MMQL: Multi-Question Learning for Medical Visual Question Answering

Qishen Chen[1(✉)], Minjie Bian[2], and Huahu Xu[1]

[1] Shanghai University
{qs-chen,huahuxu}@shu.edu.cn
[2] Shanghai Data Group
bianmj@sdata.net.cn

Abstract. Medical visual question answering (Med-VQA) aims to answer medical questions with given medical images. Current methods are all designed to answer a single question with its image. Still, medical diagnoses are based on multiple factors, so questions related to the same image should be answered together. This paper proposes a novel multi-question learning method to capture the correlation among questions. Notably, for one image, all related questions are given predictions simultaneously. For those images that already have some questions answered, the answered questions can be used as prompts for better diagnosis. Further, to deal with the error prompts, an entropy-based prompt prune algorithm is designed. A shuffle-based algorithm is designed to make the model less sensitive to the sequence of input questions. In the experiment, patient-level accuracy is designed to compare the reliability of the models and reflect the effectiveness of our multi-question learning for Med-VQA. The results show our methods on top of recent state-of-the-art Med-VQA models on both VQA-RAD and SLAKE, with a 3.77% and 4.24% improvement of overall accuracy, respectively. And a 6.90% and 15.63% improvement in patient-level accuracy. The codes are available at: https://github.com/shanziSZ/MMQL.

Keywords: Multi Question Learning · Medical Visual Question Answering · Medical Image.

1 Introduction

Medical Visual Question Answering (Med-VQA) presents a unique challenge at the crossroads of visual and language modalities, focusing on the specialized domain of medical imagery. While recent research has made significant strides in Med-VQA [2,15,16], existing approaches typically assume an one-to-one relationship between images and questions, overlooking the inherent complexity of medical diagnoses where a single image may prompt multiple questions, establishing an one-to-many relationship. Moreover, real-world medical diagnoses often require considering multiple indications simultaneously, suggesting potential correlations among questions for improved predictive accuracy. For example, a combination of questions indicating the presence of bowel abnormalities

Questions	Predictions		
	MMQL	M-Mixup	MEVF-BAN-CR
Where is the diffuse pleural thickening?	Right Lung	**Right**	Right Lung
Is the image study normal?	No	No	No
Which lung is abnormal?	Right Lung	Right Lung	**Volume Loss**
Are the lungs normal?	No	No	No
Which plane is this image taken?	PA	PA	PA

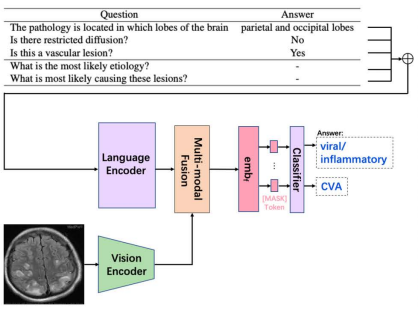

Fig. 1. The outputs of Single question methods and Our MMQL(left), wrong predictions are highlighted in red, bold and underlined. The overall architecture of our MMQL method(Right).

and high air-fluid levels might collectively suggest the diagnosis of Pneumoperitoneum. Besides, as illustrated in Fig. 1(left), traditional single question learning may yield different answers to semantically similar questions related to the same image, leading to diagnostic inconsistencies.

Multi-Question Learning (MQL), initially proposed for video question answering tasks [9], offers a means to jointly train multiple image-question pairs to generate a comprehensive visual-question representation. As far as we known, we are the first to adapt MQL mechanism in Med-VQA task, diagnose the medical image from multiple perspective. Further, previous MQL mechanisms either neglected answers in contextual [9] or enforced a specific order for question-answer(QA) pairs[22]. In contrast, our work combines with parallel answering of multiple questions and the ability to draw insights from existing QA pairs. Notably, prior work has not addressed the impact of incorrect answers in context [22].

While some visual QA methods implicitly or explicitly incorporate multiple questions into input, these methods did not formally discussing their impact. Their primary emphasis on improving inference speed [20], or establishing connections between video and subtitles [17], typically limited to answering two questions concurrently. In contrast, we contend that MQL plays a crucial role in Med-VQA tasks, conducted a detailed analysis of the MQL mechanism in Med-VQA.

