Efficient Bilinear Attention-based Fusion for Medical Visual Question Answering

Zhilin Zhang*, Jie Wang†, Ruiyang Qin†, Ruiqi Zhu† and Xiaoliang Gong†‡
*Tandon School of Engineering, New York University, New York, USA
†College of Electronic and Information Engineering, Tongji University, Shanghai, China
Email: zz10068@nyu.edu, 2054310@tongj.edu.cn, 2432008@tongji.edu.cn, zhurq@tongji.edu.cn
Corresponding Author: ‡gxllshsh@tongji.edu.cn

***Abstract*—Medical Visual Question Answering (MedVQA) is a crucial task that combines computer vision and natural language processing to assist in clinical decision-making by answering questions based on medical images. In this paper, we propose the Intra-modal and Cross-modal Bilinear Attention Network (IC-BAN), a novel model that integrates multi-scale feature extraction, self-attention mechanisms, and bilinear attention networks to effectively fuse visual and textual features. Our approach addresses the challenges of MedVQA by balancing accuracy and computational efficiency. Experimental results demonstrate that IC-BAN can achieve superior performance compared to both traditional BAN and Transformer-based fusion methods. This work highlights the potential of advanced bilinear attention mechanisms in MedVQA.**

***Index Terms*—Medical Visual Question Answering, Multi-modal Fusion, Bilinear Attention, Multi-scale Feature Extraction**

I. Introduction

Medical Visual Question Answering (MedVQA) is an emerging research area within multimodal artificial intelligence that applies the general principles of Visual Question Answering (VQA) to the medical domain. The primary objective of MedVQA is to assist healthcare professionals by automatically answering questions related to medical images, thereby supporting clinical decision-making processes. This task involves the integration of computer vision and natural language processing techniques to analyze medical images in conjunction with natural language questions, generating accurate and contextually relevant answers.

The complexity of MedVQA lies in the specialized nature of medical images and the terminology used in questions, which requires models that can capture intricate details and relationships between visual and textual data. Traditional VQA models, such as those utilizing VGGNet [1], ResNet [2], GRU [3], and LSTM [4], have been adapted for MedVQA. However, these approaches often struggle with the unique challenges posed by the medical domain, particularly the scarcity of labeled medical data and the need for models to generalize effectively across diverse medical scenarios.

Recent advances in multimodal fusion techniques have shown promise in enhancing MedVQA performance. Bilinear attention networks (BAN) and Transformer-based models have been widely explored for their ability to capture complex interactions between visual and textual modalities. While Transformer models offer powerful fusion capabilities, they come with increased computational complexity, making them less suitable for applications where efficiency is critical.

In this paper, we introduce the Intra-modal and Cross-modal Bilinear Attention Network (IC-BAN), a model designed to address the specific challenges of MedVQA. IC-BAN integrates multi-scale feature extraction, self-attention mechanisms, and bilinear attention networks to achieve a balance between performance and computational efficiency. By leveraging the strengths of BAN and enhancing it with advanced attention mechanisms, our model demonstrates competitive performance without high computational costs.

II. Related Work

A. Medical Visual Question Answering

Medical Visual Question Answering (MedVQA) is an emerging research area within multimodal artificial intelligence that applies the general principles of Visual Question Answering (VQA) to the medical domain. This field combines computer vision and natural language processing techniques to analyze and understand medical images in conjunction with natural language questions, with the goal of generating accurate answers that can assist in clinical decision-making.

Initial research efforts in MedVQA adopt VQA models that have proven effective in general VQA, and adapt them for medical applications.

In terms of image feature extraction, researchers often rely on pre-trained models like VGGNet [1] and ResNet [2], which are fine-tuned for the specific task of MedVQA. On the text side, methods like GRU [3] and LSTM [4] are commonly used to extract textual features, while some approaches incorporate additional semantic information derived from medical corpora to enhance the embeddings used for question representation.

To address the challenges specific to MedVQA, such as the scarcity of labeled medical data, various techniques have been proposed. Nguyen et al. [5] introduced the Model-Agnostic Meta-Learning (MAML) framework combined with a Convolutional Denoising Auto-Encoder (CDAE) to improve feature learning. Similarly, Liu et al. [6] utilized contrastive learning to train a pre-trained model (CPRD) that was then applied to MedVQA tasks. These approaches often use transfer learning to leverage external datasets and pre-trained models to enhance the quality of the extracted image and text features.

