Towards Robust Visual Question Answering: Making the Most of Biased Samples via Contrastive Learning

Anonymous ACL submission

Abstract

Models for Visual Question Answering (VQA) often rely on the spurious correlations, i.e., the language priors, that appear in the biased samples of training set, which make them brittle against the out-of-distribution (OOD) test data. Recent methods have achieved promising progress in overcoming this problem by reducing the impact of biased samples on model training. However, these models reveal a trade-off that the improvements on OOD data severely sacrifice the performance on the in-distribution (ID) data (which is dominated by the biased samples). Therefore, we propose a novel contrastive learning approach, MMBS, for building robust VQA models by **M**aking the **M**ost of **B**iased **S**amples. Specifically, we construct positive samples for contrastive learning by eliminating the information related to spurious correlation from the original training samples and explore several strategies to use the constructed positive samples for training. Instead of undermining the importance of biased samples in model training, our approach precisely exploits the biased samples for unbiased information that contributes to reasoning. The proposed method is compatible with various VQA backbones. We validate our contributions by achieving competitive performance on the OOD dataset VQA-CP v2 while preserving robust performance on the ID dataset VQA v2.

1 Introduction

Visual Question Answering (VQA), aiming to answer a question about the given image, is a multi-modal task that involves the intersection between vision and language. Despite the remarkable performance on many VQA datasets such as VQA v2 (Goyal et al., 2017), recent studies (Antol et al., 2015; Kafle and Kanan, 2017; Agrawal et al., 2016) find that the VQA systems rely heavily on the language priors. They are caused by the strong spurious correlation between certain question category and answers, e.g., the frequent co-occurrence of the question category 'what sport' and the answer 'tennis' (Selvaraju et al., 2019). As a result, the VQA models, which are over-reliant on the language priors of training set, fail to generalize to the OOD dataset, VQA-CP v2 (Agrawal et al., 2018).

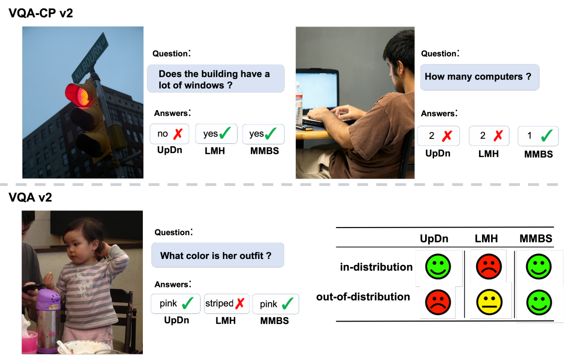

Figure 1: Qualitative comparison of our method LMH+MMBS against the plain method UpDn and the debiasing method LMH. In VQA-CP v2 (upper), the question types ('Does the' and 'How many') bias UpDn to the most common answers (see Fig. 5 for the answer distribution). LMH alleviates the language priors for yesno questions (upper left), while it fails on the more difficult non-yesno questions (upper right). Besides, LMH damages the ID performance, giving an uncommon answer to the common sample from VQA v2 (lower right). MMBS improves the OOD performance while maintains the ID performance (lower right).

Recently, several methods achieved remarkable progress in overcoming this language prior problem. They assign less importance to the biased samples that can be correctly classified with the spurious correlation. However, most of them achieve gains on VQA-CP v2 at the cost of degrading the model's ID performance on the VQA v2 dataset (see Tab. 2). This trade-off suggests that the success of these methods merely comes from biasing the models to other directions, rather than endowing them with the reasoning capability and robustness to language priors. Ideally, a robust VQA system should maintain its performance on the ID dataset while overcoming the language priors, as shown in Fig. 1.

We think the essence of both language-prior and trade-off problems is about the learning of biased samples. The former is caused by over-reliance on biased information from biased samples, while the latter is caused by undermining the importance of biased samples. Therefore, if a model can precisely exploit the biased samples for intrinsic information of the given task, both problems can be alleviated simultaneously.

Motivated by this, we propose a self-supervised contrastive learning method (MMBS) for building robust VQA systems by **M**ake the **M**ost of **B**iased **S**amples. Firstly, in view of the characteristics of the spurious correlations, we construct two kinds of positive samples for the questions of training samples to exploit the unbiased information, and then design four strategies to use the constructed positive samples. Next, we propose a novel algorithm to distinguish between biased and unbiased samples, so as to treat them differently. On this basis, we introduce an auxiliary contrastive training objective, which helps the model learn a more general representation with ameliorated language priors by narrowing the distance between original samples and positive samples in the cross-modality joint embedding space.

To summarize, our contributions are as follow: **i**) We propose a novel contrastive learning method, which effectively addresses the language prior problem and the ID-OOD performance trade-off in VQA, by making the most of biased samples. **ii**) We propose an algorithm to distinguish between biased and unbiased samples and treat them differently in contrastive learning. **iii**) Experimental results demonstrate that our method is compatible with various VQA backbones and achieve competitive performance on the language-bias sensitive VQA-CP v2 dataset while preserving the original accuracy on the in-distribution VQA v2 dataset.

2 Related Work

Overcoming Language Priors in VQA. Recently, the language biases in VQA datasets raised the attention of many researchers (Goyal et al., 2017; Antol et al., 2015; Agrawal et al., 2016; Kervadec et al., 2021) (see **App.** A.1 for details). In response to this problem, numerous methods are proposed to debias the VQA models. The most effective ones of them can be roughly divided into two categories: **Ensemble-based methods** (Grand and Belinkov, 2019; Belinkov et al., 2019; Cadene et al., 2019; Clark et al., 2019; Mahabadi and Henderson, 2019; Niu et al., 2021) introduce a biased model, which is designed to focus on the spurious features, to assist the training of the main model. For example, the recent method LPF (Liang et al., 2021) leverages the output distribution of the bias model to down-weight the biased sample when computing the VQA loss. However, these methods neglect the useful information that helps reasoning in biased samples. **Data-balancing methods** (Zhu et al., 2020; Liang et al., 2020) balance the training priors. For example, CSS and Mutant (Chen et al., 2020; Gokhale et al., 2020) generate samples by masking the critical object in images and word in questions and by semantic image mutations respectively. These methods usually outperform other debiasing methods with a large margin on VQA-CP v2, because they bypass the challenge of the imbalanced settings (Liang et al., 2021; Niu et al., 2021) by explicitly balancing the answers' distribution at the training stage. Though our method constructs the positive questions, it does not change the training answers' distribution. We also extend our method to the data-balancing method SAR (Si et al., 2021). Another line is the visual grounding methods which are shown in **App.** A.2.

Contrastive Learning in VQA. Recently, the contrastive learning is well-developed in unsupervised learning (Oord et al., 2018; He et al., 2020) while its application in VQA is still in initial stage. CL (Liang et al., 2020) is the first work to employ contrastive learning to improve VQA model's robustness. Its motivation is to learn a better relationship among the input sample and the factual and counterfactual sample which are generated by CSS. However, CL brings weak OOD performance gain and ID performance drop based on CSS. In contrast, our method attributes the key point of solving language bias to the positive-sample designs for excluding the spurious correlations. It is model-agnostic and can boost models' OOD performance significantly while retain the ID performance.

3 Method

Fig. 2 shows MMBS's overview, which includes: 1) A backbone VQA model; 2) A positive sample construction module; 3) An unbiased sample selection module; 4) A contrastive learning objective.

