BESTMVQA: A Benchmark Evaluation System for Medical Visual Question Answering

Xiaojie Hong[1], Zixin Song[1], Liangzhi Li[2(✉)], Xiaoli Wang[1(✉)], and Feiyan Liu[1]

[1] School of Informatics, Xiamen University, Xiamen, China
{xjhong,zxsong,feiyanliu}@stu.xmu.edu.cn, xlwang@xmu.edu.cn
[2] Meetyou AI Lab, Xiamen, China
liliangzhi@xiaoyouzi.com

Abstract. Medical Visual Question Answering (Med-VQA) is a task that answers a natural language question with a medical image. Existing VQA techniques can be directly applied to solving the task. However, they often suffer from (*i*) the data insufficient problem, which makes it difficult to train the state of the arts (SOTAs) for domain-specific tasks, and (*ii*) the reproducibility problem, that existing models have not been thoroughly evaluated in a unified experimental setup. To address the issues, we develop a Benchmark Evaluation SysTem for Medical Visual Question Answering, denoted by BESTMVQA. Given clinical data, our system provides a useful tool for users to automatically build Med-VQA datasets. Users can conveniently select a wide spectrum of models from our library to perform a comprehensive evaluation study. With simple configurations, our system can automatically train and evaluate the selected models over a benchmark dataset, and reports the comprehensive results for users to develop new techniques or perform medical practice. Limitations of existing work are overcome (*i*) by the data generation tool, which automatically constructs new datasets from unstructured clinical data, and (*ii*) by evaluating SOTAs on benchmark datasets in a unified experimental setup. The demonstration video of our system can be found at https://youtu.be/QkEeFlu1x4A, and the source code is shared on https://github.com/emmali808/BESTMVQA.

Keywords: Medical Visual Question Answering · Benchmark Evaluation System · Comprehensive Experimental Study

1 Introduction

Medical visual question answering is a challenging task in healthcare industry, which answers a natural language question with a medical image. Figure 1 shows an example of the Med-VQA data. It may aid doctors in interpreting medical images for diagnoses with responses to close-ended questions, or help patients

A. Bifet et al. (Eds.): ECML PKDD 2024, LNAI 14949, pp. 435–451, 2024.
https://doi.org/10.1007/978-3-031-70378-2_27

Question	Is this image abnormal?
Answer	Yes.
Question Type	Abnormality
Answer Type	CLOSED
Image Organ	LEG

Fig. 1. An example of Med-VQA

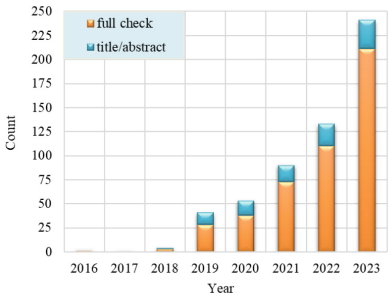


Fig. 2. Publications on Med-VQA since 2016

with urgent needs get timely feedback on open-ended questions [13]. It is a challenging problem which processes multi-modal information. Different from general VQA, Med-VQA requires substantial prior domain-specific knowledge to thoroughly understand the contents and semantics of medical visual questions.

Many exiting techniques contribute to solving this task (e.g., [9]). However, they generally suffer from the data insufficient problem. They need to be trained on well-annotated large datasets, to learn enough domain-specific knowledge for understanding medical visual questions. Several works focus on constructing Med-VQA datasets [2,11,12,15,17]. However, these datasets seem to be a drop in the bucket. Other works employ data augmentation method to tackle the problem. VQAMix [9] has focused on generating Med-VQA training samples. However, it may incur noisy samples that affect the performance of models. Current work have adopted transfer learning to pre-train a visual encoder on external medical image-text pairs to capture suitable visual representations for subsequent cross-modal reasoning [6,9,13]. They achieve success by performing pre-training using large-scale data unannotated data. However, they have not been thoroughly evaluated in benchmark settings.

To address the problems, we develop BESTMVQA, which is a benchmark evaluation system for Med-VQA. We first provide a data generation tool for users to automatically construct new datasets from self-collected clinical data. We implement a wide spectrum of SOTA models for Med-VQA in a model library. Accordingly, users can conveniently select a benchmark dataset and any model in model library for medical practice. Our system can automatically train the models and evaluate them over the selected dataset, and present a final comprehensive report to users. With our system, researchers can comprehensively study SOTA models and their applicability in Med-VQA. The impact of our contributions also can be inferred from Fig. 2, which shows the significant increase in Med-VQA publications since 2016. We provide a unified evaluation system for users to (*i*) reveal the applicability of SOTA models to benchmark datasets, (*ii*) conduct a comprehensive study of the available alternatives to develop new Med-VQA techniques, and (*iii*) perform various medical practice.

2 Research Scope and Task Description

The research scope is tailored to two types of readers: (*i*) Researchers who require Med-VQA techniques to perform downstream tasks; (*ii*) Contributors in the research community of Med-VQA who need to thoroughly evaluate the SOTAs.

Medical visual question answering is a domain-specific task that inputs a medical image and a related question, outputting an answer in natural language. It requires extensive domain knowledge, adding complexity beyond general VQA tasks. The lack of well-annotated large-scale datasets makes it hard to learn enough medical knowledge. To address the challenge, current work typically pre-train a visual encoder on large unlabeled medical image-text pairs.

In Fig. 3, Med-VQA models consist of four main components: vision encoder, text encoder, feature fusion, and answer prediction, which together process the image and question inputs to predict answers.