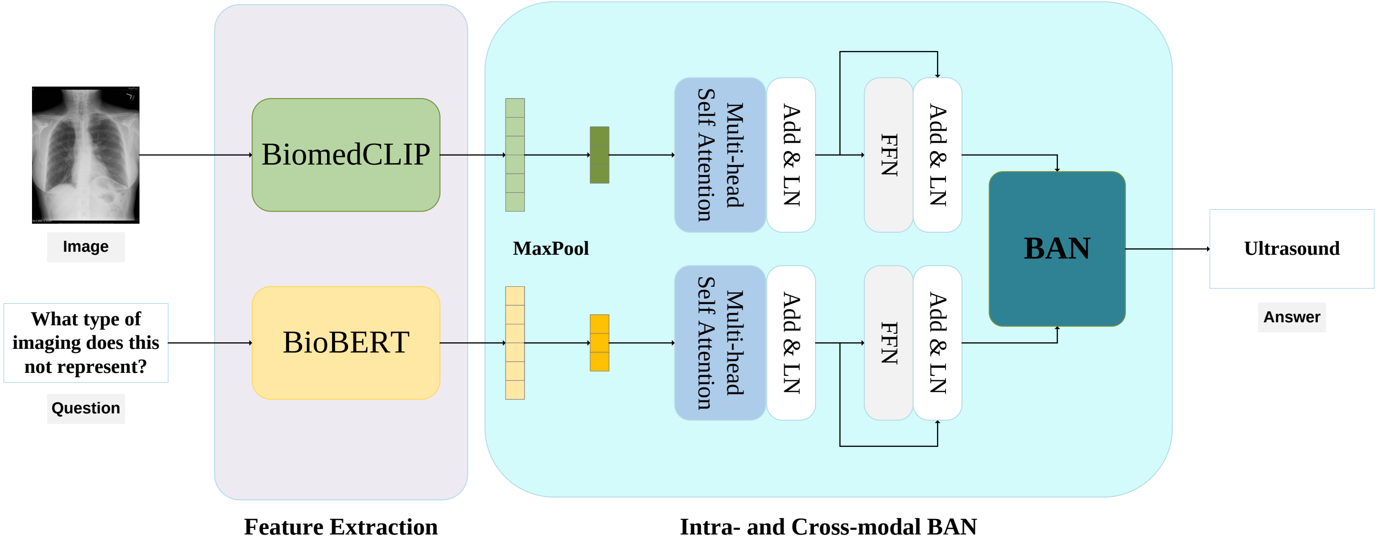
Image
What type of imaging does this not represent?
Question
BiomedCLIP
BioBERT
MaxPool
Multi-head Self Attention
Add & LN
FFN
Add & LN
Multi-head Self Attention
Add & LN
FFN
Add & LN
BAN
Ultrasound
Answer
Feature Extraction
Intra- and Cross-modal BAN

Fig. 1. Overview of our proposed Intra-modal and Cross-modal Bilinear Attention Fusion Network (IC-BAN).

B. *Multi-modal Fusion*

In the Visual Question Answering task, multi-modal fusion plays a crucial role by integrating visual and textual features to enable accurate classification. The performance of VQA models largely depends on the effective extraction of features within modalities and the subsequent fusion across modalities.

Recent approaches to multi-modal fusion in VQA have introduced various methods to enhance the integration of visual and textual features. A foundational method is SAN [7], which iteratively refines attention on relevant image regions based on the given question. Bilinear pooling has been a major focus in improving fusion by capturing complex interactions between modalities. Yu et al. [8] proposed MFB to address the high computational cost of bilinear pooling by factorizing the interaction into two low-rank matrices, preserving performance while reducing complexity. MUTAN [9] builds on this by applying Tucker decomposition to further compress the bilinear tensor, resulting in a more compact and efficient multi-modal representation. To make full use of bilinear attention maps, Kim et al. [10] proposed BAN, which capture intricate dependencies between visual and textual features efficiently.