In summary, the main contributions of this paper are as follows: (1) To our knowledge, we are the first to integrate the MQL approach into the realm of Med-VQA. We call it Medical Multi Question Learning(MMQL). Our innovative approach involves jointly training medical questions associated with a single image, demonstrating significant enhancements to Med-VQA models. (2) Our MQL module accommodates scenarios with no-answer and prompt-QA availability, effectively harnessing external information. A Shuffle-based augmentation algorithm is introduced to mitigate the sensitivity of question sequences. Additionally, we have introduced entropy-based prompt pruning methods to ad-

dress false QA instances in prompts. (3) We proposed a patient-level evaluation metric for better measuring the effectiveness of MQL methods. We validate the proposed methods on two public Med-VQA datasets. Experiment results indicate that our proposed methods surpass the state-of-the-art Med-VQA methods.

2 Methods

2.1 Notation

This paper treats the Med-VQA problem as a C-class classification problem. Letting $D = \{(v_{g(i)}, q_i, a_i)\}_1^n$ representing the training set, where n is the number of training samples, and $v_{g(i)}$, q_i, and a_i stands for the image, question, and answer of a sample, respectively. G is a mapping that maps the question index i to image index $g(i) \in N^*, g(i) < m$, where m is the number of images in training set.

2.2 Base Model

As shown in Fig 1(right), we build a base model with Swis-Transformer[13] for visual-encoder, BERT[6] for language-encoder. A 6 layer transformer with 8 attention heads and 768 hidden sizes is used as multi-modal fusion module.

2.3 Medical Multi Question Learning

In this study, we operationalize MQL mechanism by employing manually designed templates to amalgamate all questions into a cohesive, extended sentence. Equation 1 exemplifies this process, wherein all questions pertaining to a specific image are concatenated. Here, $\circ$ signifies the concatenation operation, and Q_k denotes the subset of all training questions, each element of which corresponds to the current input image. In scenarios where prompt-QA availability is ensured, if a question Q_i has an associated answer, we append the answer following the question.

$$\exists\ q_i \in Q_k,\ (v_k, q_i, a_i) \in D,\ T = q_1 \circ q_2 ... \circ q_i \circ a_i ... \circ q_{|Q_k|}. \tag{1}$$

2.4 Shuffle Based Augmentation

In prompt-QA-available setting, to ensure the model learning from prompt-QA pairs. We design a question-level mask strategy that uses pre-defined mask probability p to keep some answers masked with $[MASK]$ token in training samples. The $[MASK]$ token is borrowed from BERT[6], serves as a special indicator denoting a question that requires an answer. Specifically, for image v_k with n_{train} questions in the training set, we randomly select $\lfloor p \times n_{train} \rfloor$ questions and replace their answer with $[MASK]$ token, and keeping the rest questions provided with their answer in T. During testing, all n_{train} questions and their answers are

available as provided information, while n_{test} answers related to v_k are masked for evaluation.

Consistent with research by [11,14], the performance of our model may be impacted by the sequence of questions and masked questions. Employing a fixed question sequence for a specific image v_k during each training epoch may lead the model to overfit to it, resulting in relatively high test result divergence. Conversely, using the same group of questions masked during training may cause the model to underfit. To address this, we employ a shuffle-based augmentation (Shuffle-Aug) method that randomly alters the sequences of questions and re-selects masked questions, thereby ensuring that the text input T varies during each training epoch. The augmentation method is detailed in Algorithm 1.