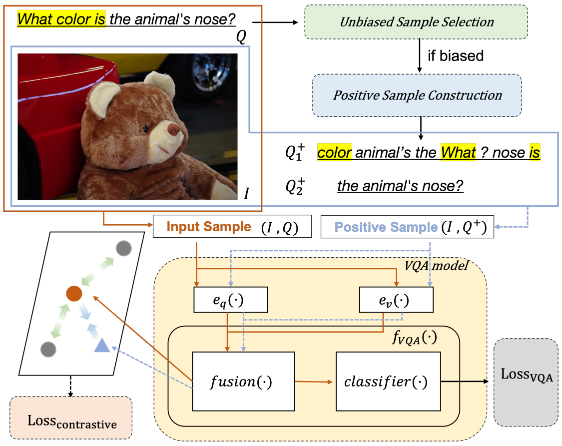

Figure 2: Overview of our method. The question category words are highlighted in yellow. The orange circle and blue triangle denote the cross-modality representations of the original sample and positive sample. The other samples in the same batch are the negative samples, which are denoted by the gray circles.

3.1 Backbone VQA Model

The backbone VQA model is a free choice in MMBS. The widely-used backbone models (Anderson et al., 2018; Mahabadi and Henderson, 2019) treat VQA as a multi-class multi-label classification task. Concretely, given a VQA dataset $D = \{I_i, Q_i, A_i\}_{i=1}^N$ with N samples, where $I_i \in I$, $Q_i \in Q$ are the image and question of the i_{th} sample. $A_i \in A$ is the ground-truth answer which is usually in multi-label form, and tgt_i is the corresponding target score of each label (see **App.** B.1 for details). Most existing VQA models consist of four parts: the question encoder $e_q(\cdot)$, the image encoder $e_v(\cdot)$, the fusion function $F(\cdot)$ and the classifier $clf(\cdot)$. For example, LXMERT (Tan and Bansal, 2019) encodes image and caption text separately to extract visual features $V_i = e_v(I_i)$, and textual features $T_i = e_q(Q_i)$, in two streams. Next, the higher co-attentional transformer layers fuse the two features and project them into the cross-modality joint embedding space, i.e., $F(V_i, T_i)$. Finally, the classifier outputs the answer prediction:

$$P(A|I_i, Q_i) = clf(F(V_i, T_i)) \tag{1}$$

The training objective minimizes the multi-label soft loss, L_{vqa} , which can be formalized as follow:

$$\begin{aligned} L_{vqa} = -\frac{1}{N}\sum\nolimits_{i=1}^{N}[tgt_i \cdot log(\delta(F(V_i, T_i))) \\ + (1 - tgt_i) \cdot log(1 - \delta(F(V_i, T_i)))] \end{aligned} \tag{2}$$

where δ denotes the sigmoid function.

3.2 Positive Sample Construction

To make the most of the unbiased information contained in the biased sample, we first construct the positive samples which exclude the biased information. According to the construction of VQA-CP v2, there is a shift between the training and test set in terms of answer distribution under the same question category (Teney et al., 2020; Agrawal et al., 2018). As a result, the frequency co-occurrence of certain answer and question category in the training set produces a major source of bias. Therefore, we construct two kinds of positive questions (Q_i^+) by corrupting the question category information of each input question (Q_i):

Shuffling: We randomly shuffle the words in the question sentence so that the question category words are mixed with the other words. This increases the difficulty of building the correlations between question category and answer.

Removal: We remove the question category words from the question sentence. It eliminates the co-occurrence of answer and question category words completely.

We notice that the construction process could induce some unexpected noise in the positive samples. To tackle this concern, we present more positive samples in **App.** B.2 and discuss their quality and potential impact on our method.

We also propose four strategies for using the constructed positive questions during training:

S: *Use the **Shuffling** positive questions.*

R: *Use the **Removal** positive questions.*

B: *Use **both** positive questions.*

SR: *Use the **Shuffling** positive questions for non-yesno (i.e., 'Num' and 'Other') questions and use the **Removal** ones for yesno (i.e., 'Y/N') questions.*

The ***SR*** strategy deals with yesno and non-yesno questions in different ways based on their characteristics. Intuitively, the question categories of the yesno questions usually contain little information, as they are mostly comprised of 'is', 'do', etc. By contrast, the question categories of non-yesno questions tend to contain more information which is important for answering correctly. Therefore, ***Removal*** is not applied to non-yesno questions.

Adopting any strategy above, we can obtain the positive samples $\{I_i, Q_i^+\}_{i=1}^B$ for input samples$\{I_i, Q_i\}_{i=1}^B$. The negative samples $\{I_b, Q_b\}_{b=1}^B$, where $b \neq i$, are the other samples in the same batch. B is the batch size of training.

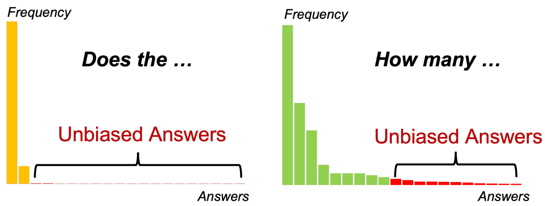

Figure 3: The answers' distributions of the yesno questions with "Does the" (left) and non-yesno questions with "How many" (right). The former has a low entropy and the latter has a high entropy.

3.3 Unbiased Sample Selection

Following Kervadec et al. (2021), we define unbiased (or OOD) samples as the infrequent samples in the answers' distribution of each question category in training set. Therefore, the unbiased samples are unlikely to contain spurious correlations, which makes them beneficial to OOD robustness. Moreover, some unexpected noise in the positive samples may negatively impact the learning of unbiased samples. For the above reasons, we do not construct positive samples for the unbiased samples. To filter out the unbiased samples, we propose a novel algorithm, consisting of three steps: (i) calculating the answer frequencies; (ii) determining the unbiased answer proportion; (iii) selecting the unbiased samples.

Answer frequencies. We denote the i_{th} sample's question category, ground truth answer and soft target score as $C_i \in C$ (65 categories in total), A_i and tgt_i respectively. We measure how frequent the answer A_j appears in the question category C_k as follows:

$$Freq_{C_k}^{A_j} = \sum\nolimits_{i=1}^{M_{C_k}} (tgt_i) \ , \ if \ A_i = A_j \tag{3}$$

where M_{C_k} is the number of all samples with the same category C_k. If a sample has a multi-label answer A_i, we count each answer's score respectively. A lower value of $Freq_{C_k}^{A_j}$ indicates weaker spurious correlations between A_j and C_k, and thus the corresponding samples are deemed as unbiased. We introduce a hyper-parameter $\beta \in [0, 1]$ to control the proportion of the unbiased samples.

Entropy-based correction factor. The answers' distributions of $|C|$ question categories are different. Empirically, when the entropy of an answers' distribution is lower, more answers will be associated with only a few samples, so that the unbiased answer proportion should be higher. Otherwise, it should be lower. An illustration is given in Fig. 3. Therefore, we propose an entropy-based correction factor W_{C_k} to dynamically adjust the β for each category C_k:

$$\begin{aligned} W_{C_k} &= 1 - sigmoid(E_{C_k} - mean(E)) \\ E_{C_k} &= Entropy(Freq_{C_k}/SUM) \end{aligned} \tag{4}$$

where E represents $\{E_{C_k}\}_{k=1}^{|C|}$ and SUM represents the sum of $Freq_{C_k}$. When the entropy is lower, the W_{C_k} is closer to 1, and otherwise W_{C_k} is closer to 0. Finally, we obtain the unbiased answer proportion $P_{C_k} = W_{C_k} * \beta$.

Selecting unbiased samples. For each question category C_k, we obtain a list of unbiased answers which rank in the last P_{C_k} in $Freq_{C_k}$. Then we determine the samples whose ground truth (highest-score) answer belongs to this list as unbiased samples. The unbiased sample statistics are shown in **App.** B.3. If a sample is biased, we adopt the strategy mentioned in previous section to construct its positive sample. If it is unbiased, we use the original sample as its positive sample.

