An Empirical Study on the Language Modal in Visual Question Answering

Daowan Peng[1,2] , **Wei Wei**[*1,2] , **Xian-Ling Mao**[3] , **Yuanyuan Fu**[2,4] and **Dangyang Chen**[2,4]
[1]Cognitive Computing and Intelligent Information Processing (CCIIP) Laboratory, School of Computer Science and Technology, Huazhong University of Science and Technology
[2]Joint Laboratory of HUST and Pingan Property & Casualty Research (HPL)
[3]Department of Computer Science and Technology, Beijing Institute of Technology
[4]Ping An Property & Casualty Insurance Company of China, Ltd
{pengdw, weiw}@hust.edu.cn, maoxl@bit.edu.cn, fuyuanyuan83@gmail.com,
chendangyang273@pingan.com.cn

Abstract

Generalization beyond in-domain experience to out-of-distribution data is of paramount significance in the AI domain. Of late, state-of-the-art Visual Question Answering (VQA) models have shown impressive performance on in-domain data, partially due to the language priors bias which, however, hinders the generalization ability in practice. This paper attempts to provide new insights into the influence of language modality on VQA performance from an empirical study perspective. To achieve this, we conducted a series of experiments on six models. The results of these experiments revealed that, 1) apart from prior bias caused by question types, there is a notable influence of postfix-related bias in inducing biases, and 2) training VQA models with word-sequence-related variant questions demonstrated improved performance on the out-of-distribution benchmark, and the LXMERT even achieved a 10-point gain without adopting any debiasing methods. We delved into the underlying reasons behind these experimental results and put forward some simple proposals to reduce the models' dependency on language priors. The experimental results demonstrated the effectiveness of our proposed method in improving performance on the out-of-distribution benchmark, VQA-CPv2. We hope this study can inspire novel insights for future research on designing bias-reduction approaches.

1 Introduction

Visuo-linguistic understanding is an important research topic in the field of multimodal machine learning. Vision Language (V+L) tasks, such as image caption [Karpathy and Fei-Fei, 2015], referring expression comprehension [Yu *et al.*, 2016], natural language for visual reasoning [Suhr *et al.*, 2017], visual entailment [Xie *et al.*, 2018], visual commonsense reasoning [Zellers *et al.*, 2019], and visual question answering (VQA) [Antol *et al.*, 2015; Anderson *et al.*, 2018], serve as proxy tasks for evaluating the capacity of a multi-modal system to achieve high-level multimodal learning and deeper visuo-linguistic understanding. This paper specifically focuses on the VQA task, which has been a long-standing challenge in the domains of computer vision and natural language processing. Previous research has shown that many state-of-the-art VQA models tend to rely excessively on easily learnable language priors instead of effectively reasoning based on the visual content within the images during training [Goyal *et al.*, 2017; Jing *et al.*, 2020; Wen *et al.*, 2021]. As a result, these VQA models can achieve decent performance on in-distribution data by capturing superficial correlations in the language modality. However, over-reliance on language priors makes these models fragile and results in poor performance on out-of-distribution (OOD) data in real-world scenarios.

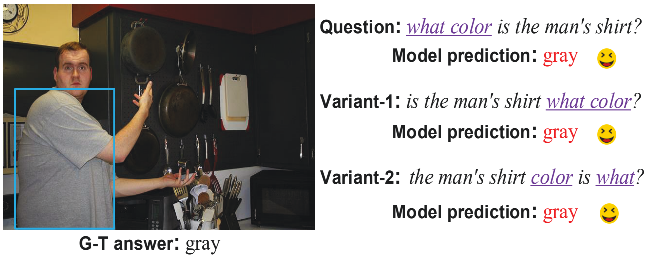

Question: what color is the man's shirt?
Model prediction: gray
Variant-1: is the man's shirt what color?
Model prediction: gray
Variant-2: the man's shirt color is what?
Model prediction: gray
G-T answer: gray

Figure 1: Here is an example that demonstrates the robustness of VQA models to question disturbances. The purple font is used to indicate the question type. Two kinds of disturbances are shown: Variant-1, which involves exchanging the positions of the prefix (question type) and postfix; and Variant-2, which randomly reorders the words.

Recently, a broad variety of bias-reduction methods [Cadene *et al.*, 2019; Clark *et al.*, 2019; Liang *et al.*, 2021; Han *et al.*, 2021; Yulei *et al.*, 2021] have been proposed, among which the commonly used approach involves adding a branch to capture language bias. For example, [Ramakrishnan *et al.*, 2018] trained a base VQA model along with a question-only adversary to mitigate bias representation by allowing the question-only model to perform poorly. However, some of these methods may introduce extra costs during the inference stage. Besides, [Yulei *et al.*, 2021] proposed

[*]Corresponding author.

a novel counterfactual inference framework based on causal effects. [Han *et al.*, 2021] introduced a greedy gradient ensemble de-bias framework, where the bias model is forced to overfit the biased data distribution, allowing the base model to learn the general patterns. Apart from the model design side, various methods from the data end have been developed to reduce language priors bias. For instance, HINT [Selvaraju *et al.*, 2019] and SCR [Wu and Mooney, 2019] utilize additional annotated data to enhance models' visual-grounding capacity for better performance [1]. CSS [Chen *et al.*, 2020a] generated counterfactual samples by masking the decisive word in the question or crucial object in the image. [Liang *et al.*, 2020] further improved the CSS method by employing contrastive learning to focus on the crucial elements. Thanks to the previous research on debiasing, some progress has been made in addressing the issue of language priors. However, in this paper, we aim to provide novel insights regarding the impact of language modality on performance in VQA tasks through empirical investigations. In this regard, we conducted a series of confirmatory experimental analysis to investigate prior bias issues. The empirical evidence revealed that, in comparison to the co-occurrence between question types and answers, the co-occurrence between objects and answers could potentially be a more significant factor in contributing to language bias. We also examine the state-of-the-art VQA models' robustness to word-sequence-related disturbance of questions and found that models are resistant to such disturbance to some extent. Figure 1 shows an example. Moreover, we found that VQA models trained with variant questions demonstrated higher accuracy in the OOD evaluation. We conducted experiments to analyze the reasons behind this phenomenon and based on these findings, we proposed bias-reduction proposals to alleviate the language bias issue. To sum up, the main contributions of this paper are as follows:

