Cause-Effect Driven Optimization for Robust Medical Visual Question Answering with Language Biases

Huanjia Zhu[1], **Yishu Liu**[2], **Xiaozhao Fang**[3], **Guangming Lu**[2] and **Bingzhi Chen**[1*]
[1]Beijing Institute of Technology, Zhuhai
[2]Harbin Institute of Technology, Shenzhen
[3]Guangdong University of Technology
bvyih3@gmail.com, liuyishu@stu.hit.edu.cn, xzhfang168@126.com, luguangm@hit.edu.cn, chenbingzhi@bit.edu.cn**

Abstract

Existing Medical Visual Question Answering (Med-VQA) models often suffer from language biases, where spurious correlations between question types and answer categories are inadvertently established. To address these issues, we propose a novel Cause-Effect Driven Optimization framework called CEDO, that incorporates three well-established mechanisms, i.e., Modality-driven Heterogeneous Optimization (MHO), Gradient-guided Modality Synergy (GMS), and Distribution-adapted Loss Rescaling (DLR), for comprehensively mitigating language biases from both causal and effectual perspectives. Specifically, MHO employs adaptive learning rates for specific modalities to achieve heterogeneous optimization, thus enhancing robust reasoning capabilities. Additionally, GMS leverages the Pareto optimization method to foster synergistic interactions between modalities and enforce gradient orthogonality to eliminate bias updates, thereby mitigating language biases from the effect side, i.e., shortcut bias. Furthermore, DLR is designed to assign adaptive weights to individual losses to ensure balanced learning across all answer categories, effectively alleviating language biases from the cause side, i.e., imbalance biases within datasets. Extensive experiments on multiple traditional and bias-sensitive benchmarks consistently demonstrate the robustness of CEDO over state-of-the-art competitors.

1 Introduction

Medical visual question answering (Med-VQA) has recently garnered significant attention [Liu *et al.*, 2021a; Liu *et al.*, 2022]. Med-VQA [Yang *et al.*, 2016; Yu *et al.*, 2017; Kim *et al.*, 2018; Chen *et al.*, 2022; Do *et al.*, 2021] aims to bridge the gap between medical images and corresponding clinical questions, enabling artificial intelligence systems to predict plausible answers. This automated diagnostic process offers significant advantages, including reduced time expenditure and lower costs, positioning it as a promising technology in the healthcare industry. However, the limited availability of medical datasets [Liu *et al.*, 2021b; Lau *et al.*, 2018] poses significant challenges for developing robust Med-VQA. Additionally, inherent biases within Med-VQA have been identified [Zhan *et al.*, 2023], stemming from the manual splitting and annotation of existing datasets. These biases lead models to learn spurious correlations between question types and answer categories, ignoring critical visual information. Hence, Med-VQA faces a formidable obstacle: *the need to develop robust reasoning capabilities despite the scarcity of medical data, while maintaining performance in out-of-distribution (OOD) training scenarios*.

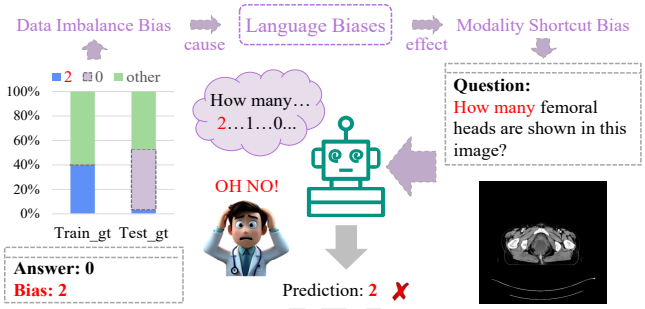


Figure 1: The causal mechanism underlying language biases involves imbalanced label distributions, which contribute to data bias and enable the model to form spurious correlations. For example, the question type "How many..." erroneously correlates with specific answers, such as "2". This leads to modality shortcut bias, where the model bypasses meaningful reasoning and instead relies on superficial patterns, ultimately generating counterfactual answers.

Previous research [Yu *et al.*, 2017; Yang *et al.*, 2016; Kim *et al.*, 2018; Chen *et al.*, 2022; Do *et al.*, 2021; Liu *et al.*, 2021a; Liu *et al.*, 2022; Ben Abacha *et al.*, 2019; Vu *et al.*, 2020] have adapted general VQA models for Med-VQA tasks. However, these studies often neglect the detrimental impact of language biases, which can lead to critical failures in clinical applications. While recent advances in general VQA [Basu *et al.*, 2023; Han *et al.*, 2021; Han *et al.*, 2023] have made strides in bias mitigation, such efforts remain underdeveloped in the medical domain. In Med-VQA, counterfactual learning [Zhan *et al.*, 2023] has emerged as a primary approach for bias mitigation, generat-

*Corresponding author: Bingzhi Chen.

ing counterfactual samples to reduce the model's dependence on language biases and refocus attention on the target information. However, this approach significantly disrupts the original data distribution, thereby compromising the robustness necessary for Med-VQA.

Addressing language bias requires a comprehensive analysis of its causes and consequences. As illustrated in Figure. 1, one major cause is **data imbalance**: frequent answer categories receive disproportionate emphasis during training, leading to overexpansion of their feature space. This amplifies spurious correlations between certain question types and answers. During training, these spurious correlations are encoded into the network through gradients. Therefore, the model disproportionately relies on the question type for predictions, neglecting critical image features. The question modality gets large gradient updates [Guo *et al.*, 2021], thus becoming a shortcut for generating answers. This phenomenon, driven by language bias, is referred to as **modality shortcut**. By identifying these challenges, this work strives to propose a novel solution to the challenges posed by inherent biases, improving the robustness of Med-VQA models.

