The Impact of Multimodal Fusion methods and Debiasing Techniques for Medical VQA

1st Aya Nuseir
Institute of Mathematics and Computer Science
University of Greifswald
Greifswald, Germany
s-a ynuse@uni-greifswald.de

2nd Marc Ebner
Institute of Mathematics and Computer Science
University of Greifswald
Greifswald, Germany
marc.ebner@uni-greifswald.de

***Abstract*—Medical Visual Question Answering (Med-VQA) combines computer vision (CV) and natural language processing (NLP) to help make clinical decisions and medical education by answering questions about medical images. However, developing powerful Med-VQA models faces several issues, including the complex nature of medical images, the availability of labeled data, and dataset biases. This research examines two critical points of Med-VQA: (1) multimodal fusion techniques for combining visual and textual information, and (2) the effectiveness of unbiased learning approaches using data augmentation and class weighting methods. The paper evaluates three fusion approaches (concatenation-based, attention-based, and bilinear attention networks) on two benchmark datasets: SLAKE and PathVQA. Our experiments show that attention-based approaches perform well, with an accuracy of 58.47% on PathVQA. The study of unbiased learning techniques shows that the adopted class weighting method does not improve Med-VQA performance as expected, while using the image augmentation technique alone outperforms combinations with the class weighting technique.**

***Index Terms*—Medical Visual Question Answering, Multimodal Fusion, Unbiased Learning, Data Augmentation, Class Weighting Methods**

I. Introduction

Visual Question Answering (VQA) is a rapidly growing field within artificial intelligence (AI) that integrates computer vision (CV) and natural language processing (NLP) fields to answer questions about images. A VQA model usually consists of four processes, namely (i) visual feature extraction, where the related features are extracted from the input image, (ii) textual feature extraction from the question, (iii) Multimodal fusion, where the extracted visual and textual representations are integrated. (iv) Answer prediction, generating the final output according to the integrated features. With the progress of VQA research in the general domain, there has been an increasing interest in adapting these techniques to the medical domain, medical visual question answering (Med-VQA) aims to aid in clinical decision-making, improve patient engagement, and can be used as a medical education tool to assist medical students in their studying [1]–[4]. However, similar to the general VQA domain, developing a robust Med-VQA model is challenging, which returns to the complex nature of medical images, limited labeled data, and inherent biases in datasets [5]-[8]. To address these challenges, this research focuses on two key research areas and explores their impacts on the Med-VQA model: (i) fusion techniques for multimodal integration and (ii) unbiased learning techniques such as data augmentation and class weighting. Fusion methods play a crucial role in Med-VQA by combining visual and textual modalities, impacting the model's ability to extract and align relevant information from both modalities. Previous research has examined various fusion techniques, but there is still no agreement on the optimal approach for Med-VQA. In this work, we investigate multiple fusion techniques, comparing their performance on two benchmark datasets: SLAKE [9], and PathVQA [10]. Medical datasets often show biases, where certain question types, medical conditions, and imaging methods are represented excessively. In addition to predicting the answers, the process is often based on shallow correlations between the questions and answers rather than including the relationship between the images and questions [11], [12]. To address this, we explored data augmentation and class weighting techniques that have shown an impressive impact in reducing bias across other deep-learning applications [13], [14]. However, our experiments show that these techniques do not improve the Med-VQA performance across all tested datasets as we expected. This suggests that these adopted unbiased methods may not fit our Medical VQA model. Our work makes several key contributions to the field of Medical Visual Question Answering:

- We adopting an attention-based fusion approach that bridges the semantic gap between medical images and clinical questions by utilizing specialized visual and textual feature encoders (DenseNet121 and BiomedVLP-CXR-BERT).
- We perform a comprehensive comparative analysis of three different fusion methods: concatenation-based, attention-based, and bilinear attention networks—across multiple Med-VQA datasets (SLAKE and PathVQA), offering a deep overview of which approaches are most

efficient for different medical imaging modalities and question types.

