

## Highlights

### Medical Visual Question Answering: A Survey

Zhihong Lin, Donghao Zhang, Qingyi Tao, Danli Shi, Gholamreza Haffari, Qi Wu, Mingguang He, Zongyuan Ge

- It is the first medical VQA survey paper as the current increasing application demand on medical VQA systems. It describes the history of medical VQA and research directions in the future.
- This survey presents an overview of the publicly available medical VQA datasets. The tasks include previous ImageCLEF VQA-Med challenges and other published datasets. This will help researchers to identify suitable research benchmarks and metrics.
- This survey gives a comprehensive summary and discussion of the published method papers for the medical VQA. It provides the comparison and discussion for the competition work notes and research papers. These will help researchers to better understand model design principles in common and conduct further research.
- It concludes some current challenges and future research directions. The common core of these challenges is the final application in the clinical scenario. The research topics cover dataset design to human-computer interaction.

# Medical Visual Question Answering: A Survey

Zhihong Lin<sup>a</sup>, Donghao Zhang<sup>b</sup>, Qingyi Tao<sup>c</sup>, Danli Shi<sup>d</sup>, Gholamreza Haffari<sup>e</sup>, Qi Wu<sup>f</sup>, Mingguang He<sup>g</sup> and Zongyuan Ge<sup>b,h,i,\*</sup>

<sup>a</sup>Faculty of Engineering, Monash University, Clayton, VIC, 3800 Australia

<sup>b</sup>eResearch Center, Monash University, Clayton, VIC, 3800 Australia

<sup>c</sup>NVIDIA AI Technology Center, 038988, Singapore

<sup>d</sup>State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-Sen University, Guangzhou, 510060 China

<sup>e</sup>Faculty of Information Technology, Monash University, Clayton, 3800, VIC, Australia

<sup>f</sup>Australian Centre for Robotic Vision, The University of Adelaide, Adelaide, SA 5005, Australia

<sup>g</sup>Eye Research Australia, Royal Victorian Eye and Ear Hospital, East Melbourne, VIC, 3002 Australia

<sup>h</sup>Airdoc Research, Melbourne, VIC, 3000 Australia

<sup>i</sup>Monash-NVIDIA AI Tech Centre, Melbourne, VIC, 3000 Australia

## ARTICLE INFO

### Keywords:

Visual Question Answering  
Medical Image Interpretation  
Computer Vision  
Natural Language Processing

## ABSTRACT

Medical Visual Question Answering (VQA) is a combination of medical artificial intelligence and popular VQA challenges. Given a medical image and a clinically relevant question in natural language, the medical VQA system is expected to predict a plausible and convincing answer. Although the general-domain VQA has been extensively studied, the medical VQA still needs specific investigation and exploration due to its task features. In the first part of this survey, we collect and discuss the publicly available medical VQA datasets up-to-date about the data source, data quantity, and task feature. In the second part, we review the approaches used in medical VQA tasks. We summarize and discuss their techniques, innovations, and potential improvements. In the last part, we analyze some medical-specific challenges for the field and discuss future research directions. Our goal is to provide comprehensive and helpful information for researchers interested in the medical visual question answering field and encourage them to conduct further research in this field.

## 1. Introduction

Visual Questing Answering (VQA) [10] is a multidisciplinary problem that incorporates computer vision (CV) and natural language processing (NLP). The VQA system is expected to answer an image-related question according to the image content. Inspired by the VQA research in the general domain, the recent exploration of medical VQA has attracted great interest. The medical VQA system is expected to assist in clinical decision-making and improve patient engagement [33, 46]. Unlike other medical AI applications often restricted to pre-defined diseases or organ types, the medical VQA can understand free-form questions in natural language and provide reliable and user-friendly answers.

In recent research, the medical VQA has been assigned to several “jobs”. The first one is the diagnostic radiologist, who acts as an expert consultant to the referring physician. A workload study [62] shows that the average radiologist has to interpret one CT or MRI image in 3 to 4 seconds. Besides the long queue of imaging studies, a radiologist must also answer an average of 27 phone calls per day from physicians and patients [22], leading to further inefficiencies and disruptions in the workflow. A medical VQA system can potentially answer the physician’s questions and help relieve

the burden of the healthcare system and improve medical professionals’ efficiency. Another application matching the advantage of VQA is to act as the pathologists who examine body tissues and help other healthcare providers make diagnoses [35].

In addition to the health professional role, the medical VQA system can also serve as a knowledgeable assistant. For example, the “second opinion” from the VQA system can support the clinicians’ opinion in interpreting medical images and decrease the risk of misdiagnosis at the same time [88].

Ultimately, a mature and complete medical VQA system can directly review patients’ images and answer any kind of questions. In some situations, such as fully automated health examinations, where medical professionals may not be available, a VQA system can provide equivalent consultation. After a hospital visit, patients search for further information online. The irregular and misleading information from the search engine might result in inappropriate answers. Alternatively, a medical VQA can be integrated into an online consultation system to provide reliable answers anytime and anywhere.

Medical VQA is technically more challenging than general-domain VQA because of the following factors. Firstly, creating a large-scale medical VQA dataset is challenging because expert annotation is expensive for its high requirement of professional knowledge, and QA pairs can not be synthetically generated directly from images. Secondly, answering questions according to a medical image also demands a specific design of the VQA model. The

\*Corresponding author

Email addresses: zhihong.lin@monash.edu (Z. Lin);

donghao.zhang@monash.edu (D. Zhang); qtao002@e.ntu.edu.sg (Q. Tao); shidli@mail2.sysu.edu.cn (D. Shi); gholamreza.haffari@monash.edu (G. Haffari); qi.wu01@adelaide.edu.au (Q. Wu); mingguang.he@unimelb.edu.au (M. He); zongyuan.ge@monash.edu (Z. Ge)

ORCID(s): 0000-0001-8499-4097 (Z. Lin)

task also needs to focus on a fine-grained scale because a lesion is microscopic. Hence, segmentation techniques may be required to locate the region of interest precisely. Finally, a question can be very professional, which requires the model to be trained with medical knowledge base rather than a general language database.

Since the first medical VQA challenge was organized in 2018 [33], an increasing number of organizations and researchers have joined to expand the tasks and propose new datasets and approaches, which have made the medical VQA task an active and inspiring field. To provide a comprehensive retrospect of these efforts, we conduct the first survey (to our best knowledge) for medical VQA.

In the first part of this survey, we overview the publicly available medical VQA datasets up-to-date. To collect the most complete information, we did an exhaustive search for the available medical VQA datasets including sources from relevant papers in google scholar, medical image computing conferences, and top-tier journals, and resulted in a total of 10 papers proposing datasets. Two dataset papers are repetitive and different versions of included datasets, and consequently, 8 datasets are analyzed and discussed in this survey. Among them, three datasets are proposed as ImageCLEF<sup>1</sup> competitions. The selected datasets are diverse in image modality and question categories. The imaging modality of those datasets covers chest X-ray, CT, MRI, and pathology. The questions include close-end questions (such as Yes/No questions) and open-end questions on a variety of topics. We also compare the data sources, the question-answer pairs creation methods, and the metrics for evaluation. These will be inspiring for researchers who are interested in designing new tasks.

In the second part of this survey, we review the published approaches to medical VQA. We gather the work notes describing the approaches used in the ImageCLEF VQA-Med competitions and collect 32 papers in total. However, the work notes are mainly simple solutions because of the time-limited situation. We also search for technical papers from conferences or journals that aim at the current pain point and we collect 13 papers. The papers' sources are from both the community conferences such as Medical Image Computing and Computer Assisted Intervention (MICCAI) and Association for Computing Machinery Multimedia (ACM-MM) and influential journals such as the IEEE Transactions on Medical Imaging. By reviewing those papers, we find that the current approaches are mostly in a framework of four components: image encoder, language encoder, feature fusion module, and answering module. The review of existing approaches will help researchers to identify the key problem in previous research and the potential hypothesis in future research.

Finally, we discuss four medical-specific challenges for the field. The medical-domain VQA is a more application-oriented problem compared with the general-domain VQA.

It has a real application scenario that will produce practical challenges. In this work, we analyze the clinical requirements to develop practical and useful applications and raise six significant challenges: the question diversity, extra medical information, interpretability, generalizability, large language models, and integration in the medical workflow. The proposed challenges will inspire researchers and develop mature and accurate medical VQA systems to support the clinical decision-making process.

## 2. Datasets and performance metrics

### 2.1. Datasets

To the best of our knowledge, there are 8 public-available medical VQA datasets up to date: VQA-MED-2018 [33], VQA-RAD [48], VQA-MED-2019 [14], RadVisDial [46], PathVQA [35], VQA-MED-2020 [13], SLAKE [57], and VQA-MED-2021 [15] (in chronological order). The dataset's details are summarized in Table 1. In the following paragraphs, we provide an overview of the QA pairs collection.

#### 2.1.1. VQA-Med-2018

VQA-Med-2018 [33] is a dataset proposed in the ImageCLEF 2018<sup>3</sup>, and it is the first publicly available dataset in the medical domain. The QA pairs were generated from captions by a semi-automatic approach. First, a rule-based question generation (QG) system<sup>4</sup> automatically generated possible QA pairs by sentence simplification, answer phrase identification, question generation, and candidate questions ranking. Then, two expert human annotators (including one expert in clinical medicine) manually checked all generated QA pairs in two passes. Respectively, one pass ensures semantic correctness, and another ensures clinical relevance to associated medical images.

#### 2.1.2. VQA-RAD

VQA-RAD [48] is a radiology-specific dataset proposed in 2018. The image set is a balanced one containing samples by the head, chest, and abdomen from MedPix<sup>5</sup>. To investigate the question in a realistic scene, the author presented the images to clinicians to collect unguided questions. The clinicians are required to produce questions in both free-from and template structures. Afterward, the QA pairs are validated and classified manually to analyze the clinical focus. The answer types are either close-ended or open-ended. Although without a large quantity, the VQA-RAD dataset has acquired essential information about what a medical VQA system should be able to answer as an AI radiologist.

#### 2.1.3. VQA-Med-2019

VQA-Med-2019 [14] is the second edition of the VQA-Med and was published during the ImageCLEF 2019 challenge. Inspired by the VQA-RAD [48], VQA-Med-2019 has addressed the four most frequent question categories:

<sup>3</sup><https://www.imageclef.org/2018>

<sup>4</sup><http://www.cs.cmu.edu/~rk/mheilman/questions/>

<sup>5</sup><https://medpix.nlm.nih.gov/home>

**Table 1**

Overview of the medical VQA datasets and their main characteristics. Visual Genome, VQA 2.0, and OK-VQA are general-domain VQA datasets listed here for comparison. The medical VQA datasets are presented in chronological order.

Dataset	# Images	# QA pairs	Source of images and content	QA Creation	Question Category
Visual Genome [47]	108K	1,773K	YFCC100M [87] Microsoft COCO [55]	Manual	- Object - Attributes - Relationships
VQA 2.0 [31]	204K	614K	Microsoft COCO	Manually	- Object - Color - Sport - Count - etc.
OK-VQA [61]	14,031	14,055	Microsoft COCO	Manual	- External knowledge
VQA-Med-2018 [33]	2,866	6,413	PubMed Central Articles <sup>2</sup>	Synthetical	- Location - Finding - Yes/No questions - Other questions
VQA-RAD [48]	315	3,515	MedPix database: - Head axial single-slice CTs or MRIs - Chest X-rays - Abdominal axial CTs	Manual	- Modality - Plane - Organ System - Abnormality - Object/Condition Presence - Positional Reasoning - Color - Size - Attribute Other - Counting - Other
VQA-Med-2019 [14]	4,200	15,292	MedPix database: - Various in 36 modalities, 16 planes, and 10 organ systems	Synthetical	- Modality - Plane - Organ system - Abnormality
RadVisDial [46] (Silver-standard)	91,060	455,300	MIMIC-CXR [39]: - Chest X-ray posterior-anterior (PA) view	Synthetical	Abnormality
RadVisDial [46] (Gold-standard)	100	500	MIMIC-CXR [39]: - Chest X-ray posterior-anterior (PA) view	Manual	Abnormality
PathVQA [35]	4,998	32,799	Electronic pathology textbooks PEIR Digital Library	Synthetical	- Color - Location - Appearance - Shape - etc. - Abnormality
VQA-Med-2020 [13] SLAKE [57]	5,000 642	5,000 14K	MedPix database Medical Segmentation Decathlon[83], NIH Chest X-ray[99], CHAOS[42]: - Chest X-rays/CTs - Abdomen CTs/MRIs - Head CTs/MRIs - Neck CTs - Pelvic cavity CTs	Synthetical Manual	- Organ - Position - Knowledge Graph - Abnormality - Modality - Plane - Quality - Color - Size - Shape
VQA-Med-2021 [15]	5,000	5,000	MedPix database	Synthetical	- Abnormality

modality, plane, organ system, and abnormality. For each category, the questions follow the patterns from hundreds of questions naturally asked and validated in the VQA-RAD [48]. The first three categories (modality, plane, and organ system) can be tackled as classification tasks, while the fourth category (abnormality) presents an answer generation problem.