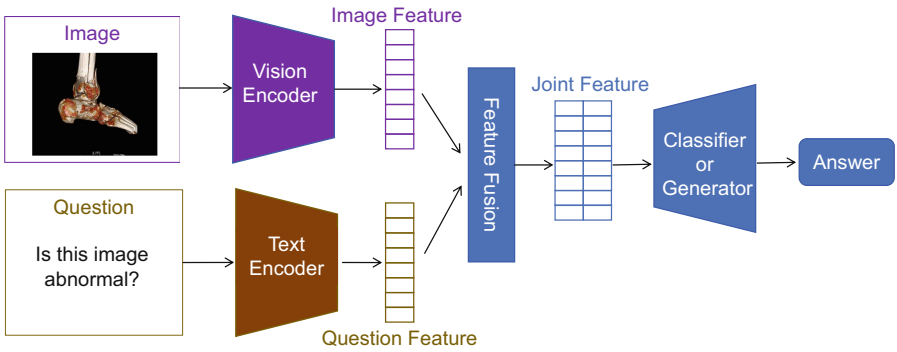


Fig. 3. The architecture of mainstream Med-VQA models

3 Related Work

Med-VQA is a challenging task that combines natural language processing and computer vision. Early work employing traditional machine learning algorithms suffers from poor performance due to significant differences between visual and textual features [26]. Inspired by the success of deep learning in information systems, deep learning models for Med-VQA are reported to have performance gains over traditional models [23]. They can be classified into four categories: joint embedding, encoder-decoder, attention-based, and large language models (LLMs). Table 1 shows the statistics of SOTAs we reproduced.

The joint embedding models combine visual and textual embeddings into a final representation. We implement some representative models such as MEVF [19] and CR [32]. MEVF uses MAML [7] and CDAE [18] to initialize

Table 1. The statistics of considered models, including the parameter size (Params), the training time (Training Time), supporting pre-training or not (Support PT), supporting fine-tuning or not (Support FT) and model category (Model Category). The left value of Training Time represents the smallest training time over all datasets, while the right value is the largest one.

Baseline	Params	Training Time	Support PT	Support FT	Model Category
MEVF [19]	15M	0.03 h–0.3 h	×	✓	Joint Embedding
CR [32]	38M	0.04 h–0.4 h	×	✓	Joint Embedding
MMQ [4]	20M	0.5 h–3.0 h	✓	✓	Joint Embedding
VQAMix [9]	19M	0.6 h–6.0 h	×	✓	Joint Embedding
CMSA [8]	88M	1.0 h–4.2 h	×	✓	Attention-Based
MMBERT [14]	117M	1.7 h–13.3 h	✓	✓	Attention-Based
PTUnifier [3]	350M	3.0 h–13.0 h	✓	✓	Attention-Based
METER [5]	320M	2.5 h–18.0 h	✓	✓	Attention-Based
TCL [29]	580M	1.3 h–8.3 h	×	✓	Encoder-Decoder
MiniGPT-4 [34]	14110M	-	×	×	LLMs
LLaVA-Med [16]	6743M	-	×	×	LLMs

the model weights for visual feature extraction, while CR proposes question-conditioned reasoning and task-conditioned reasoning modules for textual feature extraction.

For encoder-decoder models, visual and textual features are extracted separately by encoders, and fused in a feature fusion layer. The decoder generates the answer based on the fused features. NLM [21], TCL [29], and MedVInT [33] are such representative models.

The third category employs attention mechanisms to capture representative visual and textual features. MMBERT [14] employ Transformer-style architecture to extract visual and textual features. CMSA [8] introduce a cross-modal self-attention module to selectively capture the long-range contextual relevance for more effective fusion of visual and textual features. MedFuseNet [22] excels in open-ended visual question answering on recent public datasets through a BERT-based multi-modal representation, coupled with an LSTM decoder. We have implemented four representative models, including MMBERT [14], CMSA [8], PTUnifier [3] and METER [5].

Recently, motivated by the achievements of ChatGPT [27] and GPT-4 [1], alongside the efficacious deployment of open-source, instruction-tuned large language models (LLMs) within the general domain, a myriad of biomedical-oriented LLM chatbots have emerged. Notable among these are ChatDoctor [31], Med-Alpaca [10], PMC-LLaMA [25], DoctorGLM [28], and Huatuo [24]. LLMs are trained on large amounts of textual data that can help interpret complex and detailed information in medical images. Our model library also provides two recent models for generating the linguistic representation of the question

in Med-VQA: MiniGPT-4 [34] has multi-modal abilities by properly aligning visual features with advanced LLMs, and LLaVA-Med [16] performs multi-modal instruction-tuning by leveraging large-scale biomedical data.