To address these challenges, this paper proposes a novel **C**ause-**E**ffect **D**riven **O**ptimization (**CEDO**) framework, which aims to alleviate language biases from their cause and effect. Specifically, the proposed CEDO framework innovatively incorporates three mechanisms: **M**odality-driven **H**eterogeneous **O**ptimization (**MHO**), **G**radient-guided **M**odality **S**ynergy (**GMS**), and **D**istribution-adapted **L**oss **R**escaling (**DLR**). On the one hand, MHO and GMS are seamlessly integrated to address modality shortcut bias. The primary purpose of MHO is to achieve adaptive optimization for different modalities by adjusting their learning rates. Leveraging the multi-learning rate strategy, it strengthens weaker modalities while preventing dominant modalities from monopolizing the prediction process, thereby mitigating the risk of a single modality becoming a shortcut. Meanwhile, GMS fosters coordinated optimization between modalities through the Pareto method and gradient orthogonality. The Pareto method identifies a steep gradient direction that benefits all objectives (optimizing each modality), allowing the system to converge to a balanced trade-off state. Gradient orthogonality removes gradient conflicts, thereby preventing excessive updates to any single modality and preserving the reasoning capability of other modalities. On the other hand, DLR mitigates data imbalance bias. It ensures balanced learning across all ground-truth answers by adjusting the loss magnitude for each answer category. Based on statistical data distributions, DLR assigns adaptive weights to rescale individual losses, preventing the model from overemphasizing frequent answer categories at the expense of rare ones. Our main contributions are summarized as follows:

- Our work systematically addresses language biases by targeting their cause and effect. Three innovative mechanisms, i.e., MHO, GMS, and DLR, are proposed to comprehensively address shortcut biases originating from modalities and imbalance biases within datasets.
- MHO and GMS cooperate to reduce shortcut biases. MHO enables modality adaptive training, while GMS facilitates coordinated updates between modalities to ensure balanced optimization.
- DLR leverages a dynamic loss rescaling strategy to counteract dataset imbalance. By assigning adaptive weights to individual loss, DLR ensures equitable attention across categories.
- Two bias-sensitive Med-VQA datasets, SLAKE-CP and VQA-RAD-CP, are built to evaluate the debiasing performance. Extensive experiments on five datasets demonstrate the effectiveness and generalization of our CEDO method, achieving state-of-the-art performance.

2 Related Work

2.1 Visual Question Answering

Medical VQA. Existing Med-VQA research [Yu *et al.*, 2017; Yang *et al.*, 2016; Kim *et al.*, 2018; Chen *et al.*, 2022; Do *et al.*, 2021; Liu *et al.*, 2021a; Liu *et al.*, 2022; Ben Abacha *et al.*, 2019] typically applies prevalent VQA models to the Med-VQA task, with a primary focus on introducing multimodal feature fusion modules. However, due to the scarcity of medical data, these direct adaptations are often hindered by severe overfitting. To mitigate this challenge, Nguyen et al. [Nguyen *et al.*, 2019] propose a hybrid enhanced visual feature (MEVF). Despite these efforts, many Med-VQA datasets [Liu *et al.*, 2021b; Lau *et al.*, 2018] attempt to balance medical images to alleviate inherent bias dependencies, yet they fail to address the detrimental effects of *OOD* data, thus falling into the trap of language biases.

Robust Medical VQA. While research on Med-VQA debiasing has only recently gained traction, studies on bias mitigation in general VQA have flourished. Recent robust VQA approaches focus on introducing bias models [Han *et al.*, 2021; Han *et al.*, 2023] or problem-specific branches [Cadene *et al.*, 2019; Clark *et al.*, 2019] to learn and mitigate the inherent biases in modalities or datasets. However, due to the scarce medical data, these methods often demonstrate suboptimal generalization performance in the medical domain. As a first attempt, DeBCF [Zhan *et al.*, 2023] specifically targeted this issue and constructed a bias-sensitive dataset to assess debiasing performance. DeBCF employs pre-generated counterfactual samples and counterfactual causal effects to mitigate bias. However, the additional samples cause a distribution shift in the original data, thereby undermining the goal of achieving robust Med-VQA.

2.2 Ensemble-Based Methods

Addressing the elusive and multifaceted nature of bias in VQA tasks, prior research has introduced a question-only model to explicitly capture inherent biases in the dataset and subsequently subtract them from the base VQA model's predictions [Cadene *et al.*, 2019; Clark *et al.*, 2019; Ramakrishnan *et al.*, 2018; Han *et al.*, 2021; Han *et al.*, 2023]. These approaches isolate the question modality to identify and eliminate shortcuts learned by the model. In addition to these explicit bias removal techniques, data re-weighting methods have emerged as a promising solution. These methods adjust the contribution of individual classes by assigning different

