Overcoming Language Priors for Visual Question Answering via Loss Rebalancing Label and Global Context

Runlin Cao[1,2] **Zhixin Li***[1,2]

[1]Key Lab of Education Blockchain and Intelligent Technology, Ministry of Education,
Guangxi Normal University, Guilin 541004, China
[2]Guangxi Key Lab of Multi-source Information Mining and Security, Guangxi Normal University, Guilin 541004, China

Abstract

Despite the advances in Visual Question Answering (VQA), many VQA models currently suffer from language priors (i.e. generating answers directly from questions without using images), which severely reduces their robustness in real-world scenarios. We propose a novel training strategy called Loss Rebalancing Label and Global Context (LRLGC) to alleviate the above problem. Specifically, the Loss Rebalancing Label (LRL) is dynamically constructed based on the degree of sample bias to accurately adjust losses across samples and ensure a more balanced form of total losses in VQA. In addition, the Global Context (GC) provides the model with valid global information to assist the model in predicting answers more accurately. Finally, the model is trained through an ensemble-based approach that retains the beneficial effects of biased samples on the model while reducing their importance. Our approach is model-agnostic and enables end-to-end training. Extensive experimental results show that LRLGC (1) improves performance for various VQA models and (2) performs competitively in the VQA-CP v2 benchmark test.

1 INTRODUCTION

Visual Question Answering (VQA) aims to answer questions based on a given image. It is a multimodal task that combines vision and language. The fundamental approach involves identifying the image region that is most relevant to the question and generating the most suitable answer by leveraging the information within the image. In recent years, VQA has made significant progress, thanks to the ongoing advancements in computer vision and natural language processing [Antol et al., 2015, Zhang et al., 2019, Anderson et al., 2018, Yu et al., 2019, Huang et al., 2020].

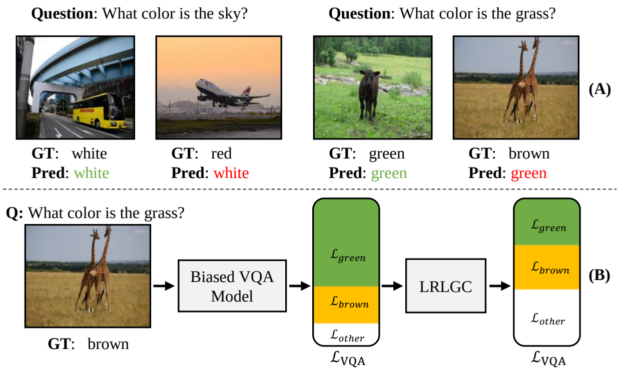

Figure 1: (A) Displays the language priors in the VQA model. The model always tends to predict answers with larger samples in the dataset, e.g., “white” sky and “green” grass. (B) LRLGC can overcome the class imbalance in the VQA dataset by rescaling the total VQA loss to a more balanced form.

Despite the extant models performing well on many VQA datasets, such as VQA v2 [Goyal et al., 2017], most machine learning datasets are inevitably biased. Some researchers have recently found that most existing VQA models rely heavily on language priors, which are apparent statistical correlations between questions and answers [Agrawal et al., 2016, Goyal et al., 2017, Zhang et al., 2016, Hudson and Manning, 2019, Agrawal et al., 2018]. They always tend to ignore the image content in predicting the correct answer. As shown in Fig. 1(A), for questions like “What color is...” the model always tends to answer the more distributed answers in the training set and ignore the content of the images. For example, in the training set, “white” skies are more common than “red” skies, and “green” grass is more common than “brown” grass. This fragile generalization becomes very poor when the distribution of answers in the training and test sets is different, which greatly limits the application of

[*] Zhixin Li is the corresponding author.

Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023), PMLR 216:249–259.

existing models in the real world. To alleviate the language priors problem in VQA models, Agrawal et al. [Agrawal et al., 2018] suggested the VQA-CP dataset, which features question-answer distributions that vary between training and test splits. As the majority of the state-of-the-art VQA models acquire language biases from the training data, their accuracy on the VQA-CP dataset is significantly reduced.

In order to overcome the adverse effects of language priors on VQA models, the research done in recent years can be broadly classified into three categories. The first are annotation-based approaches [Selvaraju et al., 2019, Wu and Mooney, 2019], which use additional visual information to increase the significance of the image, i.e. they try to match the visual attention of the VQA model with human visual attention to ensure that the model can successfully use visual information. However, annotation-based methods are expensive to annotate manually [Selvaraju et al., 2019, Wu and Mooney, 2019], and these annotations are scarce and not readily available. Moreover, recent work [Teney et al., 2020b, Shrestha et al., 2020] has demonstrated that the improvement in accuracy arises from regularization rather than an improved visual basis. The second category is data balance methods [Zhu et al., 2020, Chen et al., 2020, Teney et al., 2020a], which balance the dataset's bias by constructing new training samples. Counterfactual data augmentation techniques [Abbasnejad et al., 2020, Chen et al., 2020, Liang et al., 2020, Teney et al., 2020a] balance the training data by constructing counterfactual samples. To balance the training data, [Zhu et al., 2020, Wen et al., 2021] used a self-supervised framework to generate irrelevant question-image combinations. The data-balance methods generally perform well and do not require additional manual annotation. However, the data-balance methods are likely to introduce new biases due to the inability to guarantee the quality of data generation. Also, the increase in training samples leads to longer training time.

However, it is still a major challenge to make VQA models generalize well under unbalanced training data. Unlike the methods mentioned above, the third type of method is the ensemble-based method [Cadene et al., 2019, Niu et al., 2021, Liang et al., 2021, Lao et al., 2021, Clark et al., 2019, Mahabadi et al., 2020], which is a more efficient solution. Ensemble-based methods do not require additional manual annotations and do not require the generation of new training data. It usually uses an ensemble strategy to combine the predicted outputs of the bias-only model and the VQA model to derive the training gradient based on the fused answers. However, we believe the previous ensemble-based methods have shortcomings: 1) they tend to overcorrect for language biases. Because they do not discriminate the degree of bias of samples well, they also take larger penalties for less biased samples. 2) Few models effectively use the global context. Fusing context and local information, models can predict more accurately. 3) Most of them gain in out-of-distribution on the VQA-CP v2 dataset at the cost of degrading the model's in-distribution performance on the VQA v2 dataset. Ideally, a robust VQA model should overcome the language priors while maintaining its performance on the in-distribution dataset.

