Debiasing Medical Visual Question Answering via Counterfactual Training

Chenlu Zhan[1], Peng Peng[2], Hanrong Zhang[2], Haiyue Sun[2], Chunnan Shang[2], Tao Chen[3], Hongsen Wang[3], Gaoang Wang[1,2(✉)], and Hongwei Wang[1,2(✉)]

[1] College of Computer Science and Technology, Zhejiang University, Zhejiang, China
chenlu.22@intl.zju.edu.cn
[2] ZJU-UIUC Institute, Zhejiang University, Zhejiang, China
{pengpeng,hanrong.22,haiyue.22,chunnan.22,gaoangwang}@intl.zju.edu.cn,
hongweiwang@zju.edu.cn
[3] Department of Cardiology, Chinese PLA General Hospital, Beijing, China

Abstract. Medical Visual Question Answering (Med-VQA) is expected to predict a convincing answer with the given medical image and clinical question, aiming to assist clinical decision-making. While today's works have intention to rely on the superficial linguistic correlations as a shortcut, which may generate emergent dissatisfactory clinic answers. In this paper, we propose a novel DeBiasing Med-VQA model with CounterFactual training (DeBCF) to overcome language priors comprehensively. Specifically, we generate counterfactual samples by masking crucial keywords and assigning irrelevant labels, which implicitly promotes the sensitivity of the model to the semantic words and visual objects for bias-weaken. Furthermore, to explicitly prevent the cheating linguistic correlations, we formulate the language prior into counterfactual causal effects and eliminate it from the total effect on the generated answers. Additionally, we initiatively present a newly splitting bias-sensitive Med-VQA dataset, Semantically-Labeled Knowledge-Enhanced under Changing Priors (SLAKE-CP) dataset through regrouping and re-splitting the train-set and test-set of SLAKE into the different prior distribution of answers, dedicating the model to learn interpretable objects rather than overwhelmingly memorizing biases. Experimental results on two public datasets and SLAKE-CP demonstrate that the proposed DeBCF outperforms existing state-of-the-art Med-VQA models and obtains significant improvement in terms of accuracy and interpretability. To our knowledge, it's the first attempt to overcome language priors in Med-VQA and construct the bias-sensitive dataset for evaluating debiased ability.

C. Zhan and P. Peng—Contributed equally to this work.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43895-0_36.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14221, pp. 382–393, 2023.
https://doi.org/10.1007/978-3-031-43895-0_36

Keywords: Medical Vision Question Answering · Language Bias · Counterfactual Sample Generation · Counterfactual Training · SLAKE-CP

1 Introduction

Medical visual question answering (Med-VQA) has attracted considerable attention in recent years. It seeks to discover the plausible answer by evaluating the visual information of a medical image and a clinical query regarding the image. The Med-VQA technology can considerably enhance the efficiency of medical professionals and fulfill the growing demand for medical resources [15,25]. However, numerous researches have found that general VQA models are significantly influenced by superficial linguistic correlations in training set, lacking adequate visual grounding [9,14,26]. Since most of the existing Med-VQA datasets [12,16] are manually spitted and annotated, the spurious over-reliant bias factor also exists in Med-VQA, as the Fig. 1 shown. Recent general VQA works dedicate to reducing the language priors through enhancing the visual information [23,27] or data balancing [11,13], there is bare attempt to prevent the language priors in medical domain. Current Med-VQA works [4,5,15,17] devote to construct effective models and most Med-VQA datasets [12,16] simply balance the medical images to mitigate the inherent bias. These works all neglect the cheating factors that the Med-VQA models typically resort to linguistic distributions priors, consequently ignoring the semantic clinic objects. This problem accordingly leads to disastrous results in clinic application consequences.

Therefore, we propose a novel unbiased and interpretable Med-VQA model and preliminarily construct a bias-sensitive Med-VQA dataset to address the problems mentioned above. First, with the aim of forcing the model to focus on clinic objects rather than superficial correlations, we prepare the counterfactual samples by masking clinic words with "[MASK]" tokens and meanwhile assign the irrelevant answers for implicit bias-weaken. Further, for explicitly reducing the linguist bias, we treat the language bias as the causal effect of the clinic