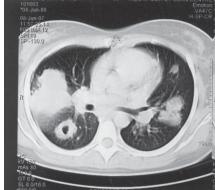
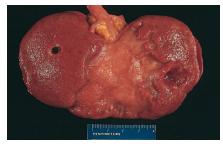
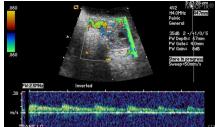
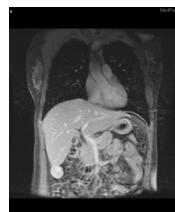
#### 2.1.4. RadVisDial

RadVisDial [46] is the first publicly available dataset for visual dialog in radiology. The visual dialogue consists of multiple QA pairs and is considered a more practical and complicated task for a radiology AI system than VQA. The images are selected from MIMIC-CXR [39]. For each

image, the MIMIC-CXR has provided a well-structured relevant report with annotations for 14 labels (13 abnormalities and one No Findings label). The RadVisDial consists of two datasets: a silver-standard dataset and a gold-standard dataset. In the silver-standard group, the dialogues are synthetically created using the plain text reports associated with each image. Each dialogue contains five questions randomly sampled from 13 possible questions. The corresponding answer is automatically extracted from the source data and limited to four choices (yes, no, maybe, or not mentioned in the report). In the gold-standard group, the dialogues are collected from two expert radiologists' conversations following detailed annotation guidelines to ensure consistency. Only 100 random images are labeled with gold-standard. The RadVisDial dataset explored a real-world scene task of AI in

**Table 2**

Samples of images and question-answer pairs from the mentioned Datasets. Q = Question, A = Answer. The datasets are presented in chronological order.

Dataset	Samples		
VQA-Med-2018 [33]	 <p><b>Q:</b> What does the ct scan of thorax show?  <b>A:</b> bilateral multiple pulmonary nodules</p>	 <p><b>Q:</b> Is the lesion associated with a mass effect?  <b>A:</b> no</p>	
VQA-RAD [48]	 <p><b>Q:</b> What is the organ system?  <b>A:</b> Gastrointestinal</p>	<p><b>Object/Condition Presence</b></p> <p><b>Q:</b> Is there gastric fullness?  <b>A:</b> yes</p>	<p><b>Positional</b></p> <p><b>Q:</b> What is the location of the mass?  <b>A:</b> head of the pancreas</p>
VQA-Med-2019 [14]	 <p><b>Q:</b> what imaging method was used?  <b>A:</b> us-d - doppler ultrasound</p>	<p><b>Modality</b></p> <p><b>Q:</b> Airspace opacity?  <b>A:</b> Yes  <b>Q:</b> Fracture?  <b>A:</b> Not in report</p>	<p><b>Plane</b></p> <p><b>Q:</b> which plane is the image shown in?  <b>A:</b> axial</p>
RadVisDial [46]	 <p><b>Q:</b> Lung lesion?  <b>A:</b> No  <b>Q:</b> Pneumonia?  <b>A:</b> Yes</p>		
PathVQA [35]	 <p><b>Q:</b> What have been stripped from the bottom half of each specimen to show the surface of the brain?  <b>A:</b> meninges</p>	 <p><b>Q:</b> Is remote kidney infarct replaced by a large fibrotic scar?  <b>A:</b> yes</p>	
VQA-Med-2020 [13]	 <p><b>Q:</b> what abnormality is seen in the image?  <b>A:</b> ovarian torsion</p>	 <p><b>Q:</b> what is abnormal in the ct scan?  <b>A:</b> partial anomalous pulmonary venous return</p>	
SLAKE [57]	 <p><b>Q:</b> Does the image contain left lung?  <b>A:</b> Yes</p>	<p><b>Q:</b> What is the function of the rightmost organ in this picture?  <b>A:</b> Breathe</p>	
VQA-Med-2021 [15]	 <p><b>Q:</b> What is most alarming about this mri?  <b>A:</b> focal nodular hyperplasia</p>	 <p><b>Q:</b> What abnormality is seen in the image?  <b>A:</b> Enhancing lesion right parietal lobe with surrounding edema</p>	

the medical domain. Moreover, the team compared the synthetical dialogue to the real-world dialogue and conducted experiments to reflect the importance of context information. The medical history of the patient was introduced and led to better accuracy.

### 2.1.5. PathVQA

PathVQA [35] is a dataset exploring VQA for pathology. The images with captions are extracted from digital resources (electronic textbooks and online libraries). The author developed a semi-automated pipeline to transfer the captions into QA pairs, and the generated QA pairs are manually checked and revised. The question can be divided into seven categories: what, where, when, whose, how, how much/how many, and yes/no. The open-ended questions account for 50.2% of all questions. For the close-ended “yes/no” questions, the answers are balanced with 8,145 “yes” and 8,189 “no”. The questions are designed according to the pathologist certification examination of the American Board of Pathology (ABP). Therefore it is an exam to verify the “AI Pathologist” in decision support. The PathVQA dataset demonstrates that medical VQA can be applied to various scenes.

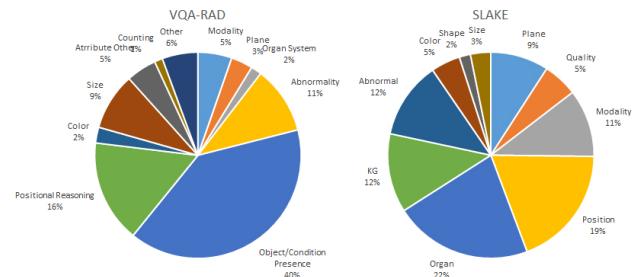
### 2.1.6. VQA-Med-2020

VQA-Med-2020 [13] is the third edition of the VQA-Med and was published in the ImageCLEF 2020 challenge. The images are selected with the limitation that the diagnosis was made according to the image content. The questions are specifically addressing on abnormality. A list of 330 abnormality problems is selected, and each problem needs to occur at least ten times in the dataset. The QA pairs are generated by patterns created.

In VQA-Med-2020, the visual question generation (VQG) task is first introduced to the medical domain. The VQG task is to generate natural language questions relative to the image content. The medical VQG dataset includes 1,001 radiology images and 2,400 associated questions. The ground truth questions are generated with a rule-based approach according to the image captions and manually revised.

### 2.1.7. SLAKE

SLAKE [57] is a comprehensive dataset with both semantic labels and a structural medical knowledge base. The images are selected from three open source datasets [83, 99, 42] and annotated by experienced physicians. The semantic labels for images provide masks (segmentation) and bounding boxes (detection) for visual objects. The medical knowledge base is provided in the form of a knowledge graph. The knowledge graph is extracted from OwnThink and manually reviewed. They are in the form of triplets (e.g., *<Heart, Function, Promote blood flow>*). The dataset contains 2,603 triplets in English and 2,629 triplets in Chinese. The introduction of a knowledge graph allows external knowledge-based questions such as organ function and disease prevention. The questions are collected from experienced doctors by selecting pre-defined questions or



**Figure 1:** The question category distribution of VQA-RAD and SLAKE

rewriting questions. Then the questions are categorized by their types and balanced to avoid bias.

### 2.1.8. VQA-Med-2021

VQA-Med-2021 [13] is published in ImageCLEF 2021 challenge. The VQA-Med-2021 is created under the same principles as those in VQA-Med-2020. The training set is the same dataset used in VQA-Med-2020. The validate set and test set are newly collected and manually reviewed by medical professionals.

### 2.1.9. Discussion

In the above sections, we present 8 medical VQA datasets about their quantity, data source, QA creation, and question categories. As shown in Table 1, we also list three general VQA datasets for comparison. The image amounts of medical VQA datasets range from 315 to 91,060, while the number of QA pairs ranges from 1 QA pair per image to 10 QA pairs per image. The imaging modality includes chest X-ray, CT, MRI, and pathology. Except for RadVisDial, the medical VQA datasets are significantly smaller than the general VQA datasets in quantity. For the QA pair creation, the medical VQA uses synthetic creation more frequently than the general domain VQA. The question categories of medical VQA and general VQA are quite different. Besides the common categories of object and attribute, the general VQA research extends their problem to objects' relationship and external knowledge, while the medical VQA research tends to image findings.

These differences between general VQA datasets and medical VQA datasets reflect the difficulties in medical dataset establishment. The three listed general VQA datasets have all utilized the Microsoft COCO[55], which provides 328K images with natural language descriptions. On the medical side, there is no such large-scale data source. One type of data source is images with a description such as a caption and medical report. The VQA-Med-2018 is the first exploration of the medical VQA dataset. It utilizes the images in articles so that the images have a corresponding textual description. The PathVQA uses images and text from textbooks and the digital library. For these two datasets, the key problem is how to perform an accurate transformation.

Another type of data source is the image with categorical attribution. The VQA-RAD, VQA-Med-2019, VQA-Med-2020, and VQA-Med-2021 are all sourced from the MedPix database, which provides images with attributions. Therefore, the key problem becomes how to acquire better questions given images and answers. The VQA-RAD starts from manual creation and collects unguided problems from clinicians. The question patterns collected from VQA-RAD leverage the construction of VQA-Med-2019, VQA-Med-2020, and VQA-Med-2021. The RadVisDial uses the data from the MIMIC-CXR dataset, which provides extracted disease labels and is the only large-scale data source in medical VQA.

Besides the data source, another difficulty is the professional knowledge required in data annotation. All three general VQA datasets adopt the Amazon Mechanical Turk workers to achieve their large-scale annotation. On the medical side, the QA creation is usually synthetical, and the manual creation is done by medical students or medical experts. The cost of manual data annotation in medical VQA is inevitably higher due to the requirement of professional knowledge.

The above difficulties in medical VQA dataset establishment have raised future challenges. With the development of technology, the existing difficulties can possibly be solved. For example, the recent Large Language Models (LLMs) have been believed good at understanding and generating natural language. The advantage of LLMs may make them ideal annotators for medical VQA. Especially for existing large-scale medical report datasets, i.e., MIMIC-CXR or FFA-IR [51], the LLMs can help with parsing the natural language reports and converting them into medical VQA data.

Another problem is the question categories. As shown in Fig. 1, the VQA-RAD and SLAKE have different question categories distribution as they are created in different ways. Some categories such as modality and plain can help the image viewer to understand the captured information of the image, while the other categories such as abnormality and abnormality attributes can help interpret the image findings. Among all medical VQA datasets, VQA-RAD is the only one collecting the natural questions and representing a question categories distribution from medical students. In contrast, other datasets are all created with pre-defined question categories and the distribution may not represent any real-world demands. In other words, currently, there is no public dataset representing a question distribution from patients in the clinical scene.

Furthermore, the task design and mission are also considerable problems. The recently proposed general VQA datasets have shown their special target, such as data balance, knowledge base, etc. In the medical domain, SLAKE provides more modalities, including segmentation, detection, and knowledge graph. This feature can improve the complexity of tasks and allow more question categories. It also raises a new mission for the approach researchers as it has more modalities to expand the method's complexity.

Despite the manual annotation limit its quantity, the golden standard annotation is prospective to benefit the community and future research.

As the medical VQA research is still in an early stage, the current datasets are only about radiology and pathology in data subjects. There is more field to discover, such as ophthalmology and dermatology, which are also popular in medical AI research and already has existing databases to create potential VQA task. Besides dataset works, there is also exploration addressing data collection efficiency. The MVQAS [11] builds an online system providing self-collected and annotation tools to allow users to upload data and semi-automatically generate VQA triplets.