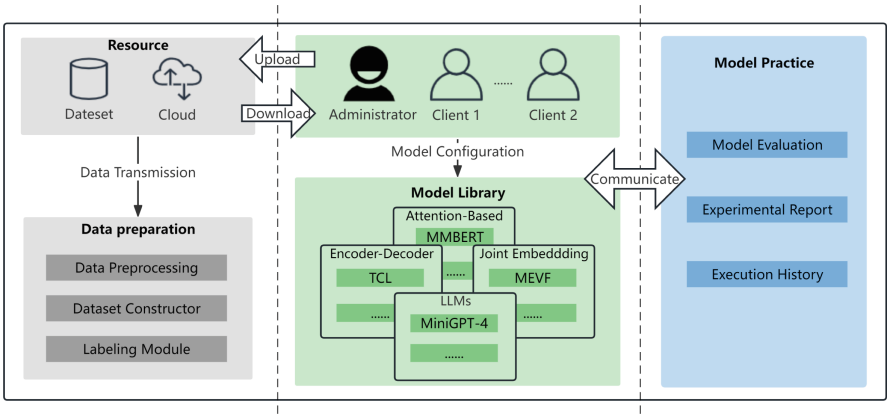


Fig. 4. System architecture of our BESTMVQA

4 System Overview

In Fig. 4, our BESTMVQA system has three components: data preparation, model library, and model practice. The data preparation component is developed based on a semi-automatic data generation tool. Users first upload self-collected clinical data. Then, medical images and relevant texts are extracted for medical concept discovery. We provide a human-in-the-loop framework to analyze and annotate medical concepts. To facilitate the effort, we first auto-label the medical concepts by employing the BioLinkBERT-BiLSTM-CRF [30]. Then, professionals can conveniently verify the medical concepts. After that, medical images, medical concepts and diagnosis texts are fed into a pre-trained language model for generating high-quality QA pairs. We employ a large-scale medical multi-modal corpus to pre-train and fine-tune an effective model, which can be easily incorporated into existing neural models for generating medical VQA pairs. our system provides a model library, to avoid duplication of efforts on implementing SOTAs for experimental evaluation. A wide spectrum of SOTAs have been implemented. The detailed statistics of the models can be seen in Sect. 3. Based our library, users can conveniently select a benchmark dataset and any number of SOTAs from our model library. Then, our system automatically performs extensive experiments to evaluate SOTAs over the benchmark dataset, and presents the final report to the user. From our report, the user can comprehensively study SOTAs and their applicability to Med-VQA. Users can also download the experimental reports and the source codes for further practice.

5 Empirical Study

Users can use our BESTMVQA system to systematically evaluate SOTAs on benchmark datasets for Med-VQA. To comprehensively evaluate the effectiveness of the models, we employ the metric of *accuracy* for open-ended, closed-ended, and overall questions. Five datasets are provided for users for model practice to investigate the applicability of models to diverse application scenarios.

Table 2. The statistics of datasets. NI, NQ and NA represent the number of images, questions and answers, respectively. MeanQL and MeanAL represent the length of questions and answers, respectively.

Dataset	NI	NQ	MeanQL	MeanAL	NA
VQA-RAD [15]	314	3515	6.49	1.61	557
MedVQA-2019 [2]	4200	15292	6.88	2.12	1749
SLAKE-EN [17]	642	7033	8.03	1.4	234
PathVQA [11]	4289	32795	6.33	1.79	4946
OVQA [12]	2000	19020	8.73	3.32	1065

5.1 Considered Models

We emphasize the utilization of "out-of-the-box" models, defining a model as "usable out of the box" if it meets the following criteria: (*i*) publicly available executable source code, (*ii*) well-defined default hyperparameters, (*iii*) no mandatory hyperparameter optimization, and (*iv*) absence of requirements for language model retraining and vocabulary adaptation. To ensure consistent evaluation and practical applicability, all models are expected to generate predictions in a standard format. Adhering to the criteria is essential for models that can help guarantee aligning with the concept of "out of the box".

Models are identified and classified as shown in Table 1, containing (*i*) those specifically tailored for Med-VQA, and (*ii*) the application of general VQA models to the medical domain.

5.2 Experimental Setup

Datasets. All models are evaluated using the following five datasets:

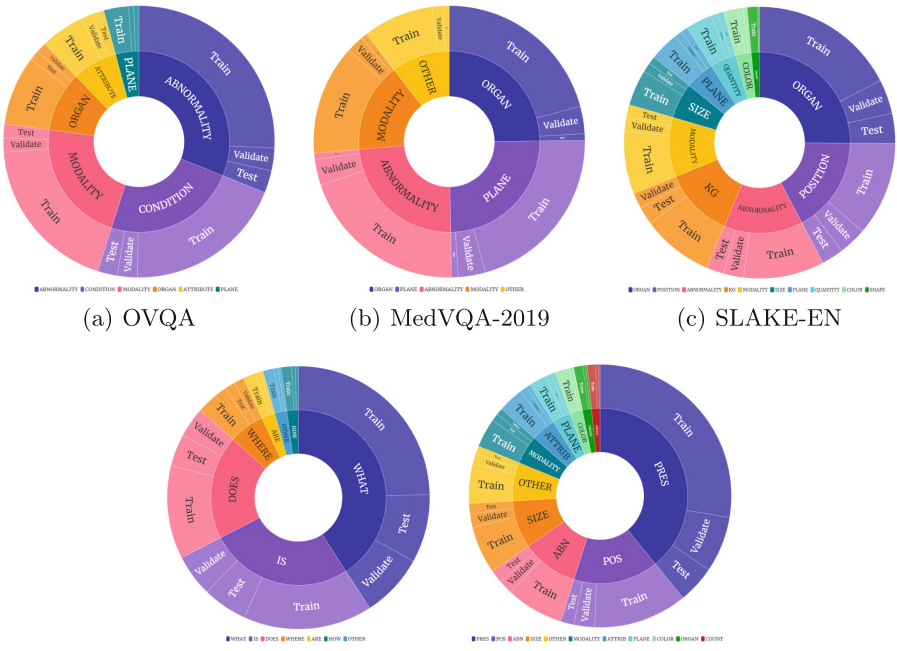
OVQA [12] has 2,001 images and 19,020 QA pairs, with each image linked to multiple QA pairs.

VQA-RAD [15] includes 314 images and 3,515 questions answered by clinical doctors, with 10 question types across the head, chest and abdomen.

SLAKE [17] is a bilingual dataset annotated by experienced doctors, which is represented as SLAKE-EN in English.