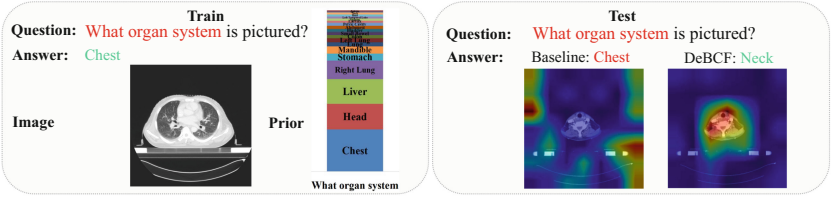

Fig. 1. The baseline generates incorrect answer "Chest" relying on the majority prior "Chest" in train-set of the publicly available SLAKE [16] dataset rather than real semantic image objects. The proposed DeBCF overcomes the language priors and generates the reliable answer with correct semantic parts.

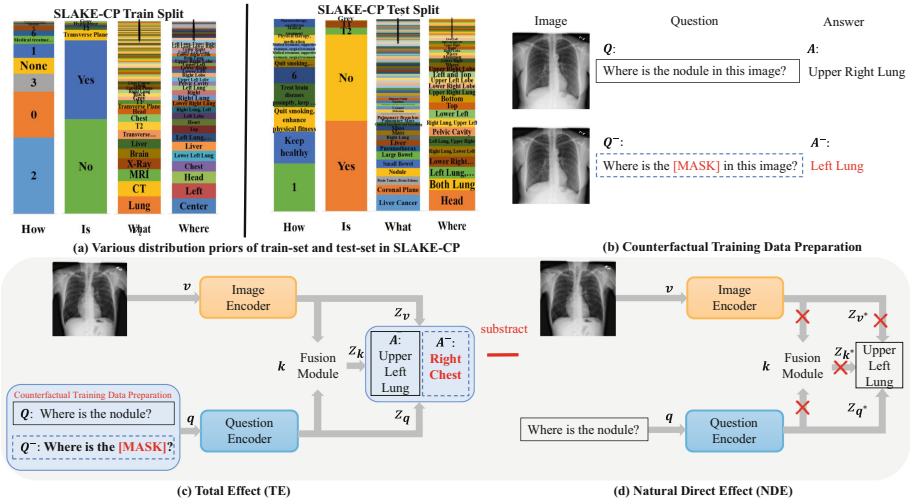

Fig. 2. (a) Various distribution priors of train-set and test-set in SLAKE-CP. (b) Counterfactual training data preparation. (c) Traditional Med-VQA causal graph with total effect. (d) Counterfactual Med-VQA causal graph with natural direct effect.

question on the generated answer and then subtract it from the total causal effect for counterfactual training. It is noted that both the original data and generated counterfactual data will be used for counterfactual training. In this way, the model may not tend to provide answers over-rely on the largest proportions of candidate answers in train-set when tested, thus concentrating on entanglement of the visual objects and language information.

Additionally, we conduct a bias-sensitive Med-VQA dataset Semantically-Labeled Knowledge-Enhanced-Changing Priors (SLAKE-CP) for evaluating the ability of disentangling the memorized linguist priors and semantic visual information. Qualitative and quantitative experimental results illustrate that our proposed model is superior to the state-of-the-art Med-VQA models on the two public benchmarks and can obtain more obvious improvements on the newly constructed SLAKE-CP.

2 Methodology

Figure 2 illustrates the proposed Med-VQA method which consists of implicit and explicit counterfactual debiased stages: the counterfactual training data preparation stage to improve the sensitivity of the critical clinic objects for implicit bias-weaken. Along with the counterfactual causal effect training stage to directly migrate the language priors.

2.1 Counterfactual Training Data Preparation

To implicitly weaken the language bias, we follow CSS [3] to prepare counterfactual training samples for improving the sensitivity of clinic objects. First, we extract the question type (e.g. "Where" in Fig. 2) of each question and calculate the importance s of the remaining words w_i in clinical question $\boldsymbol{q}$ to the label a as:

$$s(a, w_i) = S(P(a|\boldsymbol{q}, \boldsymbol{v}), w_i) := (\nabla_{w_i} P(a|\boldsymbol{q}, \boldsymbol{v}))^T \mathbf{1} \tag{1}$$

where $P(a|\boldsymbol{q}, \boldsymbol{v})$ represents the probability of predicting answer a through Med-VQA model with image $\boldsymbol{v}$ and question $\boldsymbol{q}$, ∇ is the gradient operator, S is the cosine similarity, and $\mathbf{1}$ is the all-ones vector. The top-K clinic words with the highest importance s are defined as critical words. Then, we construct counterfactual samples Q^- by replacing the critical words with "[MASK]". We also assign the Q^- with an answer A^-, and the detailed assigning procedure is as follows. We first generate the probability of predicting answer $P^+(a)$ with the question Q^+ which replaces the marginal words with "[MASK]" (all but the question type labels and the critical words), and then pick up top-N candidate answers with the highest probability as A^+. The rest answers are denoted as $A^- = \{a_i | a_i \in A, a_i \notin A^+\}$ and are assigned to Q^-.

