Counterfactual VQA: A Cause-Effect Look at Language Bias

Yulei Niu[1]   Kaihua Tang[1]   Hanwang Zhang[1]   Zhiwu Lu[2]   Xian-Sheng Hua[3]   Ji-Rong Wen[2]

[1]Nanyang Technological University

[2]Gaoling School of Artificial Intelligence, Renmin University of China

[3]Damo Academy, Alibaba Group

`yn.yuleiniu@gmail.com`, `{kaihua001@e.,` `hanwangzhang}@ntu.edu.sg`, `{luzhiwu,` `jrwen}@ruc.edu.cn`

Abstract

Recent VQA models may tend to rely on language bias as a shortcut and thus fail to sufficiently learn the multi-modal knowledge from both vision and language. In this paper, we investigate how to capture and mitigate language bias in VQA. Motivated by causal effects, we proposed a novel counterfactual inference framework, which enables us to capture the language bias as the direct causal effect of questions on answers and reduce the language bias by subtracting the direct language effect from the total causal effect. Experiments demonstrate that our proposed counterfactual inference framework 1) is general to various VQA backbones and fusion strategies, 2) achieves competitive performance on the language-bias sensitive VQA-CP dataset while performs robustly on the balanced VQA v2 dataset.

1. Introduction

Visual Question Answering (VQA) [8, 4] has become the fundamental building block that underpins many frontier interactive AI systems, such as visual dialog [15], vision-language navigation [6], and visual commonsense reasoning [48]. VQA systems are required to perform visual analysis, language understanding, and multi-modal reasoning. Recent studies [19, 3, 8, 19, 24] found that VQA models may rely on spurious linguistic correlations rather than multi-modal reasoning. For instance, simply answering "tennis" for the sport-related questions and "yes" to the questions "Do you see a ..." can achieve approximately 40% and 90% accuracy on the VQA v1.0 dataset, respectively. As a result, VQA models will fail to generalize well if they simply memorize the strong language priors in the training data [2, 50, 19], especially on the recently proposed VQA-CP [3] dataset where the priors are quite different in the training and test sets.

One straightforward solution to mitigate language bias is to enhance the training data by using extra annotations or data augmentation. Specifically, visual [14] and textual [22] explanations are used to improve the visual grounding ability [36, 43]. Besides, counterfactual training samples generation [11, 1, 52, 18, 27] helps to balance the training data, and outperform other debiasing methods by large margins on VQA-CP. These methods demonstrate the effect of *debiased training* to improve the generalizability of VQA models. It is worth noting that VQA-CP was proposed to validate whether VQA models can disentangle the learned visual knowledge and memorized language priors. Therefore, how to make unbiased inference under *biased training* still remains a major challenge. Another popular solution [10, 13] is using a separate question-only branch to learn the language prior. During the test stage, the prior is mitigated by excluding the extra branch. However, we argue that the language prior consists of both "bad" language bias (*e.g.*, binding the color of bananas with the major color "yellow") and "good" language context (*e.g.*, narrowing the answer space based on the question type "what color"). Simply excluding the question-only branch cannot make use of the good context. Indeed, it is still challenging for recent methods to disentangle the good and bad from the whole.

Motivated by causal effects [28, 29, 30], we propose a novel counterfactual inference framework called CF-VQA to reduce language bias in VQA. Overall, we formulate language bias as the direct causal effect of questions on answers, and mitigate the bias by subtracting the direct language effect from the total causal effect. As illustrated in Figure 1, we introduce two scenarios, conventional VQA and counterfactual VQA, to estimate the causal effects, which are defined as follows:

Conventional VQA: *What will answer A be, if machine hears question Q, sees image V, and extracts the multi-modal knowledge K?*

Counterfactual VQA: *What would A be, if machine hears Q, but had not extracted K or seen V?*

Intuitively, conventional VQA depicts the scenario where both Q and V are available. In this case, we can estimate the total causal effect of V and Q on A. However, conventional VQA cannot disentangle the single-modal spurious
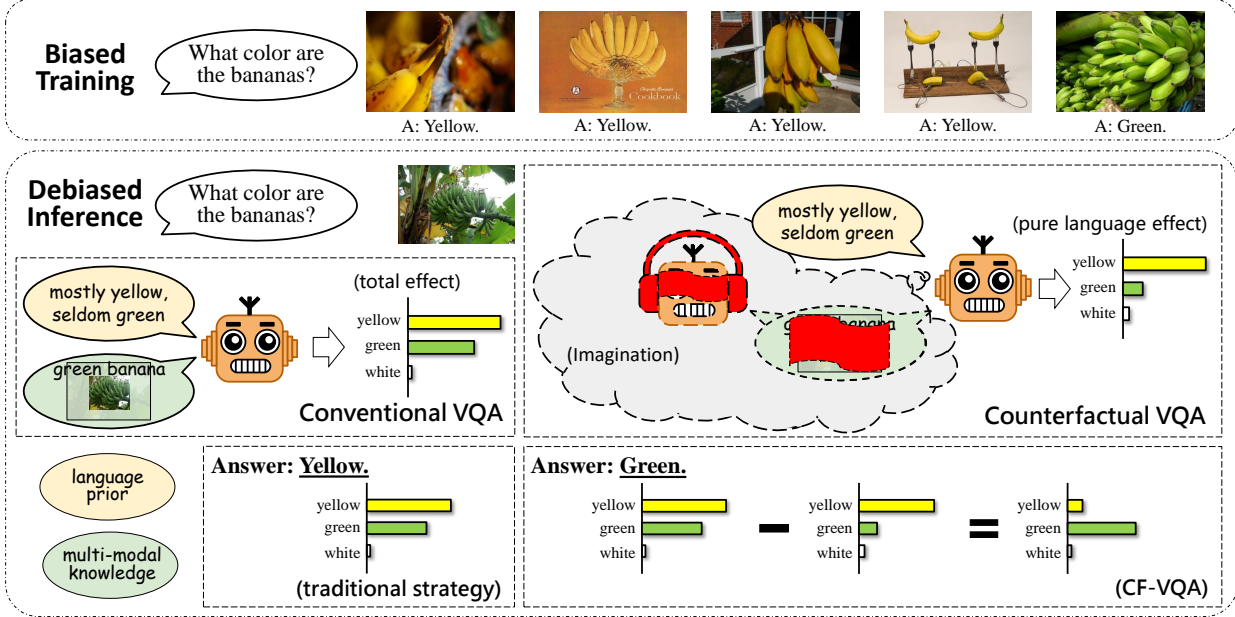
Figure 1: Our cause-effect look at language bias in VQA. Conventional VQA depicts the fact where machine hears the question and extracts the multi-modal knowledge. Counterfactual VQA depicts the scenario where machine hears the question but the knowledge is blocked. We subtract the pure language effect from the total effect for debiased inference.

linguistic correlation and multi-modal reasoning, *i.e.*, direct and indirect effects. Therefore, we consider the following counterfactual question: *"What would have happened if the machine had not performed multi-modal reasoning?"* The answer to this question can be obtained by imagining a scenario where the machine hears Q, but the multi-modal knowledge K is intervened under the no-treatment condition, *i.e.*, V and Q had not been accessible. Since the response of K to Q is blocked, VQA models can only rely on the single-modal impact. Therefore, language bias can be identified by estimating the direct causal effect of Q on A, *i.e.*, pure language effect. The training stages follows language-prior based methods [10, 13] that train an ensemble model with a prevailing VQA model and single-modal branches. During the test stage, CF-VQA uses the debiased causal effect for inference, which is obtained by subtracting the pure language effect from the total effect. Note that our inference strategy is totally different from traditional strategies that use the posterior possibility. Perhaps surprisingly, recent language-prior based methods [10, 13] can be further unified into our proposed counterfactual inference framework as special cases. Experimental results show that CF-VQA outperforms methods without data argumentation approaches by large margins on the VQA-CP dataset [3], and remains stable on the balanced VQA v2 dataset [19].

The contribution of this paper is threefold. First, our counterfactual inference framework is the first to formulate language bias in VQA as causal effects. Second, we provide a novel causality-based interpretation for recent debi-

asing VQA works [10, 13], and easily improve RUBi [10] by 7.5% with only one more learnable parameter. Third, our cause-effect look is general and suitable for different baseline VQA architectures and fusion strategies.

2. Related Work

Language Bias in VQA can be interpreted in two ways. First, there exists strong correlations between questions and answers, which reflects the *"language prior"* [19, 3]. Simply answering "tennis" to the sport-related questions can achieve approximately 40% accuracy on VQA v1.0 dataset. Second, the questioner tends to ask about the objects seen in the image, which leads to the *"visual priming bias"* [8, 19, 24]. Simply answering "yes" to all the questions "Do you see a ..." achieves nearly 90% accuracy on VQA v1.0 dataset. In both ways, machines may merely focus on the question rather than the visual content. As a result, this serious shortcut limits the generalization of VQA models [2, 50, 19], especially when the test scenario is quite different from that in training.

Debiasing Strategies in VQA. Recently, a new VQA split, Visual Question Answering under Changing Priors (VQA-CP) [3], was proposed to evaluate the generalizability of VQA models, where the distributions of answers for every question type are different during training and test stages. Most of recent solutions to reduce the language bias in VQA can be grouped into three categories, strengthening visual grounding [36, 43], weakening lan-

guage prior [32, 10, 13], and implicit/explicit data argumentation [11, 1, 39, 52]. First, human visual [14] and textual [22] explanations are exploited to strengthen the visual grounding in VQA [36, 43]. The VQA-HAT dataset collects human attention map as visual attention supervision [14], while the VQA-X dataset contains textual explanations related to question-answer pairs. Second, ensemble-based methods proposed to use a separated QA branch to capture the language prior under adversarial learning [32] or multi-task learning [10, 13]. Third, the success of data argumentation inspires recent works to automatically generate additional question-image pairs [11, 1] and balance the distribution of training data. Some other works investigate how to improve the robustness of VQA models via question representation learning [26] or language attention mechanism [52]. In this paper, the language-prior based methods [10, 13] can be unified into our proposed counterfactual inference framework as special cases.

