A Robust Dual-debiasing VQA Model based on Counterfactual Causal Effect

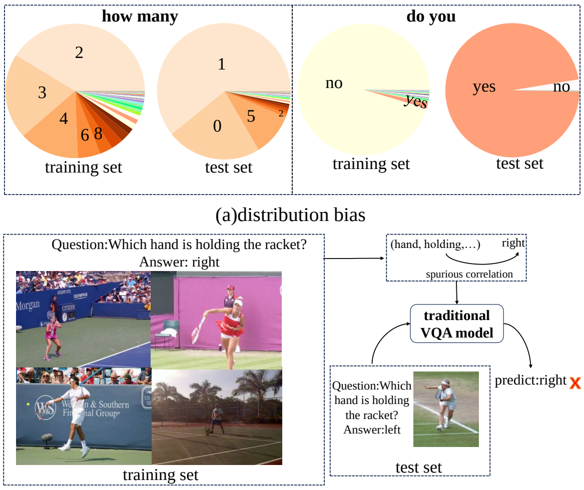
Lingyun Song[*][1], Chengkun Yang[2], Xuanyu Li[2], Xuequn Shang[†][2]
[1]Research & Development Institute of
Northwestern Polytechnical University in Shenzhen, Shenzhen, China
[2]School of Computer Science, Northwestern Polytechnical University, Xi'an, China
{lysong,shang}@nwpu.edu.cn; {yck,lxy}@mail.nwpu.edu.cn
Correspondence: lysong@nwpu.edu.cn; shang@nwpu.edu.cn

Abstract

Traditional VQA models are inherently vulnerable to language bias, resulting in a significant performance drop when encountering out-of-distribution datasets. The conventional VQA models suffer from language bias that indicates a spurious correlation between textual questions and answers. Given the outstanding effectiveness of counterfactual causal inference in eliminating bias, we propose a model-agnostic dual-debiasing framework based on Counterfactual Causal Effect (DCCE), which explicitly models two types of language bias (i.e., shortcut and distribution bias) by separate branches under the counterfactual inference framework. The effects of both types of bias on answer prediction can be effectively mitigated by subtracting direct effect of textual questions on answers from total effect of visual questions on answers. Experimental results demonstrate that our proposed DCCE framework significantly reduces language bias and achieves state-of-the-art performance on the benchmark datasets without requiring additional augmented data. Our code is available in https://github.com/sxycyck/dcce.

1 Introduction

Visual Question Answering (VQA) (Antol et al., 2015) is a challenging task that seeks to develop intelligent systems capable of generating answers to questions about a given image. Recently, Deep Neural Networks (DNN) have demonstrated robust representation learning capabilities across different types of modality data and have produced impressive results on VQA tasks (Song et al., 2023; Shao et al., 2023; Anderson et al., 2018). However, most existing VQA models only perform well on in-distribution (ID) datasets, but still are vulnerable with out-of-distribution (OOD) datasets where the answer distributions for the same question type differ between the training and test sets (Si et al., 2022; Agrawal et al., 2018). This occurs because these models often adopt shortcut solutions driven by dataset biases instead of the intended solutions derived from reasoning.



(a)distribution bias

(b)shortcut bias

Figure 1: Examples of two types of dataset bias.

In particular, as shown in Fig. 1(a), the distribution of answers for some question types is imbalanced. For example, in the VQA-CP validation set, simply answering "yes" to the questions "Do you..." can achieve 95% accuracy, whereas answering "1" to the questions "How many..." can achieve 60% accuracy. That is, they tend to directly output answers based on the answer distribution for certain question type , resulting in ***distribution bias***.

In addition, most DNN-based VQA methods exhibit over-reliance on datasets, which would induce DNNs to learn spurious correlations between questions and answers, resulting in ***shortcut bias***. This bias arises from question-answer shortcut rather than proper visual grounding. (Cadene et al., 2019), such as the frequent co-occurrence of question elements (e.g., keywords) and answers. As shown in Fig. 1(b), traditional VQA models tend to give the

[*]Corresponding author
[†]Corresponding author

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 4242–4252
November 12-16, 2024 ©2024 Association for Computational Linguistics