Algorithm 1: The pseudo-code of Shuffle-Aug algorithm.

```
Data: dataset D, mask probability p, model m.
foreach epoch in max epochs do
    foreach image v_i in D do
        Shuffle the sequence of questions related to v_i. foreach q_j in all
          questions related to v_i, do
            Generate a random number r among [0, 1]. if r < p then
                The answer of q_j is masked.
            else
                The answer of q_j is not masked.
            end
        end
        Construct the input by using the templates T.
    end
    Feed all the inputs to m. Validate the model, and keep the best.
end
Result: trained model m_best
```

2.5 False-Prompt Prune

While we maintain the correctness of prompt-QA pairs during training, it is plausible that prompt-QA pairs, whether sourced from users or models, could contain inaccuracies and potentially lead to erroneous guidance in the MQL reasoning process. To mitigate this risk, we introduce a False-Prompt Prune method based on entropy. Prompt-QA pairs can be regarded as external information aimed at reducing uncertainty. The entropy of q_i given $prompt_{QA}$ can be computed using Equation 2.

$$E(Ent(q_i)) = -\frac{1}{N}\sum_{t=1}^{N}\sum_{j=1}^{C} p(a_j|q_j; prompt_{QA}) logp(a_j|q_j; prompt_{QA}) \tag{2}$$

In Equation 2, $p(a_j|q_j; prompt_{QA})$ represents the output probability of class j given the model's target question q_i and prompt $prompt_{QA}$. To mitigate the influence of sequence-induced randomness, we compute the average entropy after shuffling the question sequence N times. Since the $prompt_{QA}$ is error-free during training, any errors introduced in the inferential stage can increase the model's uncertainties, consequently raising the entropy. Therefore, we train two models: one incorporating the prompt, denoted as f_w, and the other excluding it, denoted as $f_{w/o}$. We then calculate the increase in entropy and select the answer for q_i using Equation 3.

$$\begin{aligned} \Delta E(Ent(q_i)) &= E(Ent(q_i))_{f_w} - E(Ent(q_i))_{f_{w/o}}, \\ answer &= \begin{cases} f_w(v_{g(i)}, q_i, prompt_{QA}), & if \Delta E > 0, \\ f_{w/o}(v_{g(i)}, q_i), & else. \end{cases} \end{aligned} \tag{3}$$

3 Experiment

3.1 Evaluation Metric

In this paper, we employ both vanilla accuracy Acc and our proposed patient-level accuracy $Acc_{patient}$. The vanilla accuracy is described by Equation 4.

$$Acc = \frac{S_c}{S_{all}} \times 100, \tag{4}$$

In Equation 4, Sc denotes the number of correctly answered questions, and S_{all} denotes the total number of test questions.

The concept behind $Acc_{patient}$ is to evaluate how effectively a model can harness the MQL mechanism. Intuitively, a model with strong MQL capability should exhibit a high $Acc_{patient}$. The computation of $Acc_{patient}$ is described by Equation 5.

$$Acc_{patient} = \frac{\sum_{k=1}^{m} \mathbf{1}(S_{v_k} = |Q_{k_{test}}|)}{m} \times 100 \tag{5}$$

In Equation 5, m represents the total number of images in the test set, $Q_{k_{test}}$ denotes the subset of the test set containing test questions related to v_k, and S_{v_k} indicates the number of correctly answered test questions under v_k.

3.2 Datasets and Setting

VQA-RAD [1] is a radiological dataset manually annotated by volunteers with professional medical knowledge. It comprises 315 radiological images from three organs (head, chest, and abdomen) and 3515 clinical questions, with an average of 10 questions per image. To ensure fair comparison, we adhere to the splitting used in previous studies, allocating 3064 images for training and the remaining 451 questions for testing. **SLAKE** [10] is a bilingual (Chinese and English) dataset encompassing a wider variety of organ types compared to VQA-RAD. It includes 642 images and over 14k open-ended or close-ended questions. In our

research, we utilize the English segment, which consists of 7k questions averaging 11 questions per image. Unlike VQA-RAD, the SLAKE dataset is split at the image level, with 450 images (70%) allocated for training, 96 images (15%) for validation, and 96 images (15%) for testing.

In experiments, the maximum sequence length of the text input is set to 430 for VQA-RAD and 300 for SLAKE, corresponding to the maximum of 22 and 20 questions under one image. AdamW is used to optimize the model. The learning rate is set to $5e-5$ for the language encoder and $1e-4$ for the rest of the model, with the batch size set to 16.

3.3 Comparison with existing methods

We compare our model with general VQA models and state-of-the-art Med-VQA models. MQL-AUG[9] is a multi-module video question-answering model that predicts the results at the image level. We re-implemented this model and augment it with the same backbone used in our base model, for fair comparison.