3.4 Contrastive Learning Objective

Given input sample (I_i, Q_i), we have the positive sample (I_i, Q_i^+) and the negative samples $(I_b, Q_b)_{b=1}^B$ in the same batch, where $b \neq i$. After feeding them into the VQA model, we obtain the cross-modality fusion representation of the input sample, $F(V_i, T_i)$, positive sample $F(V_i, T_i^+)$ and negative samples $F(V_b, T_b)_{b=1}^B$, which are denoted as the anchor a, the positive p and the negative $n_{b}{}_{b=1}^{B}$ respectively. Following (Robinson et al., 2020; Liang et al., 2020), we use the cosine similarity, $cos(\cdot)$, as the scoring function. The contrastive loss (Oord et al., 2018) is formulated as:

$$L_{cl} = \mathop{\mathbb{E}}_{a,p,n_b} \left[-\log \frac{e^{cos(a,p)}}{e^{cos(a,p)} + \sum_{b=1}^{B} e^{cos(a,n_b)}} \right] \tag{5}$$

By minimizing it, the models can focus on the unbiased information from the positive question. The overall loss of MMBS is formulated as: $L = L_{vqa} + \alpha * L_{cl}$, where α is the weight of L_{cl}.

3.5 Inference Process

After training with this contrastive loss, the models can handle the question in ***original***, ***Shuffling*** and ***Removal*** forms (Sec. 3.2) in the inference phase.[1]

[1]The models without MMBS performs much worse when the question is in ***Shuffling*** or ***Removal*** forms.

		VQA-CP v2 test					VQA v2 val				
	Methods	All	Y/N	Num	Other	Gap ↑	All	Y/N	Num	Other	Gap ↑
Plain Models	BAN	37.03	41.55	12.43	41.4	**+10.60**	63.9	81.42	45.18	55.54	**+0.88**
	+MMBS	**47.63**	**66.18**	**16.36**	**46.49**		**64.78**	**82.03**	**46.48**	**56.51**	
	UpDn	39.74	42.27	11.93	46.05	**+8.45**	63.48	**81.18**	42.14	55.66	**+0.36**
	+MMBS	**48.19**	**65.00**	**14.05**	**48.75**		**63.84**	79.61	**44.23**	**57.05**	
	LXM	47.19	50.55	24.06	51.77	**+9.32**	**71.01**	88.24	54.07	**62.39**	-0.16
	+MMBS	**56.51**	**79.83**	**28.70**	**51.92**		70.85	**88.25**	**55.67**	61.63	
Debiasing Models	LMH	52.01	72.58	31.12	46.97	**+4.43**	56.35	65.06	37.63	54.69	**+5.52**
	+MMBS	**56.44**	**76.00**	**43.77**	**49.67**		**61.87**	**75.86**	**40.34**	**56.95**	
	SAR	66.73	86.00	62.34	57.84	**+1.66**	69.22	**87.46**	**51.20**	60.12	**+0.21**
	+MMBS	**68.39**	**87.30**	**65.21**	**59.36**		**69.43**	87.39	50.37	**60.82**	

Table 1: Results on VQA-CP v2 test and VQA-v2 validation set based on different VQA models. 'Gap' denotes the accuracy improvement of MMBS over the base model.

We find that in the framework of MMBS, ***Shuffling*** can further boost OOD performance for the plain models (e.g., UpDn and LXM), while ***original*** performs the best for debiasing methods (e.g., LMH, SAR). Therefore, we shuffle the question words at test time when applying MMBS to the plain models. Detailed discussions are shown in **App.** D.1.

4 Experiments

4.1 Datasets and Evaluation

We evaluate our models on the OOD VQA-CP v2 (Agrawal et al., 2018) and the ID VQA v2 (Goyal et al., 2017) with the standard evaluation metric (Antol et al., 2015) based on accuracy. Previous works (Chen et al., 2020; Si et al., 2021; Gokhale et al., 2020) think that a minor accuracy difference between VQA v2 and VQA-CP v2 shows the real robustness. This encourages the researchers to work in the direction that increases the accuracy on VQA-CP v2 by sacrificing the performance on VQA v2. However, a robust VQA model should perform well on both datasets. Therefore, we compute the relative accuracy between each method and its base method on both ID and OOD datasets.

4.2 Baselines and Implementations

Our approach is general to various VQA backbones. In the work, we evaluate MMBS based on three plain VQA models (which are not specially designed for overcoming language priors): **BAN** (Kim et al., 2018), **UpDn** (Anderson et al., 2018) and **LXMERT** (LXM), and two debiasing methods: **LMH** (Clark et al., 2019) and **SAR** (Si et al., 2021).

We also compare our methods with the state-of-the-art methods on VQA-CP v2, which contain: 1) The ensemble-based methods: **AdvReg.** (Ramakrishnan et al., 2018), **GRL** (Grand and Belinkov, 2019), **RUBi** (Cadene et al., 2019), **DLR** (Jing et al., 2020a), **LMH** (Clark et al., 2019), **CF-VQA** (Niu et al., 2021), **LPF** (Liang et al., 2021). 2) The data-balancing methods: **SSL** (Zhu et al., 2020), **CSS** (Chen et al., 2020), **CL** (Liang et al., 2020), **SAR** (Si et al., 2021) and **MUTANT** (best-performance method) (Gokhale et al., 2020). The visual-grounding methods perform much worse (see **App.** D.2 [2]), so we do not conduct further comparison with them.

Following the baselines above, the checkpoint for evaluation is also picked by the test set directly in the work due to the lack of val set (Teney et al., 2020; Agrawal et al., 2018). In this paper, we mainly report the results with ***SR*** strategy. We also conduct experiments to analyze the impact of different positive-sample construction strategies. More implementation details are shown in **App.** C.

4.3 Main Results

Performance based on different VQA models. As can be seen in Tab. 1, regardless of the backbone architectures and debiasing methods, our proposed method consistently outperforms the baselines with comfortable margin (1.66 ~10.60 absolute accuracy improvement) on OOD VQA-CP v2. For the plain models, MMBS particularly improves the performance on yesno questions (22.73 ~29.28) because the simple yesno questions are more susceptible to the influence of language bias (Zhu et al., 2020; Liang et al., 2021). In terms of the ID dataset, the baselines' performance can also be also improved or at least maintained with MMBS, while most debiasing methods sacrifice the accuracy on VQA v2 (see the corresponding column in Tab. 2). Especially, compared with LMH, LMH+MMBS gets a prominent accuracy boost of 5.52 on VQA v2. This is because making the most of biased samples can effectively alleviate the ID performance decline

[2] We list the complete results of all baselines together in **App.** D.2 for detailed comparison.

	VQA-CP v2 test					VQA v2 val		Gaps
Methods	All	Y/N	Num	Other	Gap ↑	All	Gap ↑	Sum
UpDn	39.74	42.27	11.93	46.05		63.48		
+AdvReg.	41.17	65.49	15.48	35.48	+1.43	62.75	-0.73	+0.70
+GRL	42.33	59.74	14.78	40.76	+2.59	51.92	-11.56	-9.00
+RUBi	44.23	67.05	17.48	39.61	+4.49	61.16	-2.32	+2.17
+DLR	48.87	70.99	18.72	45.57	+9.13	57.96	-5.52	+3.61
+LMH	52.01	72.58	31.12	46.97	+12.27	56.35	-7.13	+5.14
+CF-VQA	53.55	91.15	13.03	44.97	+13.81	63.54	+0.06	+13.87
+LPF	55.34	88.61	23.78	46.57	+15.60	55.01	-8.47	+7.13
+LMH+**MMBS**	**56.44**	76.00	43.77	49.67	**+16.70**	61.87	-1.61	**+15.09**
LXM	47.19	50.55	24.06	51.77		71.01		
+LMH*	63.34	78.28	65.95	54.79	+16.15	69.49	-1.52	+14.63
+U-SAR*	64.98	81.89	59.65	57.61	+17.79	69.17	-1.84	+15.95
+LMH+**MMBS**	**65.70**	81.70	61.24	58.54	**+18.51**	70.29	-0.72	**+17.79**
+U-SAR+**MMBS**	**68.01**	86.55	64.69	59.21	**+20.82**	69.29	-1.72	**+19.10**

Table 2: Comparison with the state-of-the-art ensemble-based methods. 'Gap' denotes the accuracy improvement of the debiasing methods over their base models. * denotes the strong baselines introduced in this paper.

resulting from the debiasing method LMH.

