# A Survey of Causality in Visual Question Answering

**Gaoxiang Luo, Hao Zou***
Department of Computer Science and Engineering
University of Minnesota, Twin Cites
Minneapolis, MN 55454
luo00042, zou00080@umn.edu

## Abstract

In this survey, we present a in-depth comparative analysis for causality in the Visual Question Answering (VQA) systems. The VQA systems are one of foundations for interactive AI in the open world, and the study of the causality has recently shown promises to improve the interpretability while also solves distribution shift in data. Therefore, we compared and contrasted five state-of-the-art VQA models inspired by causal thinking. We categorized the models based on the Pearl's Ladder of Causation. To compare their performance, we tested the models on the same dataset VQA v2.0. The results demonstrated the feasibility of utilizing causality to perform multi-modal reasoning (i.e., the goal of VQA tasks), and sheds light on applying causality in other deep learning domains.

## 1 Introduction

In the past 20 years, deep learning applications, including recommender systems, smart voice assistants, face recognition, etc., have been widely accepted in people's daily living. However, when it comes to the application domains that might affect human lives, such as healthcare and autonomous vehicle, people cannot fully trust Artificial Intelligence (AI) unless they're interpretable and stable [23]. Deep neural networks, for instance, can fit a non-linear function arbitrarily well by the theoretical support of the universal approximation theorem [13]. Still, it remains a black-box approach since people don't comprehensively understand what these models are learning from. More importantly, most machine learning algorithms and deep learning models make and heavily rely on the i.i.d hypothesis (i.e., the training and testing set have the same data distribution). Therefore, these trained models might not have the desired performance whenever the data distribution is different from the training set in a non-stationary environment, which is also known as the issue of generalization to out-of-distribution (OOD) data. Fortunately, the study of causality and its bridging to deep learning has recently shown promises to address both the issues of interpretability and stability above. In this survey, we will target the field of VQA and perform an in-depth analysis of how causality being involved in modern deep learning models to answer free-form and open-ended questions.

Problems at the intersection of vision and language have been a challenging research area with great significance, and VQA systems are one of the fundamental building blocks to support the frontier interactive AI systems, such as visual commonsense reasoning [27], vision-language navigation [4] and visual dialog [10]. Therefore, these VQA systems have to be capable of performing visual analysis, language understanding and multi-modal reasoning. However, most VQA systems are thrilled to find the correlations in data and hence fail to achieve human-level intelligence (see Figure 1). Also, people come to realize that scaling a data-driven language model with plenty of computational resources may only bring little performance to machine's visual understanding, so they switch their attention to causality. It is natural to think human intelligence is inseparable from their

*These authors contribute equally to this work.

Preprint. Under review.

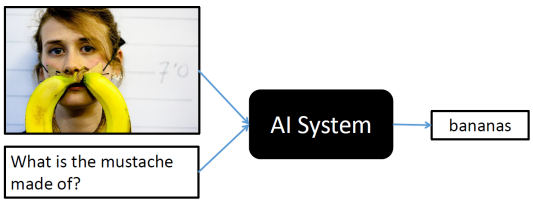


Figure 1: The key problem that VQA is aiming to solve.

causal reasoning. While the structural causal model (SCM) with the causal graph [19] remains as the most comprehensive causality framework with theoretical support, there are very few deep learning research [18] successfully implementing SCM, because raw data (e.g., images) are unstructured and do not expose any direct information about causality [22]. Nonetheless, people borrow the concepts from causality (e.g., intervention, counterfactual, etc) and obtained a significant performance boost in VQA tasks.

The main contribution of this survey is twofold. First, it's the first quantitative survey specifically concerning causality in VQA system to the authors' best knowledge. Second, we categorized almost all the state-of-the-art VQA models that deployed the understanding of causality in the past 3 years (including 2021) according to the Pearl's Ladder of Causation.

## 2 Related Work

### 2.1 Representation Learning

Learning good representation is vital to the success of deep learning models, and learning causal representation may accommodate the scenarios when the i.i.d. hypothesis doesn't hold. Ideally, people want to adopt the understanding of causal inference into deep learning models. However, it turns out to be non-trivial because the current causal models are structural [12], so they are unable to deal with the high-dimensional raw data, such as images. Therefore, learning causal representation (i.e., converting high-dimensional raw data to structural variables that can be used in the causal model) [22] will be an essential bridge to connect and combine causal inference and machine learning to build the next generation AI – Causal AI, as well as the solving the OOD and distribution shift issues to obtain better generalization.

