# Eliminating Language Bias for Medical Visual Question Answering with Counterfactual Contrastive Training

Xingyu Wan[1], Qiaoying Teng[1], Jun Chen[1†], Yonghan Lu[1], Deqi Yuan[2] and Zhe Liu[1†]

[1] School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, China
{chenjun, 1000004088}@ujs.edu.cn
[2] Department of Radiology, Zhenjiang First People's Hospital Branch, Zhenjiang, China.

**Abstract.** Medical Visual Question Answering (Med-VQA) aims to assist in clinical diagnosis, but still faces challenges with language bias. Current approaches oversimplify the causal relationship between clinical terms and answers by treating it as a binary positive/negative effect. This can lead to the persistence of bias or reduced sensitivity to questions. To address this limitation, we propose a novel approach named DeCoCT (Debiasing Med-VQA via Counterfactual Contrastive Training). We decompose the causal relationship between clinical terms and answers into two components: (1) concept localization in medical images, and (2) prior knowledge from training data. We introduce a Key Region Capture Module (KRCM), trained with counterfactual strategies. It can enhance the model's ability to capture critical information through clinical terms. Furthermore, we employ counterfactual contrastive training to eliminate spurious correlations introduced by clinical terms while enhancing the model's focus on relevant visual regions. In addition, we construct a new conditional prior dataset based on VQA-RAD, named VQA-RAD-CP. Extensive experiments demonstrate that our approach significantly mitigates language bias in Med-VQA. Our codes and VQA-RAD-CP dataset are available at https://github.com/YX542/DeCoCT.

**Keywords:** Medical Vision Question Answering · Language Bias · Counterfactual Training · Contrastive Training

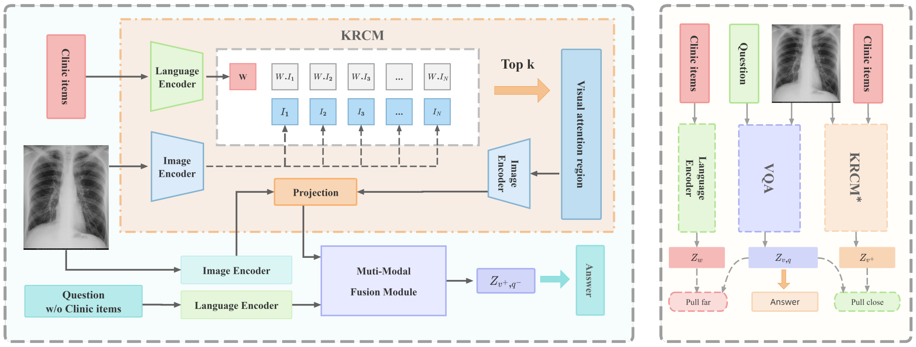
## 1 Introduction

Medical Visual Question Answering (Med-VQA) plays a crucial role in assisting with the diagnosis of medical imaging. Given a medical image and a clinical question associated with the image, it can provide a user-friendly answer. With increasing amounts of data in clinical practice, radiologists face substantial challenges in managing their workload [1]. Med-VQA can improve diagnostic efficiency, thus reducing both the misdiagnosis rate and labor costs [17]. Existing

Med-VQA still faces the challenges of language bias, which means that a model generates answers based on superficial connections between questions and answers, neglecting visual information. This problem is mainly caused by the unbalanced data distribution of the dataset and the subjective influence of human on the data processing process [27,12,13,19]. In clinical practice, language bias can weaken the generalization capabilities of models, potentially leading to misdiagnoses. Furthermore, it can also reduce the reliability of Med-VQA systems and hinder their further development [26].