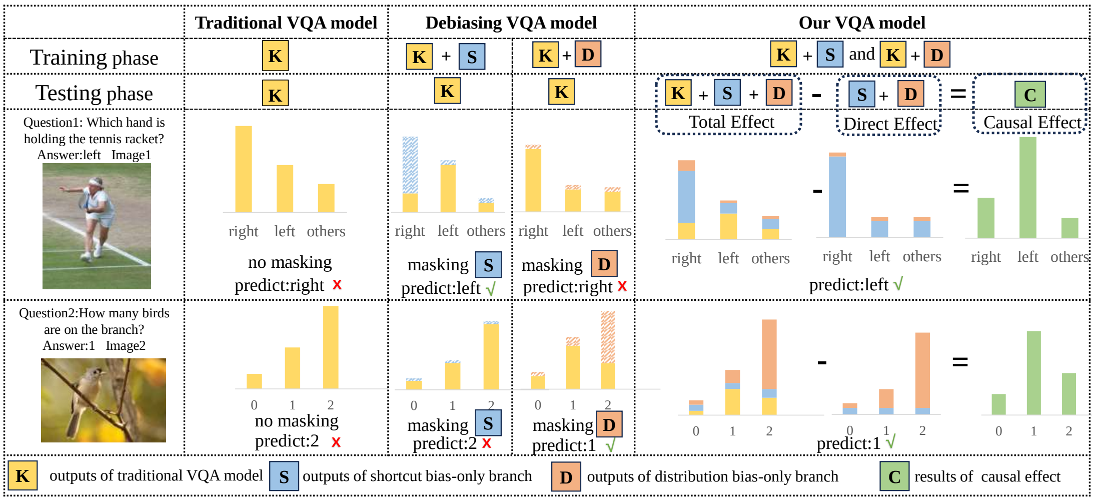


Figure 2: Comparison between different VQA models. Traditional VQA models make predictions based on the multimodal knowledge from visual questions, neglecting the language bias in datasets. Existing debiasing models mitigate the bias by simply masking the results from the branch modeling one type of bias. Our method mitigate the bias by removing the effect of two types of bias under the counterfactual inference framework. Different colors denote the predicted probability distribution obtained by different modules.

answer “right” when both words “hand” and “holding” appear in the question by shortcut correlation. This shortcut bias can result in a significant performance drop when the test set exhibits different statistical regularities. Mitigating the aforementioned language bias has been crucial for enhancing VQA performance.

Current debiasing methods for VQA often address language bias by incorporating an extra question-only branch (Cadene et al., 2019; Clark et al., 2019). Initially, this extra branch is employed to capture language bias from the training data, and subsequently, it is removed to mitigate the bias’s effect on answer predictions. However, the rough branch-excluding operation cannot efficiently disentangle the “good” language context from the captured language bias (Niu et al., 2021). Inspired by counterfactual reasoning and causal effects (Pearl and Mackenzie, 2018; Hitchcock and Pearl, 2000), the recent CF-VQA (Niu et al., 2021) addresses this issue by formulating language bias as the direct effect of questions on answers through the question-only branch, and mitigates the bias by subtracting the direct language effect from the total causal effect of visual questions on answers. Although CF-VQA shows encouraging results, it addresses language bias with a single branch. This is inadequate to capture both *shortcut bias* and *distribution bias* that represents two complementary aspects of the inherent language bias, which, if confused, would hamper performance (Han et al., 2021).

To address the above issues, we propose a model-agnostic Dual-debiasing framework based on Counterfactual Causal Effect (DCCE) (Wang and Vasconcelos, 2020; Pearl, 2001), which can effectively mitigate the shortcut bias and distribution bias, each representing distinct facets of language bias. Specifically, DCCE formulates the shortcut bias and the distribution bias as a direct causal effect of the content of the questions and the types of questions on the answers, respectively. The above two types of bias are mitigated by subtracting the corresponding direct causal effect from the total causal effect, which represents the causal effect of visual questions on answers. For our method, whether facing ID or OOD datasets, the prediction is not affected by data distribution interference, so it is robust.

To distinguish the effect of each bias type and multimodal reasoning in VQA, as shown in Fig. 2, during the training phase, DCCE develops two ensemble models that combine the VQA model with each of the two bias branches. A question-only branch is used to learn the shortcut bias, while a question type-only branch is used to learn the distribution bias. In this case, DCCE can easily mitigate the effects of two biases for the testing by subtracting the respective direct effects from the total effect. This is achieved by subtracting the logit sums of the question-only and question type-only branches from the total logit sum. Note that our DCCE is model-agnostic and can integrate different baseline VQA models. In summary, the contributions of our work can be summarized as follows.

- Our DCCE can mitigate the language bias in VQA from two different aspects (i.e., shortcut and distribution biases) simultaneously using a unified counterfactual inference framework. Specifically, shortcut bias is formulated as the direct effect of the question on the answers, while distribution bias is formulated as the direct effect of the question type on the answers.
- Our DCCE can effectively exclude the direct effect of shortcut bias and distribution bias on answers from the total causal effect. This ensures that the answer predictions are based on the causal effects of multimodal knowledge derived from the content of visual questions, rather than on shortcut or distribution biases learned from the training set.
- Our DCCE outperforms all previous state-of-the-art debiasing methods trained on biased datasets without requiring data augmentation.

2 Related Work

2.1 Language Bias in VQA

The VQA task necessitates strong abilities in both vision and language understanding. However, early well-known VQA datasets (such as VQA v1 (Antol et al., 2015) and VQAv2 (Goyal et al., 2017)) exhibit significant problems related to language bias. In particular, the distribution bias arises from the imbalanced answer distributions for various question types, whereas the shortcut bias arises from statistical regularities between questions and answers. To mitigate the language bias in VQA, researchers develop the VQA-CP dataset Agrawal et al. (2018), which features markedly different answer distributions between the training and testing sets. The goal of VQA-CP is to force VQA models to make predictions using the multimodal information contained in visual questions, rather than depending on biases learned from the training data.

2.2 Debiasing Methods in VQA

Current debiasing techniques can be divided into three categories (Ma et al., 2024): (1) enhancing visual information, emphasizing the importance of visual data in the model; (2) employing data augmentation to balance the original dataset; (3) addressing language bias by first modeling language bias and subsequently reducing their effect.

The first group of works aims to improve the performance by forcing VQA models to strengthen the usage of visual information grounded with language questions. (Goyal et al., 2017; Hendricks et al., 2018; Selvaraju et al., 2019; Wu and Mooney, 2019). Although visual information in images is crucial for reasoning answers to visual problems, the performance enhancement of the aforementioned methods is primarily due to the disruption of the learned biases by incorporating visual information((Liang et al., 2020)).

