Digging out Discrimination Information from Generated Samples for Robust Visual Question Answering

Zhiquan Wen[1,2], **Yaowei Wang**[2*], **Mingkui Tan**[1,3*], **Qingyao Wu**[1], **Qi Wu**[4]
[1]School of Software Engineering, South China University of Technology, China
[2]PengCheng Laboratory, China
[3]Key Laboratory of Big Data and Intelligent Robot (South China University of Technology), Ministry of Education
[4]School of Computer Science, University of Adelaide
sewenzhiquan@mail.scut.edu.cn, wangyw@pcl.ac.cn

Abstract

Visual Question Answering (VQA) aims to answer a textual question based on a given image. Nevertheless, recent studies have shown that VQA models tend to capture the biases to answer the question, instead of using the reasoning ability, resulting in poor generalisation ability. To alleviate the issue, some existing methods consider the natural distribution of the data, and construct samples to balance the dataset, achieving remarkable performance. However, these methods may encounter some limitations: 1) rely on additional annotations, 2) the generated samples may be inaccurate, *e.g.,* assigned wrong answers, and 3) ignore the power of positive samples. In this paper, we propose a method to **D**ig out **D**iscrimination information from **G**enerated samples (DDG) to address the above limitations. Specifically, we first construct positive and negative samples in vision and language modalities, without using additional annotations. Then, we introduce a knowledge distillation mechanism to promote the learning of the original samples by the positive samples. Moreover, we impel the VQA models to focus on vision and language modalities using the negative samples. Experimental results on the VQA-CP v2 and VQA v2 datasets show the effectiveness of our DDG.

1 Introduction

With the vigorous development of computer vision and natural language processing fields, it has promoted the vision-and-language (Gu et al., 2022; Wen et al., 2023b) field to take a forward step. As a typical task of the vision-and-language field, Visual Question Answering (VQA) (Anderson et al., 2018; Cadène et al., 2019a) requires an agent to fully comprehend the information of the questions and images, and then correctly answer the textual question according to the image. Although recent advances (Cadène et al., 2019a) have achieved impressive performance on the benchmark datasets (*e.g.,* VQA v2 (Goyal et al., 2017)), numerous studies (Agrawal et al., 2018; Kafle and Kanan, 2017) have shown that some VQA models tend to excessively rely on the superficial correlations (*i.e.,* biases) between the questions and answers, instead of adopting reasoning ability to answer the questions. For example, the VQA models can easily answer "2" and "tennis" for the questions "How many . . . " and "What sports . . . ", respectively, would obtain higher accuracy, since the corresponding answers "2" and "tennis" are occupied the most in the dataset. However, memorising the biases to answer the questions would signify flawed reasoning ability, resulting in poor generalisation ability.

To mitigate the bias issues, many methods have been proposed, which can be roughly categorised into three types: 1) enhance visual attention (Selvaraju et al., 2019; Wu and Mooney, 2019), 2) directly weaken the biases (Cadène et al., 2019b; Niu et al., 2021), and 3) balance the dataset (Chen et al., 2020; Zhu et al., 2020). Previous studies have shown the methods that balance the dataset usually outperform other types of methods, since they dig out the natural distribution of the data, and then devise a suitable strategy to overcome the biases. Specifically, CSS (Chen et al., 2020) and Mutant (Gokhale et al., 2020) methods generate counterfactual samples by masking the critical objects or words in the images and questions, respectively. However, these methods require additional annotations that are hard to obtain. To get rid of the dependence on the additional annotations, MMBS (Si et al., 2022) constructs the positive questions by randomly shuffling the question words or removing the words of question types, which destroys the grammar and semantics of the original questions. Moreover, SimpleAug (Kil et al., 2021) and KDDAug (Chen et al., 2022) build the new samples by re-combining the existing questions and images, which may be difficult to assign correct answers for the generated samples. SSL-VQA (Zhu et al.,

[*]Corresponding author

Findings of the Association for Computational Linguistics: ACL 2023, pages 6910–6928
July 9-14, 2023 ©2023 Association for Computational Linguistics

2020) and D-VQA (Wen et al., 2021) construct the negative samples by randomly sampling the images or questions in a mini-batch data. Nevertheless, these methods consider the negative samples only, but ignoring the generated positive samples would improve the diversity of the dataset and further promote the robustness of the VQA models.

To overcome the above issues, we propose a method to **D**ig out **D**iscrimination information from **G**enerated samples (DDG). As pointed out by (Wen et al., 2021), the bias issues exist in both vision and language modalities, we thus construct positive and negative samples in vision and language modalities and devise corresponding training objectives to achieve unbiased learning. Concretely, we feed the samples to the UpDn (Anderson et al., 2018) model pre-trained on the VQA-CP (Agrawal et al., 2018) v2 training set, and select k objects based on the top-k image attention weights of the UpDn as the positive images. The positive questions can be constructed by using the translate-and-back-translate mechanism, *e.g.,* English → French → English. We then combine the positive images and positive questions with original questions and original images, respectively, as positive image and question samples. Based on the positive samples, we adopt a knowledge distillation mechanism (Hinton et al., 2015) to help the learning of the original samples.

Moreover, inspired by (Wen et al., 2021), we construct mismatched image-question pairs as negative samples. Generally speaking, one cannot answer the question correctly given the mismatched image-question pairs, since missing the supporting modality information. To promote the VQA models to focus on the vision and language modalities, we devise a training objective that aims to minimise the likelihood of predicting the ground-truth answers of the original samples when given the corresponding negative samples. Besides, we further introduce the corresponding positive samples to assist the training. Based on the above debiased techniques, our DDG achieves impressive performance on the VQA-CP v2 (Agrawal et al., 2018) and VQA v2 (Goyal et al., 2017) datasets, which demonstrates the effectiveness of our DDG.

Our contributions can be summarised as follows: 1) We devise a novel positive image samples generation strategy that uses the image attention weights of the pre-trained UpDn model to guide the selection of the target objects. 2) We introduce the knowledge distillation mechanism to promote the learning of the original samples by the positive samples. 3) We adopt the positive and negative samples to impel the VQA models to focus on the vision and language modalities, to mitigate the biases.

2 Related Work

2.1 Overcoming biases in VQA

Recently, researchers have proposed vast debiased techniques (Selvaraju et al., 2019; Niu et al., 2021; Zhu et al., 2020; Wen et al., 2023a) to alleviate the bias issues in VQA, which can be roughly categorised into three types: 1) enhance the visual attention, 2) directly weaken the biases, 3) balance the dataset.

Methods that enhance visual attention. These methods seek to adopt human-annotated information to strengthen the visual attention of the VQA models. Specifically, Selvaraju *et al.* (Selvaraju et al., 2019) aligned the important image regions identified based on the gradient with the human attention maps to enhance the visual attention in the VQA models. Wu *et al.* (Wu and Mooney, 2019) introduced a self-critical training objective that matches the ground-truth answer with the most important image region recognised by human explanations. However, these methods require human annotations that are hard to obtain.

Methods that weaken the biases. Ramakrishnan *et al.* (Ramakrishnan et al., 2018) adopted adversarial learning to inhibit the VQA models capture the language biases. Inspired by (Ramakrishnan et al., 2018), Cadene *et al.* (Cadène et al., 2019b) devised a question-only model to generate weight to re-weight the samples. Moreover, Han *et al.* (Han et al., 2021) forced the biased models to capture different types of biases, and removed them step by step. Different from the above, Niu *et al.* (Niu et al., 2021; Niu and Zhang, 2021) introduced the idea of cause-effect to help alleviate the biases. Nevertheless, these methods introduce additional parameters in training or inference phrases.

Methods that balance the dataset. CSS (Chen et al., 2020) and Mutant (Gokhale et al., 2020) methods generated massive counterfactual samples by masking the critical objects and words in the images and questions, respectively. However, these methods require additional annotations to assign the answers for the generated samples. To get rid of the dependence on the annotations, KDDAug (Chen et al., 2022) and SimpleAug (Kil et al., 2021) constructed the samples by re-composing the existing questions and images, which however, is hard to assign the correct answers for the generated samples. SSL-VQA (Zhu et al., 2020) and D-VQA (Wen et al., 2021) constructed the negative samples by randomly sampling the images or questions in a mini-batch data. Nevertheless, these methods ignored the positive samples could improve the diversity of the dataset, which was helpful for improving the robustness of the VQA models. Moreover, Si *et al.* (Si et al., 2022) constructed the positive question samples by randomly shuffling the words or removing the words of question types, which destroys the semantics of the original questions.

Different from the above methods, we seek to construct positive samples and negative samples in both vision and language modalities, and devise corresponding debiased strategies to achieve unbiased learning.