However, current Med-VQA approaches [11,9,7,6,14] mainly focus on building robust models by introducing external knowledge, pre-training, etc. These works overlooked the salient issue of language bias. Consequently, this oversight leaves the issue unresolved, reducing the effectiveness of these systems. In general VQA, [2,23,5,16,10,25] and others have made progress in debiasing work. However, these methods do not consider the particularity of medical images. Fortunately, recent efforts have begun to address this gap. [27] solves the problem of language bias by training with counterfactual data and eliminating the causal effect of prior language knowledge. Likewise, [3] generates counterfactual samples using LRP, applying counterfactual causal reasoning to enhance interpretability. Despite these advances, these methods ignore a key factor in language bias, the connection between clinical terms and answers. For example, when asked "Is there a pleural effusion?", the clinical terms (pleural effusion) may directly affect the final answer distribution. "No" is the most common answer, and the model favors "no" due to this connection, even when presented with an image that suggests otherwise. The spurious correlations between clinical terms and answers is an aspect that needs to be eliminated. However, removing it directly could lead to a reduction in the question-sensitive ability of the model [5]. Given this complexity, treating the causal relationship between clinical terms and answers as a simple binary positive/negative effect is unreasonable. We propose that the main causal relationship between clinical terms and answers can be divided into concept localization in medical images and prior knowledge from training data.

In this paper, we introduce Debiasing Med-VQA via Counterfactual Contrastive Training (DeCoCT), a novel method designed to reduce language bias in Med-VQA models. This approach explores the relationship between clinical terms and regions of the medical image. This allows us to obtain a visual positive effect to complement the reduced ability to capture key information. Specifically, our approach consists of three parts: (i) We modify the question by masking the clinical terms and assign the corresponding images and answers. This allows us to generate counterfactual data that only modify the question. (ii) We introduce a Key Region Capture Module (KRCM), trained with counterfactual data. This enables the model to focus on regions related to clinical terms through counterfactual training. (iii) We employ counterfactual contrastive training to eliminate spurious correlations introduced by clinical terms, while enhancing the model's focus on relevant visual areas. In addition, to further assess bias in the Med-VQA model, we created a bias-sensitive version of VQA-RAD [13], called VQA-RAD-CP. We also assessed the performance of the latest methods on this dataset.



**(a) Counterfactual Training**

**(b) Contrastive Training**

**Fig. 1.** Overview of the DeCoCT method's two-stage training process for debiasing Med-VQA, consisting of (a) Counterfactual Training and (b) Contrastive Training. $KRCM^*$ indicates no use of a projection layer.

Extensive experiments show that DeCoCT can achieve a balance between avoiding the direct connection of clinical terms with answers and maintaining the model's question-sensitive ability. In particular, its state-of-the-art performance on SLAKE-CP [27] and VQA-RAD-CP demonstrates that our method has made significant advancements in bias mitigation.

## 2 Method

### 2.1 Counterfactual Data Preparation

To avoid ambiguity, we preprocess the data from datasets. First, we use ScispaCy [21] to recognize clinical terms in the questions. Then we normalize the terminology and unify the case. Finally, we submit the screened clinical terms to radiology experts for manual review. The expert review process is as follows: Two experts with sufficient work experience independently review ScispaCy's extraction results, and any disagreements (2.17%) are resolved by the third expert with extensive experience. Through the above operations, we construct a domain-specific lexicon. Based on this lexicon, we extract clinical terms from the questions and calculate the co-occurrence frequency of clinical terms and answers. Next, we compute the PMI (Pointwise Mutual Information) of these data as:

$$PMI(w, a) = \log \frac{P(w, a)}{P(w)P(a)} \tag{1}$$

where $w$ represents a clinical term and $a$ represents an answer. $P(w, a)$ is the joint probability of $w$ and $a$ occurring together. To avoid interference from common terms like (CT, patient and so on), we remove parts with PMI close to 0. As a result, we obtain pairs of clinical terms and answers with high relevance and separate the clinical items $W$ from them. In addition, we determine the most impactful clinical terms $w^*$ through a occlusion testing as:

$$w_a = \arg\max_i \left(P(y \mid x) - P(y \mid x_i)\right) \tag{2}$$

Where $P(y \mid x)$ represents the probability of the answer $y$ given the original input $x$. $x_i$ denotes the input $x$ with the $i$-th clinical term masked. We construct counterfactual samples $Q^-$ by replacing clinical terms $W$ with [MASK]. Then we assign the corresponding answer $A$ and image $V$ to $Q^-$ and $W$, forming the counterfactual data $(Q^-, W, I, A)$.