The second group of methods focuses on balancing datasets using augmentation techniques and employs these balanced data to train robust VQA models (Zhu et al., 2020; Chen et al., 2020; Gokhale et al., 2020). Nevertheless, data augmentation is expensive and limits the development of robust VQA models capable of generalizing well to other unbiased datasets.

The third group of methods first use a separate question-only branch to learn the bias, and then exclude the branch to mitigate the language bias in testing phase Clark et al. (2019); Cadene et al. (2019); Niu et al. (2021); Han et al. (2021). However, the crude branch-excluding technique fails to preserve the beneficial contextual information from the language bias that aids in deriving correct answers. Later, CFVQA Niu et al. (2021) addresses this issue by the counterfactual inference which disentangles the negative effect of language bias on the answers from the total effect. However, CFVQA uses a single branch to capture language bias, thereby conflating distribution bias and shortcut bias, which are complementary yet depict different aspects of language bias (Han et al., 2021). In contrast, our DCCE distinguishes between the two types of bias within the counterfactual inference framework.

3 Methodology

3.1 Language Bias Modeling

In this part, we introduce the modeling of the shortcut bias and the distribution bias.

Shortcut bias arises due to spurious correlations between the question content and the answer. We build a QA model f_Q to capture the shortcut bias.

$$z_{q_i} = f_Q(e_q(q_i)), \tag{1}$$

where q_i denotes the i-th textual question, e_q denotes text encoder, $z_{q_i} \in \mathbb{R}^{n_c}$ denotes the probability distribution of answers generated by the QA model and n_c denotes number of candidate answers.

Distribution bias arises from imbalanced answer distributions for specific question types. Specifically, the answer distribution can be obtained by a statistical calculation process (denoted by f_T), which can be formulated by

$$b_{t_i} = f_T(t_i), \tag{2}$$

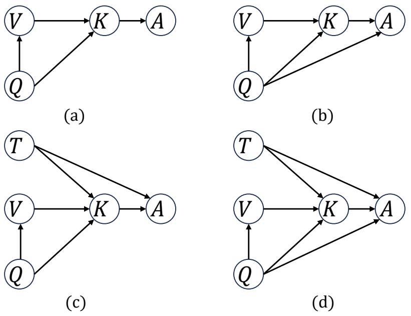
where t_i is the question type of i-th question and $b_{t_i} \in \mathbb{R}^{n_c}$ is the probability distribution of the answer corresponding to the question type t_i.

3.2 Counterfactual Causal Effect

In this part, we introduce the construction of Causal graphs for input visual questions. Specifically, given a question Q paired with image V, we construct the Causal graph for conventional VQA models as shown in Fig. 3 (a), where T denotes the question type, K is composed of V and Q, and A denotes the answer. $Q \rightarrow V$ denotes the effect of Q on V, which is implemented by applying question-guided attention over images to obtain the relevant visual feature (Anderson et al., 2018). $K \rightarrow A$ denotes effect of the multimodal knowledge K on A, indicating that the model predicts the answer using multimodal knowledge.

The causal graph of the debiasing VQA model that model the shortcut bias is shown in Fig. 3 (b). Compared with traditional VQA models, the debiasing uses an additional branch to incorporate the effect of the single-modal question Q on A (i.e., $Q \rightarrow A$). In this way, Q typically affects A through two paths: $Q \rightarrow A$ indicates the direct effect on the answer, whereas $Q \rightarrow K \rightarrow A$ indicates the indirect effect on the answer.

The Causal graph addressing the distribution bias is shown in Fig. 3 (c). The path $T \rightarrow A$ indicates that there is a strong correlation between the question type and the answer due to the distribution bias. To mitigate both types of bias, we construct the Causal graph depicted in Fig. 3 (d), where T and Q affect A by both direct and indirect paths.



(a) (b) (c) (d)

Figure 3: Simplified Causal graphs for different VQA models: (a) Conventional VQA models. (b) Debiasing VQA models considering shortcut bias. (c) Debiasing VQA models considering distribution bias. (d) Our VQA model considering both types of biases.

Inspired by counterfactual causal inference (Hitchcock and Pearl, 2000; Pearl, 2001), we exclude the pure language effect on the answer, referred to as Direct Effect (DE), to mitigate shortcut bias. The effect of multimodal knowledge K on the answer is termed the Indirect Effect (IE), encompassing both the Total Indirect Effect (TIE) and the Natural Indirect effect (NIE). As shown in Table 1, $Q = q^*$ represents a constant used to form the question feature. $K(Q = q^*)$ signifies that a constant is employed as the question feature to form the multimodal feature.

Table 1: Counterfactual Causal effect. Natural Effect (NE) represents the effect of single variable (Q or K) on A. Total Effect (TE) represents the effect of multi variables (Q and K) on A. $P(A|K, Q = q)$ represents effect of K on A given $Q = q$.

	$DE(Q \rightarrow A)$	$IE(K \rightarrow A)$
$NE(Q\ or\ K)$	$P(A\|Q, K(Q = q^*))$ $NDE(Natural\ Direct\ Effect)$	$P(A\|K, Q = q^*)$ $NIE(Natural\ Indirect\ Effect)$
$TE(Q\ and\ K)$	$P(A\|Q, K(Q = q))$ $TDE(Total\ Direct\ Effect)$	$P(A\|K, Q = q)$ $TIE(Total\ Indirect\ Effect)$

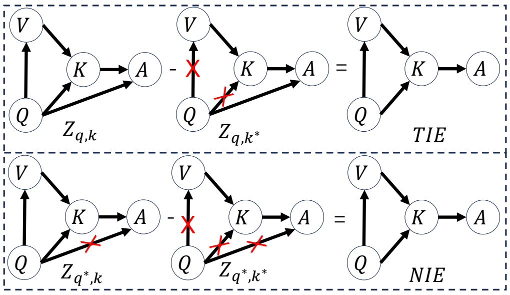


Figure 4: Two counterfactual Causal effect:TIE and NIE for mitigating shortcut bias. The red cross mean that the path is blocked.