- We comprehensively evaluate the impact of unbiased learning techniques on Med-VQA performance, revealing the unexpected result that class weighting methods do not improve results when applied to medical datasets. At the same time, image augmentation alone shows better performance.

This paper is organized as follows. Section III provides an overview of the proposed Med-VQA framework. Section IV explains the experiments and displays the evaluation of the experiments, and finally, the paper is concluded in Section VI.

II. Related Work

Recent research in Med-VQA has examined different approaches to integrating visual and textual representations and developed techniques to handle bias challenges. To overcome the data limitation of medical VQA, Nguyen et al. [5] proposed a framework that explores the use of the unsupervised Denoising Auto-Encoder (DAE) and the supervised Meta-Learning (MAML). However, their focus was only on the data limitation in the aspect of medical images while ignoring the impact of textual representation. MedFuseNet [15] is an attention-based multimodal model that aims to learn representations by an optimal fusion of the multimodal inputs using the attention mechanism by focusing on the most related part of the medical images and questions. The multi-modal representation of image and question are passed through an LSTM decoder. Van Sonsbeek et al. [16] shift away from classification-based methods to open-ended generative answers. They map the extracted visual features to a set of learnable tokens serving as a visual prefix for the language model. Then, along with the question, these visual prefixes are passed directly to the language model to generate the answer token by token. Chen et al. [8] focused on critical information interaction within each modality and relevant information interaction between modalities. They proposed a Symmetric Interaction Attention Module (SInAM) to construct dense and deep intra- and inter-modal information interaction in medical images and clinical questions. SInAM consists of multiple symmetric interaction attention blocks that contain two basic units: self-attention and interaction attention, where self-attention is utilized in the intra-modal information interaction, and interaction attention for inter-modal information interaction. Our research is inspired by these studies. While previous works have studied different fusion methods and debiasing techniques separately, we evaluate both parts across two benchmark datasets (SLAKE and PathVQA).

III. Methodology

Given an input image ***I*** and a natural language question ***Q***, the medical VQA task aims to predict an accurate answer ***a***. Formally, we define this as:

$$a = argmax P(a|I, Q) \tag{1}$$

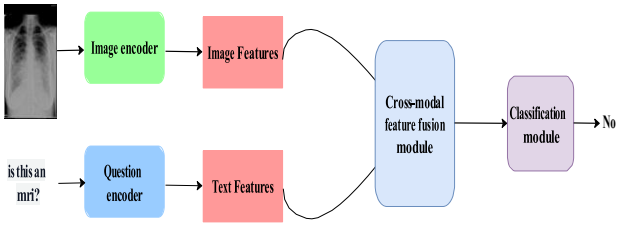

Fig. 1. Model structure.

where P(a| I, Q) represents the probability that answer a is correct given I and Q as input. The overall architecture of our baseline model is shown in Figure 1, which consists of four modules: (i) the image encoder that encodes medical images to obtain visual features, in our work, we employ the Densenet121 [17] architecture pre-trained on ImageNet as the backbone for the image encoder; the DenseNet parameters are frozen during training to prevent overfitting. (ii) Question encoder, responsible for producing textual embeddings; here, we employ BiomedVLP-CXR-BERT [18]. The process begins with tokenization, where the input question text is tokenized and padded to regulate input lengths. Then the BERT model processes the tokenized questions to generate textual representations. Textual representations are used twice in our model. They are first used in the attention mechanism in guiding the model in focusing on the relevant image regions, and they are directly concatenated with the attended image features to create the multimodal representation, as seen in Figure 2.
(iii) The cross-modal feature fusion module that aligns and combines visual and textual representations. For this module, we adopted the attention mechanism (a question-guided attention mechanism) that creates the interactions between the image and text features. This attention module calculates attention weights to focus on related image features according to the question content. The attended image features are then concatenated with text features and passed through a sequence of fully connected layers with normalization, ReLUs, and dropout for regularization. In more detail, the process begins with projecting the textual and visual features into the shared dimension:

$$V' = W_v V + b_v, \quad T' = W_t T + b_t \tag{2}$$

where: V is the visual features extracted from the Densnet121 W_v and W_t are learnable weight matrices that project the visual and textual features into the shared embedding space of dimension d. b_v and b_t represent the bias. T is the textual features extracted from Bert.

