This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Med-BiasX: Robust Medical Visual Question Answering with Language Biases

Huanjia Zhu[1,3], Yishu Liu[2†], Chengju Zhou[1], Guangming Lu[2], and Bingzhi Chen[3†]

[1] South China Normal University
[2] Harbin Institute of Technology, Shenzhen
[3] Beijing Institute of Technology, Zhuhai
liuyishu@stu.hit.edu.cn, chenbingzhi@bit.edu.cn

Abstract. In Medical Visual Question Answering (Med-VQA), accurate interpretation of clinical questions alongside medical images is crucial for reliable diagnostic support. However, conventional methods often exhibit pronounced medical language biases that stem from *imbalanced data distribution* and *question shortcut dependence*, causing models to disproportionately rely on textual priors at the expense of valuable visual semantics. To mitigate this challenge, we propose a novel Med-VQA debiasing approach called "**Med-BiasX**" that synergistically combines two strategies, i.e., Energy-aware Confidence Constraint (ECC) and Distribution-aware Dependence Calibration (DDC). Specifically, ECC aims to reinforce correct answers and adjust the energy associated with incorrect answers by leveraging the global normalization property of free energy and the intrinsic properties of energy. DDC is designed to shift the model's dependency from question shortcuts to multimodal information by explicitly measuring the similarity between predicted distributions from different branches and prior distributions. Extensive experiments on multiple medical standard benchmarks and bias-sensitive benchmarks, SLAKE-BIAS and VQA-RAD-BIAS, consistently demonstrate the robustness and superiority of our Med-BiasX approach over state-of-the-art competitors.

Keywords: Medical · Visual Question Answering · Language Biases · Energy Function · Debiasing

1 Introduction

Medical visual question answering (Med-VQA) systems [15, 13, 6, 5] operate at the intersection of artificial intelligence and healthcare, aiming to develop models that can accurately interpret medical images and answer clinical questions by integrating visual and textual data. Early research adopts cross-modal inference strategies from general VQA models [19, 21, 8]. Given the significant differences between medical and general images and scarce medical data, direct transfer

[†]Corresponding authors: Bingzhi Chen and Yishu Liu

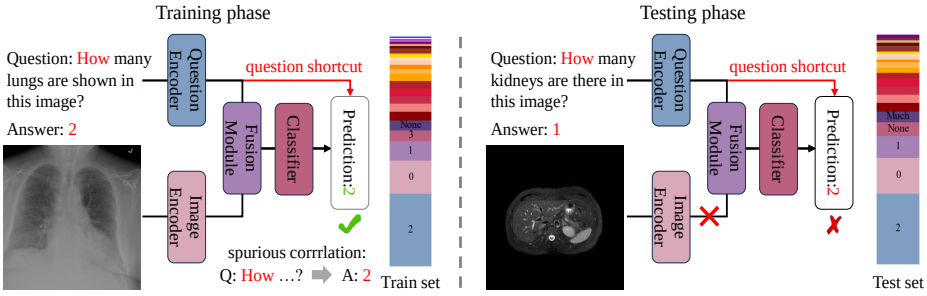

Fig. 1: During training, conventional Med-VQA models tend to learn spurious correlation between questions and answers. Therefore, during testing, they ignore the semantic visual information and rely on question shortcuts for predictions.

leads to severe overfitting. To address data scarcity, Nguyen et al. [17] propose MEVF. Besides, RUBi [3] uses question-only branch outputs as masks to down-weight biased predictions and up-weight informative ones. These approaches have yielded notable improvements in the Med-VQA task. Our approach goes further by detecting and quantifying bias through energy-aware calibration and distribution-aware measures, explicitly removing spurious correlations.

However, a recent study [22] demonstrates that traditional Med-VQA models suffer from **medical language biases**, as shown in Fig. 1. Theoretically, medical language biases stems from ***imbalanced data distribution*** and ***question shortcut dependence*** [24, 4, 9, 12, 18]. On the one hand, medical datasets often exhibit significant imbalances (the question types, such as *how* and *what*, and their corresponding answers), where rare diseases or uncommon diagnostic outcomes are underrepresented while common cases dominate the data. This skewed distribution causes models to capture frequently occurring patterns at the expense of rare but clinically critical features, substantially reducing diagnostic performance in uncommon or complex cases. On the other hand, standardized expressions and inherent semantic patterns, i.e., question types, prevalent in clinical interviews and medical records lead to spurious correlation between questions and answers in the training set when data is imbalanced. The spurious correlation causes models to rely on superficial linguistic cues for "shortcut" reasoning rather than deeply analyzing the underlying pathological information in medical images. Such shortcut dependence not only increases the risk of misdiagnosis in complex or borderline cases but also limits the system's generalizability to novel, unseen medical scenarios.

