Counterfactual VQA: A Cause-Effect Look at Language Bias

Yulei Niu[1] **Kaihua Tang**[2] **Hanwang Zhang**[2] **Zhiwu Lu**[1] **Xian-Sheng Hua**[3] **Ji-Rong Wen**[1]
[1]Gaoling School of Artificial Intelligence, Renmin University of China
[2]Nanyang Technological University
[3]Damo Academy, Alibaba Group
{niu, luzhiwu, jrwen}@ruc.edu.cn, {kaihua001@e., hanwangzhang@}ntu.edu.sg

Abstract

Visual Question Answering (VQA) models tend to rely on the language bias and thus fail to learn the reasoning from visual knowledge, which is however the original intention of VQA. In this paper, we propose a novel **cause-effect** look at the language bias, where the bias is formulated as the direct effect of question on answer from the view of **causal inference**. The effect can be captured by **counterfactual VQA**, where the image had not existed in an imagined scenario. Our proposed cause-effect look 1) is general to any baseline VQA architecture, 2) achieves significant improvement on the language-bias sensitive VQA-CP dataset, and 3) fills the theoretical gap in recent language prior based works [1, 2].

1 Introduction

Visual Question & Answering (VQA) [3, 4] has become the fundamental building block that underpins many frontier interactive AI systems, including visual dialog [5], vision-language navigation [6], and visual commonsense reasoning [7]. Over the recent years, we have witnessed the fast evolution of VQA, especially multi-modal fusion [8, 9, 10] and attention mechanism [11, 12]. However, a common observation is that most VQA models tend to exploit the language bias for answering. This serious shortcut limits the generalization of VQA in real-world scenarios where the test question language is quite different from that in training [13, 14, 15]. The VQA language bias can be interpreted in two ways. Firstly, there exists strong correlation between questions and answers, which reflects the *"language prior"* [15, 16]. For instance, simply answering "tennis" for the sport-related questions can achieve approximately 40% accuracy on VQA v1.0 dataset. Secondly, the questioner tends to ask about the objects seen in the image, which leads to the *"visual priming bias"* [3, 15]. For example, 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.

One straightforward solution is to improve the data quality by creating a more balanced dataset [15]. Although this strategy can effectively reduce the visual priming bias, models can still benefit from exploiting the language prior [16]. Most existing works have contributed to overcoming the VQA language bias by strengthening visual grounding or weakening language prior: 1) human visual [17] and textual [18] explanations are exploited as external knowledge to improve the visual grounding in VQA [19, 20]. However, such human explanation collection is expensive. 2) a separate question-only branch used to directly capture the language prior is imposed in training [21, 1, 2] and excluded in testing. The key motivation is that the whole question provides the language prior. However, we argue that the prior consists of both "bad" language bias (*e.g.*, most bananas are yellow while few are green from observations) and "good" language context (*e.g.*, "what color"). Indeed, it is challenging for the above methods to disentangle the good and bad from the whole, but it will no longer be, after we introduce the counterfactual reasoning in the proposed Counterfactual VQA (CF-VQA).

Preprint. Under review.

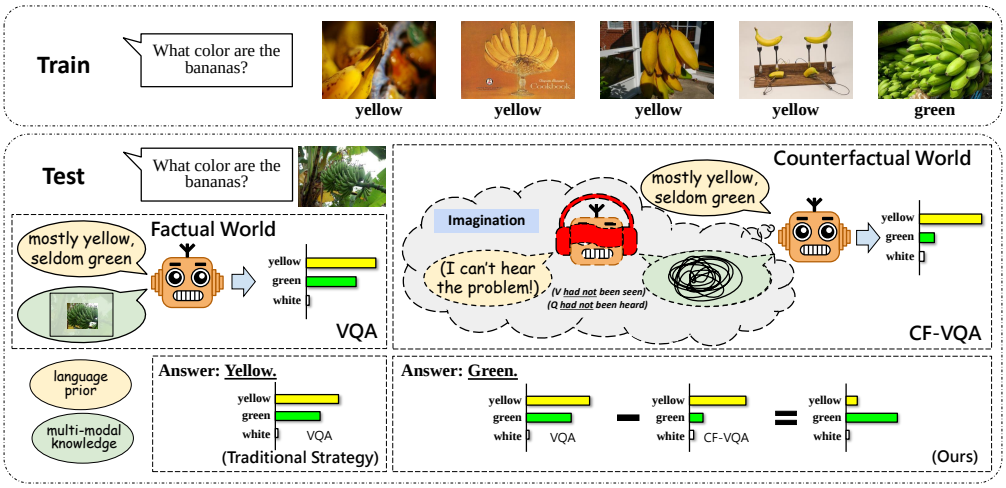


Figure 1: Our cause-effect look at the language bias in VQA by using the proposed *Counterfactual VQA* (CF-VQA). The factual world depicts the conventional VQA, and the counterfactual world depicts our proposed CF-VQA.

CF-VQA is a novel cause-effect look at the language bias in VQA, which is inspired by the counterfactual thinking in causal inference. Counterfactual thinking gifts us humans the imagination ability to introspect whether our decision is made from the bias. Suppose that a submitted paper achieves state-of-the-art performance (SOTA) but has weak novelty, will you vote for accept or reject? Junior reviewers may vote for accept because most of the accepted papers have SOTA. However, we argue that the basis for decision-making should be obtained via counterfactual analysis, *i.e.*, comparing the fact with its *counterfactual*. For example, if SOTA *would* still be achieved using the same experimental settings (*e.g.*, strong backbone) but the proposed method *had been* replaced with a simple baseline in the counterfactual world, we can draw the conclusion that the proposed method contributes little to SOTA. In that case, we may vote for reject. Indeed, this acceptance bias can be explained by the well-known *survivor bias* [22], since we can read the accepted papers while rarely access the rejected ones, *i.e.*, our observations are biased. Thanks to the counterfactual analysis, this bias can be overcome by the comparison between factual and counterfactual outcomes.

As shown in Figure 1, the factual VQA world is that machine both hears question Q and sees image V when generating the answer A. In addition, machine extracts multi-modal knowledge K (*i.e.*, fused feature representation) from both Q and V to predict A. In the counterfactual world, the machine hears Q , but K^* is from an *imagined* world where Q *had NOT been heard* and V *had NOT been seen*. Here comes the comparison between conventional VQA and our proposed *counterfactual VQA*:

VQA: *What will A be, if machine hears Q and extracts K based on V?*
Counterfactual VQA (CF-VQA): *What **would** A be, if machine hears Q, but extracts K^* when it **had NOT** heard Q and **had NOT** seen V?*

Comparing the answers generated from VQA and CF-VQA worlds, machine can identify the bad language bias and exclude its effect before answering. As a result, the pure language bias can be captured by the direct effect of Q on A (*i.e.*, the effect with K blocked) in CF-VQA. During the training stage, we simply adopt a prevailing VQA model with additional QA model for revealing the language-only effect. During the test stage shown in Figure 1, instead of answering based on the posterior $P(A|V,Q)$ in the factual world, we expect machine to use the *causal effect* of V and Q on A by removing the pure language effect captured by CF-VQA, where K is fixed as the value K^* in the imagined world without V and Q, *i.e.*, the effect via multi-modal knowledge is blocked. Experimental results show that our causal strategy CF-VQA achieves a new state-of-the-art performance on the language-bias sensitive VQA-CP [16] dataset while remains stable on the balanced VQA v2 [15].

