

Supplementary Material

Combating Visual Question Answering Hallucinations via Robust Multi-Space Co-Debias Learning

Anonymous Authors

1 DATASETS AND BENCHMARKS

The hallucination problem within the robust VQA task is challenging, and models that cleverly exploit biases or shortcuts may generate predictions that influence human decisions. Therefore, we select the VQA-CP v1 and VQA-CP v2 [1] datasets as benchmarks to evaluate the performance under artificially changed prior conditions. Additionally, to prove that our performance improvement comes from reducing hallucinations, we also evaluate our model on the VQA v2 datasets [11].

VQA-CP v2 contains three types of ~658K questions, based on ~122K Microsoft COCO images, with the same dataset splits. Following previous work, answers that appeared more than 9 times in the training set were selected as candidate answers, resulting in 3129 candidate answers. Notably, the answers to the *train* set and *test* set are inversely distributed, making it convincing to evaluate the model against the hallucination problem.

VQA-CP v1 is developed from the VQA v1 dataset and contains ~122K images and ~370K questions. Similarly, it serves as a benchmark to evaluate the robustness of the VQA model under artificially changed prior conditions.

VQA v2 is the most commonly used VQA balanced benchmark dataset, with its images identical to those in VQA-CP v2. The distribution difference between the divided *train* set and *test* set is significantly reduced. Used as a benchmark to demonstrate whether model improvements come from reducing hallucinations.

2 BASELINES AND EVALUATION METRIC

We refine the experimental section in the main paper and provide more baseline analysis. A summary of different types of baselines is as follows;

- General VQA frameworks, including UpDn, S-MRL. These models have become the basic architecture for VQA models, and we follow the UpDn framework.
- Modify vision-language models, including VGQE, DLR.
- Balancing data methods can be categorized into those with additional annotations (CSS, CSS + CL, CSS + IntroD, ECD, and Mutant) and those without annotations (SSL-VQA, D-VQA, KDDAug, and DDG). Following the previous setup [4, 8, 15, 21, 23, 24], we do not make direct comparisons with methods that add annotations to balance the dataset.
- Biases mitigation methods, such as AdvReg, RUBi, LMH, GGE-iter, AdaVQA, COB, PWVQA, CVIV, GENB, GGD, and RMLVQA, are specifically designed to enhance VQA robustness by addressing biases. These methods are particularly relevant to our work.

For evaluation on all datasets, all experiments use the standard VQA evaluation metric [3]. When provided with an image and a corresponding question, the accuracy of a predicted answer \mathcal{A} is

Table 1: Model performance on the VQA-CP v2 test set. In comparison with baselines, MSCD based on the multi-space co-debias learning paradigm shows the best performance. The * represents a reproduction that follows the official code.

Datasets	VQA-CP v2			
Methods	Overall-CP	Y/N-CP	Num-CP	Others-CP
(I) General VQA Frameworks				
UpDn [2]	39.74	42.27	11.93	46.05
S-MRL [5]	38.46	42.85	12.81	43.20
(II) Modify Vision-Language Models				
VGQE [18]	50.11	66.35	27.08	46.77
DLR [16]	48.87	70.99	18.72	45.57
(III) Biases Mitigation Methods				
UpDn [2]	39.74	42.27	11.93	46.05
AdvReg [22]	41.17	65.49	15.48	35.48
RUBi [5]	47.11	68.65	20.28	43.18
LMH [9]	52.15	70.29	44.10	44.86
GGE-iter [13]	57.12	87.35	26.16	49.77
AdaVQA* [12]	55.76	72.47	53.81	45.58
COB [15]	57.53	88.36	28.81	49.27
PWVQA [24]	59.06	88.26	52.89	45.45
GENB* [8]	59.10	87.96	40.03	49.21
GGD [14]	59.37	88.23	38.11	49.82
CVIV [21]	60.08	88.85	40.77	50.30
RMLVQA* [4]	60.42	89.86	47.93	49.72
MSCD	62.26	88.03	55.50	50.45
(IV) Balancing Data Methods				
CSS [6]	58.95	84.37	49.42	48.21
CSS+CL [19]	59.18	86.99	49.89	48.21
CSS+ IntroD[20]	60.17	89.17	46.91	48.62
ECD [17]	59.92	83.23	52.29	49.71
MUTANT [10]	61.72	88.90	49.68	50.78
SSL-VQA [27]	57.59	86.53	29.87	50.03
D-VQA* [26]	60.50	89.61	48.49	48.53
KDDAug [7]	60.24	86.13	52.29	49.71
DDG [25]	61.14	88.77	49.33	49.90

computed as follows:

$$Acc(\mathcal{A}) = \min \left(1, \frac{\# \text{humans that provide Ans}}{3} \right). \quad (1)$$

3 MORE EXPERIMENTAL ANALYSIS

To further validate the effectiveness of our robust learning paradigm in mitigating hallucinations, we compare MSCD with the latest methods outlined in Table 1. These methods encompass General VQA Frameworks, Modify Vision-Language Models, Biases Mitigation Methods, and Balancing Data Methods. To reduce the impact of equipment errors, we re-implemented AdaVQA [12], Genb [8], RMLVQA [4], and D-VQA [26] according to the official code.