To address these challenges, we propose a novel Med-VQA debiasing approach, namely "**Med-BiasX**", for improving the robust reasoning ability of Med-VQA. The proposed Med-BiasX method integrates two well-established mechanisms, i.e., **E**nergy-aware **C**onfidence **C**onstraint (**ECC**) and **D**istribution-aware **D**ependence **C**alibration (**DDC**). Specifically, inspired by free energy [16], the purpose of ECC is to enforce confidence constraints by leveraging the global normalization property of the log-sum-exp operation to comprehensively adjust

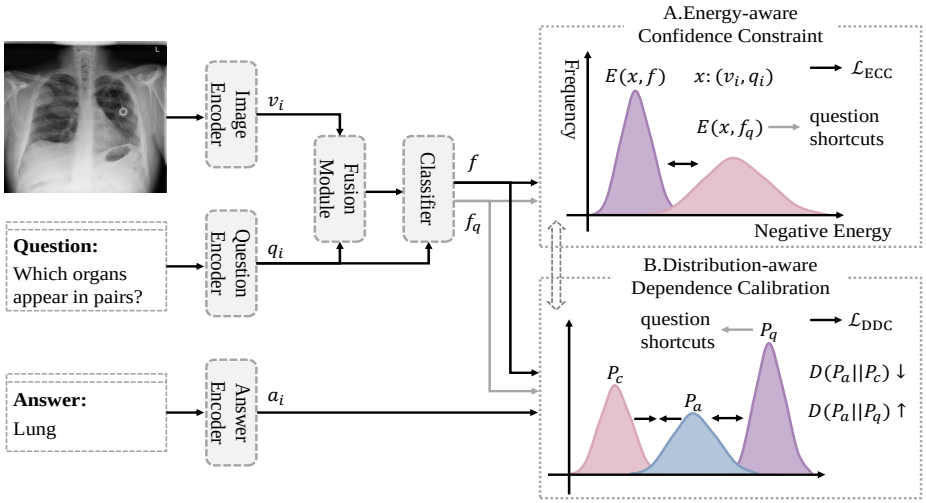

Fig. 2: Illustration of the proposed Med-BiasX approach for addressing medical language biases. Two well-established mechanisms, i.e., Energy-aware Confidence Constraint and Distribution-aware Dependence Calibration, are synergistically integrated to suppress question shortcuts and improve the model's robustness.

the energy states of all candidate answers. Based on the inherent characteristic that low energy correlates with high confidence, ECC seeks to maximize the energy associated with question shortcuts while minimizing the energy of joint predictions. Furthermore, DDC empowers the network to recalibrate the prediction dependence by measuring the similarity between predicted distributions from different branches and the prior distribution. Benefiting from the quantification of the model's reliance on question shortcuts, DDC dynamically adjusts the model's focus on multimodal features to improve robustness and accuracy.

2 Methodology

2.1 Preliminaries

The purpose of the Med-VQA task is to answer the question based on an image. Given a dataset $\mathcal{D} = \{v_i, q_i, a_i\}_{i=1}^{N}$, consisting of an image $v_i \in \mathcal{V}$, a question $q_i \in \mathcal{Q}$ and a ground-truth answer $a \in \mathcal{A}$, the goal is to optimize a mapping $f_{\mathcal{VQ}} : \mathcal{V} \times \mathcal{Q} \rightarrow \mathbb{R}^{|\mathcal{A}|}$ to generate predictions corresponding to a given image-question pair. Without loss of generality, the function can be composed as follows:

$$f(v_i, q_i) = c(g(e_v(v), e_q(q)), \tag{1}$$

where e_v and e_q are the image encoder and question encoder, respectively. $g(\cdot)$ denotes the multi-model fusion network, and $c(\cdot)$ is an answer classifier to generate the logits f.