Inspired by [Liang et al., 2021, Guo et al., 2022], we consider it crucial to rebalance the proportion of the loss value of each answer in the total VQA loss (cf. Fig. 1(B)). We propose a novel model-agnostic training scheme called Loss Rebalancing Label and Global Context (LRLGC), as shown in Fig. 2. It can overcome language priors and fully use global context. It mainly consists of three modules. (1) Loss Rebalancing Label Module: LRLGC uses the bias-only model's prediction output and ground-truth semantic similarity to determine sample bias. The Loss Rebalancing Label (LRL) is dynamically constructed based on the sample bias. It can assign lower weights to biased samples and ensure a more balanced total VQA loss. (2) Global Context Module: We propose the Global Context (GC) module to utilize the global context effectively. It focuses on globally valid information in images and questions and retains beneficial context priors in biased samples. (3) Ensemble Training Module: We use the ensemble training method [Cadene et al., 2019, Liang et al., 2021, Clark et al., 2019] to merge the debiased and global context module's predicted outputs into a single training. By training with an ensemble-based approach, the beneficial effects of biased samples on the model are preserved while their importance is reduced. Following [Cadene et al., 2019, Liang et al., 2021, Clark et al., 2019], our approach only keeps the base VQA module.

This paper's contributions are summarized as follows:

- We propose a novel model-agnostic generic framework LRLGC that enables end-to-end training and can be easily integrated into various VQA models.
- We propose LRL and Global Context Module, which can effectively help the model overcome the language priors while preserving the contextual information.
- Experimental results show that LRLGC achieves competitive performance on the bias-sensitive VQA-CP v2 (60.91%) without sacrificing performance on the in-distribution VQA v2 (60.81%).

2 RELATE WORK

2.1 VISUAL QUESTION ANSWERING

VQA attempts to understand visual content and natural language questions in order to predict appropriate answers [Antol et al., 2015]. With the increasing demand for multimodal information understanding and the potential of VQA for its powerful applications, VQA tasks have recently attracted a lot of attention in recent years [Antol et al., 2015, Anderson et al., 2018, Yang et al., 2016]. VQA has made significant progress and achieved good results in real images due to the development of deep learning techniques and the proposal of large-scale VQA datasets [Antol et al., 2015, Goyal et al., 2017, Hudson and Manning, 2019]. Existing methods include attention-based [Anderson et al., 2018, Gao et al., 2018, Kim et al., 2018], graph-based [Huang et al., 2020, Li et al., 2019, Khademi, 2020], and knowledge-based [Wu et al., 2016, Zhang et al., 2020]. However, most existing models remember language priors during training and neglect visual information, resulting in poor performance on a test set from a different domain.

2.2 OVERCOMING LANGUAGE PRIORS IN VQA

Many studies [Agrawal et al., 2018, Cadene et al., 2019, Selvaraju et al., 2019, Zhu et al., 2020, Niu et al., 2021, Ouyang et al., 2022, Clark et al., 2019, Chen et al., 2020] have found that the majority of VQA models have serious language priors, which greatly limit the ability of VQA models to understand and generalize multimodal information. Although most existing models can achieve good results on datasets with the same answer distribution for both training and test sets, applying them to real-world scenarios is difficult due to the models' fragile generalization capabilities. In recent years, much work has been proposed to overcome language priors in VQA. These methods can be divided into three categories: annotation-based methods, data-balanced methods, and ensemble-based methods.

Annotation-Based Methods. The annotation-based method has shown its effectiveness in improving model generalization under external visual supervision. The importance of image regions is increased by HINT [Selvaraju et al., 2019] using the annotation of the VQA-HAT [Das et al., 2017] dataset. SCR [Wu and Mooney, 2019] matches correct answers and influential image regions to human text interpretation, thus reducing the sensitivity of incorrect answers to influential objects. The annotation-based method strengthens the visual foundation by introducing additional human visual supervision. Annotation-based methods focus on strengthening the visual foundation by introducing additional human visual supervision, but all of them require manual annotation, which is very expensive. Furthermore, the study [Shrestha et al., 2020] showed that the performance improvement is not a result of visual basis improvement but rather a regularization effect of preventing overfitting the language priors.

Data-Balanced Methods. SSL-VQA [Zhu et al., 2020] introduces a self-supervised framework to balance data bias by replacing relevant question-image pairs with irrelevant ones to generate additional data. CSS [Chen et al., 2020] generates counterfactual training samples by masking critical objects in the images and words in the questions and assigning different ground true answers. Based on CSS [Chen et al., 2020], CL-VQA [Liang et al., 2020] constructs positive and negative samples for counterfactual samples and uses contrast learning for training. Augmenting the data to balance dataset biases does not require additional manual annotation. However, the additional data generated may introduce new biases and make it challenging to ensure the quality of the generated data.

Ensemble-Based Methods. The ensemble-based approach attempts to include an additional branch to account for language priors in order to mitigate their negative impact on the model. AdvReg [Ramakrishnan et al., 2018] uses an adversarial learning approach to prevent VQA models from capturing language biases in their question encoding. RUBi [Cadene et al., 2019] and LMH [Clark et al., 2019] are fusion-based methods. This approach combines the two predicted outputs of the VQA model and the question-only branch together and serves as the final output of the VQA model in the training phase. It effectively prevents the VQA model from using bias for answer prediction. LPF [Liang et al., 2021] and LP-Focal [Lao et al., 2021] use the bias model's output distribution to reduce the bias samples' weight when calculating the VQA loss.

However, the ensemble-based method also compromises the ability of the model to learn context to some extent [Yang et al., 2021]. To improve this problem, we propose our LRLGC training strategy that can reduce language biases while preserving the model's ability to learn context.