Causality-inspired Computer Vision. Counterfactual thinking and causal inference have inspired several studies in computer vision, including visual explanations [20, 7, 41, 46], scene graph generation [12, 37], image recognition [37], video analysis [16, 25], few-shot learning [47], representation learning [42, 51], semantic segmentation [49], and vision-language tasks [11, 38, 31, 44, 17]. Especailly, counterfactual learning has been exploited in recent VQA studies [11, 38, 1]. Different from these works that generate counterfactual samples for *debiased* training, our cause-effect look focuses on counterfactual inference with even *biased* training data. Therefore, counterfactual-learning based works are orthogonal to ours.

3. Preliminaries

In this section, we introduce the used concepts of causal inference [28, 29, 34, 30]. In the following, we represent a random variable as a capital letter (*e.g.*, X), and denote its observed value as a lowercase letter (*e.g.*, x).

Causal graph reflects the causal relationships between variables, which is represented as a directed acyclic graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V}$ denotes the set of variables and $\mathcal{E}$ represents the cause-and-effect relationships. Figure 2(a) shows an example of causal graph consisting of three variables. If the variable X has a *direct* effect on the variable Y, we say that Y is the child of X, *i.e.*, $X \rightarrow Y$. If X has an *indirect* effect on Y via the variable M, we say that M acts as a *mediator* between X and Y, *i.e.*, $X \rightarrow M \rightarrow Y$.

Counterfactual notations are used to translate causal assumptions from graphs to formulas. The value that Y would obtain if X is set to x and M is set to m is denoted as:

$$Y_{x,m} = Y(X = x, M = m)^1 \qquad (1)$$

---

¹If there is no confounder of X, then we have that $do(X = x)$. equals to $X = x$ and can omit the *do* operator.
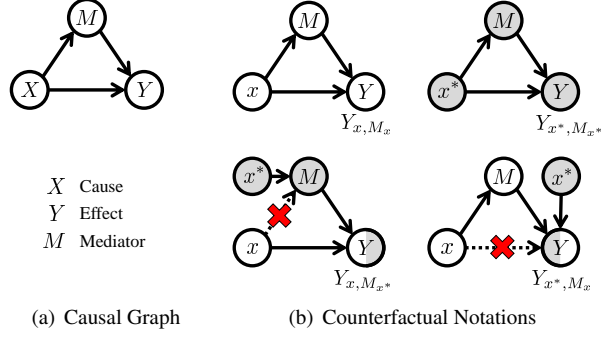

(a) Causal Graph (b) Counterfactual Notations

Figure 2: (a) Example of causal graph. (b) Examples of counterfactual notations. White nodes are at the value $X = x$ while gray nodes are at $X = x^*$.

In the factual scenario, we have $m = M_x = M(X = x)$. In the counterfactual scenario, X is set as different values for M and Y. For example, $Y_{x,M_{x^*}}$ describes the situation where X is set to x and M is set to the value when X had been x^*, *i.e.*, $Y_{x,M_{x^*}} = Y(X = x, M = M(X = x^*))$. Note that X can be simultaneously set to different values x and x^* only in the counterfactual world. Figure 3(b) illustrates examples of counterfactual notations.

Causal effects are the comparisons between two potential outcomes of the same individual given two different treatments [35, 33]. Supposed that $X = x$ represents "under treatment condition" and $X = x^*$ represents "under no-treatment condition"². The total effect (TE) of treatment $X = x$ on Y compares two hypothetical situations $X = x$ and $X = x^*$, which is denoted as:

$$TE = Y_{x,M_x} - Y_{x^*,M_{x^*}}. \qquad (2)$$

Total effect can be decomposed into natural direct effect (NDE) and total indirect effect (TIE). NDE denotes the effect of X on Y with the mediator M blocked. NDE expresses the increase in the outcome Y with X changing from x^* to x, while M is set to the value it *would have obtained* at $X = x^*$, *i.e.*, the ability of M to respond to the treatment $X = x$ is disabled:

$$NDE = Y_{x,M_{x^*}} - Y_{x^*,M_{x^*}}. \qquad (3)$$

TIE is the difference between TE and NDE, denoted as:

$$TIE = TE - NDE = Y_{x,M_x} - Y_{x,M_{x^*}}, \qquad (4)$$

TE can be also decomposed into natural indirect effect (NIE) and total direct effect (TDE). Similarly, NIE reflects the effect of X on Y through the mediator M, *i.e.*, $X \rightarrow$

---

²For example, to estimate the effect of a drug on a disease, $X = x$ represents taking the drug, while $X = x^*$ represents not taking the drug.

$M \rightarrow Y$, while the direct effect on $X \rightarrow Y$ is blocked by setting X as x^*. NIE is denoted as:

$$NIE = Y_{x^*, M_x} - Y_{x^*, M_{x^*}}. \tag{5}$$

In Section 4, we will further discuss the meanings and differences of these effects in VQA.

4. Cause-Effect Look at VQA

Following the common formulation, we define the VQA task as a multi-class classification problem. VQA models are required to select an answer from the candidate set $\mathcal{A} = \{a\}$ given an image $V = v$ and a question $Q = q$.

4.1. Cause-Effect Look

The causal graph of VQA is illustrated in Figure 3(a). The effect of V and Q on A can be divided into single-modal impact and multi-modal impact. The single-modal impact captures the *direct* effect of V or Q on A via $V \rightarrow A$ or $Q \rightarrow A$. The multi-modal impact captures the *indirect* effect of V and Q on A via the multi-modal knowledge K, *i.e.*, $V, Q \rightarrow K \rightarrow A$. We propose to exclude pure language effect on $Q \rightarrow A$ to reduce language bias.

Following the counterfactual notation in Eq. (1), we denote the score that the answer a (*e.g.*, "green") would obtain if V is set to v (*e.g.*, an image which includes green bananas) and Q is set to q (*e.g.*, a question "What color are the bananas?") as

$$Y_{v,q}(a) = Y(a; V = v, Q = q).$$

Without loss of generality, we omit a for simplicity, *i.e.*, $Y_{v,q} = Y(V = v, Q = q)$. Similarly, the counterfactual notation of K is denoted as $K_{v,q} = K(V = v, Q = q)$.

As shown in Figure 3(a), there exist three paths directly connected to A, *i.e.*, $Q \rightarrow A$, $V \rightarrow A$, and $K \rightarrow A$. Therefore, we can rewrite $Y_{v,q}$ as the function of Q, V and K:

$$Y_{v,q} = Z_{q,v,k} = Z(Q = q, V = v, K = k), \tag{6}$$

where $k = K_{v,q}$. Following the definition of causal effects in Section 3, the *total effect* (TE) of $V = v$ and $Q = q$ on $A = a$ can be written as:

$$TE = Y_{v,q} - Y_{v^*,q^*} = Z_{q,v,k} - Z_{q^*,v^*,k^*}, \tag{7}$$

where $k^* = K_{v^*,q^*}$. Here v^* and q^* represent the no-treatment condition where v and q are not given.

As we have discussed in Section 1, VQA models may suffer from the spurious correlation between questions and answers, and thus fail to conduct effective multi-modal reasoning. Therefore, we expect VQA models to exclude the direct impact of questions. To achieve this goal, we proposed counterfactual VQA to estimate the causal effect of
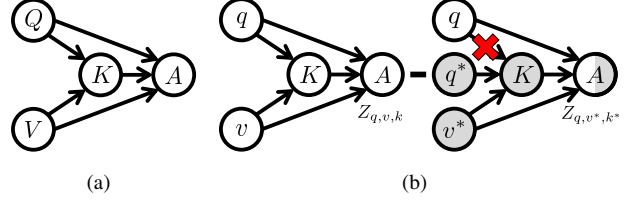

Figure 3: (a) Causal graph for VQA. Q: question. V: image. K: multi-modal knowledge. A: answer. (b) Comparison between conventional VQA and counterfactual VQA. White nodes are at the value $V = v$ and $Q = q$ while gray nodes are at the value $V = v^*$ and $Q = q^*$.

$Q = q$ on $A = a$ by blocking the effect of K and V. Counterfactual VQA describes the scenario where Q is set to q and K would attain the value k^* when Q had been q^* and V had been v^*. Since the response of mediator K to inputs is blocked, the model can only rely on the given question for decision making. Figure 3(b) shows the comparison between conventional VQA and counterfactual VQA. We obtain the *natural direct effect* (NDE) of Q on A by comparing counterfactual VQA to the no-treatment conditions:

$$NDE = Z_{q,v^*,k^*} - Z_{q^*,v^*,k^*}. \tag{8}$$

Since the effect of Q on the intermediate K is blocked (*i.e.*, $K = k^*$), NDE explicitly captures the language bias. Furthermore, the reduction of language bias can be realized by subtracting NDE from TE, which is represented as:

$$TIE = TE - NDE = Z_{q,v,k} - Z_{q,v^*,k^*}. \tag{9}$$

We select the answer with the maximum TIE for inference, which is totally different from traditional strategies that is based on the posterior probability *i.e.*, $P(a|v,q)$.

4.2. Implementation

Parameterization. The calculation of the score $Z_{q,v,k}$ in Eq. (6) is parameterized by three neural models $\mathcal{F}_Q$, $\mathcal{F}_V$, $\mathcal{F}_{VQ}$ and one fusion function h as:

$$Z_q = \mathcal{F}_Q(q), \quad Z_v = \mathcal{F}_V(v), \quad Z_k = \mathcal{F}_{VQ}(v,q),$$
$$Z_{q,v,k} = h(Z_q, Z_v, Z_k), \tag{10}$$

where $\mathcal{F}_Q$ is the language-only branch (*i.e.*, $Q \rightarrow A$), $\mathcal{F}_V$ is the vision-only branch (*i.e.*, $V \rightarrow A$), and $\mathcal{F}_{VQ}$ is the vision-language branch (*i.e.*, $V, Q \rightarrow K \rightarrow A$). The output scores are fused by h to obtain the final score $Z_{q,v,k}$.