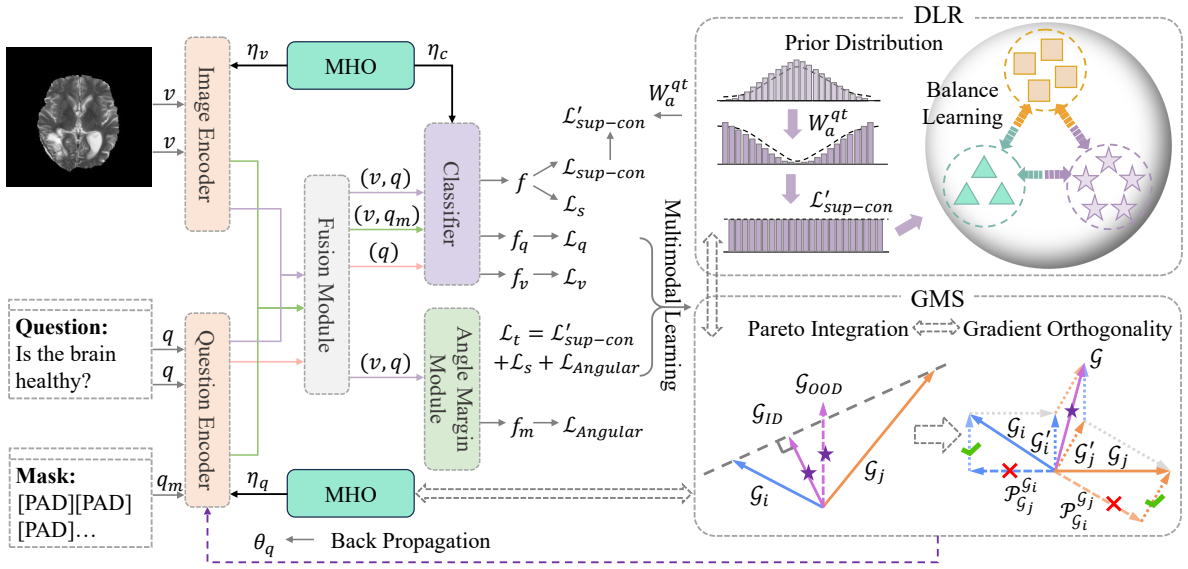


Figure 2: Illustration of the proposed Cause-Effect Driven Optimization (CEDO) framework for addressing medical language biases. Three adaptive mechanisms, i.e., Modality-driven Heterogeneous Optimization (MHO), Gradient-guided Modality Synergy (GMS), and Distribution-adapted Loss Rescaling (DLR), are dexterously established and synergistically integrated to mitigate shortcut bias and imbalance bias from the causal perspective of language biases.

weights during loss calculation. A widely used approach is to prioritize tail classes (i.e., less frequent categories) by assigning them higher weights, while head classes (i.e., more frequent categories) receive lower weights, ensuring balanced learning across the dataset. For example, Focal Loss [Ross and Dollár, 2017] introduces a dynamic weighting mechanism that emphasizes hard-to-classify and misclassified samples, making it particularly effective in imbalanced datasets. Furthermore, Guo et al. [Guo *et al.*, 2021] propose a loss rescaling strategy tailored to the dataset's distribution, further mitigating the overemphasis on dominant classes.

3 Methodology

3.1 Preliminaries

The Med-VQA task can be viewed as a multi-label classification problem. Without loss of generality, given a batch of data samples $\mathcal{D}$ having N samples, each consisting of an image $v \in \mathcal{V}$, a question $q \in \mathcal{Q}$ and a ground-truth answer $a \in \mathcal{A}$, the goal of the Med-VQA model is to optimize a mapping function $f : \mathcal{V} \times \mathcal{Q} \to \mathbb{R}^{|\mathcal{A}|}$ to predict the answer corresponding to a given image-question pair. The base Med-VQA model embeds the two features to obtain a joint representation and uses an answer classifier c to generate the logits f. Thus the problem can be formulated as follows:

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

where e_v and e_q are the image encoder and question encoder, respectively, and g is a multi-modal feature fusion network. The Med-VQA model can be trained by minimizing the Cross-Entropy (*CE*) loss function:

$$\mathcal{L}_{CE} = \sum_{i=1}^{|A|} -a_i \log \frac{\exp(f_i)}{\sum_{j=1}^{|A|} \exp(f_j)}. \tag{2}$$

In this paper, we integrate [Basu *et al.*, 2023] as our base model, and the total loss is denoted as $\mathcal{L}_t$.

3.2 Modality-driven Heterogeneous Optimization

In Med-VQA, different modalities exhibit distinct complexities and learning requirements. The question modality is prone to biases [Guo *et al.*, 2021], requiring a lower learning rate to avoid overfitting, while the image modality, especially in medical contexts, demands a higher learning rate to capture complex features [Khan *et al.*, 2022; Liu *et al.*, 2021c; Tajbakhsh *et al.*, 2020; Hesamian *et al.*, 2019]. Besides, the classifier tends to be insufficiently trained [Guo *et al.*, 2021]. Traditional single learning rate strategies fail to address these differences, leading to over-reliance on simpler modalities and underutilization of complex ones. To tackle this, we propose MHO, which assigns tailored learning rates to each modality, ensuring synchronized and effective learning aligned with their inherent complexities.

Given a multimodal model F, its parameters θ are grouped into three distinct sets corresponding to the modalities:

$$\theta = \{\theta_q, \theta_v, \theta_c\}, \tag{3}$$

where $\theta_q, \theta_v, \theta_c$ represent the parameters of the question modality, the image modality, and the classifier, respectively. To perform heterogeneous optimization, we define a learning rate η_k for each modality $k \in \{q, v, c\}$. The total parameter update for a training step is expressed as:

$$\theta_k \leftarrow \theta_k - \eta_k \nabla_{\theta_k} L, \quad \forall k \in \{q, v, c\}, \tag{4}$$

where L denotes the loss, and η_q, η_v, η_c are the hyperparameters. The MHO mechanism enhances convergence by addressing the inherent differences in modality complexity. By assigning learning rates tailored to each modality, the algorithm prevents overfitting to modality biases (e.g., question

shortcuts) and encourages robust learning from underrepresented modalities (e.g., medical images).

3.3 Gradient-guided Modality Synergy

A robust Med-VQA model is expected to possess well-trained question and image modalities with aligned optimization directions. However, the pervasive issue of modality shortcut bias has been extensively explored [Cadene *et al.*, 2019; Han *et al.*, 2021; Han *et al.*, 2023], where dominant modalities often exhibit disproportionately larger gradient norms [Guo *et al.*, 2021], leading to imbalanced learning dynamics. Inspired by [Sener and Koltun, 2018; Wang *et al.*, 2024], we propose the GMS module that applies the Pareto method and gradient orthogonality to promote synchronized updates between modalities, effectively mitigating the shortcut bias.