Table 3. Default values for Batch Size, Learning Rate, and Epoch for each model

Baseline	Batch Size	Learning Rate	Epoch
MEVF+SAN	16	1.00E-03	20
MEVF+BAN	8	1.00E-03	20
CR	64	1.00E-03	40
MMQ	64	1.00E-03	60
VQAMix+SAN	8	1.00E-03	80
VQAMix+BAN	8	1.00E-03	80
CMSA	32	1.00E-03	60
MMBERT	16	1.00E-03	80
PTUnifier	8	1.00E-05	50
METER	32	1.00E-05	25
TCL	4	2.00E-05	20



(a) OVQA (b) MedVQA-2019 (c) SLAKE-EN

(d) PathVQA (e) VQA-RAD

Fig. 5. Distribution of question types per dataset

MedVQA-2019 [2] is a radiology dataset from the ImageClef challenge, which includes 642 images with over 7,000 QA pairs.

PathVQA [11] consists of 32,795 pairs generated from pathological images.

Datasets were chosen for their diversity in sample sizes (Table 2). For VQA-RAD and SLAKE, we have reorganized the datasets in a 70%-15%-15% ratio due to the lack of validation sets. As for the other datasets, We use the proportion of the corresponding data splits. The detailed statistics for data splits are shown in Table 4. The distribution of question types is illustrated in Fig. 5.

Table 4. The statistics of data splits. NI represents the number of images. MaxQL, MinQL and MeanQL represent the max, min and mean length of questions, respectively; NCF and NOF represent the number of close-ended and open-ended questions, respectively. MedVQA-2019 is not divided into open-ended and closed-ended questions.

Dataset	Sample	NI	MaxQL	MinQL	MeanQL	Vocabulary	NCF	NOF
VQA-RAD (train)	2451	314	21	3	6.43	1114	1443	1008
VQA-RAD (valid)	613	258	19	3	6.42	625	380	233
VQA-RAD (test)	451	203	22	3	6.89	538	272	179
Total	3515	314	22	3	6.49	1288	2095	1420
MedVQA-2019 (train)	12792	3200	11	4	6.88	98	-	-
MedVQA-2019 (valid)	2000	500	11	4	6.86	94	-	-
MedVQA-2019 (test)	500	500	11	4	6.86	93	-	-
Total	15292	4200	11	4	6.88	98	-	-
SLAKE-EN (train)	4777	546	21	4	7.98	301	1905	2872
SLAKE-EN (valid)	1195	484	18	4	8.12	265	460	735
SLAKE-EN (test)	1061	96	21	4	8.11	265	416	645
Total	7033	642	21	4	8.03	306	2781	4252
PathVQA (train)	19755	2599	37	2	6.35	4161	9868	9887
PathVQA (valid)	6279	832	37	2	6.24	2537	3156	3123
PathVQA (test)	6761	858	42	2	6.33	2608	3409	3352
Total	32795	4289	42	2	6.33	5095	16433	16362
OVQA (train)	15216	2000	95	4	8.63	958	8037	7179
OVQA (valid)	1902	1235	62	4	9.04	613	830	1072
OVQA (test)	1902	1234	67	4	9.26	533	832	1070
Total	19020	2000	95	4	8.73	1005	9699	9321

Implementation Details. For pre-training, we use a large-scale publicly available dataset called by ROCO [20]. It contains image-text pairs collected from PubMed (https://pubmed.ncbi.nlm.nih.gov/). We selected 87,952 non composite radiographic images with relevant captions. For fine-tuning, we follow the training, validation, and testing data splits according to Table 4. Five benchmark Med-VQA datasets were used to train and evaluate SOTAs. Questions are divided into closed-ended and open-ended. Closed-ended questions are usually answered with "yes/no" or other limited options. Open-ended questions have

Table 5. Experimental results for discriminative models on the test set of VQA-RAD, SLAKE-EN, PathVQA, and OVQA datasets, including the *Accuracy* (ACC) of three indicators: Closed-ended, Open-ended, and Overall.

Dataset	Baseline	Closed-ended (ACC)	Open-ended (ACC)	Overall (ACC)
VQA-RAD	MEVF+SAN	75.4	40.2	61.4
	MEVF+BAN	78.3	52.5	68.1
	CR	77.2	57.6	69.4
	MMQ	75.7	56.9	68.2
	VQAMix+SAN	79.4	57	70.5
	VQAMix+BAN	80.9	57.5	71.6
	CMSA	78.5	63.7	72.5
	MMBERT	74.3	46.9	63.4
	PTUnifier	**86.4**	**68.2**	**79.2**
	METER	78.3	57	69.8
	TCL	73.5	56.4	66.7
SLAKE-EN	MEVF+SAN	78.4	75.3	76.5
	MEVF+BAN	81	75.7	77.8
	CR	76.9	78.4	77.5
	MMQ	78.4	76.7	77.4
	VQAMix+SAN	77.9	77.7	77.8
	VQAMix+BAN	83.2	78.1	80.1
	CMSA	68.3	49.1	56.6
	MMBERT	43.3	1.9	18.1
	PTUnifier	**89.4**	**81.6**	**84.6**
	METER	87.3	79.2	82.4
	TCL	87.5	78.4	82
PathVQA	MEVF+SAN	83.4	13.1	48.5
	MEVF+BAN	83.8	16.4	50.3
	CR	84.9	15.9	50.5
	MMQ	83.2	14.3	48.9
	VQAMix+SAN	83.9	9.6	46.9
	VQAMix+BAN	84.3	12.7	48.6
	CMSA	83.7	16.1	50.2
	MMBERT	83.2	13	48.1
	PTUnifier	85.5	10.1	48.1
	METER	**89.9**	29.8	60
	TCL	88.1	**36.9**	**62.7**
OVQA	MEVF+SAN	74.2	52.3	61.9
	MEVF+BAN	76.6	50.5	61.9
	CR	76.6	36.9	54.3
	MMQ	79	53.2	64.5
	VQAMix+SAN	77.6	59.1	67.2
	VQAMix+BAN	79.3	57	66.8
	CMSA	79.7	45.6	60.5
	MMBERT	80.5	48.7	62.6
	PTUnifier	**84.9**	**60.5**	**71.3**
	METER	82.1	51.7	65.1
	TCL	82.6	60.4	70.1

no restrictive structure and can have multiple correct answers. All models are trained on dual graphics NVIDIA RTX V100 GPU. We use the AdamW opti-