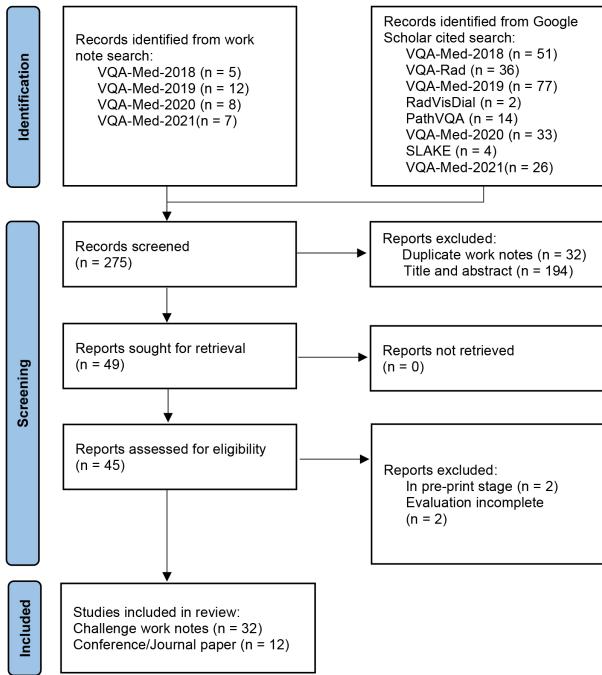
## 2.2. Performance Metrics

The performance metrics used in the proposed medical VQA tasks can be categorized into classification-based metrics and language-based metrics. The classification-based metrics are the general metrics in classification tasks such as accuracy and F1 score. They treat the answer as a classification result and calculate the exact match accuracy, precision, recall, and e.t.c. All eight tasks in this paper use classification-based metrics as part of their performance metrics. The Language metrics are the general metrics for sentence evaluation tasks (e.g., translation, image captioning). The tasks using language-based metrics include VQA-Med-2018, VQA-Med-2019, PathVQA, VQA-Med-2020, and VQA-Med-2021. All of those four tasks use the BLEU [67], which measures the similarity of the phrases (n-grams) between two sentences. However, the BLEU is originally a metric for machine translation and is also used in medical report generation tasks [51]. As shown in Table 2, the ground truth answers in medical VQA are obviously shorter than those of machine translation or medical report generation tasks. Also, for some questions, the semantically positive or negative is more important than the word match. It suggests that BLEU may be an inappropriate metric for current medical VQA datasets. However, the BLEU can still be useful when the answer corpus of the future medical VQA becomes extensive and comprehensive sentences.

Besides the general metrics, there are also custom metrics designed for the medical VQA. For example, the WBSS (Word-based Semantic Similarity) and CBSS (Concept-based Semantic Similarity) are created in the VQA-Med-2018 [33] as new language metrics. However, the metric is not a fixed component of the dataset and the approach. Researchers can alternatively use more suitable metrics to evaluate their results. For example, some researchers [79] introduce the AUC-ROC (Area under the ROC Curve) as their classification metrics to better evaluate the measure of separability.

## 3. Methods

To investigate the feature of approaches used in the medical VQA task, we reviewed the published papers evaluated on the datasets. We search for the method papers with two strategies. For the ImageCLEF competitions, we use the



**Figure 2:** The number of papers included/excluded at the literature review of conference/journal papers

corresponding overview papers [1, 14, 13, 15] to identify the participating teams and collect their work notes. This strategy helps us collect a total number of 32 papers. Then we search Google Scholar with the citation search function to find 217 papers that cite medical VQA datasets and collect a total of 12 papers. For a detailed count of the papers included and excluded at each stage, refer to Fig. 2. Finally, we select 45 published papers, including 32 work notes and 13 conference/journal papers. The 45 papers describe 46 approaches, and the performance and characteristics are shown in Table 3. The following sub-sections discuss the medical VQA methods by the framework, components, other techniques, performance comparison, and overall discussion.

### 3.1. Framework

Among the 46 approaches, 39 of them can be attributed to a common framework in the general VQA domain, the **joint embedding** [10]. It is proposed as the baseline method for the VQA v1 dataset [10] and referred to as the "LSTM Q+I". As illustrated in Fig. 3, the framework includes four components: an image encoder, a question encoder, a feature fusing algorithm, and an answering component according to the task requirement. Respectively, the image feature extractor can be the well-developed convolution neural network (CNN) backbones such as VGG Net [82] and ResNet [34], and the question encoder can be the prevalent language encoding models such as LSTM [37] and the Transformer [92]. The feature encoding models are often initialized with pre-trained weights, and they can be either frozen or fine-tuned

in an end-to-end manner during the training of the VQA task. The answering component is usually a neural network classifier or a recurrent neural network language generator. In the LSTM Q+I, the question features and image features are fused via element-wise multiplication. Then, researchers developed innovative fusing algorithms and introduced the popular attention mechanism into the system to further increase the performance.

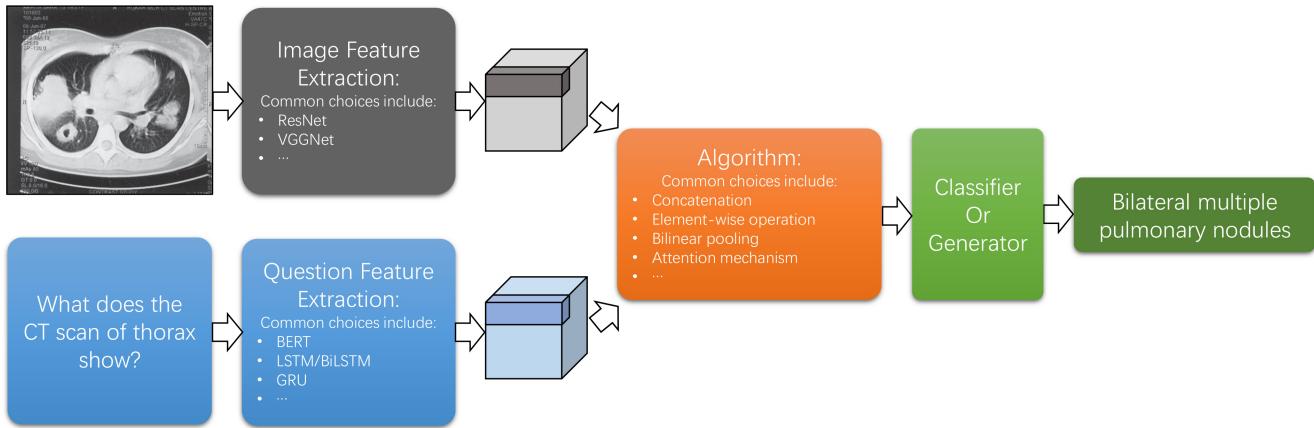
Besides the joint embedding framework, the other 7 approaches choose not to include the question feature. They find the semantic space of the questions is simply due to the data nature and use only the image feature to produce the answer. The framework has outstanding performance in VQA-Med-2020 and VQA-Med-2021, where the questions are only about abnormality and in only "Yes/No" or "What" types.

Compare to the general VQA domain, the architecture that appears in medical VQA research is less diverse. There is also other architecture in the general VQA domain, such as compositional model such as the Neural Module Network [9]. However, no compositional model has been adopted in current medical VQA approaches, and it is also potential to introduce other frameworks to the medical VQA.

### 3.2. Image Encoder

In terms of the image encoder in the public challenges, the VGG Net [82] is the most popular choice. As shown in Fig.4, the participants using VGG Net as the image encoder represent a significant proportion in all of the three VQA-MED challenges. Meanwhile, the pre-trained image encoder is often used for both public challenges and conference/journal works. According to Table 3, more than half of the teams (29 of 46) directly used a pre-trained model on the ImageNet [74]. However, ImageNet has different content compared with medical VQA. Using ImageNet pre-trained is a non-reasonable practice but a workable option when the low data quantity and lack of labels both limit the pre-training on medical datasets.

Finding better pre-training methods is a popular topic in medical VQA research as well as in the medical AI community. The image numbers of most medical VQA datasets are under 5,000. It leads to difficulty in training image representation. The solution proposed includes using other pre-trained models, using the extra dataset, contrastive learning, multi-task pre-training, and meta-learning. The LIST team [6] utilized an image encoder pre-trained on the CheXpert [99] dataset. The MMBERT [43] team uses an auxiliary dataset named ROCO [69] (images and captions) to perform pre-training in a token-masking manner. The CPRD [56] team introduces contrastive learning technology to conduct pre-training with unlabeled images in a self-supervised scheme. Self-supervised training has the advantage that it does not need image labels, which are expensive to acquire for medical images. The MTPT-CSMA [29] team uses an auxiliary dataset of segmentation tasks. On the other hand, researchers also try to further digest the original data.



**Figure 3:** A schematic diagram of the mainstream medical VQA frameworks illustrating main components, including image feature extraction, question feature extraction, feature fusion, and answering component.

The MEVF [64] team proposed the Mixture of Enhanced Visual Features (MEVF) component, which utilizes the Model-Agnostic Meta-Learning (MAML) and the Convolutional Denoising Auto-Encoder (CDAE) to initialize the model weights for the image encoder to overcome the data limitation in quantity. The various exploration on image encoder is specific in medical VQA compared that in general VQA.

Notably, the auxiliary datasets selected have three different types: captioning dataset, unlabeled image, and segmentation dataset. It indicates that the pre-training on various extra data can leverage the image decoder's performance. Hence, exploring the methods to pre-train the image encoder will be a prospective task. Another feature is that all image encoders in the reviewed papers are CNN classification models than detection ones. It restricts the application of detection-based methods such as Up-Down [8], which are popular in the general VQA domain.

### 3.3. Language Encoder

The language encoders in the reviewed works include LSTM [37] (18 of 46), Bi-LSTM [77] (5 of 46), GRU [21] (3 of 38), the Transformer [92] (including BERT [24] and BioBERT [49]) (12 of 46) and the others (8 of 46). Notably, all the top five teams are using The BERT in the VQA-Med-2019 challenge. It is shown in Fig.4 that participating teams select BERT/BioBERT instead of LSTM/Bi-LSTM/GRU over time. This indicates the advantage of the Transformer model and the BERT pre-training despite the corpus differences between the general domain and medical domain. Meanwhile, the Recurrent Neural Networks (LSTM/Bi-LSTM/GRU) users are also adopting the pre-trained word embedding component (12 of 26). It indicates that the researchers tend to use a pre-training model to process the questions. However, only the MMBERT team [43] conducts personalized pre-training with a self-supervised method, the token masking strategy on extra data. Compared with the image encoder, the language encoder does not get much research. It is potential to develop more NLP

pre-training methods or vision plus language pre-training methods for medical VQA.

Especially, 7 teams process the questions without a deep learning model but with keyword or template matching. The reason is that keyword or template matching has been powerful in their tasks. Hence, a light language encoder can be a practical choice in medical VQA applications as the task may not have a large number of question categories.

### 3.4. Fusion Algorithm

The fusion stage gathers the extracted visual feature and language feature and then models the hidden relationship between the language feature and the visual feature. It is the core component of VQA methods. The typical fusion algorithm includes the attention mechanism and the pooling module.

**Attention mechanism** is widely used in vision and language tasks. Among the Medical VQA approaches, 23 of 46 apply attention mechanisms in the fusion stage. Stacked Attention Networks (SAN) [104] is a typical attention algorithm. It uses the question feature as a query to rank the answer-related image regions. With a multiple-layer structure, the SAN can query an image several times to infer the answer progressively. The SAN introduced an incursive attention mechanism and is used as a baseline for many datasets. Besides the SAN, some other works employ the attention mechanism, such as the Bilinear Attention Networks (BAN) [44], the Hierarchical Question-Image Co-Attention (HieCoAtt) [60], e.t.c. Notably, the popular multi-head attention methods(e.g., the Transformer [92], the Modular Co-Attention Network (MCAN) [105]) are seldom applied to medical VQA.

Table 3: The Results and Characteristics of Approaches in medical VQA Tasks.