Multimodal Learning. The Med-VQA task can be viewed as multimodal learning, where models are expected to produce correct predictions by integrating information from multiple modalities. However, only utilizing such joint loss to optimize all modalities together could result in the optimization process being dominated by one modality, leaving others being severely under-optimized [Peng *et al.*, 2022; Huang *et al.*, 2022]. To overcome this imbalanced multimodal learning problem, introducing unimodal loss, which targets the optimization of each modality, is widely used and verified effective for alleviating this imbalanced multimodal learning problem [Wang *et al.*, 2020]. Therefore, we introduce a question branch and an image branch using CE loss, denoted as $\mathcal{L}_q$ and $\mathcal{L}_v$, respectively:

$$f_q(q) = c(g(e_q(q))), \tag{5}$$

$$\mathcal{L}_q = \sum_{i=1}^{|A|} -a_i \log \frac{\exp(f_{q,i})}{\sum_{j=1}^{|A|} \exp(f_{q,j})}, \tag{6}$$

where f_q denotes the logits in question branch. f_v and $\mathcal{L}_v$ are computed in the similar way. Thus, the final loss function is:

$$L = L_t + L_q + L_v \tag{7}$$

Pareto Integration. In multimodal systems, the relationships between multimodal loss and unimodal losses are intricate, as they are highly interdependent but often exhibit gradient conflicts. These conflicts stem from the fact that optimizing one modality's loss might detract from the performance of another, especially when modalities vary in representational strength. The gradients of these losses can be expressed as:

$$\mathcal{G}_k = \nabla_{\theta_q} \mathcal{L}_k(R_k, A), \quad \forall k \in \{t, q, v\}, \tag{8}$$

where R_t represents the joint features, R_q and R_v correspond to the unimodal features. Resolving how to integrate $\mathcal{G}_k$ effectively without introducing additional conflicts is critical.

The Pareto method, widely utilized in multi-task learning, provides a principled approach for managing these competing gradients [Sener and Koltun, 2018]. It adaptively assigns weights to gradients at each iteration and combines them into a single gradient vector. This ensures that the optimization direction benefits all objectives, facilitating convergence to a Pareto-optimal state. At Pareto-optimality, further improving any single objective is impossible without sacrificing the performance of others. Integrating Pareto optimization into multimodal learning frameworks naturally aligns with the need to reconcile multimodal and unimodal gradients. The corresponding optimization problem can be formalized as:

$$\begin{aligned} &\min_{\alpha_t, \alpha_q, \alpha_v \in \mathcal{R}} \left\| \alpha_t \mathcal{G}_t + \alpha_q \mathcal{G}_q + \alpha_v \mathcal{G}_v \right\|^2 \\ &\text{s.t.} \quad \alpha_t, \alpha_q, \alpha_v \geq 0, \alpha_t + \alpha_q + \alpha_v = 1, \end{aligned} \tag{9}$$

where $\|\cdot\|$ denotes the L_2-norm. This formulation seeks to minimize the norm of the weighted gradient combination within the convex hull of the gradient family $\{\mathcal{G}_k\}_{k \in \{t,q,v\}}$. The theoretical properties of this optimization are particularly compelling. As shown in [Désidéri, 2012], there are two key outcomes: (1) if the minimum norm equals zero, the parameters are Pareto-stationary, satisfying a necessary condition for Pareto-optimality; (2) otherwise, the solution identifies a descent direction that is beneficial for all learning objectives.

Gradient Orthogonality. Under an ideal unbiased setting, the Pareto method identifies the optimal gradient direction for each modality, ensuring efficient and balanced updates. However, the inherent bias present in *OOD* datasets significantly influences the model's learning process, increasing the likelihood of the Pareto method conforming to this bias rather than mitigating it. To address this limitation and suppress shortcut biases, we introduce gradient orthogonality as a corrective mechanism. Taking $\mathcal{G}_q$ and $\mathcal{G}_v$ for instance, the cosine similarity $\mathcal{S}$ between two gradients can be written as:

$$\mathcal{S}_{qv} = \frac{\mathcal{G}_q \cdot \mathcal{G}_v}{\|\mathcal{G}_q\| \|\mathcal{G}_v\|}. \tag{10}$$

If $S_{qv} > 0$, it indicates that the two modalities are well-aligned during training, with optimization progressing without bias. Conversely, a negative $\mathcal{S}_{qv}$ reveals that $\mathcal{G}_q$ and $\mathcal{G}_v$ are oriented in opposing optimization directions. In such cases, the modality with a larger gradient norm tends to dominate the training process, leading to shortcut biases. To this end, we employ gradient orthogonality to recalibrate the optimization directions, ensuring unbiased learning dynamics. The projection of $\mathcal{G}_v$ onto $\mathcal{G}_q$ can be denoted as $\mathcal{P}_{\mathcal{G}_q}^{\mathcal{G}_v}$:

$$\mathcal{P}_{\mathcal{G}_q}^{\mathcal{G}_v} = \left(\frac{\mathcal{G}_v \cdot \mathcal{G}_q}{\|\mathcal{G}_q\|^2}\right) \mathcal{G}_q. \tag{11}$$

Therefore, the biased gradient norms in the question modality can be dislodged:

$$\mathcal{G}_q^{'} = \mathcal{G}_q - \mathcal{P}_{\mathcal{G}_q}^{\mathcal{G}_v}. \tag{12}$$

The gradient norms $\mathcal{G}_v^{'}$ and $\mathcal{G}_t^{'}$ are obtained in the same manner. In the context of the general case, the ultimate gradient $\mathcal{G}$ can be represented as follows:

$$\mathcal{G} = \mathcal{G}_q^{'} + \mathcal{G}_v^{'} + \mathcal{G}_t^{'}. \tag{13}$$

This approach ensures effective optimization of the modalities, mitigating shortcut bias arising from the question modality. Notably, the method can also be extended to the image modality or fusion modules, though it necessitates careful consideration of the trade-off between computational resource requirements and potential performance gains.