Table 6. Experimental results for discriminative models on the test set of MedVQA-2019. Due to the fact that the MedVQA-2019 is not strictly divided into open-ended and closed-ended question types, the table only contains the values of Overall *Accuracy*

Dataset	Baseline	Overall(ACC)
MedVQA-2019	MEVF+SAN	50
	MEVF+BAN	47.4
	CR	46.8
	MMQ	50
	VQAMix+SAN	47.2
	VQAMix+BAN	49
	CMSA	47.4
	MMBERT	51.2
	PTUnifier	60.3
	METER	**73.9**
	TCL	63

Table 7. Experimental results for generative models on the test set of VQA-RAD, SLAKE-EN, PathVQA, OVQA and MedVQA-2019 datasets, including the *Accuracy* (ACC) of Closed-ended and the *Recall*, $METEOR$ of Open-ended.

Dataset	Baseline	Closed-ended (ACC)	Open-ended	
			Recall	$METEOR$
VQA-RAD	MiniGPT-4	56.2	32.2	0.043
	LLaVA-Med	58.8	32.1	0.238
SLAKE-EN	MiniGPT-4	53.2	36.8	0.038
	LLaVA-Med	53.6	40.7	0.308
PathVQA	MiniGPT-4	53.4	12	0.018
	LLaVA-Med	57.9	11.8	0.026
OVQA	MiniGPT-4	53.1	33.4	0.066
	LLaVA-Med	66.8	39.1	0.237
MedVQA-2019	MiniGPT-4	-	23.2	0.019
	LLaVA-Med	-	25	0.055

mizer with the same preheating steps. See Table 3 for detailed parameter settings of models.

5.3 Evaluation Metrics

To quantitatively measure the performance of models, we use the *accuracy* as an evaluation metric, and compute it for closed-ended and open-ended questions for discriminative models, as they can be defined as a classification task. Let P_i

and L_i respectively denote the prediction and ground-truth label of sample i in the test set, and T represents the test set. The *accuracy* is calculated as follows:

$$accuracy = \frac{1}{|T|} \sum_{i \in T} l(P_i = L_i) \tag{1}$$

where l equals 1 only if $P_i = L_i$, otherwise 0.

For generative models such as MiniGPT-4 and LLaVA-Med, we report the *accuracy* for closed-ended questions as we leverage prompts to guide the model in answering these questions under a specified candidate set. For open-ended questions, we adopt *recall* to evaluate the ratio that ground-truth tokens appear in the generated sequences and $METEOR$ to assess the word order consistency between generated answer and ground-truth. The *recall* can be formalized as:

$$recall = \frac{TP}{TP + FN} \tag{2}$$

where TP is the number of ground-truth tokens that correctly predicted and FN stands for the number of ground-truth tokens that didn't appear in the predicted answer.

5.4 Results

Tables 5, 7 and 6 show the *accuracy* achieved by all the considered models.

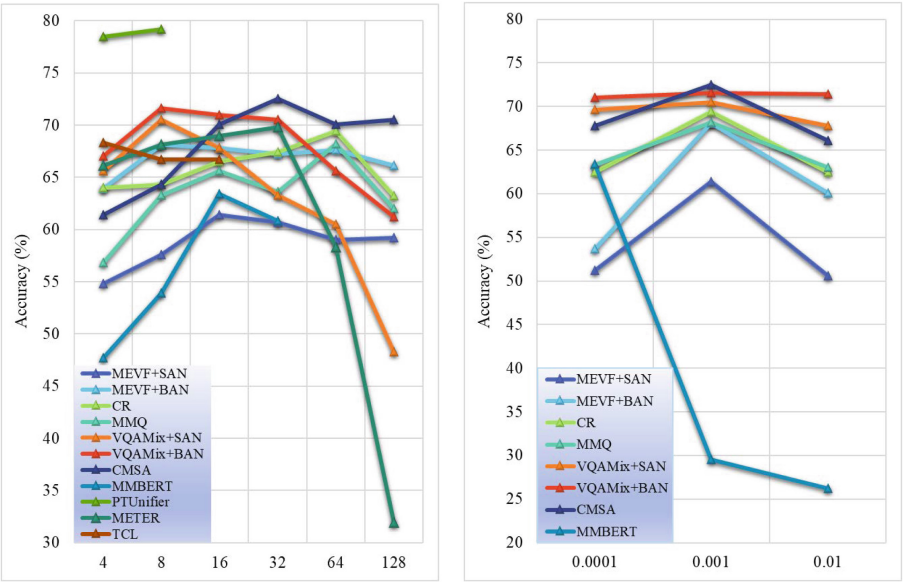
(*i*) In closed-ended questions, discriminative models (Table 5), are more applicable to Med-VQA, compared with LLMs (Table 7). This is because the generative models focus on simulating and generating data that requires broader language understanding and visual information processing capabilities. For simple closed-ended questions, they may suffer from the over-generation problem.