2.2 Energy-aware Confidence Constraint

To address the challenges posed by language biases, the ECC mechanism is meticulously designed to perform confidence constraint by dynamically adjusting the energy states of all candidate answers based on the inherent properties of energy. Specifically, we introduce a question-only branch as a bias-assisted model with corresponding logits f_q:

$$f_q(q_i) = c(g(e_q(q))). \tag{2}$$

Inspired by [16], we incorporate the free energy function into our training objectives. *Energy* refers to a non-probabilistic scalar obtained by mapping each input point x via a function $E(x) : \mathbb{R}^D \to \mathbb{R}$, with D denoting the dimensionality of the input space. The free energy function $E(x, f)$ is given by:

$$E(x, f) = -T \cdot \log \sum_{i}^{C} e^{f_i(x)/T}, \tag{3}$$

where C is the number of candidate answers and T is the temperature parameter. Free Energy converts the uncertainty and overall confidence predicted by the model into a scalar by applying a log-sum-exp operation on the logits of all candidate answers. This global normalization captures the model's overall confidence in an input, preventing it from merely focusing on the highest local logit and ignoring the contributions of other answers.

In Med-VQA, medical language biases often guide the model to rely exclusively on the question text, resulting in overly confident (low-energy) yet incorrect predictions. Free Energy thus serves as a simple and effective indicator to quantify this bias. Based on Eq. (3), the confidence constraint can be formulated using the squared hinge loss with a margin hyperparameter m:

$$\mathcal{L}_{\text{ECC}} = (\max(0, m + E(x, f) - E(x, f_q)))^2. \tag{4}$$

If the energy difference $E(x, f) - E(x, f_q) > -m$, the model penalizes question shortcuts and promotes multimodal learning, thus preventing the model from reaching a low-energy state in question-only scenarios and encouraging it to incorporate visual information for more robust predictions.

2.3 Distribution-aware Dependence Calibration

Inspired by research on question shortcuts [7, 24], our DDC mechanism implicitly calibrates the dependence of model prediction by promoting multimodal representation learning while penalizing question shortcuts. The degree of spurious correlation is reflected by applying the Kullback-Leibler divergence [20] $D(\cdot||\cdot)$:

$$D(P_a||P_q) = \sum P_a \log \frac{P_a}{P_q}, \tag{5}$$

where P_a and P_q are the probability distributions of ground-truth answers and the question-only branch outputs f_q, respectively. A small $D(P_a||P_q)$ indicates that the question-only prediction closely resembles the answer's prior distribution, implying an over-reliance on spurious correlation. We thus treat the negative of $D(P_a||P_q)$ as an indicator of question shortcuts and penalize it. To encourage the model to make predictions based on the semantic information of multimodal features, we further define:

$$D(P_a||P_c) = \sum P_a \log \frac{P_a}{P_c}, \tag{6}$$

where P_c is the probability distribution of logits f_c. Finally, we combine these divergences into the DDC loss:

$$\mathcal{L}_{\text{DDC}} = \exp(D(P_a||P_c) - D(P_a||P_q)) + \log(\frac{D(P_a||P_c)}{D(P_a||P_q)} + 1), \tag{7}$$

where $D(P_a||P_c)$ is intended to steer the model towards extracting and utilizing robust semantic joint representations, thereby ensuring that predictions are grounded in meaningful image-text correlations. In contrast, $D(P_a||P_q)$ serves as an indicator of bias by quantifying the model's reliance on spurious correlation. During the training progress, we observe a decreasing trend in $D(P_a||P_c)$, which reflects improved alignment with the semantic content, while an increasing trend in $D(P_a||P_q)$ highlights the model's increasing awareness of and correction against shortcut biases. These empirical trends validate our approach and underscore its effectiveness in promoting reliable and unbiased predictions.

2.4 Training and Optimization

Based on the above analyses, the comprehensive training objective of the proposed Med-BiasX approach encompasses a combination of various loss functions:

$$\mathcal{L}_{\text{TOTAL}} = \mathcal{L}_{\text{RMLVQA}} + \mathcal{L}_{\text{ECC}} + \mathcal{L}_{\text{DDC}}. \tag{8}$$

We adopt $\mathcal{L}_{\text{RMLVQA}}$ [2] as the base loss term. By jointly optimizing these losses, our approach can enhance the robustness and reliability of Med-VQA systems.

3 Experiments

3.1 Datasets and Implementation Details.

Datasets. We evaluate the effectiveness of our Med-BiasX on two medical benchmark datasets, SLAKE [14] and VQA-RAD [10], and two bias-sensitive benchmark evaluation protocols, SLAKE-BIAS [23] and VQA-RAD-BIAS [23].