Conventional VQA models consider the Total Effect (TE) as predictions, i.e., $TE = P(A|K, Q)$, which incorporates both $Q \rightarrow A$ and $K \rightarrow A$. What we really need to evaluate is the effect of K on A and it can be achieved through two methods: Natural Indirect Effect (NIE) and Total Indirect Effect (TIE), both of which exclude the effect of pure language on the answers. As shown in Fig. 4, the NIE and TIE of multimodal knowledge K can be derived by eliminating the path $Q \rightarrow A$ through two different ways. Specifically, NIE and TIE can be learned by

$$NIE = P(A|K, Q = q^*) = Z_{q^*,k} - Z_{q^*,k^*} \quad (3)$$

$$TIE = P(A|K, Q = q) = Z_{q,k} - Z_{q,k^*}. \quad (4)$$

$Q = q^*$ represents $Q \rightarrow A$ is blocked, $k^* = K(Q = q^*)$ represents $Q \rightarrow K$ is blocked, $Z_{q,k}$ denotes the TE, Z_{q,k^*} denotes the NDE. $Z_{q^*,k}$ represents the NIE and Z_{q^*,k^*} is actually a constant vector. The value Z denotes the effect of variables on answers. Specifically, $Z_{q,k}$ and Z_{q,k^*} can be obtained by

$$Z_{q,k} = Z(Q = q, K(Q = q)), \quad (5)$$

$$Z_{q,k^*} = Z(Q = q, K(Q = q^*)). \quad (6)$$

In $Z_{q^*,k}$ and Z_{q,k^*}, the conditions $Q = q$ and $Q = q^*$ happen at the same time, which is counterfactual since Q can only be either q or q^* in reality. Using counterfactual reasoning, we manually intervene in the question, manage whether the path is blocked, and determine the causal effect of multimodal knowledge on the answers. Similarly to shortcut bias, distribution bias is mitigated in the same way. Specifically, NIE and TIE can be formulated by

$$NIE = Z_{t^*,k} - Z_{t^*,k^*}, \quad (7)$$

$$TIE = Z_{t,k} - Z_{t,k^*}. \quad (8)$$

When considering both type of biases, TIE and NIE are represented by

$$NIE = Z_{q^*,t^*,k} - Z_{q^*,t^*,k^*}, \quad (9)$$

$$TIE = Z_{q,t,k} - Z_{q,t,k^*}. \quad (10)$$

3.3 Debiasing VQA Model

We use the results Z_k generated by the vanilla VQA model to represent the effect of multimodal knowledge on the answers. Specifically, Z_k can obtained by

$$Z_k = \begin{cases} z_k &= f_{V,Q}(v, q), \text{if } Q = q \\ z_{k^*} &= c, \qquad \text{if } Q = \varnothing \end{cases} \quad (11)$$

where $f_{V,Q}$ represents the vanilla VQA model, $Q = \varnothing$ aligns with $Q = q^*$ and c is a learnable parameter.

In this section, we introduce the process of mitigating both types of bias, as illustrated in Figure 5. To fully mitigate language bias, we establish two separate branches to train the VQA model.

Shortcut bias branch. We utilize the output Z_q produced by the QA model to denote the direct effect of the question content on answers, aiming to capture the shortcut bias. Specifically, Z_q can be learned by

$$Z_q = \begin{cases} z_q &= f_Q(q), \text{if } Q = q \\ z_{q^*} &= c, \qquad \text{if } Q = \varnothing \end{cases}, \quad (12)$$

where f_Q denotes a QA model. We combine Z_q with Z_k to denote the total effect $Z_{q,k}$, which encompasses both the direct effect $(Q \rightarrow A)$ and the indirect effect $(Q \rightarrow K \rightarrow A)$. $Z_{q,k}$ can learned by

$$Z_{q,k} = log \ \ \sigma(Z_q + Z_k). \quad (13)$$

During the training phase, samples represented by triplets (v, q, a) are given for training. The parameters of our VQA and QA models are optimized by the cross-entropy loss function $loss_{CE}$ shown below.

$$L_{VQA_q}(v, q, a) = loss_{CE}(Z_{q,k}, a), \quad (14)$$

$$L_{QA}(q, a) = loss_{CE}(Z_q, a). \quad (15)$$

The loss function for the shortcut bias branch is the sum of L_{VQA_q} and L_{QA}, i.e.,

$$L_s = \sum L_{VQA_q} + L_{QA}. \quad (16)$$

Distribution bias branch. Given the answer probability distribution b_t for each question type in Subsection 3.1, the direct effect Z_t of the question type on answers, attributed to distribution bias, is defined by

$$Z_t = \begin{cases} z_t &= h(l) * r(b_t), \text{if } T = t \\ z_{t^*} &= c, \qquad \text{if } T = \varnothing \end{cases}, \quad (17)$$

where $h(l) = Softplus(W(l))$ is a function used to control the weight of the answer probability b_t, and l is the feature output by the VQA model $f_{V,Q}$. In particular, the function $softplus$ ensures that $h(\cdot)$ remains positive, allowing Z_t to contribute a positive factor when combined with

