the pretraining data size to 50,000, 70,000, and 90,000. Since TransCap mainly generates descriptive judgmental statements, Table 3 demonstrates that our TransCap method is effective in terms of the model's performance on closed-ended questions.

Table 3. Ablation studies on different components of our method

Methods	Pretrain # images	SLAKE		
		CLOSED	OPENED	OVERALL
ours (base model)	38,800	82.91	81.47	82
baseline	38,800	82.82	79.11	80.36
ours (JTM+TranScap)	50,000	83.94	**81.87**	82.47
ours (JTM+TranScap)	70,000	83.94	81.44	82.38
ours (JTM+TranScap)	90,000	**84.51**	81.19	**83**

Implement Details: Here, we will present the experimen-

tal details of our pretraining and fine-tuning models. All of our training was conducted on a single NVIDIA RTX8000-50GB GPU. During the pretraining stage, we did not use any data augmentation techniques. We used the Adamw optimizer with cosine learning rate decay, an initial learning rate of 2e-5, weight decay of 0.05, a minimum learning rate of 0, and training for 100 epochs on the pretraining dataset. Additionally, during the pretraining stage, the input image size for our model was 224x224 pixels. In the fine-tuning stage, we used the same optimizer settings and learning rate as in the pretraining stage, with an input image size of 480x480 pixels, and trained for 200 epochs. For our baseline model, the Vit-based visual encoder consists of 12 layers of transformer, and both the JTM encoder and text decoder contain 12 layers of transformer-based layers.