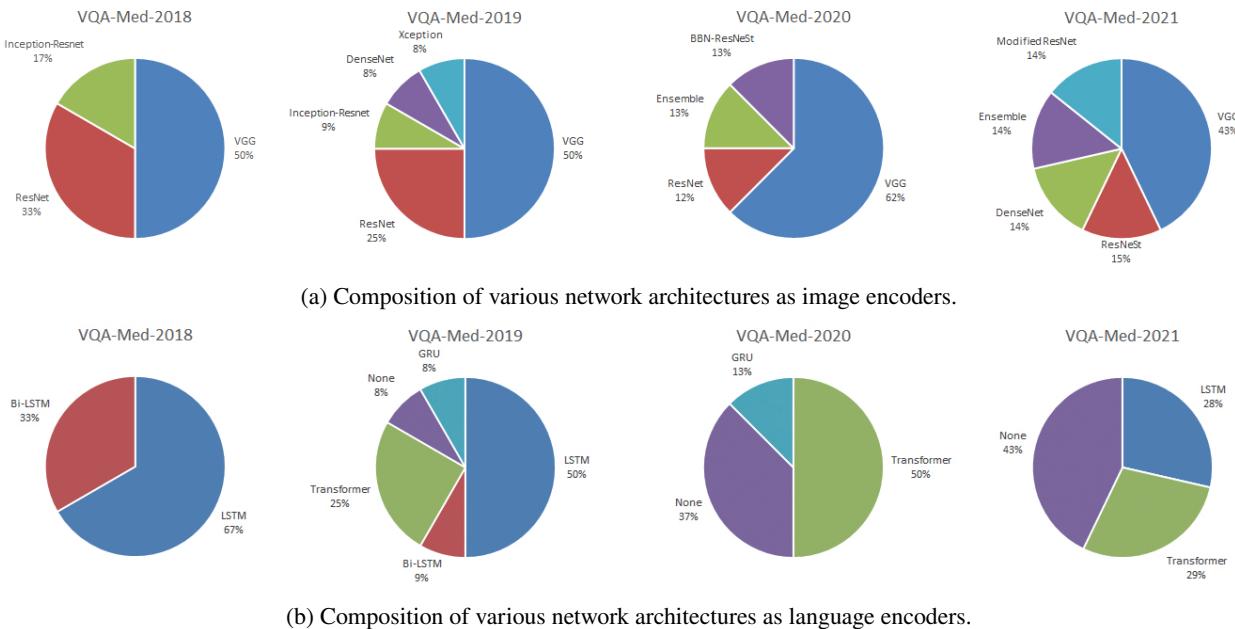
Team/Method	Image Encoder	Pre-trained (Image)	Language Encoder	Pre-trained (Language)	Fusion	Attention (Fusion)	Output Mode	Other Technique(s)	Language Score(BLEU) Accuracy	Classification
<b>VQA-MED-2018</b>										
Chakri* [7]	VGG16	ImageNet	GRU	None	No	Element-wise multiplication (GRU)	Generation		0.188	
UMMS [70]	ResNet-152	ImageNet	LSTM	Pre-trained Word Embedding	Yes	MFB with Co-attention	Classification	Embedding model	topic	0.162
TU [111]	Inception-Resnet-v2	ImageNet	Bi-LSTM	Not mentioned	Yes	Attention mechanism	Classification	Output rules		0.135
HQS* [32]	Inception-Resnet-v2	ImageNet	Bi-LSTM	Pre-trained Word Embedding	No	Concatenation	Classification	Question segregation		0.132
NLM [1]	VGG16	ImageNet	LSTM	None	Yes	SAN	Classification		0.121	
NLM JUST [85]	ResNet-152	ImageNet	LSTM	None	No	MCB	Classification	Pre-training with extra data	0.085	
FSTT [5]	VGG16	ImageNet	Bi-LSTM	Not mentioned	No	Concatenation	Classification (LSTM)	Generation	0.061	
KEML* [109]	VGG16	ImageNet	Transformer	BERT	No	BLOCK	Classification	Decision Tree Classifier	0.054	
<b>VQA-MED-2019</b>										
MedFuseNet* [79]	ResNet-152	ImageNet	LSTM	BERT Embedding	Yes	MedFuseNet	Classification/ Generation (LSTM)	Image Attention, Image-Question Attention	0.27 (Subset)	Knowledge-Based Representation Learning
MMBERT* [43]	ResNet-152	ROCO[69]	Transformer	Yes	Yes	Multi-head attention (Transformer)	Generation	Pre-training with extra data	0.69	0.789 (Subset)
CGMVQA* [73]	ResNet-152	ImageNet	Transformer	BERT	Yes	Concatenation	Classification (Transformer)	Global Average Pooling	0.659	0.64
Hanlin [102]	VGG16	ImageNet	Transformer	BERT	Yes	MFB with Co-attention	Classification (Transformer)	Global Average Pooling	0.644	0.62

Team/Method	Image Encoder	Pre-trained (Image)	Language Encoder	Pre-trained (Language)	Attention (Fusion)	Fusion	Output Mode	Other Technology(s)	Language Score (BLEU)	Classification Accuracy
minhv [93]	ResNet-152	ImageNet	Transformer	BERT	Yes	MLB, MUTAN with attention	Classification	Ensemble, Skip-thought	0.634	0.616
TUA1 [112]	Inception-Resnet-v2	ImageNet	Transformer	BERT	No	Concatenation	Classification Generation (LSTM)	Sub-models, Question classifier	0.633	0.606
QC-MLB* [94]	ResNet-152	ImageNet	Skip-thought vectors	Yes	Yes	QC-MLB	Classification	Question-centric fusion	0.603	
UMMS [80]	ResNet-152	ImageNet	Bi-LSTM	Pre-trained Word Embedding	Yes	MFH with Co-attention	Classification	SVM Question classifier	0.593	0.566
IBM Research AI [45]	VGG16	ImageNet	LSTM	Pre-trained Word Embedding	Yes	Attention mechanism	Question classifier	Image size encoder	0.582	0.558
LIST [6]	DenseNet-121	CheXpert	LSTM	Pre-trained Word Embedding	No	Classification Concatenation	Generation (LSTM)	Generation	0.583	0.556
Turner.JCE [90]	VGG19	Not mentioned	LSTM	Not mentioned	No	Concatenation	Classification	Sub-models, Question classifier	0.572	0.536
JUST19 [4]	VGG16	ImageNet	None	None	No	None	Classification/Generation	Sub-models, Question classifier	0.591	0.534
Team_PwC-Med [12]	ResNet-50	ImageNet	LSTM	Pre-trained Word Embedding	Yes	Attention mechanism	Classification/Generation (LSTM)	Sub-models, Question classifier	0.534	0.488
Techno [16]	VGG16	Not mentioned	LSTM	Not mentioned	No	Concatenation	Classification	Classification	0.486	0.462
Gasm* [28]	EfficientNet	ImageNet	Bi-LSTM	Not mentioned	No	Concatenation	Classification	Optimal parameter selection	0.42	0.391
Dear stranger [59]	Xception	Not mentioned	GRU	Yes	Attention mechanism	Classification	Element-wise multiplication	Generation (LSTM)	0.393	0.21
abhishek-thanki [86]	VGG19	ImageNet	LSTM	Pre-trained Word Embedding	No	None	Generation	Sub-models	0.462	0.16
<b>VQA-MED-2020</b>										
AIML [53]	Ensemble CNNs	Not mentioned	None	Not mentioned	No	None	Classification	Multi-scale and architecture ensemble	0.542	0.496
TheInception-Team [3]	VGG16	Not mentioned	None	None	No	None	Classification	Sub-models	0.511	0.48

Team/Method	Image Encoder	Pre-trained (Image)	Language Encoder	Pre-trained (Language)	Attention (Fusion)	Fusion	Output Mode	Other Technology(s)	Language Score (BLEU)	Classification Accuracy
bumjun-jung [40]	VGG16	ImageNet	Transformer	BioBERT	Yes	MFH with Co-attention	Global Average Pooling	0.502	0.466	
HCP-MIC [20]	BBN-ResNeSt-50	Not mentioned	Transformer	BioBERT	No	None	Bilateral-Branch Network	0.462	0.426	
NLM [75]	ResNet-50	ImageNet	None	Not mentioned	No	None	Classification	0.441	0.4	
HARENDRAKV [41]	VGG16	Not mentioned	Transformer	BERT	Yes	MFB	Generation (LSTM)	0.439	0.378	
Shengyan [58]	VGG16	ImageNet	GRU	Not mentioned	No	None	Generation (GRU)	0.412	0.376	
kdevqa [91]	VGG16	Not mentioned	Transformer	BERT	No	GLU	Classification	0.35	0.314	
<b>VQA-MED-2021</b>										
SYSU-HCP [30]	ResNets, VGGNet, and plus HAGAP	ImageNet	None	None	No	None	Classification	Hierarchically global average pooling, Model ensemble, Mixup Augment, Curriculum learning, Label smoothing	0.416	0.382
Yunnan University [10]	VGG16	ImageNet	Transformer	BioBERT	Yes	MFH with Co-attention	Global Average Pooling	0.402	0.362	
TeamS [26]	ResNeSt50	Not mentioned	None	None	No	None	Classification	Bilateral-Branch Networks	0.391	0.348
Lijie [50]	VGG8	ImageNet	Transformer	BioBERT	Yes	MFH with Co-attention	Classification	0.352	0.316	
IALab_PUC [76]	DenseNet-121	ImageNet	None	None	No	None	Classification	0.276	0.236	
TAM [52]	Modified ResNet-34	Not mentioned	LSTM	GLOVE word embeddings	Yes	MFB with a co-attention	Classification	0.255	0.222	
Sheerin [84]	VGGNet	ImageNet	LSTM	Not mentioned	No	Element-wise multiplication	Generation(LSTM)	0.227	0.196	
<b>VQA-RAD</b>										
MTPT-CMSA [29]	Multi-ResNet-34	Multi-task	LSTM	Pre-trained Word Embedding	Yes	CSMA	Classification	Cross-modal self-attention, Multi-task pre-training with extra data	0.732	

Team/Method	Image Encoder	Pre-trained (Image)	Language Encoder	Pre-trained (Language)	Fusion (Fusion)	Attention Fusion	Output Mode	Other Technology(s)	Language Score (BLEU)	Classification Accuracy
CPRD [56]	ResNet-8	Contrastive	LSTM	Pre-trained Word Embedding	Yes	BAN	Classification	Constrastive Learning, Knowledge Distillation	0.727	
MMBERT [43]	ResNet-152	ROCO	Transformer	Pre-trained Word Embedding	Yes	Multi-head attention (Transformer)	Generation (Transformer)	Pretraining with extra data	0.72	
QCR [108]	MEVF	None	LSTM	Not mentioned	Yes	BAN/SAN	Classification	Question-Conditioned Reasoning, Type-Conditioned Reasoning	0.716	
MMQ [25]	MMQ	None	LSTM	Pre-trained Word Embedding	Yes	BAN/SAN	Classification	Multiple Quantifying	0.67	
MEVF [64]	MEVF	None	LSTM	Not mentioned	Yes	BAN/SAN	Classification	Model-Agnostic Learning, Convolutional Auto-Encoder	0.439/0.751 (0.627)	
HQS [32]	Inception-Resnet-v2	ImageNet	Bi-LSTM	Pre-trained Word Embedding	No	Concatenation	Classification	Question Segregation	0.411	
<b>PathVQA</b>										
MedFuseNet [79]	ResNet-152	ImageNet	LSTM	BERT Embedding	Yes	MedFuseNet (LSTM)	Classification/ Generation	Image Attention, Image-Question Attention	0.605 (Subset)	0.636 (Subset)
MMQ [25]	MMQ	None	LSTM	Pre-trained Word Embedding	Yes	BAN/SAN	Classification	Multiple Quantifying	Meta-model	0.488
<b>SLAKE</b>										
CPRD [56]	ResNet-8	Contrastive	LSTM	Pre-trained Word Embedding	Yes	BAN	Classification	Constrastive Learning, Pre-training with extra data, Knowledge Distillation	0.821	

The “\*” means the approach is not proposed during the public challenge period.