2.2 Knowledge Distillation

Knowledge Distillation (KD) (Hinton et al., 2015) is a universal model compression method that seeks to train a small student model guided by a large teacher model. Due to the effectiveness of KD, the idea has been applied to other tasks, *e.g.,* long-tail classification (He et al., 2021; Xiang et al., 2020), object detection (Chen et al., 2017; Wang et al., 2019), and video captioning (Pan et al., 2020; Zhang et al., 2020). Recently, some debiased VQA methods (Niu and Zhang, 2021; Chen et al., 2022) introduced KD to alleviate the bias issues. Specifically, Niu *et al.* (Niu and Zhang, 2021) devised two teachers (*i.e.,* ID-teacher and OOD-teacher) to generate "soft" labels to guide the training of the student model (*i.e.,* the baseline model) with the KD mechanism. Inspired by IntroD, Chen *et al.* (Chen et al., 2022) adopted a multi-teacher KD mechanism to help generate robust pseudo labels for all newly composed image-question pairs. In our DDG, we seek to improve the reasoning ability of the VQA models with the help of the generated positive samples, via the KD mechanism.

3 Digging out Discrimination Information from Generated Samples

As shown by (Agrawal et al., 2018; Kafle and Kanan, 2017), VQA models tend to capture the biases in a dataset to answer questions, instead of adopting the reasoning ability, resulting in poor generalisation ability. Moreover, bias issues exist in both vision and language modalities (Wen et al., 2021). To address the above issues, we seek to construct both positive and negative samples in vision and language modalities, and devise corresponding debiased strategies to achieve unbiased learning. The overall framework is shown in Figure. 1.

3.1 Preliminary

Visual question answering (VQA) requires an agent to answer a textual question given a corresponding image. Traditional VQA methods (Anderson et al., 2018; Kim et al., 2018; Ben-younes et al., 2019; Cadène et al., 2019a) regard the VQA task as a multi-class classification problem, where each class corresponds to a unique answer. To be specific, given a VQA dataset $\mathcal{D} = \{(v_i, q_i, a_i)\}_{i=1}^{N}$ with N samples, where $v_i \in \mathcal{V}$ (image set), $q_i \in \mathcal{Q}$ (question set) are the i-th sample in $\mathcal{D}$, and $a_i \in \mathcal{A}$ (answer set) is a corresponding ground-truth answer, VQA methods seek to learn a multimodal mapping: $\mathcal{V} \times \mathcal{Q} \to [0, 1]^{|\mathcal{A}|}$ to generate an answer distribution over the answer set $\mathcal{A}$. Generally speaking, most VQA models usually contain four parts, namely, vision feature encoder $e_v(\cdot)$, language feature encoder $e_q(\cdot)$, multimodal feature fusion module $f(\cdot, \cdot)$, and classifier $c(\cdot)$. These modules can be formed as a traditional VQA model:

$$P(\mathcal{A}|v_i, q_i) = c(f(e_v(v_i), e_q(q_i))). \tag{1}$$

Formally, since regarding the VQA task as a multi-class classification problem, the VQA models can be optimised by a binary cross-entropy loss $\mathcal{L}_{vqa}$, which can be formulated as:

$$\begin{aligned} \mathcal{L}_{vqa} = -\frac{1}{N}\sum_{i=1}^{N} \mathbf{a}_i \log(\sigma(P(\mathcal{A}|v_i, q_i))) + \\ (1 - \mathbf{a}_i)\log(1 - \sigma(P(\mathcal{A}|v_i, q_i))), \end{aligned} \tag{2}$$

where σ denotes the sigmoid activation function, and $\mathbf{a}_i$ is the target score obtained based on the answer a_i that humans annotated for (v_i, q_i).

3.2 Sample Generation

Our method aims to adopt the generated samples to achieve unbiased learning. Hence we present how to generate the positive and negative samples at first. As pointed out by (Wen et al., 2021), biases exist in both language and vision modalities. To overcome the bias issue, we seek to generate positive and negative samples regarding the vision and language modalities for each original sample, to assist the training process.

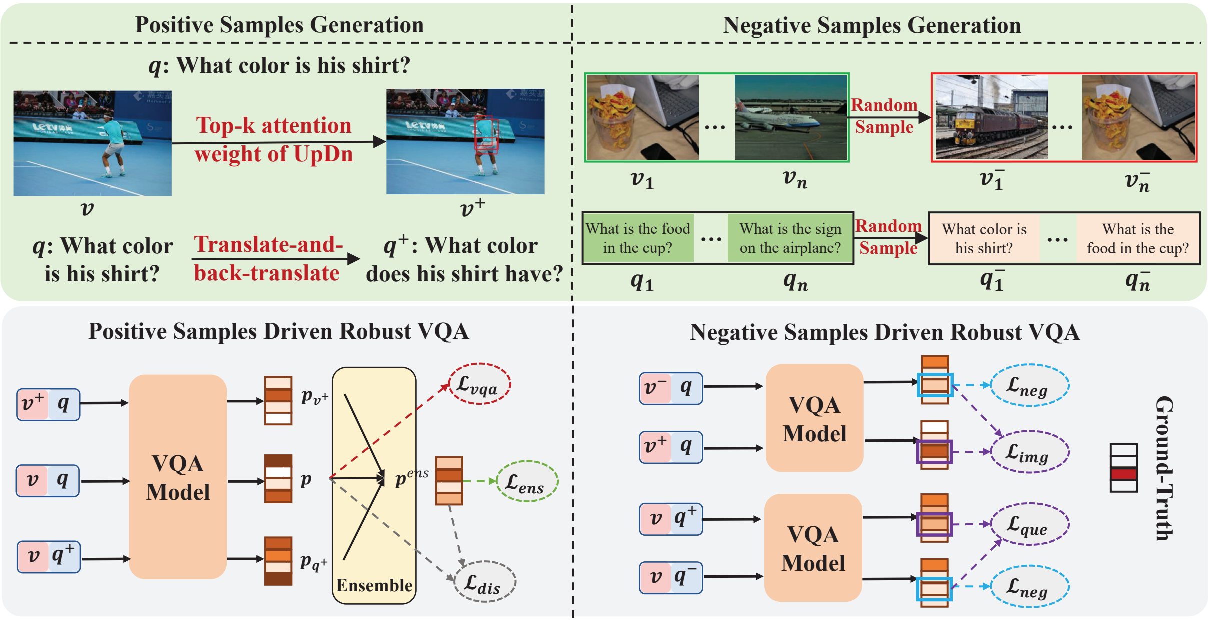


Figure 1: Overview of our DDG method. we draw support from the pre-trained UpDn (Anderson et al., 2018) model and the translate-and-back-translate mechanism to generate the positive image and question samples, respectively, and then introduce a knowledge distillation mechanism to facilitate the learning of the original samples. Moreover, we construct the mismatched image-question pairs by randomly sampling the images or questions in a mini-batch data as negative samples. By using these negative samples, we enhance the attention of VQA models towards both vision and language modalities when answering the questions, thereby mitigating biases.

Positive samples generation. To mitigate the bias issue over vision and language modalities in VQA, we build two types of positive samples, *i.e.,* a positive question sample, and a positive image sample.

Specifically, to generate the **positive image samples**, we seek to draw support from the image attention weights in a pre-trained baseline VQA model. We have empirically found that although the baseline models (*e.g.,* UpDn (Anderson et al., 2018)) achieve unsatisfactory performance in the out-of-distributions (OOD) test set (*e.g.,* VQA-CP v2 (Agrawal et al., 2018) dataset), they still obtain promising performance in the independent and identically distributed (IID) dataset (*e.g.,* VQA v2 dataset (Goyal et al., 2017)). In other words, the baseline models can identify the target objects in the images referred to in the questions to accomplish answering during the training process, regardless of whether capturing the biases. Hence, the image attention weights of the pre-trained UpDn (Anderson et al., 2018) model can help find target objects as positive image samples, which can exclude the background information of the images.

Given the sample (v_i, q_i) from the VQA- CP v2 training set, we first feed it to the UpDn model pre-trained on the VQA-CP v2 training set and would obtain the image attention weights of the UpDn model regarding the objects in image v_i. Note that we select k objects based on the top-k image attention weights as the positive image samples (v_i^+, q_i), where k is a hyper-parameters.

To generate the **positive question samples**, previous methods (Si et al., 2022) seek to adopt some data augmentation methods to expand the data, *e.g.,* randomly shuffle the question words or remove question category words. However, these methods would severely destroy the grammar and semantics of the original question, resulting in changing the semantic information of the questions. To mitigate this issue, inspired by (Tang et al., 2020), we adopt the translate-and-back-translate mechanism to generate the positive question samples. Specifically, we first use pre-trained English-to-French and English-to-German translation models to translate the original question to French and German, respectively. Then we use corresponding pre-trained back-translation models to translate them back into English. [1] Moreover, we further adopt a pre-trained sentence similarity model to choose a back-translated question sample that has the highest similarity score with the original question as the positive question sample. Note that for some sim-

[1] All pre-trained translation models are obtained from the Hugging Face repository.

ple questions, they would still keep the same even feeding them to the translate-and-back-translate process. To generate positive question samples for these questions, we substitute the words in the question with synonyms based on the pre-trained synonym word substitution model. [2] In this way, we obtain the positive question samples (v_i, q_i^+).