Our contributions are summarized as:

- We are the first to view language bias in VQA from the cause-effect look based on causal inference.
- Our cause-effect look fills the theoretical gap in recent debiasing VQA works [1, 2].
- Our cause-effect look takes a general form and is suitable for different baseline VQA architectures (*e.g.*, UpDn [23], SAN [11]) and fusion strategies (*e.g.*, RUBi [1] and our proposed variants).

2 Related Work

Bias in Vision-Language. Recent studies found that dataset bias, especially language bias, widely exists in vision-and-language tasks [14, 15, 24, 25, 26, 27, 28]. In VQA v1.0 dataset [3], the answer distribution is unbalanced over many question types, *e.g.*, sports-related questions and yes-no questions [14, 15]. In MS-COCO dataset for image captioning, there exist high co-occurring word pairs or sentence patterns, such as "sheep+field" and "man+standing" [24]. The language bias encourages machines to simply leverage language prior rather than fully exploiting the visual content. Researchers have made efforts to overcome or utilize vision-and-language language bias from the views of dataset collection [15, 16], robust training [21, 1, 2], and external knowledge [19, 20]. These works indicate that language bias has become an increasing concern to the community.

Debiasing VQA. Recently, a new split VQA-CP [16] is proposed to evaluate the generalizability of VQA models, where the distributions of answers per question type during training and test stages are different while the concepts remain the same. Recent works [19, 20, 1, 2, 21, 29, 30, 31, 32] have made efforts to overcome the VQA language bias mainly in two aspects, strengthening visual grounding and weakening language prior. Firstly, human visual [17] and textual [18] explanations are exploited to strengthen the visual grounding in VQA [19, 20]. The VQA-HAT dataset collects human attention map as visual attention supervision [17], while the VQA-X dataset contains textual explanations related to question-answer pairs. However, these human explanations require additional expensive annotations, which limits the generalization to other datasets. Secondly, recent methods proposed a separated QA branch to capture the language prior on the training set in an adversarial learning manner [21] or multi-task learning manner [1, 2], which are the most related works to us. However, they lack a theoretical explanation about the inconsistency between training and inference. In this paper, our proposed cause-effect look fills in the theoretical gap of the language prior based methods [1, 2] from a causal view, which guides us how to unify them and further improve them.

Causal Inference. Causal inference has been widely studied in statistics, economics, epidemiology, and sociology [33, 34, 35, 36, 37, 38]. The machine learning community has also made efforts in treatment effect estimation [39, 40, 41] and real-world applications [42, 43, 44, 28, 45, 46, 47]. Recently, counterfactual thinking has been exploited in various computer vision problems, including visual explanations [48, 49], scene graph generation [50], video analysis [51, 52], and vision-language tasks [29, 27, 32]. In this paper, we use counterfactual analysis in the causal inference framework.

3 Preliminaries

In this section, we formally introduce the key concepts of *causal inference* [53, 38, 54, 55] used in this paper. We represent a random variable as a capital letter (*e.g.*, X), and denote its observed value as a lowercase letter (*e.g.*, x).

The *causal graph* 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-effect relationships. If X has *direct* effect on Y, we say that there exists a direct edge from X to Y, *i.e.*, $X \rightarrow Y$. If a variable M lies on the way from X and Y, *i.e.*, $X \rightarrow M \rightarrow Y$, we say that M acts as the *mediator* between X and Y. The value that Y would obtain if X is set to x and M is set to m is denoted as:

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

which is the general *counterfactual notation*[1]. In the factual world, we have $m = M_x = M(X = x)$. In the counterfactual world, we take $Y_{x,M_{x^*}}$ as an example. The notation $Y_{x,M_{x^*}}$ describes the situation where X is set to x and M is set to the value when X had been x^*:

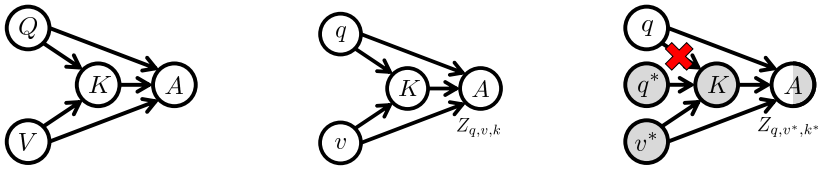
$$Y_{x,M_{x^*}} = Y(X = x, M = M(X = x^*))$$

Note that X can be simultaneously set to different values x and x^* only in the counterfactual world.

4 Cause-Effect Look at VQA

A VQA model is required to generate an answer $A = a$ based on an image $V = v$ and the related question $Q = q$. In the following, we set up the causal graph and analyze the causal effects in VQA.

[1] If 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) Conventional VQA (c) Counterfactual VQA

Figure 2: (a) Causal graph for VQA. Q: question. V: image. K: multi-modal knowledge. A: answer. (b) & (c) Comparison between conventional VQA and counterfactual VQA. White nodes are at the value $V=v$ and $Q=q$ while gray nodes are at $V=v^*$ and $Q=q^*$.

4.1 Causal Graph

Since the answer A is generated based on the image V and question Q, we call V and Q as the cause of A. The impact of V and Q on A can be divided into two parts: single-modal impact and multi-modal impact. Intuitively, the single-modal impact reflects the direct effect of V or Q on A, while the multi-modal impact captures the indirect effect of V and Q on A via multi-modal knowledge K (*e.g.*, joint feature). The detailed causal graph is shown in Figure 7(a).

4.2 Cause-Effect Look

Following the counterfactual notation in Eq. (1), we denote the logit Y that a specific answer a (*e.g.*, "green") would obtain if V had been v (*e.g.*, an image which includes green bananas) and Q had been q (*e.g.*, "What color are the bananas?") as $Y_{v,q} = Y(V=v, Q=q)$, where we omit a for simplicity. The reference values of V and Q are denoted as v^* and q^* respectively. Similarly, the counterfactual notation of the multi-modal knowledge K is denoted as $K_{v,q} = K(V=v, Q=q)$.