**Figure 4:** The encoders used in VQA-Med challenges.

**Multi-modal pooling** is another important technique used for fusing visual and language features. The basic practice includes concatenation, sum, and element-wise product. In the reviewed papers, direct concatenation is the most widely used fusion method (10 of 46) but shows average performance. As the outer product may be computation-cost expensive when input vectors are with high dimensionality, researchers have proposed more efficient pooling methods. Multi-modal Compact Bilinear (MCB) pooling [27] is a typical method for pooling. It aggregates visual and text features by embedding the image and text features into higher-dimensional vectors and then convolving them with multiplications in the Fourier space for efficiency. The attention mechanism can also be used in the pooling module. Regarding the multi-modal pooling family, there are some other works such as the Multi-modal Factorized High-Order (MFH) pooling [107], the Multi-modal Factorized Bilinear (MFB) pooling [106], e.t.c. The pooling with attention method is the second adopted method (8 of 46). Both the winners in VQA-Med-2018 and VQA-Med-2019 are using MFB pooling with attention solutions.

Among the 46 approaches, the QC-MLB [94] and MedFuseNet [79] are the only two works that proposed an innovative fusion algorithm. The QC-MLB uses a multi-glimpse attention mechanism to ensure that the question feature selects the proper image regions to answer. The MedFuseNet uses two attention modules: Image Attention and Image-Question Co-Attention to let the image feature and question feature interact twice. The QC-MLB and MedFuseNet all show performance improvement. However, the medical VQA research focusing on fusion schemes is still insufficient at this stage.

### 3.5. Answering Component

Among the 46 reviewed approaches, 33 approaches choose the classification mode for output, and 8 methods choose the generation mode. Some methods [4, 73, 12, 112, 79] use a switching strategy to adopt both classification and generation. The selection of the classification method or generation method reflects the length distribution of ground-truth answers belonging to a single phrase. The classification mode will have an advantage within a small answer space. However, it can become difficult if the answer candidates are longer and with more complex information, such as lesion descriptions.

### 3.6. Other Techniques

Other than research in the basic components, some task-specific strategies are also applied to improve performance. Sub-task strategy is the approach that divides the overall task into several sub-tasks and assigns branch models. In these cases, an extra task classification module is applied to select the corresponding model by a particular question category or a type of image modality. It is often used [45, 4, 112, 90, 80, 53, 3, 20] especially in VQA-Med-2019 (questions are only in four categories) and VQA-Med-2020 (questions are all in one category but in two types). The reason is that the questions have only distinct categories to be divided easily. These multiple model approaches may improve the single model effectiveness but lead to another risk that they may suffer from inappropriate model allocation due to the erroneous results from the sub-task classification module. Another problem is that these approaches cannot be generalized to other tasks if the question categories are different.

Another frequently applied technique is Global Average Pooling(5 of 46). Global Average Pooling [54] is using the average of feature maps to replace the last fully connected

layers. It is believed to produce better image representation. Other techniques include Embedding-based Topic Model, Question-Conditioned Reasoning, and Image Size encoder.

### 3.7. Performance Comparison

As the performance shown in Table 3, the keys of state-of-the-art approaches on each dataset are quite different. Among the VQA-Med-2018 approaches, the top team Chakri [7] is the only team that uses generative output. For the VQA-Med-2019, the key factor is the language encoder model. The BERT users significantly achieve a higher performance than the LSTM users. In VQA-Med-2020 and VQA-Med-2021, the top teams AIML [53] and SYSU-HCP [30] all removed the language encoders and adopted ensemble strategy for image encoder, as the question categories were reduced. For VQA-RAD, the top three teams all have extra pre-training on their image encoders. The above comparison shows the enhancement of the image encoder is the essential ingredient of achieving state-of-the-art performance in most medical VQA datasets.

Another finding is that over most datasets, the approaches with attention-based fusion algorithms are averagely outperforming the ones without attention. For example in VQA-Med-2019, the approaches with attention-based fusion have an average accuracy of 0.576, while the others have an average accuracy of 0.522. A similar conclusion can also be drawn for other datasets. It suggests that attention-based fusion algorithms are suitable for medical VQA datasets.

### 3.8. Overall Discussion

In the method survey, we reviewed 45 papers on medical VQA approaches, including 32 challenges work notes and 13 conference/journal papers. Although medical VQA research just started in 2018, there have already been various methods proposed and explored. Since the challenges only have restricted time for problem-solving, the work notes tend to make direct applications of well-verified deep learning models. They have a high proportion of using pre-trained components. On the other hand, they also develop some task-oriented techniques according to the intrinsic data property. Notably, there are two teams [20, 26] observing the long-tailed distribution in their tasks and introducing the Bilateral-Branch Network (BBN) [110] to deal with the imbalanced data. Despite insufficient innovations, the challenge teams' observations and strategies are valuable and inspiring.

The conference/journal papers explore further problem solutions. According to the 13 conference/journal papers, the current research is more focused on the image encoder than the other component, which is quite different from the general VQA research. Researchers introduce popular CV field ideas such as meta-learning and contrastive learning to enhance the image encoder. Pre-training is commonly applied in both the image encoder and the language decoder for answering. Auxiliary data is also a simple but efficient solution. Current research shows that acquiring a generalized

image encoder is a high-priority and specific task in medical VQA research.

Another research topic that appeared in those technical papers is the generalizability of proposed methods. Among the 13 papers, only 5 papers are aware of showing their generalizability and evaluating their approaches on multiple datasets. And only one team [94] evaluate their approach not only on a medical VQA dataset but also on general VQA datasets. It will be a future standard that an approach should be evaluated on multiple datasets.

Interpretability is also a demand in medical AI research. Among the 13 papers, 5 papers have illustrated the model visualization with techniques such as GradCAM and attention visualization. Although the SLAKE dataset has the annotation for semantic segmentation and object bounding box, none of them use the SLAKE to evaluate their visualization.

## 4. Medical VQA v.s. General VQA

Medical VQA and general VQA are two approaches that harness the power of visual content to answer questions but in distinct domains and contexts. While both aim to interpret and provide meaningful responses based on visual information, they differ significantly in their applications, objective, datasets, methods, domain knowledge, and evaluation. By understanding the unique characteristics and requirements of each approach, we can gain insights into their respective contributions and impact in their respective fields. In this section, we undertake a comparative analysis of medical VQA and general VQA across various dimensions.

**Application** Commonly, general VQA and medical VQA have potential applications of human-computer interaction such as image interpreting and education. The difference is the general VQA can be embedded into information retrieval such as search engines and virtual assistants. Also, it has potential applications in further human-computer interaction such as navigation and robotics. Medical VQA focuses on specific areas such as clinical decision support, telemedicine, and patient empowerment. Overall, the modes of interaction can be more diverse in general VQA applications.

**Objective** While both medical VQA and general VQA aim to bridge the gap between visual perception and natural language understanding, they differ in their functions and the specific information they focus on. medical VQA is designed to assist in medical diagnosis, treatment, and decision-making by leveraging visual medical data and providing accurate answers to medical-related questions. Hence, the objective of medical VQA systems is to gain ability in medical image understanding, abnormality locating, and terminology communication. On the other hand, the objective of general VQA is to enable machines to understand and respond to questions about general visual content, such as everyday images or videos. General VQA systems aim to comprehend the visual scene, recognize objects, infer relationships and attributes, and generate accurate answers in natural language.

**Datasets** As discussed in Sec. 2.1.9, the medical VQA datasets have a smaller quantity because of the limit of data source and high knowledge requirement for annotation. Also, most of the current medical VQA datasets are less diverse in the question category. Besides, each dataset is designed with specific objectives in mind, which can vary depending on the dataset’s focus and intended applications. The general VQA datasets have already focused on specific problems such as answer balance, while most of the medical VQA datasets are at the stage of expanding topics and application scenes.

**Method** The development of methods has been hindered by the complexity of the dataset. General VQA methods have expanded to encompass various topics, such as multi-task learning, logical reasoning, and interaction with the environment. In the general domain, the research on image and language understanding has primarily been undertaken by the upstream community, namely computer vision and NLP. General methods often rely on universal pre-trained models from the upstream community, with a primary focus on multi-modal fusion and reasoning. Conversely, the medical community lacks such universal resources, resulting in medical VQA works placing less emphasis on fusion algorithms and instead prioritizing image encoder pre-training as discussed in Sec. 3.8.

**Domain knowledge** In general VQA, domain knowledge is instrumental in designing effective models and algorithms. For example, different strategies may be needed for object recognition, scene understanding, or reasoning about relationships between objects in images. In the case of medical VQA, domain knowledge is crucial for understanding the specific medical concepts, terminology, and context involved in the questions and images. Understanding the unique characteristics and challenges of medical imaging data, such as different modalities (e.g., X-ray, MRI), image quality variations, and anatomical complexities, is essential for developing robust and accurate medical VQA systems. It is challenging to integrate medical domain knowledge into medical VQA design.

**Evaluation** Due to the critical nature of medical information, medical VQA systems are expected to provide highly accurate and reliable answers. Evaluation metrics for Medical VQA might emphasize medical relevance, precision, and recall, considering the importance of accurate medical information. General VQA systems may prioritize a wider range of plausible answers rather than strictly aiming for accuracy. General VQA evaluation metrics may prioritize overall answer correctness, language understanding, and image understanding. Especially, medical datasets are naturally imbalanced distribution and long-tail distribution. Traditional metrics may not fully capture the performance and provide misleading results. The medical VQA can introduce metrics such as Average Precision (AP) for bias problems, while the general VQA pursues balanced distribution.

## 5. Challenge and Future Works

After reviewing medical VQA datasets and approaches, we identify some existing problems. Compared to the general domain VQA, the specific requirement and practical implementation scene also lead to unique and new challenges. Besides, the works on general-domain VQA provide some inspiration for future research direction.

### 5.1. Question Diversity

Question diversity is one of the most significant challenges of medical VQA. The VQA-RAD [48] investigates the natural questions in clinical conversation. The questions can be categorized into modality, plane, organ system, abnormality, object/condition presence, positional reasoning, color, size, attribute other, counting, and other. In other datasets such as VQA-Med-2019 [14] and PathVQA [35], the question categories tend to be less diverse than the VQA-RAD dataset. In the RadVisDial [46], VQA-Med-2020 [13], and VQA-Med-2021 [15], the question category is reduced to the abnormality presence only. However, most questions about an abnormality in existing datasets are about presence without further inquiry, like the location of tumors or tumor size. Therefore, questions remain to be added to diversify the medical VQA dataset.

To create better medical VQA and benefit clinical workflow integration, future research should be conducted about identifying the useful question categories in practical requirements. For example, the imaging modality and examined organ are totally not required because they are already in the study record information. This kind of information may be delivered to the end users outside the VQA system. First, the data source can be expanded. The synthesized QA pairs in current medical VQA datasets are usually created from image captions and medical reports. However, the information in captions and reports is restricted to specific topics. Therefore, more data sources, such as the textbook, should be considered to provide more textual corpus. Second, collecting real-world clinical conversations, especially with patients, will help researchers better understand the practical requirements. Third, the restriction that the question must be relevant to the image content should be broken. However, in a realistic clinical scene, the conversation content often exceeds the presented image content, for example, explaining the future risks of the abnormality or predicting disease progression. Overall, the question diversity is an essential consideration in dataset design to empower medical VQA with comprehensive coverage, real-world relevance, and user engagement.

Besides dataset design, question diversity raises a challenge in method development. To answer the diverse questions, the medical VQA systems require various reasoning abilities. For example, as a sample shown in Table 2, to answer question “*What is the function of the rightmost organ in this picture?*”, the model should understand the region described, identify the organ in the region of interest, and finally answer the function of the organ. Besides the basic

image and language understanding, medical domain knowledge is a critical ability required, which includes knowledge of anatomical structures, medical procedures, diseases, medical imaging modalities, treatment options, and clinical practices. More specially, for the question categories in Table 1, *Modality, Plane* need knowledge about radiology examination; *Organ, System, Abnormality, Object/Condition Presence* need knowledge about human anatomy and medicine; *Positional Reasoning, Color, Size, Attribute Other, Counting* need general knowledge and reasoning; *Knowledge Graph* need to combine the upon knowledge with the knowledge triplets given in dataset. Therefore, to address question categories correspondingly, the future medical VQA should be equipped with computer-aid diagnosis, general language understanding, reasoning, knowledge integration, and contextual understanding.

Finally, the evaluation of the model performance should also consider the question of diversity. The imbalanced question category distribution can provide misleading results under overall evaluation. To mitigate the impact of imbalanced data on overall evaluation, it is important to consider specialized evaluation strategies. This can involve using category-specific metrics or weighted overall metrics. Furthermore, incorporating language-based evaluation metrics will be a challenge with the presence of verbose answers. Both the correctness of the answers and the quality of the language used should be evaluated to encourage a more comprehensive medical VQA system.