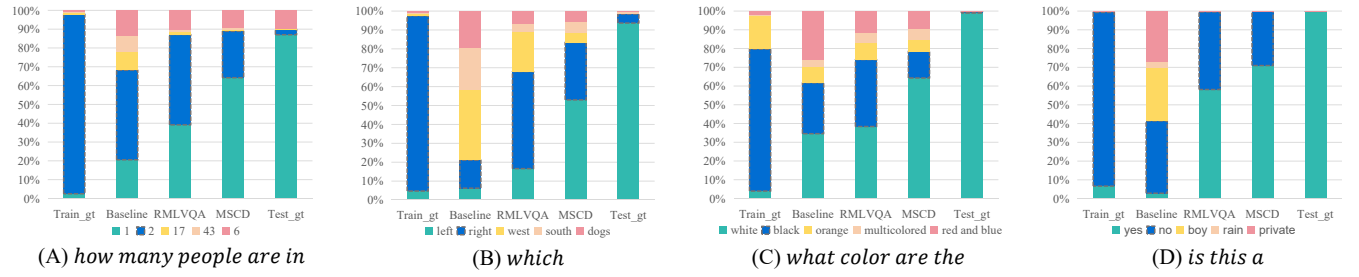
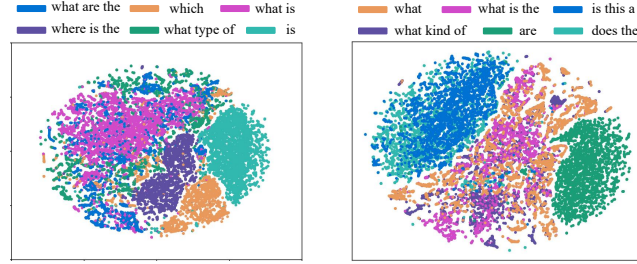


Figure 1: Illustrations of combating VQA hallucinations for more question types. We plot more types of hallucination distribution legends to demonstrate the effect of combating hallucinations, and MSCD still maintains continuous improvement.

Table 2: Performance on VQA-CP v2 test set with various proportions of training data. The MSCD learning method can achieve excellent performance with low data volume.

Methods	Proportion of Training Set				
	20%	40%	60%	80%	100%
Updn	36.22	38.90	39.40	40.61	41.53
Genb	48.08	53.98	56.22	57.78	59.10
Adavqa	49.54	52.95	53.67	54.88	55.76
RMLVQA	54.77	57.84	58.95	59.72	60.42
MSCD	56.79	59.67	61.12	61.86	62.26
SSL-VQA	52.71	54.42	56.83	57.31	57.59
D-VQA	52.76	56.67	58.39	59.37	60.50



(A) TSNE visualization of MSCD with different question types (Left) and (Right)

Figure 2: T-SNE Visualization of the discriminative space within different question types on the VQA-CP v2 test set.

From the comparative analysis, we draw the following conclusions: 1) Our proposed multi-space co-debias paradigm outperforms all other methods, achieving state-of-the-art performance. 2) Compared with the latest methods, such as RMLVQA (based on feature learning), GenB (based on ensemble models) and CVIV (based on causal inference), we have achieved at least 1.85% improvement, which is an encouraging performance for the multi-space co-debias paradigm. 3) Even compared with balancing data methods, which often involve adding artificial annotations or entail cost-free data, the MSCD model stands out. These methods primarily rely on data balancing rather than addressing hallucinations inherent in the data patterns. While the MSCD model incorporates cost-free examples during the counterexample learning phase, its main purpose is to

enhance the prior independence of Spherical debias learning rather than just balancing the data set.