Table 1. Comparison of different Med-VQA models on the VQA-RAD and SLAKE.

Models	VQA-RAD				SLAKE			
	Open(%)	Closed(%)	Overall(%)	Patient-Level(%)	Open(%)	Closed(%)	Overall(%)	Patient-Level(%)
BAN[8]	34.64	74.63	58.76	38.92	74.57	79.09	76.34	6.25
MQL-AUG∗[9]	39.66	74.26	60.53	40.39	73.64	69.23	71.91	5.21
MEVF-BAN-CR[19]	55.87	80.88	70.95	53.21	78.76	81.97	80.02	7.29
MMBERT[7]	63.1	77.9	72.0	-	-	-	-	-
CMSA[4]	61.45	80.88	73.17	56.16	73.80	81.97	77.00	3.00
RAMM[18]	63.69	79.80	73.39	-	-	-	-	-
M-Mixup[12]	55.31	79.04	69.62	47.77	80.16	86.06	82.47	13.54
PubMedCLIP[3]	58.66	80.88	72.06	54.68	77.83	81.49	79.26	11.46
MMQL(Ours)	**64.25**(+0.56)	**85.66**(+4.78)	**77.16**(+3.77)	**61.58**(+6.90)	**84.19**(+2.9)	**90.63**(+4.57)	**86.71**(+4.24)	**29.17**(+15.63)

Note: * stands for re-implementation

The experimental results comparing our proposed MMQL method with state-of-the-art methods are presented in Table 1. Our method achieves the highest accuracy on both the VQA-RAD and SLAKE datasets, demonstrating improved performance on both open and closed questions. We do not include MedVInT-TD[21] in our comparison because this method benefits from pre-training on the PMC-VQA dataset, which contains 177k samples, while our method does not utilize such extensive pre-training. Notably, MedVInT-TD-S (MedVInT-TD without further pre-training) still underperforms compared to our method. By comparison, the MQL in video QA domain does not perform well in Med-VQA tasks, indicates the necessity of adapting the MQL mechanism to Med-VQA backbones. Additionally, through a comparison of overall accuracy and patient-level accuracy, we observe that patient-level accuracy does not necessarily exhibit a positive correlation with overall accuracy. The more detailed cased study is presented in the supplementary material.

For further analysis, we report the patient-level accuracy of our proposed model on the SLAKE dataset after each epoch during training. We assume that models in the early epochs represent inferior models, while those in the later

stages represent superior models, enabling us to examine how patient-level accuracy reflects model performance. The results are depicted in Fig. 2. Incorporating the MQL method leads to an steady increase in patient-level accuracy during training, indicating that our method successfully captures the relationships among questions associated with the same image. Conversely, the results of M-Mixup demonstrate that while overall accuracy improvement is observed, there is no guarantee that patient-level accuracy will also increase if the model processes image-question pairs independently.

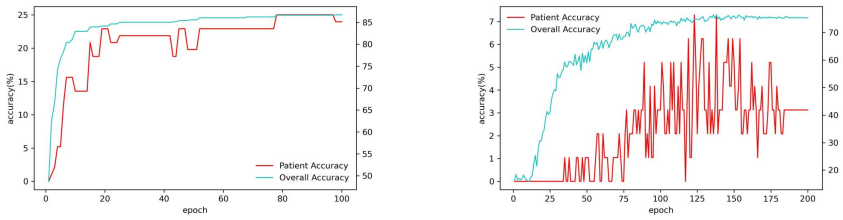

Fig. 2. The accurate curve of our MMQL(left) and M-Mixup(right) on SLAKE dataset.

3.4 Ablation Study

We study the effectiveness of our MMQL from three aspects. First, we train on single question. Then, MQL is added. Next, we introduced Shuffle-Aug. The result is shown in Table 2, the base model does not perform well under single question setting. Introducing the MQL mechanism improved the model's performance, which had a 3% increase. Then, Shuffle-Aug algorithm can improve another 7.26% performance. We also report the standard deviation of test results due to the different sequences of questions, which shows that Shuffle-Aug can decrease the negative effect of different permutations of questions.