The advent of Transformer [11] has brought a significant shift in multi-modal fusion strategies. Methods like LXMERT [12], hi-VQA [13] and MCAN [14] utilize cross-modality Transformers with separate encoders for vision and language, followed by a cross-modality encoder to fuse the extracted features. In the domain of Medical Visual Question Answering, Liu et al. [15] employs a Transformer-based architecture that directly fuse image and text features to generate joint representation. While Transformers are effective for fusion, they require more computational resources and are more complex compared to bilinear pooling-based methods. This makes balancing performance and computational cost particularly important in Medical Visual Question Answering, where medical data is often limited.

III. Method

A. *Problem Formulation*

Medical Visual Question Answering is regarded as a classification task, and the objective is to identify the most probable answer a from a predefined set of possible answers $A = \{a_1, a_2, a_3, \ldots, a_n\}$. This prediction can be expressed as:

$$\hat{a} = \arg\max_{a \in A} P(a \mid v_i, q_i) \tag{1}$$

where $P(a \mid v_i, q_i)$ denotes the probability of a given answer a being correct given the image v_i and the question q_i, and $\hat{a}$ is the predicted answer that maximizes this probability.

B. *Multi-modal Feature Extraction*

Image Encoder In this work, we employ the pre-trained BiomedCLIP Image Encoder [16] as our image encoder. This choice is motivated by the unique challenges posed by Med-VQA compared to general VQA tasks. Medical images and questions are inherently more complex, necessitating a model with enhanced feature representation capabilities. CLIP [17], originally designed to learn image representations through natural language supervision, has demonstrated strong zero-shot performance across various domains. BiomedCLIP further fine-tunes CLIP on the PMC-15M dataset, which consists of medical images and associated text, thereby improving its ability to handle medical visual data. We utilize the pre-trained BioMedCLIP to extract image features, where the input image $I_i \in \mathbb{R}^{H \times W \times C}$ is processed through the model, and the output of its projection layer serves as the image representation v_i with a dimensionality of 512-D:

$$v_{hid} = CLIP(I_i) \tag{2}$$

$$v_i = Proj(v_{hid}) \tag{3}$$

Question Encoder For the question encoder, we utilize the pre-trained BioBERT model [18], which has been fine-tuned on a large-scale biomedical corpus. BioBERT is based on the BERT architecture but is specifically adapted to handle the complex and specialized language used in the biomedical domain. This fine-tuning allows BioBERT to capture the nuances of medical terminology more effectively than general-purpose language models.

Compared to traditional approaches in VQA that often rely on LSTM [?] and GRU [?] as text encoders, BERT-based models like BioBERT offer significant advantages, as BERT employs a Transformer architecture, which enables it to capture long-range dependencies and contextual relationships within the text more effectively. This makes representations of the input questions richer and more accurate, making it particularly well-suited for the sophisticated language demands of Medical Visual Question Answering. In our work, BioBERT encodes each question into a 768-D vector representation, denoted as q_i.

C. Intra-modal and Cross-modal Bilinear Attention Fusion Network

Our proposed Intra-modal and Cross-modal Bilinear Attention Network (IC-BAN) integrates multi-scale feature extraction, self-attention mechanisms, and bilinear attention networks to effectively fuse visual and textual features in Medical Visual Question Answering. This comprehensive approach allows the model to capture complex interactions across different levels and modalities without incurring high computational complexity.

1) Multi-scale Feature Fusion: First, the image features $v_i \in \mathbb{R}^{N_v \times d_v}$ and question features $q_i \in \mathbb{R}^{N_q \times d_q}$ are passed through a multi-scale feature extractor, where $N_v = 1$ for the image features (as the CLIP output does not contain region-level information) and N_q represents the sequence length for the question features. The extractor splits each modality into high- and low-scale features:

$$\tilde{v}_i^{\text{low}}, \tilde{v}_i^{\text{high}} = \text{MaxPool}(v_i), v_i \tag{4}$$

$$\tilde{q}_i^{\text{low}}, \tilde{q}_i^{\text{high}} = \text{MaxPool}(q_i), q_i \tag{5}$$

Here, MaxPooling is applied to reduce the dimensionality of the features, producing low-scale features while retaining the original features as high-scale. The use of multi-scale features enables the model to process information at different levels of detail, which is particularly useful for tasks requiring nuanced understanding. The low-scale features capture broader contextual information, while the high-scale features focus on finer details.