Based on the above, we would obtain two types of positive samples (*i.e.,* (v_i^+, q_i) and (v_i, q_i^+)) for each sample (v_i, q_i), in which the positive image samples have foreground information in the image and the positive question samples are semantic equivalent to the original question.

Negative samples generation. Inspired by (Wen et al., 2021), we construct the negative samples over language and vision modalities by randomly sampling one question and one image in a mini-batch data for each sample. Specifically, given a mini-batch data $\{(v_b, q_b)\}_{b=1}^B$, for each sample (v_i, q_i), we randomly sample one image v_i^- and one question q_i^- from $\{(v_b, q_b)\}_{b=1}^B$ to form the negative samples, namely, negative question sample (v_i, q_i^-) and negative image sample (v_i^-, q_i).

3.3 Generated Samples Driven Robust VQA

Positive samples driven robust VQA. We attempt to achieve robust VQA with the help of the generated positive samples. Specifically, given an original sample (v_i, q_i) and its counterpart positive samples (v_i^+, q_i) and (v_i, q_i^+), we first feed them into the VQA models to obtain the predictions $P(\mathcal{A}|v_i, q_i)$, $P(\mathcal{A}|v_i^+, q_i)$, and $P(\mathcal{A}|v_i, q_i^+)$. Generally speaking, the ensemble predictions usually perform better than the predictions before the ensemble. We thus adopt a simple ensemble strategy (*i.e.,* averaging these predictions) to obtain ensemble predictions $P_{ens} = (P(\mathcal{A}|v_i, q_i) + P(\mathcal{A}|v_i^+, q_i) + P(\mathcal{A}|v_i, q_i^+))/3$. One intuitive way to make the VQA models achieve better performance with the help of the positive samples is to adopt a knowledge distillation mechanism (Hinton et al., 2015). Concretely, we regard the ensemble prediction P_{ens} and the original prediction $P(\mathcal{A}|v_i, q_i)$ as a teacher and a student, respectively, and then introduce a Kullback-Leibler (KL) Divergence $\mathcal{L}_{dis}$ as the objective to optimise the VQA models, which can be formulated as:

$$\mathcal{L}_{dis} = \sum_{i=1}^{N} P_{ens} \log \frac{P_{ens}}{P(\mathcal{A}|v_i, q_i)}. \tag{3}$$

By minimising the KL divergence, the VQA models can extract discrimination information from the positive samples to help better answer the original questions q_i correctly based on the images v_i.

To guarantee the teacher (*i.e.,* the ensemble prediction P_{ens}) performs better than the student (*i.e.,* the original prediction $P(\mathcal{A}|v_i, q_i)$), we still use the binary cross-entropy loss $\mathcal{L}_{ens}$ on P_{ens} to further optimise the VQA models.

Negative samples driven robust VQA. Besides adopting the positive samples to assist the training process, we also introduce the debiased strategy on negative samples to alleviate the bias issues. As shown by (Agrawal et al., 2018), the bias issue usually denotes the VQA models tend to capture the superficial correlations between one modality and the answers to make a prediction on the questions. To mitigate this issue, one direct solution is to improve the attention on both language and vision modalities information when the VQA models answer the questions. We thus consider adopting the negative samples to achieve this aim.

Intuitively, given a mismatched image-question pair, the VQA models even the human being cannot make a correct prediction. Drawing from this insight, when given original samples and the counterpart negative samples, we can alleviate the biases by giving contrary training objectives to the negative samples. This encourages the VQA models to answer the questions by paying more attention to the information of each modality. Concretely, inspired by (Wen et al., 2021), given an original sample (v_i, q_i, a_i) and its counterpart negative samples (v_i^-, q_i) and (v_i, q_i^-), the VQA models cannot answer correctly when feeding the negative samples, which can be achieved by minimising the possibility of predicting the ground-truth answer:

$$\mathcal{L}_{neg} = \delta(P(\mathcal{A}|v_i^-, q_i))[x] + \delta(P(\mathcal{A}|v_i, q_i^-))[x], \tag{4}$$

where x is the index of ground-truth answer a_i in the answer set $\mathcal{A}$, and δ is the softmax activation function. Minimising the training objective $\mathcal{L}_{neg}$ encourages the VQA models not to give the ground-truth answer when feeding the mismatched image-question pairs. Thus, the VQA models are able to consider both image and question information before making a prediction, which implicitly alleviates the bias issue.

Moreover, to further enhance the attention of VQA models towards both vision and language modalities, we introduce positive samples into the

[2] The pre-trained model is from the GitHub repository.

training process. Specifically, given positive image samples (v_i^+, q_i, a_i) and negative image samples (v_i^-, q_i), when feeding them to the VQA models, one hopes the VQA models can answer correctly with high confidence on the positive samples, while having low prediction confidence to the ground-truth answer with the negative samples. This can be formulated as:

$$\max\ \delta(P(\mathcal{A}|v_i^+, q_i))[x] - \delta(P(\mathcal{A}|v_i^-, q_i))[x].$$

Maximising the objective encourages the VQA models to make accurate predictions for the matched image-question pairs, while discouraging the models from generating the ground-truth answer when provided with negative image samples. This impels the VQA models to allocate more attention to the vision modality.

By leveraging the monotonicity property of the logarithmic function log, we convert the maximisation problem into an equivalent minimisation problem. This transformation can be mathematically formulated as follows:

$$\begin{aligned}\mathcal{L}_{img} = -\log(\sigma(\delta(P(\mathcal{A}|v_i^+, q_i))[x] -& \\ \delta(P(\mathcal{A}|v_i^-, q_i))[x]))&\end{aligned} \tag{5}$$

Moreover, with regard to the negative samples, the prediction scores associated with the ground-truth answer of the corresponding positive samples can serve as an indicator of the extent to which VQA models capture biases. The higher the prediction score $\delta(P(\mathcal{A}|v_i^-, q_i))[x]$), the greater the degree of bias it represents. Therefore, it should be subject to a higher penalty in the loss $\mathcal{L}_{img}$. Inspired by the focal loss (Lin et al., 2017), we consider the prediction score $\delta(P(\mathcal{A}|v_i^-, q_i))[x]$) as a measure of the degree of bias, and thus the loss $\mathcal{L}_{img}^{weight}$ can be reformulated as:

$$\mathcal{L}_{img}^{weight} = -\delta(P(\mathcal{A}|v_i^-, q_i))[x] * \mathcal{L}_{img} \tag{6}$$

We would obtain $\mathcal{L}_{que}^{weight}$ in the same way. Thus the weighted loss is formulated as follows:

$$\mathcal{L}^{weight} = \mathcal{L}_{img}^{weight} + \mathcal{L}_{que}^{weight}. \tag{7}$$

3.4 Overall Training Objective

In total, our overall training objective can be formulated as:

$$\mathcal{L} = \mathcal{L}_{vqa} + \mathcal{L}_{ens} + \mathcal{L}_{dis} + \mathcal{L}_{neg} + \lambda * \mathcal{L}^{weight}, \tag{8}$$

where λ is a hyper-parameter.

4 Experiments

4.1 Datasets

We evaluate our DDG on the OOD dataset VQA-CP v2 (Agrawal et al., 2018) and IID dataset VQA v2 (Goyal et al., 2017) validation set based on the standard evaluation metric (Antol et al., 2015). Due to the page limitation, we put the implementation details and compared methods into the Appendix.

4.2 Quantitative results

We report the experimental results on the VQA-CP v2 and VQA v2 datasets in Table 1. From these results, we have the following observations: 1) On the whole, the methods that balance the datasets outperform the other two types of methods *i.e.,* enhance visual attention and directly weaken the biases. This demonstrates that alleviating the biases by paying more attention to the natural distribution of the data would obtain higher performance. 2) Our DDG outperforms most compared methods. Specifically, our DDG surpasses SCR (Wu and Mooney, 2019), GGE-DQ (Han et al., 2021), SSL-VQA (Zhu et al., 2020), and KDDAug (Chen et al., 2022) by approximately 12%, 3%, 3%, and 1%, respectively. These results demonstrate the effectiveness of our DDG. 3) Although our method performs slightly worse than the Mutant (Gokhale et al., 2020) and D-VQA (Wen et al., 2021), our DDG achieves higher performance on the VQA v2 dataset. Moreover, Mutant constructed the counterfactual samples highly relying on the additional annotations, while our method build the samples without introducing additional annotations. Meanwhile, compared to the D-VQA method, our method performs better when the data is limited, which can be shown in Table 2. These results further demonstrate the effectiveness of our DDG.