Then the attention weights α are computed according to the following formula:

$$A = W_{att}\, tanh(T' + V') \tag{3}$$

$$\alpha = softmax(A) \tag{4}$$

where W_{att} is a learnable attention matrix

The attended visual features V_{att} are computed as:

$$V_{att} = \sum(\alpha_i V_i) \quad (5)$$

Finally, the attended visual features are combined with the textual embeddings to form a joint representation. Figure 2 illustrates the architecture of the model developed using this method in fusing the extracted visual and textual features.

(iv)The classification module receives the fused multimodal features, where this concatenated representation goes through several transformations to predict the final answer. First, the fused features are projected through a fully connected layer to a lower-dimensional space(N):

$$fc = Wf + b \quad (6)$$

where $W \in \mathrm{R}^{N\times(d_v+d_t)}$ is the weight matrix and f is the fused features. $b \in \mathrm{R}^N$ represents the bias. Then the projected features are passed through a ReLU activation function to introduce non-linearity:

$$u = ReLU(fc) = max(0, fc) \quad (7)$$

A dropout layer with probability p=0.5 is applied during training to prevent overfitting:

$$z = Dropout(u, p = 0.5) \quad (8)$$

Finally, the model makes use of a classification layer to map the features to answer probabilities:

$$y' = softmax(W_c z + b_c) \quad (9)$$

Where $W_c \in \mathrm{R}^c \times N$is the classification weight matrix, $b_c \in \mathrm{R}^c$ is the bias term, and c is the number of answer classes. During training, the model minimizes the cross-entropy loss between predicted probabilities y' and ground truth answers y:

$$L = -\sum_i y_i \log(y_i') \quad (10)$$

IV. Experiments and Results

We use SLAKE [9], and PathVQA [10] in our experiments. SLAKE is an English-Chinese bilingual dataset containing 642 images and 14,028 question-answer pairs. This dataset includes 12 diseases and 39 organs of the whole body. In our work, we use the English subset of the SLAKE dataset.

The PathVQA dataset consists of 32,799 question-answer pairs generated from 4,998 pathology images collected from two pathology textbooks and the PEIR digital library. The questions in the dataset are divided into open-ended and closed-ended (yes/no) questions. Each one of these two datasets shows a different kind of challenge, represented by imbalanced question types and limited training samples, making them ideal for testing the fusion strategies and unbiased learning techniques. These two datasets provide a comprehensive view within their domains but represent a subset of medical imaging modalities and clinical techniques used in practice. In this work, we define the Med-VQA task as a multiclass classification problem, and to evaluate the models, we adopt accuracy to analyze the exact matches of predictions.

For a more comprehensive understanding of the efficiency in combining visual and textual modalities, we analyze three fusion techniques: Concatenation-based, Attention-based, and Bi-linear attention network. The first method, the Concatenation-based method, simply concatenates the feature vectors of text and image along the feature dimension. Attention-based fusion employs an additive attention mechanism to allow our model to focus on relevant image regions based on the question content. It computes attention weights that emphasize important visual features according to the textual query. The attended visual features are then concatenated with textual features to form a joint representation, as described in detail in Section 3. Bilinear Attention Networks (BAN) [19] extend traditional co-attention mechanisms to bilinear attention. Rather than relying on finding separate attention distributions for each modality, BANs examine every possible pair of interactions between visual and textual features through bilinear attention maps. The technique employs low-rank bilinear pooling to compute these interactions. The second key point is unbiased learning techniques: we investigate whether the adopted unbiased strategies (image augmentation and class weighting) improve Med-VQA performance. The techniques adopted are described below.