Comparison with ensemble-based SOTAs. The upper part of Tab. 2 compares the methods based on the UpDn backbone. We can observe that: 1) Compared with UpDn, most ensemble-based methods suffer from obviously performance drops on VQA v2. This phenomenon attests to the trade-off between the ability to overcome the language priors and the ability to memorize the knowledge of in-distribution samples. Though to a certain extent, CF-VQA alleviates the phenomenon, its accuracy on VQA-CP v2 is prominently lower than our method. 2) LMH+MMBS performs the best on VQA-CP v2 and rivals the accuracy of the backbone on VQA v2, clearly surpassing the previous best in 'GapsSum'. This shows that the trade-off problem is effectively alleviated by the propose method. 3) The previous methods, e.g., CF-VQA and LPF, achieve high accuracy on the simple yesno question where the language biases are more likely to exist. By contrast, our method substantially improves over them on the more challenging non-yesno question, while achieves relatively good performance on the yesno questions.

The methods in the lower part of Tab. 2 are based on the LXM backbone. LXM is a cross-modal pre-trained model that has been used as backbone in some data-balancing method to further boost performance (Si et al., 2021; Gokhale et al., 2020). However, the performance of LXM with ensemble-based methods has not been fully investigated. We introduce two strong baselines based on LXM, i.e., LXM+LMH and U-SAR. LXM+LMH represents the LXM model trained with LMH method, which is widely used as an essential component by existing methods (Chen et al., 2020; Liang et al., 2020; Si et al., 2021). U-SAR is a variants of the

		VQA-CP v2 test		VQA v2 val		Gaps
Methods	Base	All	Gap↑	All	Gap↑	Sum
SSL	UpDn	57.59	+17.85	63.73	+0.25	+18.10
LMH+CCS	UpDn	58.95	+19.21	59.91	-3.57	+15.64
LMH+CCS+CL	UpDn	59.18	+19.44	57.29	-6.19	+13.25
SAR	LXM	66.73	+19.54	69.22	-1.79	+17.75
MUTANT	LXM	69.52	+22.33	70.24	-0.77	+21.56
SAR+**MMBS**	LXM	68.39	+21.20	69.43	-1.58	+19.62

Table 3: Comparison with the state-of-the-art data-balancing methods.

Method	Strategy	All	Y/N	Num	Other
UpDn	Base*	41.06	43.13	13.71	47.48
	S	42.26	45.11	13.99	48.52
	R	42.83	57.74	12.25	43.41
	B	44.37	51.58	**14.94**	48.67
	SR	**48.19**	**65.00**	14.05	**48.75**
LXM	Base*	47.19	50.55	24.06	51.77
	S	47.90	52.71	26.48	51.26
	R	52.11	63.65	27.89	**52.72**
	B	50.76	61.33	**29.21**	51.14
	SR	**56.51**	**79.83**	28.70	51.92
LMH	Base*	52.58	67.10	36.59	49.36
	S	55.89	76.67	37.64	50.01
	R	55.87	**76.79**	34.96	**50.65**
	B	55.62	76.47	35.71	50.15
	SR	**56.44**	76.00	**43.77**	49.67

Table 4: Results of different positive-sample construction strategies on the VQA-CP v2 test set.

two-stage method **SAR**, with the data-balancing method SSL replaced with UpDn. We can see that MMBS further promotes the two strong baselines, enhancing the OOD performance and relieving the ID performance drop. Moreover, the LXM-based MMBS is even competitive with the data-balancing methods that generate samples.

Comparison with data-balancing SOTAs. We can derive three observations from the results in Tab. 3: 1) Most data-balancing methods also hurt the ID performance, which is the result of a mismatch between the balanced training priors and the biased test priors. 2) Another existing contrastive learning model LMH+CSS+CL (Liang et al., 2020), which can only be applied to the data-balancing method LMH+CSS, achieves a mild improvement of 0.23 on VQA-CP v2 and sacrifices the accuracy on VQA v2. Compared with it, our MMBS is general to various VQA backbones and does not hurt the ID performance. 3) Our SAR+MMBS brings encouraging performance gain over the strong baseline SAR and achieves competitive performance against the best-performing method MUTANT without utilizing extra manual annotations to construct extensive data.

4.4 Analysis on Individual Components and Hyper-Parameters

The effect of positive sample construction strategies. As shown in Tab. 4, we conduct experiments based on three widely used methods, i.e., the

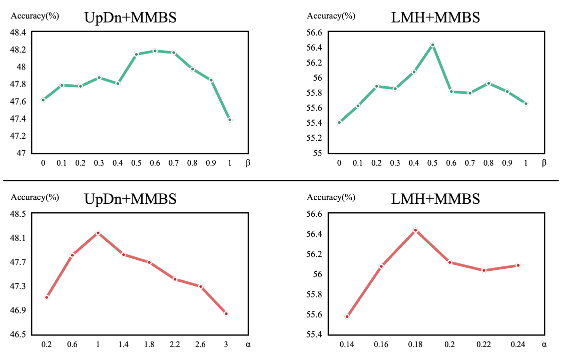

Figure 4: Results of UpDn+MMBS and LMH+MMBS on VQA-CP v2 with varying of β (upper) and α (lower).

plain model UpDn, pre-trained model LXM and UpDn with the debiasing method LMH. From the results UpDn and LXM, we can observe that: 1) Both ***S*** and ***R*** strategies gain performance boost. This shows that the designs of both of them are sound and effective, and their benefits outweigh the potential semantic noise. 2) ***R*** strategy has a better overall performance than ***S*** because the model may still learn the superficial correlation between answer and the question category even when the category words are shuffled with the other words of the sentence. 3) ***SR*** strategy performs the best among the four strategies, especially on the yesno questions. The reason is that ***R*** strategy significantly outperforms ***S*** strategy on the yesno questions while the ***S*** strategy performs well on the non-yesno questions. ***SR*** strategy combines the advantages of both strategies. 4) ***B*** strategy is obviously inferior to the ***SR*** strategy. This is because learning from two positive samples for each sample simultaneously may confuse the model.

From the results of LMH, we find that all the strategies considerably boost the performance, including the ***S*** strategy. This is because the unbiased information contained in biased samples, which is useful for reasoning, is also being neglected by the ensemble-based methods. Through the contrastive learning objective, both ***Shuffling*** and ***Removal*** positive samples give them another channel to learn and utilize the useful information. ***SR*** strategy still has the best performance among all the strategies.

The effect of β and α. As shown in the upper plots of Fig. 4, the accuracy rises first and then decreases as β increases. There is a trade-off behind this phenomenon: when β is too small, the method will construct the positive samples for the unbiased samples, which may affect the learning of robust information from the unbiased samples. When β is too large, the method will not construct positive samples for some biased samples. This demeans the profits from the contrastive learning objective.

The lower plots of Fig. 4 also revel a trade-off with the increase of α. This suggests that the contrastive learning objective is beneficial but paying too much attention to this objective hurts the final performance. we also find that the best α for LMH+MMBS is smaller than that for UpDn+MMBS. This is because LMH itself already has certain ability to alleviate language priors.