Both the high- and low-scale features are then processed through intra-modal self-attention layers to enhance their contextual relationships within single modality. The self-attention mechanism, combined with a feed-forward network (FFN) and layer normalization (LN), is applied separately to each scale:

$$\tilde{v}_i^{\text{att}} = \text{LN}\left(\tilde{v}_i + \text{Dropout}\left(\text{FFN}\left(\text{SelfAttention}(\tilde{v}_i)\right)\right)\right) \tag{6}$$

$$\tilde{q}_i^{\text{att}} = \text{LN}\left(\tilde{q}_i + \text{Dropout}\left(\text{FFN}\left(\text{SelfAttention}(\tilde{q}_i)\right)\right)\right) \tag{7}$$

This process is applied independently to both the low- and high-scale features, producing refined visual and textual features $\tilde{v}_i^{\text{low-att}}, \tilde{v}_i^{\text{high-att}}$ and $\tilde{q}_i^{\text{low-att}}, \tilde{q}_i^{\text{high-att}}$.

2) Cross-modal Bilinear Attention: Following the intra-modal refinement, the IC-BAN model applies a bilinear attention mechanism to fuse the refined visual and textual features. This mechanism computes the interactions between features from different modalities by evaluating attention distributions across all pairs of input channels.

For the high-scale features, the bilinear attention map $\mathbf{A}^{\text{high}}$ is computed as:

$$\mathbf{A}^{\text{high}} = \text{softmax}\left(\left(\mathbf{U}^{\text{high}}\tilde{v}_i^{\text{high-att}}\right) \circ \left(\mathbf{V}^{\text{high}}\tilde{q}_i^{\text{high-att}}\right)\right) \tag{8}$$

where $\mathbf{U}^{\text{high}} \in \mathbb{R}^{d_v \times K}$ and $\mathbf{V}^{\text{high}} \in \mathbb{R}^{d_q \times K}$ are learnable projection matrices, and $\circ$ denotes the Hadamard product. The same process is applied to the low-scale features to compute $\mathbf{A}^{\text{low}}$.

The bilinear attention features for each attention head g are then calculated:

$$\mathbf{f}_g^{\text{high}} = \sum_{j=1}^{N_v}\sum_{k=1}^{N_q} \mathbf{A}_{jk}^{\text{high}} \left(\tilde{v}_i^{j,\text{high-att}}\right)^T \mathbf{U}_g^{\text{high}}\mathbf{V}_g^{\text{high}}\tilde{q}_i^{k,\text{high-att}} \tag{9}$$

$$\mathbf{f}_g^{\text{low}} = \sum_{j=1}^{N_v}\sum_{k=1}^{N_q} \mathbf{A}_{jk}^{\text{low}} \left(\tilde{v}_i^{j,\text{low-att}}\right)^T \mathbf{U}_g^{\text{low}}\mathbf{V}_g^{\text{low}}\tilde{q}_i^{k,\text{low-att}} \tag{10}$$

3) Joint Feature Representation: To form the final joint representation, our method aggregates the bilinear features from all attention maps and from both scales. The features are concatenated and processed through a series of linear projections to produce the final representation:

$$\mathbf{h}^{\text{joint}} = \text{Concat}\left(\sum_{g=1}^{\gamma}\mathbf{f}_g^{\text{high}}, \sum_{g=1}^{\gamma}\mathbf{f}_g^{\text{low}}\right) \tag{11}$$

where γ denotes the number of glimpses for Bilinear Attention Network.

4) Theoretical Analysis and Computational Complexity: In terms of computational complexity, both the IC-BAN model and purely cross-modal Transformer-based models compute attention across all pairs of elements in both modalities, but they differ in how they handle these interactions and in the subsequent processing steps.

In a typical cross-modal Transformer-based model, the attention mechanism has a complexity of $O(N_v \times N_q \times d)$ per layer, where N_v and N_q are the number of visual and textual elements, respectively, and d is the feature dimension. This complexity arises because the Transformer computes attention weights for every possible pair of elements in the two modalities, followed by a weighted sum that produces a new set of feature representations.