4 DATA DEPENDENCY EXPERIMENT

As shown in Table 2, our method performs better when data is limited, indicating that the MSCD is less dependent on data and can focus on instance semantic reasoning with fewer samples. These results further demonstrate the effectiveness of our MSCD. Benefiting from the robust training process based on multi-space co-debias, our MSCD outperforms all comparison methods on the VQA-CP v2 dataset. Specifically, in the 60% data scenario, our MSCD outperforms other state-of-the-art models by at least 0.7% when compared to their performance at full data volume. This indicates that MSCD holds promise for extending to low-data tasks, achieving superior generalization performance with smaller data sizes and faster training speeds. Simultaneously, it surpasses balanced dataset methods such as SSL-VQA and D-VQA by approximately 4.67% and 1.76%, respectively. This highlights that MSCD isn't reliant on data and addresses inherent shifts phenomena through multi-space collaboration to enhance semantic reasoning capabilities.

5 MORE ANSWERS DISTRIBUTION

The Fig. 1 displays the answer distributions for more question types in VQA-CP v2, encompassing the *train* set, *test* set, and prediction distributions resulting from three methods. In fact, the answer distribution between the *train* set and the *test* set almost exhibits an inverse distribution. This disparity significantly undermines the robustness of the models and underscores the formidable challenge of mitigating hallucinations. Due to its simple design lacking debias constraints, the baseline model produces many incorrect predictions. Even state-of-the-art RMLVQA models struggle to effectively mitigate biases present in the training data, as evidenced by the persistence of numerous unreasonable distributions in the test set. Notably, MSCD employs multi-space co-debias learning to generate predictions that align more closely with real semantics, effectively combating hallucinations. It's apparent from the legends that for classes appearing only a few times in the *train* set, MSCD can produce accurate predictions through semantic reasoning during the testing process. Similarly, MSCD's answer distribution closely mirrors the ground-truth answers from the *test* set, indicating that MSCD doesn't rely on priors and biases in the *train* set, but genuinely comprehends instance semantics to mitigate hallucinations.

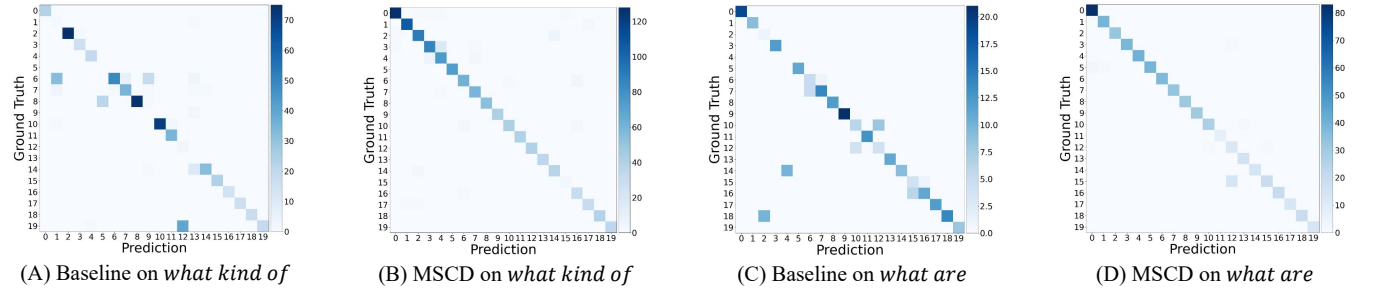


Figure 3: Confusion matrix between the predictions and ground-truth labels of top 20 class accuracy based on question types. Two question types were randomly selected for comparison to verify the effectiveness of MSCD in mitigating instance shift.