(*ii*) Among discriminative models, the PTUnifier which is pre-trained in the medical domain performs the best on VQA-RAD, SLAKE-EN and OVQA, but not so well on PathVQA and MedVQA-2019. As for the pre-trained models in general domain, TCL and METER achieve better performance on PathVQA and MedVQA-2019. The possible reason is that PathVQA is collected from a wide range of sources, including textbooks and literature, while MedVQA-2019 is artificially generated and cannot represent formal clinical data. PTUnifier adopts a visual language pre-training framework and unifies the fused encoder and dual encoder, thereby excelling on multi-modal tasks.

(*iii*) For generative models, MiniGPT-4 performs worst in terms of both the accuracy and the word order of generating answer on every dataset. Although utilizing massive amounts of data for training, it is still unable to effectively mine the domain-specific knowledge to answer a medical question, then over-generate lots of irrelevant text, and finally resulting in poor performance. In addition, the usage of inappropriate prompts may further degrade the model performance.

(*iv*) The performance of lightweight models such as MEVF, CR, MMQ, and CMSA is significantly inferior to complex models like PTUnifier, TCL, and METER. This is because models like PTUnifier have more parameters and adopt

a deeper neural network structure, which is beneficial for learning the alignment between images and texts.



(a) Batch Size (b) Learning Rate

Fig. 6. Model performance varies with batch size and learning rate

5.5 Detailed Analysis

Figure 6 shows that the values of hyperparameters are determined based on the values set with the best performance on the validation dataset. The results of each model are obtained by changing the Batch Size (BZ) and Learning Rate (LR). Due to limited computing power, we only show parts of the results: (*i*) The results of MiniGPT-4 and LLaVA-Med are eliminated as they cannot be fine-tuned; (*ii*) We show part of results for PTUnifier in Fig. 6(a), as it requires more computing power for larger values of BZ; (*iii*) Similarly, we show part of results for PTUnifier, TCL, and METER with larger number of parameters in Fig. 6(b), as the value range of LR is not comparable to that of other models.

In Fig. 6(a), the performance of each model gradually increases when the BZ values increase, and then decrease after reaching a saddle point, due to the gradient calculation. However, when BZ is set to a large value, some models converge to local stationary points, such as METER and VQAMix-SAN. In Fig. 6(b), (*i*) with the increase of LR values, the performance of MMBERT shows a significant decline, and (*ii*) the performance of MEVF, CR, and CMSA first increase and then decrease with the increase of LR values.

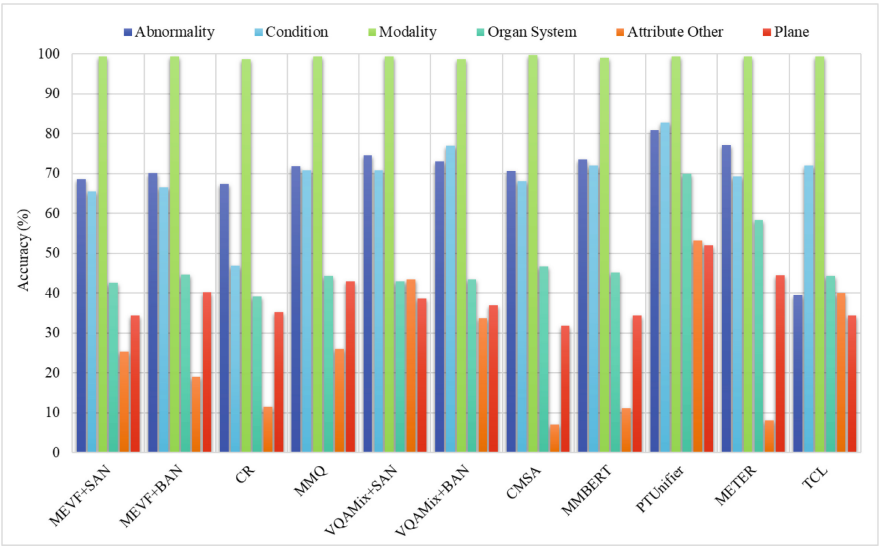
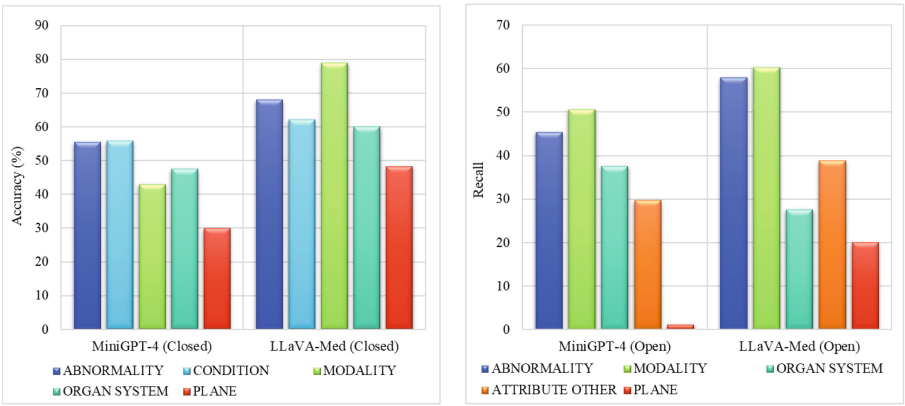


Fig. 7. The *Accuracy* of different question types for discriminative models in OVQA



(a) Closed-ended Question (b) Open-ended Question

Fig. 8. The performance of different question types for LLMs in OVQA

Figures 7 and 8 show the results on various question types for discriminative and generative models over the OVQA dataset, respectively. In Fig. 7, we can derive that:(*i*) All discriminative models perform well on the *Modality* type of questions because MRI or CT image features are obvious, enabling the image encoder to effectively extract image features. (*ii*) All models have unsatisfactory performance on the *Attribute Other* type of questions, as descriptive questions are not suitable for label classification tasks. (*iii*) PTUnifier and VQAMix perform well on most types of questions. PTUnifier introduces visual and textual

prompts for feature representation and improves the diversity of the prompts by constructing prompt pools, which enable different types of questions to select the appropriate prompts and enhance the image-text alignment in the fusion encoder. VQAMix incorporates a conditional label combination strategy for data augmentation, allowing for extracting more comprehensive image features.