Implementations Details. We employ a popular VQA architecture UpDn as our baseline [1]. We implemented our Med-BiasX model in PyTorch with a single RTX 3090 GPU and used the AdamW optimizer with a weight decay of 0.001. The batch size B is set to 64. The learning rate is set to 0.002. The value of the margin hyper-parameter m is set to 1.0.

Table 1: Comparisons with SOTAs on the SLAKE-BIAS and VQA-RAD-BIAS.

Methods	Reference	SLAKE-BIAS			VQA-RAD-BIAS		
		All	Open	Closed	All	Open	Closed
SAN [19]	CVPR'16	26.02	48.30	6.42	16.29	59.73	6.05
MFB [21]	ICCV'17	30.56	55.70	8.44	22.53	72.12	10.84
BAN [8]	NIPS'18	17.50	30.90	5.72	17.30	67.70	5.42
UpDn [2]	CVPR'18	31.45	59.90	6.42	26.67	74.78	15.33
MEVF+SAN [17]	MICCAI'19	18.62	32.60	6.33	22.11	68.14	11.26
MEVF+BAN [17]	MICCAI'19	19.33	35.00	5.54	19.07	62.39	8.86
RUBi [3]	NIPS'19	33.88	60.30	10.64	81.27	60.62	86.13
LPF [11]	SIGIR'21	40.34	43.70	37.38	41.52	65.04	35.97
GGE-iter [7]	ICCV'21	35.05	61.30	11.96	21.60	51.33	14.60
RMLVQA [2]	CVPR'23	76.42	60.50	**90.41**	89.45	69.03	94.26
Med-BiasX	Ours	**78.33**	**64.80**	90.24	**91.48**	**76.11**	**95.10**
	Increased ↑	+1.91	+3.50	-0.17	+2.03	+1.33	+0.84

Table 2: Comparisons with SOTAs on the SLAKE and VQA-RAD.

Methods	Reference	SLAKE			VQA-RAD		
		All	Open	Closed	All	Open	Closed
SAN [19]	CVPR'16	76.00	74.00	79.10	52.89	31.64	65.50
MFB [21]	ICCV'17	73.89	71.63	77.40	54.10	41.90	62.13
BAN [8]	NIPS'18	76.25	75.97	76.68	55.43	48.60	59.93
UpDn [2]	CVPR'18	81.34	79.84	83.65	66.74	**51.40**	76.47
MEVF+SAN [17]	MICCAI'19	75.97	74.72	77.88	60.71	40.65	74.05
MEVF+BAN [17]	MICCAI'19	77.76	75.97	80.53	62.34	43.09	75.14
RUBi [3]	NIPS'19	78.42	76.43	81.49	51.22	36.87	60.66
LPF [11]	SIGIR'21	75.59	73.33	79.09	56.32	49.72	60.66
GGE-iter [7]	ICCV'21	79.83	79.22	80.77	65.19	49.16	75.74
RMLVQA [2]	CVPR'23	81.43	80.47	82.93	65.41	49.16	76.10
Med-BiasX	Ours	**82.47**	**80.62**	**85.34**	**67.42**	50.64	**78.46**
	Increased ↑	+1.04	+0.15	+1.69	+0.68	-0.76	+1.99

Table 3: Ablation studies for different modules of Med-BiasX on SLAKE-BIAS.

Methods	ECC	DDC	All	Open	Closed
Baseline	-	-	76.42	64.80	90.41
w/ ECC	✓	-	77.82	63.30	**90.59**
w/ DDC	-	✓	77.12	62.50	89.97
Med-BiasX	✓	✓	**78.33**	**64.80**	90.24

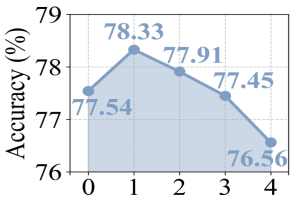

Fig. 3: Comparison of accuracy on SLAKE-BIAS with different parameter configurations.

3.2 Comparisons with State-of-The-Arts

To evaluate the effectiveness of our Med-BiasX in mitigating language biases, we conduct comprehensive experiments against a variety of state-of-the-art baselines on SLAKE-BIAS, VQA-RAD-BIAS, SLAKE, and VQA-RAD datasets. The comparative results are summarized in Table 1 and Table 2.