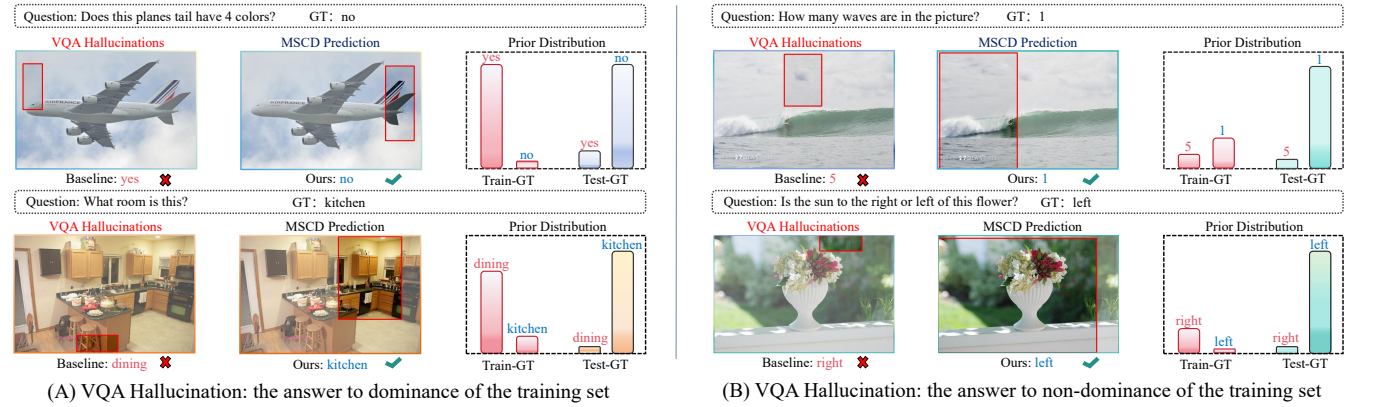


Figure 4: Attention region visualization and hallucination analysis for instances. By classifying the hallucinations of these four image-question instances into two types: dominant answers and non-dominant answers.

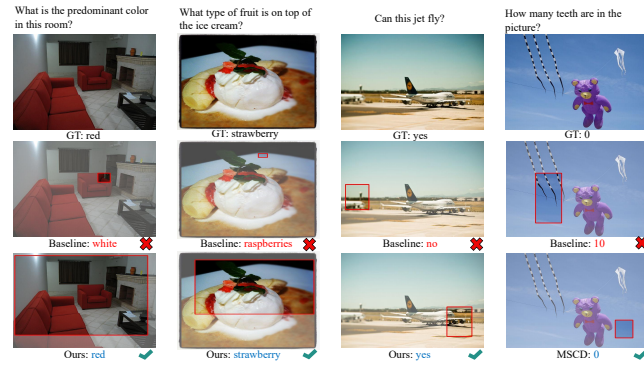


Figure 5: Visualizations for more question types showing predictions and attention regions. We demonstrate model attentional regions under the influence of hallucinations.

6 VISUALIZATION OF DIFFERENT QUESTION TYPES

To verify the robustness of the spatial learning representation, we randomly visualize the distribution of MSCD answers to six different question types on the VQA-CP v2 test set. As shown in Fig. 2, we observe that MSCD’s representation exhibits a discriminative

space across various question types. This shows that our multi-space co-debias paradigm effectively utilizes the representation space to differentiate answers for each instance. It is worth noting that: 1) Among the different types of Fig. 2 (Left), there are obvious boundaries between question types composed of different ground truths (such as "is", "where is the", "which" and "what is"), there is no entanglement between these types. Likewise, in the different types of Fig. 2 (Right), there are yes/no question types closely clustered together with an overwhelming majority of "no" answers ("is this a" and "does the"), while the question types that give overwhelmingly "yes" ground-truth answers are further away ("are"). 2) In a question type space with multiple ground truth answer classes (e.g. "what", "what is the", "what kind of", etc.), question types with higher diversity exhibit a loose distribution. In summary, the above analysis shows that the MSCD model can better distinguish instances in the discriminative space.

7 CONFUSION MATRIX FOR ROBUST VQA

To delve deeper into the phenomenon of instance shift and distribution shift, we present Fig. 3, which randomly visualizes the confusion matrices of two question types. We select the answer predictions of the top twenty classes to assess the degree of shift calibration of MSCD.

In summary, among the question types, the distribution of the top twenty classes can provide insight into the model's answer predictions. It's evident that MSCD effectively calibrates the shift between the predictions of the top twenty classes and the ground truth labels, thereby enhancing model robustness.

8 MORE DE-HALLUCINATION EXAMPLES

In this section, Fig. 4 shows how MSCD focuses on key regions and makes correct predictions under different prior conditions. The main reason for this success is that MSCD reduces the detrimental effects of priors and biases in both spaces from a homeomorphic perspective, effectively addressing biases and avoiding potential prior traps that lead to wrong answers. We further explored the internal mechanism of hallucinations and found that whether the answer is dominant or not, it may lead to model hallucinations. For Fig. 4 (A), during the training process, many instances will shift to dominant feature areas, and these dominant answers in the training set will seriously affect the robustness performance in testing. Obviously, even if the prior distributions are very different, MSCD still effectively avoids this phenomenon. It is worth noting that Fig. 4 (B) shows that even non-dominated answers can lead to model hallucination, which may be an entanglement effect between some semantically similar spaces. However, MSCD's proficiency in disentangling instance-semantic entanglements and prioritizing semantic reasoning can further enhance robustness. Furthermore, we list more different types of examples in Fig. 5, which reflects the consistent improvement of our model for different types of image-question instances.