- We provide empirical evidence demonstrating that language bias in VQA tasks is not solely caused by the co-occurrence of question types and answers, but also by the co-occurrence of visually-grounded concepts and answers, with the latter having a greater impact. Additionally, there may also exist multimodal bias.
- Extensive experiments reveal that models trained with variant questions outperform those trained with original questions. This improvement is attributed to the disruptions in the word sequence of questions, which impact the model's learning of prior knowledge related to question types, leading to reduced bias-dependency learning.
- In light of the above findings, we propose de-biasing methods into multiple base VQA models by incorporating variant questions during training. The experimental results demonstrate significant performance enhancements on the VQA-CPv2 benchmark for the base models equipped with our proposed method.

2 Related Work

Visual Question Answering. As a high-level task that bridges the gap between computer vision and natural language processing, VQA [Antol *et al.*, 2015; Yang *et al.*, 2016; Agrawal *et al.*, 2017; Anderson *et al.*, 2018; Kim *et al.*, 2018; Liu *et al.*, 2022a] has received considerable attention from both the computer vision and natural language processing communities. Since the proposal of bottom-up and top-down (UpDn) attention mechanism [Anderson *et al.*, 2018], it has been the de-facto standard baseline for the VQA task. [Kim *et al.*, 2018] proposed a bilinear attention network (BAN) to efficiently compute multimodal representations. Additionally, [Yu *et al.*, 2019] developed a deep modular co-attention network (MCAN) on top of the powerful Transformer [Vaswani *et al.*, 2017], which models both intra- and inter-modal interactions simultaneously, making it a powerful baseline for the VQA task. In addition, pre-trained Vision Language Models (VLMs) [Tan and Bansal, 2019; Chen *et al.*, 2020b; Su *et al.*, 2020; Zhang *et al.*, 2021; Zeng *et al.*, 2022; Wang *et al.*, 2022] that learn high-level multi-modal representations from large-scale data via a variety of pre-training tasks have demonstrated state-of-the-art performance in many Vision Language (V+L) tasks, including VQA. For instance, the DPT model [Liu *et al.*, 2022b], which aligns the objectives of the pre-trained visual-language model with the specific requirements of the VQA task, has demonstrated improved generalizability and performance.

Bias and Robustness in VQA. The study of robustness in VQA is an important topic, particularly the issue of language bias, which significantly affects the OOD performance in VQA task. As such, an increasing number of bias-mitigation approaches [Cadene *et al.*, 2019; Guo *et al.*, 2021; Chen *et al.*, 2020a; Han *et al.*, 2021] and benchmarks [Agrawal *et al.*, 2017; Agrawal *et al.*, 2018; Kervadec *et al.*, 2021a] have been proposed. [Cadene *et al.*, 2019] built a question-only branch to capture the unwanted regularities by dynamically adjusting the loss. [Clark *et al.*, 2019] trained a naive model that relied solely on dataset biases and then used an ensemble approach to incorporate a robust model that focused on other generalized patterns. [Liang *et al.*, 2021] added a question-only branch to measure the intensity of language priors and then reshaped the objective function based on the loss of the question-only branch. [Lao *et al.*, 2021] proposed the LP-Focal loss, which endows the cross-entropy loss with sample-level loss re-weights by building a question-only branch to capture language priors. [Yang *et al.*, 2021] proposed a CCB method by building content and context branches to focus on local content and global context, respectively. On top of these two branches, a joint loss function with language bias optimizes the prediction. [Kervadec *et al.*, 2021a] suggested that the standard evaluation metric is misleading by the overall accuracy under the unbalanced concepts and questions, thus they proposed a new benchmark consisting of a dataset and a new evaluation metric. Apart from the language bias issue, [Gokhale *et al.*, 2020] found that VQA models could answer single questions but struggled to answer logical compositions of multiple such questions. Therefore, they constructed an augmentation of the VQA dataset by collecting logical com-

[1] However, it has been revealed that the accuracy improvements of these methods result from the regularization effects [Shrestha *et al.*, 2020]. Besides, collecting such human annotations can be expensive and burdensome.

position questions, including negation, conjunction, disjunction, and antonyms. [Shah *et al.*, 2019] proposed a training scheme by exploiting cycle consistency to regularize the training process, which allows VQA models to become robust to linguistic variations. Besides, [Kervadec *et al.*, 2021b] argued that noise and uncertainties in visual inputs are the main bottlenecks in VQA, which prevent the successful learning of reasoning capacities. SwapMix [Gupta *et al.*, 2022] investigated the robustness of VQA models from the perspective of visual context. They swapped some irrelevant objects in the image and found VQA models are not robust for such visual context perturbation, indicating models over-rely on them to make predictions.