As introduced in Section 4.1, the no-treatment condition is defined as blocking the signal from vision or language, *i.e.*, v or q is not given. We represent the no-treatment conditions as $V = v^* = \varnothing$ and $Q = q^* = \varnothing$. Note that neural models cannot deal with no-treatment conditions where
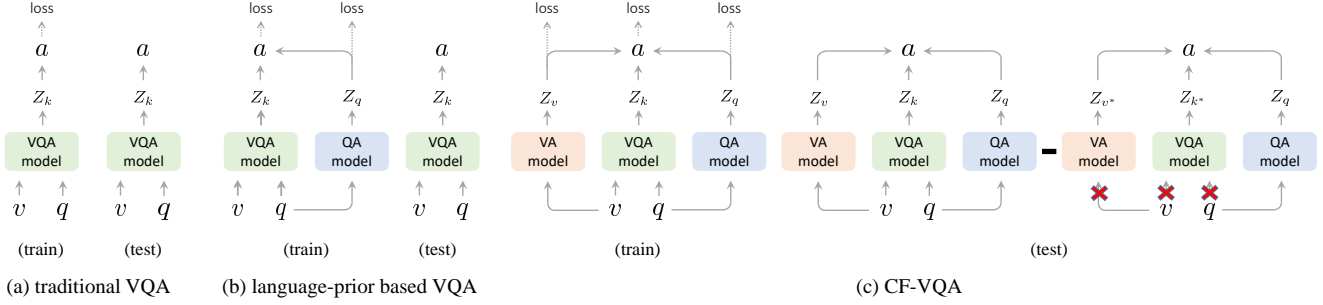
Figure 4: Comparison between our CF-VQA and other VQA methods. (a) Traditional VQA methods use a single VQA model. (v) Language-prior based methods [10, 13] use an additional QA model to capture the language prior during the training stage. The QA model is not used during testing. (c) Our proposed CF-VQA maintains the QA model during the test stage, and makes inference based on the debiased causal effect.

the inputs are void. Therefore, we assume that the model will randomly guess with equal probability under the no-treatment conditions. Therefore, Z_q, Z_v and Z_k in Eq. (10) can be represented as:

$$Z_q = \begin{cases} z_q = \mathcal{F}_Q(q) & \text{if } Q = q \\ z_q^* = c & \text{if } Q = \varnothing \end{cases}, \quad (11)$$

$$Z_v = \begin{cases} z_v = \mathcal{F}_V(v) & \text{if } V = v \\ z_v^* = c & \text{if } V = \varnothing \end{cases}, \quad (12)$$

$$Z_k = \begin{cases} z_k = \mathcal{F}_{VQ}(v,q) & \text{if } V = v \text{ and } Q = q \\ z_k^* = c & \text{if } V = \varnothing \text{ or } Q = \varnothing \end{cases}, \quad (13)$$

where c denotes a learnable parameter. We use the uniform distribution assumption for two reasons. Firstly, as for us human, we would like to make a wild guess if we have absolutely no idea about the specific treatments, including question types or topics. Secondly, as z_{v^*} and z_{k^*} are used to estimated NDE of Q, the uniform distribution can guarantee a safe estimation. We further empirically validate the uniform distribution assumption in the ablation study.

Fusion Strategies. We expect that the fused score $Z_{q,v,k}$ is a combination of Z_q, Z_v and Z_k. Specifically, we proposed two non-linear fusion variants, Harmonic (HM) and SUM:

$$\text{(HM)} \qquad h(Z_q, Z_v, Z_k) = \log \frac{Z_{\text{HM}}}{1 + Z_{\text{HM}}}, \quad (14)$$

where $Z_{\text{HM}} = \sigma(Z_q) \cdot \sigma(Z_v) \cdot \sigma(Z_k)$.

$$\text{(SUM)} \qquad h(Z_q, Z_v, Z_k) = \log \sigma(Z_{\text{SUM}}), \quad (15)$$

where $Z_{\text{SUM}} = Z_q + Z_v + Z_k$.

Training. The training strategy follows [10]. As illustrated in Figure 4(c), given a triplet (v, q, a) where a is the ground-truth answer of image-question pair (v, q), the branches are

optimized by minimizing the cross-entropy losses over the scores $Z_{q,v,k}$, Z_q and Z_v:

$$\mathcal{L}_{cls} = \mathcal{L}_{VQA}(v,q,a) + \mathcal{L}_{QA}(q,a) + \mathcal{L}_{VA}(v,a), \quad (16)$$

where $\mathcal{L}_{VQA}$, $\mathcal{L}_{QA}$ and $\mathcal{L}_{VA}$ are over $Z_{q,v,k}$, Z_q and Z_v. Note that we introduce a learnable parameter c in Eq. (11)-(13), which controls the sharpness of the distribution of Z_{q,v^*,k^*} like the softmax temperature [21]. We hypothesize that the sharpness of NDE should be similar to that of TE. An improper c would lead to the result that TIE in Eq. (4) is dominated by TE or NDE. Therefore, we use a Kullback-Leibler divergence to estimate c:

$$\mathcal{L}_{kl} = \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} -p(a|q,v,k) \log p(a|q,v^*,k^*), \quad (17)$$

where $p(a|q,v,k) = \text{softmax}(Z_{q,v,k})$ and $p(a|q,v^*,k^*) = \text{softmax}(Z_{q,v^*,k^*})$. Only c is updated by minimizing $\mathcal{L}_{kl}$. The final loss is the combination of $\mathcal{L}_{cls}$ and $\mathcal{L}_{kl}$:

$$\mathcal{L} = \sum_{(v,q,a) \in \mathcal{D}} \mathcal{L}_{cls} + \mathcal{L}_{kl} \quad (18)$$

Inference. As discussed in Section 4.1, we use the debiased effect for inference, which is implemented as:

$$\begin{aligned} TIE = TE - NDE &= Z_{q,v,k} - Z_{q,v^*,k^*} \\ &= h(z_q, z_v, z_k) - h(z_q, z_v^*, z_k^*). \end{aligned} \quad (19)$$

4.3. Revisiting RUBi and Learned-Mixin

Note that language-prior based methods RUBi [10] and Learned-Mixin [13] use an ensemble model that consists a vision-language branch $\mathcal{F}_{VQ}$ and a question-only branch $\mathcal{F}_Q$. Besides, they simply exclude $\mathcal{F}_Q$ and use $z_k = \mathcal{F}_{VQ}(v,q)$ for answer selection. The conceptual comparison between our CF-VQA and language-prior based methods methods are illustrated in Figure 4. These methods

Table 1: **Comparison on VQA-CP v2 test set and VQA v2 val set**. <u>Best</u> and **second best** results obtained without extra generated training samples are highlighted in each column. "Base." indicates the VQA base model. We report the average accuracy of CF-VQA over 5 experiments with different random seeds.

| Test set		VQA-CP v2 test				VQA v2 val			
Methods	Base.	All	Y/N	Num.	Other	All	Y/N	Num.	Other
GVQA [3]	–	31.30	57.99	13.68	22.14	48.24	72.03	31.17	34.65
SAN [45]	–	24.96	38.35	11.14	21.74	52.41	70.06	39.28	47.84
UpDn [6]	–	39.74	42.27	11.93	46.05	63.48	81.18	42.14	<u>55.66</u>
S-MRL [10]	–	38.46	42.85	12.81	43.20	63.10	–	–	–
methods based on modifying language module follow:									
DLR [23]	UpDn	48.87	70.99	18.72	45.57	57.96	76.82	39.33	48.54
VGQE [26]	UpDn	48.75	–	–	–	<u>64.04</u>	–	–	–
VGQE [26]	S-MRL	50.11	66.35	**27.08**	46.77	63.18	–	–	–
methods based on strengthening visual attention follow:									
AttAlign [36]	UpDn	39.37	43.02	11.89	45.00	63.24	80.99	42.55	55.22
HINT [36]	UpDn	46.73	67.27	10.61	45.88	63.38	81.18	42.99	**55.56**
SCR [43]	UpDn	49.45	72.36	10.93	<u>48.02</u>	62.2	78.8	41.6	54.5
methods based on weakening language prior follow:									
AdvReg. [32]	UpDn	41.17	65.49	15.48	35.48	62.75	79.84	42.35	55.16
RUBi [10]	UpDn	44.23	67.05	17.48	39.61	–	–	–	–
RUBi [10]	S-MRL	47.11	68.65	20.28	43.18	61.16	–	–	–
LM [13]	UpDn	48.78	72.78	14.61	45.58	63.26	81.16	42.22	55.22
LM+H [13]	UpDn	52.01	72.58	<u>31.12</u>	**46.97**	56.35	65.06	37.63	54.69
CF-VQA (HM)	UpDn	$49.74^{\pm0.26}$	$74.81^{\pm0.51}$	$18.46^{\pm0.39}$	$45.19^{\pm0.27}$	$\underline{63.73}^{\pm0.07}$	$82.15^{\pm0.08}$	$\underline{44.29}^{\pm0.54}$	$54.86^{\pm0.07}$
CF-VQA (HM)	S-MRL	$51.27^{\pm0.25}$	$77.80^{\pm0.83}$	$20.64^{\pm1.01}$	$45.76^{\pm0.21}$	$62.49^{\pm0.06}$	$81.19^{\pm0.09}$	$\underline{44.64}^{\pm0.20}$	$52.98^{\pm0.14}$
CF-VQA (SUM)	UpDn	$\mathbf{53.55}^{\pm0.10}$	$\mathbf{91.15}^{\pm0.06}$	$13.03^{\pm0.21}$	$44.97^{\pm0.20}$	$63.54^{\pm0.09}$	$\mathbf{82.51}^{\pm0.12}$	$43.96^{\pm0.17}$	$54.30^{\pm0.09}$
CF-VQA (SUM)	S-MRL	$\underline{55.05}^{\pm0.12}$	$90.61^{\pm0.28}$	$21.50^{\pm0.86}$	$45.61^{\pm0.16}$	$60.94^{\pm0.15}$	$81.13^{\pm0.15}$	$43.86^{\pm0.40}$	$50.11^{\pm0.16}$
methods based on balancing training data follow:									
CVL [1]	UpDn	42.12	45.72	12.45	48.34	–	–	–	–
Unshuffling [39]	UpDn	42.39	47.72	14.43	47.24	61.08	78.32	42.16	52.81
RandImg [40]	UpDn	55.37	83.89	41.60	44.20	57.24	76.53	33.87	48.57
SSL [52]	UpDn	57.59	86.53	29.87	50.03	63.73	–	–	–
CSS [11]	UpDn	58.95	84.37	49.42	48.21	59.91	73.25	39.77	55.11
CSS+CL [27]	UpDn	59.18	86.99	49.89	47.16	57.29	67.27	38.40	54.71
Mutant [18]	UpDn	61.72	88.90	49.68	50.78	62.56	82.07	42.52	53.28
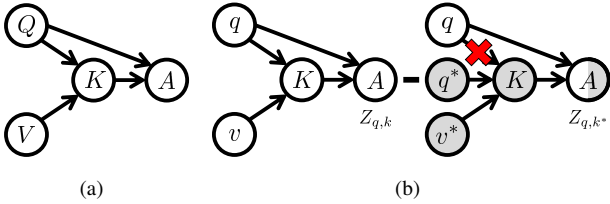
Figure 5: (a) Simplified VQA causal graph. (b) Comparison between conventional and counterfactual VQA.