## 5.2. Integrating the Extra Medical Information

Another challenge for the medical VQA is to integrate the extra information into the inference procedure. For example, Kovaleva et al. found that incorporating the medical history of the patient leads to better performance in answering questions [46].

### 5.2.1. EHR

An electronic health record (EHR) contains a patient's medical history, diagnoses, medications, treatment plans, etc. In the medical AI domain, the EHR has been proven useful for disease prediction [81]. Hence it should also be helpful in medical VQA. To support the research of EHR in medical VQA, a new dataset containing the original EHR or synthetical EHR is required. Furthermore, the VQA model should also be modified because the EHR contains a lot of metric variables that should be treated as numbers other than natural language. As far as we know, no previous research has investigated combining EHR and computer vision. Using numeric variables as VQA input and the corresponding fusion algorithm also has not been researched. Therefore encoding EHR in VQA is meaningful and worth studying.

### 5.2.2. Multiple Images

In the medical scene, especially radiology, it is general that the study is based on multiple images other than a single image. Multiple images can contain different scan planes and sequential slices to support the decision-making of medical

professionals. For example, in MIMIC-CXR [39], a radiology report is often associated with two images of postero-anterior view and lateral view. However, in RadVisDial [46], the authors keep only the postero-anterior view, although the abandoned images are also informative. The reason may be that the common VQA or VisDial models only support single-image input. Therefore, the VQA datasets with multiple input images are required to support the model research. Future research should also be devoted to developing the VQA model by taking multiple images as input.

## 5.3. Interpretability and Reliability

Interpretability is a long-standing problem of deep learning. As the VQA models are commonly based on deep learning methods, the medical VQA has to deal with interpretability. To the medical VQA system, interpretability determines the reliability of the predicted answer, and it is more important than that in the general domain because the wrong decision may lead to catastrophic consequences. The general-domain VQA researchers have addressed this problem and investigated several directions to evaluate the inference ability of a model.

### 5.3.1. Unimodal Bias of VQA Models

In the VQA field, the unimodal bias means that the model may answer the question based on statistical regularities from one modality without considering the other modality. Particularly, the bias to language input is also called language prior. The researchers have noticed bias since the first VQA dataset was proposed. They tested the question-only model "LSTM Q", which has only a slightly lower performance compared with the standard model "LSTM Q+I" [10] and indicates the language prior. To better measure the effect of language prior problem of VQA models, Agrawal et al. presented new splits of the VQA v1 and VQA v2 datasets with changing priors (respectively VQA-CP v1 and VQA-CP v2). They also proposed a Grounded Visual Question Answering model (GVQA) to prevent the model from "cheating" by primarily relying on priors [2]. Alternatively, Ramakrishnan et al. proposed a novel regularization scheme that poses training as an adversarial game between the VQA model and a question-only model to discourage the VQA model from capturing language biases [72]. Cadene et al. proposed the RUBi to reduce biases in VQA models. It minimizes the importance of the most biased examples and implicitly forces the VQA model to use the two input modalities instead of relying on statistical regularities between the question and the answer [19].

The unimodal bias in general-domain VQA has been discussed in many aspects, such as benchmark and training schemes. These provide a good starting point for further work in medical VQA. Benchmarks similar to VQA-CP are required to investigate the unimodal bias problem of models. As the quantity of medical VQA datasets is not so large as general-domain VQA datasets, the solution for unimodal bias also needs research.

### 5.3.2. External Knowledge

In VQA research, external knowledge means the model may need to incorporate an external knowledge base to infer the answer besides the image and the question representation. In general-domain VQA, an “adult-level common sense” is required to support the cognition and inference in question answering [100]. Wang et al. proposed an approach named “Ahab” to provide explicit knowledge-base reasoning [95]. They also provide a dataset KB-VQA and a protocol to evaluate the knowledge-based methods. Furthermore, Wang et al. proposed the FVQA dataset providing a supporting fact that is critical for answering each visual questions [96]. On the other hand, to test the VQA methods’ ability to retrieve relevant facts without a structured knowledge base, Marino et al. proposed the OK-VQA dataset, including only questions that require external knowledge resources. They also proposed a knowledge-based baseline named “ArticleNet”, which retrieved some articles from Wikipedia for each question image pair and then trained a network to find the answer in the retrieved articles.

The challenge in medical VQA is that the knowledge requirement is much higher than “adult-level common sense”. The future research includes identifying questions that required external knowledge, exploiting the existing medical knowledge base such as [63], and building a medical VQA dataset with a structured or unstructured knowledge base. There are also additional studies required to determine whether the medical knowledge-based VQA needs specific approaches.

Moreover, external knowledge can also expand the function of medical VQA in clinic workflow. In the practical environment, the questions from the patient can start from image findings and move further to topics like possible disease, potential risk, and disease control. To answer the questions of those topics, an external knowledge base is required. A medical VQA system equipped with external knowledge will have a strong topic capability and become a powerful tool in clinic workflow.

### 5.3.3. Evidence Verification

Evidence verification is a common method to verify model interpretability. In the medical domain, illustrating evidence of the abnormality will gain the trust of medical professionals and patients. A typical method to collect model evidence is through visualization, such as feature attribution or saliency map. In terms of the multi-modal models with attention mechanism, an attention map can be visualized to show the image regions that lead to the answer [104]. By comparing the attention map with annotated human visual attention, the researchers can verify whether the VQA methods use the correct evidence to get the answer. In general-domain VQA, Das et al. proposed the VQA-HAT dataset containing annotated human visual attention to evaluate the attention maps both qualitatively (via visualizations) and quantitatively (via rank-order correlation) [23]. To extend the verification to textual evidence, Park et al. proposed the VQA-X dataset containing human-annotated both textual

explanations and visual explanations [68]. To address text-based question-answer pairs in the VQA task, Wang et al. proposed the EST-VQA dataset annotating the bounding boxes of correct text clues [98]. Alternatively, Jiang et al. proposed the IQVA dataset annotating the eye-tracking data as human visual attention [38].

Although those studies have established a complete scheme, including the annotated evidence, reasoning model, and evaluation metrics, several problems are to be explored when applying evidence verification in medical VQA. First, the previous studies showed different annotation modes such as bounding box and eye-tracking. It is unknown what is the most suitable mode for medical VQA. For some question categories, such as the modality, the answer evidence can be hard to explain with a certain region and may need a text-based explanation. Second, manual annotation for a medical task is expensive because it requires the professional skill of radiologists. In this situation, the existing medical multi-task datasets such as [89] can be beneficial by providing comprehensive knowledge of medical terms. Therefore, the evidence verification approaches for medical VQA will need to consider generating a text-based explanation and utilizing the existing annotated medical datasets.

### 5.3.4. Summarization

These three directions examine the different reasoning procedures of VQA models and provide performance evaluation methods. The VQA models have to look into the image, learn the extra knowledge, and find the correct evidence to pass the “exams”. To our best knowledge, there are only a few works considering interpretability in medical VQA, such as the SLAKE providing lesion annotation and the CPRD, MedFuseNet providing visualization. However, there is no work really performing quantified measurement of interpretability. Some currently feasible ways include giving the attribution of images against problems and using the SLAKE dataset to calculate the overlapping rate between saliency maps and the annotations. It remains to be done building more quantified benchmarks to evaluate medical VQA interpretability.

If the medical VQA system can be verified with its inference ability, it will become a more convincing and reliable tool. It is also helpful to present the knowledge and evidence used in inference explicitly. Hence the evaluation of inference ability should be regarded as a more important benchmark than the answer accuracy. Hopefully, the medical VQA will answer the question “why” in the nearest future.

## 5.4. Generalizability

Generalizability is a universal topic in medical AI research and an inevitable issue for applications running in practical scenarios. The cause of the Generalizability problem is the practical input can be out of the distribution (OOD) of the training data. The factors can be various, such as the patient race and imaging devices. The gap between different data distributions is named *domain shift*. For downstream tasks like VQA, the generalizability problem is two-fold and more comprehensive. Firstly, as discussed

in Sec. 3, the VQA models usually consist of several sub-models which may have a pre-trained weight. At this stage, a domain shift between pre-train data and current training data is introduced. Secondly, after a VQA model is developed and deployed, there will be a domain shift between training data and practical data, which is usually evaluated by cross-dataset validation.

Among the reviewed methods, several approaches have considered the domain shift in sub-model pre-training and acquiring medical pre-trained models as image encoders [56]. However, there is no study considering the potential domain shift in language encoders. Also, the domain shift crossing medical VQA datasets has not yet been considered and studied. With the growing number of medical VQA datasets works, there has been sufficient material for transferring learning studies such as domain adaptation and domain generalization. Measuring and improving model generalizability will be a feasible and meaningful research topic.

## 5.5. Large Language Models

Large Language Models (LLMs) are highly complex artificial intelligence systems that have the capability to learn from the vast amounts of available text data [71]. Discriminative LLMs (e.g., BERT) have been well-studied for answering questions in a classification manner. However, the generative capability of LLMs enables them to answer diverse questions more flexibly and effectively with human-like responses. Generative LLMs have already started to show their potential and remarkable results in the field of the medical domain, where models like Open AI's GPT-3 [17], GPT-4 [18] have successfully passed a part of the US medical licensing exam [65]. Thus, it is interesting to explore the potential approaches of how these LLMs could be of aid in the task of VQA. As the task of medical VQA involves understanding the spatial relationships in a medical image, it is challenging for LLMs that have been primarily trained on text data. However, LLMs have the potential to be integrated into improving the QA capability of a model. Below we describe certain directions of how LLMs could be beneficial to the task of medical VQA.

As a straightforward way, LLMs can be used to process the questions and generate responses based on the extracted features from the images provided by an image-based model. For example in the recent ChatCAD framework [97], the authors converted the outputs provided by different image-based models (classification, segmentation) into textual information and provided it to ChatGPT [78] for refinement and understanding. Then, follow-up queries were asked by the authors about the disease condition of the patient and ChatGPT provided comprehensive answers by combining the given outputs from the image-based models and its learned knowledge from the huge corpus of data. Another recent framework Prophet [103] improved the answer generation capability of a trained VQA model by encoding answer heuristics generated by the VQA model into a prompt for GPT-3 to better comprehend the task thus making better use of the potential of GPT-3. Prophet framework was able

to achieve a new state-of-the-art performance on two challenging knowledge-based VQA datasets. Some works [66] have also employed LLMs to improve the interpretability of the existing image-based models by generating specific attributes that could be analyzed in an image. In addition to the above directions, LLMs can also adapt to multiple modalities, such as patient history, demographics, lab results, etc. which if combined with the current medical VQA models could lead to a more comprehensive analysis and better answering of the related questions.

Although LLMs have a huge potential to be integrated into the task of medical VQA, there are some associated potential pitfalls. First, although LLMs are trained on vast amounts of text corpus data, they might not possess the same domain-specific knowledge expertise as medical professionals in specific domains. Second, LLMs have learned biases or misinformation inherently present in their training data, which could lead to incorrect conclusions. Third, LLMs are not able to express uncertainty for their answers, and thus LLMs could confidently generate incorrect responses or hallucinations [65], which could be fatal in medical decision-making. So, careful attention and high-level collaboration are required to mitigate the above issues. Specifically, involving medical experts in the framework development and fine-tuning process can help to tailor the model responses corresponding to specific medical domains. Moreover, diverse, accurate, and representative real-world medical training data could help to minimize the above biases and generate factually correct responses.

## 5.6. Integration in Medical Workflow

The community has a long history of trying to integrate AI decision-supporting systems into the clinical flow. Integrating the medical VQA system into clinical practice will provide efficient communication and serve as an effective assistant. Hekler et al. found that the combination of human and artificial intelligence can achieve superior results over the independent results of both of these systems in skin cancer classification [36]. Furthermore, Tschandl et al. found that physicians with AI-based support outperformed either AI or physicians. Meanwhile, the least experienced clinicians gain the most from AI-based support [88]. Tschandl et al. also experimented with the accuracy improvements of human-computer collaboration in skin cancer recognition and found multiclass probabilities outperformed either content-based image retrieval (CBIR) or malignancy probability in the mobile technolog [88]. These findings indicate that many factors must be considered in developing successful outcomes for medical AI-supporting system integration, such as clinicians' cognitive style, cognitive error, personality, experience, and acceptance of AI.