3 Empirical Analysis

3.1 Task Definition

The VQA task has been cast as a classification problem, where given an image, I and a question Q, the objective is to predict an answer $\hat{a}$ from all the candidate answers A. This prediction is based on the image content and the context of the question. Without loss of generality, a VQA model can be formulated as a function transformation $F : (Q, I) \mapsto A$. The objective function $p(.)$ is formulated as:

$$\hat{a} = \arg\max_{a \in A} p(a|Q, I; \Theta), \tag{1}$$

where Θ denotes the model parameters. The common solution to predict the answer is via the cross-entropy loss

$$\begin{aligned} \mathcal{L}_{ce} &= -\frac{1}{N}\sum_{i}^{N} a_i log(p_i) \\ p_i &= Softmax(Wh_i + b), \end{aligned} \tag{2}$$

where N denotes the number of samples, W and b are the learnable matrix and bias, h_i is the fused multi-modal feature.

3.2 Revisiting Question in VQA

Which Contributes More Bias?

The language priors bias in the VQA task is generally attributed to the co-occurrence of certain question-types and answers [Agrawal *et al.*, 2018], and most bias-reduction methods are designed based on this hypothesis. In this section, we attempt to verify that the language bias issue is not solely due to the co-occurrence of question-types and answers through empirical analysis. To begin, we decompose the question into two parts: the question type (also known as the prefix) and the concepts (which include objects and other visually-grounded words or phrases in the question, also known as the postfix). We then examine their respective contributions to the final accuracy of the model. Intuitively, it is difficult to answer a question correctly if the question is incomplete (*i.e.*, only the prefix or postfix is given). If an incomplete question is answered correctly, it suggests that there is some co-occurrence correlation between the incomplete portion and the corresponding answer, which indicates the presence of bias. These experiment settings are as follows. **Dataset**: We selected the widely-used VQAv2 benchmark [Goyal *et al.*, 2017] and its OOD benchmark, VQA-CPv2 [Agrawal *et al.*, 2018]. **Base VQA models**: In the experiment, we chose the most frequently used base models in the VQA task, which include attention-based models such as SAN [Yang *et al.*, 2016] and UpDn [Anderson *et al.*, 2018]), bilinear attention network, BAN [2] [Kim *et al.*, 2018], co-attention based model, MCAN [Yu *et al.*, 2019]), multi-modal pre-trained model, LXMERT [Tan and Bansal, 2019] and a question-only model (henceforth, Q-only). Among them, the LXMERT model uses BERT [Devlin *et al.*, 2019] as the question encoder, while the other models use LSTM [Hochreiter and Schmidhuber, 1997] or GRU [Cho *et al.*, 2014] as question encoders. **Validation mode**: We conducted two types of verification. The first type involved training models with original questions and evaluating them on either the prefix or postfix. The second type involved training models with either prefix or postfix and evaluating them on the original questions. All text inputs were padded or truncated to a fixed length using a predetermined character. Moreover, for the sake of simplicity in implementation, we used the questions with the prefix removed as postfix in our experiments.

The experimental results are presented in Table 1, and several important findings can be derived from these results. The results for the first type of verification mode are displayed in columns highlighted with a light green background in Table 1. We observed that all models with postfix inputs achieved better performance than those with prefix inputs on the VQA-CPv2 test split, particularly for the BAN, LXMERT, and MCAN models, where the postfix inputs contributed significantly more than the prefix. On the VQAv2 dataset, the BAN, LXMERT, and MCAN models performed slightly better with postfix inputs, while the prefix inputs resulted in better performance for the remaining models. The experimental results for the second type of verification mode are shown in columns against a light yellow background in Table 1. As can be seen, when models were trained with the postfix, their accuracy performances were better than those trained with the prefix for all the models except MCAN, which showed slightly lower performance on both VQA-CPv2 and VQAv2. The results from both verification modes in Table 1 indicate that postfix contribute more to the bias issue than prefixes. Moreover, we noticed that certain models (*e.g.,* Q-only, UpDn and LXMERT) trained with postfix even outperformed those trained with the complete question on the VQA-CPv2 dataset. We conjecture that this could be attributed to the differences in question-type distributions between the training and test splits. Consequently, some models that do not learn information about the prefix of questions during the training process may exhibit better performance on VQA-CPv2.

Are There Any Other Kinds of Bias?

Also, Table 1 demonstrates that all models perform better than the Q-only model. Apart from the difference in model design, the Q-only model only takes the question as input, whereas others incorporate both the question and image as inputs. This allows these models to potentially depend on the co-occurrence between visual objects in the image and the keywords in the question. Therefore, we speculate that the bias issue exists not only in the language modality but also

[2] We use the 4-layers version of BAN in this paper.

Model	VQA-CPv2					VQAv2				
	ques	*pre-train*	*post-train*	*pre-test*	*post-test*	*ques*	*pre-train*	*post-train*	*pre-test*	*post-test*
Q-only	21.37	17.34	**24.87**	15.22	**16.85**	45.09	33.33	**35.68**	**31.73**	25.9
SAN	40.7	22.75	**40.35**	20.07	**27.16**	62.78	44.93	**50.71**	**39.9**	36.77
UpDn	41.53	26.12	**42.3**	21.7	**28.75**	65.56	45.19	**52.65**	**40.78**	37.32
BAN	41.73	26.6	**28.18**	22.18	**37.76**	67.07	37.28	**37.87**	39.17	**41.19**
LXMERT	40.96	28.44	**43.61**	21.05	**36.77**	64.51	44.91	**54.37**	40.24	**42.23**
MCAN	43.73	**26.31**	26.11	20.5	**34.41**	68.65	**41.46**	40.56	39.17	**39.87**

Table 1: The performance in terms of accuracy (Acc. %) on the VQA-CPv2 test split and VQAv2 validation split. The *ques* columns indicate models trained and tested with original question; *pre-train* and *post-train* denote models trained on prefix and postfix respectively and tested on the original question; *pre-tes*t and *post-test* refer to models trained on the original question and tested on the prefix and postfix respectively. The better results are bold. Note that the LXMERT was fine-tuned for 10 epochs.

across multimodalities due to the fact that incomplete questions cannot be answered.