Method	All	Y/N	Num	Other
UpDn	41.06	43.13	13.71	47.48
UpDn+*SR*	47.62	62.72	13.92	**48.95**
UpDn+*SR*+β	48.00	64.06	14.10	48.89
UpDn+*SR*+β+W_C	**48.19**	**65.00**	14.05	48.75
LXM	47.19	50.55	24.06	51.77
LXM+*SR*	55.26	77.13	27.33	51.47
LXM+*SR*+β	55.66	78.64	28.10	51.17
LXM+*SR*+β+W_C	**56.51**	**79.83**	28.70	51.92
LMH	52.01	72.58	31.12	46.97
LMH+*SR*	55.41	76.50	37.20	49.35
LMH+*SR*+β	56.15	**77.46**	37.90	50.00
LMH+*SR*+β+W_C	**56.44**	76.00	**43.77**	49.67

Table 5: Results of ablation study on VQA-CP v2.

Ablation study. Tab. 5 investigates the effect of each component of MMBS, i.e., the backbone models, the positive-sample construction module (***SR***) and the unbiased sample selection module (β) which includes the correction factor W_C. We find that: 1) +*SR* constantly outperforms the base models significantly, especially on the yesno questions where the language biases tend to exist. We also conduct experiments for further validation of the effectiveness of the ***SR*** strategy in **App.** D.3. 2) Comparing the performance of +*SR* and +*SR*+β, we can find that the unbiased sample selection module always benefits MMBS. This attests to the intuition that we do not need to construct the positive samples for the unbiased samples. 3) The correction factor W_C consistently has a positive impact on the model performance. This further demonstrates that dynamically adjusting the unbiased sample proportion for each question category is a useful strategy.

4.5 Qualitative Analysis on the Effectiveness of MBSS

Visualization of the answers' distribution. To better understand the effectiveness of MBSS, we compare the distribution of the predicted answers by three methods, i.e., UpDn, LMH and LMH+MMBS, and the real answer distribution of the training and test sets of VQA-CP v2 (left) and VQA v2 (right) in Fig. 5. From the left part, we

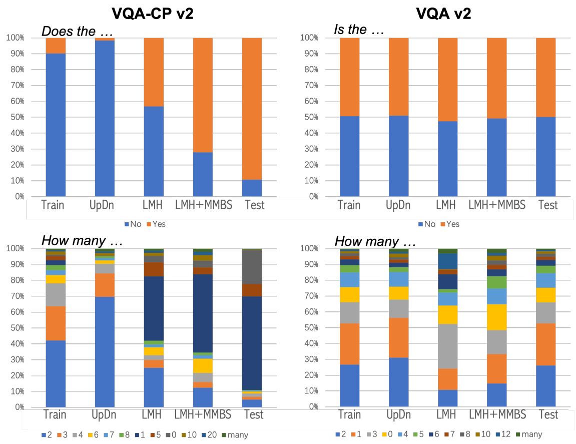

Figure 5: The answer distribution of the training sets, test sets, and three methods.

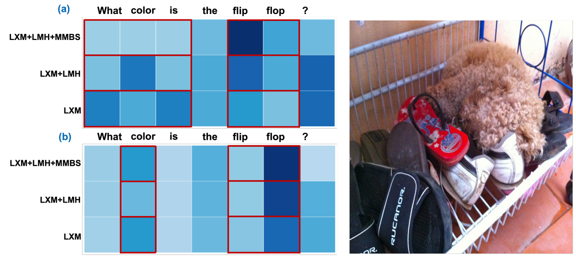

Figure 6: (a) The attention graph of the last cross-attention of cross-modality encoder, which averages the attention of all visual regions to each question word. (b) The attention graph of the last self-attention layer of the language encoder.

find that UpDn tends to output the most frequent answers of training set, which demonstrates that it overfits the training priors. In comparison, LMH alleviates the domination of the biased answers and MBSS further mitigates the impact training priors, resulting in answer distributions that are closet to the test set. This explains why MBSS generalizes the best to the OOD VQA-CP v2 test set.

From the upper right plot, we see that for the relatively easy yesno question 'Is the', when the training set is balanced in answer distribution, the three methods can also produce balanced answer distributions similar to the test set. For the question type 'How many' on VQA v2, the most frequent answers in the training set, i.e., '2' and '1', account for much smaller proportion in the answer distribution of LMH. This is because that LMH diminishes the training signal from biased samples. Consequently, LMH performs worse on VQA v2 where most questions can be correctly answered by the common answers. By contrast, our method exploits the biased samples using contrastive learning rather than undermining them like LMH, and thus MBSS recovers the answers' distribution of ID test set.

Attention graph of question words. The attention graphs of LXM+LMH+MMBS, LXM+LMH and LXM are shown in Fig 6. As highlighted in the red boxes, we focus on the question category words, i.e., 'What color is' or 'color', and the subject words, i.e., 'flip flop'. We observe that: 1) For the cross-modality encoder (a) that extracts higher level representation for classification, LXM pays low attention to the subject words and high attention to the question category words, which is the source of language bias. In comparison, the introduction of LMH alleviates this problem and MBSS further shifts the attention to the subject words, which contain less biased information and have more specific visual groundings. 2) For the question encoder (b) that summarizes information from the textual domain, LXM+LMH pays less attention to the question category word 'color', as compared with the other two methods. We conjecture that this can partly explain the poor performance of LMH on the ID dataset that contains strong language priors, because the word 'color' is essential to the meaning of the question. LXM pays more attention to 'color' but relatively less attention to the subject words. By contrast, our method assigns sufficient attention to both the question category and subject words, which can produces a better question representation.

5 Conclusion

In this paper, we propose a novel contrastive learning method to ameliorate the ID-OOD trade-off problem faced by most existing debaising methods for VQA models. Instead of undermining the importance of the biased samples, our method makes the most of them via contrastive learning. Considering the characteristics of language priors, we design the positive samples which eliminate the biased information. On this basis, we investigate several strategies to use the positive samples and design an algorithm that treat biased and unbiased samples differently in contrastive learning. The proposal is compatible with multiple backbone models and debiasing methods, and achieves competitive performance on OOD VQA-CP v2 while maintaining the performance on ID VQA v2. Meanwhile, our approach provides insights on how to avert the trade-off between in-distribution and out-of-distribution performance.

References

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

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

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

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

Yonatan Belinkov, Adam Poliak, Stuart M Shieber, Benjamin Van Durme, and Alexander M Rush. 2019. Don't take the premise for granted: Mitigating artifacts in natural language inference. *arXiv preprint arXiv:1907.04380*.

Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. 2019. Rubi: Reducing unimodal biases for visual question answering. *Advances in neural information processing systems*, 32:841–852.

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

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. *arXiv preprint arXiv:1909.03683*.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2020. Learning to model and ignore dataset bias with mixed capacity ensembles. *arXiv preprint arXiv:2011.03856*.

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

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

Gabriel Grand and Yonatan Belinkov. 2019. Adversarial regularization for visual question answering: Strengths, shortcomings, and side effects. *NAACL HLT 2019*, page 1.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738.

C. Jing, Y. Wu, X. Zhang, Y. Jia, and Q. Wu. 2020a. Overcoming language priors in vqa via decomposed linguistic representations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(7):11181–11188.

Chenchen Jing, Yuwei Wu, Xiaoxun Zhang, Yunde Jia, and Qi Wu. 2020b. Overcoming language priors in vqa via decomposed linguistic representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11181–11188.

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

Corentin Kervadec, Grigory Antipov, Moez Baccouche, and Christian Wolf. 2021. Roses are red, violets are blue... but should vqa expect them to? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2776–2785.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks. *Advances in Neural Information Processing Systems*, 31.

Mingrui Lao, Yanming Guo, Yu Liu, Wei Chen, Nan Pu, and Michael S Lew. 2021. From superficial to deep: Language bias driven curriculum learning for visual question answering. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3370–3379.

Zujie Liang, Haifeng Hu, and Jiaying Zhu. 2021. Lpf: A language-prior feedback objective function for debiased visual question answering. *arXiv preprint arXiv:2105.14300*.

