Eliminating Language Bias in Visual Question Answering with Potential Causality Models

Anonymous ACL submission

Abstract

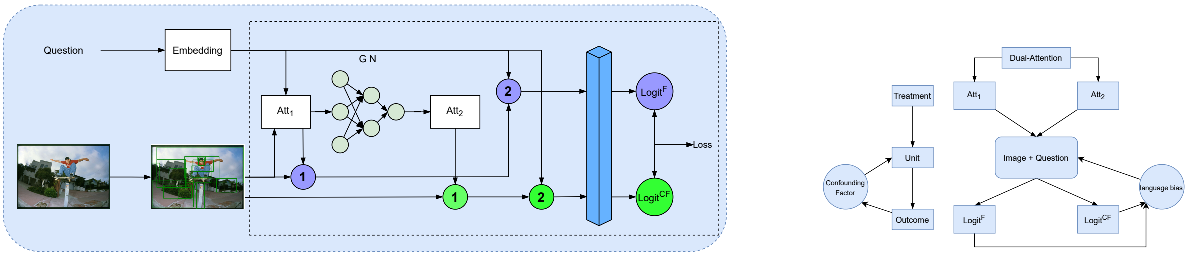
The main goal of Visual Question Answering (VQA) is to effectively learn useful information from vision and language to perform answer reasoning. However, recent studies have shown that VQA models often have language bias, which is the false correlations between questions and answers, rather than truly extracting answers from multi-modal knowledge. Existing methods mainly focus on modeling the question part to capture the language bias, while ignoring the influence of visual content on the model. To address this issue, in this paper, we combine potential causal models with VQA models, using dual-attention as treatment, and treating language bias as a confounding factor in the model. We enhance the role of visual information in the VQA model through the construction of observed and counterfactual outcomes, thus eliminating the impact of language bias on the VQA model. We conduct experiments on the VQA-CP v2 and VQA v2 datasets to demonstrate the effectiveness of our proposed method.

1 Introduction

In recent years, visual question answering (VQA) has gained significant attention in deep learning research (Hudson and Manning, 2019). This task (Goyal et al., 2017; Antol et al., 2015) aims to answer questions about a given image through a model, which is required to handle multi-modal information from vision and language (Tan and Bansal, 2019). However, recent studies (Kafle and Kanan, 2017; Agrawal et al., 2016) have shown that VQA are often vulnerable to language biases, which cause models to rely on false associations between questions and answers when responding to questions. When the same question is presented with a different image, the model may still incorrectly rely on the previous biased language association to provide a wrong answer. Therefore, many research studies have aimed to mitigate language biases in VQA.

Several groundbreaking methods have been developed to address language bias. The most straightforward method is using data augmentation methods to address language biases (Chen et al., 2020). These methods (Liang et al., 2020; Gokhale et al., 2020) build counterfactual samples based on the VQA-CP v2 dataset (Agrawal et al., 2018), and adopt strategies such as expanding training samples to balance this dataset and improve the VQA model's data generalization ability. However, it is worth noting that the VQA-CP v2 dataset is used solely to verify whether VQA models can recognize language biases. Therefore, while these methods can enhance the model's generalization ability under unbiased sample conditions, how to perform unbiased inference without data augmentation remains a challenging task. Another category of methods (Jing et al., 2020; Cadene et al., 2019; Clark et al., 2019; Han et al., 2021) is ensemble-based regularization methods. These methods train the two models as a whole, using a weak model to capture shallow or spurious patterns. The main model can then focus on the more difficult examples, thus removing bias effects. However, it should be noted that this method does not utilize visual information fully for true connections between images and questions to perform answer reasoning.

Thus, we believe that solving the language bias can be achieved by enhancing the involvement of visual information in the model. Similar to the CF-VQA (Niu et al., 2021), we choose to leverage causal models in our method. The main purpose of causal models is to infer causal relationships between variables in large datasets, seeking links between results and causes. Therefore, for VQA, causal models can efficiently capture the correlations between inputs and outputs. In the CF-VQA, the authors define language bias as a direct causal



(a) The proposed DAP model uses a dual-attention mechanism. First, it leverages the Top-Down attention method from UpDn to generate attention weight distribution att_1. This att_1 is then fed into an Attention Generation Network (GN) to produce att_2. The visual information is fused with these two attention types (depicted by the purple and green circles marked 1), and the final feature fusion is done (purple and green circles marked 2). The acquired features are then input to a classification layer to obtain $Logit^F$ and $Logit^C F$, which are used for causality estimation.

(b) The Correspondence Between the Three Elements in PCM and Bias Issues in VQA

Figure 1: Overview of the model.

effect of question-answer pairs using Structural Causal Models(SCMs), and remove the causal effect of language bias from the overall model. However, the accuracy of a VQA model requires not only correctly responding to visual content, but also on the ability to answer incorrectly based on incorrect visual information (Han et al., 2021). This demands not only a full understanding of the visual content by the model, but also an ability to establish the correspondence between visual and language information. Therefore, simply removing the influence of language from the total causal effect would greatly weaken the model's understanding of language.

Compared to the prior works, we propose using Potential Causal Models (PCMs) (Yao et al., 2021) to capture the language biases in VQA models. PCMs are typically used to observe the outcomes of applying specific treatments to samples, such as doctors observing symptoms to prescribe treatment after a patient has taken a certain medication. PCMs make inferences by obtaining the causal effects between observed data. For VQA, we utilize attention mechanisms to focus on the visual content and construct observed and counterfactual outcomes using PCMs. This allows us to optimize attention's ability to understand images, enhancing the involvement of visual content in VQA. Meanwhile, prior to language bias removal in the model, we evaluate the current questions' significance in the model during inference through a score for language bias evaluation, and dynamically adjust language bias removal in the model. Based on this, we attempt to assess the causal effects of VQA by measuring the relationship between the potential outcomes of different treatments in PCMs. Therefore, we propose a de-biasing algorithm based on PCMs (DAP), which utilizes dual-attention to combine PCMs with VQA to achieve the goal of de-biasing.