The Robustness to Variant Question

Robustness has always been a crucial concern in machine learning. [Cui *et al.*, 2022] proposed a novel pre-training paradigm for language models. Their approach involves predicting the original order of perturbed words in text, aiming to enhance the model's resilience to the text modality and improve its ability to comprehend text semantics. Inspired by their work, this paper investigated the robustness of state-of-the-art VQA models to disruptions in the questions. To achieve this, a series of confirmatory experiments were conducted to evaluate the robustness of VQA models. Specifically, we conducted the experiments on three types of variant questions. Given a question such as "*what color is the flower?*" with the prefix "*what color is*", we defined three kinds of variant questions as follows:

- variant-1: = `Concate`(postfix, prefix), *i.e.,* exchange the positions of the prefix and postfix, resulting in a variant such as "*the flower what color is?*"
- variant-2: = `Random`(question), *i.e.,* shuffle the order of the words in the question randomly, resulting one of the possibilities as "*the flower color is what?*"
- variant-3: = `Inverse`(question), *i.e.,* inverse the word sequence of the question, making the variant as "*flower the is color what?*"

To evaluate the model's robustness to the variant questions, we define an evaluation metric Rob as follow,

$$\%Rob = \frac{N_{rv,rq}}{N_{rq}} \times 100\%, \tag{3}$$

where N_{rq} represents the number of correct predictions for the original questions, and $N_{rv,rq}$ the number of both original questions and their variants that are correctly answered.

Model	tested with	VQA-CPv2		VQAv2	
		Acc.	*Rob*	Acc.	*Rob*
Q-only	question	**21.37**	–	**45.09**	–
	variant-1	18.30	59.5	33.26	61.7
	variant-2	19.10	53.6	32.73	61.0
	variant-3	15.22	40.7	27.75	51.0
SAN	question	**40.7**	–	**62.78**	–
	variant-1	30.42	61.6	50.05	73.7
	variant-2	28.52	53.9	47.59	70.0
	variant-3	27.43	51.4	44.87	65.2
UpDn	question	**41.53**	–	**65.56**	–
	variant-1	36.86	69.1	55.71	79.7
	variant-2	31.39	55.3	49.83	70.8
	variant-3	28.38	47.7	47	66.4
BAN	question	41.73	–	**67.07**	–
	variant-1	**42.29**	79.9	59.52	84.6
	variant-2	39.87	70.1	55.26	78.3
	variant-3	35.09	60.6	48.42	68.1
LXMERT	question	40.96	–	**64.51**	–
	variant-1	**43.06**	87.4	60.76	90.4
	variant-2	40.17	74.7	56.28	82.6
	variant-3	39.12	69.9	53.96	78.6
MCAN	question	43.73	–	**68.65**	–
	variant-1	**44.13**	85.2	64.8	91.7
	variant-2	42.73	74.7	59.17	82.6
	variant-3	41.68	68.9	57.88	80.4

Table 2: The accuracy (Acc.%) and Rob in terms of different variant models on VQA-CPv2 test split and VQAv2 validation split.

For the selection of VQA models, we have used the ones chosen in the previous section. The experimental results are presented in Table 2. The results on the VQAv2 validation split demonstrate that all models experience varying degrees of performance degradation when evaluated on variant questions. Among them, MCAN and LXMERT show comparable performance on this in-distribution dataset. Furthermore, the results regarding robustness, Rob indicate that all models exhibit the best robustness for variant-1, followed by variant-2, and the worst for variant-3. Meanwhile, from the results of the VQA-CPv2 test split, an intriguing observation was that when tested with variant-1 questions, the test accuracy was even higher than that of the original questions in terms of the BAN, MCAN, and LXMERT models. The above experimental results lead us to consider a hypothesis, that is, variant questions may help alleviate language priors dependency. To further verify this hypothesis, we conducted another series of experiments. In this experimental setup, we trained VQA models using variant questions as text inputs, and the resulting models are referred to as variant models. During

Model	trained with	VQA-CPv2				VQAv2			
		All	Yes/No	Num	Other	All	Yes/No	Num	Other
Q-only	question	21.37	41.01	12.14	13.61	**45.09**	69.57	32.37	29.81
	variant-1	**27.80**+6.43	53.64	39.32	11.09	41.81	68.21	30.21	24.78
	variant-2	22.41+1.04	42.47	12.4	14.65	43.9	68.55	31.99	28.29
	variant-3	25.90+4.53	66.02	1.34	11.62	32.38	65.69	1.09	15.32
SAN	question	40.70	41.62	13.14	47.77	**62.78**	78.69	41.52	56.31
	variant-1	40.95+0.25	56.03	15.5	40.02	57.25	75.69	35.62	48.96
	variant-2	**41.32**+0.62	43.21	13.17	48.06	61.47	76.96	40.76	55.17
	variant-3	31.41-9.29	40.18	12.46	32.01	48.45	68.09	24.21	39.94
UpDn	question	41.53	42.91	13.56	48.55	**65.56**	82.87	44.9	57.87
	variant-1	**44.83**+3.30	60.45	20.84	43.23	60.37	79.86	35.21	52.21
	variant-2	42.33+0.80	44.89	13.35	48.94	64.22	81.5	43.75	56.5
	variant-3	33.89-7.64	41.22	24.9	32.52	50.35	71.52	29.68	39.72
BAN	question	41.73	42.72	13.51	48.95	**67.07**	84.11	48.2	59.11
	variant-1	47.19+5.46	61.21	18.41	47.74	63.81	81.39	44.67	55.51
	variant-2	**49.92**+8.19	67.92	20.72	48.49	63.04	81.1	42.91	54.62
	variant-3	45.87+4.14	66.81	15.04	43.37	55.19	75.51	33.01	45.61
LXMERT	question	43.29	46.37	15.38	49.34	**65.67**	83.31	46.69	57.29
	variant-1	**53.66**+10.37	75.21	21.4	51.22	65.34	83.14	45.82	56.96
	variant-2	43.57+0.28	46.49	15.71	49.68	65.29	83.19	46.35	56.7
	variant-3	45.33+2.04	51.93	22.2	48.21	59.82	76.78	43.18	51.33
MCAN	question	43.73	42.6	15.69	52.02	**68.65**	85.91	51.05	60.17
	variant-1	48.57+4.84	52.53	26.15	52.64	66.45	81.4	50.73	59.23
	variant-2	**48.79**+5.06	55.79	22.00	52.48	66.65	81.93	50.61	59.27
	variant-3	48.47+4.74	65.75	18.19	47.73	58.97	79.85	38.37	48.54