Benefiting from the training process based on the positive samples, our DDG performs better than all the compared methods on the VQA v2 dataset. Specifically, our DDG outperforms SSL-VQA (Zhu et al., 2020) and D-VQA (Wen et al., 2021) by around 1.8% and 0.6%, respectively, which demonstrates our DDG is able to improve the model performance on both IID (*i.e.,* VQA v2 dataset) and OOD (*i.e.,* VQA-CP v2) datasets, further implying the superiority of our DDG.

4.3 Qualitative results.

To further demonstrate the effectiveness of our DDG on alleviating the biases, we provide the

Case	Model	VQA-CP v2 test (%)				VQA v2 val (%)			
		All	Yes/No	Num	Other	All	Yes/No	Num	Other
–	SAN (Yang et al., 2016)	24.96	38.35	11.14	21.74	52.41	70.06	39.28	47.84
	GVQA (Agrawal et al., 2018)	31.30	57.99	13.68	22.14	48.24	72.03	31.17	34.65
	UpDn (Anderson et al., 2018)	39.74	42.27	11.93	46.05	63.48	81.18	42.14	55.66
I	AttAlign (Selvaraju et al., 2019)	39.37	43.02	11.89	45.00	63.24	80.99	42.55	55.22
	HINT (Selvaraju et al., 2019)	46.73	67.27	10.61	45.88	63.38	81.18	42.99	55.56
	SCR (Wu and Mooney, 2019)	48.47	70.41	10.42	47.29	62.30	77.40	40.90	56.50
II	AdvReg (Ramakrishnan et al., 2018)	41.17	65.49	15.48	35.48	62.75	79.84	42.35	55.16
	RUBi (Cadène et al., 2019b)	44.23	67.05	17.48	39.61	-	-	-	-
	Re-Scaling (Guo et al., 2022)	47.09	68.42	21.71	42.88	55.50	64.22	39.61	53.09
	DLR (Jing et al., 2020)	48.87	70.99	18.72	45.57	57.96	76.82	39.33	48.54
	VGQE (KV and Mittal, 2020)	48.75	-	-	-	64.04	-	-	-
	LMH (Clark et al., 2019)	52.01	72.58	31.12	46.97	56.35	65.06	37.63	54.69
	IntroD (Niu and Zhang, 2021)	51.31	71.39	27.13	47.41	62.05	77.65	40.25	55.97
	CF-VQA (Niu et al., 2021)	53.55	91.15	13.03	44.97	63.54	82.51	43.96	54.30
	RMFE (Gat et al., 2020)	54.55	74.03	49.16	45.82	-	-	-	-
	CKCL (Pan et al., 2022)	55.05	90.33	18.99	46.46	62.55	79.17	41.94	55.38
	LPF (Liang et al., 2021)	55.34	88.61	23.78	46.57	55.01	64.87	37.45	52.08
	GGE-DQ (Han et al., 2021)	57.32	87.04	27.75	49.59	59.11	73.27	39.99	54.39
	D-VQA (Wen et al., 2021)	61.91	88.93	52.32	50.39	64.96	82.18	44.05	57.54
III	CSS (Chen et al., 2020)	58.95	84.37	49.42	48.21	59.91	73.25	39.77	55.11
	CSS+CL (Liang et al., 2020)	59.18	86.99	49.89	47.16	57.29	67.27	38.40	54.71
	CSS^{+} (Chen et al., 2021)	59.54	83.37	52.57	48.97	59.96	73.69	40.18	54.77
	ECD (Kolling et al., 2022)	59.92	83.23	52.29	49.71	57.38	69.06	35.74	54.25
	Mutant (Gokhale et al., 2020)	61.72	88.90	49.68	50.78	62.56	82.07	42.52	53.28
IV	CVL (Abbasnejad et al., 2020)	42.12	45.72	12.45	48.34	-	-	-	-
	Unshuffling (Teney et al., 2021)	42.39	47.72	14.43	47.24	61.08	78.32	42.16	52.71
	MMBS (Si et al., 2022)	48.19	65.00	14.05	48.75	63.84	79.61	44.23	57.05
	SimpleAug (Kil et al., 2021)	52.65	66.40	43.43	47.98	64.34	81.97	43.91	56.35
	RandImg (Teney et al., 2020)	55.37	83.89	41.60	44.20	57.24	76.53	33.87	48.57
	SSL-VQA (Zhu et al., 2020)	57.59	86.53	29.87	**50.03**	63.73	-	-	-
	KDDAug (Chen et al., 2022)	60.24	86.13	**55.08**	48.08	62.86	80.55	41.05	55.18
	DDG (Ours)	**61.14**	**88.77**	49.33	49.90	**65.54**	**82.92**	**44.80**	**57.80**

Table 1: Comparison with the state-of-the-art methods on the VQA-CP v2 test set and VQA v2 validation set.The best scores are **bold**, and the second best scores of ours are underlined. The backbone model is UpDn (Anderson et al., 2018). **I** – **IV** denote methods that enhance visual attention, directly weaken the biases, balance the dataset using additional annotations, and balance the dataset without introducing additional annotations, respectively.

qualitative results on the VQA-CP v2 dataset in Figures. 2 and 3. From the results in Figure 2, UpDn (Anderson et al., 2018) and SSL-VQA (Zhu et al., 2020) fail to find the target objects mentioned in the question within the image, leading to erroneous predictions. In contrast, our DDG demonstrates a remarkable ability to accurately localize the target objects with a high degree of confidence, resulting in precise answers to the posed questions. These visualisation results demonstrate the effectiveness of our DDG. Moreover, in Figure. 3, we provide visualisations of the answer distributions obtained by various approaches for different question types, namely "How many . . . ", "Is this . . . ", and "How many people are in . . . ". From the results, we have the following observations: 1) the training answer distribution is different from that in the test set, which is very challenging. 2) The UpDn model excessively fits the biases in the training set, and thus outputs a similar answer distribution with the training set given the test set, resulting in poor performance. 3) SSL-VQA seeks to alleviate the bias issue, which however is limited. Our DDG is able to alleviate the biases effectively, and thus achieves similar answer distributions with the test set, embodying the better generalisation ability.

4.4 Ablation studies

Effect of the scale of the training set. To demonstrate the effectiveness of our method in the data-limited scenario, we conduct experiments on different scales of the training data. Specifically, on the VQA-CP v2 dataset, we manually split the training set into different proportions (*i.e.,* from 20% to 80% of the original training data), while the test set is unchanged. From the experimental results in Table 2,

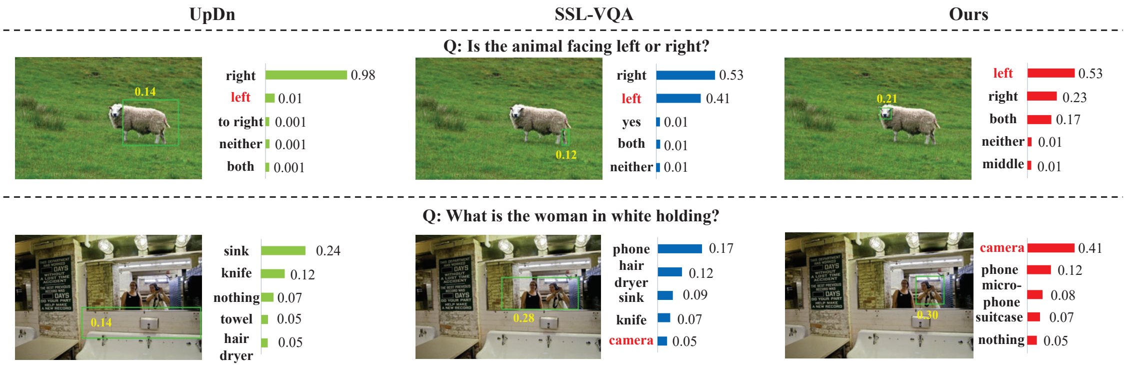


Figure 2: Qualitative comparison among UpDn (Anderson et al., 2018), SSL-VQA (Zhu et al., 2020), and our DDG on the VQA-CP v2 test set. For each example, we put the bounding box with the highest attention weight in the image and show the answers with the top-5 predictions. The bold, red answer is the ground-truth answer.