3 METHOD

3.1 BASE VQA MODULE

We denote the VQA dataset with N training instances as $\mathcal{D} = \{I_i, Q_i, a_i\}_{i=1}^{N}$, where $I_i \in \mathcal{I}$, $Q_i \in \mathcal{Q}$ and $a_i \in \mathcal{A}$ denote the i^{th} instance image, question, and ground truth answer. The visual encoder e_v and the question encoder e_q encode I_i and Q_i to generate the embedding vectors $v_i = e_v\left(I_i\right)$ and $q_i = e_q\left(Q_i\right)$, respectively. The goal of the VQA model is to train a mapping function $f_{VQA} : \mathcal{I} \times \mathcal{Q} \rightarrow \mathcal{R}^{\mathcal{A}}$ that produces a correct distribution across answer space $\mathcal{A}$. The VQA work can generally be considered a multi-class classification task Anderson et al. [2018], Kim et al. [2018]. We train the VQA model using binary cross-entropy loss to optimize its learning parameters:

$$P_{VQA}\left(\mathcal{A}|v_i, q_i\right) = Softmax\left(f_{VQA}\left(\mathcal{A}|v_i, q_i\right)\right) \quad (1)$$

$$\begin{aligned} \mathcal{L}_{VQA} = -\frac{1}{N}\sum_{i=1}^{N} t_i log\left(\sigma\left(f_{VQA}\left(\mathcal{A}|v_i, q_i\right)\right)\right) \\ + \left(1 - t_i\right) log\left(1 - \sigma\left(f_{VQA}\left(\mathcal{A}|v_i, q_i\right)\right)\right) \end{aligned} \quad (2)$$

where soft target score t_i is denoted by $t_i \in [0, 1]^{\|\mathcal{A}\|}$ for a_i, and $\sigma\left(\cdot\right)$ denotes the sigmoid function.

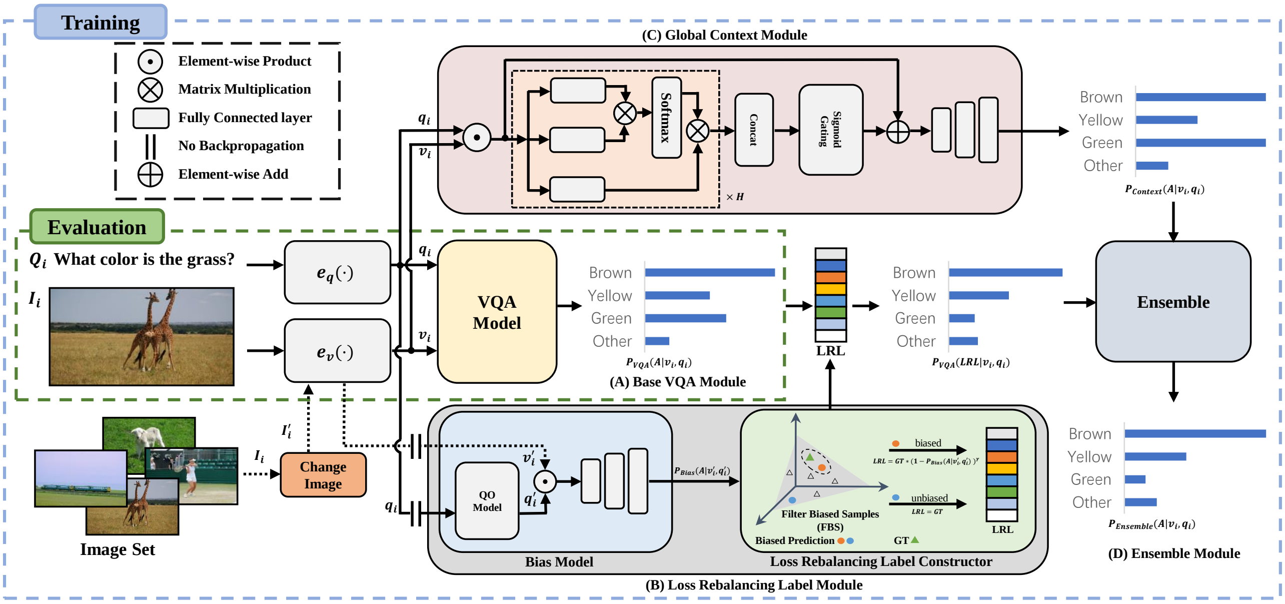

Figure 2: Overview of LRLGC training strategy. (A) An arbitrary VQA model. (B) A Bias Model captures language biases, and the Loss Rebalancing Label Constructor dynamically generates Loss Rebalancing Labels (LRL) for each biased sample. (C) A gated multi-headed self-attention mechanism captures global context. (D) Learning by the ensemble.

3.2 LOSS REBALANCING LABEL MODULE

In this part, we try to construct Loss Rebalancing Labels (LRL) for each biased sample and use the LRL to train the VQA model to reduce the negative impact of biased samples, i.e., to reduce the negative bias of multimodal features $f_{VQA}\left(\mathcal{A}|v_i, q_i\right)$. Before that, we need to capture the presence of language biases in the training samples.

Bias Model captures the biases. An intuitive method is to train a unimodal model that accepts only one of the two modes as input to capture the biases in the VQA dataset [Cadene et al., 2019]. It is common practice to use the question-only model as a branch of the VQA model [Cadene et al., 2019, Liang et al., 2021, Lao et al., 2021], but this unimodal feature contains only language modality information and lacks the use of image information. Inspired by Zhu et al. [2020], we swap the original image I with image I', which is chosen randomly from the image set $\mathcal{I}$. Considering the vast size of $\mathcal{I}$, the probability that (Q, I') are related is extremely remote, i.e., the input question and the image are unrelated. A comparison of the effects of the question-only model and the bias model is shown in Table 4. Specifically, our bias model $f_{Bias} : \mathcal{R}^{d_v} \times \mathcal{R}^{d_q} \to \mathcal{R}^{\mathcal{A}}$ consists of a question-only model $f_{QO} : \mathcal{R}^{d_q} \to \mathcal{R}^{d_q}$, which can be formalized as:

$$q_i^{'} = f_{QO}(q_i), v_i^{'} = e_v(I_i^{'}) \tag{3}$$

$$f_{Bias}(\mathcal{A}|v_i^{'}, q_i^{'}) = clf_{Bias}(m_{Bias}(v_i^{'} \odot q_i^{'})) \tag{4}$$

where $\odot$ denotes element-wise product and $m_{Bias} : \mathcal{R}^{d_v} \times$ $\mathcal{R}^{d_v} \to \mathcal{R}^{d_m}$ denotes multi-layer perceptron (MLP), and $clf_{Bias} : \mathcal{R}^{d_m} \to \mathcal{R}^{A}$ denotes the classifier. Formally, binary cross-entropy loss is used to optimise the parameters of the bias model and the question-only model:

$$\begin{aligned}\mathcal{L}_Q = -\frac{1}{N}\sum_{i=1}^{N} t_i log\left(\sigma\left(f_{QO}\left(q_i\right)\right)\right) \\ + \left(1 - t_i\right) log\left(1 - \sigma\left(f_{QO}\left(q_i\right)\right)\right)\end{aligned} \tag{5}$$

$$\begin{aligned}\mathcal{L}_B = -\frac{1}{N}\sum_{i=1}^{N} t_i log\left(\sigma\left(f_{Bias}\left(\mathcal{A}|v_i', q_i'\right)\right)\right) \\ + \left(1 - t_i\right) log\left(1 - \sigma\left(f_{Bias}\left(\mathcal{A}|v_i', q_i'\right)\right)\right)\end{aligned} \tag{6}$$

Filter Biased Samples (FBS). According to [Cadene et al., 2019, Ouyang et al., 2022], we classify the training samples as biased and unbiased. Among them, biased samples are those in which the model relies on the questions alone to predict the answer, while unbiased samples require images and questions to infer the answer.

$$sim(\mathrm{f}_{Bias}, \mathrm{f}_{GT}) = \frac{{\mathrm{f}_{\mathrm{Bias}}}^{\top}\mathrm{f}_{GT}}{\|\mathrm{f}_{Bias}\| \, \|\mathrm{f}_{GT}\|} \tag{7}$$

where $sim\left(\cdot, \cdot\right)$ denotes a similarity score function, and a higher value indicates a higher probability that the sample is biased. In practice, we use the similarity function made by cosine similarity. f_{Bias} and f_{GT} denote the vectors for word embedding using Glove 300, where f_{Bias} is the predicted answer for the bias model and f_{GT} is the ground true answer. We assume that all candidate answers are independent of each other. Therefore they cannot be effectively modelled by a recurrent neural network (e.g., GRU) similar to the question sentence. For answers containing multiple words (e.g., “knife and spoon”), we use the sum of these word vectors to represent them.

The VQA dataset answer types are “yes/no”, “number” and “other”. When the answer type is “other”, we consider a sample biased if f_{Bias} has a similar semantic space distance to f_{GT}. Unlike the “other” questions, the predicted answers to the “yes/no” and “number” questions have a strong semantic correlation with the underlying ground true answers, which is not conducive to determining biased samples through similarity in the semantic space. Therefore, the sample is biased when the answer type is “y/n” or “num” and f_{Bias} equals f_{GT}.

$$\text{biased sample} \in \begin{cases} sim(\mathrm{f}_{Bias}, \mathrm{f}_{GT}) > \alpha, \; at \text{ is other} \\ \mathrm{f}_{Bias} = \mathrm{f}_{GT}, \; at \text{ is y/n or num} \end{cases} \tag{8}$$

where at denotes the answer type and α is a hyperparameter.

Construct LRL and train the model using LRL. The bias model captures the language biases, and we want to use the language biases to construct LRLs for biased samples. LRL helps the model to focus more on the training samples that only make the language modal and cannot answer correctly. If the training samples are unbiased, the LRLs are ground true answers. In contrast, for the biased samples, we use LRL as the label to train the model, thus reducing the harmful effects of biased samples on the model.

$$LRL = \begin{cases} t_i \left(1 - P_{Bias}\left(\mathcal{A}|v_i', q_i'\right)\right)^{\beta}, & biased \\ t_i, & unbiased \end{cases} \tag{9}$$

$$P_{Bias}(\mathcal{A}|v_i^{'}, q_i^{'}) = Softmax(f_{Bias}(\mathcal{A}|v_i^{'}, q_i^{'})) \tag{10}$$

where β is a hyperparameter. A larger β indicates a stronger penalty for that biased sample. Using LRLs for training, we can obtain the model debiased predicted output $P_{VQA}\left(LRL|v_i, q_i\right)$.

3.3 GLOBAL CONTEXT MODULE

Multi-head Self-Attention. We use a multi-headed self-attention mechanism [Vaswani et al., 2017] to capture the correlation between the image and the question.

$$Att(Q, K, V) = Softmax(\frac{QK^{\top}}{\sqrt{d_k}})V \tag{11}$$

where $Att\left(\cdot,\cdot,\cdot\right)$ indicates self-attention mechanism. $Softmax\left(\cdot\right)$ operation is performed for each row. Q, K and V denote query, key and value respectively and $\sqrt{d_k}$ is the channel number of Q and K.

$$h_i = Att\left(XW_i^Q, XW_i^K, XW_i^V\right) \tag{12}$$

$$f_{MSA}\left(X\right) = Concat\left(h_1, ..., h_H\right) + X \tag{13}$$

where h_i refers to the output of the i^{th} head, and H denotes the number of heads. $Concat\left(\cdot\right)$ represents concatenating the results of multiple heads.