### 2.2 Counterfactual Data Training

To enhance the model's ability to capture key information, we train the model using counterfactual data $(Q^-, W, I, A)$. We introduce a key region capture module ($KRCM$), which consists mainly of a CLIP model and a visual encoder. The CLIP model is used to calculate the relevance between $W$ and each patch of $V$, using cosine similarity.

$$\text{sim}(\mathbf{W}, \mathbf{V}) = \left[\frac{\mathbf{W} \cdot \mathbf{v}_1}{\|\mathbf{W}\|\|\mathbf{v}_1\|}, \ldots, \frac{\mathbf{W} \cdot \mathbf{v}_N}{\|\mathbf{W}\|\|\mathbf{v}_N\|}\right] \tag{3}$$

where $v_j$ denotes the j-th patch in V, and N is the total number of patches. Then, the visual encoder merges and encodes patches with the highest relevance to obtain the feature of $V^*$.

$$\mathbf{V}^* = \{\mathbf{v}_i \mid i \in \text{TopK}\left(\text{sim}(\mathbf{W}, \mathbf{V}), k\right)\} \tag{4}$$

Next, a projection layer maps the features of $V^*$ and $V$ to the feature size of V, obtaining $V^+$. The feature fusion module fuses $V^+$ and $Q^-$ to obtain an answer distribution $A^*$. In the initial stages of training, we focus on the Key Region Capture Module $KRCM$ by freezing the other weights and performing partial training on this module. We define $P_{v^+,q^-} = \text{Softmax}(Z_{v^+,q^-})$ as the answer distribution, where $Z_{v^+,q^-}$ is obtained by the model when entering $V^+$ and $Q^-$. We use binary cross-entropy loss $L_{BCE}$ as training objective :

$$L_{\text{BCE}} = -\frac{1}{N} \sum_{i=1}^{N} \left[t_i \log \hat{y} + (1 - t_i) \log(1 - \hat{y})\right] \tag{5}$$

where $\hat{y_i}$ denotes the $P_{v^+,q^-}$ of the i-th sample. Through counterfactual training, we enable the model to utilize the visual information provided by $KRCM$ to answer questions in the absence of clinical terms.

### 2.3 Counterfactual Contrastive Training

To conduct comparative training, we designed a positive and negative sample selection strategy. We define $P_{v,q} = \text{Softmax}(Z_{v,q})$, where $Z_{v,q}$ is obtained by the model when entering $V$ and $Q$. We consider the output $P_{v,q}$ of the baseline network as the anchor output. Next, we further define $P_w = \text{Softmax}(Z_w)$ and $P_{v^*} = \text{Softmax}(Z_{v^*})$, where $Z_w$ and $Z_{v^*}$ are the outputs of the model when only entering $W$ and $V^*$, respectively. Similarly to [2,23,27], we consider the effect of $w$ on $A$ (pure language effect) as language bias. Therefore, our objective is to increase the distance between $P_w$ and $P_{v,q}$ to reduce the impact of language bias. Meanwhile, we regard the effect of $v^*$ on $A$ as the aid of regions of visual attention to answer questions. Thus, our objective is to decrease the distance between $P_{v^*}$ and $P_{v,q}$ to enhance the contribution of visual information. The model is updated through the loss function $\mathcal{L}$:

$$\mathcal{L} = D_{\mathrm{KL}}(P_{v,q} \parallel P_{v^*}) + \lambda \cdot \max(0, \gamma - D_{\mathrm{KL}}(P_{v,q} \parallel P_w)) \tag{6}$$

where $D_{\mathrm{KL}}$ denotes the KL divergence between the anchor distribution $P_{v,q}$ and the positive sample distribution $P_{v^*}$. $\lambda$ is the weight coefficient that balances the positive and negative sample terms. $\max(0, \gamma - D_{\mathrm{KL}}(P_{v,q} \parallel P_w))$ ensures that the KL divergence between the anchor distribution$P_{v,q}$ and the negative sample distribution $P_w$ is at least $\gamma$.