The first technique is the class weight; in this research, we adopt the Fernando and Tsokos [14] method to find the class weight and study its impact on the performance of the model. The weighting follows a logarithmic formula based on class frequencies, where classes with fewer instances receive higher weights.

$$w_i = \log\left(\frac{\max(n_i | i \in c)}{n_i}\right) + 1 \quad (11)$$

where: w_i is the weight for class i, n_i is the frequency of class I, and $\max(n_i | i \in c)$ represents the class frequency of the majority class, and c is the set of classes.

Data Augmentation is adopted as the second technique; the augmentation pipeline includes resizing, random rotation, horizontal flipping, and colour jittering. These augmentations are applied only during training. All models are implemented in PyTorch and trained using the Adam optimizer with a learning rate of 1e-4 and a batch size of 16. We fine-tune each model for 25 epochs, and all computations are performed on an NVIDIA GeForce RTX 4090 GPU.

V. Results and Discussion

From Table I and Figure 3, our model (which uses the attention-based fusion method illustrated in Figure 2) achieves an overall accuracy of 78.98% on the SLAKE dataset, which came in second place compared with the state-of-the-art models, and 58.47% on the PathVQA dataset, which out-performs the other models, particularly the performance on close-ended questions reaching 88.28% accuracy. Our model shows an impressive performance in closed-ended questions for the PathVQA dataset, but shows a performance gap within open-ended questions, where the accuracy is 28.62%. This difference shows a challenge that faces the Med-VQA models in general, which is the difficulty in generating the exact match

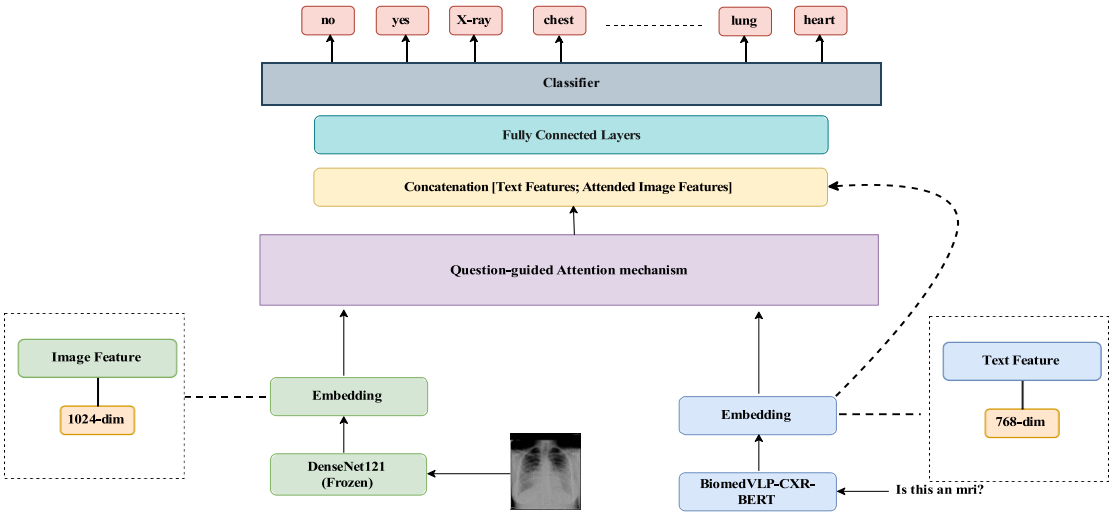

Fig. 2. Developed model architecture utilizing the question-guided attention mechanism.

Of free-form answers. Additionally, closed-ended questions require binary classification (yes/no), which can be a simpler task compared to open-ended questions, which require the precise generation of exact medical terms, anatomical structures, or pathological findings from a large vocabulary space. Further, pathology images contain complex visual features requiring specialized domain knowledge to analyze correctly [20]. When compared to state-of-the-art models like BiomedGPT-M (which achieves 78.3% on open-ended SLAKE questions) and Van Sonsbeek et al.'s approach (with 87.00% on closed-ended PathVQA questions), our developed model shows competitive performance.