As shown in Figure 7(a), there exist three paths directly connected to A, *i.e.*, $Q \rightarrow A$ from the exposure Q, $V \rightarrow A$ from the exposure V, and $K \rightarrow A$ from the mediator K. Therefore, we rewrite $Y_{v,q}$ as the function of Q, V and K:

$$Z_{q,v,k} = Z(Q=q, V=v, K=k) = Y_{v,q} \tag{2}$$

where k represents $K_{v,q}$. Following the definition of causal effects [53, 38, 54, 55], the *total effect* (TE) of v and q on a can be written as:

$$TE = Y_{v,q} - Y_{v^*,q^*} = Z_{q,v,k} - Z_{q^*,v^*,k^*} \tag{3}$$

where $k^* = K_{v^*,q^*}$. The total effect reflects the causal effect in conventional VQA. Recall that counterfactual VQA (CF-VQA) describes the world where Q is set to q and K would attain the value k^* when Q had been q^* and V had been v^* (see Section 1). Figure 7(b) and 7(c) show the comparison between VQA and CF-VQA. Compared to the reference state, we obtain the *natural direct effect* (NDE) of q on a (*i.e.*, pure language effect) as:

$$NDE = Z_{q,v^*,k^*} - Z_{q^*,v^*,k^*} \tag{4}$$

Since and the effect of Q on the intermediate K is blocked (*i.e.*, $K=k^*$), *e.g.*, answering the question "What color are the bananas?" without any multi-modal knowledge, NDE explicitly captures the language bias from the cause-effect view. Therefore, removing language bias can be realized by reducing NDE from TE, which is represented as:

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

During the inference stage, we **select the answer with the maximum effect in Eq. (5) for response**, which is totally different from traditional inference strategies that pick the answer based on the maximum posterior (*i.e.*, $P(a|v,q) \propto Y_{v,q} = Z_{q,v,k}$).

4.3 Implementation

Parameterization. As for VQA, we define the reference value as blocking the signal from vision ($V=v^*=\varnothing$) or language ($Q=q^*=\varnothing$), *i.e.*, v or q is not given. Note that this definition is counter

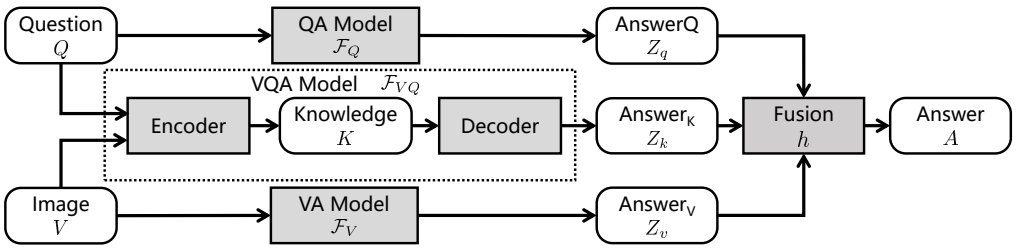


Figure 3: The overall architecture for the implementation of VQA causal graph.

to the conventional VQA where both V and Q are always valid, and VQA models cannot extract knowledge from the empty input based on the actual definition. We assume that the knowledge would be void representation if V or Q had not been given, and rewrite K as

$$K = \begin{cases} k = K_{v,q} & \text{, if } V = v \text{ and } Q = q \\ k^* = \varnothing & \text{, if } V = v^* \text{ or } Q = q^* \end{cases} \tag{6}$$

The calculation of the single logit $Z_{q,v,k}$ in Eq. (2) is parameterized by three neural models $\mathcal{F}_Q$, $\mathcal{F}_V$, $\mathcal{F}_{VQ}$ and one fusion function h as:

$$Z_q = \mathcal{F}_Q(q), \quad Z_v = \mathcal{F}_V(v), \quad Z_k = \mathcal{F}_{VQ}(v, q), \quad Z_{q,v,k} = h(Z_q, Z_v, Z_k) \tag{7}$$

where $\mathcal{F}_Q$ is the language-only branch (*i.e.*, $Q \to A$) with the question-based logit $Z_q \in \mathbb{R}^1$ as output, $\mathcal{F}_V$ is the vision-only branch (*i.e.*, $V \to A$) with the vision-based logit $Z_v \in \mathbb{R}^1$ as output, and $\mathcal{F}_{VQ}$ is the vision-language branch (*i.e.*, $V, Q \to K \to A$) with the knowledge-based logit $Z_k \in \mathbb{R}^1$ as output. These three logits are fused by function $h : \mathbb{R}^1 \times \mathbb{R}^1 \times \mathbb{R}^1 \to \mathbb{R}^1$ to obtain the final logit $Z_{q,v,k}$. Figure 3 illustrates the combination of these components.

Since the neural models cannot deal with void input (*e.g.*, $K = \varnothing$), we define the outcome of void input as the same constant for all the logits. As for us human, if we have no knowledge for VQA (*i.e.*, vision-language branch) or QA (*i.e.*, language-only branch), we would like to make inference by random guess, which means that each candidate answer has the same chance (*i.e.*, logit) to be picked. Therefore, we represent Z_q, Z_v and Z_k in Eq. (7) as:

$$\begin{gathered} Z_q = \begin{cases} z_q = \mathcal{F}_Q(q) & \text{if } Q = q \\ z_q^* = c & \text{if } Q = \varnothing \end{cases} \qquad Z_v = \begin{cases} z_v = \mathcal{F}_V(v) & \text{if } V = v \\ z_v^* = c & \text{if } V = \varnothing \end{cases} \\ Z_k = \begin{cases} z_k = \mathcal{F}_{VQ}(v, q) & \text{if } V = v \text{ and } Q = q \\ z_k^* = c & \text{if } V = \varnothing \text{ or } Q = \varnothing \end{cases} \end{gathered} \tag{8}$$

where c denotes the constant. Note that *our cause-effect look is model-agnostic*. Commonly, the vision-language branch can be any VQA model, *e.g.*, UpDn [23] and SAN [11]. The language-only branch can be a simple QA model. More details can be found in the supplementary materials.

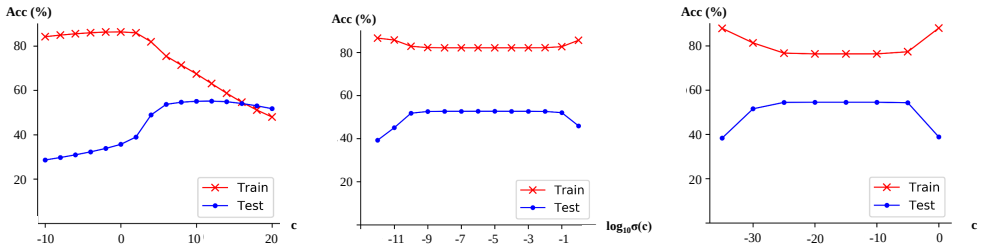
Fusion Strategies. Apart from being model-agnostic, our cause-effect look is also *fusion-agnostic*. We expect that the fused logit $Z_{q,v,k}$ is non-linear combination of Z_q, Z_v and Z_k, *i.e.*, $Z_{q,v,k} = h(Z_q, Z_v, Z_k)$. Recently, RUBi [1] implements the fusion function as $h(Z_q, Z_k) = Z_k \cdot \sigma(Z_q)$, where $\sigma(\cdot)$ represents the sigmoid activation function. Considering the additional vision-only branch, we extend RUBi as:

$$h(Z_q, Z_v, Z_k) = Z_k \cdot (\sigma(Z_q) + \sigma(Z_v)) \qquad \text{(RUBi)} \tag{9}$$

In order to evaluate the generalization of our cause-effect look, we further propose another two non-linear fusion variants:

$$h(Z_q, Z_v, Z_k) = \log \frac{\sigma(Z_q) \cdot \sigma(Z_v) \cdot \sigma(Z_k)}{1 + \sigma(Z_q) \cdot \sigma(Z_v) \cdot \sigma(Z_k)} \qquad \text{(Harmonic)} \tag{10}$$

$$h(Z_q, Z_v, Z_k) = \log \sigma(Z_q + Z_v + Z_k) \qquad \text{(Sum)} \tag{11}$$



(a) RUBi (b) Harmonic (c) Sum

Figure 4: Accuracies (%) on VQA-CP v2 with S-MRL and different c.