Table 3: The performance (in %) in terms of different variant models on VQA-CPv2 test split and VQAv2 validation split.

	Q-only			SAN			UpDn			BAN			MCAN			LXMERT		
	var1	var2	var3	var1	var2	var3	var1	var2	var3	var1	var2	var3	var1	var2	var3	var1	var2	var3
✓→ ✗	28.5	**11.2**	42.1	23.0	**10.5**	44.3	20.0	**11.3**	44.6	15.5	**14.7**	23.4	14.1	**14.1**	21.6	**8.5**	9.0	22.6
✗→ ✓	17.5	6.0	**17.8**	**19.1**	11.7	16.6	**23.1**	13.1	21.2	23.7	**28.0**	26.3	24.1	24.2	**28.9**	**27.6**	10.6	23.4

Table 4: The ratio (in %) of prediction changes of variant models on VQA-CPv2 test split. The mark ✓→ ✗(lower is better) measures the ratio (based on the predictions of original models) of questions that the original model can answer correctly while the variant models cannot; ✗→ ✓(higher is better) represents the opposite case. The var(x) is abbreviated of variant-(x).

the inference stage, we evaluated the performance of the variant models on original questions. The experimental results are presented in Table 3. As shown, some variant models achieved comparable performance to that of the original models on VQAv2 dataset. Furthermore, almost all the variant models achieve better performance on VQA-CPv2, except for SAN and UpDn, whose performances degrade when trained with variant-3. The LXMERT, fine-tuned for 20 epochs, even achieved a 10-point gain without adopting any debiasing methods when trained with variant-1 questions. Besides, a notable phenomenon can be observed in Table 3, which is that the accuracy changes of different variant models mainly occur in the "Yes/No" metric, although the changes demonstrated by different models vary.

3.3 Why Did the Performance Improve?

To figure out the reasons behind the accuracy improvements from these variant models on VQA-CPv2, we conducted further experimental analysis. Firstly, we performed a fine-grained analysis of the experimental results to investigate how the predictions of the variant models differed from those of the original models. To be more specific, we aimed to determine the number of predictions that changed from correct to incorrect and vice versa for the variant models.

The results are presented in Table 4, from which it can be seen that almost all of the variant-2 models have the smallest proportion of samples that were predicted correctly by the original model but predicted incorrectly by the variant models. The corresponding variant-1 model of the LXMERT model performs the best, with the smallest proportion of correct predictions flipped to incorrect ones, and it can also flip 27.6% of the incorrect predictions to correct ones. Besides, the BAN and MCAN models also demonstrate good performance regarding converting incorrect predictions to correct ones. Additionally, we conducted further analysis to examine the question types that correspond to the change in performance of the variant models. The results are depicted in Figure 2. It can be observed that the accuracy improvement

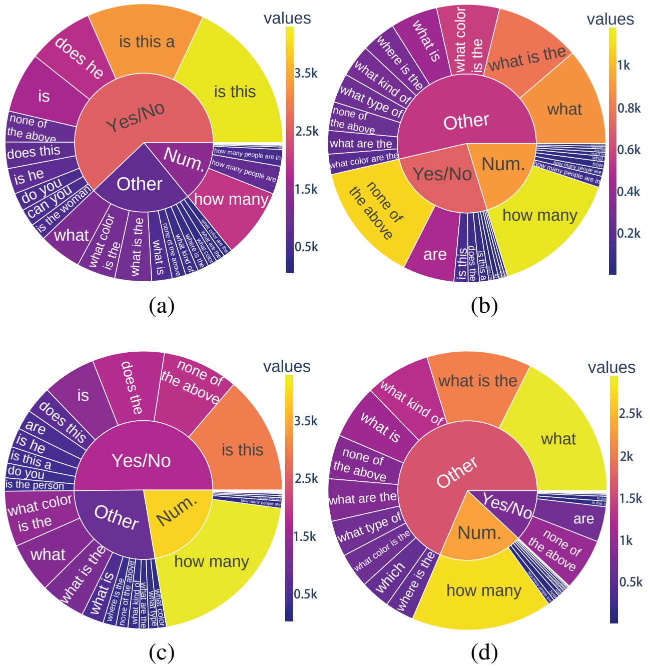


Figure 2: Statistics on the top-10 question types for each answer type corresponding to prediction-flip samples. The subplots (a) and (b) are statistics with respect to best-performing variant model of LXMERT, while (c) and (d) are with respect to best-performing variant model of UpDn. The first column represents the distribution of question types where the model changed its incorrect predictions to correct predictions, while the second column represents the opposite case.