Table II illustrates the impact of unbiased learning techniques on the three fusion methods for Med-VQA: Starting with the effects of not applying any of the adopted unbiased learning techniques, for the Slake dataset, the Attention-Based Fusion method performs best for open-ended questions and closed-ended ones at 77.19% and 82.53% respectively. The concatenation-based method performs better on closed-ended questions than open-ended questions. At the same time, BAN notably performs worse, especially on open-ended questions, with an accuracy of 49.00%. For the PathVQA Dataset, Attention-Based Fusion performs best with an overall accuracy equal to 58.47%, 28.62%, and 88.28% for open-ended questions and closed-ended questions, respectively. BAN performs poorly, especially on open-ended questions (9.94%). The results show that the first two fusion methods outperform the BAN method. This may be related to many factors, such as the limited size of the two datasets and the fact that the BAN method requires more data to learn cross-modal interactions effectively. Also, the other fusion methods can be considered less complex than BAN, which means they have fewer parameters and thus show better performance with limited medical datasets. Using image augmentation without a class weighting technique, the overall accuracy for both concatenation-based and Attention-Based Fusion methods is quite similar on the Slake dataset: 77.37%. The attention-based fusion method shows a slight improvement in the open-ended questions type compared with the concatenation-based method. The best accuracy for the PathVQA dataset is 58.32% using the Attention-Based Fusion method. The BAN continues to underperform, with 56.92% accuracy on SLAKE and 48.47% on the PathVQA dataset. The class weighting method (logarithmic) with image augmentation shows a drop in performance compared to adopting only the image augmentation technique. For the Slake dataset, the overall accuracy for both concatenation-based and Attention-Based Fusion methods are the same and equal to 75.68%, but this time the concatenation-based method's performance within the open-ended questions is slightly better than the performance of the Attention-Based Fusion method. For the PathVQA dataset, the results show that using the Attention-Based Fusion method outperforms the other two methods with an accuracy of 57.82%. Removing image augmentation while keeping the class weighting method shows a few different results, but continues the overall negative impact on performance.

The most remarkable finding is the symmetric pattern across both datasets, where adopting the image augmentation technique alone outperforms combinations with a class weighting method. While the research study addresses class imbalance through weighting techniques, we acknowledge that medical VQA datasets contain other important bias types beyond our current coverage. Different kinds of bias exist in the medical field, which could all impact performance [21]–[23], such as depending on the specific contexts of the dataset rather than the image-question correlations. Also, some models show bias toward certain imaging types. Our finding that class weighting did not enhance performance indicates that other bias types might have greater impacts in medical VQA. Future work should explore techniques specifically designed to address

TABLE I. Results on SLAKE and PathVQA datasets. Our model outperforms all previous models for the PathVQA dataset and came in second place for the SLAKE dataset (represented in red). While blue represents the best-performing medical model in each dataset. The symbol (-) indicates that the results were not reported in the original papers.

Model	SLAKE Dataset			PathVQA Dataset		
	Open-End	Close-End	Acc	Open-End	Close-End	Acc
MedFuseNet [15]	-	-	-	-	63.6%	
Van Sonsbeek et al. [16]	-	82.01%	-	-	87.00%	-
BiomedGPT-M [28]	78.3%	86.80%	-	12.5%	85.7%	-
Chen et al. [8]	-	-	-	13.9%	83.4%	-
Our Model	77.19%	82.53%	78.98%	28.62%	88.28%	58.47%