Training and Inference. We follow the training strategy used by [1]. Specifically, given a dataset $\mathcal{D} = \{(v, q, a)\}$ where a is the ground-truth answer of image-question pair (v, q), the model is optimized by minimizing the sum of cross-entropy losses $\mathcal{L}$ over the final logit $z_{q,v,k}$, question-based logit z_q, and vision-based logit z_v:

$$\mathcal{L} = \sum_{\{(v,q,a)\}\in\mathcal{D}} \underbrace{-\log \text{softmax}(z_{q,v,k})[a]}_{\text{VQA}} \underbrace{-\log \text{softmax}(z_q)[a]}_{\text{QA}} \underbrace{-\log \text{softmax}(z_v)[a]}_{\text{VA}} \tag{12}$$

where $z_{q,v,k} = h(z_q, z_v, z_k)$. As discussed in Section 4.2, we remove NDE from TE for inference. Considering the above implementation, we have

$$TE - NDE = Z_{q,v,k} - Z_{q,v^*,k^*} = h(z_q, z_v, z_k) - h(z_q, z_v^*, z_k^*) \quad \text{(CF-VQA)} \tag{13}$$

Note that Cadene *et al.* [1] and Clark *et al.* [2] also use an ensemble model with the vision-language branch $\mathcal{F}_{VQ}$ and question-only branch $\mathcal{F}_Q$. During the test stage, they simply exclude $\mathcal{F}_Q$ and use $z_k = \mathcal{F}_{VQ}(v, q)$ for inference. From our cause-effect look, *these language prior based methods are the special cases of ours*, which (1) remove the direct path $V \rightarrow A$ in the causal graph, and (2) use natural indirect effect for inference. The detailed proof is provided in the supplementary materials.

5 Experiments

The main experiments are conducted on the language-bias sensitive VQA-CP v2 (Visual Question Answering under Changing Priors) [16] dataset. VQA-CP v2 is proposed to evaluate the robustness of VQA models when the answer distributions of training and test splits are significantly different (see examples in Figure 6). In addition, we also report the performance on the balanced VQA v2 dataset to see whether the approach over-corrects the language bias. The models are evaluated via the standard evaluation metric (*i.e.*, accuracy). Following recent works, we conduct experiments with three baseline VQA architectures: Stacked Attention Network (**SAN**) [11], Bottom-up and Top-down Attention (**UpDn**) [23], and a simplified version of MUREL [56] (**S-MRL**) proposed by Cadene *et al.* [1]. Detailed experimental settings and other experimental results (*e.g.*, on VQA-CP v1) are provided in the supplementary materials.

5.1 Ablation study

How does the constant impact? Recall that we use a constant c in case of void input as Eq. (8). Figure 4 shows the influence of c on our proposed causal inference strategy (CF-VQA) with the S-MRL base model. Interestingly, Harmonic and Sum fusion strategies are less sensitive than RUBi to the constant c. The reason may be that the sigmoid activation works as a normalization over Z_k. In addition, the bottom of the training accuracy curve and the top of the test accuracy curve are reached simultaneously for Harmonic and Sum, which indicates that we can select c based on the training performance for these two strategies. In the following experiments, we set c as 10 for RUBi, logit(10^{-3}) for Harmonic, and -10 for Sum. Considering the hyper-parameter selection based on the training set, we prefer Harmonic and Sum in practice. It is worth noting that the constant is only used during the test stage. Therefore, tuning the constant does not require re-training the model.

Table 1: Ablations of (1) baseline models (SAN, UpDn and S-MRL), (2) fusion strategies (RUBi, Harmonic, Sum) on VQA-CP v2 test set. We reimplement the baselines for fair comparison.

SAN	All	Y/N	Num.	Other
Baseline	33.18	38.57	12.25	36.10
RUBi	40.48	68.99	17.75	31.79
+ CF-VQA	**47.85**	**88.71**	**18.79**	**34.41**
Harmonic	45.89	70.37	**23.99**	39.07
+ CF-VQA	**48.10**	**77.68**	22.19	**39.71**
Sum	43.98	68.98	**17.32**	38.19
+ CF-VQA	**50.15**	**87.95**	16.46	**39.59**

UpDn	All	Y/N	Num.	Other
Baseline	37.69	43.17	12.53	41.72
RUBi	43.90	64.20	16.84	**40.69**
+ CF-VQA	**51.21**	**90.41**	**27.25**	37.24
Harmonic	47.97	69.19	18.80	44.86
+ CF-VQA	**49.94**	**74.82**	**18.93**	**45.42**
Sum	47.29	72.26	12.54	43.74
+ CF-VQA	**53.69**	**91.25**	**12.80**	**45.23**

S-MRL	All	Y/N	Num.	Other
Baseline	37.09	41.39	12.46	41.60
RUBi	47.35	69.98	21.53	42.58
+ CF-VQA	**55.10**	**89.83**	**26.63**	**44.70**
Harmonic	49.37	73.20	**20.10**	44.92
+ CF-VQA	**51.43**	**78.62**	20.07	**45.79**
Sum	48.27	74.60	20.96	41.96
+ CF-VQA	**54.95**	**90.56**	**21.88**	**45.36**

How to deal with void input? Recall that we expect the model to make a random guess with void input, *i.e.*, uniform distribution over answers in Eq. (8). Another choice is to define the logits as a prior distribution over answers, where the prior is computed based on the frequencies of answers during training. Table 2 shows that the prior distribution assumption decreases the accuracy compared to using the constant, which indicates that our uniform distribution assumption, *i.e.*, random guess in the case of none knowledge, is both theoretically and practically correct for the void input scenario.

Table 2: Ablation of logit with void input on VQA v2 test split.

S-MRL	All	Y/N	Num.	Other
Har. w/ prior	49.07	74.95	**23.79**	42.44
Har. w/ unif.	**51.43**	**78.62**	20.07	**45.79**
Sum w/ prior	50.40	89.09	14.91	39.86
Sum w/ unif.	**54.95**	**90.56**	**21.88**	**45.36**

How general is CF-VQA? Table 1 demonstrates that CF-VQA is very general to both *baseline VQA architectures* and *fusion strategies*. Here, "RUBi/Harmonic/Sum" means using the language prior based strategies [1, 2] that train the ensemble model and test with only the vision-language branch. It can be clearly seen that CF-VQA improves the language prior based strategies by over 2% in all cases. In particular, RUBi and Sum are improved by over 6% with different baseline architectures.