As shown in the Fig. 1, to avoid instability during the training process, we employed three branches for training. The weights related to the baseline network were used to initialize these branches. Through cross-model contrastive training, we can utilize $V^+$ and W to eliminate language bias brought by clinical terms.

## 3 Experiment

### 3.1 Construction of VQA-RAD-CP

Based on the design paradigm of VQA-CP [13] and SLAKE-CP [27], we construct a bias-sensitive Med-VQA dataset, called VQA-RAD-CP. First, we integrate the train set and test set of VQA-RAD into a set. Then we group the samples according to question type (determined by the prefix word of question) and answer label. If a sample group has a corresponding group with different question types or different answers in the test set, the group is completely divided into the test set. We break the consistency of the prior distribution between the train set and the test set by establishing a transfer pattern of the joint question-answer distribution. The partition was completed when the size of the test set reached 451 cases (aligned with the size of test set in VQA-RAD).

### 3.2 Datasets and Implementation Details

**Datasets** We validated our model on VQA-RAD [13] and SLAKE [19]. VQA-RAD contains 315 radiological images with 3,515 question-answer pairs, including 451 pairs reserved for testing. SLAKE comprises 14,028 samples, divided into 70% training subset, 15% validation subset, and 15% test subset. Additionally, we evaluated our model's bias mitigation performance on VQA-RAD-CP and SLAKE-CP [27] datasets to verify its debiasing effectiveness.

**Table 1.** Comparison results on VQA-RAD and SLAKE datasets, with partial data experimented using five different random seeds for the analysis.

| Methods | **VQA-RAD** | | | **SLAKE** | | |
|---|---|---|---|---|---|---|
| | Open | Closed | Overall | Open | Closed | Overall |
| MEVF+SAN[22] | 49.20 | 73.90 | 64.10 | 75.30 | 78.40 | 76.50 |
| MEVF+BAN[22] | 49.20 | 77.20 | 66.10 | 77.80 | 79.80 | 78.60 |
| CPRD+BAN[18] | 52.50 | 77.90 | 67.80 | 79.50 | 83.40 | 81.10 |
| M2I2[15] | 61.80 | 81.60 | 73.70 | 74.70 | **91.10** | 81.20 |
| M3AE[6] | 67.23 | 83.46 | 77.01 | 80.31 | 87.82 | 83.25 |
| MISS[4] | **71.81** | 80.35 | 76.05 | 81.47 | 82.91 | 82.00 |
| CCIS-MVQA[3] | 68.78 | 79.24 | 75.06 | 80.12 | 86.72 | 84.08 |
| LPF[16] | $41.7_{\pm 1.3}$ | $72.1_{\pm 1.1}$ | $60.9_{\pm 1.3}$ | $74.8_{\pm 1.4}$ | $77.0_{\pm 1.1}$ | $74.9_{\pm 1.3}$ |
| RUBI[2] | $42.4_{\pm 1.2}$ | $73.2_{\pm 1.0}$ | $61.5_{\pm 1.2}$ | $75.1_{\pm 1.2}$ | $77.6_{\pm 1.3}$ | $75.8_{\pm 1.3}$ |
| GGE[10] | $44.6_{\pm 1.4}$ | $74.5_{\pm 1.1}$ | $63.8_{\pm 1.1}$ | $76.4_{\pm 1.1}$ | $78.7_{\pm 1.2}$ | $76.6_{\pm 1.2}$ |
| CLIPQCR[8] | $58.0_{\pm 1.4}$ | $79.6_{\pm 1.1}$ | $71.1_{\pm 1.2}$ | $78.2_{\pm 1.3}$ | $82.6_{\pm 1.5}$ | $80.1_{\pm 1.3}$ |
| DeBCF[27] | $58.6_{\pm 1.1}$ | $80.9_{\pm 0.8}$ | $71.6_{\pm 1.0}$ | $80.8_{\pm 0.9}$ | $84.9_{\pm 0.7}$ | $82.6_{\pm 0.9}$ |
| **DeCoCT(Ours)** | $67.1_{\pm 0.5}$ | $\mathbf{85.7}_{\pm 0.4}$ | $\mathbf{78.3}_{\pm 0.5}$ | $\mathbf{82.5}_{\pm 0.3}$ | $87.0_{\pm 0.6}$ | $\mathbf{84.9}_{\pm 0.5}$ |