Gate Mechanism. We can make visual modality and text modality interact and integrate better through $f_{MSA}\left(\cdot\right)$. However, the visual and textual modalities of $f_{MSA}\left(\cdot\right)$ projections may contain noisy or meaningless information. We added a sigmoid-gating mechanism to pass information adaptively and suppress useless details [Cadene et al., 2019]. It consists in multiplying the output of a newly added multi-head self-attention layer by $sigmoid\left(w_g\right)$ before adding it to the input representation from the residual connection, where w_g is a learnable scalar and is initialized at 0.

$$\begin{aligned} f_{Context}\left(\mathcal{A}|v_i, q_i\right) &= y + \sigma\left(w_g\right) \\ &\times f_{MSA}\left(y\right), where\; y = v_i \odot q_i \end{aligned} \tag{14}$$

$$P_{Context}\left(\mathcal{A}|v_i, q_i\right) = Softmax\left(f_{Context}\left(\mathcal{A}|v_i, q_i\right)\right) \tag{15}$$

This gating mechanism improves both the stability of the training and the final performance. Note that here v_i is not calculated with q_i for local attention, because we want to have access to all information that may affect context features $f_{Context}$.

We convert $sim(\mathrm{f}_{Bias}, \mathrm{f}_{GT})$ to a binary vector b_i as the label for computing $\mathcal{L}_C$ to learn context priors with language biases, as shown below:

$$\begin{aligned} \mathcal{L}_C = -\frac{1}{N}\sum_{i=1}^{N} & b_i log\left(\sigma\left(f_{Context}\left(\mathcal{A}|v_i, q_i\right)\right)\right) \\ & + \left(1 - b_i\right) log\left(1 - \sigma\left(f_{Context}\left(\mathcal{A}|v_i, q_i\right)\right)\right) \end{aligned} \tag{16}$$

$$b_i = \begin{cases} 1, & sim(\mathrm{f}_{Bias}, \mathrm{f}_{GT}) \geq \gamma \\ 0, & sim(\mathrm{f}_{Bias}, \mathrm{f}_{GT}) < \gamma \end{cases} \tag{17}$$

Based on empirical values, we fixed γ to 0.1.

3.4 ENSEMBLE TRAINING MODULE

To combine the debiased prediction output with contextual information, inspired by [Clark et al., 2019], we trained an ensemble of $f_{VQA}\left(LRL|v_i, q_i\right)$ and $f_{Context}\left(\mathcal{A}|v_i, q_i\right)$ and computed a new prediction distribution $P_{Ensemble}\left(\mathcal{A}|v_i, q_i\right)$.

$$\begin{aligned} f_E\left(\mathcal{A}|v_i, q_i\right) &= log\left(f_{VQA}\left(LRL|v_i, q_i\right)\right) \\ &+ log\left(f_{Context}\left(\mathcal{A}|v_i, q_i\right)\right) \end{aligned} \tag{18}$$

$$P_{Ensemble}\left(\mathcal{A}|v_i, q_i\right) = Softmax\left(f_E\left(\mathcal{A}|v_i, q_i\right)\right) \tag{19}$$

A binary cross-entropy loss can be used to optimize the parameters of the $\mathcal{L}_E$:

$$\begin{aligned}\mathcal{L}_E = -\frac{1}{N}\sum_{i=1}^{N} t_i log\left(\sigma\left(f_E\left(\mathcal{A}|v_i, q_i\right)\right)\right) \\ +\left(1-t_i\right) log\left(1-\sigma\left(f_E\left(\mathcal{A}|v_i, q_i\right)\right)\right)\end{aligned} \tag{20}$$

Note that all other modules are involved in the training stage. All extra modules are removed during the inference stage, and we use only $f_{VQA}\left(\cdot\right)$ to make accurate predictions.

Finally, the total loss function can be defined as follows:

$$\mathcal{L} = \mathcal{L}_{VQA} + \mathcal{L}_Q + \mathcal{L}_B + \mathcal{L}_C + \mathcal{L}_E \tag{21}$$

4 EXPERIMENTS

4.1 DATASETS

We evaluate our method using standard evaluation metrics Antol et al. [2015] on the most widely used out-of-distribution benchmark VQA-CP v2 Agrawal et al. [2018] test set and the standards-based in-distribution VQA v2 Goyal et al. [2017] validation set. The VQA-CP v2 training and test sets contain around 121k and 98k images and 245k and 220k questions, respectively.

4.2 BASELINES

To demonstrate the effectiveness of our LRLGC, we used different backbones, including UpDn [Anderson et al., 2018], SAN [Yang et al., 2016], BAN [Kim et al., 2018]. None of these three models is designed to mitigate language priors. In addition, several different methods are compared to our approach: (1) annotation-based methods: AttAlig [Selvaraju et al., 2019], HINT [Selvaraju et al., 2019], SCR [Wu and Mooney, 2019]. (2) data-balanced methods: Unshuffling[Teney et al., 2021], RandImg [Teney et al., 2020b], CSS [Chen et al., 2020], CL-VQA [Liang et al., 2020], SSL-VQA [Zhu et al., 2020]. (3) ensemble-based methods: AdvReg[Ramakrishnan et al., 2018], RUBi [Cadene et al., 2019], LMH [Clark et al., 2019], CF-VQA [Niu et al., 2021],GGE-DQ [Han et al., 2021], LPF [Liang et al., 2021], Loss-Rescaling [Guo et al., 2022], LP-Focal [Lao et al., 2021], CCB-VQA [Yang et al., 2021], SBS [Ouyang et al., 2022].

4.3 IMPLEMENTATION DETAILS

We use pre-trained Faster-RCNN to extract object features. Specifically, we extract 36 object features with dimensions of 2048 for each image. All questions are padded to the same length 14. Each word vector is embedded with 300-dimensional Glove. Then, they are fed into a single-level GRU to produce a 1280-dimensional representation of the sentence level. Inspired by Zhu et al. [2020], we set a Batch Normalization layer in front of each classifier and use a binary cross-entropy loss to train all branches during training.The Adam optimizer is used with an initial learning rate of 0.001. After ten epochs, we halve the learning rate every five epochs. The batch size is set to 512, and we train our LRLGC for 30 epochs. The $\alpha = 0.5$ and $\beta = 4$ settings are used in all tests in this work. In later sections, we will also study the hyperparameter setting. Besides, we set the Global Context Module to use one layer and 64 headers.

4.4 PERFORMANCE COMPARISON

On the VQA-CP v2 test set and the VQA v2 validation set, our LRLGC and state-of-the-art approaches are compared, and the experimental outcomes are presented in Table 1. The following conclusions can be drawn from these results.