The key points of DAP lie in how to apply the PCMs to the VQA, and to model the causal relationship among vision, language, and answer. First, we employ dual-attention as a treatment to estimate causality using observed outcomes and counterfactual outcomes. Secondly, we define language bias as a confounding factor in PCMs and aim to remove the bias by eliminating its effect on the model. As depicted in Figure 1(a), we obtain two forms of attention during the model inference phase and use them to calculate corresponding potential results to perform causal estimation. Figure 1(b) delineates the process of the dashed line in detail. We show that our method achieves substantial performance improvement via extensive experiments. In summary, the main contributions of this paper are twofold. First, we use PCMs to capture the causal relationships between observed data and evaluate the causal effects of VQA by measuring the relationships between different potential outcomes in PCMs. Second, our method employs dual-attention mechanism to enhance the model's ability to understand visual information and eliminate language bias.

2 Related Work

2.1 Language Bias

In recent studies, multiple researchers have focused on addressing language biases that arise in VQA, which stands for VQA (Agrawal et al., 2016; Jabri et al., 2016; Manjunatha et al., 2019; Teney et al., 2020b). Some of the forms of language biases in VQA have been widely defined by scholars such as Han *et al.* (Han et al., 2021) and Wen *et al.* (Wen et al., 2021). One of the types of biases is the statistical distribution bias that exists between the train and test sets, which leads to a long-tail phenomenon in the answer distribution of certain question types in the dataset. Consequently, models can achieve high accuracy by simply providing a "Yes" answer to a given question. Another type is the shortcut bias between the question and its answer, where there is a strong correlation between the two. For example, when asked the question "What sport?", the model can easily answer "tennis" and still achieve a high accuracy rate. Finally, Wen *et al.* (Wen et al., 2021) had introduced the concept of visual bias, which refers to the tendency of VQA models to focus on prominent objects in the image when answering questions, leading to the provision of incorrect answers.

2.2 De-bias Method

Recently, many researches have focused on mitigating the impact of language biases on models, and a new VQA dataset (VQA-CP) has been proposed by Agrawal et al. (Agrawal et al., 2018) to evaluate the generalization ability of VQA models. Currently, Debiasing methods applied to this dataset can be classified into several categories: First, methods of architectural bias force the model to predict answers by incorporating VQA knowledge into the model's architecture, as in the works of Agrawal et al. (Agrawal et al., 2018) and Kumar and Verma (Kv and Mittal, 2020). Second, adversarial methods (Ramakrishnan et al., 2018) use adversarial losses to reduce known sources of bias by inducing errors in the model when presented with only the question. Third, regularization methods include ensemble-based regularization (Cadene et al., 2019; Clark et al., 2019; Niu et al., 2021; Han et al., 2021; Han et al., 2023; Guo et al., 2019; Gat et al., 2020) and data-augmentation based regularization (Chen et al., 2020; Liang et al., 2020; Agarwal, 2020; Teney et al., 2020a; Si et al., 2022; Wen et al., 2021; Chen et al., 2022; Zhu et al., 2020; Teney et al., 2021) to improve the robustness of the main model. Fourth, Structuring output loss methods, such as Dancette and Lebret (Dancette et al., 2020), Kervadec et al. (Kervadec et al., 2020), and Guo et al. (Guo et al., 2021), helped the model understand its errors by providing additional information rather than correcting them.

2.3 Causal Inference

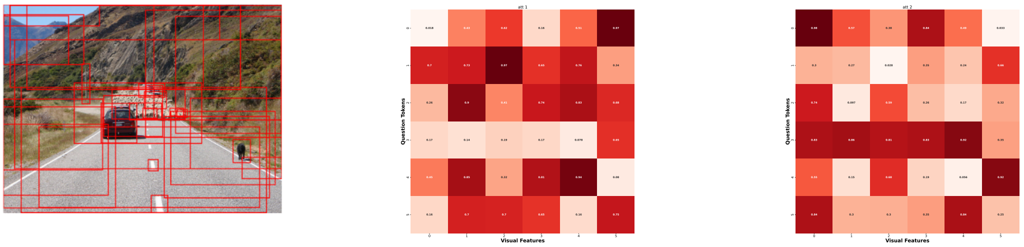
Deep causal models have become a core method based on unbiased estimation in the field of deep learning (Li and Zhu, 2022). Similarly, causal inference has made progress in visual-language tasks. Agarwal et al. (Agarwal, 2020) proposed a new metric for analyzing and measuring model robustness based on the dependence of VQA on language-relatedness, and generate an additional VQA dataset through semantic operations. Shah et al. (Shah et al., 2019) proposed a model-agnostic cyclic consistent training scheme to increase the model's robustness by semantic change. Ray et al. (Ray et al., 2019) introduced a new VQA dataset and a quantitative metric to evaluate VQA consistency. Wang et al. (Wang et al., 2020) proposed an unsupervised region feature learning method that captures inter-object relationships through causal relationships and applies them to VQA. Meanwhile, Niu et al. (Niu et al., 2021) proposed a counterfactual VQA method using SCMs that indicate essential elements of VQA in causal graphs, removing language biases through direct and indirect effects. In this paper, we strengthen the involvement of visual content in multi-modal inference through attention mechanisms, balance confounding factors and enhance the model's understanding of the visual content to mitigate language biases.

3 Method

This section will cover how we apply the PCMs to VQA task. For information related to PCMs, please refer to Appendix 6. For the VQA, we need to provide the model with inputs for vision and language, denoted as V and Q respectively. As illustrated in Figure 1(b), we juxtapose the concepts in PCMs with the key elements in VQA, and use dual-attention as treatment to obtain the corresponding observed and counterfactual outcomes for the data V and Q, enabling the estimation of causality.

3.1 Dual Attention

In the previous section, we discuss the impact of confounding factors in the PCMs, which can lead to false effects on the outcomes and affect the overall assessment of the model. Thus, we also take into account the impact of confounding factors on the VQA model when combining it with the PCMs. Specifically, we define the language bias in the



(a) Bottom-Up attention feature representation

(b) att_1

(c) att_2

Figure 2: Diagram of dual attention, where Visual Features indicate the visual features processed by Fast R-CNN, and Question Tokens indicate the question vector with a maximum length. (a) indicates the attention mechanism in the UpDn model, and (b) indicates a simple example of the opposite attention generated in our method. The depth of the colors indicates the degree that the model pays attention. (c) indicates the feature representation formed by the image under the Bottom-Up attention mechanism.