### 2.2 Pearl's Ladder of Causation

In this survey, we will not thoroughly cover causality in a philosophical sense but provide some necessary priories to draw a connection between causality and modern machine learning. Judea Pearl, as a founder of Causal AI, describes the exploring of causality using the ladder of causation in *The Book of Why* [21]. The first level of the ladder is association, which is where most deep learning models are located, and it emphasizes what you see. The idea behind is the passive observation, which is essentially the possibility of the occurrence of event Y when event X occurs $P(Y|X)$. Association, in other words, is also identified as correlation within the machine learning community, but correlation doesn't imply causation [25]. For example, sociologists found both the sales of ice cream and the rate of crime increase during the hot weather, but it doesn't imply that selling more ice cream would result in more crimes. In fact, the false causation is caused by the ignorance of the confounder [12] – hot weather. By a loose definition, confounder is a variable that influences both the dependent variable and independent variable, and hence causing a spurious association.

The second level of the ladder of causation is intervention, and it emphasizes what you do. Besides passive observation, operational changes may intervene and remove the effect of confounders so that a true causal inference can be derived. Randomized controlled trial (RCT) [2] has been a gold standard to analyze cause and effect, but sometimes it may be unethical to apply particular intervention such as randomly selecting a group of people to smoke cigarettes for years. Therefore, do-calculus proposed by Pearl [20] is designed to remove the effect of confounder by adjusting the formulation of observation data.

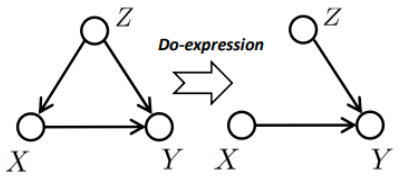


Figure 2: Do-calculus Causal Intervention. Nodes X, Y and Z refer to variables and arrows refer to the direct causal effects.

The third level of the ladder of causation is counterfactual, which is also the top of the ladder. The key difference between counterfactual and intervention is that counterfactual emphasis hindsight, which is to inference based on counterfactual, such as "Would this patient have lower blood sugar if she had received a different medication?" [14].

### 2.3 Visual Question Answering (VQA)

The largest two communities of deep learning today are Computer Vision (CV) and Natural Language Processing (NLP), and VQA is a task proposed by Parikh et al. [5], bringing vision and language people together to achieve higher intelligence of AI in the open world. The original problem definition is that by training on the dataset $\mathcal{D} = \{I_i, Q_i, A_i\}_i^N$ a model needs to predict the answer $A'$ accurately when given $\{I', Q'\}$ during the testing stage. In order to easily evaluate a proposed algorithm, most questions tend to seek specific information and hence their answers are sufficient by simply consisting of one-to-three words.

## 3 Preliminaries

### 3.1 Causal Graph

A causal graph is a directed graph $G = (\mathcal{V}, \mathcal{E})$ that describes the causal effect between variables, where $\mathcal{V}$ represents the vertices and $\mathcal{E}$ stands for the edges. The graph in Figure 2 before Do-expression can be a example of a typical causal graph, where the variable $Z$ has a direct causal effect on the variable $Y$ donoted by $Z \rightarrow Y$ and an indirect causal effect on the variable $Y$ via the variable $X$ (i.e., $Z \rightarrow X \rightarrow Y$).

### 3.2 Do-Calculus

According to Figure 2, for all causal inferences between X and Y, all potential direct or indirect common causes Z can be observed. However, in order to infer that the occurrence of X can lead to the occurrence of Y, X and Z should be conditional independent. Do-calculus eliminate the effect of Z to force X incorporating every Z fairly.

$$\text{Original Bayes rule: } P(Y|X) = \sum_z P(Y|X,z)P(z|X) = \frac{P(Y,X)}{P(X)} \tag{1}$$

$$\text{After Do-Expression: } P(Y|do(X)) = \sum_z P(Y|X,z)P(z) = \frac{P(Y,X,z)P(z)}{P(X,z)} \tag{2}$$

## 4 Dataset

### 4.1 VQA v1.0

The original VQA v1.0 dataset [5] proposed in 2015 consists of 204,721 images from the MS COCO dataset [15] and a abstract scene dataset [29, 6] with around 50,000 scenes (see Figure 3). In addition, VQA v1.0 contains more than 760K questions with around 10M answers, where around 40% of questions requires a "yes/no" answer and 13% of questions require a "number" answer.