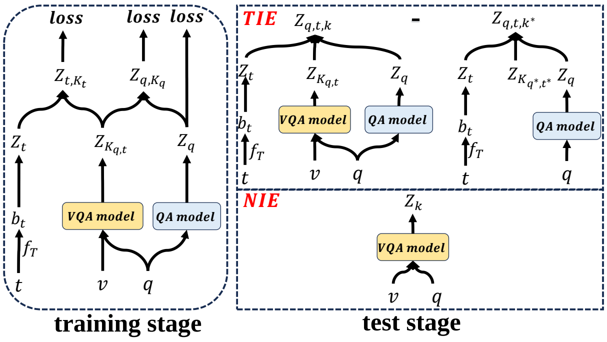


Figure 5: The pipeline of our method. In the train stage, VQA model and QA model are trained. In the test stage, Z_{k^*} is constants.

Z_k. W is a parameter matrix that transforms the multidimensional l into a single dimension. $r(b_t) = \log(b_t + \sigma(s))$, where b_t is derived from f_T according to Eq.(2) and represents the answer distribution for the question type t. s is a trainable parameter employed to smooth the training process. $Z_{t,k}$ denotes the total effect, encompassing both the direct effect $(T \to A)$ and the indirect effect $(T \to K \to A)$. Specifically, $Z_{t,k}$ can be obtained by fusing the distribution prior with the outputs produced by the VQA model $f_{V,Q}$, which can be formulated by

$$Z_{t,k} = log \ \ \sigma(Z_t + Z_k). \tag{18}$$

During the training stage, we optimize the parameters of VQA model by the following cross-entropy loss.

$$L_{VQA_t}(v, q, b_t, a) = loss_{CE}(Z_{t,k}, a), \tag{19}$$

where the probability distribution b_t is a prior probability which is precomputed. However, Clark et al. (2019) observed that $h(\cdot)$ in Eq. (17) tends to 0 during the training phase, which is an unexpected result. Therefore, a penalty term loss function is appended to solve this issue. The loss function for penalty is represented by

$$L_{penalty} = w * g(softmax(Z_t)), \tag{20}$$

where $g(x) = -\sum_i x_i \log(x_i)$, which denotes the entropy of the vector x. The parameter w is a hyper-parameter that controls the weight of $L_{penalty}$. Thus, the loss function for the distribution bias branch is the sum of L_{VQA_t} and $L_{entropy}$.

$$L_d = \sum L_{VQA_t} + L_{penalty}. \tag{21}$$

Since we need to subtract DE from TE, it is crucial that Z_{q,t,k^*} align with $Z_{q,t,k}$ in the probability distribution as closely as possible. The KL divergence loss is used to quantify the difference between Z_{q,t,k^*} and $Z_{q,t,k}$. Thus, we use the KL divergence loss function to optimize the parameters, which can be represented by

$$L_{kl} = \frac{1}{N} \sum -p(a|q,t,k) logp(a|q,t,k^*), \tag{22}$$

where $p(a|q,t,k) = softmax(Z_{t,q,k})$. It is important to note that solely the parameter c within the KL divergence loss will be subject to training and optimization. The loss function for the shortcut bias branch is denoted as L_s, while the loss function for the distribution bias branch is represented as L_d. Overall, the final loss function is

$$L = L_s + L_d + L_{kl}. \tag{23}$$

During the inference, the total effect $Z_{q,t,k}$ is derived by merging the indirect effect and the direct effect resulting from the shortcut bias with the direct effect caused by the distribution bias. The computation of $Z_{q,k,t}$ can be formulated by

$$Z_{q,t,k} = log \ \ \sigma(Z_q + Z_t + Z_k). \tag{24}$$

We denote the pure language direct effect by Z_{q,t,k^*}. Answers can be derived in two ways: NIE and TIE, which can be obtained by

$$\hat{a}_{TIE} = arg \max_{a \in A}(Z_{q,t,k} - Z_{q,t,k^*}), \tag{25}$$

$$\begin{aligned} \hat{a}_{NIE} &= arg \max_{a \in A}(Z_{q^*,t^*,k} - Z_{q^*,t^*,k^*}) \\ &= arg \max_{a \in A}(Z_k), \end{aligned} \tag{26}$$

where $\hat{a}_{TIE}$ and $\hat{a}_{NIE}$ denotes answer id with max probability obtained via TIE and NIE, respectively. In the equation for $\hat{a}_{NIE}$, both q^*, t^* and Z_{q^*,t^*,k^*} are constant, so NIE and Z_k have a similar relative order. Given the features of visual and textual questions, Z_k can be derived using $f_{V,Q}(v, q)$.

4 Experiment Results and Analysis

4.1 Experimental Setup

Following the baseline methods, we train our model on the VQA-CP v2 dataset (Agrawal et al., 2018) and evaluate its performance on both the VQA-CP v2 and VQA v2 datasets. The answer distribution for each type of question in VQA-CP v2 varies between the training and validation sets. The VQA-CP v2 training set consists of 121K images, 438K

Table 2: Comparison on VQA v2 and VQA-CP v2 dataset between some advanced methods and our method. Our methods include Ours (TIE) and Ours (NIE). The best results without extra augmented data is highlighted in bold.