3.4 Distribution-adapted Loss Rescaling

Data imbalance is a primary factor contributing to language biases. Frequent answer categories disproportionately impact the loss function, leading to excessive focus and spurious correlations. To address this issue, inspired by [Guo *et al.*, 2021], we propose an interpretable weighting mechanism that dynamically assigns category-specific weights based on the distribution of question types, ensuring balanced learning across all answer categories. The weight w_i^j is obtained via:

$$w_i^j = \frac{1}{\mathcal{M}_j \times m_i^j}, \tag{14}$$

where $\mathcal{M}_j$ is the number of samples under question type qt_j, and m_i^j is the number of answer a_i under qt_j. To achieve a finer balance between rare and common answers, we incorporated the softplus function into the weighting process:

$$\mathcal{W}_i^j = \log(1 + \exp w_i^j). \tag{15}$$

The softplus function not only smooths extreme weight values but also enhances the stability of the training process. This weighting mechanism is specifically applied to the supervised contrastive loss [Khosla *et al.*, 2020] in [Basu *et al.*, 2023], addressing the adverse effects of imbalanced data on feature spaces in contrastive learning:

$$\mathcal{L}'_{sup-con} = \sum_{i \in I} \frac{-1}{|P_i|} \sum_{p \in P_i} \mathcal{W}_i^j \log \frac{\exp(x_i^T x_p / \tau)}{\sum_{n \in N_i} \exp(x_i^T x_n / \tau)}, \tag{16}$$

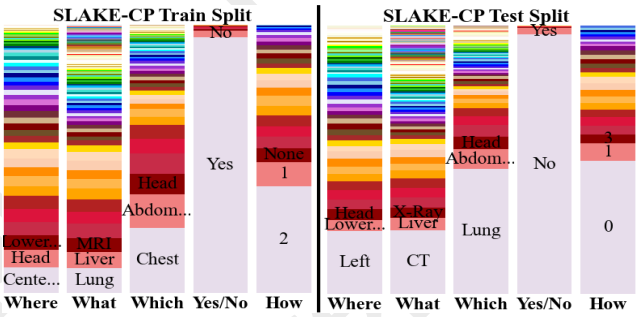
where i is the index of the current sample in a mini-batch of size I of fused features denoted as $\{x_1, x_2, ..., x_I\}$. a_i and qt_j are the ground truth and question type of the sample x_i, respectively. The set of positive examples in the mini-batch is represented as $P_i : \{p \in I \; s.t. \; a_p = a_j\}$, and the set of negative examples is denoted by $N_i : \{n \in I \; s.t. \; a_n \neq a_i\}$. The temperature τ is set to 1 following [Basu *et al.*, 2023].

The proposed mechanism tackles data imbalance by ensuring balanced contributions across all answer categories, effectively mitigating distributional shifts and improving robustness in *OOD* scenarios. Additionally, this mechanism maintains training stability through moderate weight adjustments while remaining computationally efficient, enabling seamless integration into clinical workflows and widespread adoption without disrupting operational efficiency.

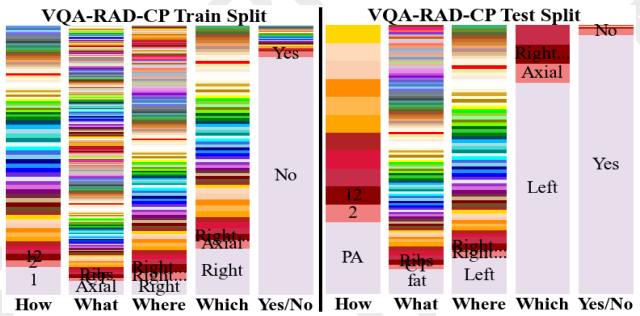
4 Experiments

4.1 Datasets

We evaluate our approach on two classical Med-VQA datasets, SLAKE [Liu *et al.*, 2021b] and VQA-RAD [Lau *et al.*, 2018], along with two *OOD* constructed medical benchmark evaluation protocols, SLAKE-CP and VQA-RAD-CP. To address the scarcity of medical data, we further test our method on a large-scale *OOD* natural benchmark, VQA-CE [Dancette *et al.*, 2021], to verify the scalability and generalization of the proposed approach. All experiments follow the standard VQA evaluation metric [Antol *et al.*, 2015]. Implementation details can be found in the supplementary material.



(a) SLAKE-CP



(b) VQA-RAD-CP

Figure 3: (a) Designed distribution bias of training set and testing set for all question types in SLAKE-CP. (b) Constructed distribution biases of the training set and the testing set in VQA-RAD-CP.

4.2 Bias Reconstruction

Following [Zhan *et al.*, 2023], we propose a novel method to construct two biased Med-VQA datasets, SLAKE-CP and VQA-RAD-CP, designed to serve as valuable benchmarks in future research. Similar to [Zhan *et al.*, 2023], the training and test sets were first merged. Each question was labeled based on its question type, determined by its initial words. Samples with binary answers ("Yes" or "No") were categorized as "Yes/No". Finally, samples sharing the same question type and answer were grouped into distinct clusters. As illustrated in Figure. 3, the samples are redistributed to introduce controlled biases. For each question type, the most frequent answers are allocated to the training and test sets in a 39:1 ratio, while the second most frequent answers are split in a 1:39 ratio. The remaining samples are divided at a 3:1 ratio to ensure the training set contains twice as many samples as the test set, maintaining consistency with the VQA-CP v2 [Agrawal *et al.*, 2018] ratio.

4.3 Comparisons with State-of-the-art

Evaluation on Medial Language Biases Benchmark. Table 1 demonstrates the superiority of the CEDO framework on the SLAKE-CP and VQA-RAD-CP datasets, specifically designed to evaluate sensitivity to language biases in the medical domain. Key observations include: 1) Most methods exhibit a substantial performance decline on the biased SLAKE-CP and VQA-RAD-CP datasets compared with SLAKE and VQA-RAD. 2) The proposed model outperforms all the state-of-the-art approaches, achieving improvements