VQA model as the confounding factor and eliminate its effect on the model by balancing confounding factors using the method.

The presence of attention allows models in VQA to extract key information from different modalities efficiently and reason to produce answers. In previous work, visual-guided attention is usually used for multi-modal reasoning. Inspired by the concept of information theory (Saxe et al., 2019), we believe that something less likely to happen contains much more information than something certain. Analogous to the VQA, we believe that when a model applies a certain forward attention, it pays extra attention to certain objects, whereas its focus may shift dramatically when applying a reversed attention.

Therefore, the difference in the information contained between the two attention distributions is greater when they are opposite, and the model can learn more knowledge from the reversed attention. Therefore, to estimate the impact of language bias in VQA, we relate the predicted results generated by both the forward and reversed attention distributions to the relationship between observed outcomes and counterfactual outcomes in PCMs. Consequently, we introduce the notion of dual-attention. As shown in Figure 2(a), the model generates k sets of visual features $V = \{v_1, ..., v_k\}$ via the Bottom-Up attention mechanism (Anderson, 2018), and a set of attention distributions att_1 via the Top-Down attention mechanism after weighting. As shown in Figure 2(b), given the visual features and text length, att_1 indicates the attention weight distribution of each feature in the image under normal conditions. In Figure 2(c), att_2 indicates the attention distribution completely opposite to att_1, where objects worth attention in att_1 receive less attention in att_2. The color scheme in the diagram indicates the level of focus the model should have on a given feature following a proper comprehension of the question and image content. Accordingly, darker colors signify greater focus the model should direct toward the feature, thus yielding more information.

Specifically, our method generates two forms of attention during the model inference phase. Firstly, att_1 is calculated by the UpDn model, which weights each feature using the Top-Down attention mechanism. Secondly, after obtaining att_1, we introduce an attention generation network to obtain att_2, which has the opposite attention weight distribution to att_1. The attention calculation is defined in Eq.(1):

$$\begin{aligned} att_1 &= f(V, Q) \\ att_2 &= \mathbb{F}(att_1), \end{aligned} \tag{1}$$

where f denotes the Top-Down attention mechanism, and $\mathbb{F}$ represents the attention generation network.

Finally, we utilize the two forms of attention as treatments to strengthen the visual information, establish the relationships between entities across modalities, balance the confounding factors, and weaken the false correlation between question and answer.

3.2 Eliminating Bias

After the previous description, we can clearly define the main objective of our method: to use dual-attention to balance confounding factors and remove the impact of language bias on the VQA model.

Initially, we indicate the data in VQA as binary pairs (v_i, q_i), which are the basic samples in our model, where $v_i \in V = \{v_1, ..., v_n\}$ and $q_i \in Q = \{q_1, ..., q_n\}$. The dual-attention is defined as $W = \{W_1, W_2\}$, which denotes the treatments applied to the samples. $Logit_i^F$ and $Logit_i^{CF}$ denote the observed and counterfactual outcomes, respectively. In the VQA model, they indicate the predicted answer probabilities under the corresponding attentions. The causal effect of a single sample under different treatments can be given by Eq.(2):

$$\begin{aligned} Logit_i^F &= \mathcal{H}((W_1 * v_i) * q_i) \\ Logit_i^{CF} &= \mathcal{H}((W_2 * v_i) * q_i) \\ I\hat{T}E_i &= Logit_i^F - Logit_i^{CF}, \end{aligned} \tag{2}$$

where $\mathcal{H}$ indicates the linear layer for answer classification based on multi-modal knowledge. $I\hat{T}E_i$ stands for the difference in the results between the observed outcome and the counterfactual outcome under different treatments for a given sample.

3.2.1 Language Bias Score

To eliminate confounding factors, we allocate attention to the given sample to focus on visual information. Our method directly applies att_1 and att_2 treatments simultaneously to the sample (v_i, q_i) and considers the corresponding results as the observed and counterfactual outcomes. Moreover, we redefine the propensity score mentioned in Section A.4 as $e(v_i, q_i)$, and rewrite Eq.(11) as Eq.(3):

$$\begin{aligned} Q_{pred} &= \mathcal{Z}(V = \varnothing, Q = q_i) \\ e(v_i, q_i) &= Softmax(Q_{pred}), \end{aligned} \tag{3}$$

where $V = \varnothing$ and $Q = q_i$ indicate that we only input the question branch into our model, ignoring visual information, to evaluate the impact of the question branch alone on the model. This evaluation provides us with the score reflecting the impact of biases on answer distribution. Q_{pred} denotes the answer distribution of the question branch after prediction by the model.

3.2.2 Causal Estimation

After defining the language bias scores, we proceed the $A\hat{T}E$ on our samples to eliminate their impact on VQA data. Firstly, for a given dataset (Q, V), we rephrase the $A\hat{T}E$ calculation formula from Eq.(13) to Eq.(4):

$$A\hat{T}E = \frac{1}{N}\sum_{i=1}^{N}(\frac{Logit_i^F - Logit_i^{CF}}{e(v_i, q_i)}) - Q_{pred}, \tag{4}$$

where $\frac{1}{e(v_i,q_i)}$ denotes the bias-to-answer's impact transformed into weight. We perform causal inference for our sample by calculating the value of $A\hat{T}E$ with visual information enhancement weighting for model prediction. Specifically, $Logit_i^F - Logit_i^{CF}$ indicates the $I\hat{T}E$, which is the difference between the answer distribution in the presence and absence of counterfactual attention. This method increases the contribution of essential visual information. Finally, to eliminate the effect of bias, we subtract Q_{pred} from the overall model prediction.