can be unified into our counterfactual inference framework, which (1) follow a simplified causal graph (Fig. 5(a)) without the direct path $V \rightarrow A$, and (2) use natural indirect effect (NIE) in Eq. (5) for inference. The detailed analysis is provided in the supplementary materials.

5. Experiments

We mainly conduct the experiments on the VQA-CP [3] dataset. VQA-CP is proposed to evaluate the robustness of VQA models when the answer distributions of training and test splits are significantly different. In addition, we also report the performance on the balanced VQA v2 dataset to see whether the approach over-corrects language bias. The models are evaluated via accuracy. For fair comparisons, we conduct experiments with three baseline VQA architectures: Stacked Attention Network (**SAN**) [45], Bottom-up and Top-down Attention (**UpDn**) [5], and a simplified version of MUREL [9] (**S-MRL**) [10].

5.1. Quantitative Results

We first compare CF-VQA with state-of-the-art methods. Recent approaches can be grouped as follows. (1) Methods that *modify language modules* proposed to decouple the linguistic concepts (**DLR**) [23] or generate visually-grounded question representations (**VGQE**) [26]. (2) Methods that *strengthen visual attention* exploit human visual [14] or textual [22] explanations, including **AttAlign** [36], **HINT** [36] and **SCR** [43]. (3) Methods that *weaken language prior* proposed to directly formu-

Table 2: **Ablation of CF-VQA** on VQA-CP v2 test set. "SAN/UpDn/S-MRL" denotes the baseline VQA model. "HM/SUM" represents the strategies that train the ensemble model and test with only the vision-language branch following ensemble-based method [10, 13]. * represents the reproduced results.

	All	Y/N	Num.	Other		All	Y/N	Num.	Other		All	Y/N	Num.	Other
SAN*	33.18	38.57	12.25	36.10	UpDn*	37.69	43.17	12.53	41.72	S-MRL*	37.09	41.39	12.46	41.60
HM	45.89	70.37	**23.99**	39.07	HM	47.97	69.19	**18.80**	44.86	HM	49.37	73.20	20.10	44.92
+ CF-VQA	**48.10**	**77.68**	22.19	**39.71**	+ CF-VQA	**49.74**	**74.81**	18.46	**45.19**	+ CF-VQA	**51.27**	**77.80**	**20.64**	**45.76**
SUM	43.98	68.98	**17.32**	38.19	SUM	47.29	72.26	12.54	43.74	SUM	48.27	74.60	20.96	41.96
+ CF-VQA	**50.15**	**87.95**	16.46	**39.59**	+ CF-VQA	**53.55**	**91.15**	**13.03**	**44.97**	+ CF-VQA	**55.05**	**90.61**	**21.50**	**45.61**

Table 3: **Ablation of CF-VQA with the simplified causal graph** on VQA-CP v2 test set. "SAN/UpDn/S-MRL" denotes the baseline VQA model. "HM/SUM" represents the strategies that train the ensemble model and test with only the vision-language branch following ensemble-based method [10, 13]. * represents the reproduced results.

	All	Y/N	Num.	Other		All	Y/N	Num.	Other		All	Y/N	Num.	Other
SAN*	33.18	38.57	12.25	36.10	UpDn*	37.69	43.17	12.53	41.72	S-MRL*	37.09	41.39	12.46	41.60
HM	45.48	71.32	17.19	39.71	HM	46.50	67.54	12.83	44.72	HM	49.57	72.31	20.28	45.68
+ CF-VQA	**49.43**	**83.82**	**17.52**	**40.16**	+ CF-VQA	**49.53**	**77.02**	**12.86**	**45.18**	+ CF-VQA	**52.68**	**82.05**	**20.76**	**46.04**
SUM	42.35	62.33	**16.64**	38.94	SUM	47.10	70.00	12.80	44.51	SUM	49.42	74.43	20.52	44.24
+ CF-VQA	**49.85**	**87.75**	16.15	**39.24**	+ CF-VQA	**53.55**	**91.15**	**12.81**	**45.02**	+ CF-VQA	**54.52**	**90.69**	**21.84**	**44.53**

Table 4: The comparison between CF-VQA and RUBi.

	VQA-CP v2 test				VQA v2 val
	All	Y/N	Num.	Other	All
S-MRL [10]	38.46	42.85	12.81	**43.20**	**63.10**
RUBi [10]	47.11$^{\pm0.51}$	68.65	20.28	43.18	61.16
+ CF-VQA	**54.69**$^{\pm0.98}$	**89.90**	**32.39**	42.01	60.53

Table 5: **Ablation of assumptions for counterfactual outputs** on VQA-CP v2 test set.

		All	Y/N	Num.	Other
S-MRL [10]		38.46	42.85	12.81	43.20
HM	random	31.27	29.69	**42.87**	28.91
	prior	46.29	61.88	20.03	45.33
	uniform	**51.27**	**77.80**	20.64	**45.76**
SUM	random	27.52	28.00	**37.88**	24.42
	prior	38.06	41.43	14.90	42.64
	uniform	**55.05**	**90.61**	21.50	**45.61**

late the language prior by a separate question-only branch, including **AdvReg.** [32], **RUBi** [10] and **Learned-Mixin (LM)** [13]. Note that *RUBi is the most related work to us*. (4) Methods that ***balance training data*** proposed to change the training distribution for unbiased training, including **CVL** [1], **Unshuffling** [39], **RandImg** [40], **SSL** [52], **CSS** [11], **CL** [27], and **Mutant** [18]. Unshuffling [39] partitions the training set into multiple invariant subsets. Other methods generate counterfactual training samples by masking or transforming critical words and objects [11, 18] or replacing the image with a different one [1, 40, 52].

The results on VQA-CP v2 and VQA v2 are reported in Table 1. Most of the methods that explicitly generate training samples [40, 52, 11, 27, 18] outperform others by large margins. However, these methods explicitly change the training priors, which violates the original intention of VQA-CP, *i.e.*, to evaluate whether VQA models are driven by memorizing priors in training data [3]. Therefore, we do not directly compare CF-VQA with these methods for fairness. Overall, compared to non-augmentation approaches, our proposed CF-VQA achieves a new state-of-the-art performance on VQA-CP v2 dataset. With a deep look at the question type on VQA-CP, we find that the improvement on "Yes/No" questions is extremely large (from ∼70% to ∼90%), which indicates that language bias have different effects on different types of questions. Besides, methods

with extra annotations or generated training samples effectively improves the accuracy on "Other" questions, while CF-VQA achieves comparable performance to other methods. It is worth noting that LM [13] achieves a competitive performance on VQA-CP v2 test set with an additional language entropy penalty (LM+H). However, the accuracy drops significantly by ∼7% on VQA v2, which indicates that the entropy penalty forces the model to over-correct language bias, especially on "Yes/No" questions. As a comparison, CF-VQA perform robustly on VQA v2, which indicates that CF-VQA does not over-correct language bias.