Moreover, most AI decision support systems used in the clinical setting are good at addressing prepared questions only. In contrast, the advantage of medical VQA is that it understands free-form questions and provides real-time communication. To address this advantage, the workflow

can remove redundant querying and get simplified. For example, the presence of abnormality may be always asked, and it can be given before the QA session. Hence, selecting useful question categories and keeping online learning may help to maintain a functional question database. The ultimate goal of a medical VQA system is answering open-ended questions (what, which, e.t.c.). This goal will introduce more complexity and uncontrolled factors under the “human-in-the-loop” hypothesis. For example, the medical VQA system may need to resolve a dispute when its opinion is different from the clinician’s. Hence, the training corpus may also prepare for a negotiation. More efforts and studies need to be conducted in this field for a successful deployment of medical VQA into the clinical pipeline. Improving workflow efficiency and service quality is also required to study multiple aspects, such as time spent, collaborative answer accuracy, and user satisfaction.

## 6. Conclusion

This article presented a survey of the datasets and approaches of medical VQA. We collect information about 8 medical VQA datasets and 45 papers describing medical VQA approaches. We conduct a comprehensive discussion on dataset creation, approach framework, approach components, and corresponding techniques. In addition to the descriptive review, we identified some challenges worth exploring in future research. The medical VQA system mainly faces the following four challenges: first, how to get the system answers a range of more comprehensive question categories; second, how to combine medical features with the task; third, how to verify the evidence of an answer to make it more convincing; forth, how to make the system not bias to any modality; finally, how to maximize the benefit of the medical VQA in the workflow. For future work, we propose the following directions. To investigate potential implementation in a real-world scene is necessary. Furthermore, we should pay attention to the conversation between medical professionals and non-professionals. The advantage of VQA is to understand natural language questions and help non-professionals. We should also introduce meaningful ideas in the general domain, such as evidence verification, bias analysis, external knowledge database, etc. These will help us build a practicable and convincing medical VQA system toward medical AI’s ultimate goal.

## 7. Declaration of Interests

We declare that we have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

- [1] Abacha, A.B., Gayen, S., Lau, J.J., Rajaraman, S., Demner-Fushman, D., 2018. NLM at ImageCLEF 2018 visual question answering in the medical domain., in: CLEF (Working Notes).
- [2] Agrawal, A., Batra, D., Parikh, D., Kembhavi, A., 2018. Don’t just assume; look and answer: Overcoming priors for visual question answering, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Los Alamitos, CA, USA. pp. 4971–4980.
- [3] Al-Sadi, A., Al-Theiabat, H., Al-Ayyoub, M., 2020. The inception team at VQA-Med 2020: Pretrained VGG with data augmentation for medical vqa and vqg, in: CLEF 2020 Working Notes.
- [4] Al-Sadi, A., Talafha, B., Al-Ayyoub, M., Jararweh, Y., Costen, F., 2019. JUST at ImageCLEF 2019 visual question answering in the medical domain., in: CLEF (Working Notes).
- [5] Allaouzi, I., Ahmed, M.B., 2018. Deep neural networks and decision tree classifier for visual question answering in the medical domain., in: CLEF (Working Notes).
- [6] Allaouzi, I., Ahmed, M.B., Benamrou, B., 2019. An encoder-decoder model for visual question answering in the medical domain., in: CLEF (Working Notes).
- [7] Ambati, R., Reddy Dudyala, C., 2018. A sequence-to-sequence model approach for imageclef 2018 medical domain visual question answering, in: 2018 15th IEEE India Council International Conference (INDICON), pp. 1–6. doi:[10.1109/INDICON45594.2018.8987108](https://doi.org/10.1109/INDICON45594.2018.8987108).
- [8] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L., 2018. Bottom-up and top-down attention for image captioning and visual question answering, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Los Alamitos, CA, USA. pp. 6077–6086.
- [9] Andreas, J., Rohrbach, M., Darrell, T., Klein, D., 2016. Neural module networks, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Los Alamitos, CA, USA. pp. 39–48.
- [10] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C., Parikh, D., 2015. VQA: Visual question answering, in: 2015 IEEE International Conference on Computer Vision (ICCV), IEEE Computer Society, Los Alamitos, CA, USA. pp. 2425–2433.
- [11] Bai, H., Shan, X., Huang, Y., Wang, X., 2021. MVQAS: A Medical Visual Question Answering System. Association for Computing Machinery, New York, NY, USA. p. 4675–4679.
- [12] Bansal, M., Gadgil, T., Shah, R., Verma, P., 2019. Medical visual question answering at Image CLEF 2019-VQA Med, in: CLEF (Working Notes).
- [13] Ben Abacha, A., Datla, V.V., Hasan, S.A., Demner-Fushman, D., Müller, H., 2020. Overview of the VQA-Med task at ImageCLEF 2020: Visual question answering and generation in the medical domain, in: CLEF 2020 Working Notes, CEUR-WS.org, Thessaloniki, Greece.
- [14] Ben Abacha, A., Hasan, S.A., Datla, V.V., Liu, J., Demner-Fushman, D., Müller, H., 2019. VQA-Med: Overview of the medical visual question answering task at imageclef 2019, in: CLEF2019 Working Notes, CEUR-WS.org, Lugano, Switzerland.
- [15] Ben Abacha, A., Sarouti, M., Demner-Fushman, D., Hasan, S.A., Müller, H., 2021. Overview of the VQA-Med task at ImageCLEF 2021: Visual question answering and generation in the medical domain, in: CLEF 2021 Working Notes, CEUR-WS.org, Bucharest, Romania.
- [16] Bounaama, R., Abderrahim, M.E.A., 2019. Tlemcen University at ImageCLEF 2019 visual question answering task., in: CLEF (Working Notes).
- [17] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., 2020. Language models are few-shot learners. Advances in neural information processing systems 33, 1877–1901.
- [18] Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S., et al., 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. arXiv preprint arXiv:2303.12712 .
- [19] Cadene, R., Dancette, C., Cord, M., Parikh, D., et al., 2019. RUBi: Reducing unimodal biases for visual question answering. Advances in Neural Information Processing Systems 32, 841–852.
- [20] Chen, G., Gong, H., Li, G., 2020. HCP-MIC at VQA-Med 2020: Effective visual representation for medical visual question answering,

- in: CLEF 2020 Working Notes.
- [21] Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar. pp. 1724–1734. doi:[10.3115/v1/D14-1179](https://doi.org/10.3115/v1/D14-1179).
- [22] Cross, N.M., Wildenberg, J., Liao, G., Novak, S., Bevilacqua, T., Chen, J., Siegelman, E., Cook, T.S., 2020. The voice of the radiologist: Enabling patients to speak directly to radiologists. Clinical imaging 61, 84–89.
- [23] Das, A., Agrawal, H., Zitnick, C.L., Parikh, D., Batra, D., 2016. Human attention in visual question answering: Do humans and deep networks look at the same regions?, in: Conference on Empirical Methods in Natural Language Processing.
- [24] Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186.
- [25] Do, T., Nguyen, B.X., Tjiputra, E., Tran, M., Tran, Q.D., Nguyen, A., 2021. Multiple meta-model quantifying for medical visual question answering, in: de Bruijne, M., Cattin, P.C., Cotin, S., Padov, N., Speidel, S., Zheng, Y., Essert, C. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, Springer International Publishing, Cham. pp. 64–74.
- [26] Eslami, S., de Melo, G., Meinel, C., 2021. TeamS at VQA-Med 2021: BBN-Orchestra for long-tailed medical visual question answering. Working Notes of CLEF 201.
- [27] Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M., 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, Texas. pp. 457–468.
- [28] Gasmi, K., Ltaifa, I.B., Lejeune, G., Alshammari, H., Ammar, L.B., Mahmood, M.A., 2021. Optimal deep neural network-based model for answering visual medical question. Cybernetics and Systems 0, 1–22. doi:[10.1080/01969722.2021.2018543](https://doi.org/10.1080/01969722.2021.2018543).
- [29] Gong, H., Chen, G., Liu, S., Yu, Y., Li, G., 2021a. Cross-modal self-attention with multi-task pre-training for medical visual question answering, in: Proceedings of the 2021 International Conference on Multimedia Retrieval, Association for Computing Machinery, New York, NY, USA. p. 456–460. doi:[10.1145/3460426.3463584](https://doi.org/10.1145/3460426.3463584).
- [30] Gong, H., Huang, R., Chen, G., Li, G., 2021b. SYSU-HCP at VQA-Med 2021: A data-centric model with efficient training methodology for medical visual question answering. Working Notes of CLEF 201.
- [31] Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D., 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering, in: Conference on Computer Vision and Pattern Recognition (CVPR).
- [32] Gupta, D., Suman, S., Ekbal, A., 2021. Hierarchical deep multimodal network for medical visual question answering. Expert Systems with Applications 164, 113993. doi:<https://doi.org/10.1016/j.eswa.2020.113993>.
- [33] Hasan, S.A., Ling, Y., Farri, O., Liu, J., Müller, H., Lungren, M.P., 2018. Overview of ImageCLEF 2018 medical domain visual question answering task., in: CLEF (Working Notes).
- [34] He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Los Alamitos, CA, USA. pp. 770–778.
- [35] He, X., Zhang, Y., Mou, L., Xing, E., Xie, P., 2020. PathVQA: 30000+ questions for medical visual question answering. arXiv preprint arXiv:2003.10286 .
- [36] Hekler, A., Utikal, J.S., Enk, A.H., Hauschild, A., Weichenthal, M., Maron, R.C., Berking, C., Haferkamp, S., Klode, J., Schadendorf, D., et al., 2019. Superior skin cancer classification by the combination of human and artificial intelligence. European Journal of Cancer 120, 114–121.
- [37] Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Computation 9, 1735–1780.
- [38] Jiang, M., Chen, S., Yang, J., Zhao, Q., 2020. Fantastic answers and where to find them: Immersive question-directed visual attention, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Los Alamitos, CA, USA. pp. 2977–2986.
- [39] Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S., 2019. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 .
- [40] Jung, B., Gu, L., HaradaAl-Sadi, T., 2020. bumjun\_jung at VQA-Med 2020: VQA model based on feature extraction and multi-modal feature fusion, in: CLEF 2020 Working Notes.
- [41] K. Verma, H., Ramachandran S., S., 2020. HARENDRAKV at VQA-Med 2020: Sequential VQA with attention for medical visual question answering, in: CLEF 2020 Working Notes.
- [42] Kavur, A.E., Selver, M.A., Dicle, O., Barış, M., Gezer, N.S., 2019. CHAOS - Combined (CT-MR) Healthy Abdominal Organ Segmentation Challenge Data. doi:[10.5281/zenodo.3362844](https://doi.org/10.5281/zenodo.3362844).
- [43] Khare, Y., Bagal, V., Mathew, M., Devi, A., Priyakumar, U.D., Jawa-har, C., 2021. Mmbert: Multimodal bert pretraining for improved medical vqa, in: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), pp. 1033–1036. doi:[10.1109/ISBI48211.2021.9434063](https://doi.org/10.1109/ISBI48211.2021.9434063).
- [44] Kim, J., Jun, J., Zhang, B., 2018. Bilinear attention networks, in: Bengio, S., Wallach, H.M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, Montréal, Canada. pp. 1571–1581.
- [45] Kornuta, T., Rajan, D., Shivade, C., Asseman, A., Ozcan, A.S., 2019. Leveraging medical visual question answering with supporting facts, in: CLEF (Working Notes).
- [46] Kovaleva, O., Shivade, C., Kashyap, S., Kanjaria, K., Wu, J., Ballah, D., Coy, A., Karargyris, A., Guo, Y., Beymer, D.B., et al., 2020. Towards visual dialog for radiology, in: Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing, pp. 60–69.
- [47] Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al., 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision 123, 32–73.
- [48] Lau, J.J., Gayen, S., Abacha, A.B., Demner-Fushman, D., 2018. A dataset of clinically generated visual questions and answers about radiology images. Scientific Data 5, 1–10.
- [49] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J., 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36, 1234–1240.
- [50] Li, J., Liu, S., 2021. Lijie at ImageCLEFmed VQA-Med 2021: Attention model based on efficient interaction between multimodality. Working Notes of CLEF 2021.
- [51] Li, M., Cai, W., Liu, R., Weng, Y., Zhao, X., Wang, C., Chen, X., Liu, Z., Pan, C., Li, M., et al., 2021a. FFA-IR: Towards an explainable and reliable medical report generation benchmark, in: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2).
- [52] Li, Y., Yang, Z., Hao, T., 2021b. TAM at VQA-Med 2021: A hybrid model with feature extraction and fusion for medical visual question answering. Working Notes of CLEF 2021.
- [53] Liao, Z., Wu, Q., Shen, C., van den Hengel, A., Verjans, J., 2020. AIML at VQA-Med 2020: Knowledge inference via a skeleton-based sentence mapping approach for medical domain visual question answering, in: CLEF 2020 Working Notes.