**Implementation Details** We adopt M3AE [6] as the baseline framework. Following the original implementation, the vision encoder utilizes CLIP-ViT-B [24] while the language encoder employs RoBERTa-base [20]. The multi-modal fusion module consists of a 6-layer Transformer. In KRCM, the clip model employs biomedclip [28], which has undergone sufficient pre-training. In our experiments, separate models with distinct weights were trained and tested on each dataset. The model is trained for 150 epochs with a batch size of 64 using AdamW optimization. Learning rates use 1e-5, whereas the fusion module adopts 5e-5. In Equation 6, the parameters $\lambda$ and $\gamma$ are set to 1 and 0.5 respectively. We chose 20% as the proportion of visual attention regions. We used five different random seeds for the experiments. The experiments were conducted on a server with 2 NVIDIA A6000 48GB GPUs.

### 3.3 Comparison with the State-of-the-arts

We compare DeCoCT with existing VQA models on the VQA-RAD and SLAKE benchmarks, as summarized in Table 1. Our model achieves significant advantages over state-of-the-art approaches, with average accuracies of 78.3% and 84.8% on the two datasets, respectively. Compared to the baseline M3AE, DeCoCT improves accuracy by 1.3% on VQA-RAD and 1.5% on SLAKE. Notably, DeCoCT outperforms existing debias methods (e.g., RUBi and DeBCF) by a clear margin. Further analysis on the conditional prior datasets (VQA-RAD-CP and SLAKE-CP) is presented in Table 2. DeCoCT achieves the best debiasing performance, attaining average accuracies of 61.9% and 49.2%, respectively. This represents 4.5% and 3.8% improvements over M3AE on VQA-RAD-CP and SLAKE-CP. The superiority of DeCoCT in bias reduction is particularly evident

**Table 2.** The additional comparison of experimental results on the VQA-RAD-CP and SLAKE-CP dataset. (∗) indicates that no pre-trained weights are used.

| Methods | VQA-RAD-CP | | | SLAKE-CP | | |
|---|---|---|---|---|---|---|
| | Open | Closed | Overall | Open | Closed | Overall |
| LPF[16] | $36.5_{\pm1.2}$ | $44.6_{\pm1.1}$ | $33.8_{\pm1.3}$ | $13.1_{\pm1.4}$ | $29.7_{\pm1.4}$ | $29.2_{\pm1.3}$ |
| RUBI[2] | $36.9_{\pm1.1}$ | $45.0_{\pm1.2}$ | $34.2_{\pm1.3}$ | $12.2_{\pm1.3}$ | $26.9_{\pm1.2}$ | $26.4_{\pm1.3}$ |
| GGE[10] | $38.3_{\pm1.2}$ | $45.2_{\pm1.1}$ | $35.4_{\pm1.3}$ | $13.9_{\pm1.1}$ | $30.9_{\pm1.3}$ | $30.2_{\pm1.3}$ |
| MEVF+SAN[22] | $38.5_{\pm1.2}$ | $40.8_{\pm1.4}$ | $35.6_{\pm1.4}$ | $12.6_{\pm1.1}$ | $29.6_{\pm1.0}$ | $28.7_{\pm1.0}$ |
| MEVF+BAN[22] | $39.7_{\pm1.1}$ | $43.9_{\pm1.2}$ | $36.7_{\pm1.3}$ | $13.0_{\pm1.4}$ | $29.8_{\pm1.2}$ | $29.1_{\pm1.3}$ |
| CLIPQCR[8] | $42.7_{\pm1.2}$ | $41.4_{\pm1.1}$ | $39.5_{\pm1.3}$ | $13.4_{\pm1.2}$ | $30.5_{\pm1.1}$ | $30.0_{\pm1.2}$ |
| CPRD+BAN[18] | $40.7_{\pm1.0}$ | $42.6_{\pm1.2}$ | $37.7_{\pm1.1}$ | $13.9_{\pm1.3}$ | $31.2_{\pm1.5}$ | $30.4_{\pm1.5}$ |
| DeBCF[27] | - | - | - | $18.6_{\pm1.1}$ | $35.4_{\pm1.0}$ | $34.2_{\pm1.2}$ |
| MISS(∗)[4] | $31.8_{\pm1.2}$ | $21.6_{\pm0.9}$ | $25.9_{\pm1.1}$ | $17.3_{\pm1.1}$ | $49.2_{\pm1.0}$ | $33.8_{\pm1.1}$ |
| M2I2(∗)[15] | $35.4_{\pm1.1}$ | $23.0_{\pm1.0}$ | $29.5_{\pm0.9}$ | $17.3_{\pm1.1}$ | $51.9_{\pm0.9}$ | $35.2_{\pm1.0}$ |
| M3AE[6] | $59.9_{\pm0.9}$ | $55.6_{\pm1.1}$ | $57.4_{\pm1.0}$ | $24.4_{\pm1.0}$ | $65.1_{\pm0.9}$ | $45.4_{\pm1.0}$ |
| **DeCoCT(Ours)** | $\mathbf{63.0}_{\pm0.5}$ | $\mathbf{61.0}_{\pm0.3}$ | $\mathbf{61.9}_{\pm0.4}$ | $\mathbf{27.3}_{\pm0.5}$ | $\mathbf{69.6}_{\pm0.5}$ | $\mathbf{49.2}_{\pm0.6}$ |