On the VQA-CP v2 test set, (1) data-balanced and ensemble-based methods perform similarly (59.18$\%$ vs. 59.57$\%$), significantly outperforming HINT Selvaraju et al. [2019] and SCR Wu and Mooney [2019] that require additional manual annotation. Although data-balanced methods perform well, they change the training prior, making it difficult to tell if the VQA model is still driven by a memory prior. (2) LRLGC outperforms all compared methods, achieving an advanced performance of 60.91$\%$. Specifically, LRLGC outperforms LMH Clark et al. [2019], CL-VQA Liang et al. [2020], and SBS Ouyang et al. [2022] by approximately 9$\%$,1.7$\%$, and 1.3$\%$, respectively. Results showed LRLGC improved (+21.17$\%$) compared to UpDn Anderson et al. [2018]. Notably, our method doesn't use additional data. (3) For the "Other" questions, LRLGC is consistent with SSL-VQA Zhu et al. [2020] and outperforms the other methods. For the "Yes/No" questions, LRLGC is second to CF-VQA Niu et al. [2021]. These results further validate the effectiveness of our LRLGC training strategy.

The VQA v2 results validate the debiasing strategy on the in-distribution dataset. On VQA v2, most ensemble-based methods do worse than UpDn Anderson et al. [2018]. Although LRLGC performs less well than UpDn Anderson et al. [2018], it still outperforms the majority of ensemble-based models, including LMH Clark et al. [2019], RUBi Cadene et al. [2019], and LPF Liang et al. [2021]. This suggests that LRLGC has some potential to address the overcorrection issue. To avoid significant performance degradation, LRLGC can reduce most VQA CP v2 statistical priors while retaining most VQA v2 global information.

From the combined results of the two datasets, (1) our method can effectively reduce the performance gap on both datasets to 0.1$\%$. (2) Among all the compared methods, our LRLGC has the highest average score of 60.86$\%$ on both datasets. All these results further show that our LRLGC

Table 1: Comparison results for the VQA-CP v2 test split and the VQA v2 validation split. The highest score is displayed in **bold**, while the second-highest score is underlined. All models below use UpDn [Antol et al., 2015] as the backbone. I – IV denote plain methods, methods based on strengthening visual information (annotation-based), methods based on data augmentation (data-balanced), and methods based on training strategies (ensemble-based), respectively.

Case	Model	VQA-CP v2 test				VQA v2 val				Comparison	
		Overall	Yes/No	Number	Other	Overall	Yes/No	Number	Other	Gap↓	Mean
I	SAN [Yang et al., 2016]	24.96	38.35	11.10	21.74	52.41	70.06	39.28	47.84	27.45	38.69
	BAN [Kim et al., 2018]	37.03	41.55	12.43	41.40	63.90	81.42	**45.18**	55.54	26.87	50.47
	UpDn [Anderson et al., 2018]	39.74	42.27	11.93	46.05	63.48	81.18	42.14	**55.66**	23.74	51.61
II	AttAlign [Selvaraju et al., 2019]	39.37	43.02	11.89	45.00	63.24	80.99	42.55	55.22	23.87	51.31
	HINT [Selvaraju et al., 2019]	46.73	67.27	10.61	45.88	63.38	81.18	42.99	55.56	16.65	55.06
	SCR [Wu and Mooney, 2019]	49.45	72.36	10.93	48.02	62.20	78.80	41.60	54.50	12.75	55.83
III	Unshuffling [Teney et al., 2021]	42.39	47.72	14.43	47.24	61.08	78.32	42.16	52.71	18.69	51.74
	RandImg [Teney et al., 2020b]	55.37	83.89	41.60	44.20	57.24	76.53	33.87	48.57	1.87	56.31
	CSS [Chen et al., 2020]	58.95	84.37	49.42	48.21	59.91	73.25	39.77	55.11	0.96	59.43
	CL-VQA [Liang et al., 2020]	59.18	86.99	49.89	47.16	57.29	67.27	38.40	54.71	1.89	58.24
	SSL-VQA [Zhu et al., 2020]	57.59	86.53	29.87	**50.03**	**63.73**	-	-	-	6.14	60.66
IV	AdvReg [Ramakrishnan et al., 2018]	41.17	65.49	15.48	35.48	62.75	79.84	42.35	55.16	21.58	51.96
	RUBi [Cadene et al., 2019]	45.42	63.03	11.91	44.33	58.19	63.04	41.00	54.43	12.77	51.81
	LMH [Clark et al., 2019]	52.01	72.58	31.12	46.97	56.35	65.06	37.63	54.69	4.34	54.18
	CF-VQA [Niu et al., 2021]	53.55	**91.15**	13.03	44.97	63.54	**82.51**	43.96	54.30	9.99	58.55
	GGE-DQ [Han et al., 2021]	57.32	87.04	27.75	49.59	59.11	73.27	39.99	54.39	1.79	58.22
	LPF [Liang et al., 2021]	55.34	88.61	23.78	46.57	55.01	64.87	37.45	52.08	0.33	55.18
	Loss-Rescaling [Guo et al., 2022]	53.26	72.82	48.00	44.46	56.81	68.21	36.37	52.29	3.55	55.04
	LP-Focal [Lao et al., 2021]	58.45	88.34	34.67	49.32	62.45	-	-	-	4.00	60.45
	CCB-VQA [Yang et al., 2021]	59.12	89.12	51.04	45.62	59.17	77.28	33.71	52.14	**0.05**	59.15
	SBS [Ouyang et al., 2022]	59.57	87.44	**52.96**	46.79	61.97	78.80	42.17	54.41	2.40	60.77
	LRLGC (Ours)	**60.91**	89.95	45.13	**50.03**	60.81	77.65	39.25	53.71	0.10	**60.86**

not only decreases training bias but also enhances model robustness.