mainly stems from the “Yes/No” answer type, and the most frequent question type is *“is there”*. On the other hand, the decreased performance mainly comes from the “Other” answer type, and the most frequent corresponding question type is *“what”*. Based on the previous results, we aimed to investigate further which words in the question the variant models focus on and how this differs from the attention of the original models. To achieve this, one of the most straightforward ways is to compare the feature representations produced by the models and their corresponding variant models for the same question. We visualized the attention weights of the questions with respect to the models and their variant models. Specifically, the question was first fed into the model to obtain its feature representation. Then, we mapped the attention to each word of the question. Figure 3 presents two toy examples. The subplots (a) and (c) in Figure 3 show that the model trained using the original question input mode places a higher weight on question type, such as the examples *“is this”* and *“what color is”*. The abundance of questions that start with the phrase *“is this”* in the training dataset makes it easier for the models to learn these simple patterns. In contrast, for the models trained with variant questions, the aforementioned scenario does not occur, resulting in slightly smoother visualization feature representations, as shown in subplots (b) and (d) of Figure 3.

In addition, more detailed results are presented in Table 3. For instance, the results of the trained variant models on the in-distribution VQAv2 dataset reveal that almost every model’s performance, in terms of each answer type, has decreased to varying degrees when compared to the performance of the original models. This demonstrates that learning the pattern of the variant questions would negatively affect the performance of the original question on in-distribution data. However, the performance of variant models on the OOD test set is quite different. Almost all models showed improvements in the All metric, primarily due to the improvements in “Yes/No” and “Num” answer types. However, the models trained with variant-1 questions, including Q-only, SAN, BAN, and UpDn, showed a decrease in performance on the “Other” metric. In contrast, LXMERT improved its performance when trained with variant-1 questions, with an accuracy improvement of +1.88% on “Other” metric. The reason for these experimental results may lie in the different question encoders used by these models.

3.4 Other Property

Furthermore, we found that the trained variant models exhibit better semantic robustness than the original models. We calculated the semantic similarity between the encoded original questions and encoded the variant questions by:

$$simi = 1 - \frac{1}{N}\sum_{i}^{N} cos < q_i, var_i > \tag{4}$$

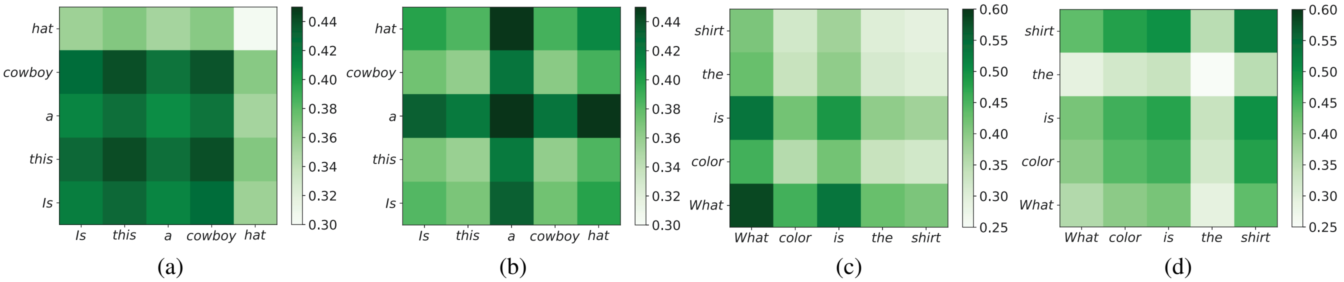
where N is the number of the samples, q_i and var_i denote the encoded original question and variant question, respectively. The results are shown in Figure 4.

4 How to Utilize These Traits?

The variant models exhibit more promising results on VQA-CPv2 compared to the original model. As demonstrated in the previous section, the variant models can avoid learning the inherent prior knowledge related to question types from the questions, allowing them to focus on other useful patterns. However, it should be noted that the syntactic structure of the variant questions may be incomplete or incorrect, and the semantics of the variant questions may even be entirely different. Therefore, the variant questions can only be utilized to aid in the design of de-biasing methods.

4.1 Proposals

To take advantage of the trait of variant models while preserving the semantics of the original question, one approach is to use the contrastive learning paradigm to combine the two encodings. In this approach, the variant question is treated as the positive sample, while negative samples are randomly sampled from the mini-batch during training. The process can be formulated as $\mathcal{L}_{con} = -\frac{1}{N}\sum_{i}^{N} log(\frac{e^{sim(h_i,h_i^{pos})}}{e^{sim(h_i,h_i^{pos})}+\sum e^{sim(h_i,h_i^{neg})}})$, where h_i is the joint feature representation of two kinds modality, h_i^{pos} and h_i^{neg} represent the positive feature and negative feature, respectively, $sim(h_i, h_i^{pos}) = \frac{h_i \cdot h_i^{pos}}{||h_i^{pos}|| \cdot ||h_i||}$. During the training stage, the total optimization objective includes the VQA loss



(a) (b) (c) (d)

Figure 3: Examples of the visualization concerning the weight mapping to each word of the questions. The subplots (a) and (b) are the visualizations with respect to the question *"Is this a cowboy hat?"*, (a) is the result of the UpDn model trained with original questions, (b) is the result of UpDn model trained with variant questions. (c) and (d) are the visualizations with respect to the question *"What color is the shirt?"*, (c) is the result of UpDn model trained with original questions, (d) is the result of UpDn model trained with variant questions.