3.2.3 Training and Testing

Given (Q, V), as shown in Figure 3(a), the model optimizes the entire network by minimizing the cross-entropy loss between the predicted results and labels. During the training phase, the causal relationship $A\hat{T}E$ obtained from different attention, shown in Figure 3(b), serves as the output of the proposed method and is used as the prediction result. Simultaneously, our optimization strategy involves minimizing the cross-entropy loss between $A\hat{T}E$ and the label. Therefore, The final loss is the combination of $\mathcal{L}_{cls}$ and $\mathcal{L}_{ate}$:

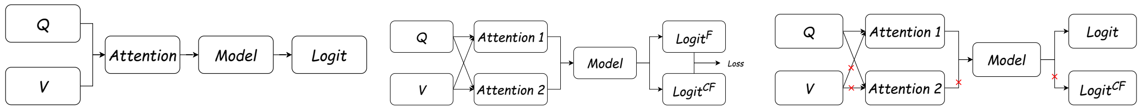
$$Loss = \mathcal{L}_{cls} + \mathcal{L}_{ate}, \tag{5}$$

where $\mathcal{L}_{cls}$ refers to the cross-entropy loss between the model's predicted probability and the ground truth label. $\mathcal{L}_{ate}$ refers to the cross-entropy obtained by $A\hat{T}E$ and the labels.

As shown in Figure 3(b), during the training phase, DAP generates dual-attention for a given (Q, V). This process entails generating observed and counterfactual outcomes. Additionally, we allow att_1 to learn the differences in information quantity between the two , thereby allowing it to pay closer attention to the visual region that is most relevant to the question. However, as shown in Figure 3(c), during the testing phase, we no longer generate att_2 and instead only use the same network structure as the UpDn model for testing.

4 Experiments

Our experiments are conducted primarily on the VQA-CP v2 (Agrawal et al., 2018) and VQA v2



(a) UpDn (b) DAP-train (c) DAP-test

Figure 3: The diagram illustrates the DAP during both the training and testing stages. (a) UpDn model using attention for reasoning. (b) DAP generates two types of attention during the training phase, to enhance visual information. (c) During the testing phase, the model remains consistent with UpDn by using only one type of attention.

(Goyal et al., 2017) datasets. These experiments aim to validate the effectiveness of our method in mitigating language bias and its adaptability on general datasets. To facilitate comparison, we choose the UpDn model (Anderson, 2018) as the baseline and evaluate results according to the VQA evaluation metrics (Antol et al., 2015). Due to our use of attention methods, we conduct experiments and comparisons on the UpDn model.

4.1 Quantitative Analysis

4.1.1 Comparison with Other Methods

We conduct experiments on the VQA-CP v2 and VQA v2 datasets using our method and compare them with other state-of-the-art methods on these datasets. The experimental results are presented in Table1. Specifically, we show on the VQA-CP v2 dataset:

(1) Our method shows significant improvement over the UpDn baseline model, increasing the overall performance by 19%.

(2) Additionally, compared to the CF-VQA model, which also utilizes a causal model and the same baseline model as ours, our method outperforms CF-VQA by approximately 5% in terms of the overall performance. However, for specific question types, CF-VQA is better than our method in answering "Y/N" questions, it may be attributed to CF-VQA's stronger ability to correct distributional biases in the dataset. However, for other question types that require more visual information, our method significantly outperforms CF-VQA. This is because our dual-attention design continually optimizes attention and enhances visual content involvement in understanding visual information during training.

(3) Moreover, we also compare our method with some other visual-aware methods. AdaVQA (Guo et al., 2021) achieves better results in the "Num." question type, but is inferior to our method in the "Other" question type. This is because AdaVQA eliminates answer bias from a feature perspective but overlooks the impact of visual content and answer diversity on the model.

(4) In addition, our method demonstrates competitiveness when compare with other methods that use data augmentation and balanced datasets.

Finally, in experiments on the VQA v2 dataset, our method also achieves reasonable performance.

4.1.2 Analysis of Other Metrics

In our approach, we aim to increase the role of visual content in reasoning. To assess its effectiveness, we use additional metrics. In Table 2, we compare our results with other methods using the CGD metric. For a more detailed understanding of CGD, please refer to (Han et al., 2021; Shrestha et al., 2020). As shown in Table 1, compared to GGE, our method performs better in "Num." question type but slightly worse in "Other" question type. In these two types of questions that require more visual information, our method demonstrates strong competitiveness compared to them. Furthermore, we compare our method with GGE using the CGD evaluation metrics, as shown in Table 2. As the CGD is used to evaluate whether visual information is utilized for answer prediction, the results indicate that our method score higher than the GGE method in CGD evaluations. Therefore, this suggests some improvements in our method's ability to utilize visual information for answer prediction.

4.2 Abalation Experiments

In this section, we design a series of ablation experiments to verify the effectiveness of our DAP method in bias mitigation. These experiments are primarily conducted on the VQA-CP v2 dataset.

The Efficacy of Dual-Attention In the first group of experiments, we aim to verify the impact of dual-attention method on VQA models. Using att_2 in conjunction with two opposing weight distributions, dual-attention enables att_1 to contain more information. We compare our method with

Table 1: The results of VQA-CP v2 test set and VQA v2 validation set are presented in the following table. Each column illustrates the **Best** performances of each method, excluding data augmentation techniques. Our DAP method has been compared with state-of-the-art methods on both datasets.