- [54] Lin, M., Chen, Q., Yan, S., 2013. Network in network. arXiv preprint arXiv:1312.4400 .
- [55] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft COCO: Common objects in context, in: European conference on computer vision, Springer. pp. 740–755.
- [56] Liu, B., Zhan, L.M., Wu, X.M., 2021a. Contrastive pre-training and representation distillation for medical visual question answering based on radiology images, in: de Bruijne, M., Cattin, P.C., Cotin, S., Padov, N., Speidel, S., Zheng, Y., Essert, C. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, Springer International Publishing, Cham. pp. 210–220.
- [57] Liu, B., Zhan, L.M., Xu, L., Ma, L., Yang, Y., Wu, X.M., 2021b. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. arXiv:2102.09542.
- [58] Liu, S., Ding, H., Zhou, X., 2020. Shengyan at VQA-Med 2020: An encoder-decoder model for medical domain visual question answering task, in: CLEF 2020 Working Notes.
- [59] Liu, S., Ou, X., Che, J., Zhou, X., Ding, H., 2019. An Xception-GRU model for visual question answering in the medical domain, in: CLEF (Working Notes).
- [60] Lu, J., Yang, J., Batra, D., Parikh, D., 2016. Hierarchical question-image co-attention for visual question answering, in: Advances in Neural Information Processing Systems, pp. 289–297.
- [61] Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R., 2019. OK-VQA: A visual question answering benchmark requiring external knowledge, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Los Alamitos, CA, USA. pp. 3190–3199.
- [62] McDonald, R.J., Schwartz, K.M., Eckel, L.J., Diehn, F.E., Hunt, C.H., Bartholmai, B.J., Erickson, B.J., Kallmes, D.F., 2015. The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload. Academic radiology 22, 1191–1198.
- [63] Müller, L., Gangadharaiyah, R., Klein, S.C., Perry, J., Bernstein, G., Nurkse, D., Wailes, D., Graham, R., El-Kareh, R., Mehta, S., et al., 2019. An open access medical knowledge base for community driven diagnostic decision support system development. BMC medical informatics and decision making 19, 93.
- [64] Nguyen, B.D., Do, T.T., Nguyen, B.X., Do, T., Tjiputra, E., Tran, Q.D., 2019. Overcoming data limitation in medical visual question answering, in: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.T., Khan, A. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, Springer International Publishing, Cham. pp. 522–530.
- [65] Nori, H., King, N., McKinney, S.M., Carignan, D., Horvitz, E., 2023. Capabilities of gpt-4 on medical challenge problems. arXiv preprint arXiv:2303.13375 .
- [66] Oikarinen, T., Das, S., Nguyen, L.M., Weng, T.W., . Label-free concept bottleneck models, in: International Conference on Learning Representations.
- [67] Papineni, K., Roukos, S., Ward, T., Zhu, W.J., 2002. BLEU: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA. pp. 311–318.
- [68] Park, D., Hendricks, L., Akata, Z., Rohrbach, A., Schiele, B., Darrell, T., Rohrbach, M., 2018. Multimodal explanations: Justifying decisions and pointing to the evidence, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Los Alamitos, CA, USA. pp. 8779–8788.
- [69] Pelka, O., Koitka, S., Rückert, J., Nensa, F., Friedrich, C.M., 2018. Radiology objects in COntext (ROCO): a multimodal image dataset, in: Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis. Springer, pp. 180–189.
- [70] Peng, Y., Liu, F., Rosen, M.P., 2018. UMass at ImageCLEF medical visual question answering (Med-VQA) 2018 task, in: CLEF (Working Notes).
- [71] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al., 2018. Improving language understanding by generative pre-training .
- [72] Ramakrishnan, S., Agrawal, A., Lee, S., 2018. Overcoming language priors in visual question answering with adversarial regularization, in: Advances in Neural Information Processing Systems, pp. 1541–1551.
- [73] Ren, F., Zhou, Y., 2020. CGMVQA: A new classification and generative model for medical visual question answering. IEEE Access 8, 50626–50636.
- [74] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., 2015. ImageNet large scale visual recognition challenge. International Journal of Computer Vision 115, 211–252.
- [75] Sarouti, M., 2020. NLM at VQA-Med 2020: Visual question answering and generation in the medical domain, in: CLEF 2020 Working Notes.
- [76] Schilling, R., Messina, P., Parra, D., Lobel, H., 2021. PUC chile team at VQA-Med 2021: approaching vqa as a classification task via fine-tuning a pretrained cnn. Working Notes of CLEF 201.
- [77] Schuster, M., Paliwal, K.K., 1997. Bidirectional recurrent neural networks. IEEE transactions on Signal Processing 45, 2673–2681.
- [78] Shao, Z., Yu, Z., Wang, M., Yu, J., 2023. Prompting large language models with answer heuristics for knowledge-based visual question answering. arXiv preprint arXiv:2303.01903 .
- [79] Sharma, D., Purushotham, S., Reddy, C.K., 2021. Medfusenet: An attention-based multimodal deep learning model for visual question answering in the medical domain. Scientific Reports 11, 1–18.
- [80] Shi, L., Liu, F., Rosen, M.P., 2019. Deep multimodal learning for medical visual question answering, in: CLEF (Working Notes).
- [81] Shickel, B., Tighe, P.J., Bihorac, A., Rashidi, P., 2018. Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE Journal of Biomedical and Health Informatics 22, 1589–1604. doi:10.1109/JBHI.2017.2767063.
- [82] Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition, in: Proceedings of the 3rd International Conference on Learning Representations.
- [83] Simpson, A.L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., van Ginneken, B., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., Bilic, P., Christ, P.F., Do, R.K.G., Golub, M., Golia-Pernicka, J., Heckers, S.H., Jarnagin, W.R., McHugo, M.K., Napel, S., Vorontsov, E., Maier-Hein, L., Cardoso, M.J., 2019. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv:1902.09063.
- [84] Sitara, N.M.S., Kavitha, S., 2021. SSN MLRG at VQA-Med 2021: An approach for VQA to solve abnormality related queries using improved datasets. Working Notes of CLEF 201.
- [85] Talafha, B., Al-Ayyoub, M., 2018. JUST at VQA-Med: A VGG-Seq2Seq model, in: CLEF (Working Notes).
- [86] Thanki, A., Makkithaya, K., 2019. MIT Manipal at ImageCLEF 2019 visual question answering in medical domain, in: CLEF (Working Notes).
- [87] Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J., 2016. YFCC100M: The new data in multimedia research. Communications of the ACM 59, 64–73.
- [88] Tschandl, P., Rinner, C., Apalla, Z., Argenziano, G., Codella, N., Halpern, A., Janda, M., Lallas, A., Longo, C., Malvehy, J., et al., 2020. Human-computer collaboration for skin cancer recognition. Nature Medicine 26, 1229–1234.
- [89] Tschandl, P., Rosendahl, C., Kittler, H., 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data 5. doi:10.1038/sdata.2018.161.
- [90] Turner, A., Spanier, A., 2019. LSTM in VQA-Med, is it really needed? JCE study on the ImageCLEF 2019 dataset, in: CLEF (Working Notes).

- [91] Umada, H., Aono, M., 2020. kdevqa at VQA-Med 2020: focusing on GLU-based classification, in: CLEF 2020 Working Notes.
- [92] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need, in: Advances in Neural Information Processing Systems, pp. 5998–6008.
- [93] Vu, M., Sznitman, R., Nyholm, T., Löfstedt, T., 2019. Ensemble of streamlined bilinear visual question answering models for the ImageCLEF 2019 challenge in the medical domain, in: CLEF 2019.
- [94] Vu, M.H., Löfstedt, T., Nyholm, T., Sznitman, R., 2020. A question-centric model for visual question answering in medical imaging. *IEEE Transactions on Medical Imaging* 39, 2856–2868. doi:10.1109/TMI.2020.2978284.
- [95] Wang, P., Wu, Q., Shen, C., Dick, A., van den Hengel, A., 2017a. Explicit knowledge-based reasoning for visual question answering, in: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17, pp. 1290–1296.
- [96] Wang, P., Wu, Q., Shen, C., Dick, A., van den Hengel, A., 2018. FVQA: Fact-based visual question answering. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 40, 2413–2427.
- [97] Wang, S., Zhao, Z., Ouyang, X., Wang, Q., Shen, D., 2023. Chat-CAD: Interactive computer-aided diagnosis on medical image using large language models. arXiv preprint arXiv:2302.07257.
- [98] Wang, X., Liu, Y., Shen, C., Ng, C., Luo, C., Jin, L., Chan, C., van den Hengel, A., Wang, L., 2020. On the general value of evidence, and bilingual scene-text visual question answering, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Los Alamitos, CA, USA. pp. 10123–10132.
- [99] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M., 2017b. ChestX-Ray8: Hospital-scale chest X-Ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Los Alamitos, CA, USA. pp. 3462–3471. doi:10.1109/CVPR.2017.369.
- [100] Wu, Q., Teney, D., Wang, P., Shen, C., Dick, A., van den Hengel, A., 2017. Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding* 163, 21–40. Language in Vision.
- [101] Xiao, Q., Zhou, X., Xiao, Y., Zhao, K., 2021. Yunnan university at VQA-Med 2021: Pretrained BioBERT for medical domain visual question answering. Working Notes of CLEF 201.
- [102] Yan, X., Li, L., Xie, C., Xiao, J., Gu, L., 2019. Zhejiang University at ImageCLEF 2019 visual question answering in the medical domain, in: CLEF (Working Notes).
- [103] Yang, Y., Panagopoulou, A., Zhou, S., Jin, D., Callison-Burch, C., Yatskar, M., 2022. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. arXiv preprint arXiv:2211.11158 .
- [104] Yang, Z., He, X., Gao, J., Deng, L., Smola, A., 2016. Stacked attention networks for image question answering, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Los Alamitos, CA, USA. pp. 21–29.
- [105] Yu, Z., Yu, J., Cui, Y., Tao, D., Tian, Q., 2019. Deep modular co-attention networks for visual question answering, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Los Alamitos, CA, USA. pp. 6274–6283.
- [106] Yu, Z., Yu, J., Fan, J., Tao, D., 2017. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering, in: 2017 IEEE International Conference on Computer Vision (ICCV), IEEE Computer Society, Los Alamitos, CA, USA. pp. 1839–1848.
- [107] Yu, Z., Yu, J., Xiang, C., Fan, J., Tao, D., 2018. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE Transactions on Neural Networks and Learning Systems* 29, 5947–5959. doi:10.1109/TNNLS.2018.2817340.
- [108] Zhan, L.M., Liu, B., Fan, L., Chen, J., Wu, X.M., 2020. Medical visual question answering via conditional reasoning, in: Proceedings of the 28th ACM International Conference on Multimedia (MM '20), ACM.
- [109] Zheng, W., Yan, L., Wang, F.Y., Gou, C., 2020. Learning from the guidance: Knowledge embedded meta-learning for medical visual question answering, in: Yang, H., Pasupa, K., Leung, A.C.S., Kwok, J.T., Chan, J.H., King, I. (Eds.), *Neural Information Processing*, Springer International Publishing, Cham. pp. 194–202.
- [110] Zhou, B., Cui, Q., Wei, X.S., Chen, Z.M., 2020. BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition , 1–8.
- [111] Zhou, Y., Kang, X., Ren, F., 2018. Employing Inception-Resnet-v2 and Bi-LSTM for medical domain visual question answering, in: CLEF (Working Notes).
- [112] Zhou, Y., Kang, X., Ren, F., 2019. TUA1 at ImageCLEF 2019 vqa-med: a classification and generation model based on transfer learning, in: CLEF (Working Notes).