Figure 3: A example of the original VQA v1.0 dataset (the third image is an abstract scene).


Figure 4: A example of the VQA v2.0 balanced dataset (i.e., each question has two answers and images respectively).

### 4.2 VQA v2.0

Among the VQA datasets, there exists strong correlation between questions and answers, which is interpreted as "language prior" by Niu et al. [17]. For instance, a model that answers "tennis" to the sport-related questions can easily achieve 40% of accuracy on the VQA v1.0 dataset. Additionally, simply answering "yes" to all "Do you see a ..." questions will obtain approximately 90% of accuracy on the VQA v1.0 dataset. In both cases, the models mainly focus on the question more than visual content, which is somewhat similar with "cheating" and fails to produce a good generalization for VQA models.

Hence, the VQA v2.0 dataset [11] collects another set of question-answer pair to ensure for each question $Q$ it has two different answers $A, A'$ for two different images $I, I'$ respectively (see Figure 4), while keeping the same problem definition as the VQA v1.0 dataset. In this way, it forces the VQA models to focus on the visual information rather than the spurious correlation existing in the questions (i.e., language prior), and all the state-of-the-art VQA models around the birth of the VQA v2.0 dataset perform significantly worse than on the VQA v1.0 dataset, which proves that they've indeed learned to exploit language prior.

### 4.3 VQA-CP v2.0

To further produce a good generalization, a new dataset VQA-CP v2.0 (i.e., VQA under Changing Priors) proposed by Agrawal et al. [1] has different distributions of answers for each question type across training and testing set (see Figure 5). The authors also point out that the performance of the most promising VQA models significantly degrades compared to the VQA v2.0 dataset. The VQA-CP v2.0 is essentially the same as VQA v2.0, with a purposely different split of the dataset so that the distribution of answers per question type ("how many", "what color is", etc) are other from the testing set (i.e., introducing a distribution shift). The intuition behind this dataset is to expect the model to answer the questions with the right reason (i.e., visual knowledge instead of language prior) to recognize "black" color on the testing set. In contrast, the most popular answer in the training set for color is "white".


Figure 5: An illustration of how a model biased toward language prior makes wrong prediction on VQA-CP v2.0 dataset.

# 5 Models

## 5.1 Inspired by Association – The First Level of Pearl's Ladder of Causation

### 5.1.1 ViLBERT

Performing the out-of-domain pretraining then transferring the learned knowledge into downstream tasks like VQA with unknown data distribution can lead to several issues such as spurious correlation between objects in the image and questions. Specifically, the conditional probability of one token (visual object or word) given another one can be relatively high without strong causation. ViLBERT[16] used the traditional association-based learning for pretrain-then-transfer approach, which would definitely lead to the spurious correlation issue as illustrated above.

## 5.2 Inspired by Intervention – The Second Level of Pearl's Ladder of Causation

### 5.2.1 Visual Commonsense R-CNN

The correlation does not totally equal to sense-making due to observational bias issue, which means we cannot directly reach sense-making by correlations of tokens (visual object or word). Meanwhile, spurious correlations can be created by the confounder as well. Thus VC R-CNN [24] tries to achieve sense-making by causal intervention to solve these problems.

For sense-making, we aim to solve the problem that how likely would the occurrence of object A causes the occurrence of Object B, which is to learn the visual commonsense about A causing B. The confounder between A and B can lead to a spurious correlation, which can be any objects in the image context when dealing with the causal inference. Therefore, VC R-CNN [24] applies a causal intervention to eliminate the effects of confounder by Do-calculus. The key insight of the causal intervention is to make a graph surgery that cut off the causal link between X and Z, see the Figure 2 for reference. In this way, the Bayes rule is applied on the new causal graph.

### 5.2.2 DeVLBert

Many frameworks use likelihood-based methods to solve out-of-domain visio-liguistic pretraining problems where the pretrained data distribution is different from the downstream task datasets. However, these kinds of likelihood-based methods could lead to spurious correlation issue as illustrated above. Hence, DeVLBert [28] performs intervention-based learning to tend to solve this problem. The crucial difference between traditional associated-based learning with causal intervention-based learning is that the intervention can block the path from $z \rightarrow X$ to eliminate the spurious correlation, which can be illustrated by the Figure 6. In this way, the condition $X$ can be controlled, and the traditional association-based pretraining becomes the causal intervention-based pretraining. Therefore, the model will achieve better causal inference for visio-linguistic representations that can be used for downstream tasks with unknown data distribution. Comparing to VC R-CNN, they are both intervention-based learning, but VC R-CNN only concerns intervention for visual domain while DeVLBert can deal with the spurious correlations between vision and language, which is essential for cross-model downstream tasks like VQA.