Methods	VQA-CP v2 val				VQA v2 val			
	All	Y/N	Num	Other	All	Y/N	Num	Other
UpDn (Anderson et al., 2018)	39.74	42.27	11.93	46.05	63.48	81.18	42.14	**55.66**
methods based on strengthening visual information:								
HINT (Selvaraju et al., 2019)	46.73	67.27	10.61	45.88	63.38	81.18	42.99	55.56
SCR (Wu and Mooney, 2019)	49.45	72.36	10.93	48.02	62.2	78.8	41.6	54.5
AdvReg (Ramakrishnan et al., 2018)	41.17	65.49	15.48	35.48	62.75	79.84	42.35	55.16
methods based on using extra augmented data:								
RandImg (Teney et al., 2020)	55.37	83.89	41.60	44.20	57.24	76.53	33.87	48.57
SSL (Zhu et al., 2020)	57.59	86.53	29.87	50.03	63.73	-	-	-
CSS (Chen et al., 2020)	58.95	84.37	49.42	48.21	59.91	73.25	39.77	55.11
CSS+CL (Liang et al., 2020)	59.18	86.99	49.89	47.16	57.29	67.27	38.40	54.71
Mutant (Gokhale et al., 2020)	61.72	88.90	49.68	50.78	62.56	82.07	42.52	53.28
methods based on mitigating language bias:								
RUBi (Cadene et al., 2019)	44.23	67.05	17.48	39.61	61.16	78.96	40.85	54.14
LMH (Clark et al., 2019)	52.01	72.58	31.12	46.97	56.35	65.06	37.63	54.69
CFVQA (Niu et al., 2021)	53.55	**91.15**	13.03	44.97	**63.54**	**82.51**	**43.96**	54.30
CIKD (Pan et al., 2021)	54.05	90.01	15.10	45.88	61.29	76.34	40.2	55.43
GGE (Han et al., 2021)	57.32	87.04	27.75	**49.59**	59.11	73.27	39.99	54.39
Ours(NIE)	56.14	82.15	43.94	45.89	60.83	77.33	40.92	53.55
Ours(TIE)	**59.12**	91.12	**45.74**	46.11	61.01	77.99	40.88	53.49

Table 3: Eight configurations of the variant models for bias mitigation.“S-bias” stands for addressing shortcut bias. “D-bias” stands for addressing distribution bias.

S-bias	D-bias	Result	Work
TIE	-	$Z_{q,k}$-Z_{q,k^*}	CFVQA
NIE	-	$Z_{q^*,k}$-Z_{q^*,k^*}	RUBi
-	TIE	$Z_{t,k}$-Z_{t,k^*}	-
-	NIE	$Z_{t^*,k}$-Z_{t^*,k^*}	LMH
NIE	NIE	$Z_{t^*,q^*,k}$-Z_{t^*,q^*,k^*}	-
NIE	TIE	$Z_{t,q^*,k}$-Z_{t,q^*,k^*}	-
TIE	NIE	$Z_{t^*,q,k}$-Z_{t^*,q,k^*}	-
TIE	TIE	$Z_{t,q,k}$-Z_{t,q,k^*}	-

questions and 4.4M answers, whereas the validation set consists of 98K images, 220K questions and 2.2M answers.

For fair comparisons, we use the UpDn (Anderson et al., 2018) model as the backbone to implement the baseline models. Our models are trained using a single NVIDIA RTX4090 24GB GPU for 22 epochs, with a batch size of 256 and a learning rate of 1×10^{-5}. Each epoch takes approximately twenty five minutes when using UpDn as baseline. We set the hyper-parameter w to dynamically change to prevent model over-correct distribution bias by following formula:

$$w = \begin{cases} 2 * cos(\frac{i}{n} * \frac{\pi}{2}), \text{if } i < n \\ 0, \qquad\qquad others \end{cases}, \tag{27}$$

where i denotes i-th epoch, $n = \frac{2}{3}N$, N denotes total number of epochs.

4.2 Comparison with Advanced Methods

We compare the experimental results of our proposed approach with state-of-the-art techniques, using methods such as HINT(Selvaraju et al., 2019) and SCR (Wu and Mooney, 2019) to improve visual information usage, techniques such as AdvReg (Ramakrishnan et al., 2018), RUBi (Cadene et al., 2019), LMH (Clark et al., 2019), CFVQA (Niu et al., 2021) and GGE (Han et al., 2021) to mitigate language bias, and methods such as SSL (Zhu et al., 2020), CSS(Chen et al., 2020), CL (Liang et al., 2020), and Mutant (Gokhale et al., 2020) to balance and augment data sets.

The experimental results are shown in Table 2. We conducted experimental validation under two cases: TIE and NIE, and our approach (TIE) demonstrates strong performance even without the use of extra augmented data. Previous studies Niu et al. (2021); Clark et al. (2019) have noted that shortcut bias mainly influences “Y/N” questions, while distribution bias primarily affects both “Y/N” and “number” questions. Compared to LMH, which is particularly effective at addressing distribution bias in “number” questions, our approach demonstrates a substantial enhancement,

Table 4: Ablation studies of various cases on VQA-CP v2 dataset. “S-bias” represents that mitigating shortcut bias. “D-bias” represents that mitigating distribution bias. The best results is highlighted.