In Fig. 8, LLaVA-Med performs better than MiniGPT-4 on almost all types of questions, as it contains extensive domain-specific knowledge by pre-training and instruction tuning based on a large-scale biomedical dataset. Especially, LLaVA-Med greatly outperforms MiniGPT-4 on the *Plane* type of open-ended questions, as these specialized questions require models to fully capture the medical image features and exert domain knowledge to generate answers.

5.6 Qualitative Analysis

We provide a qualitative comparison of all models. Two examples from the OVQA dataset in Fig. 9 show that early discriminative models such as MEVF, CR, MMBERT, CMSA, and VQAMix, fail to answer Med-VQA questions, compared to the latest discriminative models such as TCL, METER, and PTUnifier. In Fig. 9, the Red Cross indicates that the prediction is wrong, and the green check indicates that the prediction is correct. The given question is to consult the abnormal position of orthopedic images. We observed that traditional models such as MEVF predict wrong abnormal positions. While TCL, and other

Q1:Is this image abnormal?
A1:Yes.
MiniGPT-4:It is difficult to determine whether the image is abnormal without more context. It appears to be a medical image, but the condition of the patient's foot is unclear. The image does not show any visible injuries or deformities. ✗
LLaVA-Med: No, the image is not abnormal. It shows a normal CT image of the tibial malleolus. ✗
MEVF+SAN:No. ✗
MEVF+BAN:No. ✗
CR:No. ✗
MMQ:No. ✗
VQAMix+SAN:No. ✗
VQAMix+BAN:No. ✗
CMSA:No. ✗
MMBERT:No. ✗
PTUnifier:Yes.✔
METER:Yes.✔
TCL:Yes.✔

Q2:Are there signs of Comminuted fracture of the left distal radius, Avulsion fracture of the left ulnar styloid process, Fracture of the right radius head, or there's no abnormality?
A2:Comminuted fracture of left distal radius; Avulsion fracture of left ulnar styloid process; Fracture of right radial head.
MiniGPT-4:The distal radius is the lower and larger of the two bones that make up the forearm. It is located at the end of the upper arm bone (humerus) and forms the wrist joint with the carpals (the two bones in the lower arm). The distal radius is a curved bone that forms part of the elbow joint and helps to stabilize the joint during movement. It is also involved in wrist movements such as flexion and extension. ✗
LLaVA-Med:The CT image does not show any signs of a comminuted fracture of the left distal radius, avulsion fracture of the left ulnar styloid process, or fracture of the right radius head. This suggests that the bones in the left forearm appear to be intact and without any significant fractures or damage. ✗
MEVF+SAN:comminuted fracture of left radius head, avulsion fracture of left ulna coronoid process.. ✗
MEVF+BAN:comminuted fracture of the left proximal humerus, avulsion fracture of the left lateral humerus, comminuted fracture of the left proximal ulna, and left distal radius fracture. ✗
CR:fracture of the fifth metacarpal bone of the right hand, with the broken end shifted at an angle. ✗
MMQ:comminuted fracture of left olecranon, fracture of left coronal process, dislocation of left radial head. ✗
VQAMix+SAN:comminuted fracture of left distal radius; avulsion fracture of left ulnar styloid process. ✗
VQAMix+BAN:comminuted fracture of left distal radius, left ulnar styloid fracture. ✗
CMSA:comminuted fracture of left distal radius; avulsion fracture of left ulnar styloid process. ✗
MMBERT:comminuted fracture of right distal radius, avulsion fracture of right ulnar styloid process. ✗
PTUnifier/METER/TCL:Comminuted fracture of left distal radius; Avulsion fracture of left ulnar styloid process; Fracture of right radial head.✔

Fig. 9. Two testing examples selected from OVQA

advanced models can locate the abnormality to the correct position. This also indicates that the advanced VQA deep learning models with large parameters can not only correctly understand the image content, but also capture the region of interest related to the question, leading to predicting the correct answer.

6 Conclusion

Deep learning models for Med-VQA face unique challenges, necessitating urgent comprehensive empirical studies on SOTAs to advance techniques and medical practice. To address this, we implemented a benchmark evaluation system that compares user-selected models and reports detailed experimental results. Additionally, users can download datasets, reports, and source codes for further exploration. Our system provides a unified platform to facilitate diverse medical practices.

Acknowledgement. This work was done when Xiaojie Hong worked for the project in Meetyou AI Lab. Xiaoli Wang was supported by the Natural Science Foundation of Fujian Province of China (No. 2021J01003).