As introduced in Section 4.3, RUBi [10] and Learned-Mixin [13] can be unified into our counterfactual inference framework. Based on our cause-effect look, CF-VQA can improve RUBi by replacing NIE with TIE and introducing only one learnable parameter c. The details are provided in the supplementary. As shown in Table 4, CF-VQA improves RUBi by over 7.5% (*i.e.*, from 47.11 to 54.69) on VQA-CP v2 test split, while the accuracy slightly drops on VQA v2 val split. Compared to our proposed symmetric fusion strategies Harmonic and SUM, the standard deviation of RUBi's overall accuracy is larger, which indicates that
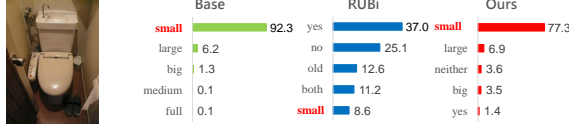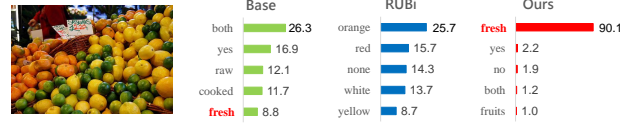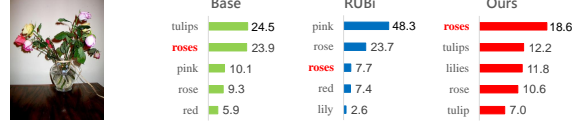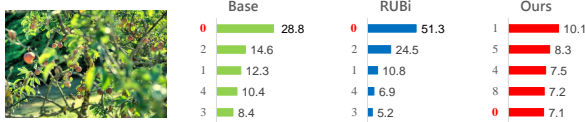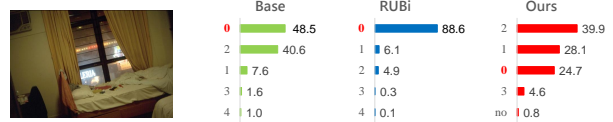
Figure 6: **Qualitative comparison** on VQA-CP v2 test split. Red bold answer denotes the ground-truth one.
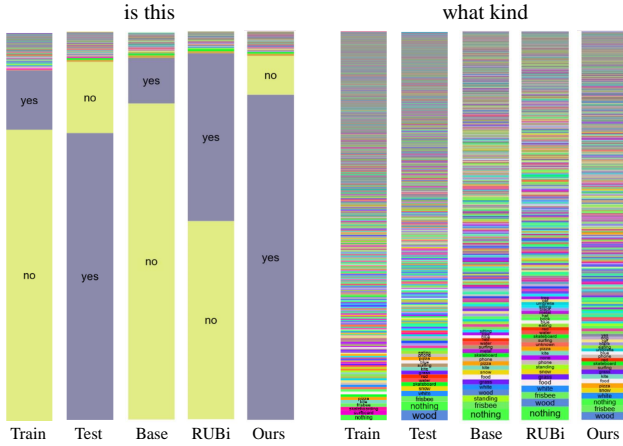

Figure 7: **Answer distributions** on VQA-CP v2.

our proposed symmetric fusion strategies are more stable and robust in implementation.

We further conduct ablation studies to validate (1) the generalizability of CF-VQA to both baseline VQA architectures, fusion strategies and causal graphs, and (2) the uniform distribution assumption for the no-treatment conditions. As shown in Table 2 and 3, Cf-VQA outperforms the ensemble-based strategies by over 2% for Harmonic and over 5% for SUM in all cases. To empirically validate the uniform distribution assumption, we proposed two candidate assumptions. "Random" represents the assumption that c_i for answer a_i are learned without any constraint. "Prior" denotes the assumption that $\{c_i\}$ obey the prior distribution of the training set. As shown in Table 5, "random"

and "prior" even perform worse than the baselines. As we discussed in Section 4.2, the reason might be that the uniform distribution assumption guarantees a safe estimation of NDE, *i.e.*, the biased linguistic causal effect.

5.2. Qualitative Results

The qualitative results are provided to validate whether CF-VQA can effectively reduce language bias and retain language context. As illustrated in Figure 6, CF-VQA can successfully overcome language bias on yes/no questions compared to RUBi, while the baseline model suffers from the memorized language prior on the training set. Besides, for "what kind" questions, RUBi prefers the meaningless answer "none" rather than specific ones. Although CF-VQA cannot recover the answer distribution very well, it attempts to respond with more meaningful answers (*e.g.*, wood, frisbee). Examples in Figure 6 further illustrate how CF-VQA preserves language context for inference. For the left top example, CF-VQA recognizes the correct context "large or small", while RUBi tends to answer yes/no based on the wrong context "is this". For the right example at the second row, although RUBi successfully locates the flowers, it wrongly focuses on visual attributes (*i.e.*, pink) rather than categories (*i.e.*, "what type"). The examples at the third row show the failure cases on number-related questions. These results highlight the importance of language context, which is not considered by language prior based approaches.

6. Conclusion

In this paper, we proposed a novel counterfactual inference framework CF-VQA to reduce language bias in VQA.

The bias is formulated as the direct causal effect of questions on answers, and is further captured by a counterfactual scenario. The reduction of language bias is realized by subtract the direct linguistic effect from the total causal effect. Experimental results demonstrate the effectiveness and generalizability of our proposed inference strategy CF-VQA. Furthermore, recent debiasing studies [10, 13] can be unified into our proposed counterfactual inference framework. Based on our cause-effect look, We can further improve RUBi [10] by simply changing a few lines of code and including only one more learnable parameter.

Supplementary Material

This supplementary document is organized as follows:

7. Revisiting RUBi and Learned-Mixin

As mentioned in Section 4.3, RUBi [10] and Learned-Mixin [13] can be unified into our counterfactual inference framework, which (1) follow a simplified causal graph without the direct path $V \to A$, and (2) use natural indirect effect (NIE) for inference. The detailed analysis is provided as follows.

7.1. Cause-Effect Look

Recent works RUBi [10] and Learned-Mixin [13] apply an ensemble architecture with a vision-language branch $\mathcal{F}_{\mathcal{VQ}}$ and a question-only branch $\mathcal{F}_{\mathcal{Q}}$, while the direct relation between vision and answer is not formulated. The architecture is shown in Figure 8 (a).

Note that total effect can be decomposed into natural direct effect (NDE) and total indirect effect (TIE). As introduced in the main paper, we remove language bias by subtracting the natural direct effect from the total effect. The TIE is calculated by:

$$
\begin{aligned}
TE &= Z_{q,k} - Z_{q^*,k^*}, \\
NDE &= Z_{q,k^*} - Z_{q^*,k^*}, \\
TIE &= TE - NDE = Z_{q,k} - Z_{q,k^*},
\end{aligned}
\tag{20}
$$

which corresponds to Eq. (4) in the main paper. An alternative option to reduce language bias is to substract the

Algorithm 1 Improving RUBi [10] using CF-VQA

```
 1: function RUBI(v, q, is_Training; θ, c)
 2:     z_q = F_Q(q)
 3:     z_k = F_VQ(v, q)
 4:     if is_Training then
 5:         z = z_k · σ(z_q)
 6:         updating θ according to L_cls
 7:         updating c according to L_kl
 8:     else
 9:         z ̶=̶ ̶z̶_̶k̶  z = (z_k − c) · σ(z_q)
10:     end if
11:     return z
12: end function
```

total direct effect (TDE) of questions on answers from total effect, which is formulated as:

$$
\begin{aligned}
TDE &= Z_{q,k} - Z_{q^*,k}, \\
NIE &= TE - TDE = Z_{q^*,k} - Z_{q^*,k^*}.
\end{aligned}
\tag{21}
$$

Intuitively, both TIE and NIE reflect the increase of confidence for the answer given the visual knowledge, *i.e.*, from k^* to k. The difference between TIE and NIE is the existence of question q. The question q is block to calculate NIE (*i.e.*, q^*), while q is given to calculate TIE. We use TIE to reserve q as the language context. In addition, both TDE and NDE reflect the increase of confidence for the answer given the question, *i.e.*, from q^* to q. The difference between TDE and NDE is also the existence of question q. Note that we hope to exclude the effect directly caused by question. Therefore, the mediator knowledge should be blocked when estimating the pure language effect, which is captured by NDE.

7.2. Implementation

RUBi [10] and Learned-Mixin (LM) [13] use the following fusion strategies for ensemble-based training:

$$
\text{(RUBi)} \qquad h(Z_q, Z_k) = Z_k \cdot \sigma(Z_q) \tag{22}
$$

$$
\text{(LM)} \quad h(Z_q, Z_k) = \log \sigma(Z_k) + g(k) \cdot \log \sigma(Z_q) \tag{23}
$$

where $\sigma(\cdot)$ represents the sigmoid function, and $g(\cdot)$ is a learned function $\mathbb{R}^d \to \mathbb{R}^1$ with the knowledge representation $k \in \mathbb{R}^d$ as input and a scalar weight as output. During the test stage, they use Z_k for inference.

Perhaps supering As for RUBi, NIE is calculated as:

$$
NIE = \underbrace{z_k \cdot \sigma(c)}_{Z_{q^*,k}} - \underbrace{c \cdot \sigma(c)}_{Z_{q^*,k^*}} \propto z_k \tag{24}
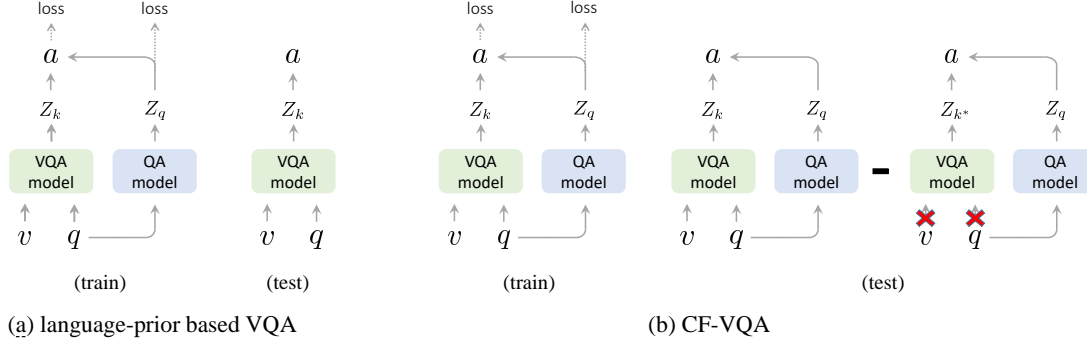$$

9

Figure 8: Comparison between our CF-VQA and Language-prior based methods [10, 13] based on the simplified causal graph for VQA.
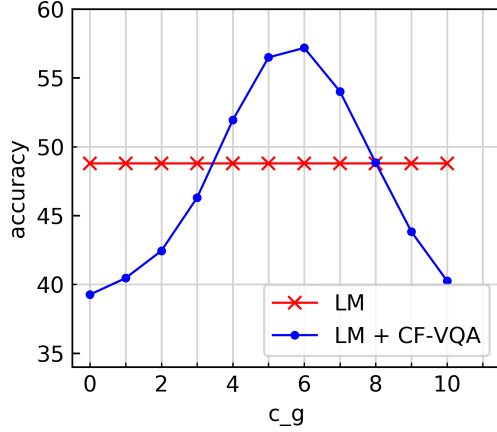

Figure 9: Accuracies of Learned-Mixin (LM) and LM+CF-VQA on VQA-CP v2 test set.