REFERENCES

- [1] 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*. 4971–4980.
- [2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, and Mark Johnson. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6077–6086.
- [3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, and Dhruv Batra. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*. 2425–2433.
- [4] Abhipsa Basu, Sravanti Addepalli, and R. Venkatesh Babu. 2018. RMLVQA: A Margin Loss Approach for Visual Question Answering With Language Biases. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11671–11680.
- [5] Remi Cadene, Corentin Dancette, Hedi Ben-younes, Matthieu Cord, and Devi Parikh. 2019. Rubi: Reducing unimodal biases in visual question answering. In *Advances in Neural Information Processing Systems*.
- [6] 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*. 10800–10809.
- [7] Long Chen, Yuhang Zheng, and Jun Xiao. 2022. Rethinking data augmentation for robust visual question answering. In *European Conference on Computer Vision*, Vol. 13696. 95–112.
- [8] Jae Won Cho, Dong-Jin Kim, Hyeon-gon Ryu, and In So Kweon. 2023. Generative Bias for Robust Visual Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11681–11690.
- [9] 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*.
- [10] Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020. Mutant: A training paradigm for out-of-distribution generalization in visual question answering. In *2020 Conference on Empirical Methods in Natural Language Processing*. 878–892.
- [11] 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*. 6904–6913.
- [12] Yangyang Guo, Liqiang Nie, Zhiyong Cheng, Feng Ji, and Ji Zhang. 2021. Overcoming Language Priors with Adapted Margin Cosine Loss. In *Proceedings of the International Joint Conference on Artificial Intelligence*.
- [13] Xinzhe Han, Shuhui Wang, and Chi Su. 2021. Greedy gradient ensemble for robust visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1584–1593.
- [14] Xinzhe Han, Shuhui Wang, Chi Su, Qingming Huang, and Qi Tian. 2023. General greedy de-bias learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45 (2023), 1–17.
- [15] Abhishek Jha, Badri Patro, Luc Van Gool, and Tinne Tuytelaars. 2023. Barlow constrained optimization for visual question answering. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1084–1093.
- [16] Chenchen Jing, Yuwei Wu, Xiaoxun Zhang, Yunde Jia, and Qi Wu. 2020. Overcoming language priors in vqa via decomposed linguistic representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 11181–11188.
- [17] Camila Kolling, Martin D. More, Nathan Gavenski, Eduardo H. P. Pooch, Otávio Parraga, and Rodrigo C. 2023. Efficient counterfactual debiasing for visual question answering. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 3001–3010.
- [18] Gouthaman KV and Anurag Mittal. 2020. Reducing language biases in visual question answering with visually-grounded question encoder. In *European Conference on Computer Vision*. 18–34.
- [19] 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*. 3285–3292.
- [20] Yulei Niu and Hanwang Zhang. 2021. Introspective distillation for robust question answering. In *Advances in Neural Information Processing Systems*.
- [21] Yonghua Pan, Jing Liu, Lu Jin, and Zechao Li. 2024. Unbiased Visual Question Answering by Leveraging Instrumental Variable. *IEEE Transactions on Multimedia* (2024).
- [22] Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. 2018. Overcoming language priors in visual question answering with adversarial regularization. In *Advances in Neural Information Processing Systems*.
- [23] Damien Teney, Ehsan Abbasnejad, Simon Lucey, and Anton van den Hengel. 2022. Evading the Simplicity Bias: Training a Diverse Set of Models Discovers Solutions With Superior OOD Generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16761–16772.
- [24] Ali Vosoughi, Shijian Deng, Songyang Zhang, Yapeng Tian, Chenliang Xu, and Jiebo Luo. 2024. Cross Modality Bias in Visual Question Answering: A Causal View with Possible Worlds VQA. *IEEE Transactions on Multimedia* (2024).
- [25] Zhiqian Wen, Yaowei Wang, Mingkui Tan, Qingyao Wu, and Qi Wu. 2023. Digging out Discrimination Information from Generated Samples for Robust Visual Question Answering. In *Findings of the Association for Computational Linguistics*. 6910–6928.
- [26] Zhiqian Wen, Guanghui Xu, Mingkui Tan, Qingyao Wu, and Qi Wu. 2021. Debiased Visual Question Answering from Feature and Sample Perspectives. In *Proceedings of the Advances in Neural Information Processing Systems*. 3784–3796.
- [27] Xi Zhu, Zhendong Mao, Chunxiao Liu, Peng Zhang, Bin Wang, and Yongdong Zhang. 2021. Overcoming language priors with self-supervised learning for visual question answering. In *Proceedings of the International Joint Conference on Artificial Intelligence*.