References

1. Achiam, J., et al.: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Ben Abacha, A., Hasan, S.A., Datla, V.V., Demner-Fushman, D., Müller, H.: VQA-med: Overview of the medical visual question answering task at imageclef 2019. In: CLEF, 9–12 September 2019 (2019)
3. Chen, Z., Diao, S., Wang, B., Li, G., Wan, X.: Towards unifying medical vision-and-language pre-training via soft prompts. arXiv preprint arXiv:2302.08958 (2023)
4. Do, T., Nguyen, B.X., Tjiputra, E., Tran, M., Tran, Q.D., Nguyen, A.: Multiple meta-model quantifying for medical visual question answering. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12905, pp. 64–74. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87240-3_7
5. Dou, Z.Y., et al.: An empirical study of training end-to-end vision-and-language transformers. In: CVPR, pp. 18166–18176 (2022)
6. Eslami, S., de Melo, G., Meinel, C.: Does clip benefit visual question answering in the medical domain as much as it does in the general domain? arXiv preprint arXiv:2112.13906 (2021)
7. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: ICML, pp. 1126–1135. PMLR (2017)
8. Gong, H., Chen, G., Liu, S., Yu, Y., Li, G.: Cross-modal self-attention with multi-task pre-training for medical visual question answering. In: ACM ICMR, pp. 456–460 (2021)
9. Gong, H., Chen, G., Mao, M., Li, Z., Li, G.: VQAMix: conditional triplet mixup for medical visual question answering. IEEE Trans. Med. Imaging **41**(11), 3332–3343 (2022)
10. Han, T., et al.: Medalpaca–an open-source collection of medical conversational AI models and training data. arXiv preprint arXiv:2304.08247 (2023)
11. He, X., et al.: Pathological visual question answering. arXiv preprint arXiv:2010.12435 (2020)

12. Huang, Y., Wang, X., Liu, F., Huang, G.: OVQA: a clinically generated visual question answering dataset. In: ACM SIGIR, pp. 2924–2938 (2022)
13. Huang, Y., Wang, X., Su, J.: An effective pre-trained visual encoder for medical visual question answering. In: Yang, X., et al. (eds.) ADMA. LNCS, vol. 14180, pp. 466–481. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-46677-9_32
14. Khare, Y., Bagal, V., Mathew, M., Devi, A., Priyakumar, U.D., Jawahar, C.: Mmbert: Multimodal BERT pretraining for improved medical VQA. In: ISBI, pp. 1033–1036. IEEE (2021)
15. Lau, J.J., Gayen, S., Ben Abacha, A., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. Sci. Data **5**(1), 1–10 (2018)
16. Li, C., et al.: LLaVA-med: training a large language-and-vision assistant for biomedicine in one day. arXiv preprint arXiv:2306.00890 (2023)
17. Liu, B., Zhan, L.M., Xu, L., Ma, L., Yang, Y., Wu, X.M.: Slake: a semantically-labeled knowledge-enhanced dataset for medical visual question answering. In: ISBI, pp. 1650–1654. IEEE (2021)
18. Masci, J., Meier, U., Cireşan, D., Schmidhuber, J.: Stacked convolutional auto-encoders for hierarchical feature extraction. In: Honkela, T., Duch, W., Girolami, M., Kaski, S. (eds.) ICANN 2011. LNCS, vol. 6791, pp. 52–59. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-21735-7_7
19. Nguyen, B.D., Do, T.-T., Nguyen, B.X., Do, T., Tjiputra, E., Tran, Q.D.: Overcoming data limitation in medical visual question answering. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11767, pp. 522–530. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32251-9_57
20. Pelka, O., Koitka, S., Rückert, J., Nensa, F., Friedrich, C.M.: Radiology objects in COntext (ROCO): a multimodal image dataset. In: Stoyanov, D., et al. (eds.) LABELS/CVII/STENT -2018. LNCS, vol. 11043, pp. 180–189. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01364-6_20
21. Sarrouti, M.: NLM at VQA-med 2020: visual question answering and generation in the medical domain (2020)
22. Sharma, D., Purushotham, S., Reddy, C.K.: Medfusenet: an attention-based multimodal deep learning model for visual question answering in the medical domain. Sci. Rep. **11**(1), 19826 (2021)
23. Srivastava, Y., Murali, V., Dubey, S.R., Mukherjee, S.: Visual Question Answering using Deep Learning: A Survey and Performance Analysis, pp. 75–86 (2021)
24. Wang, H., et al.: Huatuo: tuning llama model with Chinese medical knowledge. arXiv preprint arXiv:2304.06975 (2023)
25. Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: PMC-llama: further finetuning llama on medical papers. arXiv preprint arXiv:2304.14454 (2023)
26. Wu, Q., Teney, D., Wang, P., Shen, C., Dick, A., Hengel, A.: Visual question answering: a survey of methods and datasets. Cornell University - arXiv, Cornell University - arXiv (2016)
27. Wu, T., et al.: A brief overview of ChatGPT: the history, status quo and potential future development. IEEE/CAA J. Automatica Sinica **10**(5), 1122–1136 (2023)
28. Xiong, H., et al.: Doctorglm: fine-tuning your Chinese doctor is not a herculean task. arXiv preprint arXiv:2304.01097 (2023)
29. Yang, J., et al.: Vision-language pre-training with triple contrastive learning. In: CVPR, pp. 15671–15680 (2022)
30. Yasunaga, M., Leskovec, J., Liang, P.: LinkBERT: pretraining language models with document links. In: ACL, pp. 8003–8016 (2022)

31. Yunxiang, L., Zihan, L., Kai, Z., Ruilong, D., You, Z.: Chatdoctor: a medical chat model fine-tuned on llama model using medical domain knowledge. arXiv preprint arXiv:2303.14070 (2023)
32. Zhan, L.M., Liu, B., Fan, L., Chen, J., Wu, X.M.: Medical visual question answering via conditional reasoning. In: ACM MM, pp. 2345–2354 (2020)
33. Zhang, X., et al.: PMC-VQA: visual instruction tuning for medical visual question answering. arXiv preprint arXiv:2305.10415 (2023)
34. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)