2.2 Counterfactual Cause Effect Training Procedure

For explicitly subtracting the language priors, following [24], we introduce casual effect [21] to translate priors into quantified expressions. The causal effects can directly reflect the comparisons between the outputs with different treatments (e.g. $X = \boldsymbol{x}$ represents with-treatment and $X = \boldsymbol{x}^*$ represents the counterfactual situation where is without the treatment). The total effect (TE) of $X = \boldsymbol{x}$ on Y can be defined as two different conditions that with or without the input:

$$TE = Y_{X=\boldsymbol{x}, M(X=\boldsymbol{x})} - Y_{X=\boldsymbol{x}^*, M(X=\boldsymbol{x}^*)} \tag{2}$$

where M is the mediator between the variables X and Y. Note that the total effect can be composed of the natural direct effect (NDE) and total indirect effect(TIE). Between them, the NDE concentrates the exclusive effect of $X = \boldsymbol{x}$ on Y and prohibit the effect through M:

$$NDE = Y_{X=\boldsymbol{x}, M(X=\boldsymbol{x}^*)} - Y_{X=\boldsymbol{x}^*, M(X=\boldsymbol{x}^*)} \tag{3}$$

Thus TIE can reflect the reduction of language bias by subtracting the NDE from the TE:

$$TIE = TE - NDE = Y_{X=\boldsymbol{x}, M(X=\boldsymbol{x})} - Y_{X=\boldsymbol{x}, M(X=\boldsymbol{x}^*)} \tag{4}$$

Based on the above definition, we translate the Med-VQA task into a causal effect graph as Fig. 2 (c)(d) shown, aiming to directly formulate the language bias and subtract it. The answer set $A = \{a\}$ is caused by direct effect from medical image $V = \boldsymbol{v}$ and clinic question $Q = \boldsymbol{q}$, also the indirect effect of fusion

knowledge $K(Q = \boldsymbol{q}, V = \boldsymbol{v})$ through the cross-modal fusion module. We define the notations that: $Y_{\boldsymbol{q},\boldsymbol{v},\boldsymbol{k}} = Y(Q = \boldsymbol{q}, V = \boldsymbol{v}, K = \boldsymbol{k})$. Through subtracting NDE of $Q = \boldsymbol{q}$ on A from the TE of $V = \boldsymbol{v}$, $Q = \boldsymbol{q}$ and $K = \boldsymbol{k}$ on the answer, we can explicitly capture language bias and remove it via TIE, which is defined below. In the inference stage, we choose the answer with the maximum TIE as the prediction.

$$TIE = TE - NDE = Y_{\boldsymbol{q},\boldsymbol{v},\boldsymbol{k}} - Y_{\boldsymbol{q},\boldsymbol{v}^*,\boldsymbol{k}^*} \tag{5}$$

where $\boldsymbol{k}^* = K(V = \boldsymbol{v}^*, Q = \boldsymbol{q}^*)$, $\boldsymbol{v}^*$ and $\boldsymbol{q}^*$ is the counterfactual situation where model is without $\boldsymbol{v}$, $\boldsymbol{q}$ as inputs. The $Y_{\boldsymbol{q},\boldsymbol{v},\boldsymbol{k}} = \log \sigma(Z_{\boldsymbol{q}} + Z_{\boldsymbol{v}} + Z_{\boldsymbol{k}})$, where the $Z_{\boldsymbol{v}} = E_V(\boldsymbol{v})$, $Z_{\boldsymbol{q}} = E_Q(\boldsymbol{q})$, $Z_{\boldsymbol{k}} = E_F(\boldsymbol{q}, \boldsymbol{v})$ are calculated from the image encoder E_V, question encoders E_Q and the fusion module E_F respectively. The E_V, E_Q, E_F can be updated by L_{cls} [20]:

$$L_{cls}(\boldsymbol{q}, \boldsymbol{v}, a) = L_{VQA}(\boldsymbol{q}, \boldsymbol{v}, a) + L_{QA}(\boldsymbol{q}, \mathrm{a}) + L_{VA}(\boldsymbol{v}, a) \tag{6}$$

where L_{VQA}, L_{QA} and L_{VA} are corss-entropy losses over $Y_{\boldsymbol{q},\boldsymbol{v},\boldsymbol{k}}$, $Z_{\boldsymbol{q}}$ and $Z_{\boldsymbol{v}}$.