when compared to existing debiasing frameworks, highlighting its effectiveness in addressing spurious correlations in Med-VQA.

**Table 3.** Ablation results for the DeCoCT method: w/o $V^+$ indicates the performance when adjustments to the visual attention regions are not made; w/o $W$ indicates the performance when the influence of clinical terms is not removed.

| Methods | VQA-RAD-CP | | | SLAKE-CP | | |
|---|---|---|---|---|---|---|
| | Open | Closed | Overall | Open | Closed | Overall |
| Baseline | $59.9_{\pm0.9}$ | $55.6_{\pm1.1}$ | $57.4_{\pm1.0}$ | $24.4_{\pm1.0}$ | $65.1_{\pm0.9}$ | $45.4_{\pm1.0}$ |
| w/o $V^+$ | $60.4_{\pm0.7}$ | $56.8_{\pm0.9}$ | $58.3_{\pm0.8}$ | $25.4_{\pm0.9}$ | $65.6_{\pm1.0}$ | $46.2_{\pm1.0}$ |
| w/o $W$ | $61.5_{\pm1.0}$ | $58.3_{\pm0.9}$ | $59.7_{\pm0.9}$ | $25.2_{\pm1.0}$ | $66.9_{\pm1.0}$ | $46.8_{\pm1.1}$ |
| **DeCoCT** | $\mathbf{63.0}_{\pm0.5}$ | $\mathbf{61.0}_{\pm0.3}$ | $\mathbf{61.9}_{\pm0.4}$ | $\mathbf{27.3}_{\pm0.5}$ | $\mathbf{69.6}_{\pm0.5}$ | $\mathbf{49.2}_{\pm0.6}$ |

### 3.4 Ablation Analysis

Table 3 presents ablation studies validating the effectiveness of our designed approach. Both (w/o $V^+$) and (w/o $W$) showed improvements over the baseline, indicating that reducing direct associations with clinical objects and introducing visual attention regions both help mitigate language bias in Med-VQA. Notably, (w/o $W$) demonstrated superior performance compared to (w/o $V^+$). The results of DeCoCT reveal that combining both strategies can more effectively reduce bias.