5.2 Comparisons with State-of-the-Art Approaches

Recent studies for robust VQA can be mainly grouped in two classes. The **visual attention based** approaches exploit human visual [17] or textual [18] explanations to strengthen visual attention mechanisms in VQA systems, including HINT [19] and SCR [20]. The **language prior based** methods directly focus on overcoming language prior for robust training and inference, including AdvReg. [21], RUBi [1] and Learned-Mixin (LM) [2]. All these works explicitly model the language prior using a separated language-only branch. For fair comparison, all the comparing methods are without data augmentation [29]. We conducted three experimental settings to evaluate the ability of overcoming language bias and robustness: **out of distribution** (trained on the VQA-CP v2 train split and evaluated on the VQA-CP v2 test split where the language priors are changed), **in distribution** (trained on the VQA v2 train split and evaluated on the VQA v2 val set where the priors are similar and countered), and our proposed **independent distribution** (trained on the biased VQA-CP v2 train split and evaluated on the balanced VQA v2 test-std split where the priors are independent). Note that the last two settings use the same model trained on the biased VQA-CP v2 train split.

Can CF-VQA overcome language bias? The out-distribution setting aims to evaluate whether VQA models effectively reduce language bias. As shown in Table 3, our CF-VQA with S-MRL and Sum achieves a state-of-the-art performance on VQA-CP v2 test split as 54.95%, over 7% improvement compared to the most related method RUBi [1]. With a deep look at the question type, we find that the improvement on "Yes/No" questions is extremely large (from $\sim$70% to $\sim$90%), which indicates that the language bias affects variously on different types of questions. The example for "is this" questions in Figure 6 also illustrates that our CF-VQA effectively reduce the language bias.

Is CF-VQA robust? The in-distribution and independent-distribution settings are conducted to evaluate whether VQA models over-correct language bias. As shown in Table 3, "UpDn+Sum" for our CF-VQA achieves a relatively good trade-off among the three settings. Note that LM [2] achieves a competitive performance on VQA-CP v2 test split ($\sim$52%) with an additional language entropy penalty (LM+H). However, the accuracy drops significantly by $\sim$7% on the in-domain setting (*i.e.*, VQA v2 val split), which indicates that the entropy penalty forces the model to over-correct the language bias, especially on "Yes/No" questions.

Table 3: Experimental results on VQA-CP v2 and VQA v2. **Best** and **second best** numbers are highlighted in each column. The results on VQA v2 test-std are obtained via our reimplementation.

Training set		VQA-CP v2 train								VQA v2 train			
Test set		VQA-CP v2 test				VQA v2 test-std				VQA v2 val			
Methods	Base.	All	Y/N	Num.	Other	All	Y/N	Num.	Other	All	Y/N	Num.	Other
GVQA [16]	–	31.30	57.99	13.68	22.14	–	–	–	–	48.24	72.03	31.17	34.65
SAN [11]	–	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	59.89	76.23	34.14	51.56	63.48	81.18	42.14	**55.66**
S-MRL [1]	–	38.46	42.85	12.81	43.20	58.82	75.16	34.39	50.20	63.10	–	–	–
Visual attention based													
AttAlign [19]	UpDn	39.37	43.02	11.89	45.00	–	–	–	–	63.24	80.99	42.55	55.22
HINT [19]	UpDn	46.73	67.27	10.61	45.88	–	–	–	–	63.38	81.18	42.99	**55.56**
SCR [20]	UpDn	49.45	72.36	10.93	**48.02**	–	–	–	–	62.2	78.8	41.6	54.5
Language prior based													
AdvReg. [21]	UpDn	41.17	65.49	15.48	35.48	–	–	–	–	62.75	79.84	42.35	55.16
RUBi [1]	UpDn	44.23	67.05	17.48	39.61	60.73	79.21	**35.62**	50.38	–	–	–	–
RUBi [1]	S-MRL	47.11	68.65	20.28	43.18	60.40	**80.32**	34.60	48.97	61.16	–	–	–
LM [2]	UpDn	48.78	72.78	14.61	45.58	**62.11**	80.29	33.11	**52.94**	63.26	81.16	42.22	55.22
LM+H [2]	UpDn	52.01	72.58	**31.12**	**46.97**	57.71	70.07	30.98	**53.07**	56.35	65.06	37.63	54.69
CF-VQA (Har.)	UpDn	49.94	74.82	18.93	45.42	**62.42**	**80.68**	35.21	52.75	**63.83**	**82.23**	**44.72**	54.88
CF-VQA (Har.)	S-MRL	51.43	78.62	20.07	45.79	61.24	79.51	**35.85**	51.15	62.54	81.21	**44.66**	53.09
CF-VQA (Sum)	UpDn	**53.69**	**91.25**	12.80	45.23	59.82	75.48	33.37	52.25	**63.65**	**82.63**	44.10	54.38
CF-VQA (Sum)	S-MRL	**54.95**	**90.56**	**21.88**	45.36	58.06	74.28	34.85	49.25	60.76	81.11	43.48	49.85

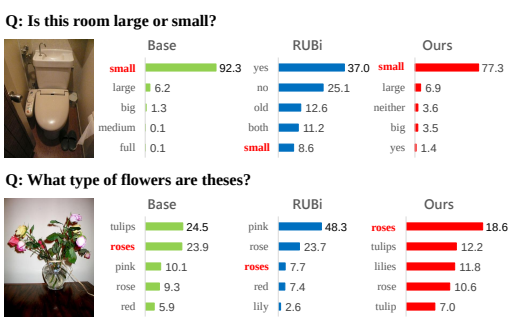


Figure 5: Example results on VQA-CP v2 test split. Red bold answer denotes the ground-truth one.

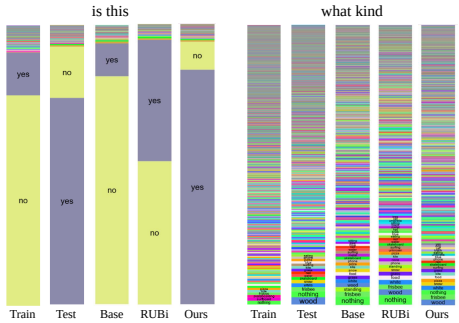


Figure 6: Examples of answer distributions on VQA-CP v2.

Can CF-VQA retain language context? Example results in Figure 5 and 6 illustrate how CF-VQA preserves language context for inference compared to the language prior based approach RUBi [1]. For the first example in Figure 5, CF-VQA recognizes the context "large or small", while RUBi tends to answer yes/no based on "is this". For the second example, although RUBi successfully locates the flowers, it wrongly focuses on visual attributes (*i.e.*, pink) rather than classes (*i.e.*, "what type"). Meanwhile, CF-VQA may highlight category-related options. As shown in Figure 6, for "what brand" questions, RUBi prefers the meaningless answer "none" rather than specific ones. Although CF-VQA cannot recover the answer distribution very well, it attempts to generate answers about kind (*e.g.*, wood, frisbee). These examples highlight the importance of language context, which is the main difference between language prior based approaches and our causal inference based CF-VQA.

6 Conclusion

In this paper, we propose a novel cause-effect look at the language bias in VQA. The language bias is formulated as the direct effect of question on answer, which can be captured by Counterfactual VQA. Experimental results demonstrate the effectiveness and generalizability of our proposed inference strategy CF-VQA. In addition, our proposed cause-effect look fills the theoretical gap in recent debiasing studies [1, 2]. In the future, we will (1) consider a more complex causal graph with external knowledge; (2) extend our cause-effect look to other vision-and-language applications.