Zujie Liang, Weitao Jiang, Haifeng Hu, and Jiaying Zhu. 2020. 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.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Rabeeh Karimi Mahabadi and James Henderson. 2019. Simple but effective techniques to reduce biases. *arXiv preprint arXiv:1909.06321*, 9.

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

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 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. *arXiv preprint arXiv:1810.03649*.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2020. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*.

Victor Sanh, Thomas Wolf, Yonatan Belinkov, and Alexander M Rush. 2020. Learning from others' mistakes: Avoiding dataset biases without modeling them. *arXiv preprint arXiv:2012.01300*.

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

Robik Shrestha, Kushal Kafle, and Christopher Kanan. 2020. 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.

Qingyi Si, Zheng Lin, Mingyu Zheng, Peng Fu, and Weiping Wang. 2021. Check it again: Progressive visual question answering via visual entailment. *arXiv preprint arXiv:2106.04605*.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

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

split	images	questions	answers
train	121K	438K	4.4M
test	98K	220K	2.2M

Table 6: The statistics about VQA-CP v2.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020. Towards debiasing nlu models from unknown biases. *arXiv preprint arXiv:2009.12303*.

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

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29.

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. *arXiv preprint arXiv:2012.11528*.

A More Related Work

A.1 Language biases in VQA datasets.

To develop the robustness and generalization capability of VQA models, Agrawal et al. (2018) carefully constructed the VQA-CP v2 dataset by re-organizing the training and val sets of VQA v2[3] dataset to introduce the distribution shift of question categories. VQA-CP v2 is an out-of-distribution dataset where the answer distributions of a same question category are different in the training set and test set while VQA v2 dataset is in-distribution. The performance of VQA models relying on the biases often drops significantly on VQA-CP v2 dataset, which has become the standard benchmark for evaluating the capability of overcoming the language priors. The data statistics of VQA-CP v2 are shown in Tab. 6. The images of both datasets are from COCO (Lin et al., 2014), and the questions are in English. Recently, another OOD benchmark GQA-OOD (Kervadec et al., 2021) was proposed. However, it is built on the GQA dataset, which is generated automatically and is not 'in the wild'.

Although VQA-CP v2 has become the current OOD benchmark in the VQA community, it introduces only one specific type of controlled distributional shift, i.e., question category, and thus its

[3]Both VQA-CP v2 and VQA v2 datasets are licensed under *Commons Attribution 4.0 International License*.

Type	original	***Shuffle***	***Removal***
Y/N	Is this indoors or outside ?	Is ? indoors outside or this	indoors or outside ?
Y/N	Are these buildings new ?	new these buildings ? Are	buildings new ?
Y/N	Does this person eat healthily ?	this ? person healthily eat Does	person eat healthily ?
Num	How many people will be dining ?	? be many people How will dining	people will be dining ?
Num	How many small zebra are there ?	there zebra small ? are How many	small zebra are there ?
Other	What is the smallest kid holding ?	the is smallest What ? holding kid	smallest kid holding ?
Other	Who is on the screen ?	Who screen ? the is on	on the screen ?
Other	What are people wearing on their heads ?	their are wearing ? on people heads What	people wearing on their heads ?
Other	What animals are walking on the road ?	road the are on What animals ? walking	animals are walking on the road ?
Other	What color is the food inside the bowl ?	the color the food What is bowl inside ?	food inside the bowl ?

Table 7: More examples of two types of positive samples.

Type	n(C_{qtype})	m(Z_C)	m(W_C)%	m(P_C)%	m(Z_C^{unb})
Y/N	28	209	92.60	18.52	39
Num	4	156	56.84	11.37	19
Other	33	836	3.76	0.75	10

Table 8: The statistics about the question type (e.g., Y/N) and the corresponding unbiased samples with the setting of β=20%. For all question categories (e.g, what color) in each question type, (C_{qtype}) represents the number of them; m(Z_C) represents the mean value of their label space size; m(W_C) represents the mean value of their correction factors which are used to dynamically adjust β; m(P_C) represents the mean value of their unbiased answer proportions after being adjusted; m(Z_C^{unb}) represents the mean value of their unbiased answer number.

OOD settings can only evaluate the model's reasoning ability beyond the single type of biases, rather than the true robustness beyond multiple types of biases. Besides, most of recent works (Cadene et al., 2019; Clark et al., 2019; Grand and Belinkov, 2019; Jing et al., 2020b; Ramakrishnan et al., 2018; Si et al., 2021; Wu and Mooney, 2019), including ours, design the de-biasing methods specifically for the known biases (i.e., language priors) and known construction of OOD splits of VQA-CP v2 (i.e., the handcrafted inverse shifts of answer distribution between test and training sets). Therefore, once the bias is unknown, or the training and test sets do not conform to such construction procedure, these models may fail to generalize. Admittedly, more OOD datasets with unknown biases (which is well-developed in NLU (Clark et al., 2020; Sanh et al., 2020; Utama et al., 2020)) and multiple types of distribution shifts are needed to promote the de-biasing research.

A.2 Visual grounding methods.

Apart from these two kinds of effective methods, another line is the visual-grounding method (Selvaraju et al., 2019; Wu and Mooney, 2019). They use extra human visual annotations to force model to answer according to the right reason, i.e., relevant image regions. However, Shrestha et al. (2020) finds that their improvements is caused by a regularization effect which hinders over-fitting to the training priors, rather than the better visual grounding. Besides, the visual-grounding methods perform much worse than the other two kinds of methods, so we do not provide discussion about them in the main paper due to the space limitation.

B More Details of the Proposed Method

B.1 Definition of target score tgt.

The tgt_i is the soft target score of each answer for the i_{th} sample. Following (Zhu et al., 2020), it is annotated by the annotators for the dataset, and obtained by:

$$tgt_i = \frac{votes}{|A_i|} \tag{6}$$

where $|A_i|$ is the number of valid answers in total for the i_{th} sample, and $votes$ denotes that, for each answer in A_i, how many annotators annotated it as ground truth for the given i_{th} question.

B.2 Discussion about the positive samples.

We give more examples of ***Shuffling*** and ***Removal*** positive questions in Tab. 7. We can see that the intention of the 'Y/N' questions can still be inferred from the ***Removal*** questions. By contrast, the intention of the ***Removal*** questions for non-'Y/N' questions is ambiguous. This attests to the rationality of the proposed ***SR*** strategy, which treats 'Y/N' and non-'Y/N' questions differently.

Although the positive samples could cause some confusion/ambiguity, it may not impact our method too much, because: 1) In MBSS, the model only makes prediction on the original samples during training, and thus it does not directly associate the answers with the positive questions, which are only