Data set		VQA-CP v2 test				VQA v2 val			
Method	Base	All	Y/N	Num.	Other	All	Y/N	Num.	Other
GVQA	-	31.30	57.99	13.68	22.14	48.24	72.03	31.17	34.65
SAN	-	24.96	38.35	11.14	21.74	52.41	70.06	39.28	47.84
UpDn	-	39.96	43.01	12.07	45.82	63.48	81.18	42.14	**55.66**
S-MRL	-	38.46	42.85	12.81	43.20	63.10	-	-	-
HINT	UpDn	46.73	67.27	10.61	45.88	63.38	81.18	42.99	55.56
SCR	UpDn	49.45	72.36	10.93	48.02	62.2	78.8	41.6	54.5
RUBi	UpDn	44.23	67.05	17.48	39.61	-	-	-	-
LMH	UpDn	52.01	72.58	31.12	46.97	56.35	65.06	37.63	54.69
DLP	UpDn	48.87	70.99	18.72	45.57	57.96	76.82	39.33	48.54
DLR	UpDn	48.87	70.99	18.72	45.57	57.96	76.82	39.33	48.54
AttAlign	UpDn	39.37	43.02	11.89	45.00	63.24	80.99	42.55	55.22
CF-VQA(SUM)	UpDn	53.55	**91.15**	13.03	44.97	**63.54**	**82.51**	**43.96**	54.30
GGE-DQ-tog	UpDn	57.32	87.04	27.75	49.59	59.11	73.27	39.99	54.39
AdaVQA	UpDn	54.67	72.47	**53.81**	45.58	-	-	-	-
DAP(Ours)	UpDn	**59.00**	86.23	47.70	48.19	60.54	75.34	40.92	54.48
Methods of data augmentation and additional annotation:									
CSS	UpDn	58.95	84.37	49.42	48.24	59.91	7.25	39.77	55.11
Mutant	UpDn	61.72	88.90	49.68	50.78	62.56	82.07	42.52	53.28
D-VQA	UpDn	61.91	88.93	52.32	50.39	64.96	82.18	44.05	57.54
KDDAug	UpDn	60.24	86.13	55.08	48.08	62.86	80.55	41.05	55.18
OLP	UpDn	57.59	86.53	29.87	50.03	-	-	-	-

the HINT (Selvaraju et al., 2019), SCR (Wu et al., 2019) methods on an attention level perspective.

As shown in Table 3, applying the dual-attention method solely to the UpDn baseline model also yield performance improvement. Additionally, compared to other attention-based methods, our method demonstrate certain performance advancements, particularly in question types that require visual context.

The dual-attention method is also shown to be competitive in these scenarios.

PCMs Ablation Experiments In this section, we conduct ablation experiments to verify the impact of PCMs on debiasing. As mentioned earlier, the key idea of DAP is to make full use of visual information to reduce the influence of bias. Therefore, we consider comparing other works to validate the effectiveness of PCMs. Specifically, we transform from causal inference at the visual level to causal inference at the language level. Similar to other works, we model bias in the question-answer branch, and bias is considered as the counterfactual outcomes in PCMs.

Therefore, we combine the PCMs with the VQA framework using the aforementioned method and compare our method to CF-VQA, which also employs causal models. As shown in Table 4, our method exhibit similar overall performance to CF-VQA. However, for specific question types, such as "Y/N", PCMs underperform compared to CF-VQA. Conversely, for "Num." and "Other" types, PCMs outperform CF-VQA.

4.3 Visual Qualitative Analysis

In this section, we will present visual experimental results to demonstrate the effectiveness of our method in mitigating language bias. As shown in Figure 4, in the first image, DAP not only provide the correct answer when ask about the shape of an object, but also increase the credibility of the response. For judgment-type questions, the model must effectively understand the image's content to

Table 2: Experiment on the evaluation metric CGD using the DAP method on the VQA-CP v2 dataset. **Best** results are displayed in each column.

Method	CGR	CGW	CGD
UpDn	44.27	40.63	3.91
HINT	45.21	34.87	10.34
RUBi	39.60	33.33	6.27
LM	**47.30**	35.97	11.33
LMH	46.44	35.84	10.60
CSS	46.70	37.89	8.87
GGE-D	38.79	**24.48**	14.31
GGE-DQ-iter	44.35	27.91	16.44
GGE-DQ-tog	42.74	27.47	15.27
DAP(Ours)	46.83	30.21	**16.62**

Table 3: For the ablation experiments with dual-attention on VQA-CP v2 dataset, DAP method is compared with other attention-related VQA debiasing techniques, using UpDn as the baseline model. **Best** results are displayed in each column.

	All	Y/N	Num.	Other
UpDn	39.96	43.01	12.07	45.82
LMH	52.01	72.58	31.12	46.97
HINT	46.73	67.27	10.61	45.88
SCR	49.45	72.36	10.93	**48.02**
DAP(Ours)	**51.15**	**83.25**	**24.02**	41.78

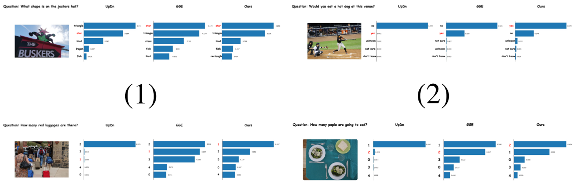
provide an accurate answer. DAP correctly comprehend the scene of the second image and provide the correct response. Counting problems are the most challenging types of questions in VQA. In the third image, DAP accurately identify the red luggage object and correctly answer the question. Additionally, in the fourth image, the model needs to understand the image's content and respond to the counting question. The UpDn baseline model's reasoning process leads to a low possibility of obtaining the correct response. However, our method generates precise reasoning and provides accurate responses. This method increases the participation of visual information in multi-modal reasoning.

5 Conclusion

In this paper, we propose a de-biasing method based on PCMs (DAP). We aim to eliminate language bias in VQA models while enhancing the influence of the visual content on the model. We consider language bias as a confounding factor in the PCMs and propose to use dual-attention to construct observed and counterfactual outcomes. Through balancing the confounding factors, we are able to eliminate the influence of the language bias on the model. The effectiveness of our method is demonstrated through extensive experiments. In addition, we believe that enhancing the model's understanding of visual content is a future research direction for the elimination of language bias. Our study demonstrates the significant experimental significance of the causal model in visual language tasks with reasonable experimental designs.

Table 4: For the ablation experiments of the PCMs on the VQA-CP v2 dataset, the DAP method indicates using only the PCMs, while the DAP(att) method indicates the simplified version with dual-attention. **Best** results are displayed in each column.

	All	Y/N	Num.	Other
UpDn	39.82	42.40	12.23	46.05
CF-VQA(SUM)	53.72	**90.86**	13.08	44.98
DAP(Ours)	**53.77**	79.02	**22.69**	**49.07**



(1) (2)

(3) (4)

Figure 4: The results of qualitative analysis show the flow of our model when making predictions by masking different image regions so that the model focuses on the effective ones

6 Limitations

Firstly, our method's performance depends on the dual-attention algorithm, and when this algorithm fails to effectively focus on relevant areas of the image, it significantly impacts model performance. Secondly, the correspondence between biases in VQA and causal effects may be more complex in real situations. Different types of biases may affect the model's causal effects differently, thus requiring distinct considerations.