The complete objective of our method is optimized to minimize the L_{DeBCF} which combines the L_{cls} over both the original and the counterfactual data:

$$L_{DeBCF} = \alpha L_{cls}(V, Q, A) + (1 - \alpha) L_{cls}(V, Q^-, A^-) \tag{7}$$

where α is the hyperparameter which control the ratio of counterfactual samples.

3 SLAKE-CP: Construction and Analysis

For further evaluating the debiasing ability of Med-VQA, we follow VQA-CP [1] to create a bias-sensitive Med-VQA dataset which can be called SLAKE-CP. The SLAKE-CP can be further adopted by future debiased Med-VQA researches.

Grouping. We first construct all image-question-answer samples of train-set and test-set in SLAKE [16] into a whole set together. We start by labeling each question with a question type (first few words). If the samples have the same question type and answer, then these samples can be divided into same group.

Re-Splitting. We re-split the SLAKE [16] dataset to construct disparate distribution as Fig. 2 (a) shows. In detail, we first assign 1 group to the test-set. Among the remaining groups, if there is a group with a different question type or answer from the groups in test-set, this group will be assigned to test-set otherwise to train-set, aiming to vary the prior distributions of the train and test while remaining unchanged distributions of the images. The iteration stops when the test-set approximately reaches 1/7rd of the whole set, and the remaining are added to the train-set. We ensure the newly constructed test-set and train-set cover the majority of question types (*"Is", "What", "Where", "Which", etc.*) after these procedures. Most of the data attributes of SLAKE-CP are consistent with SLAKE, such as the train-test splitting, and open-close type splitting.

Table 1. The comparison results. $*$ indicates our re-implemented result, including the mean accuracy and standard deviation by 5 runs under 5 different seeds.

Methods	SLAKE			VQA-RAD		
	Open	Closed	All	Open	Closed	All
MFB [29]	72.2	75.0	73.3	14.5	74.3	50.6
SAN [28]	74.0	79.1	76.0	31.3	69.5	54.3
BAN [10]	74.6	79.1	76.3	37.4	72.1	58.3
LPF(*) [13]	$74.8_{\pm 1.4}\%$	$77.0_{\pm 1.1}\%$	$74.9_{\pm 1.3}\%$	$41.7_{\pm 1.3}\%$	$72.1_{\pm 1.1}\%$	$60.9_{\pm 1.3}\%$
RUBi(*) [2]	$75.1_{\pm 1.2}\%$	$77.6_{\pm 1.3}\%$	$75.8_{\pm 1.3}\%$	$42.4_{\pm 1.2}\%$	$73.2_{\pm 1.0}\%$	$61.5_{\pm 1.2}\%$
GGE(*) [8]	$76.4_{\pm 1.1}\%$	$78.7_{\pm 1.2}\%$	$76.6_{\pm 1.2}\%$	$44.6_{\pm 1.4}\%$	$74.5_{\pm 1.1}\%$	$63.8_{\pm 1.1}\%$
MEVF+SAN [19]	75.3	78.4	76.5	49.2	73.9	64.1
MEVF+BAN [19]	77.8	79.8	78.6	49.2	77.2	66.1
CLIPQCR(*) [6]	$78.2_{\pm 1.3}\%$	$82.6_{\pm 1.5}\%$	$80.1_{\pm 1.3}\%$	$58.0_{\pm 1.4}\%$	$79.6_{\pm 1.1}\%$	$71.1_{\pm 1.2}\%$
CPRD+BAN [15]	79.5	83.4	81.1	52.5	77.9	67.8
Ours	$\mathbf{80.8}_{\pm 0.9}\%$	$\mathbf{84.9}_{\pm 0.7}\%$	$\mathbf{82.6}_{\pm 0.9}\%$	$\mathbf{58.6}_{\pm 1.1}\%$	$\mathbf{80.9}_{\pm 0.8}\%$	$\mathbf{71.6}_{\pm 1.0}\%$

Table 2. The additional comparison of experimental results on the SLAKE-CP dataset.