			VQA-CP v2 test					VQA v2 val					Gaps
Methods	Base.	Venue	All	Y/N	Num	Other	Gap ↑	All	Y/N	Num	Other	Gap ↑	Sum ↑
Plain Models follow:													
SAN (Yang et al., 2016)	-	CVPR	24.96	38.35	11.14	21.74	-	52.41	70.06	39.28	47.84	-	-
GVQA (Agrawal et al., 2018)	-	CVPR	31.30	57.99	13.68	22.14	-	48.24	72.03	31.17	34.65	-	-
BAN (Kim et al., 2018)	-	NeurIPS	37.03	41.55	12.43	41.4	-	63.9	81.42	45.18	55.54	-	
UpDn (Anderson et al., 2018)	-	CVPR	39.74	42.27	11.93	46.05	-	63.48	**81.18**	42.14	55.66	-	-
LXM (Tan and Bansal, 2019)	-	EMNLP	47.19	50.55	24.06	51.77	-	**71.01**	88.24	54.07	62.39	-	-
MMBS	BAN	**Ours**	**47.63**	**66.18**	**16.36**	**46.49**	**+10.60**	**64.78**	**82.03**	**46.48**	**56.51**	**+0.88**	**+11.48**
MMBS	UpDn	**Ours**	**48.19**	**65.00**	**14.05**	**48.75**	**+8.45**	**63.84**	79.61	**44.23**	**57.05**	**+0.36**	**+8.81**
MMBS	LXM	**Ours**	**56.51**	**79.83**	**28.70**	**51.92**	**+9.32**	70.85	**88.25**	**55.67**	61.63	**-0.16**	**+9.16**
Visual-grounding Methods follow:													
AttAlign(Selvaraju et al., 2019)	UpDn	ICCV	39.37	43.02	**11.89**	45.00	-0.37	63.24	80.99	42.55	55.22	-0.24	-0.61
HINT (Selvaraju et al., 2019)	UpDn	ICCV	46.73	67.27	10.61	45.88	+6.99	**63.38**	**81.18**	**42.99**	**55.56**	**-0.1**	+6.89
SCR (Wu and Mooney, 2019)	UpDn	NeurIPS	**49.45**	**72.36**	10.93	**48.02**	**+9.71**	62.20	78.80	41.60	54.50	-1.28	**+8.43**
Ensemble-based Methods follow:													
AdvReg. (Ramakrishnan et al., 2018)	UpDn	NeurIPS	41.17	65.49	15.48	35.48	+1.43	62.75	79.84	42.35	55.16	-0.73	+0.70
GRL (Grand and Belinkov, 2019)	UpDn	NAACL	42.33	59.74	14.78	40.76	+2.59	51.92	-	-	-	-11.56	-9.00
RUBi (Cadene et al., 2019)	UpDn	NeurIPS	44.23	67.05	17.48	39.61	+4.49	61.16	-	-	-	-	-2.32
DLR (Jing et al., 2020a)	UpDn	AAAI	48.87	70.99	18.72	45.57	+9.13	57.96	76.82	39.33	48.54	-5.52	+3.61
LMH (Clark et al., 2019)	UpDn	EMNLP	52.01	72.58	31.12	46.97	+12.27	56.35	65.06	37.63	54.69	-7.13	+5.14
CF-VQA (Niu et al., 2021)	UpDn	CVPR	53.55	**91.15**	13.03	44.97	+13.81	**63.54**	**82.51**	**43.96**	54.30	**+0.06**	+13.87
LPF (Liang et al., 2021)	UpDn	SIGIR	55.34	88.61	23.78	46.57	+15.60	55.01	64.87	37.45	52.08	-8.47	+10.13
LMH+MMBS	UpDn	**Ours**	**56.44**	76.00	**43.77**	**49.67**	**+16.70**	61.87	75.86	40.34	**56.95**	-1.61	**+15.09**
LMH(Clark et al., 2019)	LXM	EMNLP	63.34	78.28	65.95	54.79	+16.15	69.49	84.16	54.07	62.41	-1.52	+14.63
UpDn-SAR (Si et al., 2021)	LXM	ACL	61.71	78.69	47.67	56.66	+14.52	-	-	-	-	-	-
UpDn-SAR+LMH (Si et al., 2021)	LXM	ACL	64.98	81.89	59.65	57.61	+17.79	69.17	88.08	51.04	59.58	-1.84	+15.95
LMH+MMBS	LXM	**Ours**	65.70	81.70	61.24	58.54	+18.51	**70.29**	84.14	**55.20**	**63.75**	**-0.72**	+17.79
UpDn-SAR+LMH+MMBS	LXM	**Ours**	**68.01**	**86.55**	**64.69**	**59.21**	**+20.82**	69.29	**88.31**	50.81	59.65	-1.72	**+19.10**
Data-balancing Methods follow:													
SSL (Zhu et al., 2020)	UpDn	IJCAI	57.59	86.53	29.87	50.03	+17.85	63.73	-	-	-	+0.25	+18.10
CSS (Chen et al., 2020)	UpDn	CVPR	41.16	43.96	12.78	47.48	+1.42	-	-	-	-	-	-
LMH+CSS (Chen et al., 2020)	UpDn	CVPR	58.95	84.37	49.42	48.21	+19.21	59.91	73.25	39.77	55.11	-3.57	+15.64
LMH+CSS+CL (Liang et al., 2020)	UpDn	EMNLP	59.18	86.99	49.89	47.16	+19.44	57.29	67.27	38.40	54.71	-6.19	+13.25
LBCL (Lao et al., 2021)	UpDn	MM	60.74	88.28	45.77	50.14	+21.00	-	-	-	-	-	-
MUTANT (Gokhale et al., 2020)	UpDn	EMNLP	61.72	88.90	49.68	50.78	+21.98	62.56	82.07	42.52	53.28	-0.92	+21.06
SSL-SAR+LMH (Si et al., 2021)	LXM	ACL	66.73	86.00	62.34	57.84	+19.54	69.22	**87.46**	**51.20**	60.12	-1.79	+17.75
MUTANT (Gokhale et al., 2020)	LXM	EMNLP	69.52	93.15	67.17	57.78	+22.33	70.24	89.01	54.21	59.96	-0.77	+21.56
SSL-SAR+LMH+MMBS	LXM	**Ours**	**68.39**	**87.30**	**65.21**	**59.36**	**+21.20**	**69.43**	87.39	50.37	**60.82**	**-1.58**	**+19.62**

Table 9: Results on VQA-CP v2 test and VQA-v2 validation set. For each base model, i.e., UpDn and LXM, the best scores are bold. 'Gap' denotes the accuracy improvement of language-prior methods over their base models. * indicates our reimplementation. We report MMBS results with ***SR*** strategy here.

used in contrastive learning. 2) ***Shuffling*** could change the original questions to a conflicting meanings, e.g., , 'How many bananas are next to the apples?' and 'How many apples are next to the bananas?'. However, such special cases are very rare. For a question whose length is 7[4], the probability of shuffling to a conflicting meaning is $\frac{1}{7!}$. In most cases, the ***Shuffling*** just eliminates the sequential information of the questions, but basically conveys the same meaning. 3) In terms of ***Removal***, we only construct this kind of positive questions for the 'Y/N' questions, which does not change the intended meaning of the original question as discussed in the above paragraph. 4) Additionally, the proposed unbiased sample selection module prevents the potential noise in positive questions from affecting the unbiased samples, which are beneficial to OOD generalization.

B.3 Unbiased sample statistics.

To further investigate how the unbiased-sample-selection algorithm treats different types of questions , i.e. 'Y/N', 'Num' and 'Other' questions, we roughly divide all the question categories into the three types according their semantics, and then do some statistical analysis about the question types and the corresponding unbiased samples. We set the initial unbiased answer proportion (hyper-parameter) β = 20%. As the detail statistics shown in Tab. 8, we find that: 1) the 'Other' questions have the largest answer space while the 'Num' questions have the smallest one. Counter-intuitively, the 'Y/N' questions also have a relatively large number of candidate answers. For example, 'red' is also annotated as the answer to the question 'Is this flower red?'. However, this rarely happens compared with the answer 'yes'. 2) The proposed correction factor W_C is close to 1 when the question is a 'Y/N' question and the W_C is close to 0 when the question is a 'Other' question. Correspondingly, the adjusted unbiased answer proportion P_C is close to β for 'Y/N' questions while it is relative smaller for 'Other' questions. This is consistent with the phenomenon that most ground truth of 'Y/N' questions concentrate on much fewer answers (e.g., 'Yes') than that of 'Other' questions.

[4]The average length of questions in the training set is 7.14

Model	Epo	α	β	Lr	N'
BAN+Ours	25	1	0.5	1e-4	-
UpDn+Ours	60	1	0.6	1e-4	-
LXM+Ours	40	1	0.2	5e-6/5e-5	-
LMH+Ours	60	0.18	0.5	1e-4	-
LXM+LMH+Ours	40	0.18	0.2	5e-6/5e-5	-
U-SAR+Ours	10	0.18	0.5	1e-5	2,20 / 2,2
SAR+Ours	10	0.18	0.5	1e-5	2,20/ 2,20

Table 10: The detailed hyper-parameter settings of our methods. The Epo represents the number of training epochs. Lr represents the initial learning rate of Adam optimizer on VQA-CP v2/VQA v2. N', is a SAR-specific hyper-parameter, represents the number of candidate answers for yesno, non-yesno questions during test on VQA-CP v2/VQA v2.