References

Vedika et al. Agarwal. 2020. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9690–9698.

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

Aishwarya Agrawal et al. 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 et al. Anderson. 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.

Remi Cadene et al. 2019. Rubi: Reducing unimodal biases for visual question answering. *Advances in neural information processing systems*, 32.

Long Chen, Yuhang Zheng, and Jun Xiao. 2022. Rethinking data augmentation for robust visual question answering. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, pages 95–112. Springer.

Long Chen et al. 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*.

Corentin Dancette, Remi Cadene, Xinlei Chen, and Matthieu Cord. 2020. Overcoming statistical shortcuts for open-ended visual counting. *arXiv preprint arXiv:2006.10079*.

Itai Gat, Idan Schwartz, Alexander Schwing, and Tamir Hazan. 2020. Removing bias in multi-modal classifiers: Regularization by maximizing functional entropies. *Advances in Neural Information Processing Systems*, 33:3197–3208.

Tejas Gokhale et al. 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.

Yangyang Guo, Zhiyong Cheng, Liqiang Nie, Yibing Liu, Yinglong Wang, and Mohan Kankanhalli. 2019. Quantifying and alleviating the language prior problem in visual question answering. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 75–84.

Yangyang Guo, Liqiang Nie, Zhiyong Cheng, Feng Ji, Ji Zhang, and Alberto Del Bimbo. 2021. Adavqa: Overcoming language priors with adapted margin cosine loss. *arXiv preprint arXiv:2105.01993*.

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*.

Xinzhe Han et al. 2021. Greedy gradient ensemble for robust visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1584–1593.

Drew Hudson and Christopher D Manning. 2019. Learning by abstraction: The neural state machine. *Advances in Neural Information Processing Systems*, 32.

Guido W Imbens. 2004. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29.

Allan Jabri, Armand Joulin, and Laurens Van Der Maaten. 2016. Revisiting visual question answering baselines. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 727–739. Springer.

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*, 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. 2020. Estimating semantic structure for the vqa answer space. *arXiv preprint arXiv:2006.05726*.

Gouthaman Kv and Anurag Mittal. 2020. Reducing language biases in visual question answering with visually-grounded question encoder. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 18–34. Springer.

Zongyu Li and Zhenfeng Zhu. 2022. A survey of deep causal model. *arXiv preprint arXiv:2209.08860*.

Zujie Liang et al. 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.

Varun Manjunatha, Nirat Saini, and Larry S Davis. 2019. Explicit bias discovery in visual question answering models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9562–9571.

Yulei Niu et al. 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.

Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. 2018. Overcoming language priors in visual question answering with adversarial regularization. *Advances in Neural Information Processing Systems*, 31.

Arijit Ray, Karan Sikka, Ajay Divakaran, Stefan Lee, and Giedrius Burachas. 2019. Sunny and dark outside?! improving answer consistency in vqa through entailed question generation. *arXiv preprint arXiv:1909.04696*.

Paul R Rosenbaum. 1987. Model-based direct adjustment. *Journal of the American statistical Association*, 82(398):387–394.

Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Donald B Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Andrew M Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan D Tracey, and David D Cox. 2019. On the information bottleneck theory of deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124020.

R. R. Selvaraju et al. 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.

Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. 2019. Cycle-consistency for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6649–6658.

Robik Shrestha et al. 2020. A negative case analysis of visual grounding methods for vqa. *arXiv preprint arXiv:2004.05704*.

Qingyi Si et al. 2022. Towards robust visual question answering: Making the most of biased samples via contrastive learning. *arXiv preprint arXiv:2210.04563*.

Jerzy Splawa-Neyman, Dorota M Dabrowska, and Terrence P Speed. 1990. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, pages 465–472.

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

Damien Teney, Ehsan Abbasnedjad, and Anton van den Hengel. 2020a. Learning what makes a difference from counterfactual examples and gradient supervision. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pages 580–599. Springer.

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

Damien Teney, Ehsan Abbasnejad, and Anton van den Hengel. 2021. Unshuffling data for improved generalization in visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1417–1427.

Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. 2020. Visual commonsense representation learning via causal inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 378–379.

Zhiquan Wen et al. 2021. Debiased visual question answering from feature and sample perspectives. *Advances in Neural Information Processing Systems*, 34:3784–3796.

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

Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. 2021. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(5):1–46.

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 Appendix

A.1 The Three Fundamental Elements

This section introduces the three fundamental elements of PCMs: unit, treatment, and outcome. A unit is the smallest physical entity used in causal inference. Consistent with the original paper (Yao et al., 2021), we use the terms "sample" and "unit" interchangeably in this paper. Treatment refers to the action or policy that the model applies to a unit, and the outcome indicates the state of the unit after a certain treatment. That is, by applying some treatment to a unit, we can observe the corresponding outcome.

The outcomes under PCMs can be classified into three categories based on the effects of treatment (Yao et al., 2021):

- Potential outcomes: For any sample, there exist potential outcomes corresponding to any treatment, denoted as $Y(w_i)$, where w_i indicates one of the treatments, and the set of all treatments is denoted by $W = \{w_1, ...w_n\}$.

- Observed outcomes: Observed outcome Y^F indicates the actual manifestation of potential outcomes when a treatment is applied to a sample. The relationship between observed and potential outcomes is defined by Eq.(6):

$$Y^F = Y(w_i), \tag{6}$$

  where Y^F indicates the observed outcomes, and $Y(w_i)$ is the outcome of actually applying a specific treatment w_i.

- Counterfactual outcomes: The outcome of a sample under an alternative treatment is defined by Eq.(7):

$$Y^{CF} = Y(w_j), \tag{7}$$

  where Y^{CF} indicates the potential outcomes that are not observed. $Y(w_j)$ denotes the outcome under a different treatment w_j, where $w_j \in W$ and $j \neq i$.