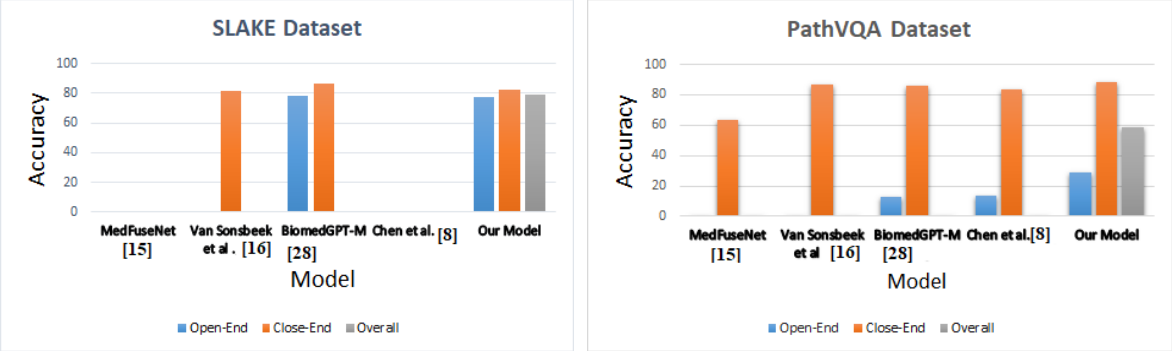

Fig. 3. Results on SLAKE and PathVQA datasets.

these additional biases. Additionally, the results show the impact of adopting the Attention-Based fusion method.

A. Ablation study

We conduct an ablation study to analyze the impact of different text encoders on Med-VQA's performance. The experiments are performed on the VQA-SLAKE dataset, and the results are reported in Table III. Specifically, we replaced Biomedvlp-CXR-BERT with two different textual encoders, namely T5 [24] and BioT5 [25], while maintaining our baseline model architecture. Table III presents the results, highlighting how different pre-trained language models affect model size and accuracy across the SLAKE dataset. Our experiments reveal that BioT5 performs better than T5, suggesting that its pre-training on textual data from the biological domain and the BioT5 size provides helpful domain knowledge for Med-VQA tasks.

VI. Conclusion and Future Work

We have evaluated fusion methods and unbiased learning techniques in the context of Medical Visual Question Answering. The study revealed several key findings: First, simpler fusion methods (concatenation-based and attention-based fusion) outperform the more complex BAN across both SLAKE and PathVQA datasets. The attention-based fusion method achieved the highest overall accuracy of 58.47% on PathVQA and 78.98% on SLAKE. Through our analysis, we find that more complex fusion mechanisms, such as BAN, may not be necessary or effective for Med-VQA tasks with limited training data sizes. Second, the study of using unbiased learning techniques reveals that the adopted class weighting methods do not enhance Med-VQA performance as expected. In particular, image augmentation alone outperforms combinations with the class-weighting method. Although SLAKE and PathVQA provide practical frameworks for our experiments, we acknowledge their limitations in representing the full coverage of medical imaging modalities (e.g., ultrasound, PET scans). This may impact the generalizability of our findings across the wider medical domain. While the study offers valuable insights into fusion methods and unbiased learning for Med-VQA, there are still several valuable directions to be explored in future research: (1) developing specialized debiasing methods specifically designed for medical datasets rather than adapting general-purpose techniques and (2) investigating the integration of large language models (LLMs) pre-trained on medical datasets to better understand medical textual data, particularly for improving performance on open- ended questions. (3) evaluating model performance across a wider range of imaging modalities and anatomical regions by including further datasets such as P-VQA [26], OVQA [27].

References

[1] S. A. Hasan, Y. Ling, O. Farri, J. Liu, H. Müller, and M. Lungren, "Overview of ImageCLEF 2018 medical domain visual question answering task," in Proceedings of CLEF 2018 Working Notes, 2018.

[2] F. Ren and Y. Zhou, "CGMVQA: A new classification and generative model for medical visual question answering," IEEE Access, vol. 8, pp. 50626-50636, 2020.

[3] S. Al-Hadhrami, MEB. Menai, S. Al-Ahmadi, A. Alnafessah, "An effective Med-VQA method using a transformer with weights fusion of multiple fine-tuned models," Applied Sciences, pp. 9735, 2023.

[4] Y. Liu, Z. Wang, D. Xu, and L. Zhou, "Q2ATransformer: Improving medical VQA via an answer querying decoder," in International Conference on Information Processing in Medical Imaging, 2023, pp. 445-456.