Table 2. Ablation study on Slake.

Base Model	MQL	Shuffle-Aug	Overall Acc(%)
BERT+Swin-T +CR			76.91±(0.00)
	✓		79.45±(0.65)
	✓	✓	86.71±(0.31)

Table 3. Effectiveness of Error Prompt Prune strategy on SLAKE*.

Error Rate	With Error	Prune	Ensemble
0%	85.07±(0.34)	85.07	85.07
20%	83.97±(0.42)	85.29	83.75
40%	83.85±(0.40)	85.43	83.75
60%	83.77±(0.47)	85.16	83.75
80%	83.67±(0.48)	85.56	83.75
100%	83.49±(0.53)	85.43	83.62
Real	83.78±(0.38)	85.16	83.35
No Prompt-QA	82.86±(0.36)	-	-
w/o MQL	76.91±(0.00)	-	-

Furthermore, we analyze the effectiveness of the False-Prompt Prune algorithm. We randomly select 30% of the test data in SLAKE as prompt-QA. We gradually increase the error rate in prompt-QA with a step size of 20%. As for Real, it uses pseudo-labels outputs by the base model. We denote this modified version of SLAKE as SLAKE*. According to the findings in Table 3, error prompt-QAs can negatively impact model predictions. However, employing the Error Prompt Prune method has proven effective in reducing this impact. In fact, it outperforms the Soft Ensemble method in both simulated and real scenarios.

3.5 Calibration Error

The alignment between a model's confidence and its accuracy holds paramount importance in the medical domain; overly confident results can potentially mislead medical professionals. To assess the confidence calibration of various methods, we employ Expected Calibration Error (ECE) and Maximum Calibration Error (MCE) metrics [5]. The results are illustrated in Fig. 3. Our model ranks second-best, suggesting that it exhibits better calibration compared to typical Med-VQA models, including the general MQL model. However, our model falls short of outperforming the M-Mixup, indicating that it may exhibit overconfidence in rare samples, whereas Mixup can augment such samples.

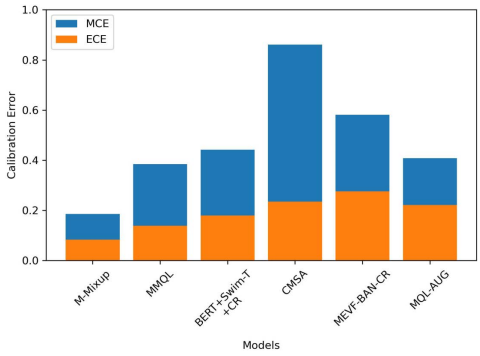

Fig. 3. The calibration error of different models.

4 Conclusion

In our paper, multi-question learning is integrated in Med-VQA for the first time. This innovative approach involves consolidating multiple medical inquiries pertaining to the same image into a singular, comprehensive sentence. This not only uses the inherent correlation among questions but also leverages past QA pairs to enhance predictive accuracy, a feature notably absent in prior MQL

methodologies. To gauge performance, we introduce patient-level accuracy, which provide a nuanced assessment of Med-VQA models and the efficacy of MQL. Experimental results demonstrate that our proposed method surpasses existing state-of-the-art Med-VQA models. Moreover, ablation study confirms the efficacy of our proposed Shuffle-Aug algorithm and False-Prompt Prune algorithm. These techniques not only make the model less susceptible to the sequence of input questions but also bolster its resilience against error prompt-QA pairs, ultimately enhancing its robustness.