Evaluation on Medial Language Biases Benchmark: We can obtain the following observations: 1) The proposed Med-BiasX method outperforms all state-of-the-art baselines with significant improvements, underscoring its robustness in addressing medical language biases. 2) In particular, most methods exhibit a substantial performance decline on SLAKE-BIAS and VQA-RAD-BIAS compared to SLAKE and VQA-RAD. 3) Natural debiasing models [11, 3, 7] yield suboptimal results, highlighting limited generalizability in the medical domain.

Evaluation on Medial Standard Benchmark: To verify the ID performance of our Med-BiasX, we conduct reliable experiments on SLAKE and VQA-RAD. Importantly, the proposed Med-BiasX approach consistently achieves the best performance and surpasses the second-best method by 1.04% and 0.68%, respectively. It is crucial to note that many existing methods perform well on ID data and experience pronounced degradation under OOD conditions.

3.3 Ablation Studies

We conduct comprehensive ablation studies by systematically evaluating the impact of each component on SLAKE-BIAS. Four variations are involved, including 1) **BASE** is regarded as the base loss [2]. 2) **BASE w/ ECC** adds the ECC mechanism based on the basic model. 3) **BASE w/ DDC** adds the DDC mechanism to the basic model. 4) **Med-BiasX** is considered as the "full" model. The comparative results are presented in Table 3. It is observed that both ECC and DDC components can impair spurious correlation and enhance robustness.

3.4 Parameter Analysis

We conduct an exhaustive parameter analysis of the proposed Med-BiasX method under different hyper-parameter configurations. Specifically, we focus on analyzing the effects of the margin hyper-parameter m, as shown in Eq. (4). Through

systematic experimentation and detailed analysis in Fig. 3, it can be observed that our Med-BiasX method achieves peak performance when $m = 1$. This analysis highlights that an optimal selection of the margin hyper-parameter can achieve superior performance of the Med-BiasX model.

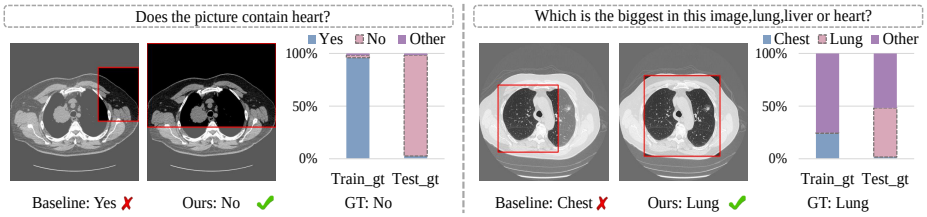

Fig. 4: Qualitative analysis of our Med-BiasX. We present two representative examples for the question type "Yes/No" (left) and "Which" (right).

3.5 Qualitative analysis

To further demonstrate the effectiveness of our Med-BiasX method, we conduct a qualitative analysis on SLAKE-BIAS for Med-BiasX and the baseline UpDn, as depicted in Fig. 4. In both examples, regardless of whether the correct region in the image was overlooked (left) or identified (right), UpDn predicted the most frequent answer for corresponding question type in the train set, revealing its reliance on spurious correlation. In contrast, our Med-BiasX model focuses on the correct regions and successfully predicts rare answer categories. These examples underscore the effectiveness and robustness of our method.

4 Conclusion

In this paper, we identified the challenge of medical language biases posed by imbalanced data distribution and question shortcut dependence. To address this challenge, we proposed a Med-VQA debiasing approach called "Med-BiasX", which integrated energy-aware confidence constraint and distribution-aware dependence calibration for robust Med-VQA. The proposed Med-BiasX method leverages the inherent properties of energy to implicitly conduct confidence constraint and measures the similarity between prediction distributions from different branches and prior distributions to recalibrate dependence on multimodal representation. Comprehensive experiments prove the effectiveness and robustness of our Med-BiasX method.

Acknowledgments. This work was supported in part by the Shenzhen Fundamental Research Fund (No. JCYJ20240813105900002), in part by the Guangdong Basic and Applied Basic Research Foundation (No. 2025A1515010225), and in part by the National Natural Science Foundation of China (No. 62302172).

