Appendix

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

1 More Details of the Proposed Method

1.1 Discussion about the positive samples.

We give more examples of ***Shuffling*** and ***Removal*** positive questions in Tab. 1. 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 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[1], 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.

1.2 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) $\beta = 20\%$. As the detail statistics shown in Tab. 2, 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.

2 More Experimental Setups

2.1 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

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

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 1: 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 2: 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.

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 3: 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.

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. 3. The details of computational experiments of our method based on UpDn and LXMERT are shown in Tab. 4. 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.

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

2.2 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. 3.

3 More Experiments and Analysis

3.1 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. 5 that: 1) *+orig.* constantly out-

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**	**3**7.20	49.35

Table 5: Results on VQA-CP v2 for validating the effectiveness of ***SR*** strategy. The models here do not contain the unbiased sample selection module.

performs 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***.

References

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.

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.

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*.

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*.