As for Learned-Mixin, NIE is calculated as:

$$
NIE = \underbrace{(\log \sigma(z_k) + g(k) \cdot \log \sigma(c))}_{Z_{q^*,k}}
$$
$$
- \underbrace{(\log \sigma(c) + g(k^*) \cdot \log \sigma(c))}_{Z_{q^*,k^*}} \propto z_k \qquad (25)
$$

where c, $g(k)$ and $g(k^*)$ are constants for the same sample. Therefore, we have $NIE \propto z_k$ for both RUBi and Learned-Mixin, which is exactly the output score of the vision-language branch $\mathcal{F}_{VQ}$. Note that RUBi and Learned-Mixin simply preserve the vision-language branch and uses z_k for inference. From our cause-effect view, *RUBi and Learned-Mixin use natural indirect effect for inference.*

7.3. Improving RUBi [10] and Learned-Mixin [13]

Thanks to our cause-effect look, RUBi [10] can be improved using CF-VQA, *i.e.*, using TIE for inference.

Specifically, TIE for RUBi is calculated as:

$$
TIE = \underbrace{z_k \cdot \sigma(z_q)}_{Z_{q,k}} - \underbrace{c \cdot \sigma(z_q)}_{Z_{q,k^*}} \qquad (26)
$$

where c denotes a learnable parameter. Table 4 in the main paper demonstrates that CF-VQA can outperform RUBi by 7% on VQA-CP v2. The red notes in Algorithm 1 show how CF-VQA improves RUBi by only changing several lines of code.

Similarly, TIE for Learned-Mixin [13] is calculated as:

$$
TIE = \underbrace{(\log \sigma(z_k) + g(k) \cdot \log \sigma(z_q))}_{Z_{q,k}}
$$
$$
- \underbrace{(\log \sigma(c) + g(k^*) \cdot \log \sigma(z_q))}_{Z_{q,k^*}} \qquad (27)
$$
$$
= \log \sigma(z_k) + (g(k) - c_g) \cdot \log \sigma(z_q)
$$

where the constant c_g denotes $g(k^*)$. Figure 9 shows that CF-VQA with an optimal c_g can improve Learned-Mixin by over 8% on VQA-CP v2, which means that there leaves a large space to promote Learned-Mixin. However, our strategy that updating c by optimizing $\mathcal{L}_{kl}$ in Eq. (17) fails to find the optimal c_g. The reason may be that c_g is totally different from c. Specifically, c is the answer score, while c_g is a weight for the combination of two scores from question-only branch and vision-language branch. We leave the problem of finding the best c_g as future work.

8. Analysis of Estimating NDE

In Section 4.2 in the main paper, we claimed that the learnable parameter c controls the sharpness of Z_{q,v^*,k^*} for estimating NDE. We give an intuitive analysis here.

For Harmonic (HM), we have:

$$
\text{(HM)} \qquad Z_{q,v^*,k^*} = \log \frac{\sigma(z_q) \cdot c_{\text{HM}}}{1 + \sigma(z_q) \cdot c_{\text{HM}}}, \qquad (28)
$$

10

Table 6: Details of VQA-CP and VQA datasets

Dataset	VQA-CP v1		VQA-CP v2		VQA v2		
Split	train	test	train	test	train	val	test
# of images	118,442	87,400	120,932	98,226	82,783	40,504	81,434
# of questions	244,547	125,314	438,183	219,928	443,757	214,354	447,793

Table 7: **Comparison on VQA-CP v2 val set**. "Base." indicates the VQA base model.

		VQA-CP v2					VQA v2
		val (in-domain)				test (OOD)	val (in-domain)
	Base.	All	Y/N	Num.	Other	All	All
GRLS [?]	–	56.90	69.23	42.50	49.36	42.33	51.92
GradSup[38]	–	62.4	77.8	43.8	53.6	46.8	–
RandImg [40]	UpDn	54.24	64.22	34.40	50.46	55.37	57.24
CF-VQA (HM)	UpDn	65.47	79.09	45.86	57.86	49.74	63.73
CF-VQA (SUM)	UpDn	60.29	66.32	47.48	57.96	51.27	62.49
CF-VQA (HM)	S-MRL	63.08	75.76	44.88	55.99	53.55	63.54
CF-VQA (SUM)	S-MRL	57.86	66.24	44.98	53.38	55.05	60.94

where $c_{\mathrm{HM}} = (\sigma(c))^2 \in (0, 1)$. We approximate the limits of Z_{q,v^*,k^*} and $TIE = Z_{q,v,k} - Z_{q,v^*,k^*}$ as:

(HM)
$$\lim_{c_{\mathrm{HM}} \to 0} Z_{q,v^*,k^*} = -\infty$$
$$\lim_{c_{\mathrm{HM}} \to 0} TIE = z_{q,v,k} - C \quad (29)$$
$$\propto z_{q,v,k},$$

where we use a extremely negative number C to replace $-\infty$ for valid estimation of TIE. In this case, NDE is estimated as the same constant for all the answers, and TIE is dominated by $z_{q,v,k}$, which means that the language bias is not reduced. For $c_{\mathrm{HM}} \to 1$, we have

(HM)
$$\lim_{c_{\mathrm{HM}} \to 1} Z_{q,v^*,k^*} = \log \frac{\sigma(z_q)}{1 + \sigma(z_q)}$$
$$\lim_{c_{\mathrm{HM}} \to 1} TIE = \log \frac{\sigma(z_v) \cdot \sigma(z_k) \cdot (1 + \sigma(z_q))}{1 + \sigma(z_q) \cdot \sigma(z_v) \cdot \sigma(z_k)}. \quad (30)$$

For SUM, we have

(SUM)
$$Z_{q,v^*,k^*} = \log \sigma(z_q + 2c), \quad (31)$$

where $c \in (-\infty, +\infty)$. We approximate the limits of Z_{q,v^*,k^*} and $TIE = Z_{q,v,k} - Z_{q,v^*,k^*}$ as:

(SUM)
$$\lim_{c \to -\infty} Z_{q,v^*,k^*} = -\infty$$
$$\lim_{c \to -\infty} TIE = z_{q,v,k} - C \quad (32)$$
$$\propto z_{q,v,k}.$$

Similar to HM, TIE is dominated by $z_{q,v,k}$. For $c \to +\infty$, we have

(SUM)
$$\lim_{c \to +\infty} Z_{q,v^*,k^*} = 0$$
$$\lim_{c \to +\infty} TIE = z_{q,v,k}. \quad (33)$$

Also, TIE is dominated by $z_{q,v,k}$. In both cases, the language bias cannot be excluded. This analysis shows that a extremely large or small c will fail to estimate NDE and TIE, and it is necessary to control the sharpness of NDE by selecting a optimal c. In the main paper, we use a KL-divergence in Eq. (17) to force the sharpness of NDE similar to that of TE.

9. Implementation Details

We use the same implementation of RUBi [10] for fair comparison, including feature representation, baseline architectures, and optimization.

Image Representation. Following the popular bottom-up attention mechanism [5], we use a Faster R-CNN based framework to extract visual features. We select top-K region proposals for each image, where K is fixed as 36.

Question Representation. Following [9, 10], we first lowercase all the questions and remove the punctuation, and then use the pretrained Skip-thought encoder [?] with fine-tuning. The size of final embedding is set as 4800.

Vision-Language Branch. The vision-language branch consists of the image representation, question representation, and a visual knowledge encoder. The baseline models for encoding visual knowledge includes SAN [45], UpDn [5], and a simplified version of the recent architecture MUREL [9] (S-MUREL) proposed in [10]. In short, S-MUREL consists of a BLOCK [?] bilinear fusion between image and question representations for each region, and a

Table 8: **Ablation of CF-VQA** on VQA-CP v1 test set. "SAN/UpDn/S-MRL" denotes the baseline VQA model. "HM/SUM" represents the strategies that train the ensemble model and test with only the vision-language branch following ensemble-based method [10, 13]. * represents the reproduced results.

	All	Y/N	Num.	Other		All	Y/N	Num.	Other		All	Y/N	Num.	Other
SAN*	32.50	36.86	12.47	36.22	UpDn*	37.08	42.46	12.76	41.50	S-MRL*	36.68	42.72	12.59	40.35
Harmonic	49.29	72.73	**20.57**	37.51	Harmonic	**55.75**	80.65	**24.72**	43.46	Harmonic	53.55	79.38	17.39	42.38
+ CF-VQA	**52.06**	**80.38**	16.88	**38.04**	+ CF-VQA	55.16	**82.27**	16.14	**43.87**	+ CF-VQA	**55.26**	**82.13**	**18.03**	**43.49**
SUM	38.34	49.88	**15.82**	35.91	SUM	52.78	78.71	14.30	42.45	SUM	49.44	76.49	16.23	35.90
+ CF-VQA	**52.87**	**84.94**	14.85	**36.26**	+ CF-VQA	**57.39**	**88.46**	**14.80**	**43.61**	+ CF-VQA	**57.03**	**89.02**	**17.08**	**41.27**

Table 9: **Ablation of CF-VQA with the simplified causal graph** on VQA-CP v1 test set. "SAN/UpDn/S-MRL" denotes the baseline VQA model. "HM/SUM" represents the strategies that train the ensemble model and test with only the vision-language branch following ensemble-based method [10, 13]. * represents the reproduced results.