Methods	SLAKE-CP		
	Open	Closed	All
MFB(*) [29]	$10.9_{\pm 1.0}\%$	$22.1_{\pm 0.8}\%$	$21.5_{\pm 0.8}\%$
SAN(*) [28]	$11.2_{\pm 1.2}\%$	$22.7_{\pm 1.1}\%$	$23.2_{\pm 1.1}\%$
BAN(*) [10]	$11.9_{\pm 1.2}\%$	$24.4_{\pm 0.9}\%$	$24.5_{\pm 1.0}\%$
RUBi(*) [2]	$12.2_{\pm 1.3}\%$	$26.9_{\pm 1.2}\%$	$26.4_{\pm 1.3}\%$
LPF(*) [13]	$13.1_{\pm 1.4}\%$	$29.7_{\pm 1.4}\%$	$29.2_{\pm 1.3}\%$
GGE(*) [8]	$13.9_{\pm 1.1}\%$	$30.9_{\pm 1.3}\%$	$30.2_{\pm 1.3}\%$
MEVF+SAN(*) [19]	$12.6_{\pm 1.1}\%$	$29.6_{\pm 1.0}\%$	$28.7_{\pm 1.0}\%$
MEVF+BAN(*) [19]	$13.0_{\pm 1.4}\%$	$29.8_{\pm 1.2}\%$	$29.1_{\pm 1.3}\%$
CLIPQCR(*) [6]	$13.4_{\pm 1.2}\%$	$30.5_{\pm 1.1}\%$	$30.0_{\pm 1.2}\%$
CPRD+BAN(*) [15]	$13.9_{\pm 1.3}\%$	$31.2_{\pm 1.5}\%$	$30.4_{\pm 1.5}\%$
Ours	$\mathbf{18.6}_{\pm 1.1}\%$	$\mathbf{35.4}_{\pm 1.0}\%$	$\mathbf{34.2}_{\pm 1.2}\%$

4 Experiments

4.1 Datasets and Implementation Details

Datasets. SLAKE [16] is a knowledge-augmented Med-VQA dataset, consisting of 642 images and 7033 question-answer samples. The VQA-RAD [12] is a manually annotated dataset validated by clinicians, which contains 315 radiographic images and 3,515 question-answer samples. We followed the original data partition, where questions are divided into closed-ended and open-ended types.

Implementation Details. For implementation, we apply Pytorch library with 6 NVIDIA TITAN 24 GB Xp GPUs. We employ the MEVF [19] as baseline.

Table 3. Ablation results. "CTD''": counterfactual training data preparation. "CCE''": counterfactual cause effect training procedure.

Index	CTD	CCE	SLAKE-CP			SLAKE		
			Open	Closed	Overall	Open	Closed	Overall
1	×	×	$13.0_{\pm 0.6}\%$	$29.8_{\pm 0.9}\%$	$29.1_{\pm 0.7}\%$	$78.6_{\pm 1.2}\%$	$80.5_{\pm 1.0}\%$	$79.8_{\pm 1.0}\%$
2	✓	×	$14.2_{\pm 1.3}\%$	$31.3_{\pm 0.9}\%$	$30.6_{\pm 1.0}\%$	$79.4_{\pm 1.3}\%$	$81.0_{\pm 1.1}\%$	$80.4_{\pm 1.1}\%$
3	×	✓	$16.7_{\pm 0.8}\%$	$33.9_{\pm 0.7}\%$	$32.9_{\pm 0.8}\%$	$80.1_{\pm 1.1}\%$	$81.9_{\pm 1.3}\%$	$81.5_{\pm 1.2}\%$
4	✓	✓	$\mathbf{18.6}_{\pm 1.1}\%$	$\mathbf{35.4}_{\pm 1.0}\%$	$\mathbf{34.2}_{\pm 1.2}\%$	$\mathbf{80.6}_{\pm 0.9}\%$	$\mathbf{84.4}_{\pm 0.7}\%$	$\mathbf{82.6}_{\pm 0.9}\%$

Table 4. Comparisons of different top-K critical words in Sect. 2.1.

top-K	SLAKE		
	Open	Closed	All
1	$\mathbf{80.8}_{\pm 1.1}\%$	$\mathbf{84.9}_{\pm 0.8}\%$	$\mathbf{82.6}_{\pm 1.0}\%$
2	$80.5_{\pm 1.0}\%$	$84.5_{\pm 0.8}\%$	$82.4_{\pm 0.9}\%$
3	$79.9_{\pm 1.2}\%$	$83.7_{\pm 1.1}\%$	$82.1_{\pm 1.2}\%$

Table 5. Evaluations of hyperparameter α.