Supplementary Materials

This supplementary document is organized as follows:

- Section 7 introduces that RUBi [1] and Learned-Mixin [2] are the special cases of ours from cause-effect view, and how to further improve these methods using our CF-VQA.
- Section 8 describes the implementation details.
- Section 9 describes the supplementary quantitative and qualitative results.

7 Cause-effect View on RUBi [1] and Learned-Mixin [2]

As mentioned in Section 4.3, *RUBi [1] and Learned-Mixin [2] are the special cases of ours* from the cause-effect view, which (1) remove the direct path $V \rightarrow A$ in the causal graph, and (2) use natural indirect effect for inference. The detailed proof is provided as follows.

7.1 Causal Graph

Recent works RUBi [1] and Learned-Mixin [2] 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. Corresponds to this architecture, Figure 7 shows the simplified causal graph by removing the direct path $V \rightarrow A$.

7.2 Cause-Effect Look

According to the simplified causal relations, the total effect (TE) can be decomposed into the natural/pure direct effect (NDE or PDE) 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:

$$\begin{aligned} TE &= Z_{q,k} - Z_{q^*,k^*} \\ NDE &= Z_{q,k^*} - Z_{q^*,k^*} \\ TIE &= TE - NDE = Z_{q,k} - Z_{q,k^*} \end{aligned} \tag{14}$$

which corresponds to Eq. (5) in the main paper. An alternative option is to select the answer with the maximum natural/pure indirect effect (NIE or PDE), which is formulated as:

$$NIE = Z_{q^*,k} - Z_{q^*,k^*} \tag{15}$$

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, which acts as language context.

Note that RUBi [1] and Learned-Mixin (LM) [2] use the following fusion strategies:

$$h(Z_q, Z_k) = Z_k \cdot \sigma(Z_q) \quad \text{(RUBi)} \tag{16}$$

$$h(Z_q, Z_k) = \log \sigma(Z_k) + g(k) \cdot \log \sigma(Z_q) \quad \text{(LM)} \tag{17}$$

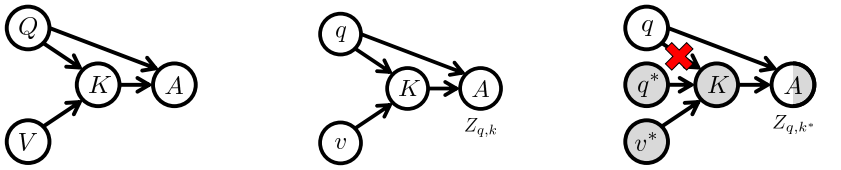
where $\sigma(\cdot)$ represents the sigmoid function and $g(k)$ is a learned function with the knowledge k as input and a weight as output. As for RUBi, NIE is calculated as:

$$NIE = \underbrace{z_k \cdot \sigma(c)}_{Z_{q^*,k}} - \underbrace{c \cdot \sigma(c)}_{Z_{q^*,k^*}} \propto z_k \quad \text{(RUBi)} \tag{18}$$

As for Learned-Mixin, NIE is calculated as:

$$NIE = \underbrace{(\log \sigma(z_k) + g(k) \cdot \log \sigma(c))}_{Z_{q^*,k}} - \underbrace{(\log \sigma(c) + g(k^*) \cdot \log \sigma(c))}_{Z_{q^*,k^*}} \propto z_k \quad \text{(LM)} \tag{19}$$

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 logit of the vision-language branch $\mathcal{F}_{VQ}$. Note that RUBi simply preserves the vision-language branch and uses the logit z_k for inference. Therefore, *RUBi and Learned-Mixin actually use natural indirect effect for inference* from our cause-effect view.



(a) Causal Graph (b) Conventional VQA (c) Counterfactual VQA

Figure 7: (a) Simplified Causal graph for VQA without $V \rightarrow A$. Q: question. V: image. K: multi-modal knowledge. A: answer. (b) & (c) Comparison between conventional VQA and counterfactual VQA. White nodes are at the value $V=v$ and $Q=q$ while gray nodes are at $V=v^*$ and $Q=q^*$.

Algorithm 1 Improving RUBi [1] using CF-VQA

function RUBI(v, q, c, is_Training)
  $z_q = \mathcal{F}_Q(q)$
  $z_k = \mathcal{F}_{VQ}(v,\ q)$
  if is_Training **then**
    $z = z_k \cdot \sigma(z_q)$
  else
    $z = z_k$ ▷ $z = (z_k - c) \cdot \sigma(z_q)$
  end if
  return z
end function

7.3 Improving RUBi [1] and Learned-Mixin [2]

Thanks to our cause-effect look, RUBi and Learned-Mixin can be improved using CF-VQA, *i.e.*, using TIE for inference. Specifically, TIE for RUBi is calculated as:

$$TIE = \underbrace{z_k \cdot \sigma(z_q)}_{Z_{q,k}} - \underbrace{c \cdot \sigma(z_q)}_{Z_{q,k^*}} \quad \text{(RUBi)} \quad (20)$$

where c denotes the constant. TIE for Learned-Mixin is calculated as:

$$\begin{aligned} TIE &= \underbrace{(\log \sigma(z_k) + g(k) \cdot \log \sigma(z_q))}_{Z_{q,k}} - \underbrace{(\log \sigma(c) + g(k^*) \cdot \log \sigma(z_q))}_{Z_{q,k^*}} \\ &= \log \sigma(z_k) + (g(k) - c_g) \cdot \log \sigma(z_q) \end{aligned} \quad \text{(LM)} \quad (21)$$

where $c_g = g(k^*)$ is the constant output for $g(\cdot)$ with void input k^*. The red notes in Algorithm 1 and 2 show how to improve RUBi [1] and Learned-Mixin [2] in implementation. It is worth noting that CF-VQA only changes the inference strategies of RUBi and Learned-Mixin. Therefore, CF-VQA does not require re-training the model.

Algorithm 2 Improving Learned-Mixin [2] using CF-VQA

function LEARNED-MIXIN(v, q, c_g, is_Training)
  $z_q = \mathcal{F}_Q(q)$
  $z_k = \mathcal{F}_{VQ}(v,\ q)$
  $k = K(v, q)$
  if is_Training **then**
    $z = \log \sigma(z_k) + g(k) \cdot \log \sigma(z_q)$
  else
    $z = z_k$ ▷ $z = \log \sigma(z_k) + (g(k) - c_g) \cdot \log \sigma(z_q)$
  end if
  return z
end function

Table 4: 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 5: Ablations of (1) baseline models (SAN, UpDn and S-MRL), (2) fusion strategies (RUBi, Harmonic, Sum) on VQA-CP v1 test split. We reimplement the baselines for fair comparison.