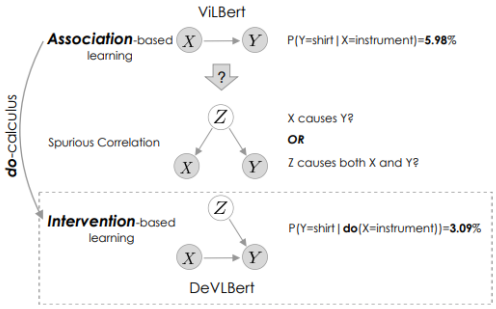


Figure 6: Causal Intervention of DeVLBert comparing to traditional association-based learning [28]. X, Y refers to visual objects while Z refers to language words.

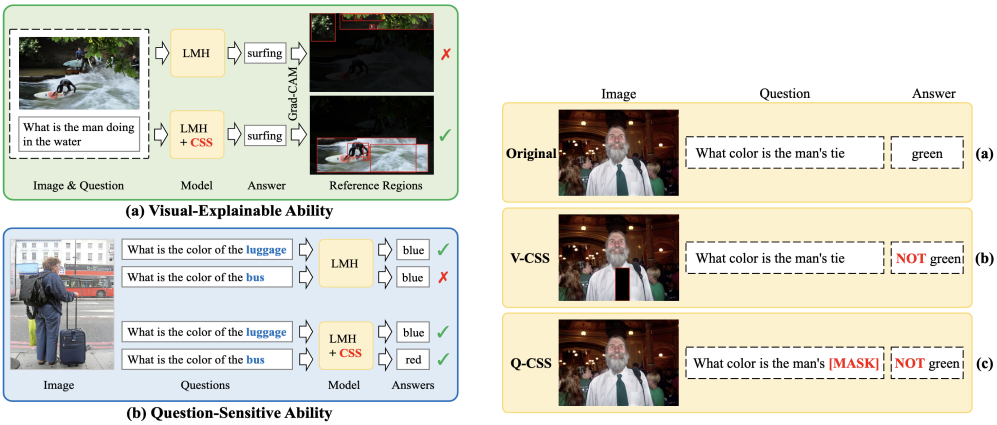


Figure 7: The visualization of visual-explainable and question-sensitive ability (left), and the synthesizing of counterfactual samples during training (right).

### 5.3 Inspired by Counterfactual – The Third Level of Pearl's Ladder of Causation

#### 5.3.1 CSS-VQA

Counterfactual Samples Synthesizing (CSS) is a novel training scheme introduced by Chen et al. [8], and it's aiming to improve the both visual-explainable and question-sensitive aspects (see Figure 7 left) of VQA models by training with counterfactual samples inspired by causality. On the right side of Figure 7, V-CSS masks out the critical objects (e.g., tie in this figure) to answer a certain question and assigns a different answer than the ground truth, while Q-CSS masks out the critical words. Both "critical" here mean these objects or words are important for the model to answer the question. This is similar to ask differently as the following:
**Conventional VQA:** *What will answers $A$ be, if machine hears question $Q$, sees image $V$?*
**V-CSS:** *What will answer $A$ be, if machine hears the critical words in $Q$, but had not seen the critical objects in the image $V$?*
**Q-CSS:** *What will answer $A$ be, if machine sees the critical objects in $V$, but had not heard the critical words in $Q$?*
After adopting the new training scheme on both original samples and synthesized samples, one of the current VQA models LMH [9] achieves a record-breaking state-of-the-art performance on the VQA-CP v2.0 dataset. Notably, since the training scheme serves as a plug-and-play component, it can be applied to many existing VQA models.