  Notably, counterfactual outcomes are also potential outcomes, and potential outcomes comprise both observed and counterfactual outcomes. Furthermore, for binary treatments ($w \in \{0, 1\}$), observed and counterfactual outcomes can be defined by Eq.(8):

$$\begin{aligned} Y^F &= Y(w) \\ Y^{CF} &= Y(1-w), \end{aligned} \tag{8}$$

  where $w \in \{0, 1\}$ denotes the treatment applied to the sample, while $1 - w$ indicates the alternative potential treatment applied to the sample at the same time.

A.2 Treatment Effect

Treatment effect defines the performance of a unit before and after taking a treatment, and is typically estimated by the outcomes (Yao et al., 2021; Rubin, 1974; Splawa-Neyman et al., 1990).

- Individual Treatment Effect ($I\hat{T}E$): The difference between the observed and counterfactual outcomes of the i-th unit. The $I\hat{T}E$ of unit i is defined as Eq.(9):

$$I\hat{T}E_i = Y_i(W = 1) - Y_i(W = 0) \tag{9}$$

  where $Y_i(W = 1)$ and $Y_i(W = 0)$ denote the observed and counterfactual outcomes for unit i respectively.

- Average Treatment Effect ($A\hat{T}E$): The difference between the observed and counterfactual outcomes in the overall sample, as defined as Eq.(10):

$$\begin{aligned} A\hat{T}E &= \mathbb{E}[Y(W = 1) - Y(W = 0)] \\ &= \frac{1}{N}\sum_{i=1}^{N}(Y_i(W = 1) - Y_i(W = 0)) \\ &= \frac{1}{N}\sum_{i=1}^{N} ITE_i, \end{aligned} \tag{10}$$

  where $Y(W = 1)$ and $Y(W = 0)$ indicate the observed and counterfactual outcomes respectively, and N denotes the total number of samples.

A.3 Confounding Factor

The essence of a confounding factor is mixing the effects of various factors. When multiple factors intertwine with the effects of the outcome, correctly assessing the true impact of a specific factor on the outcome can be challenging. Specifically, in the context of the PCMs, confounding factors are a special type of variables that influence treatment allocation and the final outcome, resulting in spurious effects.

For instance, "age" can be regarded as a confounding factor when evaluating the effect of a particular drug treatment on a given disease, and failing to account for age can lead to biased outcomes in the final assessment. This bias is a spurious effect of confounding factors on evaluating the treatment. Therefore, in the PCMs, it is necessary to account for the influence of confounding factors to obtain a correct estimate of the treatment effect.

A.4 Eliminating Confounding Factors

Accounting for the influence of confounding factors is a crucial element of causal inference models. This subsection primarily describes the methods for eliminating confounding factors (Yao et al., 2021; Imbens, 2004; Rosenbaum and Rubin, 1983; Rosenbaum, 1987).

- Selection bias: The observed outcomes cannot indicate the outcomes of interest, due to the influence of confounding factors on the choice of treatment, leading to a biased phenomenon.

- Propensity score: The probability of taking a specific treatment under a given background condition is defined as Eq.(11):

  $$e(x_i) = Pr(W = 1|X = x_i), \qquad (11)$$

  where Pr represents conditional probability, $W = 1$ indicates taking a specific treatment, X and x_i respectively refer to the sample set and a specific sample i, and $e(x_i)$ denotes the probability of sample i taking a specific treatment given the sample set X.

- Inverse propensity weighting: The reweighting of samples based on the propensity score allotting a new weight to each sample, defined as Eq.(12):

  $$r = \frac{W}{e(x_i)} + \frac{1-W}{1-e(x_i)}, \qquad (12)$$

  where W denotes some treatment, $e(x_i)$ is the propensity score, and r indicates inverse propensity weighting. Specifically, when the given sample x_i is inclined to select a certain treatment, it implies that this treatment would achieve better results under the model. Therefore, we use weight allocation to balance the effect of this treatment.

- Inverse propensity weighting $A\hat{T}E$: The $A\hat{T}E$ after re-weighting the sample, defined as Eq.(13):

  $$A\hat{T}E = \frac{1}{N}\sum_{i=1}^{N}\frac{W_iY_i^F}{e(x_i)} - \frac{1}{N}\sum_{i=1}^{N}\frac{(1-W_i)Y_i^{CF}}{1-e(x_i)} \qquad (13)$$

  After re-weighting the sample using propensity scores, it is sufficient to eliminate the influence of selection bias when taking different treatments. At this point, the difference between the observed outcomes and the counterfactual outcomes can be utilized to eliminate the impact of confounding factors.

A.5 Language Bias Score

In Sections A.4 and 3.2.1, we introduce the conversion from propensity score to language bias score, but we do not maintain their formal unity. Therefore, we will give a detailed introduction to Eq.(11) and Eq.(4).

First, in Eq.(11), $e(x)$ represents the conditional probability that any study subject is allocated to the treatment group or control group given the conditions, ultimately achieving a balanced sample between different groups. However, in our method, we adopt the dual-attention treatment for each batch of samples simultaneously, so no sample grouping is done. Secondly, we redefine the propensity score as a language bias score in Eq.(3), which is the score for predicting the answer, in the case of modeling the language bias. Therefore, in Eq.(4), we only use $\frac{1}{e(v_i,q_i)}$ as the re-weighted score. Additionally, to maintain consistency with the re-weighting method in Eq.(13) and the computation formula for $A\hat{T}E$ is defined as Eq.(14):

$$A\hat{T}E = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{Logit_i^F}{e(v_i,q_i)} - \frac{Logit_i^{CF}}{1-e(v_i,q_i)}\right) - Q_{pred}. \qquad (14)$$

We also conduct experiments using the same weighting method, as shown in Table 5.

A.6 Daul-Attention Simplified Version

In addition, to demonstrate the effectiveness of dual-attention on PCMs, we conduct a simplified version which is called DAP(att). Rather than using an attention generation network to generate att_2, here we obtain att_2 by simply using Eq. (15):

$$att_2 = 1 - att_1. \qquad (15)$$

Alternatively, we obtain att_2 directly by subtracting the weights from att_1 and combine it with the PCMs for the experiment. As shown in Table 6, DAP(att) indicates a simplified version of our dual attention approach. It is evident that the model achieves higher overall accuracy, particularly for question types that require more visual content.