Model	Param.	Training Time	Infrastructure
UpDn+Ours	36M	0.38h/epo	TITAN RTX 24GB GPU
LXM+Ours	213M	1.73h/epo	2 x TITAN RTX 24GB GPUs

Table 11: The details of computational experiments of our methods based on UpDn and LXM.

C More Experimental Setups

C.1 Implementation details.

Following existing works, we use the Faster R-CNN (Ren et al., 2015) to extract fixed 36 objects feature embeddings with 2048 dimensions for each image. All the questions are trimmed or padded to 14 words. For the UpDn backbone model, we apply a single-layer GRU to encode the word embeddings(initialized with Glove (Pennington et al., 2014)) of the question into a 1280-dimensional question embeddings. We follow (Zhu et al., 2020) and adopt a multi-step learning rate that halves every 5 epochs after 10 epochs. For the LXMERT backbone, we use the tokenizer of LXMERT to segment each input question into words. We adopt the cosine learning rate decay following the warmup in the first 5 epochs. We train the models with batch size of 128. The detailed hyper-parameter settings of our methods in the main results are shown in Tab. 10. The details of computational experiments of our method based on UpDn and LXMERT are shown in Tab. 11. We keep the same random seed during training and testing for ***Shuffling*** method. As the change of seed has little effect on each method, following most of previous works, we also report the results with a single run.

Method	Form	***S***	***R***	***B***	***SR***
UpDn	***original***	42.20	42.38	42.69	42.80
	Shuffling	**42.26**	33.68	**44.37**	**48.19**
	Removal	26.15	**42.83**	43.19	22.67
LMH	***original***	**55.89**	**55.87**	**55.62**	**56.44**
	Shuffling	54.14	39.93	52.3	52.64
	Removal	31.46	49.4	47.48	32.43

Table 12: Results of UpDn+MMBS and LMH+MMBS with three question forms at test on VQA-CP v2. ***S***, ***R***, ***B*** and ***SR*** are the four strategies to use positive sample in training.

C.2 Positive sample construction for SAR.

SAR (Si et al., 2021) is a two-stage framework: it first selects the most relevant candidate answers, and then combines the question and each candidate answer to produce *dense captions*, and finally, reranks the dense captions based on visual entailment. They design two ways to construct the dense captions, including 1) replacing the question category prefix with answer and 2) concatenating question and answer directly. To apply MMBS to SAR, we construct the positive dense captions for the rerank stage. Specifically, we directly use the first kind of captions as ***S*** positive captions, because the question category prefix has already been removed. For the second kind of captions, we randomly shuffle the words to construct the ***R*** positive captions. The input dense caption during training and test are the second kind of captions. Following Si et al. (2021), we set the number of candidate answers for training to 20. During test, we set the number of the candidate answers to N' shown in Tab. 10.

D More Experiments and Analysis

D.1 Performance with different question forms at test.

After contrastive learning using the positive questions, the models trained with MMBS can also take the positive question as input in the inference phase, while normal models cannot. For more comprehensive analysis, we report the results of three question forms here. Because the annotation of question categories should not be available at test, the ***Removal*** questions are not used in the other experiments. From the results shown in Tab. 12, we find that: 1) For UpDn with the ***S***, ***B*** and ***SR*** strategies (which involve the ***Shuffling*** positive sample), the performance is the best when the test question is in the ***Shuffling*** form. This shows that the ***Shuffling*** form input question, when used in the test stage, may further prevent the model from relying on the superficial correlations. 2) For LMH, when the input

Method	All	Y/N	Num	Other
UpDn	41.06	43.13	13.71	47.48
UpDn+*orig.*	41.39	42.23	13.7	48.54
UpDn+*rand-SR*	44.21	51.19	**15.05**	48.56
UpDn+*SR*	**47.62**	**62.72**	13.92	**48.95**
LXM	47.19	50.55	24.06	51.77
LXM+*orig.*	48.14	51.25	25.63	**52.69**
LXM+*rand-SR*	51.07	62.22	**29.68**	51.09
LXM+*SR*	**55.26**	**77.13**	27.33	51.47
LMH	52.01	72.58	31.12	46.97
LMH+*orig.*	55.25	74.84	**41.11**	48.87
LMH+*rand-SR*	**55.50**	75.36	35.67	**50.54**
LMH+*SR*	55.41	**76.50**	37.20	49.35

Table 13: Results on VQA-CP v2 for validating the effectiveness of ***SR*** strategy. The models here do not contain the unbiased sample selection module.

question during test is ***original***, the models always perform the best. This is probably because the LMH+MMBS method is robust enough and will not be easily biased by the superficial correlations in the ***original*** questions. On the in-distribution settings, all the models obtain the best performance on VQA v2 when the test questions are in the ***original*** form.

D.2 Full results.

For detailed comparison, we list the complete results of all mentioned baselines and the proposed methods in Tab. 9.

D.3 Further validation of the effectiveness of ***SR*** strategy.

To better validate the effectiveness of ***SR*** strategy, we also evaluate the model performance directly using the original sample as positive sample (+*orig.*), or randomly adopting one of ***S*** and ***R*** as positive sample (+*rand-SR*) for each sample. We can observe from Tab. 13 that: 1) +*orig.* constantly outperforms the backbone models because the contrastive learning itself is helpful for learning a better feature representation. 2) It is worth noting that when we apply +*orig.* on LMH, the performance improvement is much more obvious. This is because ensemble-based methods have relieved the language priors to some extent at the cost of almost entirely attenuating the positive information from the biased samples. Our method makes up for this drawback and forces the model to pay attention again to this information by minimizing contrastive learning loss which does not cause superficial correlations, unlike the normal VQA loss. This can also explain that the performance of +*orig.*, +*rand-SR* and +*SR* is similar based on the ensemble-based methods. 3) For UpDn and LXM: a) +*rand-SR* outperforms +*orig.* considerably, which demonstrates that the design of positive samples by excluding the correlations between the question category and answer benefits MMBS in overcoming language priors; b) Compared with +*rand-SR*, +*SR* achieves prominent performance boost on 'Y/N' questions, and slightly improves the performance or maintains competitive performance on the other two types of questions, which attests to the soundness of the motivation of strategy ***SR*** (refer to Sec. 3.2);

Method	All	Y/N	Num	Other
LXM+LMH	63.34	78.28	65.95	54.79
LXM+LMH+*orig.*	65.56	81.35	61.17	58.50
LXM+LMH+*rand-SR*	65.22	78.55	**62.13**	**59.09**
LXM+LMH+*SR*	65.66	81.39	60.69	58.78
LXM+LMH+*SR*+β	65.63	81.64	60.62	58.72
LXM+LMH+*SR*+β+W_C	**65.70**	**81.70**	61.24	58.54
LXM+LMH+*S*+β+W_C	64.98	79.67	59.90	58.68
LXM+LMH+*R*+β+W_C	65.34	78.37	**62.63**	**59.25**
LXM+LMH+*B*+β+W_C	65.03	78.38	61.05	59.14
LXM+LMH+*SR*+β+W_C	**65.70**	**81.70**	61.24	58.54

Table 14: Upper: Ablation study of MMBS based on LXM+LMH. Lower: Comparison of different positive-sample construction strategies based on LXM+LMH.

D.4 Ablation study of MMBS based on LXM+LMH.

From Tab. 14, we find that the phenomenon of the ablation study based on LXM+LMH is in line with the that of LMH in Tab. 5, Tab. 13 and Tab. 4.