In contrast, the Bilinear Attention Network used in IC-BAN also calculates interactions across all pairs, but it does so in bilinear space by factorizing the interaction into lower-dimensional spaces. Specifically, BAN employs learnable projection matrices $\mathbf{U}$ and $\mathbf{V}$, which project the image and text features into a joint space where the interactions are computed. While this operation still involves all pairs of elements, the subsequent bilinear pooling and the use of lower-dimensional projections (i.e., K instead of d) make the computation more efficient in practice. The complexity of this bilinear interaction is $O(N_v \times N_q \times K)$, where K is typically smaller than d, particularly in cases where the dimensionality reduction is significant.

Additionally, in Transformer-based models, after the attention computation, the features are passed through multiple layers of feed-forward networks, which further add to the computational burden. In contrast, IC-BAN benefits from bilinear attention's direct aggregation of multimodal features (except for integrated single self-attention layer), reducing the need for extensive post-attention processing.

Therefore, while both models compute interactions across all pairs, IC-BAN achieves a lower overall complexity by reducing the dimensionality of the interaction space and simplifying the subsequent processing steps.

D. Classifier and Loss Function

Finally, the joint representation output by the IC-BAN is fed into the classifier to predict the most probable answer. We utilize a simple feed-forward neural network as the classifier. It consists of two fully connected layers with an intermediate activation function.

$$\hat{a} = \text{Classify}(\mathbf{h}^{\text{joint}}) \tag{12}$$

For the loss function, we employ the Binary Cross Entropy with Logits Loss, which is commonly used in multi-label classification tasks.

IV. Experiments

A. Dataset and Metric

We use VQA-RAD dataset in our experiments, which contains 3,515 QA pairs based on 315 radiology images. The questions in this dataset are categorized into two types: Closed and Open. Closed questions are those with a limited set of possible answers, most commonly yes/no, while Open questions are more varied and do not restrict the type of response. The dataset is split into a training set with 3,064 question–answer pairs and a test set with 451 pairs.

For the evaluation metric, we primarily use accuracy, as MedVQA can be viewed as a multi-class classification task. Accuracy measures the proportion of correctly predicted answers out of the total number of questions, providing a straightforward and widely accepted indicator of model performance in this context.

B. Experimental Setup

We conduct the experiments on a single NVIDIA RTX 3090 (24GB) GPU. The learning rate is set to 0.0005, with a batch size of 32. We train the model for 40 epochs, and the best-performing model on the validation set is saved as the representative model. The Adamax optimizer is used to optimize the model parameters.

TABLE I
Comparison of Accuracy (%) on VQA-RAD test set.

Methods	**Accuracy**		
	Open	***Closed***	***All***
MEVF+SAN [5]	40.7	74.1	60.8
MEVF+BAN [5]	43.9	75.1	62.7
MMQ [20]	53.7	75.8	67.0
QCR [19]	60.0	79.3	71.6
MEVF+PubMedCLIP [21]	48.6	78.1	66.5
QCR+PubMedCLIP [21]	60.1	80.0	72.1
BiomedCLIP [16]	67.6	79.8	75.2
BiomedCLIP+IC-BAN(Ours)	67.6	**80.9**	**75.6**

At the model level, based on the architecture described in III, we integrate the TCR module [19], which includes a pre-trained question classifier. This classifier is used to distinguish between open and closed questions. For each type of question, we train and validate separate models, allowing for specialized handling of the unique characteristics inherent in open and closed question types.

C. Results and Analysis

The results of our experiments on the VQA-RAD test set are summarized in Table I. As shown, the performance of different methods is evaluated across open, closed, and overall accuracy categories. The baseline method MEVF+SAN [5] achieves an overall accuracy of 60.8%, while MEVF+BAN [5] slightly improves this to 62.7%. The MMQ [20] method further increases the overall accuracy to 67.0%. The QCR [19] model demonstrates a significant improvement, reaching 71.6% overall accuracy. When PubMedCLIP [21] is integrated into the QCR model, the accuracy further improves to 72.1%. BiomedCLIP [16] achieves 75.2% accuracy, with 67.6% on open questions and 79.8% on closed questions. Our proposed BiomedCLIP+IC-BAN method achieves the highest accuracy on the closed questions at 80.9% and an overall accuracy of 75.6%, which is a modest improvement over the original BiomedCLIP model.