S-bias	D-bias	VQA-CP v2 val				VQA v2 val			
		All	Y/N	Num	Other	All	Y/N	Num	Other
-	-	39.74	42.27	11.93	46.05	63.48	81.18	42.14	**55.66**
NIE	-	46.47	69.95	12.83	43.39	**63.71**	**82.6**	**44.08**	54.55
TIE	-	53.55	**91.15**	13.03	44.97	63.54	82.51	43.96	54.3
-	NIE	55.98	79.44	50.15	45.29	50.61	53.18	36.97	52.28
-	TIE	58.93	89.02	**51.05**	45.32	49.96	51.52	36.88	52.26
NIE	NIE	55.15	78.89	43.92	45.92	60.93	77.75	40.84	53.52
NIE	TIE	55.29	79.23	43.74	45.92	60.93	77.65	40.92	53.52
TIE	NIE	56.14	82.15	43.94	45.89	60.83	77.33	40.92	53.55
TIE	TIE	**59.12**	91.12	45.74	**46.11**	61.01	77.99	40.88	53.49

surpassing LMH by 14.62%. This underscores our method’s ability to enhance the mitigation of distribution bias. Compared to CFVQA, our approach is only 0.03% less accurate for “Y/N” questions, but it shows superior accuracy for “number” questions. This indicates that our proposed model can effectively addresses one type of bias without diminishing the mitigation effect on another bias. TIE outperforms NIE when it comes to “Y/N” questions. Compared to the dataset balancing approach, our accuracy exceeds that of SSL and CSS without requiring any modifications to the dataset and is almost on par with CSS+CL.

4.3 Ablation Studies

This part aims to evaluate the performance of the variants of our models that mitigate bias using several different configurations, which are listed in Table 3. Specifically, there are eight variant models that target shortcut bias, distribution bias, or a combination of both. Each type of bias can be addressed by the NIE or TIE.

We have identified several recent studies that theoretically align with one of the configurations of our variants. For example, CFVQA (Niu et al., 2021) uses TIE to derive answers, reducing the shortcut bias. Similarly, RUBi (Cadene et al., 2019) utilizes NIE to derive answers, addressing the shortcut bias. Additionally, LMH (Clark et al., 2019) leverages NIE to derive answers, mitigating distribution bias.

The experimental results are presented in Table 4. We can observe that addressing distribution bias greatly enhances the accuracy of both the “number” and “Y/N” questions. Likewise, addressing shortcut bias leads to a significant improvement in the accuracy of “Y/N” questions. This indicates that shortcut bias and distribution bias may overlap when dealing with “Y/N” questions. The accuracy obtained using NIE to derive answers is lower compared to TIE, indicating that TIE is a more efficient approach. When we address both shortcut bias and distribution bias together and use TIE to derive answers, we achieve the maximum accuracy on the VQA-CP v2 datasets. In addition, the inferior results of our method relative to the baselines on VQA v2 highlight the capability of our method to mitigate bias. The higher performance of baselines lie in their memorization of the bias in the training data, which can be demonstrated by their dramatic performance drop on VQA-CP v2. Shortcut bias and distribution bias are two aspects of language bias(Han et al., 2021), a single branch is unable to model such two types of biases. Mitigating both bias can achieve a better trade-off between accuracy and robustness.

4.4 Quantitative and Qualitative Analysis

We conducted a quantitative analysis of our model, as shown in table 5, “ST” represents the samples corrected by shortcut bias. “SF” represents the samples not corrected by shortcut bias.The meaning of “DT” and “DF” is similar, which is for distribution bias. For example, “ST&DF” represents that the number of samples corrected by shortcut bias but not corrected by distribution bias. We find that mitigating distribution bias can effectively improve accuracy on “Num” questions. Each bias can correct some samples that not be corrected by another bias for the “Y/N” questions and “Other” questions. So the samples they can correct don’t completely overlap. Simply mitigating single bias is not enough to achieve language bias mitigation.

We conducted a qualitative analysis of our model, as shown in Fig. 6. In the first example, it is evident that the probability of “no” produced by UpDn is significantly greater than “yes”, suggesting that UpDn disregards visual information

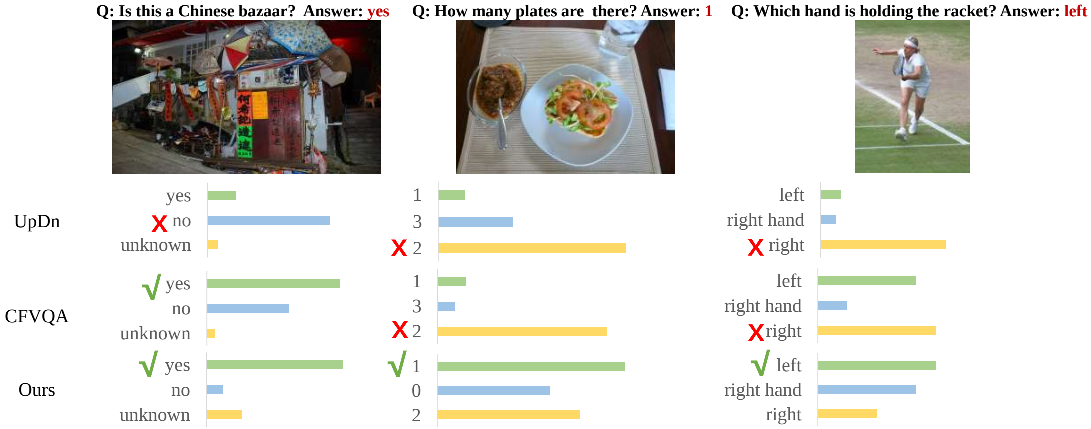


Figure 6: Qualitative comparison between UpDn, CFVQA and Ours.