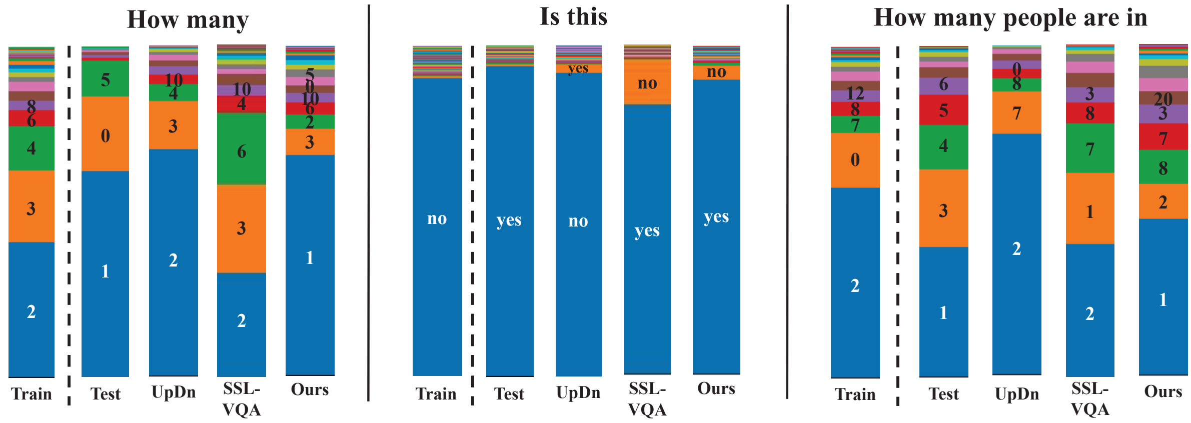


Figure 3: Qualitative comparison among UpDn (Anderson et al., 2018), SSL-VQA (Zhu et al., 2020) and our DDG on the VQA-CP v2 test set about the answer distributions.

Model	Proportion of Training Set				
	20%	40%	60%	80%	100%
UpDn†	36.22	38.90	39.40	40.61	41.53
SSL-VQA	52.71	54.42	56.83	57.31	57.59
D-VQA	52.94	56.74	58.31	59.05	**61.91**
Ours	**55.74**	**57.42**	**58.99**	**59.69**	61.14

Table 2: Effect of different scales of the training data of VQA-CP v2 on the model performance. We report the results in terms of Accuracy (%).

we find that our method performs better when the training data is limited, which is more practical and suitable for the real world. Specifically, our method performs better than SSL-VQA (Zhu et al., 2020) and D-VQA (Wen et al., 2021) on any proportions of the training data, especially when only remains 20% of the training data, our DDG outperforms SSL-VQA and D-VQA by around 3%. These results demonstrate the superiority of our DDG in the data-limited scenarios.

Effect of each component of our DDG. We conduct ablation studies on the VQA-CP v2 dataset to evaluate each component in our DDG, and show the experimental results in Table 3. From these results, we have the following observations: 1) when introducing the ensemble binary cross-entropy loss $\mathcal{L}_{ens}$ with the positive samples, the model performance improves by around 5% compared with the UpDn (Anderson et al., 2018) model (*i.e.,* 41.53% *vs.* 46.63%), which demonstrates the positive samples are able to assist the training process to alleviate the bias issue. 2) By incorporating the KL loss $\mathcal{L}_{dis}$, the performance would be further improved (*i.e.,* 46.63% *vs.* 47.77%), which highlights the ensemble prediction is able to guide the training of the original prediction. 3) Upon introducing $\mathcal{L}_{neg}$ and $\mathcal{L}^{weight}$, which leverage negative samples, the performance would improve substantially (*i.e.,* 47.77% *vs.* 61.14%). This significant enhancement underscores the significance of promoting the attention of VQA models towards both vision and language modalities when answering the questions.

$\mathcal{L}_{ens}$	$\mathcal{L}_{dis}$	$\mathcal{L}_{neg}$	$\mathcal{L}^{weight}$	VQA-CP v2 (%)
				41.53
√				46.63
√	√			47.77
√	√	√		60.85
√	√	√	√	**61.14**

Table 3: Effect of each component of our DDG on the model performance. We show the results on the VQA-CP v2 dataset in terms of Accuracy (%). We use UpDn (Anderson et al., 2018) as the backbone model.

Model	k	VQA-CP v2 test set (%)			
		All	Yes/No	Number	Other
DDG	3	59.73	86.98	42.36	50.22
	6	60.24	88.28	41.77	**50.62**
	8	60.74	88.41	45.54	50.41
	10	**61.14**	88.77	**49.33**	49.90
	12	60.65	88.79	45.92	49.95
	14	60.38	**88.88**	41.55	50.61

Table 4: Effect of different k. We report the experimental results in terms of Accuracy (%).

Model	λ	VQA-CP v2 test set (%)			
		All	Yes/No	Number	Other
DDG	0.01	60.86	88.89	47.21	49.93
	0.03	60.93	88.93	47.59	49.92
	0.05	**61.14**	88.77	**49.33**	49.90
	0.07	60.74	**88.97**	44.96	50.27
	0.1	60.58	88.89	43.11	**50.53**

Table 5: Effect of different λ. We report the experimental results in terms of Accuracy (%).

Model	Yes/No	Number	Other	Overall	Gap$\Delta \uparrow$
SAN†	38.44	12.91	46.65	39.11	**+16.41**
+ DDG	**85.59**	**24.62**	**48.24**	**55.52**	
UpDn†	43.45	13.64	48.18	41.53	**+19.61**
+ DDG	**88.77**	**49.33**	**49.90**	**61.14**	

Table 6: Effect of different backbones (*i.e.,* SAN (Yang et al., 2016) and UpDn (Anderson et al., 2018)) on the model performance. We report the experimental results on the VQA-CP v2 dataset in terms of Accuracy (%). † denotes the re-implementation of the baseline.

These results further demonstrate the effectiveness of each component in our DDG.

Effect of k. k denotes the number of target objects that are selected as the positive samples, which can be referred to in Section 3.2. To demonstrate the effect of the k on the model performance, we conduct experiments on the VQA-CP v2 dataset regarding different k. From the results in Table 4, we have the following observations: 1) with the increase of k (*e.g.,* from 3 to 10), the model performance exhibits a gradual improvement.The results indicate that a higher value of k increases the likelihood that the positive image samples indeed encompass the target objects mentioned in the corresponding questions. 2) Once k exceeds 10, the model performance starts to drop, thereby illustrating that an excessively high value of k introduces extraneous background information that adversely affects the model's performance. These results demonstrate that an appropriate k helps to obtain the best performance on our DDG.

Evaluation of λ. λ is the weight of the loss $\mathcal{L}^{weight}$. We conduct ablation studies about different λ on the model performance, and show the experimental results in Table 5. From the results, the best performance is obtained in our DDG when $\lambda = 0.05$, and the performance is drop whenever λ is higher or lower than 0.05. These results demonstrate that a suitable weight of loss $\mathcal{L}^{weight}$ helps to obtain better performance.

Evaluation of different backbones. Our DDG is model-agnostic. To demonstrate the effectiveness of our DDG on different backbones (*i.e.,* SAN (Yang et al., 2016) and UpDn (Anderson et al., 2018)), we conduct experiments on the VQA-CP v2 dataset, and show the results in Table 6. From the results, our DDG consistently achieves a substantial improvement in model performance, regardless of which backbone it is. These results further embody the superiority of our DDG.

5 Conclusion

In this paper, we have proposed a novel method named DDG to alleviate the bias issues in VQA from vision and language modalities. Specifically, we construct both positive and negative samples in vision and language modalities without using additional annotations, in which the positive questions have similar semantics to the original questions, while the positive images contain foreground information. Based on the positive samples, we heuristically introduce the knowledge distillation mechanism to facilitate the training of the original samples through guidance from positive samples. Moreover, we put forth a strategy that encourages VQA models to focus more on the vision and language modalities when answering the questions, aided by the negative samples. Extensive experiments on the VQA-CP v2 and VQA v2 datasets show the effectiveness of our DDG.

Acknowledgements

We would like to thank all the anonymous reviewers for their constructive comments and suggestions. This work was partially supported by STI 2030—Major Projects (2022ZD0208900), Key Realm R&D Program of Guangzhou 202007030007, Program for Guangdong Introducing Innovative and Enterpreneurial Teams 2017ZT07X183, National Natural Science Foundation of China (NSFC) 62072190.

Limitations

The paper focuses on the VQA task only, we will extend our method to other multimodal tasks in future works, *e.g.,* referring expression comprehension (REC). Moreover, although our DDG outperforms most of the state-of-the-art methods, the performance is still a long way from humans.

Ethics Statement

The authors declare that they have no conflict of interest. This paper introduces a novel method named DDG to overcome the bias issue in Visual Question Answering (VQA). Mitigating the biases can impel the VQA models to adopt real reasoning ability to answer the questions, instead of using the captured biases. Hence, this research can promote the development of the AI robot, *e.g.,* dialogue robots, and facilitate people's daily lives. The failure of the debiased technique may result in the collapse of the VQA system in environments that have seen less or even never seen. Moreover, we evaluate our DDG on the benchmark out-of-distribution (OOD) dataset and demonstrate the remarkable debiased ability.

References

Ehsan Abbasnejad, Damien Teney, Amin Parvaneh, Javen Shi, and Anton van den Hengel. 2020. Counterfactual vision and language learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10041–10051.

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

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

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433.

Hedi Ben-younes, Rémi Cadène, Nicolas Thome, and Matthieu Cord. 2019. BLOCK: bilinear superdiagonal fusion for visual question answering and visual relationship detection. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 8102–8109.