	All	Y/N	Num.	Other		All	Y/N	Num.	Other		All	Y/N	Num.	Other
SAN*	32.50	36.86	12.47	36.22	UpDn*	37.08	42.46	12.76	41.50	S-MRL*	36.68	42.72	12.59	40.35
Harmonic	46.83	66.64	19.45	38.13	Harmonic	54.13	80.60	15.75	43.24	Harmonic	54.51	80.82	17.30	43.29
+ CF-VQA	**54.48**	**83.73**	**22.73**	**38.15**	+ CF-VQA	**56.19**	**85.08**	**16.00**	**43.61**	+ CF-VQA	**56.82**	**86.01**	**17.38**	**43.63**
SUM	40.08	54.15	15.53	**35.95**	SUM	51.20	74.70	13.61	42.94	SUM	52.54	78.42	16.77	**41.18**
+ CF-VQA	**52.73**	**84.64**	**16.02**	35.75	+ CF-VQA	**56.80**	**87.76**	**13.89**	**43.25**	+ CF-VQA	**57.07**	**89.28**	**17.39**	41.00

MLP classifier composed of three fully connected layers with ReLU activations. The dimension are 2,048, 2,048, and 3,000. More details can be found in [10].

Language-Only Branch. The language-only branch consists of the question representation and a question-only classifier. The question-only classifier is implemented by a MLP with three fully connect layers with ReLU activations. Note that this MLP has the same structure with the classifier for vision-language branch with different parameters.

Vision-Only Branch. The vision-only branch is composed of the question representation and a vision-only classifier. The vision-only classifier has the same structure as the language-only classifier with different parameters.

Optimization. All the experiments are conducted with the Adam optimizer for 22 epochs. The learning rate linearly increases from 1.5×10^{-4} to 6×10^{-4} for the first 7 epochs, and decays after 14 epochs by multiplying 0.25 every two epochs. The batch size is set as 256.

Datasets. The experiments are conducted on VQA-CP [3] and VQA [19] datasets. VQA-CP v1 and v2 are created by re-organizing the train and val splits of the VQA v1 and v2 datasets, respectively [3]. Therefore, there is no overlap between VQA-CP v2 train split and VQA v2 test-std split, which are used in our proposed independent-distribution setting. The details of datasets are shown in Table 6.

10. Supplementary Experimental Results

We have conducted the ablation study and compared CF-VQA with state-of-the-art methods in the main paper. In this section, we show supplementary experimental results.

10.1. Quantitative Results

As suggested by [40, **?**, 38, 39], we further hold out 8,000 instances from the training set (*i.e.*, VQA-CP v2 val) to measure the in-domain performance. Note that the results on VQA v2 val set also measure the in-domain performance. The results are given in Table 7. Compared to GRLS [**?**], all of our variants outperform GRLS by large margins for both in-domain and out-of-distribution (OOD) settings. Compared tp GradSup [38], CF-VQA (HM) achieves better results on both VQA-CP val set and test set. Compared to RandImg [40], CF-VQA (SUM) achieves competitive results on VQA-CP v2 test set, and outperforms RandImge on in-domain settings by over 3%. These results demonstrate that CF-VQA not only effectively reduces language bias, but also performs robustly.

Table 8 shows the ablation study on VQA-CP v1 test split. As shown in Table 8, CF-VQA is general to both *baseline VQA architectures* and *fusion strategies*, which is also demonstrated by the results on VQA-CP v2. Table 9 shows the ablation study on VQA-CP v1 test split using the simplified causal graph. Similarly, CF-VQA achieves significant improvement for different baseline VQA architectures and fusion strategies.

10.2. Qualitative Results

Figure 10 illustrates examples to show how CF-VQA improves RUBi by simply replacing natural indirect effect with total indirect effect for inference following Algorithm 1. The examples show that CF-VQA benefits from language context, *e.g.*, "large or small", "deep or shallow", and "real or a statue" in the first row. Some failure cases are shown in the last two rows. First, CF-VQA may tend

Is this area large or small?

RUBi + CF-VQA		RUBi	
small	94.6%	old	36.3%
large	3.5%	yes	32.8%
big	1.5%	no	18.7%
medium	0.3%	both	5.8%
huge	0.1%	small	3.0%

Is this water deep or shallow?

RUBi + CF-VQA		RUBi	
deep	97.0%	no	40.3%
shallow	2.5%	yes	35.7%
yes	0.3%	unknown	12.1%
ascending	0.1%	expert	2.3%
choppy	0.1%	old	1.8%

Is this horse real or a statue?

RUBi + CF-VQA		RUBi	
statue	90.8%	horse	84.8%
toy	5.9%	no	13.0%
real	2.7%	none	1.0%
model	0.2%	neither	0.5%
big	0.1%	yes	0.4%

Is this natural or artificial light?

RUBi + CF-VQA		RUBi	
natural	50.4%	no	72.6%
yes	25.6%	yes	18.5%
both	7.7%	none	3.2%
curved	4.8%	old	1.3%
main	3.9%	unknown	1.3%

Is this bus going or coming?

RUBi + CF-VQA		RUBi	
going	78.4%	police	36.3%
stopped	8.6%	going	30.6%
coming	6.1%	forward	11.9%
leaving	1.1%	coming	6.2%
city	0.9%	no	4.3%

Is this airport in the city or country?

RUBi + CF-VQA		RUBi	
city	93.1%	New York	25.3%
country	6.2%	yes	22.7%
yes	0.5%	unknown	18.4%
both	0.1%	not sure	7.9%
United States	0.1%	no	6.6%

Is this fruit fresh or frozen?

RUBi + CF-VQA		RUBi	
fresh	63.9%	orange	25.7%
frozen	8.3%	red	15.7%
half	3.8%	none	14.3%
salt	2.2%	white	13.7%
wheat	1.6%	yellow	8.7%

What brand is on the coffee cup?

RUBi + CF-VQA		RUBi	
starbucks	89.6%	none	37.7%
dunkin donuts	5.1%	unknown	29.6%
coca cola	4.7%	can't tell	7.8%
coke	0.5%	not sure	4.2%
jones	0.1%	not possible	3.5%

What brand of truck is shown?

RUBi + CF-VQA		RUBi	
ford	58.5%	no	29.2%
volvo	9.0%	yes	27.7%
toyota	8.8%	unknown	10.9%
dodge	5.4%	expert	6.2%
chevy	4.0%	old	5.0%

What brand is the man's shirt?

RUBi + CF-VQA		RUBi	
nike	60.1%	billabong	42.8%
adidas	12.9%	none	19.1%
billabong	10.4%	blue	12.1%
hurley	8.8%	white	6.6%
polo	3.3%	hurley	4.1%

What brand of phone is this?

RUBi + CF-VQA		RUBi	
iphone	25.9%	nintendo	68.2%
motorola	19.0%	wii	26.3%
samsung	19.0%	none	4.3%
htc	10.2%	unknown	0.3%
apple	3.3%	iphone	0.3%

What brand of soda is on the nightstand?

RUBi + CF-VQA		RUBi	
coke	42.2%	none	45.2%
coca cola	38.9%	nothing	34.5%
coca-cola	8,9%	unknown	9.0%
starbucks	1.6%	not possible	3.3%
fanta	1.6%	white	1.5%

What brand is the racket?

RUBi + CF-VQA		RUBi	
wilson	50.2%	adidas	83.0%
nike	24.1%	wilson	6.5%
prince	9.1%	nike	4.8%
head	8.5%	w	2.8%
adidas	5.6%	white	0.8%

What brand is the box?

RUBi + CF-VQA		RUBi	
hp	32.2%	dell	64.2%
canon	21.5%	hp	26.5%
dell	15.1%	windows	6.9%
toshiba	13.4%	adidas	0.9%
head	8.0%	toshiba	0.6%

What brand is shown?

RUBi + CF-VQA		RUBi	
harley	55.7%	yamaha	34.8%
yamaha	20.6%	harley	33.9%
honda	7.8%	harley davidson	18.2%
toyota	6.8%	honda	4.6%
nike	2.8%	kawasaki	2.7%

What type of herb is on the left?

RUBi + CF-VQA		RUBi	
parsley	52.1%	orange	73.7%
ginger	9.2%	red	18.6%
lily	6.0%	none	2.7%
pepper	5.7%	white	1.7%
maple	2.8%	yellow	0.9%

What type of seeds are stuck to the outside of the bun?

RUBi + CF-VQA		RUBi	
sesame	99.5%	unknown	41.7%
pepper	0.1%	none	30.1%
regular	0.1%	yellow	8.6%
sunflower	0.1%	sesame	6.6%
0	0.1%	white	2.9%

What type of hairstyle is the girl wearing?

RUBi + CF-VQA		RUBi	
ponytail	19.5%	striped	26.2%
braid	18.7%	curly	14.1%
straight	11.4%	stripes	11.7%
long	11.0%	blue	10.0%
bob	8.6%	sweater	5.2%

What type of sneakers are the players playing in?

RUBi + CF-VQA		RUBi	
cleats	40.7%	baseball	92.5%
baseball	10.9%	cleats	6.1%
tennis shoes	8.6%	baseball cap	0.6%
converse	5.9%	don't know	0.1%
giants	5.7%	yes	0.1%

What type of sink is seen in the picture?

RUBi + CF-VQA		RUBi	
pedestal	49.6%	bathroom	48.5%
bathroom	10.3%	pedestal	40.0%
porcelain	8.0%	white	5.6%
ceramic	7.1%	ceramic	5.2%
white	4.1%	porcelain	1.0%

What type of RUBi + CF-VQA does this man

RUBi + CF-VQA		RUBi	
bow RUBi + CF-VQA	74.7%	striped	50.6%
bow	15.1%	curly	21.1%
bowRUBi + CF-VQA	8.0%	stripes	7.8%
regular	0.8%	blue	3.9%
horizontal	0.6%	sweater	3.5%

What type of flower is in the vase?

RUBi + CF-VQA		RUBi	
rose	31.0%	pink	69.2%
daisy	28.3%	rose	21.8%
carnation	16.0%	red	4.6%
lily	6.2%	roses	2.1%
lilly	3.0%	purple	0.8%

What type of building is pictured in the photo?