SAN	All	Y/N	Num.	Other
Baseline	32.50	36.86	12.47	36.22
RUBi	49.29	80.43	13.79	**32.58**
+ CF-VQA	**51.15**	**85.75**	**15.02**	31.27
Harmonic	49.29	72.73	**20.57**	37.51
+ CF-VQA	**52.06**	**80.38**	16.88	**38.04**
Sum	38.34	49.88	**15.82**	35.91
+ CF-VQA	**52.87**	**84.94**	14.85	**36.26**

UpDn	All	Y/N	Num.	Other
Baseline	37.08	42.46	12.76	41.50
RUBi	47.90	68.97	13.69	**40.69**
+ CF-VQA	**55.19**	**87.93**	**18.05**	37.54
Harmonic	**55.75**	80.65	**24.72**	43.46
+ CF-VQA	55.16	**82.27**	16.14	**43.87**
Sum	52.78	78.71	14.30	42.45
+ CF-VQA	**57.39**	**88.46**	**14.80**	**43.61**

S-MRL	All	Y/N	Num.	Other
Baseline	36.68	42.72	12.59	40.35
RUBi	40.53	55.94	**16.08**	35.53
+ CF-VQA	**53.34**	**84.15**	15.56	**37.87**
Harmonic	53.55	79.38	17.39	42.38
+ CF-VQA	**55.26**	**82.13**	**18.03**	**43.49**
Sum	49.44	76.49	16.23	35.90
+ CF-VQA	**57.03**	**89.02**	**17.08**	**41.27**

8 Implementation Details

We use the same implementation of RUBi [1] for fair comparison, including feature representation, baseline architectures, and optimization.

Image Representation. Following the popular bottom-up attention mechanism [23], 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 [56, 1], we first lowercase all the questions and remove the punctuation, and then use the pretrained Skip-thought encoder [57] 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 [11], UpDn [23], and a simplified version of the recent architecture MUREL [56] (S-MUREL) proposed in [1]. In short, S-MUREL consists of a BLOCK [58] bilinear fusion between image and question representations for each region, and a 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 [1].

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.

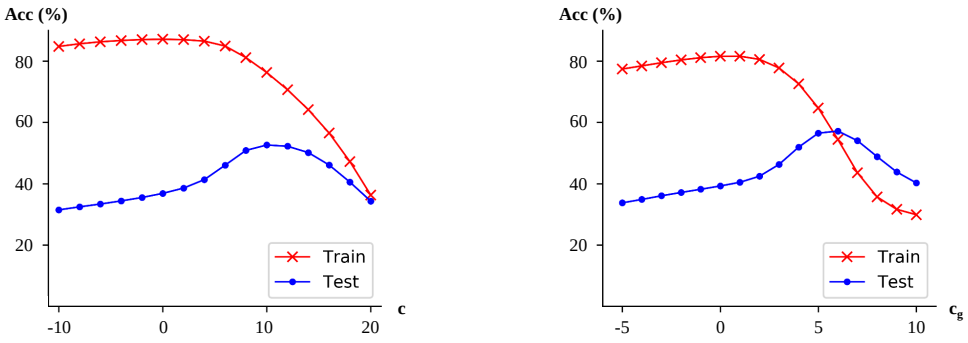
Datasets. The experiments are conducted on VQA-CP [16] and VQA [15] datasets. VQA-CP v1 and v2 are created by re-organizing the train and val splits of the VQA v1 and v2 datasets, respectively [16]. Therefore, there is no overlap between VQA-CP v2 train split and VQA v2 test-std split, which are used in our proposed independent-distribution setting. The details of datasets are shown in Table 4.

Table 6: Ablations of (1) baseline models (SAN, UpDn and S-MRL), (2) fusion strategies (RUBi, Harmonic, Sum) on VQA-CP v2 test split using the simplified VQA causal graph. We reimplement the baselines for fair comparison.

SAN	All	Y/N	Num.	Other
Baseline	33.18	38.57	12.25	36.10
RUBi	37.37	53.08	17.29	34.65
+ CF-VQA	**49.35**	**88.43**	**18.65**	**37.28**
Harmonic	45.48	71.32	17.19	39.71
+ CF-VQA	**49.43**	**83.82**	**17.52**	**40.16**
Sum	42.35	62.33	**16.64**	38.94
+ CF-VQA	**49.85**	**87.75**	16.15	**39.24**

UpDn	All	Y/N	Num.	Other
Baseline	37.69	43.17	12.53	41.72
RUBi	42.78	56.48	17.55	**42.52**
+ CF-VQA	**51.33**	**91.13**	**23.16**	38.20
Harmonic	46.50	67.54	12.83	44.72
+ CF-VQA	**49.53**	**77.02**	**12.86**	**45.18**
Sum	47.10	70.00	12.80	44.51
+ CF-VQA	**53.55**	**91.15**	**12.81**	**45.02**

S-MRL	All	Y/N	Num.	Other
Baseline	37.09	41.39	12.46	41.60
RUBi	47.43	69.11	19.78	**43.66**
+ CF-VQA	**52.62**	**87.14**	**23.69**	42.47
Harmonic	49.57	72.31	20.28	45.68
+ CF-VQA	**52.68**	**82.05**	**20.76**	**46.04**
Sum	49.42	74.43	20.52	44.24
+ CF-VQA	**54.52**	**90.69**	**21.84**	**44.53**



(a) RUBi (b) Learned-Mixin

Figure 8: Accuracies (%) of RUBi and Learned-Mixin on VQA-CP v2 with different c and c_g.

9 Supplementary Experimental Results

We have conducted the ablation study and sensitivity analysis in the main paper to evaluate the generalisability and robustness of our proposed CF-VQA. In this section, we show additional quantitative and qualitative results.

9.1 Ablations

Table 5 shows the ablation study on VQA-CP v1 test split. As shown in Table 5, CF-VQA is very general to both *baseline VQA architectures* and *fusion strategies*, which is also concluded from the results on VQA-CP v2. Specifically, CF-VQA improves RUBi, Harmonic and Sum in most cases. The exception is using UpDn baseline architecture with Harmonic fusion strategy, where using CF-VQA slightly decreases the accuracy from 55.75% to 55.16%.

Table 6 shows the ablation study on VQA-CP v2 test split using the simplified causal graph. Similar to the results using the complete causal graph, CF-VQA achieves significant improvement for different baseline VQA architectures and fusion strategies.

Table 7: Improving RUBi [1] and Learned-Mixin (LM) [2] using our CF-VQA on VQA-CP v2 test split. We reimplement these methods for fair comparison.