Rémi Cadène, Hedi Ben-younes, Matthieu Cord, and Nicolas Thome. 2019a. MUREL: multimodal relational reasoning for visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1989–1998.

Rémi Cadène, Corentin Dancette, Hedi Ben-younes, Matthieu Cord, and Devi Parikh. 2019b. Rubi: Reducing unimodal biases for visual question answering. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 839–850.

Guobin Chen, Wongun Choi, Xiang Yu, Tony X. Han, and Manmohan Chandraker. 2017. Learning efficient object detection models with knowledge distillation. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 742–751.

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

Long Chen, Yuhang Zheng, Yulei Niu, Hanwang Zhang, and Jun Xiao. 2021. Counterfactual samples synthesizing and training for robust visual question answering. *Arxiv*.

Long Chen, Yuhang Zheng, and Jun Xiao. 2022. Rethinking data augmentation for robust visual question answering. In *European Conference on Computer Vision (ECCV)*, volume 13696, pages 95–112.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.

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

Itai Gat, Idan Schwartz, Alexander G. Schwing, and Tamir Hazan. 2020. Removing bias in multi-modal classifiers: Regularization by maximizing functional entropies. In *Conference on Neural Information Processing Systems (NeurIPS)*, volume 33, pages 3197–3208.

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

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6325–6334.

Jing Gu, Eliana Stefani, Qi Wu, Jesse Thomason, and Xin Wang. 2022. Vision-and-language navigation: A survey of tasks, methods, and future directions. In *The Association for Computational Linguistics (ACL)*, pages 7606–7623.

Yangyang Guo, Liqiang Nie, Zhiyong Cheng, Qi Tian, and Min Zhang. 2022. Loss re-scaling VQA: revisiting the language prior problem from a class-imbalance view. *IEEE Transactions on Image Processing (TIP)*, 31:227–238.

Xinzhe Han, Shuhui Wang, Chi Su, Qingming Huang, and Qi Tian. 2021. Greedy gradient ensemble for robust visual question answering. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1564–1573.

Yin-Yin He, Jianxin Wu, and Xiu-Shen Wei. 2021. Distilling virtual examples for long-tailed recognition. In *IEEE International Conference on Computer Vision (ICCV)*, pages 235–244.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *ArXiv*.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*, pages 448–456.

Chenchen Jing, Yuwei Wu, Xiaoxun Zhang, Yunde Jia, and Qi Wu. 2020. Overcoming language priors in VQA via decomposed linguistic representations. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 11181–11188.

Kushal Kafle and Christopher Kanan. 2017. An analysis of visual question answering algorithms. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1983–1991.

Jihyung Kil, Cheng Zhang, Dong Xuan, and Wei-Lun Chao. 2021. Discovering the unknown knowns: Turning implicit knowledge in the dataset into explicit training examples for visual question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6346–6361.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 1571–1581.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Camila Kolling, Martin D. More, Nathan Gavenski, Eduardo H. P. Pooch, Otávio Parraga, and Rodrigo C. Barros. 2022. Efficient counterfactual debiasing for visual question answering. In *Winter Conference on Applications of Computer Vision (WACV)*, pages 2572–2581.

Gouthaman KV and Anurag Mittal. 2020. Reducing language biases in visual question answering with visually-grounded question encoder. In *European Conference on Computer Vision (ECCV)*, pages 18–34.

Zujie Liang, Haifeng Hu, and Jiaying Zhu. 2021. LPF: A language-prior feedback objective function for debiased visual question answering. In *ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 1955–1959.

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

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988.

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

Yulei Niu and Hanwang Zhang. 2021. Introspective distillation for robust question answering. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 16292–16304.

Boxiao Pan, Haoye Cai, De-An Huang, Kuan-Hui Lee, Adrien Gaidon, Ehsan Adeli, and Juan Carlos Niebles. 2020. Spatio-temporal graph for video captioning with knowledge distillation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10867–10876.

Yonghua Pan, Zechao Li, Liyan Zhang, and Jinhui Tang. 2022. Causal inference with knowledge distilling and curriculum learning for unbiased VQA. *ACM Trans. Multim. Comput. Commun. Appl. (ACM TOMMCCAP)*, 18(3):67:1–67:23.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 8024–8035.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. 2018. Overcoming language priors in visual question answering with adversarial regularization. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 1548–1558.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 91–99.

Ramprasaath Ramasamy Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry P. Heck, Dhruv Batra, and Devi Parikh. 2019. Taking a HINT: leveraging explanations to make vision and language models more grounded. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2591–2600.

Qingyi Si, Yuanxin Liu, Fandong Meng, Zheng Lin, Peng Fu, Yanan Cao, Weiping Wang, and Jie Zhou. 2022. Towards robust visual question answering: Making the most of biased samples via contrastive learning. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 6650–6662.

Ruixue Tang, Chao Ma, Wei Emma Zhang, Qi Wu, and Xiaokang Yang. 2020. Semantic equivalent adversarial data augmentation for visual question answering. In *European Conference on Computer Vision (ECCV)*, volume 12364, pages 437–453.

Damien Teney, Ehsan Abbasnejad, Kushal Kafle, Robik Shrestha, Christopher Kanan, and Anton van den Hengel. 2020. On the value of out-of-distribution testing: An example of goodhart's law. In *Conference on Neural Information Processing Systems (NeurIPS)*, volume 33, pages 407–417.

Damien Teney, Ehsan Abbasnejad, and Anton van den Hengel. 2021. Unshuffling data for improved generalization in visual question answering. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1417–1427.

Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. 2019. Distilling object detectors with fine-grained feature imitation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4933–4942.

Zhiquan Wen, Shuaicheng Niu, Ge Li, Qingyao Wu, Mingkui Tan, and Qi Wu. 2023a. Test-time model adaptation for visual question answering with debiased self-supervisions. *IEEE Transactions on Multimedia (TMM)*.

Zhiquan Wen, Qi Wu, Leyuan Fang, and Mingkui Tan. 2023b. Transformer-based relational inference network for complex visual relational reasoning. *ACM Trans. Multimedia Comput. Commun. Appl. (ACM TOMMCCAP)*.

Zhiquan Wen, Guanghui Xu, Mingkui Tan, Qingyao Wu, and Qi Wu. 2021. Debiased visual question answering from feature and sample perspectives. In *Conference on Neural Information Processing Systems (NeurIPS)*, volume 34, pages 3784–3796.

Jialin Wu and Raymond J. Mooney. 2019. Self-critical reasoning for robust visual question answering. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 8601–8611.

Liuyu Xiang, Guiguang Ding, and Jungong Han. 2020. Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In *European Conference on Computer Vision (ECCV)*, volume 12350, pages 247–263.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander J. Smola. 2016. Stacked attention networks for image question answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21–29.

Ziqi Zhang, Yaya Shi, Chunfeng Yuan, Bing Li, Peijin Wang, Weiming Hu, and Zheng-Jun Zha. 2020. Object relational graph with teacher-recommended learning for video captioning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13275–13285.

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

6 Appendix

6.1 Datasets

We conduct experiments on the VQA-CP (Agrawal et al., 2018) v2 and VQA (Goyal et al., 2017) v2 datasets. Specifically, the training set of VQA-CP v2 contains approximately 121k images and 483k questions, while the test set contains around 98k images and 220k questions.

6.2 Compared methods

We compare our DDG with existing state-of-the-art methods, including 1) methods that enhance visual attention: HINT (Selvaraju et al., 2019) and SCR (Wu and Mooney, 2019). 2) Methods that weaken the biases: AdvReg (Ramakrishnan et al., 2018), RUBI (Cadène et al., 2019b), Re-Scaling (Guo et al., 2022), DLR (Jing et al., 2020), VGQE (KV and Mittal, 2020), LMH (Clark et al., 2019), IntroD (Niu and Zhang, 2021), CF-VQA (Niu et al., 2021), RMFE (Gat et al., 2020), CKCL (Pan et al., 2022), LPF (Liang et al., 2021), GGE-DQ (Han et al., 2021), and D-VQA (Wen et al., 2021). 3) Methods that balance the dataset using additional annotations: CSS (Chen et al., 2020), CSS+CL (Liang et al., 2020), CSS$^+$ (Chen et al., 2021), ECD (Kolling et al., 2022), and Mutant (Gokhale et al., 2020). (4) Methods that balance the dataset without introducing additional annotations: CVL (Abbasnejad et al., 2020), Unshuffling (Teney et al., 2021), MMBS (Si et al., 2022), SimpleAug (Kil et al., 2021), RandImg (Teney et al., 2020), SSL-VQA (Zhu et al., 2020), and KDDAug (Chen et al., 2022). Our DDG generates positive and negative samples without introducing additional annotations to help alleviate the biases, which belongs to the methods in the fourth part.