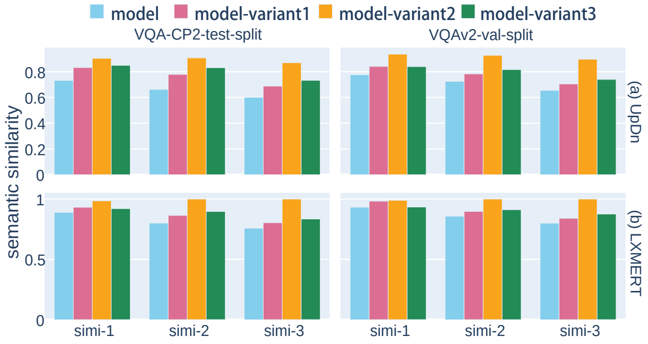


Figure 4: The semantic similarity between the encoded original questions and the encoded variant questions under different models. simi-(i) means the semantic similarity between the original question and the variant-(i) question.

$\mathcal{L}_{ce}$ and the contrastive loss $\mathcal{L}_{con}$. Besides, another straightforward way is to combine the features of the original question with those of the variant question directly as a kind of data augmentation. Specifically, the joint features were encoded by a weighted combination of two kinds of features. Therefore, the resulting features cover richer patterns.

4.2 Experiments

In this section, we validate the effectiveness of the latter debiasing method on the OOD benchmark, *i.e.,* VQA-CPv2. The proposed method is model-agnostic and can be combined with any other VQA model. Here, we also choose the most widely used base VQA models, SAN, UpDn, BAN, LXMERT, MCAN, and the Q-only model, as the baseline models. Regarding the implementation details in the training process, we adhere to the experimental settings of the open-source codes and do not modify other parameters such as learning rate, batch size, or optimizer. The experimental results are presented in Table 5. As evident from the results, all base models exhibited performance improvements when integrated with the method proposed in this paper. Moreover, the majority of models demonstrated enhancements in the "Other" metric, with only the UpDn and LXMERT models experiencing a slight decrease. This indicates that the base models combined with the proposed method can not only improve the simple pattern but also learn more difficult patterns. In addition, while the performance of some models combined with the proposed method may not be as good as the results of the variant models, the overall accuracy has improved significantly compared to the original models.

Model	VQA-CPv2			
	All	Yes/No	Num	Other
Q-only	21.37	41.01	12.14	13.61
Q-only+ours	**26.3**	42.59	11.7	16.23
SAN	40.70	41.62	13.14	47.77
SAN+ours	**41.41**	43.37	12.82	48.23
UpDn	41.53	42.91	13.56	48.55
UpDn+ours	**44.95**	54.51	14.89	48.18
BAN	41.73	42.72	13.51	48.95
BAN+ours	**43.63**	45.96	15.04	50.25
MCAN	43.73	42.6	15.69	52.02
MCAN+ours	**44.89**	44.58	16.24	52.92
LXMERT	43.29	46.37	15.38	49.43
LXMERT+ours	**50.85**	71.54	17.19	49.24

Table 5: The experimental results (Acc.%) of base VQA models combined with the proposed data augmentation method. Note that the LXMERT was fine-tuned for 20 epochs.

5 Conclusion and Future Work

In this paper, we investigate language modality in the VQA task through experimental analysis. The empirical findings indicated that the issue of language priors bias is not only related to question types alone, the postfix of questions even has a greater impact on language bias. Furthermore, we observed that variant models outperform original models on the VQA-CPv2 benchmark. We identified the underlying reasons for these results and proposed new debiasing methods based on these findings. The experimental results demonstrated that our method enhances the VQA models' generalization ability. Our main purpose is not to pursue state-of-the-art results but to gain insights for designing bias-reduction methods. However, we only present some novel experimental findings, and we plan to provide in-depth theoretical analysis and probe other methods to leverage these traits in future work.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62276110, No. 62172039 and in part by the fund of Joint Laboratory of HUST and Pingan Property & Casualty Research (HPL). The authors would also like to thank the anonymous reviewers for their comments on improving the quality of this paper.

References

[Agrawal *et al.*, 2017] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C Zitnick, Devi Parikh, and Dhruv Batra. Vqa: Visual question answering. *IJCV*, 123(1):4–31, 2017.

[Agrawal *et al.*, 2018] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *CVPR*, pages 4971–4980, 2018.

[Anderson *et al.*, 2018] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, pages 6077–6086, 2018.

[Antol *et al.*, 2015] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. In *ICCV*, pages 2425–2433, 2015.

[Cadene *et al.*, 2019] Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. Rubi: Reducing unimodal biases for visual question answering. In *NeurIPS*, pages 839–850, 2019.

[Chen *et al.*, 2020a] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *CVPR*, pages 10800–10809, 2020.

[Chen *et al.*, 2020b] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, pages 104–120, 2020.

[Cho *et al.*, 2014] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *EMNLP*, pages 1724–1734, 2014.

[Clark *et al.*, 2019] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *EMNLP-IJCNLP*, pages 4067–4080, 2019.

[Cui *et al.*, 2022] Yiming Cui, Ziqing Yang, and Ting Liu. Pert: Pre-training bert with permuted language model. *CoRR*, abs/2203.06906, 2022.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186, 2019.

[Gokhale *et al.*, 2020] Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. Vqa-lol: Visual question answering under the lens of logic. In *ECCV*, pages 379–396, 2020.

[Goyal *et al.*, 2017] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, pages 6325–6334, 2017.