α	SLAKE		
	Open	Closed	All
0.3	$80.1_{\pm 1.0}\%$	$83.2_{\pm 0.9}\%$	$81.9_{\pm 1.1}\%$
0.4	$80.2_{\pm 1.2}\%$	$83.5_{\pm 1.1}\%$	$82.0_{\pm 1.2}\%$
0.5	$80.7_{\pm 1.2}\%$	$84.6_{\pm 0.9}\%$	$82.5_{\pm 1.1}\%$
0.6	$\mathbf{80.8}_{\pm 1.1}\%$	$\mathbf{84.9}_{\pm 1.0}\%$	$\mathbf{82.6}_{\pm 1.0}\%$
0.7	$80.6_{\pm 1.3}\%$	$84.2_{\pm 1.0}\%$	$82.3_{\pm 1.0}\%$

The vision encoders are initialized by MAML [7] and CDAE [18], and LSTM is adopted as question encoder. The BAN [10] is adopted as the fusion module E_F. The medical images are resized into 224×224, and questions are cut to 12 words and then embed into 300 dimensions through Golve. The proposed model is trained for 200 epochs with 64 batch size and optimized with Adam whose learning rate is $1e^{-3}$. In Sect. 2.1, we choose top-1 candidate answer as A^+ and mask top-1 critical clinic word, the hyperparameter α is set to 0.6.

4.2 Experimental Results

Comparison with State-of-the-Art Methods. We compare our DeBCF with 10 state-of-the-art Med-VQA models on the SLAKE [16] and VQA-RAD [12] public benchmarks as the Table 1 shown. The proposed model obviously outperform the existing advanced models, attaining 82.6% and 71.6% mean accuracy respectively. Specifically, the results of our proposed model have prominent improvements over the attention-based models MFB [29], SAN [28], BAN [10]. Further, the improvements over MEVF+BAN [30] and CPRD+BAN [15] which adopt the same fusion model BAN [10] as ours are 4.0%, 1.5% overall accuracy on SLAKE respectively. In particular, the proposed model conspicuously improved the overall accuracy by 2.5% and compared with the advanced CLIPQCR [6]. Moreover, our model has significant superior with other debiasing models, including RUBi [2] and LPF [13], GGE [8]. Although these works can effectively reduce language bias, they reckon without visual-linguist explicable information and contrarily weaken the inference ability. For

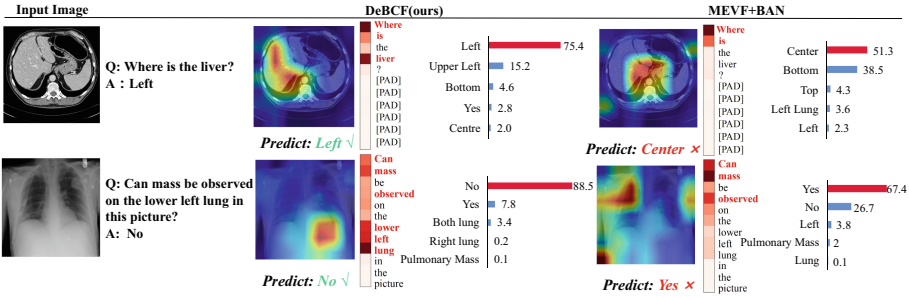

Fig. 3. Quantitative comparison analysis. The darker parts, the more contributions.

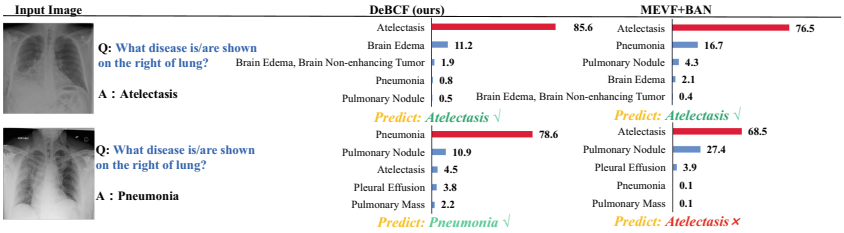

Fig. 4. The comparison analysis of sensitivity to the visual grounds.

ours, we explicitly subtract the language bias through causal effect and generate counterfactual samples to implicitly improve the sensitivity of clinical words and visual objects for inference.