**Table 4.** Comparison of DeCoCT performance with different selections of keyword quantities.

| Methods | VQA-RAD | | |
|---|---|---|---|
| | Open | Closed | Overall |
| Top1 | $65.9_{\pm 0.6}$ | $84.9_{\pm 0.5}$ | $77.4_{\pm 0.5}$ |
| All | $\mathbf{67.1}_{\pm 0.5}$ | $\mathbf{85.7}_{\pm 0.4}$ | $\mathbf{78.3}_{\pm 0.5}$ |

**Table 5.** Comparison of Removing Language Priors from Entire Questions Versus Isolating Removal to Clinical Terms.

| Methods | VQA-RAD | SLAKE |
|---|---|---|
| Question | $77.0_{\pm 0.3}$ | $83.7_{\pm 0.4}$ |
| Clinic terms | $\mathbf{78.3}_{\pm 0.5}$ | $\mathbf{84.9}_{\pm 0.5}$ |

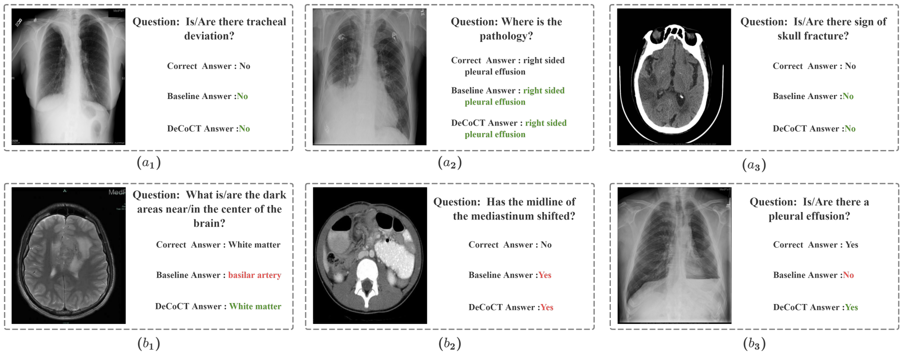

Question: Is/Are there tracheal deviation?
Correct Answer : No
Baseline Answer :No
DeCoCT Answer :No
(a1)
Question: Where is the pathology?
Correct Answer : right sided pleural effusion
Baseline Answer : right sided pleural effusion
DeCoCT Answer : right sided pleural effusion
(a2)
Question: Is/Are there sign of skull fracture?
Correct Answer : No
Baseline Answer :No
DeCoCT Answer :No
(a3)
Question: What is/are the dark areas near/in the center of the brain?
Correct Answer : White matter
Baseline Answer : basilar artery
DeCoCT Answer : White matter
(b1)
Question: Has the midline of the mediastinum shifted?
Correct Answer : No
Baseline Answer :Yes
DeCoCT Answer :Yes
(b2)
Question: Is/Are there a pleural effusion?
Correct Answer : Yes
Baseline Answer :No
DeCoCT Answer :Yes
(b3)

**Fig. 2.** Qualitative analysis of six VQA-RAD cases: Group A cases have strong prior associations between clinical items and answers, whereas Group B cases do not.

As shown in Table 4, selecting all clinical terms performs better than selecting only one of the most influential clinical terms. Normally, 1-3 clinical words are valid for a question. As shown in Table 5, we try to remove the language priors of the entire question from the baseline model. The results show that the effect of removing the whole language prior may be ineffective in the context of medical VQA datasets.

### 3.5 Qualitative Analysis

To further illustrate the effectiveness of our DeCoCT model in debiasing, we performed a qualitative analysis on six Med-VQA cases from VQA-RAD, as shown in Fig 2. The results for Group A indicate that, in scenarios with strong priors, both DeCoCT and the baseline can provide the correct answers. However, the results for Group B show that DeCoCT achieves superior performance in the absence of sufficient priors, while the baseline struggles to handle such scenarios effectively. These six cases collectively demonstrate that our method can effectively eliminate language bias in Med-VQA.

## 4 Conclusion

In this paper, we propose a new method (DeCoCT) to eliminate language bias in Med-VQA. We argue that the causal relationships between clinical terms and answers should be decomposed. To help the model focus more on image regions related to these clinical terms, we introduce the Key Region Capture Module (KRCM), which is trained with counterfactual strategies. Building on these strategies, DeCoCT achieves debiasing effects through counterfactual contrastive training. Experiments show that our method has made significant progress. Our current focus is on basic experiments on the VQA-RAD and Slake. In the future, we will conduct a more fine-grained analysis between questions and answers.