4.5 ABLATION STUDIES

Effect of different backbones. To demonstrate that our LRLGC works effectively on a variety of VQA models, we built LRLGC frameworks on SAN [Yang et al., 2016], BAN [Kim et al., 2018], and UpDn [Anderson et al., 2018] and ran experiments on the VQA-CP v2. From the results in Table 2, LRLGC significantly improves the model's accuracy regardless of the backbone, indicating that LRLGC is model-agnostic.

Performance on different scales of the dataset. We ran experiments on VQA-CP v2 with varying training sizes to further prove our method's superiority. As shown in Table 3, the percentage results of the training split variables show that our LRLGC improves the three benchmark models by 18.8% on average. Even with less training data (20% and 40%), LRLGC can overcome the language priors and exploit the global context to improve overall average performance (12.7% and 17.4%).

Each LRLGC module's effect on the model performance. We conducted an ablation study on the VQA-CP v2 to demonstrate the effectiveness of each component in our LRLGC. The results are shown in Table 4. The following findings can be drawn from these results: (1) LRL can help the model overcome the language priors (rows 1-3), reducing the proportion of biased samples in the total loss. (2) Including random images in the bias model is better than question-only (rows 2-3), proving the importance of the visual modality in capturing language priors. (3) LRL+FBS has essentially no effect on performance (rows 2-5), demonstrating that unbiased samples need not be overly penalized. (4) Global context can help the model perform better (rows 4-7). In particular, FBS+Context works better (rows 6-9), indicating that biased samples need more context. Overall, these results are evidence of the effectiveness of each component of our LRLGC in the improvement of model performance.

Effect of hyperparameters α and β. As Table 5 shows, we tested different combinations of α and β on the VQA-CP v2 split. α is used to control the judgment threshold of biased samples, and β indicates the strength of penalizing biased samples. An appropriate ratio between α and β will lead to better performance of LRLGC. From the experimental results, the highest performance is the combination of α = 0.5 and β = 4.

Table 2: The effect of different backbones on model performance on the VQA-CP v2 test set.

Model	Yes/No	Number	Other	Overall	Gap↑
SAN† [Yang et al., 2016]	40.86	13.43	46.98	40.08	**+18.48**
SAN+LRLGC	88.03	42.05	47.65	58.56	
BAN† [Kim et al., 2018]	43.53	13.60	46.35	40.53	**+18.66**
BAN+LRLGC	89.85	42.74	47.64	59.19	
UpDn† [Anderson et al., 2018]	43.32	13.41	48.32	41.54	**+19.37**
UpDn+LRLGC	89.95	45.13	50.03	60.91	

Table 3: Results of the LRLGC on the VQA-CP v2 test set with different proportions of training split. † denotes the model we have re-implemented.

Model	Proportion of Training Set				
	20%	40%	60%	80%	100%
SAN† [Yang et al., 2016]	33.15	36.62	39.11	39.71	40.08
SAN+LRLGC	43.80	53.19	56.67	57.13	**58.56**
BAN† [Kim et al., 2018]	33.05	37.28	38.52	40.00	40.53
BAN+LRLGC	42.66	54.16	56.91	58.65	**59.19**
UpDn† [Anderson et al., 2018]	36.37	38.72	39.91	40.53	41.54
UpDn+LRLGC	54.10	57.57	59.02	59.96	**60.91**

Table 4: Each LRLGC module's effect on the model performance. UpDn† as the backbone. And q denotes question-only, qv denotes question and random image.

	LRL	FBS	GC	VQA-CP v2 test (%)
1				41.54
2	q			57.83
3	qv			58.90
4	q	✓		58.17
5	qv	✓		58.77
6	q		✓	59.43
7	qv		✓	59.81
8	q	✓	✓	59.84
9	qv	✓	✓	60.91

Table 5: Results for various α and β combinations.

Model	α vs. β	VQA-CP v2 test (%)
LRLGC	0.1 : 4	59.58
	0.3 : 4	60.08
	0.5 : 4	**60.91**
	0.7 : 4	60.28
	0.5 : 3	60.65
	0.5 : 5	60.08

Table 6: Comparison of LRLGC and other re-weighting methods

Model	Adaptive	q	v	FBS	GC	VQA-CP v2 test (%)
Loss-Rescaling [Guo et al., 2022]		✓				53.26
LPF [Liang et al., 2021]	✓	✓				55.34
LP-Focal [Lao et al., 2021]	✓	✓				58.45
LRLGC (Ours)	✓	✓	✓	✓	✓	60.91

4.6 LRLGC VS. OTHER RE-WEIGHTING METHODS

In this part, we further compare our LRLGC with other re-weighting methods, and the results are shown in Table 6. Loss-Rescaling [Guo et al., 2022] takes full advantage of the spurious statistical relationship between question types and answers, from which bias values are calculated for each sample. However, the bias values are pre-calculated based on the dataset and are not adaptively adjusted during training. Both LPF [Liang et al., 2021] and LP-Focal [Lao et al., 2021] use question-only branches to dynamically capture language biases during training and are able to adaptively adjust loss values for each sample. However, LPF and LP-Focal lack the filtering of biased samples, which can easily lead to over-correction of unbiased samples and degrade the performance in in-distribution datasets. In addition, the introduction of the visual modality captures the bias in the sample more adequately than language modality alone (rows 2-3 of Table 4). Moreover, our LRLGC incorporates global context and, through ensemble-based training, can minimize the negative effects of biased samples on the model while retaining their useful information.

4.7 QUALITATIVE ANALYSIS

We provide qualitative results from the VQA-CP v2 in Fig. 3 to demonstrate our LRLGC's validity further. The prediction results are based on UpDn and LRLGC. These examples cover "yes/no," "num," and "other". We visualize the model's top 3 important regions, output attention weights, and show the top 4 answers. For the question "What color is the mouse pad?", LRLGC accurately located the key visual object and answered. For the "yes/no" questions, "yes" has relatively more priors than other answers. For example,

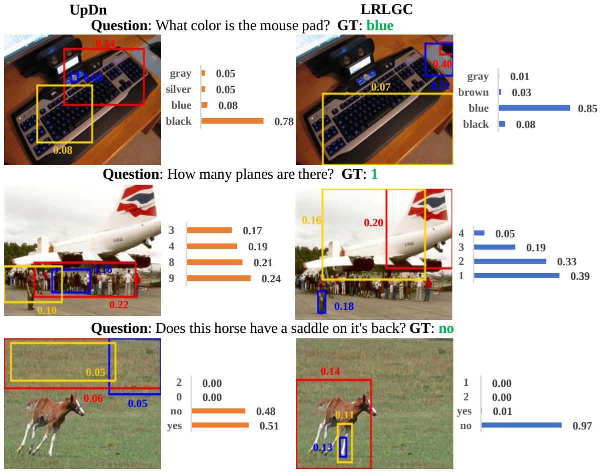

Figure 3: Qualitative comparison results of UpDn [Anderson et al., 2018] and LRLGC on VQA-CP v2 test set.