6.3 Implementation Details

Following existing VQA methods (Anderson et al., 2018; Cadène et al., 2019b; Zhu et al., 2020), we extract the top-36 object features with a dimension of 2048 in each image by the Faster-RCNN (Ren et al., 2015) model that is pre-trained by (Anderson et al., 2018). Moreover, each question is first truncated or padded into the same length (*i.e.,* 14), and then encoded by the Glove (Pennington et al., 2014) embedding with a dimension of 300. The dimension of the question encoder (*i.e.,* single layer GRU (Cho et al., 2014)) is 1280.

Inspired by SSL-VQA (Zhu et al., 2020), we introduce one Batch Normalisation (Ioffe and Szegedy, 2015) layer before the classifier of UpDn (Anderson et al., 2018). We train our method for 30 epochs with the Adam (Kingma and Ba, 2015) optimiser. Specifically, we adopt $\mathcal{L}_{ens}$ and $\mathcal{L}_{dis}$ to train the baseline model for 12 epochs, and introduce $\mathcal{L}_{neg}$ and $\mathcal{L}^{weight}$ at the 13-th epoch. The learning rate is set to $1e$-3, and decreases by half every 5 epochs after 10 epochs. The batch size is set to 256. We set k and λ to 10 and 0.05, respectively. We implement our method based on PyTorch (Paszke et al., 2019), and the model is trained with one Titan Xp GPU. Moreover, our method does not introduce additional parameters except the backbone model.

Note that our method is model-agnostic and can be applied to different backbones of VQA models. To better demonstrate the effectiveness of our DDG, we conduct experiments based on different backbones, including UpDn (Anderson et al., 2018), and SAN (Yang et al., 2016) in the same settings. Moreover, we perform experiments over three rounds using varying seeds, and present the results in terms of mean values. The source code and the pre-trained models are available at DDG.

6.4 Training Method

We provide the training method of our DDG in Algorithm 1. Specifically, when the training epoch is lower than the threshold τ, we forward the base model $\mathcal{M}_b$ with the original and positive samples to calculate the knowledge distillation loss $\mathcal{L}_{dis}$ and binary cross-entropy loss $\mathcal{L}_{ens}$. Moreover, when the training epoch is higher than threshold τ, we construct the negative image and question samples without introducing additional annotations, and then forward $\mathcal{M}_b$ with the negative samples and obtain the loss $\mathcal{L}_{neg}$ and $\mathcal{L}^{weight}$. Finally, we update $\mathcal{M}_b$ based on the overall loss $\mathcal{L}$.

6.5 More Ablation Studies

Effect of the training strategy. As shown in Algorithm 1, we adopt knowledge distillation loss $\mathcal{L}_{dis}$ and ensemble binary cross-entropy loss $\mathcal{L}_{ens}$ to train the base model for 12 epochs. To demonstrate the effectiveness of the training strategy, we conduct experiments about the training strategies on the VQA-CP v2 dataset, and the results are shown in Table 7. From the results, the strategy that trains the model with the KL loss in the whole training process performs worse than that training for 12 epochs. We infer that the KL loss may accelerate the fitting of the training dataset, which hinders the

Algorithm 1 Training method of our DDG.

```
Require: Training data $\{(v_i, q_i, a_i)\}_{i=1}^{N}$, generated positive image samples $\{(v_i^{+}, q_i, a_i)\}_{i=1}^{N}$, generated positive question samples $\{(v_i, q_i^{+}, a_i)\}_{i=1}^{N}$ a base model $\mathcal{M}_b$, batch size $b$, threshold $\tau$.
Randomly initialise the parameters of $\mathcal{M}_b$.
while not converge do
    Randomly sample a mini-batch data $\{(v_i, q_i, a_i)\}_{i=1}^{b}$ from the training data, and obtain the corresponding positive samples $\{(v_i^{+}, q_i, a_i)\}_{i=1}^{b}$, and $\{(v_i, q_i^{+}, a_i)\}_{i=1}^{b}$.
    Forward $\mathcal{M}_b$ with the training data and the positive samples, and then obtain the predictions $P(\mathcal{A}|v_i, q_i)$, $P(\mathcal{A}|v_i^{+}, q_i)$, $P(\mathcal{A}|v_i, q_i^{+})$, and the ensemble prediction $P_{ens}$.
    Calculate the binary cross-entropy loss $\mathcal{L}_{vqa}$ for $P(\mathcal{A}|v_i, q_i)$ by Eq.(2).
    if the training epoch lower than $\tau$ then
        // Introduce Knowledge Distillation Mechanism
        Calculate the Knowledge Distillation Loss $\mathcal{L}_{dis}$ based on $P(\mathcal{A}|v_i, q_i)$ and $P_{ens}$ via Eq. (3).
        Calculate the binary cross-entropy loss $\mathcal{L}_{ens}$ for $P_{ens}$ by Eq. (2).
    else
        // Introduce Negative sample loss
        Randomly sample images and questions from the mini-batch data $\{(v_i, q_i, a_i)\}_{i=1}^{b}$ to form the negative samples as $\{(\bar{v}_i, q_i, a_i)\}_{i=1}^{b}$ and $\{(v_i, \bar{q}_i, a_i)\}_{i=1}^{b}$.
        Forward $\mathcal{M}_b$ with negative samples, and obtain the predictions $P(\mathcal{A}|v_i^{-}, q_i)$ and $P(\mathcal{A}|v_i, q_i^{-})$.
        Calculate the loss $\mathcal{L}_{neg}$ based on the predictions of the negative samples via Eq. (4).
        Calculate the loss $\mathcal{L}^{weight}$ based on the predictions of both positive and negative samples by Eqs. (6) and (7).
    end if
    Update $\mathcal{M}_b$ by minimising the overall loss $\mathcal{L}$ (obtained via Eq. (8)).
end while
```

training process of the negative sample losses $\mathcal{L}_{neg}$ and $\mathcal{L}^{weight}$.

The objective of the KL loss is to make the two distributions close. In our method, we seek to make the predictions of the original samples approach to the ensemble predictions. Thanks to the generated high quality positive samples, the KL loss can improve the robustness of the VQA models (Refer to in Line 1-2 of Table 7), to some extent. However, if we adopt the KL loss in the whole training process, the VQA models will fit the data distributions excessively, and thus may hinder the training process of the negative sample losses. The experimental results in Table 7 also confirm it. For example, our DDG with the training strategy that introduces KL loss in the overall training process still performs better than that training using only ($\mathcal{L}_{dis}$ and $\mathcal{L}_{ens}$), but performs worse than that training the model using KL loss for 12 epochs.

Strategy	VQA-CP v2 test set (%)			
	All	Yes/No	Number	Other
UpDn†	41.53	43.45	13.64	48.18
+ $\mathcal{L}_{dis} + \mathcal{L}_{ens}$ (all epochs)	47.77	63.49	13.91	48.83
+ DDG (KL for all epochs)	56.35	86.05	21.30	**50.40**
+ DDG (KL for 12 epochs)	**61.14**	**88.77**	**49.33**	49.90

Table 7: Effect of the training strategy. We report the experimental results in terms of Accuracy (%). "$\mathcal{L}_{dis} + \mathcal{L}_{ens}$ (all epochs)" denotes we additionally introduce $\mathcal{L}_{dis}$ and $\mathcal{L}_{ens}$ losses to train the UpDn model in the whole training process. "DDG (KL for all epochs)" means we train the UpDn model with the DDG method, where the KL loss exists in the whole training process. "DDG (KL for 12 epochs)" denotes we train the UpDn model with the DDG method, and the KL loss exists in the first 12 epochs.

Comparison with the state-of-the-art methods regarding of the number of training samples. As shown in Table 8, the VQA-CP v2 dataset comprises 438k training samples, while KDDAug, SimpleAug, and our DDG generate an additional 4088k, 3081k, and 1752k augmented training samples, respectively. Despite using fewer augmented samples than KDDAug and SimpleAug, our DDG outperforms these methods by around 8% and 1%, respectively, which demonstrates the effectiveness of our DDG.

Model	VQA-CP v2 test set (%)	# Samples
UpDn (Anderson et al., 2018)	41.53	438k
SimpleAug (Kil et al., 2021)	52.65	+3081k
KDDAug (Chen et al., 2022)	60.24	+4088k
DDG (Ours)	**61.14**	**+1752k**

Table 8: Comparison with the state-of-the-art data augmentation based methods (*e.g.,* SimpleAug and KDDAug) on the VQA-CP v2 dataset.

6.6 More Visualisation Results

Qualitative results. We provide more visualisation results in Figure. 4 to present the effectiveness of our DDG. From the results, our DDG localise the target objects more accurately than the UpDn (Anderson et al., 2018) model and SSL-VQA method, and thus makes a more correct prediction than the compared methods. These visualisation results demonstrate the effectiveness of our DDG.