Approaches	Methods	Reference	SLAKE-CP			VQA-RAD-CP		
			All	Open	Closed	All	Open	Closed
Classical	SAN [Yang *et al.*, 2016]	CVPR'16	26.02	48.30	6.42	16.29	59.73	6.05
	MFB [Yu *et al.*, 2017]	ICCV'17	30.56	55.70	8.44	22.53	72.12	10.84
	BAN [Kim *et al.*, 2018]	NIPS'18	17.50	30.90	5.72	17.30	67.70	5.42
	UpDn [Basu *et al.*, 2023]	CVPR'18	31.45	59.90	6.42	26.67	**74.78**	15.33
Med-Debias	MEVF+SAN [Nguyen *et al.*, 2019]	MICCAI'19	18.62	32.60	6.33	22.11	68.14	11.26
	MEVF+BAN [Nguyen *et al.*, 2019]	MICCAI'19	19.33	35.00	5.54	19.07	62.39	8.86
Natural-Debias	RUBi [Cadene *et al.*, 2019]	NIPS'19	33.88	60.30	10.64	81.27	60.62	86.13
	LPF [Liang *et al.*, 2021]	SIGIR'21	40.34	43.70	37.38	41.52	65.04	35.97
	GGE-iter [Han *et al.*, 2021]	ICCV'21	35.05	61.30	11.96	21.60	51.33	14.60
	RMLVQA [Basu *et al.*, 2023]	CVPR'23	76.42	60.50	**90.41**	89.45	69.03	94.26
Ours	CEDO	—	**79.27**	**66.70**	90.33	**92.07**	73.45	**96.45**
		Increased ↑	2.85	6.20	-0.08	2.62	-1.33	2.19

Table 1: Comparisons with state-of-the-art methods are conducted on the SLAKE-CP and VQA-RAD-CP datasets.

Approaches	Methods	Reference	SLAKE			VQA-RAD		
			All	Open	Closed	All	Open	Closed
Classical	SAN [Yang *et al.*, 2016]	CVPR'16	76.00	74.00	79.10	52.89	31.64	65.50
	MFB [Yu *et al.*, 2017]	ICCV'17	73.89	71.63	77.40	54.10	41.90	62.13
	BAN [Kim *et al.*, 2018]	NIPS'18	76.25	75.97	76.68	55.43	48.60	59.93
	UpDn [Basu *et al.*, 2023]	CVPR'18	81.34	79.84	83.65	66.74	51.40	**76.47**
Med-Debias	MEVF+SAN [Nguyen *et al.*, 2019]	MICCAI'19	75.97	74.72	77.88	60.71	40.65	74.05
	MEVF+BAN [Nguyen *et al.*, 2019]	MICCAI'19	77.76	75.97	80.53	62.34	43.09	75.14
Natural-Debias	RUBi [Cadene *et al.*, 2019]	NIPS'19	78.42	76.43	81.49	51.22	36.87	60.66
	LPF [Liang *et al.*, 2021]	SIGIR'21	75.59	73.33	79.09	56.32	49.72	60.66
	GGE-iter [Han *et al.*, 2021]	ICCV'21	79.83	79.22	80.77	65.19	49.16	75.74
	RMLVQA [Basu *et al.*, 2023]	CVPR'23	81.43	80.47	82.93	65.41	49.16	76.10
Ours	CEDO	—	**83.41**	**81.09**	**87.02**	**67.41**	**58.66**	73.16
		Increased ↑	1.98	0.62	3.37	0.67	7.62	-3.31

Table 2: Comparisons with state-of-the-art methods are conducted on the SLAKE and VQA-RAD datasets.

Datasets		VQA-CE		
Methods		Overall	Counter	Easy
SAN [Yang *et al.*, 2016]	CVPR'16	55.61	26.64	24.96
UpDn [Basu *et al.*, 2023]	CVPR'18	63.52	33.91	76.69
RMLVQA [Basu *et al.*, 2023]	CVPR'23	58.05	35.01	68.21
MSCD [Zhu *et al.*, 2024]	MM'24	58.82	35.67	69.12
CEDO	Ours	**60.05**	**35.71**	71.03

Table 3: The performance comparison on the VQA-CE dataset proves the satisfactory scalability and generalization of the CEDO.

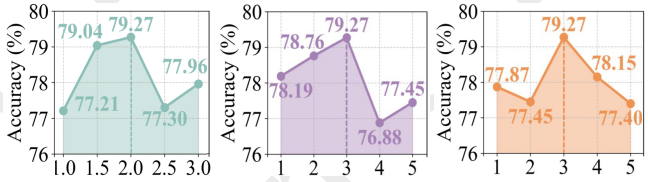
Methods	MHO	GMS	DLR	All	Open	Closed
Baseline	-	-	-	76.42	60.50	90.41
w/ MHO	✓	-	-	78.01	64.10	90.24
w/ GMS	-	✓	-	78.33	64.60	90.41
w/ DLR	-	-	✓	78.94	65.70	**90.59**
CEDO	✓	✓	✓	**79.27**	**66.70**	90.33

Table 4: Ablation experiments for different modules of the CEDO model on the biased SLAKE-CP dataset.

of 2.85% and 2.62% on SLAKE-CP and VQA-RAD-CP, respectively, underscoring its robustness in addressing medical bias challenges. 3) Debiasing models designed for natural scene datasets [Cadene *et al.*, 2019; Liang *et al.*, 2021; Han *et al.*, 2021] yield suboptimal results, highlighting their limited generalizability to the medical domain.

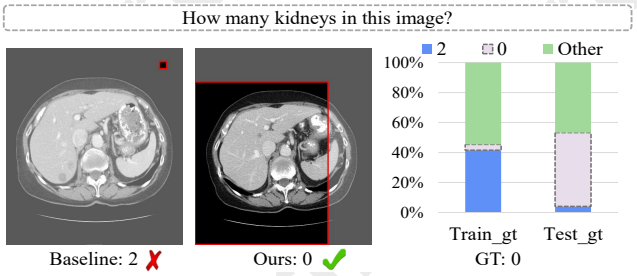
Evaluation on Medial Standard Benchmark. Table 2 highlights the significant performance improvements achieved by our CEDO method on the SLAKE [Liu *et al.*, 2021b] and VQA-RAD [Lau *et al.*, 2018] datasets. CEDO demonstrates notable gains of at least 1.98% and 0.67% over state-of-the-art approaches, respectively. It is important to note that while many existing methods perform well on in-distribution data, they exhibit pronounced performance degradation when exposed to varying prior conditions.