**Acknowledgements.** This work was partially supported by the National Natural Science Foundation of China (Grant 62276116), the Six Talent Peaks Project in Jiangsu Province (Grant DZXX-122), the Science and Technology Project of Zhenjiang City (Grant SH2024021), the Postgraduate Research & Practice Innovation Program of Jiangsu Province (Grant KYCX24_3982), and the China Postdoctoral Science Foundation (Grant 2024M751192).

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

## References

1. Aiello, M., Cavaliere, C., D'Albore, A., Salvatore, M.: The challenges of diagnostic imaging in the era of big data. Journal of clinical medicine **8**(3), 316 (2019)
2. Cadene, R., Dancette, C., Cord, M., Parikh, D., et al.: Rubi: Reducing unimodal biases for visual question answering. Advances in neural information processing systems **32** (2019)
3. Cai, L., Fang, H., Xu, N., Ren, B.: Counterfactual causal-effect intervention for interpretable medical visual question answering. Authorea Preprints (2024)
4. Chen, J., Yang, D., Jiang, Y., Lei, Y., Zhang, L.: Miss: A generative pre-training and fine-tuning approach for med-vqa. In: International Conference on Artificial Neural Networks. pp. 299–313. Springer (2024)
5. Chen, L., Zheng, Y., Niu, Y., Zhang, H., Xiao, J.: Counterfactual samples synthesizing and training for robust visual question answering. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(11), 13218–13234 (2023)
6. 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: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 679–689. Springer (2022)
7. Chen, Z., Li, G., Wan, X.: Align, reason and learn: Enhancing medical vision-and-language pre-training with knowledge. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 5152–5161 (2022)
8. 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)

9. Gu, T., Yang, K., Liu, D., Cai, W.: Lapa: Latent prompt assist model for medical visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4971–4980 (2024)
10. 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)
11. Huang, J., Chen, Y., Li, Y., Yang, Z., Gong, X., Wang, F.L., Xu, X., Liu, W.: Medical knowledge-based network for patient-oriented visual question answering. Information Processing & Management **60**(2), 103241 (2023)
12. Kiener, M.: Artificial intelligence in medicine and the disclosure of risks. AI & society **36**(3), 705–713 (2021)
13. 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)
14. Li, P., Liu, G., He, J., Zhao, Z., Zhong, S.: Masked vision and language pre-training with unimodal and multimodal contrastive losses for medical visual question answering. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 374–383. Springer (2023)
15. Li, P., Liu, G., Tan, L., Liao, J., Zhong, S.: Self-supervised vision-language pretraining for medial visual question answering. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2023)
16. 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)
17. Lin, Z., Zhang, D., Tao, Q., Shi, D., Haffari, G., Wu, Q., He, M., Ge, Z.: Medical visual question answering: A survey. Artificial Intelligence in Medicine **143**, 102611 (2023)
18. 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 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24. pp. 210–220. Springer (2021)
19. 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)
20. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
21. Neumann, M., King, D., Beltagy, I., Ammar, W.: ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In: Proceedings of the 18th BioNLP Workshop and Shared Task. pp. 319–327. Association for Computational Linguistics, Florence, Italy (Aug 2019). https://doi.org/10.18653/v1/W19-5034, https://www.aclweb.org/anthology/W19-5034
22. Nguyen, B.D., Do, T.T., Nguyen, B.X., Do, T., Tjiputra, E., Tran, Q.D.: Overcoming data limitation in medical visual question answering. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22. pp. 522–530. Springer (2019)

23. 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)
24. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
25. Vosoughi, A., Deng, S., Zhang, S., Tian, Y., Xu, C., Luo, J.: Cross modality bias in visual question answering: A causal view with possible worlds vqa. IEEE Transactions on Multimedia (2024)
26. Yuan, D.: Language bias in visual question answering: A survey and taxonomy. arXiv preprint arXiv:2111.08531 (2021)
27. 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: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 382–393. Springer (2023)
28. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)