From the experiments, it is evident that our method shows a slight improvement in accuracy compared to the original BiomedCLIP approach. Although the improvement is limited, it is important to note that the original method used a Transformer-based fusion model (as described in [22]). This indicates that BAN's capability in cross-modal fusion is not necessarily inferior to that of Transformers. The results suggest that BAN can effectively capture the interactions between visual and textual modalities, and in some cases, it may even outperform Transformer-based models. The ability of BAN to model complex interactions in a more computationally efficient manner could be one of the contributing factors to this performance. Additionally, the integration of the IC-BAN model with BiomedCLIP demonstrates the potential of combining advanced fusion techniques with domain-specific feature extractors to achieve high performance in challenging tasks like MedVQA.

V. Conclusion and Future Work

In this paper, we proposed the Intra-modal and Cross-modal Bilinear Attention Network (IC-BAN) for the task of Medical Visual Question Answering (MedVQA). Our model integrates multi-scale feature extraction, self-attention mechanisms, and bilinear attention networks to effectively fuse visual and textual features. Compared to traditional BAN fusion methods and Transformer-based fusion methods, IC-BAN achieves a balanced performance in terms of accuracy and computational efficiency. Specifically, our experimental results demonstrate that IC-BAN, when combined with BiomedCLIP, slightly outperforms the original Transformer-based BiomedCLIP model. This indicates that BAN, with its ability to efficiently model complex interactions between modalities, remains a competitive approach even when compared to more recent Transformer-based methods.

For future work, there are several directions worth exploring. One potential avenue is to modify the internal attention computation structure of BAN to further enhance its ability to capture cross-modal interactions. Additionally, incorporating more advanced attention mechanisms or integrating external medical knowledge could further improve the model's performance on MedVQA tasks.

References

[1] K. Simonyan, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[3] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[4] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[5] B. D. Nguyen, T.-T. Do, B. X. Nguyen, T. Do, E. Tjiputra, and Q. D. Tran, "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*. Springer, 2019, pp. 522–530.

[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 *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*. Springer, 2021, pp. 210–220.

[7] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 21–29.

[8] Z. Yu, J. Yu, J. Fan, and D. Tao, "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1821–1830.

[9] H. Ben-Younes, R. Cadene, M. Cord, and N. Thome, "Mutan: Multimodal tucker fusion for visual question answering," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2612–2620.

[10] J.-H. Kim, J. Jun, and B.-T. Zhang, "Bilinear attention networks," *Advances in neural information processing systems*, vol. 31, 2018.

[11] A. Vaswani, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[12] H. Tan and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," *arXiv preprint arXiv:1908.07490*, 2019.

[13] C. Pellegrini, M. Keicher, E. Özsoy, and N. Navab, "Rad-restruct: A novel vqa benchmark and method for structured radiology reporting," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 409–419.

[14] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, "Deep modular co-attention networks for visual question answering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6281–6290.

[15] 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*. Springer, 2023, pp. 445–456.

[16] S. Zhang, Y. Xu, N. Usuyama, H. Xu, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri *et al.*, "Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs," *arXiv preprint arXiv:2303.00915*, 2023.

[17] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[18] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[19] L.-M. Zhan, B. Liu, L. Fan, J. Chen, and X.-M. Wu, "Medical visual question answering via conditional reasoning," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2345–2354.

[20] T. Do, B. X. Nguyen, E. Tjiputra, M. Tran, Q. D. Tran, and A. Nguyen, "Multiple meta-model quantifying for medical visual question answering," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*. Springer, 2021, pp. 64–74.

[21] S. Eslami, C. Meinel, and G. De Melo, "Pubmedclip: How much does clip benefit visual question answering in the medical domain?" in *Findings of the Association for Computational Linguistics: EACL 2023*, 2023, pp. 1181–1193.

[22] Z.-Y. Dou, Y. Xu, Z. Gan, J. Wang, S. Wang, L. Wang, C. Zhu, P. Zhang, L. Yuan, N. Peng *et al.*, "An empirical study of training end-to-end vision-and-language transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 166–18 176.