Evaluation on Natural Standard Benchmark. To evaluate the generalization capability of our CEDO approach in realistic application scenarios, we conduct experiments on the challenging *OOD* benchmark, VQA-CE. This benchmark is designed to test the robustness of VQA models under distributional shifts. As shown in Table 3, the proposed CEDO achieves a minimum improvement of 1.23%, demonstrating

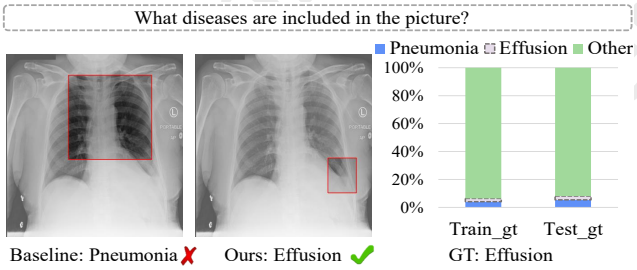


(a) $\eta_q \times 1000$ (b) $\eta_v \times 1000$ (c) $\eta_c \times 1000$

Figure 4: Comparison of Accuracy on the SLAKE-CP dataset with different parameter configurations.



(a) Prior distribution



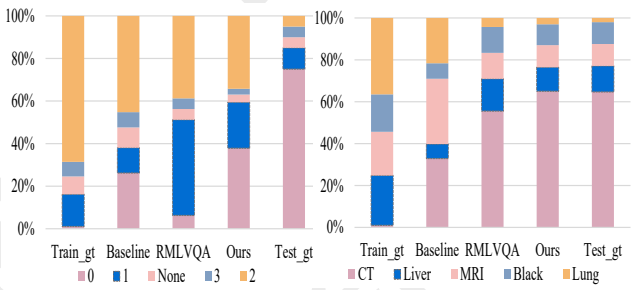
(b) Rare diseases distribution

Figure 5: Visualization analysis of CEDO. Our proposed approach combines robust reasoning capabilities for rare diseases with an effective strategy for mitigating bias-related challenges.

its effectiveness in handling *OOD* conditions.

4.4 Ablation Study

To assess the effectiveness of each component, ablation studies were conducted on SLAKE-CP, as presented in Table 4. The results are summarized as follows: 1) **Baseline:** We integrate [Basu *et al.*, 2023] as the base model. 2) **Baseline w/ MHO:** Adding the MHO mechanism improves performance by 1.59%, demonstrating its effectiveness in tailoring learning rates for different modalities. This mechanism mitigates language bias while enhancing the learning of complex visual features, leading to better multimodal integration. 3) **Baseline w/ GMS:** The GMS component yields a performance gain of 1.91%, effectively facilitating collaborative optimization between modalities, thus significantly reducing shortcut bias. 4) **Baseline w/ DLR**: The addition of DLR achieves a notable performance improvement of 2.52%, highlighting its capability to prevent sparse categories from being neglected and to mitigate data imbalance bias. 5) **Baseline w/ CEDO:** The full CEDO model achieves the best performance, demonstrating its robustness in tackling language bias comprehensively from both cause and effect sides.



(a) How (b) What

Figure 6: Bias Mitigation: Answer distributions of "How" and "What" question types on SLAKE-CP demonstrate that CEDO maintains prominent improvements over baseline and RMLVQA.

4.5 Parameter Analysis

We conduct an in-depth parameter analysis of the proposed CEDO method by exploring its behavior under various hyperparameter configurations. Our investigation focuses on three key hyperparameters: η_q, η_v, and η_c, as specified in Eqn. (4). Through systematic experimentation and detailed analysis, as depicted in Figure. 4, we observe that the model achieves peak performance when $\eta_q = 0.002$, $\eta_v = 0.003$, and $\eta_c = 0.003$. This analysis highlights that an optimal combination of these hyperparameters can achieve superior performance of the CEDO model.

4.6 Visualization Results

From the analysis in Figure 5, our CEDO method demonstrates dual advantages in Med-VQA by prioritizing critical visual information. It achieves outstanding debiasing performance, reducing language biases and modality shortcuts, while ensuring high diagnostic accuracy for rare diseases—an often overlooked yet vital aspect of medical AI. Additionally, Figure 6 demonstrates that our CEDO method significantly alleviates medical language bias by shifting the focus from biased patterns to essential information within the data. This enables the model to accurately predict unbiased answers, even in challenging scenarios where language shortcuts or distributional biases may otherwise dominate.

5 Conclusion

In this paper, we identified the cause and effect of medical language biases, i.e., shortcut bias and imbalance bias. To overcome these challenges, we proposed an innovative Cause-Effect Driven Optimization (CEDO) framework that addresses language biases from a causal perspective. The proposed method introduces a multi-learning rate strategy to heterogeneously optimize each modality, then applies the Pareto method and gradient orthogonal technology to achieve inter-modality coordination, thereby mitigating shortcut bias, combined with the loss rescaling mechanism to alleviate the imbalance bias. Two bias-sensitive Med-VQA datasets are constructed to evaluate the debiasing performance.

Ethical Statement

There are no ethical issues. Code is available at https://github.com/bvyih3/CEDO

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).

References

[Agrawal *et al.*, 2018] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 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 (CVPR)*, pages 4971–4980, 2018.

[Antol *et al.*, 2015] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433, 2015.

[Basu *et al.*, 2023] Abhipsa Basu, Sravanti Addepalli, and R. Venkatesh Babu. 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)*, pages 11671–11680, 2023.

[Ben Abacha *et al.*, 2019] Asma Ben Abacha, Sadid A Hasan, Vivek V Datla, Dina Demner-Fushman, and Henning Müller. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In *Proceedings of Conference and Labs of the Evaluation Forum (CLEF)*, 2019.