Discussion of SLAKE-CP. Table 2 illustrates the superiority of the DeBCF on the newly constructed SLAKE-CP datasets which is the linguistic-bias sensitive evaluation. In particular, the DeBCF yields 34.2% mean overall accuracy on SLAKE-CP datasets. The performance of all the models has prominently dropped in the newly unbiased SLAKE-CP datasets compared with the SLAKE. It is obviously observed that the DeBCF significantly outperforms the baselines [10,19,28,29] and the debiasing methods [2,8,13]. The proposed model is also superior to the advanced models CLIPQCR [6] and CPRD+BAN [15], over-passing 4.2% and 3.8% overall accuracy. Within the bias-sensitive benchmarks, the comparisons demonstrate that our model may have the superiority to overcome the linguistic priors and force the model to generate more creditable answers rather than taking the superficial linguistic correlations as a shortcut.

Ablation Analysis. Table 3 demonstrates the ablation study which verifies the effectiveness of devised methods. We adopt MEVF+BAN [10] as the baseline in index 1. The baseline equipped with the counterfactual data preparation stage gains 1.5% and 0.6% overall accuracy on SLAKE-CP and SLAKE datasets. This

illustrates that masking critical clinical objects contributes to the implicit suppression of linguistic bias in Med-VQA. In addition, the comparison between index 3 and 1 demonstrates that subtracting the cause-effect of the question can explicitly weaken the prior and modifies the model to focus on intrinsically meaningful objects rather than superficial counterfactual correlations. Moreover, the model which combines the counterfactual masking samples into the counterfactual causal effect training procedure in index 4 obtains significant gains by up to 5.2% and 2.8% overall accuracy on SLAKE-CP and SLAKE, illustrating that we have built a robust unbiased Med-VQA model to overcome language priors.

Influence of Hyperparameters. The influence results of the top-K and the hyperparameter α are conducted in Table 4 and 5, which reveal that choosing top-1 critical words and $\alpha = 0.6$ achieves the best performance respectively. Crucially, masking top-1 critical clinic word can disentangle the linguistic bias and redundancy masking may result in interference.

Quantitative Analysis. As Fig. 3 shown, we conduct a quantitative comparison analysis to illustrate the ability to disengage the language prior to our proposed model through Grad-CAM maps [22]. For example 1 in row 1, the proposed DeBCF sensitively recognizes the precise critical keywords *"Where, is, liver"* and corresponding visual image objects to predict the correct answer with the highest probability score, while the advanced model MEVF+BAN [19] is subjected to the language prior that generate the wrong answer according to the superficial context *"Where is"* and ignore the reliable visual objects. Additionally, we also conduct the comparison of sensitivity to the visual grounds in Fig. 4. Given the same question but different medical images and answers, the proposed model correctly predict the various answers while the MEVF+BAN [19] fails. The detailed comparisons illustrate the debiased ability of the proposed model to overcome language priors and ingeniously grasp the critical parts (clinic keywords and visual objects) for a precise explanation.

5 Conclusion

In this paper, we propose a novel debiasing Med-VQA model that prepares the counterfactual data by masking critical clinic words and combines it into the counterfactual training stage which subtracting the causal effect of language priors directly, aiming to migrate the linguistic-bias in Med-VQA. Additionally, we construct a linguistic-bias sensitive Med-VQA dataset SLAKE-CP by disintegrating the language priors from training. Experimental results demonstrate the superior debiasing and interpretive performance of the proposed model. It's the first attempt to construct a preliminary bias-sensitive Med-VQA dataset, which will be elaborated in our future work. The codes will be released.

Acknowledgements. This work was supported in part by Zhejiang Provincial Natural Science Foundation of China (LDT23F02023F02).