Table 5: Quantitative analysis for the number of samples corrected by shortcut bias or distribution bias.

Answer Type	ST&DF	SF&DT	ST&DT
Y/N	5072	2661	18407
Num	748	12746	802
Other	5267	6245	6118

and depends solely on the dataset bias. In contrast, our model successfully mitigates bias, resulting in a higher probability for “yes” compared to “no”. For the second example, UpDn and CFVQA predict the answer "2" according to distribution prior. In contrast, our model actually counts the objects based on the visual questions and gives the correct answer. In the third example, our model effectively mitigates shortcut bias and correctly predicts the answer “left” based on the multimodal information of visual questions.

5 Conclusion

This research introduces a novel VQA model designed to mitigate shortcut bias and distribution bias in VQA tasks by leveraging counterfactual causal effects. Furthermore, we explore how various configurations can mitigate language bias in VQA within the framework of counterfactual inference. Experimental results show show that answers obtained through the causal effect TIE exhibit superior accuracy, highlighting the efficacy of the method.

Limitations

The stochastic intersection of shortcut bias and distribution bias has not been taken into account, but this minimally affects our approach to addressing both biases at the same time. The performance on the VQA v2 val dataset which is insensitive to language bias, slightly declined. We think that it may be due to over-fitting of the original VQA model on the VQA v2 dataset with same distribution between training and val datasets. And performance of our robust method returns to normal levels after debiasing.

Acknowledgment

The research was supported in part by National Nature Science Foundation of China under Grant Nos. 62102321, Guangdong Basic and Applied Basic Research Foundation under Grant No. 2024A1515011715, and the Fundamental Research Funds for the Central Universities under Grant No. D5000230095. We would like to thank the anonymous reviewers and meta-reviewer for their insightful suggestions.

References

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don’t just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 4971 – 4980.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 6077 – 6086.

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

Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. 2019. Rubi: Reducing unimodal biases for visual question answering. *Advances in Neural Information Processing Systems*, 32:841–852.

Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. 2020. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 10797 – 10806.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing*, pages 4069 – 4082.

Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020. Mutant: A training paradigm for out-of-distribution generalization in visual question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 878 – 892.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, volume 2017-January, pages 6325 – 6334.

Xinzhe Han, Shuhui Wang, Chi Su, Qingming Huang, and Qi Tian. 2021. Greedy gradient ensemble for robust visual question answering. In *Proceedings of the IEEE international conference on computer vision*.

Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. 2018. Grounding visual explanations. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11206 LNCS, pages 269 – 286.

Hitchcock and Pearl. 2000. *Causality: Models, Reasoning and Inference*. Cambridge University Press.

Zujie Liang, Weitao Jiang, Haifeng Hu, and Jiaying Zhu. 2020. Learning to contrast the counterfactual samples for robust visual question answering. In *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pages 3285 – 3292.

Jie Ma, Pinghui Wang, Dechen Kong, Zewei Wang, Jun Liu, Hongbin Pei, and Junzhou Zhao. 2024. Robust visual question answering: Datasets, methods, and future challenges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1 – 20.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Yonghua Pan, Zechao Li, Liyan Zhang, and Jinhui Tang. 2021. Distilling knowledge in causal inference for unbiased visual question answering. In *Proceedings of the 2nd ACM International Conference on Multimedia in Asia*, pages 1–7.

Judea Pearl. 2001. Direct and indirect effects. *Probabilistic and Causal Inference: The Works of Judea Pearl*, page 373.

Judea Pearl and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect*. Basic books.

Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. 2018. Overcoming language priors in visual question answering with adversarial regularization. In *Advances in Neural Information Processing Systems*, volume 2018-December, pages 1541 – 1551.

Ramprasaath Ramasamy Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry Heck, Dhruv Batra, and Devi Parikh. 2019. Taking a hint: Leveraging explanations to make vision and language models more grounded. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2019-October, pages 2591 – 2600.

Zhenwei Shao, Zhou Yu, Meng Wang, and Jun Yu. 2023. Prompting large language models with answer heuristics for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14974–14983.

Qingyi Si, Fandong Meng, Mingyu Zheng, Zheng Lin, Yuanxin Liu, Peng Fu, Yanan Cao, Weiping Wang, and Jie Zhou. 2022. Language prior is not the only shortcut: A benchmark for shortcut learning in vqa. *arXiv preprint arXiv:2210.04692*.

Lingyun Song, Jianao Li, Jun Liu, Yang Yang, Xuequn Shang, and Mingxuan Sun. 2023. Answering knowledge-based visual questions via the exploration of question purpose. *Pattern Recognition*, 133:109015.

Damien Teney, Ehsan Abbasnejad, Kushal Kafle, Robik Shrestha, Christopher Kanan, and Anton Van Den Hengel. 2020. On the value of out-of-distribution testing: An example of goodhart's law. *Advances in neural information processing systems*, 33:407–417.

Pei Wang and Nuno Vasconcelos. 2020. Scout: Self-aware discriminant counterfactual explanations. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 8978 – 8987.

Jialin Wu and Raymond Mooney. 2019. Self-critical reasoning for robust visual question answering. *Advances in Neural Information Processing Systems*, 32.

Xi Zhu, Zhendong Mao, Chunxiao Liu, Peng Zhang, Bin Wang, and Yongdong Zhang. 2020. Overcoming language priors with self-supervised learning for visual question answering. In *IJCAI International Joint Conference on Artificial Intelligence*, volume 2021-January, pages 1083 – 1089.