[Cadene *et al.*, 2019] Remi Cadene, Corentin Dancette, Hedi Ben-younes, Matthieu Cord, and Devi Parikh. Rubi: Reducing unimodal biases in visual question answering. In *Proceedings of the Conference and Workshop on Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

[Chen *et al.*, 2022] Zhihong Chen, Yuhao Du, Jinpeng Hu, Yang Liu, Guanbin Li, Xiang Wan, and Tsung-Hui Chang. Multi-modal masked autoencoders for medical vision-and-language pre-training. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 679–689. Springer, 2022.

[Clark *et al.*, 2019] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

[Dancette *et al.*, 2021] Corentin Dancette, Remi Cadene, Damien Teney, and Matthieu Cord. Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1574–1583, 2021.

[Désidéri, 2012] Jean-Antoine Désidéri. Multiple-gradient descent algorithm (mgda) for multiobjective optimization. *Comptes Rendus Mathematique*, 350(5-6):313–318, 2012.

[Do *et al.*, 2021] Tuong Do, Binh X Nguyen, Erman Tjiputra, Minh Tran, Quang D Tran, and Anh Nguyen. Multiple meta-model quantifying for medical visual question answering. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 64–74. Springer, 2021.

[Guo *et al.*, 2021] Yangyang Guo, Liqiang Nie, Zhiyong Cheng, Qi Tian, and Min Zhang. Loss re-scaling vqa: Revisiting the language prior problem from a class-imbalance view. *IEEE Transactions on Image Processing (TIP)*, 31:227–238, 2021.

[Han *et al.*, 2021] Xinzhe Han, Shuhui Wang, and Chi Su. Greedy gradient ensemble for robust visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1584–1593, 2021.

[Han *et al.*, 2023] Xinzhe Han, Shuhui Wang, Chi Su, Qingming Huang, and Qi Tian. General greedy de-bias learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45:1–17, 2023.

[Hesamian *et al.*, 2019] Mohammad Hesam Hesamian, Wenjing Jia, Xiangjian He, and Paul Kennedy. Deep learning techniques for medical image segmentation: achievements and challenges. *Journal of Digital Imaging (JDI)*, 32:582–596, 2019.

[Huang *et al.*, 2022] Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. Modality competition: What makes joint training of multi-modal network fail in deep learning?(provably). In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 9226–9259. PMLR, 2022.

[Khan *et al.*, 2022] Tariq M Khan, Syed S Naqvi, and Erik Meijering. Leveraging image complexity in macro-level neural network design for medical image segmentation. *Scientific Reports*, 12(1):22286, 2022.

[Khosla *et al.*, 2020] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Proceedings of the Conference and Workshop on Neural Information Processing Systems (NeurIPS)*, volume 33, pages 18661–18673, 2020.

[Kim *et al.*, 2018] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *Proceedings of the Conference and Workshop on Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

[Lau *et al.*, 2018] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018.

[Liang *et al.*, 2021] Zujie Liang, Haifeng Hu, and Jiaying Zhu. Lpf: A language-prior feedback objective function for de-biased visual question answering. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 1955–1959, 2021.

[Liu *et al.*, 2021a] Bo Liu, Li-Ming Zhan, and Xiao-Ming 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)*, pages 210–220. Springer, 2021.

[Liu *et al.*, 2021b] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654. IEEE, 2021.

[Liu *et al.*, 2021c] Xiangbin Liu, Liping Song, Shuai Liu, and Yudong Zhang. A review of deep-learning-based medical image segmentation methods. *Sustainability*, 13(3):1224, 2021.

[Liu *et al.*, 2022] Bo Liu, Li-Ming Zhan, Li Xu, and Xiao-Ming Wu. Medical visual question answering via conditional reasoning and contrastive learning. *IEEE Transactions on Medical Imaging (TMI)*, 42(5):1532–1545, 2022.

[Nguyen *et al.*, 2019] Binh D Nguyen, Thanh-Toan Do, Binh X Nguyen, Tuong Do, Erman Tjiputra, and Quang D Tran. Overcoming data limitation in medical visual question answering. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 522–530. Springer, 2019.

[Peng *et al.*, 2022] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8238–8247, 2022.

[Ramakrishnan *et al.*, 2018] Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. Overcoming language priors in visual question answering with adversarial regularization. In *Proceedings of the Conference and Workshop on Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

[Ross and Dollár, 2017] T-YLPG Ross and GKHP Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2980–2988, 2017.

[Sener and Koltun, 2018] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Proceedings of the Conference and Workshop on Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

[Tajbakhsh *et al.*, 2020] Nima Tajbakhsh, Laura Jeyaseelan, Qian Li, Jeffrey N Chiang, Zhihao Wu, and Xiaowei Ding. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical Image Analysis (MedIA)*, 63:101693, 2020.

[Vu *et al.*, 2020] Minh H Vu, Tommy Löfstedt, Tufve Nyholm, and Raphael Sznitman. A question-centric model for visual question answering in medical imaging. *IEEE Transactions on Medical Imaging (TMI)*, 39(9):2856–2868, 2020.

[Wang *et al.*, 2020] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12695–12705, 2020.

[Wang *et al.*, 2024] Hao Wang, Shengda Luo, Guosheng Hu, and Jianguo Zhang. Gradient-guided modality decoupling for missing-modality robustness. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, pages 15483–15491, 2024.

[Yang *et al.*, 2016] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21–29, 2016.

[Yu *et al.*, 2017] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. 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)*, pages 1821–1830, 2017.

[Zhan *et al.*, 2023] Chenlu Zhan, Peng Peng, Hanrong Zhang, Haiyue Sun, Chunnan Shang, Tao Chen, Hongsen Wang, Gaoang Wang, and Hongwei Wang. Debiasing medical visual question answering via counterfactual training. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 382–393. Springer, 2023.

[Zhu *et al.*, 2024] Jiawei Zhu, Yishu Liu, Huanjia Zhu, Hui Lin, Yuncheng Jiang, Zheng Zhang, and Bingzhi Chen. Combating visual question answering hallucinations via robust multi-space co-debias learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 955–964, 2024.