[5] B. D. Nguyen, TT. Do, BX. Nguyen, T. Do, E. Tjiputra, QD. Tran, "Overcoming data limitation in medical visual question answering,"in MICCAI2019,pp.522-530,2019.

TABLE II. Impact of Unbiased Learning Techniques on the Three Fusion Methods for MedVQA. The best-performing results are represented in blue, while the second-performing results are represented in red.

Fusion Method	SLAKE Dataset			PathVQA Dataset		
	Open-End	Close-End	Acc	Open-End	Close-End	Acc
No Image Augmentation, No Class Weighting						
Concatenation-based method	76.06%	79.15%	77.09%	27.73%	87.50%	57.64%
Attention-Based Fusion	**77.19**%	**82.53**%	**78.98**%	**28.62**%	**88.28**%	**58.47**%
Bilinear Attention Network	49.00%	65.35%	54.47%	09.94%	87.15%	48.57%
Image Augmentation, No Class Weighting						
Concatenation-based method	**75.49**%	**81.12**%	**77.37**%	27.16%	88.69%	57.95%
Attention-Based Fusion	**75.92**%	**80.28**%	**77.37**%	**27.91**%	**88.69**%	**58.32**%
Bilinear Attention Network	50.00%	70.70%	56.92%	09.68%	87.21%	48.47%
Image Augmentation, Class Weighting Method						
Concatenation-based method	76.62%	73.80%	75.68%	26.36%	86.97%	56.69%
Attention-Based Fusion	75.63%	75.77%	75.68%	27.97%	87.62%	57.82%
Bilinear Attention Network	47.45%	71.26%	55.41%	11.31%	87.15%	49.26%
No Image Augmentation, Class Weighting Method						
Concatenation-based method	75.35%	69.01%	73.23%	28.12%	87.98%	58.07%
Attention-Based Fusion	76.06%	74.64%	75.58%	27.37%	85.12%	56.27%
Bilinear Attention Network	46.60%	70.42%	54.57%	09.88%	86.07%	48.01%

TABLE III. Ablation study on the SLAKE dataset with different text encoders.

Text Encoder	Open-End	Close-End	Acc	Model Size(MB)
T5	70.01%	66.19%	68.73%	167
BioT5	71.85%	69.01 %	70.97%	460
Biomedvlp-CXR-BERT	77.19 %	82.53 %	78.98 %	455

[6] B. Liu, L.-M. Zhan, and X.-M. Wu, "Contrastive pre-training and representation distillation for medical visual question answering based on radiology images," in MICCAI 2021, 2021, pp. 210-220.

[7] B. Koçak, A. Ponsiglione, A. Stanzione, C. Bluethgen, J. Santinha, L. Ugga, M. Huisman,…, and R.Cuocolo,"Bias in artificial intelligence for medical imaging: Fundamentals, detection, avoidance, mitigation, challenges, ethics, and prospects," Diagnostic and Interventional Radiology, pp. 75, 2025.

[8] Z. Chen, , B. Zou, Y. Dai, C. Zhu, G. Kong, W. Zhang, "Medical visual question answering with symmetric interaction attention and cross-modal gating," Biomedical Signal Processing and Control, pp. 105049, 2023.

[9] B. Liu, L.-M. Zhan, L. Xu, L. Ma, Y. Yang, and X.-M. Wu, "SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering," in *2021 IEEE 18th ISBI*, pp. 1650-1654, 2021.

[10] X. He, Y. Zhang, L. Mou, E. Xing, and P. Xie, "PathVQA: 30,000+ questions for medical visual question answering," *arXiv preprint arXiv:2003.10286*, 2020.

[11] Z. Sun, A. Harit, AI. Cristea, J. Yu, N. Al. Moubayed, L. Shi,"Is unimodal bias always bad for visual question answering? A medical domain study with dynamic attention," in *2022 IEEE International Conference on Big Data (Big Data)*, pp. 5352-5360, 2022.