Visualisation of the generated samples. As shown in Section 3.2, we have generated both positive image and question samples. To evaluate the generated methods, we provide some visualisation results about the augmented questions and selected target objects in Table 9 and Figure. 5, respectively. From the results in Table 9, our augment questions have similar semantics to the original questions, which demonstrates our generated questions are reasonable as the positive samples. Moreover, although the baseline model UpDn (Anderson et al., 2018) trained on the VQA-CP (Agrawal et al., 2018) v2 dataset achieves poor performance on the test set, the UpDn model still can obtain good performance on the training set. Thus, we adopt the image attention weights of the pre-trained UpDn model to help find the objects that are relevant to the questions. As shown in Figure. 5, we show the objects with the top-3 attention weights of the pre-trained UpDn model in the images. From the results, the pre-trained UpDn model can localise the target objects referred to in the questions, which demonstrates that selecting the objects with top-k image attention weights of the pre-trained UpDn model is reasonable, and can exclude background information.

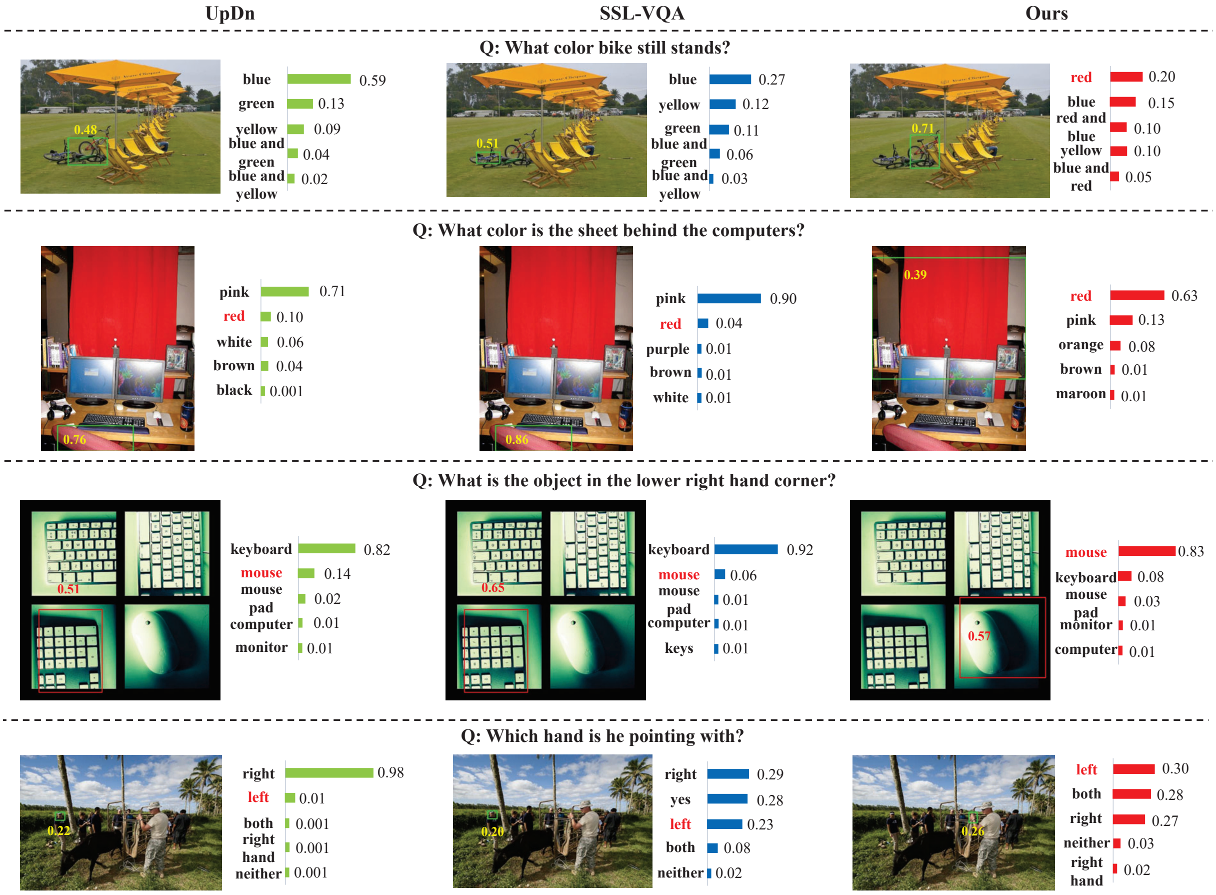
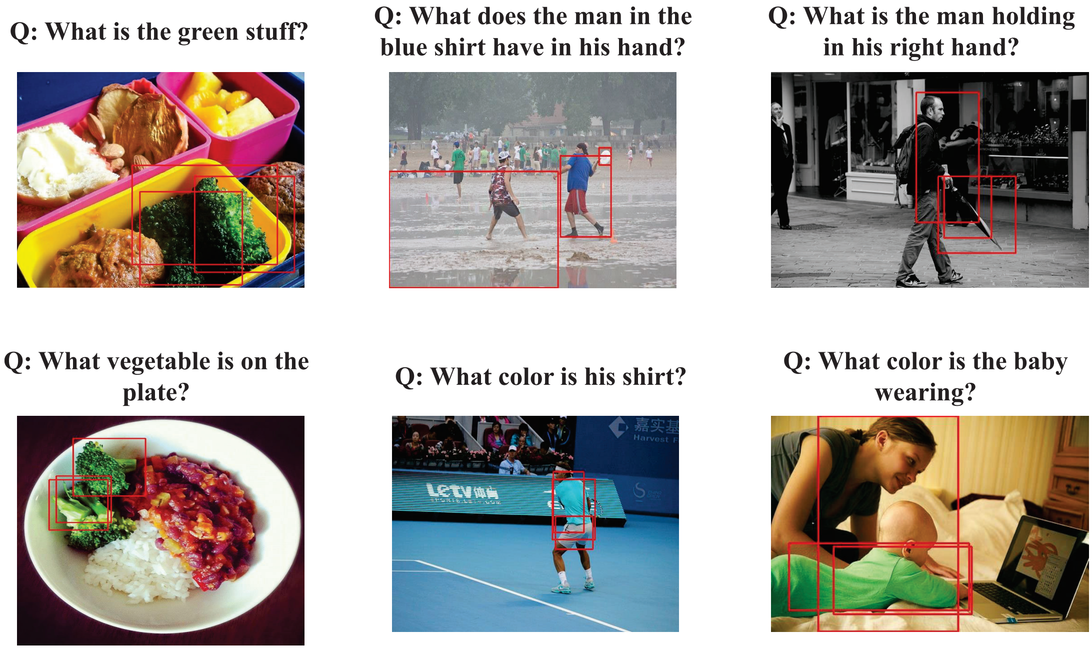


Figure 4: Qualitative comparison among UpDn (Anderson et al., 2018), SSL-VQA (Zhu et al., 2020), and our DDG on the VQA-CP v2 test set. For each example, we put the bounding box with the highest attention weight in the image and show the answers with the top-5 predictions. The bold, red answer is the ground-truth answer.

Original Question	Augment Question
What color is the bike?	What color does the bike have?
Is this plane landing?	Is the plane going to land?
How many pans are visible?	How many pans can be seen?
Is the zebra's tail up?	Is the tail of the zebra raised?
How do you turn on the cold water?	How do you turn the cold water on?
What is the woman in the room doing?	What does the woman do in the room?
What type of silverware is on the plates?	What kind of silverware is there on plates?
What direction are the animals heading?	In what direction are animals going?
Are there lots of healthy options on the table?	Are there many healthy options on the table?
What does the black machine next to the man produce?	What is the black machine producing next to the man?
How many floors do you think the highest building has?	How many floors does the tallest building have in your opinion?

Table 9: We provide some visualisation results of the augmented questions based on our generation technique referred to in Section 3.2.



Q: What is the green stuff?

Q: What does the man in the blue shirt have in his hand?

Q: What is the man holding in his right hand?

Q: What vegetable is on the plate?

Q: What color is his shirt?

Q: What color is the baby wearing?

Figure 5: We provide some visualisation results of the augmented images based on our generation technique referred to in Section 3.2. We put the bounding box with the top-3 attention weight of the pre-trained UpDn (Anderson et al., 2018) model in the image.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
We discuss the limitations of our work in Section Limitations

☑ A2. Did you discuss any potential risks of your work?
We discuss the potential risks of our work in Section Ethics Statement

☑ A3. Do the abstract and introduction summarize the paper's main claims?
In Section Abstract and Introduction

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

We conduct experiments on the VQA v2 and VQA-CP v2 datasets. In overall Sections

☑ B1. Did you cite the creators of artifacts you used?
In overall sections

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
In Section Appendix

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
In Section Appendix

C ☑ Did you run computational experiments?

In Section Experiments

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
In Section Appendix

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
In Section Experiments and Section Appendix

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
In Section Experiments and Appendix

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
In Section Appendix

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.