	All	Y/N	Num.	Other
RUBi [1]	47.43	69.11	19.78	**43.66**
+ CF-VQA	**52.62**	**87.14**	**23.69**	42.47
LM [2]	48.36	71.83	14.68	45.31
+ H [2]	52.63	72.76	33.19	47.42
+ CF-VQA	**57.18**	**80.18**	**45.62**	**48.31**

9.2 Improving RUBi [1] and Learned-Mixin [2]

As discussed in Section 7.3, CF-VQA can further improve language prior based works RUBi [1] and Learned-Mixin (LM) [2] by replacing NIE with TIE for inference. Table 7 shows that CF-VQA could improve RUBi by over 5% and improve LM by nearly 9%. As shown in Algorithm 1 and 2,

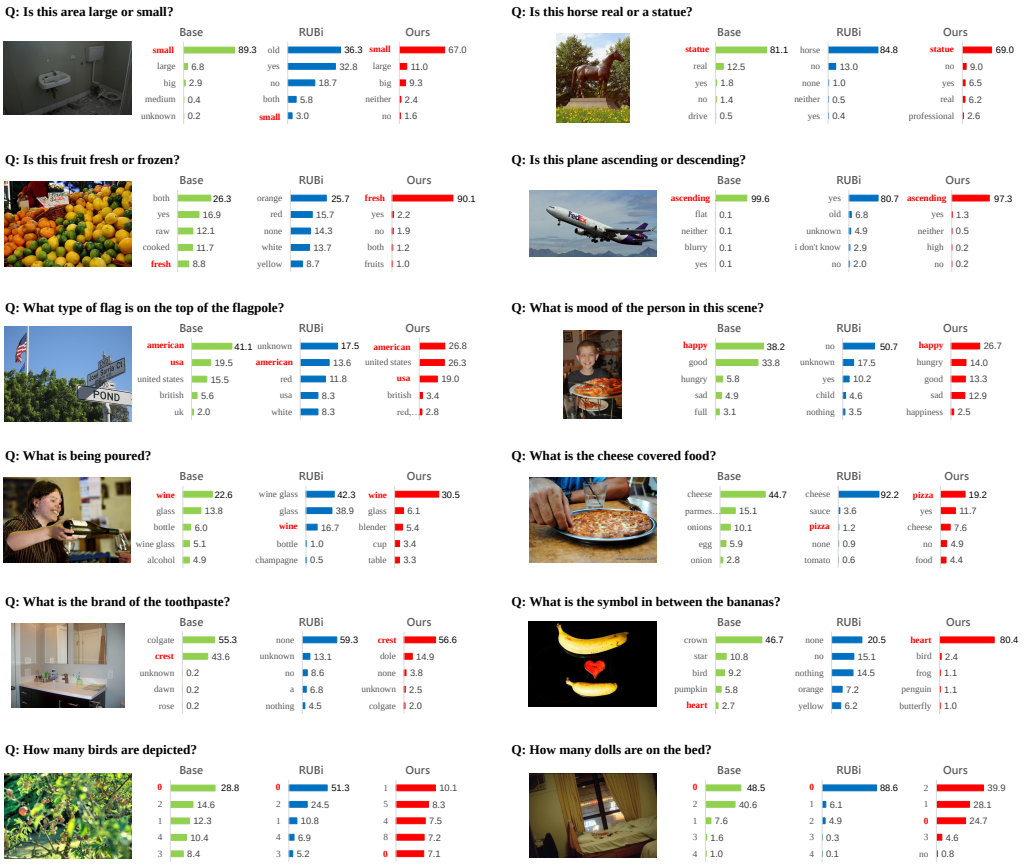


Figure 9: Supplementary example results on VQA-CP v2 test split. Red bold answer denotes the ground-truth one.

CF-VQA only makes influences during the test stage. Therefore, CF-VQA could improve RUBi and LM without re-training the model, and the hyper-parameter can be selected on one trained model. As a comparison, LM+H [2] needs to re-train LM model with different entropy penalty, which leads to a high computation cost. However, as shown in Figure 8(a) and 8(a), we cannot select the constant c and c_g based on the training performance. Therefore, these two fusion strategies are not preferred in practice.

9.3 Additional Qualitative Results

Figure 9 shows supplementary examples results about how our CF-VQA preserves language context compared to RUBi [1]. Our CF-VQA can recognize language context like "fresh or frozen", "mood", "poured" for answer selection. Although RUBi successfully focuses on the right visual objects, *e.g.*, orange, wine glass, and cheese, the produced answer does not match the question well. These examples further show the importance of language context for answering the questions. Therefore, it is important to simultaneously reducing bad language bias and preserving good language context, which is the shortcoming of language prior based methods. The last row shows two failure cases about "Number" questions, especially when there is no related objects in the image. These examples and quantitative results indicate that "Number" questions are still difficult in VQA.

References

[1] 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.

[2] 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.

[3] 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.

[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.

[5] 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.

[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.

[7] 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.

[8] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016.

[9] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 1821–1830, 2017.

[10] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2612–2620, 2017.

[11] 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.

[12] Yang Shi, Tommaso Furlanello, Sheng Zha, and Animashree Anandkumar. Question type guided attention in visual question answering. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 151–166, 2018.

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

[21] 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.

[22] Martin Rohleder, Hendrik Scholz, and Marco Wilkens. Survivorship bias and mutual fund performance: Relevance, significance, and methodical differences. *Review of Finance*, 15(2):441–474, 2011.

[23] 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.

[24] Xu Yang, Hanwang Zhang, and Jianfei Cai. Learning to collocate neural modules for image captioning. *arXiv preprint arXiv:1904.08608*, 2019.

[25] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

[26] Varun Manjunatha, Nirat Saini, and Larry S Davis. Explicit bias discovery in visual question answering models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9562–9571, 2019.

[27] Jiaxin Qi, Yulei Niu, Jianqiang Huang, and Hanwang Zhang. Two causal principles for improving visual dialog. *arXiv preprint arXiv:1911.10496*, 2019.

[28] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. *arXiv preprint arXiv:2002.11949*, 2020.

[29] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. *arXiv preprint arXiv:2003.06576*, 2020.

[30] 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.

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

[32] 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.

[33] Stephen L Morgan and Christopher Winship. *Counterfactuals and causal inference*. Cambridge University Press, 2015.

[34] James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology, 2000.

[35] Victor Chernozhukov, Iván Fernández-Val, and Blaise Melly. Inference on counterfactual distributions. *Econometrica*, 81(6):2205–2268, 2013.

[36] Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

[37] Maya L Petersen, Sandra E Sinisi, and Mark J van der Laan. Estimation of direct causal effects. *Epidemiology*, pages 276–284, 2006.

[38] 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.

[39] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

[40] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3076–3085. JMLR. org, 2017.

[41] Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation learning for treatment effect estimation from observational data. In *Advances in Neural Information Processing Systems*, pages 2633–2643, 2018.

[42] Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *The Journal of Machine Learning Research*, 14(1):3207–3260, 2013.

[43] Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione. Regret minimization in games with incomplete information. In *Advances in neural information processing systems*, pages 1729–1736, 2008.

[44] Michael Oberst and David Sontag. Counterfactual off-policy evaluation with gumbel-max structural causal models. *arXiv preprint arXiv:1905.05824*, 2019.

[45] Vedika Agarwal, Rakshith Shetty, and Mario Fritz. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. *arXiv preprint arXiv:1912.07538*, 2019.

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

[47] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense r-cnn. *arXiv preprint arXiv:2002.12204*, 2020.

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

[49] 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.

[50] 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.

[51] 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.

[52] Atsushi Kanehira, Kentaro Takemoto, Sho Inayoshi, and Tatsuya Harada. Multimodal explanations by predicting counterfactuality in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8594–8602, 2019.

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

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

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

[56] 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.

[57] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302, 2015.

[58] Hedi Ben-Younes, Rémi Cadene, Nicolas Thome, and Matthieu Cord. Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection. *arXiv preprint arXiv:1902.00038*, 2019.