[Guo *et al.*, 2021] Yangyang Guo, Liqiang Nie, Zhiyong Cheng, Feng Ji, Ji Zhang, and Del Alberto Bimbo. Adavqa: Overcoming language priors with adapted margin cosine loss. In *IJCAI*, pages 708–714, 2021.

[Gupta *et al.*, 2022] Vipul Gupta, Zhuowan Li, Adam Kortylewski, Chenyu Zhang, Yingwei Li, and Alan Yuille. Swapmix: Diagnosing and regularizing the over-reliance on visual context in visual question answering. In *CVPR*, pages 5078–5088, 2022.

[Han *et al.*, 2021] Xinzhe Han, Shuhui Wang, Chi Su, Qingming Huang, and Qi Tian. Greedy gradient ensemble for robust visual question answering. In *ICCV*, pages 1564–1573, 2021.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[Jing *et al.*, 2020] Chenchen Jing, Yuwei Wu, Xiaoxun Zhang, Yunde Jia, and Qi Wu. Overcoming language priors in vqa via decomposed linguistic representations. In *AAAI*, pages 11181–11188, 2020.

[Karpathy and Fei-Fei, 2015] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137, 2015.

[Kervadec *et al.*, 2021a] Corentin Kervadec, Grigory Antipov, Moez Baccouche, and Christian Wolf. Roses are red, violets are blue... but should vqa expect them to? In *CVPR*, pages 2776–2785, 2021.

[Kervadec *et al.*, 2021b] Corentin Kervadec, Theo Jaunet, Grigory Antipov, Moez Baccouche, Romain Vuillemot, and Christian Wolf. How transferable are reasoning patterns in vqa? In *CVPR*, pages 4207–4216, 2021.

[Kim *et al.*, 2018] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *NeurIPS*, pages 1571–1581, 2018.

[Lao *et al.*, 2021] Mingrui Lao, Yanming Guo, Yu Liu, and Michael S Lew. A language prior based focal loss for visual question answering. In *ICME*, pages 1–6. IEEE, 2021.

[Liang *et al.*, 2020] Zujie Liang, Weitao Jiang, Haifeng Hu, and Jiaying Zhu. Learning to contrast the counterfactual samples for robust visual question answering. In *EMNLP*, pages 3285–3292, 2020.

[Liang *et al.*, 2021] Zujie Liang, Haifeng Hu, and Jiaying Zhu. Lpf: A language-prior feedback objective function for de-biased visual question answering. In *SIGIR*, pages 1955–1959, 2021.

[Liu *et al.*, 2022a] Yuhang Liu, Wei Wei, Daowan Peng, Xian-Ling Mao, Zhiyong He, and Pan Zhou. Depth-aware and semantic guided relational attention network for visual question answering. *TMM*, pages 1–14, 2022.

[Liu *et al.*, 2022b] Yuhang Liu, Wei Wei, Daowan Peng, and Feida Zhu. Declaration-based prompt tuning for visual question answering. In *IJCAI*, pages 3264–3270, 2022.

[Ramakrishnan *et al.*, 2018] Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. Overcoming language priors in visual question answering with adversarial regularization. In *NeurIPS*, pages 1548–1558, 2018.

[Selvaraju *et al.*, 2019] Ramprasaath R Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry Heck, Dhruv Batra, and Devi Parikh. Taking a hint: Leveraging explanations to make vision and language models more grounded. In *ICCV*, pages 2591–2600, 2019.

[Shah *et al.*, 2019] Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. Cycle-consistency for robust visual question answering. In *CVPR*, pages 6649–6658, 2019.

[Shrestha *et al.*, 2020] Robik Shrestha, Kushal Kafle, and Christopher Kanan. A negative case analysis of visual grounding methods for vqa. In *ACL*, pages 8172–8181, 2020.

[Su *et al.*, 2020] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *ICLR*, 2020.

[Suhr *et al.*, 2017] Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *ACL*, pages 217–223, 2017.

[Tan and Bansal, 2019] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP-IJCNLP*, pages 5099–5110, 2019.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017.

[Wang *et al.*, 2022] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. In *ICLR*, 2022.

[Wen *et al.*, 2021] Zhiquan Wen, Guanghui Xu, Mingkui Tan, Qingyao Wu, and Qi Wu. Debiased visual question answering from feature and sample perspectives. In *NeurIPS*, pages 3784–3796, 2021.

[Wu and Mooney, 2019] Jialin Wu and J. Raymond Mooney. Self-critical reasoning for robust visual question answering. In *NeurIPS*, pages 8601–8611, 2019.

[Xie *et al.*, 2018] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment task for visually-grounded language learning. *CoRR*, abs/1811.10582, 2018.

[Yang *et al.*, 2016] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *CVPR*, pages 21–29, 2016.

[Yang *et al.*, 2021] Chao Yang, Su Feng, Dongsheng Li, Huawei Shen, Guoqing Wang, and Bin Jiang. Learning content and context with language bias for visual question answering. In *ICME*, pages 1–6. IEEE, 2021.

[Yu *et al.*, 2016] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, pages 69–85, 2016.

[Yu *et al.*, 2019] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *CVPR*, pages 6281–6290, 2019.

[Yulei *et al.*, 2021] Niu Yulei, Tang Kaihua, Zhang Hanwang, Lu Zhiwu, Hua Xian-Sheng, and Wen Ji-Rong. Counterfactual vqa: A cause-effect look at language bias. In *CVPR*, pages 12700–12710, 2021.

[Zellers *et al.*, 2019] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, pages 6720–6731, 2019.

[Zeng *et al.*, 2022] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. In *ICML*, pages 25994–26009, 2022.

[Zhang *et al.*, 2021] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *CVPR*, pages 5579–5588, 2021.