Acknowledgments. This work is Supported by Shanghai Technical Service Center of Science and Engineering Computing, Shanghai University.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ben Abacha, A., Hasan, S.A., Datla, V.V., Demner-Fushman, D., Müller, H.: Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In: Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes. 9-12 September 2019 (2019)
2. Cong, F., Xu, S., Guo, L., Tian, Y.: Caption-aware medical vqa via semantic focusing and progressive cross-modality comprehension. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 3569–3577 (2022), doi:10.1145/3503161.3548122
3. Eslami, S., Meinel, C., De Melo, G.: Pubmedclip: How much does clip benefit visual question answering in the medical domain? In: Findings of the Association for Computational Linguistics: EACL 2023. pp. 1151–1163 (2023)
4. Gong, H., Chen, G., Liu, S., Yu, Y., Li, G.: Cross-modal self-attention with multi-task pre-training for medical visual question answering. In: Proceedings of the 2021 international conference on multimedia retrieval. pp. 456–460 (2021), doi:10.1145/3460426.3463584
5. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International conference on machine learning. pp. 1321–1330. PMLR (2017)
6. Kenton, J.D.M.W.C., Toutanova, L.K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT. pp. 4171–4186 (2019), doi:10.18653/v1/n19-1423
7. Khare, Y., Bagal, V., Mathew, M., Devi, A., Priyakumar, U.D., Jawahar, C.: Mmbert: Multimodal bert pretraining for improved medical vqa. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). pp. 1033–1036. IEEE (2021), doi:10.1109/ISBI48211.2021.9434063
8. Kim, J.H., Jun, J., Zhang, B.T.: Bilinear attention networks. Advances in neural information processing systems **31** (2018)
9. Lei, C., Wu, L., Liu, D., Li, Z., Wang, G., Tang, H., Li, H.: Multi-question learning for visual question answering. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 11328–11335 (2020), doi:10.1609/aaai.v34i07.6794

10. Liu, B., Zhan, L.M., Xu, L., Ma, L., Yang, Y., Wu, X.M.: Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). pp. 1650–1654. IEEE (2021), doi:10.1109/ISBI48211.2021.9434010
11. Liu, J., Shen, D., Zhang, Y., Dolan, B., Carin, L., Chen, W.: What makes good in-context examples for gpt-3? DeeLIO 2022 p. 100 (2022), doi:10.18653/v1/2022.deelio-1.10
12. Liu, L., Su, X.: How well apply multimodal mixup and simple mlps backbone to medical visual question answering? In: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 2648–2655. IEEE (2022), doi:10.1109/BIBM55620.2022.9995347
13. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021), doi:10.1109/ICCV48922.2021.00986
14. Lu, Y., Bartolo, M., Moore, A., Riedel, S., Stenetorp, P.: Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 8086–8098 (2022), doi:2022.acl-long.556
15. Nguyen, B.D., Do, T.T., Nguyen, B.X., Do, T., Tjiputra, E., Tran, Q.D.: Overcoming data limitation in medical visual question answering. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22. pp. 522–530. Springer (2019), doi:10.1007/978-3-030-32251-9$_5$7
16. Pan, H., He, S., Zhang, K., Qu, B., Chen, C., Shi, K.: Amam: An attention-based multimodal alignment model for medical visual question answering. Knowledge-Based Systems **255**, 109763 (2022), doi:10.1016/j.knosys.2022.109763
17. Yang, Z., Garcia, N., Chu, C., Otani, M., Nakashima, Y., Takemura, H.: Bert representations for video question answering. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1556–1565 (2020)
18. Yuan, Z., Jin, Q., Tan, C., Zhao, Z., Yuan, H., Huang, F., Huang, S.: Ramm: Retrieval-augmented biomedical visual question answering with multi-modal pre-training. arXiv e-prints pp. arXiv–2303 (2023), doi:10.48550/arXiv.2303.00534
19. Zhan, L.M., Liu, B., Fan, L., Chen, J., Wu, X.M.: Medical visual question answering via conditional reasoning. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2345–2354 (2020), doi:10.1145/3394171.3413761
20. Zhang, C., Gupta, A., Zisserman, A.: Temporal query networks for fine-grained video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4486–4496 (2021)
21. Zhang, X., Wu, C., Zhao, Z., Lin, W., Zhang, Y., Wang, Y., Xie, W.: Pmc-vqa: Visual instruction tuning for medical visual question answering. arXiv preprint arXiv:2305.10415 (2023)
22. Zhao, Z., Jiang, X., Cai, D., Xiao, J., He, X., Pu, S.: Multi-turn video question answering via multi-stream hierarchical attention context network. In: IJCAI. vol. 2018, p. 27th (2018)