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

References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6077–6086 (2018)
2. Basu, A., Addepalli, S., Babu, R.V.: Rmlvqa: A margin loss approach for visual question answering with language biases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11671–11680 (2023)
3. Cadene, R., Dancette, C., Ben-younes, H., Cord, M., Parikh, D.: Rubi: Reducing unimodal biases in visual question answering. In: Proceedings of the Conference and Workshop on Neural Information Processing Systems (NeurIPS). vol. 32 (2019)
4. Chen, L., Yan, X., Xiao, J., Zhang, H., Pu, S., Zhuang, Y.: Counterfactual samples synthesizing for robust visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10800–10809 (2020)
5. Chen, Z., Du, Y., Hu, J., Liu, Y., Li, G., Wan, X., Chang, T.H.: Multi-modal masked autoencoders for medical vision-and-language pre-training. In: Medical Image Computing and Computer Assisted Intervention (MICCAI). pp. 679–689. Springer (2022)
6. Do, T., Nguyen, B.X., Tjiputra, E., Tran, M., Tran, Q.D., Nguyen, A.: Multiple meta-model quantifying for medical visual question answering. In: Medical Image Computing and Computer Assisted Intervention (MICCAI). pp. 64–74. Springer (2021)
7. Han, X., Wang, S., Su, C.: Greedy gradient ensemble for robust visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 1584–1593 (2021)
8. Kim, J.H., Jun, J., Zhang, B.T.: Bilinear attention networks. In: Proceedings of the Conference and Workshop on Neural Information Processing Systems (NeurIPS). vol. 31 (2018)
9. Kolling, C., More, M., Gavenski, N., Pooch, E., Parraga, O., Barros, R.C.: Efficient counterfactual debiasing for visual question answering. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 3001–3010 (2022)
10. Lau, J.J., Gayen, S., Ben Abacha, A., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. Scientific data **5**(1), 1–10 (2018)
11. Liang, Z., Hu, H., Zhu, J.: Lpf: A language-prior feedback objective function for debiased visual question answering. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR). pp. 1955–1959 (2021)
12. Liang, Z., Jiang, W., Hu, H., Zhu, J.: Learning to contrast the counterfactual samples for robust visual question answering. In: Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 3285–3292 (2020)
13. Liu, B., Zhan, L.M., Wu, X.M.: Contrastive pre-training and representation distillation for medical visual question answering based on radiology images. In: Medical

Image Computing and Computer Assisted Intervention (MICCAI). pp. 210–220. Springer (2021)
14. 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)
15. Liu, B., Zhan, L.M., Xu, L., Wu, X.M.: Medical visual question answering via conditional reasoning and contrastive learning. IEEE Transactions on Medical Imaging (TMI) **42**(5), 1532–1545 (2022)
16. Liu, W., Wang, X., Owens, J., Li, Y.: Energy-based out-of-distribution detection. In: Proceedings of the Conference and Workshop on Neural Information Processing Systems (NeurIPS). vol. 33, pp. 21464–21475 (2020)
17. 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). pp. 522–530. Springer (2019)
18. Wen, Z., Xu, G., Tan, M., Wu, Q., Wu, Q.: Debiased visual question answering from feature and sample perspectives. In: Proceedings of the Conference and Workshop on Neural Information Processing Systems (NeurIPS). vol. 34, pp. 3784–3796 (2021)
19. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21–29 (2016)
20. Yao, Z., Lai, Z., Liu, W.: A symmetric kl divergence based spatiogram similarity measure. In: 2011 18th IEEE International Conference on Image Processing. pp. 193–196. IEEE (2011)
21. Yu, Z., Yu, J., Fan, J., Tao, D.: Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1821–1830 (2017)
22. Zhan, C., Peng, P., Zhang, H., Sun, H., Shang, C., Chen, T., Wang, H., Wang, G., Wang, H.: Debiasing medical visual question answering via counterfactual training. In: Medical Image Computing and Computer Assisted Intervention (MICCAI). pp. 382–393. Springer (2023)
23. Zhu, H., Liu, Y., Zhou, C., Lu, G., Chen, B.: Cause-effect driven optimization for robust medical visual question answering with language biases. arXiv preprint arXiv:2506.17903 (2025)
24. Zhu, J., Liu, Y., Zhu, H., Lin, H., Jiang, Y., Zhang, Z., Chen, B.: Combating visual question answering hallucinations via robust multi-space co-debias learning. In: Proceedings of the ACM International Conference on Multimedia (MM). pp. 955–964 (2024)