[12] S. Ye, U. Naseem, M. Meng, D. Feng, J. Kim, "A causal approach to mitigate modality preference bias in medical visual question answering," in *Proceedings of the First International Workshop on Vision-Language Models for Biomedical Applications*, pp. 13-17, 2024.

[13] H. Gong, , G. Chen, M. Mao, Z. Li, G. Li ,"VQAMix: Conditional triplet mixup for medical visual question answering," *IEEE Transactions on Medical Imaging*, pp. 3332-3343, 2022.

[14] K. R. M. Fernando and C. P. Tsokos, "Dynamically weighted balanced loss: Class imbalanced learning and confidence calibration of deep neural networks," IEEE Transactions on Neural Networks and Learning Systems, pp. 2940-2951, 2021.

[15] D. Sharma, S. Purushotham, and C. K. Reddy, "MedFuseNet: An attention-based multimodal deep learning model for visual question answering in the medical domain," *Scientific Reports*, pp. 19826, 2021.

[16] T. Van Sonsbeek, MM. Derakhshani, I. Najdenkoska, CGM. Snoek, M. Worring, "Open-ended medical visual question answering through prefix tuning of language models," in *MICCAI*, pp. 726-736, 2023.

[17] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *CVPR*, 2017.

[18] B. Boecking, N. Usuyama, S. Bannur, DC. Castro, A. Schwaighofer, S. Hyland, M. Wetscherek,…and O.Oktay, "Making the most of text semantics to improve biomedical vision-language processing," *European conference on computer vision. Cham: Springer Nature Switzerland, 2022*.

[19] J.-H. Kim, J. Jun, and B.-T. Zhang, "Bilinear attention networks," *Advances in Neural Information Processing Systems*, 2018.

[20] T. Do, BX. Nguyen, E. Tjiputra, M. Tran, QD. Tran, A. Nguyen,"Multiple meta-model quantifying for medical visual question answering," in *Medical Image Computing and Computer Assisted Intervention -- MICCAI 2021: 24th International Conference*,pp. 64-74,2021.

[21] X. Hu, L. Gu, Q. An, M. Zhang, L. Liu, K. Kobayashi, T. Harada, RM. Summers, Y. Zhu, "Expert knowledge-aware image difference graph representation learning for difference-aware medical visual question answering," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023, pp. 4156-4165.

[22] X. Hu, L. Gu, K. Kobayashi, L. Liu, M. Zhang, T. Harada, RM. Summers, Y. Zhu,"Interpretable medical image visual question answering via multi-modal relationship graph learning," *Medical Image Analysis*, pp. 103279, 2024.

[23] Y. Bi, X. Wang, Q. Wang, and J. Yang, "Overcoming data limitations and cross-modal interaction challenges in medical visual question answering," in *IJCNN*, pp. 1-8, 2024.

[24] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S.Narang, M. Matena,Y. Zhou, W. Li, P. Liu "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, pp. 1-67, 2020.

[25] Q. Pei, W. Zhang, J. Zhu, K. Wu, K. Gao, L. Wu, Y. Xia, R. Yan,"BioT5: Enriching cross-modal integration in biology with chemical knowledge and natural language associations," *arXiv preprint*, 2023.

[26] J. Huang, Y. Chen, Y. Li, Z. Yang, X. Gong, FL. Wang, X. Xu, W. Liu ,"Medical knowledge-based network for patient-oriented visual question answering," *Information Processing & Management*, pp. 103241, 2023.

[27] Y. Huang, X. Wang, F. Liu, and G. Huang, "OVQA: A clinically generated visual question answering dataset," in Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information *Retrieval*, pp. 2924-2938, 2022.

[28] K. Zhang, J.Yu , E. Adhikarla, R. Zhou , Z.Yan, Y.Liu , Z. Liu , L. He, B. Davison, X. Li, H. Ren. "BiomedGPT: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks." arXiv e-prints, 2023