References

1. Agrawal, A., Batra, D., Parikh, D., Kembhavi, A.: Don't just assume; look and answer: overcoming priors for visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4971–4980 (2018)
2. Cadene, R., et al.: RUBi: reducing unimodal biases for visual question answering. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
3. 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/CVF Conference on Computer Vision and Pattern Recognition, pp. 10800–10809 (2020)
4. Chen, Z., et al.: Multi-modal masked autoencoders for medical vision-and-language pre-training. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. Lecture Notes in Computer Science, vol. 13435. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16443-9_65
5. Do, T., Nguyen, B.X., Tjiputra, E., Tran, M., Tran, Q.D., Nguyen, A.: Multiple meta-model quantifying for medical visual question answering. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12905, pp. 64–74. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87240-3_7
6. Eslami, S., de Melo, G., Meinel, C.: Does clip benefit visual question answering in the medical domain as much as it does in the general domain? arXiv preprint: arXiv:2112.13906 (2021)
7. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 1126–1135. PMLR (2017)
8. Han, X., Wang, S., Su, C., Huang, Q., Tian, Q.: Greedy gradient ensemble for robust visual question answering. In: Proceedings of the IEEE/CVF International Conference on Computer vision, pp. 1584–1593 (2021)
9. Jing, C., Wu, Y., Zhang, X., Jia, Y., Wu, Q.: Overcoming language priors in VQA via decomposed linguistic representations. Proc. AAAI Conf. Artif. Intell. **34**(07), 11181–11188 (2020)
10. Kim, J.H., Jun, J., Zhang, B.T.: Bilinear attention networks. In: Advances in Neural Information Processing Systems, vol. 31 (2018)
11. KV, G., Mittal, A.: Reducing language biases in visual question answering with visually-grounded question encoder. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12358, pp. 18–34. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58601-0_2
12. Lau, J.J., Gayen, S., Ben Abacha, A., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. Sci. Data **5**, 180251 (2018). https://doi.org/10.1038/sdata.2018.251
13. 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, pp. 1955–1959 (2021)
14. Liang, Z., Jiang, W., Hu, H., Zhu, J.: Learning to contrast the counterfactual samples for robust visual question answering. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 3285–3292 (2020)

15. Liu, B., Zhan, L.-M., Wu, X.-M.: Contrastive pre-training and representation distillation for medical visual question answering based on radiology images. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12902, pp. 210–220. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87196-3_20
16. 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 (2021). https://doi.org/10.1109/ISBI48211.2021.9434010
17. Liu, B., Zhan, L.M., Xu, L., Wu, X.M.: Medical visual question answering via conditional reasoning and contrastive learning. IEEE Trans. Med. Imaging **42**, 1532–1545 (2022). https://doi.org/10.1109/TMI.2022.3232411
18. Masci, J., Meier, U., Cireşan, D., Schmidhuber, J.: Stacked convolutional auto-encoders for hierarchical feature extraction. In: Honkela, T., Duch, W., Girolami, M., Kaski, S. (eds.) ICANN 2011. LNCS, vol. 6791, pp. 52–59. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-21735-7_7
19. 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: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11767, pp. 522–530. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32251-9_57
20. Niu, Y., Tang, K., Zhang, H., Lu, Z., Hua, X.S., Wen, J.R.: Counterfactual VQA: a cause-effect look at language bias. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12700–12710 (2021)
21. Pearl, J.: Direct and indirect effects. In: Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence, 2001, pp. 411–420 (2001)
22. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626 (2017)
23. Selvaraju, R.R., et al.: Taking a hint: Leveraging explanations to make vision and language models more grounded. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2591–2600 (2019)
24. Tang, K., Niu, Y., Huang, J., Shi, J., Zhang, H.: Unbiased scene graph generation from biased training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3716–3725 (2020)
25. Tascon-Morales, S., Márquez-Neila, P., Sznitman, R.: Consistency-preserving visual question answering in medical imaging. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention - MICCAI 2022. Lecture Notes in Computer Science, vol. 13438, pp. 386–395. Springer Nature Switzerland, Cham (2022). https://doi.org/10.1007/978-3-031-16452-1_37
26. Teney, D., Abbasnedjad, E., van den Hengel, A.: Learning what makes a difference from counterfactual examples and gradient supervision. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12355, pp. 580–599. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58607-2_34
27. Wu, J., Mooney, R.: Self-critical reasoning for robust visual question answering. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 32 (2019)
28. 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, pp. 21–29 (2016)

29. 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 International Conference on Computer Vision, pp. 1821–1830 (2017)
30. 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)