RUBi + CF-VQA		RUBi	
clock tower	28.8%	unknown	91.0%
church	15.1%	none	3.7%
apartment	9.5%	yellow	3.0%
school	7.9%	sesame	0.7%
tower	7.1%	white	0.5%

Where on the cow's body is there a tag?

RUBi + CF-VQA		RUBi	
ear	48.8%	yes	76.8%
yes	17.3%	no	6.5%
back	9.3%	left	3.0%
head	5.0%	unknown	2.9%
legs	3.2%	bowl	2.5%

What are these machines used for?

RUBi + CF-VQA		RUBi	
money	39.2%	parking	90.0%
transportation	28.2%	money	8.6%
parking	13.4%	picture	0.6%
parking meter	3.6%	parking meter	0.2%
riding	2.0%	driving	0.1%

What are these buildings?

RUBi + CF-VQA		RUBi	
houses	26.8%	church	73.1%
skyscrapers	13.8%	city	8.4%
office	6.7%	building	5.5%
tower	6.4%	london	1.8%
church	6.2%	castle	1.6%

What are those round green things?

RUBi + CF-VQA		RUBi	
vegetables	27.3%	peas	97.3%
peas	19.5%	grapes	1.2%
grapes	10.2%	vegetables	1.2%
beets	6.6%	fruit	0.1%
fruit	5.9%	blueberries	0.1%

Why are they wearing wetsuits?

RUBi + CF-VQA		RUBi	
safety	67.9%	surfing	77.5%
surf	7.7%	surf	11.9%
yes	5.1%	yes	3.9%
protection	4.0%	sunny	2.1%
surfing	3.5%	walking	1.0%

Why are the people gathered?

RUBi + CF-VQA		RUBi	
parade	82.3%	polo	40.3%
riding horses	5.8%	rodeo	27.1%
jousting	2.3%	horseback riding	12.2%
race	2.1%	riding horses	9.3%
horseback riding	1.7%	jousting	7.2%

What kind of window covering is shown?

RUBi + CF-VQA		RUBi	
blinds	96.5%	curtain	16.8%
shade	2.9%	canopy	15.1%
sheet	0.3%	fan	14.9%
curtains	0.2%	curtains	7.9%
cloth	0.1%	sheet	7.7%

Figure 10: Qualitative comparison of RUBi and RUBi+CF-VQA on VQA-CP v2 test split. Red bold answer denotes the ground-truth one.

to generate broad answers, *e.g.*, "houses" v.s "church", and "vegetables" v.s "peas". Second, CF-VQA may ignore visual content like traditional likelihood strategy. Therefore, there remains the challenge about how to balance visual understanding and language context.

References

[1] Ehsan Abbasnejad, Damien Teney, Amin Parvaneh, Javen Shi, and Anton van den Hengel. Counterfactual vision and language learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10044–10054, 2020. 1, 3, 6, 7

[2] Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. Analyzing the behavior of visual question answering models. *arXiv preprint arXiv:1606.07356*, 2016. 1, 2

[3] 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*, pages 4971–4980, 2018. 1, 2, 6, 7, 12

[4] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C Lawrence Zitnick, Devi Parikh, and Dhruv Batra. Vqa: Visual question answering. *International Journal of Computer Vision*, 123(1):4–31, 2017. 1

[5] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018. 6, 11

[6] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3674–3683, 2018. 1, 6

[7] Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. Grounding visual explanations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 264–279, 2018. 3

[8] 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*, pages 2425–2433, 2015. 1, 2

[9] Remi Cadene, Hedi Ben-Younes, Matthieu Cord, and Nicolas Thome. Murel: Multimodal relational reasoning for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1989–1998, 2019. 6, 11

[10] Remi Cadene, Corentin Dancette, Hedi Ben-younes, Matthieu Cord, and Devi Parikh. Rubi: Reducing unimodal biases in visual question answering. *arXiv preprint arXiv:1906.10169*, 2019. 1, 2, 3, 5, 6, 7, 9, 10, 11, 12

[11] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10800–10809, 2020. 1, 3, 6, 7

[12] Long Chen, Hanwang Zhang, Jun Xiao, Xiangnan He, Shiliang Pu, and Shih-Fu Chang. Scene dynamics: Counterfactual critic multi-agent training for scene graph generation. *arXiv preprint arXiv:1812.02347*, 2018. 3

[13] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. *arXiv preprint arXiv:1909.03683*, 2019. 1, 2, 3, 5, 6, 7, 9, 10, 12

[14] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90–100, 2017. 1, 3, 6

[15] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335, 2017. 1

[16] Zhiyuan Fang, Shu Kong, Charless Fowlkes, and Yezhou Yang. Modularized textual grounding for counterfactual resilience. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6378–6388, 2019. 3

[17] Tsu-Jui Fu, Xin Eric Wang, Scott Grafton, Miguel Eckstein, and William Yang Wang. Sscr: Iterative language-based image editing via self-supervised counterfactual reasoning. *arXiv preprint arXiv:2009.09566*, 2020. 3

[18] Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. Mutant: A training paradigm for out-of-distribution generalization in visual question answering. *arXiv preprint arXiv:2009.08566*, 2020. 1, 6, 7

[19] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017. 1, 2, 12

[20] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. *arXiv preprint arXiv:1904.07451*, 2019. 3

[21] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 5

[22] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8779–8788, 2018. 1, 3, 6

[23] Chenchen Jing, Yuwei Wu, Xiaoxun Zhang, Yunde Jia, and Qi Wu. Overcoming language priors in vqa via decomposed linguistic representations. In *AAAI*, pages 11181–11188, 2020. 6

[24] Kushal Kafle and Christopher Kanan. An analysis of visual question answering algorithms. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1965–1973, 2017. 1, 2

[25] Atsushi Kanehira, Kentaro Takemoto, Sho Inayoshi, and Tatsuya Harada. Multimodal explanations by predicting coun-

terfactuality in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8594–8602, 2019. 3

[26] Gouthaman KV and Anurag Mittal. Reducing language biases in visual question answering with visually-grounded question encoder. *arXiv preprint arXiv:2007.06198*, 2020. 3, 6

[27] Zujie Liang, Weitao Jiang, Haifeng Hu, and Jiaying Zhu. Learning to contrast the counterfactual samples for robust visual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3285–3292, 2020. 1, 6, 7

[28] Judea Pearl. *Causality: models, reasoning and inference*, volume 29. Springer, 2000. 1, 3

[29] Judea Pearl. Direct and indirect effects. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*, pages 411–420. Morgan Kaufmann Publishers Inc., 2001. 1, 3

[30] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic Books, 2018. 1, 3

[31] Jiaxin Qi, Yulei Niu, Jianqiang Huang, and Hanwang Zhang. Two causal principles for improving visual dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10860–10869, 2020. 3

[32] Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. Overcoming language priors in visual question answering with adversarial regularization. In *Advances in Neural Information Processing Systems*, pages 1541–1551, 2018. 3, 6, 7

[33] James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7(9-12):1393–1512, 1986. 3

[34] James M Robins. Semantics of causal dag models and the identification of direct and indirect effects. *Oxford Statistical Science Series*, pages 70–82, 2003. 3

[35] Donald B Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pages 34–58, 1978. 3

[36] Ramprasaath R Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Dhruv Batra, and Devi Parikh. Taking a hint: Leveraging explanations to make vision and language models more grounded. *arXiv preprint arXiv:1902.03751*, 2019. 1, 2, 3, 6

[37] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3716–3725, 2020. 3

[38] Damien Teney, Ehsan Abbasnedjad, and Anton van den Hengel. Learning what makes a difference from counterfactual examples and gradient supervision. *arXiv preprint arXiv:2004.09034*, 2020. 3, 11, 12

[39] Damien Teney, Ehsan Abbasnejad, and Anton van den Hengel. Unshuffling data for improved generalization. *arXiv preprint arXiv:2002.11894*, 2020. 3, 6, 7, 12

[40] Damien Teney, Kushal Kafle, Robik Shrestha, Ehsan Abbasnejad, Christopher Kanan, and Anton van den Hengel. On the value of out-of-distribution testing: An example of goodhart's law. *arXiv preprint arXiv:2005.09241*, 2020. 6, 7, 11, 12

[41] Pei Wang and Nuno Vasconcelos. Scout: Self-aware discriminant counterfactual explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8981–8990, 2020. 3

[42] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10760–10770, 2020. 3

[43] Jialin Wu and Raymond J Mooney. Self-critical reasoning for robust visual question answering. *arXiv preprint arXiv:1905.09998*, 2019. 1, 2, 3, 6

[44] Xu Yang, Hanwang Zhang, and Jianfei Cai. Deconfounded image captioning: A causal retrospect. *arXiv preprint arXiv:2003.03923*, 2020. 3

[45] 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*, pages 21–29, 2016. 6, 11

[46] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019. 3

[47] Zhongqi Yue, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. Interventional few-shot learning. *arXiv preprint arXiv:2009.13000*, 2020. 3

[48] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6720–6731, 2019. 1

[49] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xiansheng Hua, and Qianru Sun. Causal intervention for weakly-supervised semantic segmentation. *arXiv preprint arXiv:2009.12547*, 2020. 3

[50] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and yang: Balancing and answering binary visual questions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5014–5022, 2016. 1, 2

[51] Shengyu Zhang, Tan Jiang, Tan Wang, Kun Kuang, Zhou Zhao, Jianke Zhu, Jin Yu, Hongxia Yang, and Fei Wu. Devlbert: Learning deconfounded visio-linguistic representations. *arXiv preprint arXiv:2008.06884*, 2020. 3

[52] Xi Zhu, Zhendong Mao, Chunxiao Liu, Peng Zhang, Bin Wang, and Yongdong Zhang. Overcoming language priors with self-supervised learning for visual question answering. IJCAI, 2020. 1, 3, 6, 7