"Does this horse have a saddle on its back?" the baseline predicted answer is "yes" but it does not locate the critical visual object (there is no "look" image), indicating language priors interference. Even for counting questions like "How many planes are there?" that require visual understanding, LRLGC gives the right answer. By using our LRLGC, the VQA model can avoid overfitting of data biases and show better results on out-of-distribution datasets.

5 CONCLUSION

This paper proposes a general training strategy called LRLGC to address the language priors problem in VQA. LRLGC applies dynamic weighting to each biased sample and integrates global context to guide the model in answering questions. Experimental results show that our method achieves promising results on both VQA-CP v2 and VQA v2. In the future, we plan to improve our method and use it for other multimodal deep-learning tasks with single-peak bias.

Acknowledgements

This work is supported by National Natural Science Foundation of China (Nos. 62276073, 61966004), Guangxi Natural Science Foundation (No. 2019GXNSFDA245018), Innovation Project of Guangxi Graduate Education (YCSW2023141), Guangxi "Bagui Scholar" Teams for Innovation and Research Project, and Guangxi Collaborative Innovation Center of Multi-source Information Integration and Intelligent Processing. (Corresponding author: Zhixin Li.)

References

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.

Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. Analyzing the behavior of visual question answering models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 1955–1960, 2016.

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.

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.

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.

Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. Rubi: Reducing unimodal biases for visual question answering. *Advances in neural information processing systems*, 32:839–850, 2019.

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.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 4067–4080, 2019.

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.

Lianli Gao, Pengpeng Zeng, Jingkuan Song, Xianglong Liu, and Heng Tao Shen. Examine before you answer: Multi-task learning with adaptive-attentions for multiple-choice vqa. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1742–1750, 2018.

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.

Yangyang Guo, Liqiang Nie, Zhiyong Cheng, Qi Tian, and Min Zhang. Loss re-scaling vqa: Revisiting the language prior problem from a class-imbalance view. *IEEE transactions on image processing: a publication of the IEEE Signal Processing Society*, 31:227–238, 2022.

Xinzhe Han, Shuhui Wang, Chi Su, Qingming Huang, and Qi Tian. Greedy gradient ensemble for robust visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1584–1593, 2021.

Qingbao Huang, Jielong Wei, Yi Cai, Changmeng Zheng, Junying Chen, Ho-fung Leung, and Qing Li. Aligned dual channel graph convolutional network for visual question answering. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7166–7176, 2020.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

Mahmoud Khademi. Multimodal neural graph memory networks for visual question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7177–7188, 2020.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. *Advances in neural information processing systems*, 31:1571–1581, 2018.

Mingrui Lao, Yanming Guo, Yu Liu, and Michael S Lew. A language prior based focal loss for visual question answering. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2021.

Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. Relation-aware graph attention network for visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10313–10322, 2019.

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.

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

Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. End-to-end bias mitigation by modelling biases in corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8706–8716, 2020.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12700–12710, 2021.

Ninglin Ouyang, Qingbao Huang, Pijian Li, Yi Cai, Bin Liu, Ho-fung Leung, and Qing Li. Suppressing biased samples for robust vqa. *IEEE Transactions on Multimedia*, 24: 3405–3415, 2022.

Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. Overcoming language priors in visual question answering with adversarial regularization. *Advances in Neural Information Processing Systems*, 31:1548–1558, 2018.

Ramprasaath R Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry Heck, Dhruv Batra, and Devi Parikh. Taking a hint: Leveraging explanations to make vision and language models more grounded. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2591–2600, 2019.

Robik Shrestha, Kushal Kafle, and Christopher Kanan. A negative case analysis of visual grounding methods for VQA. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8172–8181, 2020.

Damien Teney, Ehsan Abbasnedjad, and Anton van den Hengel. Learning what makes a difference from counterfactual examples and gradient supervision. In *Proceedings of Computer Vision–ECCV 2020*, pages 580–599, 2020a.

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

Damien Teney, Ehsan Abbasnejad, and Anton van den Hengel. Unshuffling data for improved generalization in visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1417–1427, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30:5998–6008, 2017.

Zhiquan Wen, Guanghui Xu, Mingkui Tan, Qingyao Wu, and Qi Wu. Debiased visual question answering from feature and sample perspectives. *Advances in Neural Information Processing Systems*, 34:3784–3796, 2021.

Jialin Wu and Raymond Mooney. Self-critical reasoning for robust visual question answering. *Advances in Neural Information Processing Systems*, 32:8601–8611, 2019.

Qi Wu, Peng Wang, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4622–4630, 2016.

Chao Yang, Su Feng, Dongsheng Li, Huawei Shen, Guoqing Wang, and Bin Jiang. Learning content and context with language bias for visual question answering. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2021.

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.

Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6281–6290, 2019.

Liyang Zhang, Shuaicheng Liu, Donghao Liu, Pengpeng Zeng, Xiangpeng Li, Jingkuan Song, and Lianli Gao. Rich visual knowledge-based augmentation network for visual question answering. *IEEE Transactions on Neural Networks and Learning Systems*, 32(10):4362–4373, 2020.

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.

Wenqiao Zhang, Siliang Tang, Yanpeng Cao, Shiliang Pu, Fei Wu, and Yueting Zhuang. Frame augmented alternating attention network for video question answering. *IEEE Transactions on Multimedia*, 22(4):1032–1041, 2019.

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