A.7 Attention Visualization

Figure 5 displays the selected attention visualization examples for analysis. In these specific exam-

Table 5: The experimental results of re-weighting method $\hat{ATE}$ calculation. **Best** results are displayed in each column.

	All	Y/N	Num.	Other
UpDn	39.82	42.40	12.23	**46.05**
DAP(Ours)	**57.61**	**80.66**	**48.94**	45.15

Table 6: The experimental results of simplified version of dual-attention. **Best** results are displayed in each column.

	All	Y/N	Num.	Other
UpDn	39.82	42.40	12.23	**46.05**
DAP(att)	**57.82**	**80.82**	**49.24**	45.16

ples, att_1 accurately captures visual information, but att_2 only focuses on irrelevant image objects.

The attention weight of the corresponding image is illustrated in Figure 6.

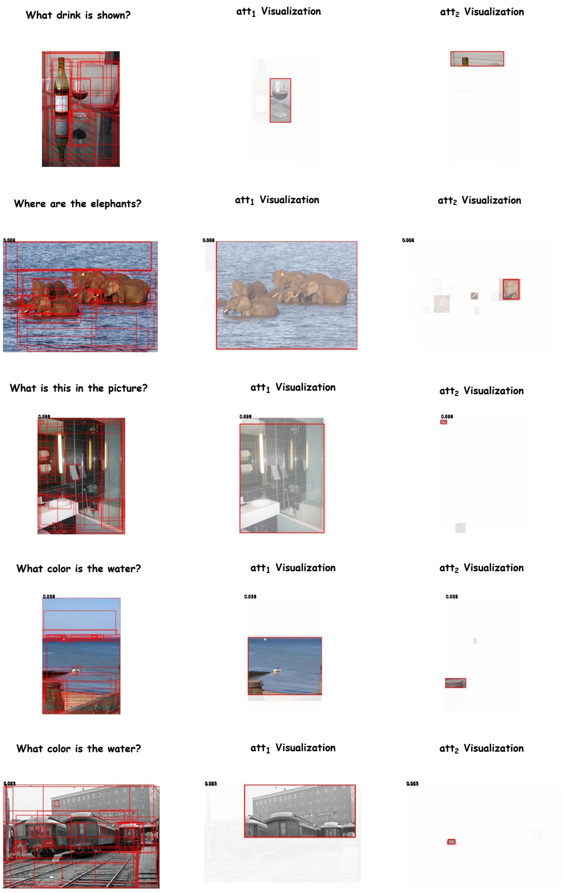

What drink is shown?
att_1 Visualization
att_2 Visualization
Where are the elephants?
att_1 Visualization
att_2 Visualization
What is this in the picture?
att_1 Visualization
att_2 Visualization
What color is the water?
att_1 Visualization
att_2 Visualization
What color is the water?
att_1 Visualization
att_2 Visualization

Figure 5: The visualization results of dual-attention indicate the original image and the visualized results of the two types of attention, respectively.

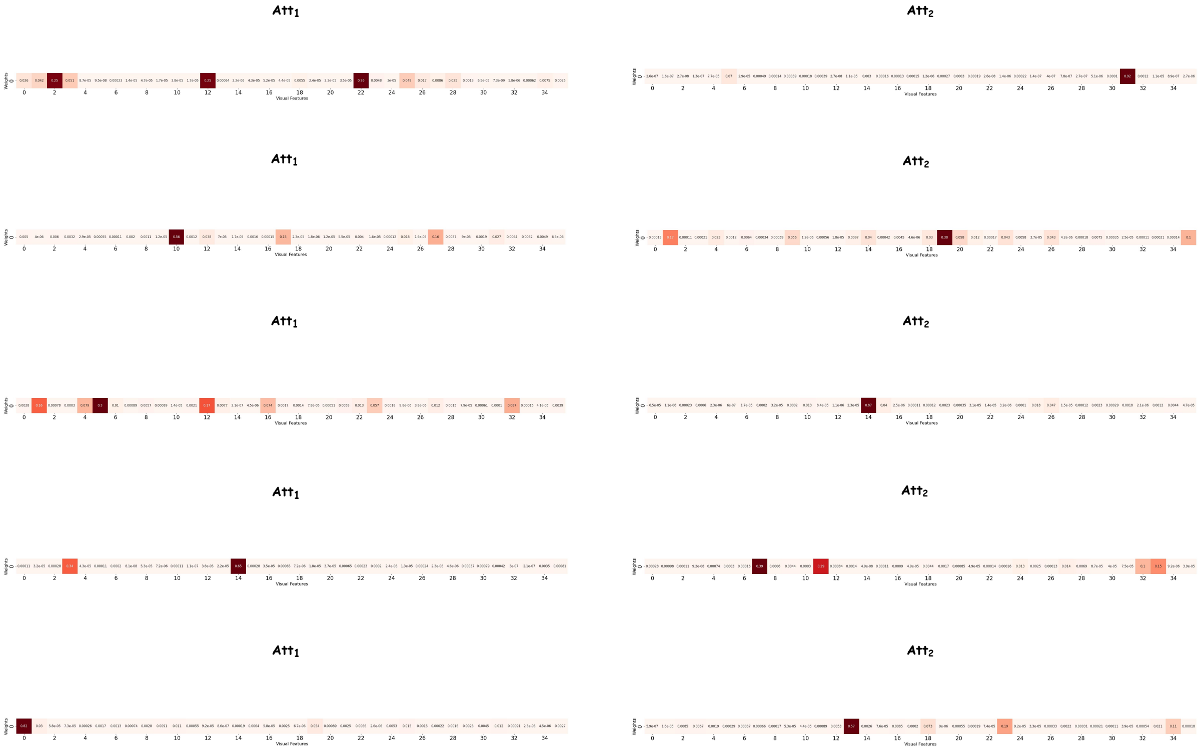


Figure 6: Example of Attention Weight Heatmap Distribution