#### 5.3.2 CF-VQA

Counterfactual VQA (CF-VQA) [17] tackles the language prior issue from a causal inference perspective, and it endeavors to make unbiased inference even under biased training, which is often

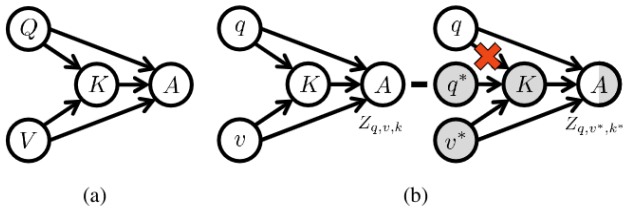


Figure 8: The causal graphs of conventional VQA (a) and counterfactual VQA (b).

unavoidable in most cases. Niu et al. argue that conventional VQA fails to disentangle the single-modal linguistic correlation and multi-modal reasoning (i.e., direct and indirect effects), so they design a causal graph for inferencing, which uses the total causal effect to reduce the direct causal effect caused the language bias. The total causal effect can be obtained while both $Q$ and $V$ are available, and hence multi-modal knowledge $K$ is assessed (see Figure 8 (a)). On the other hand, the direct causal effect of language bias can be assessed by blocking $K$ under the no-treatment condition (see Figure 8 (b)); therefore, the model only depends on the single-modal effect from hearing question $q$. A running example can be found in Appendix B. Since CF-VQA remains as a inference framework, it yields the potential to be unified with other methods for future research.

# 6 Experiments

## 6.1 Experimental Setup

Since we were conducting a survey, we attempted to reproduce the results of these models in Section 5 by using the publicly available code (see Appendix A) on the VQA v2.0 and VQA-CP v2.0 datasets, and we were running our experiments on Nvidia Tesla K40 and V100 GPUs funded by Minnesota Supercomputing Institute (MSI), and granted access by CSCI 5525 instruction team.

## 6.2 Evaluation Metric

The evaluation accuracy is defined as the following:

$$Acc(\text{answer}) = \min\{\frac{\text{\# of human said answer}}{3}, 1\} \tag{3}$$

The intuitive idea is that if the answer produced by the VQA model matches at least 3 of the annotators, it will be assigned the maximum score of 1 to award the production of a popular answer. On the other hand, if the answer doesn't match any of the 10 annotators, it will get a 0 score. It is worth mentioning that some papers also report the sub-level accuracies according to the answer types (e.g., Y/N, number, etc.).

## 6.3 Results

**Limitations** Due to the expensive computing resource (e.g., 8 GPUs) requirement for language model, we did not reproduce the results of ViLBERT and DeVLBert in time, so we directly use the results in the paper.

# 7 Discussion

Causality is originated as a philosophical concept and has been well-studied in the statistic community. People from the machine learning community gradually started to pay attention to causality in the past several years. We collect the causality-related papers from the CV conferences in recent years, and most of them are conceptual convey. The causality papers that have a working solution with modern neural networks to a particular task are concentrating the area of VQA models; hence, that becomes our motivation to conduct this survey.

Table 1: The quantitative comparison across different VQA models on the dataset VQA v2.0 and VQA-CP v2.0 in terms of accuracy.

| Dataset | | VQA v2.0 | | | | VQA-CP v2.0 | | | |
|---|---|---|---|---|---|---|---|---|---|
| Models | Conference | All | Y/N | Num. | Other | All | Y/N | Num. | Other |
| CSS-VQA [8] + LHM [9] | CVPR 2020 | 60.71 | 86.53 | 33.83 | 45.82 | **58.95** | 84.37 | 49.42 | 48.24 |
| CF-VQA [17] + S-MRL [7] | CVPR 2021 | 60.94 | 81.13 | 43.86 | 50.11 | 55.05 | 90.61 | 21.50 | 45.61 |
| VC-RCNN [24] + UP-Down [3] | CVPR 2020 | 68.15 | 84.26 | 48.50 | 58.86 | - | - | - | - |
| VC-RCNN [24] + MCAN [26] | CVPR 2020 | **71.21** | **87.41** | **53.28** | **61.44** | - | - | - | - |
| ViLBERT [16] | NIPS 2019 | 70.60 | - | - | - | - | - | - | - |
| DeVLBert [28] | ACM MM 2020 | 71.10 | - | - | - | - | - | - | - |

The task of VQA has a long-standing problem of language prior, which results of VQA models are not answering the question with a right reason (i.e., multi-modal knowledge) but try to cheat by finding the spurious correlation in language contexts. Even after the distribution-shift dataset VQA-CP v2.0 was released, the language prior issue still remains to some extent. CSS-VQA utilized the counterfactual thinking to generate counterfactual samples during the training state, which is somewhat similar to doing data augmentation, and it obtained a record-breaking No.1 performance of overall accuracy of 58.95% with LHM baseline (see Table 1) on the benchmark VQA-CP v2.0 dataset.

While CSS-VQA applies causality on the training state, CF-VQA employs causality during inference. The reduction of direct causal effect from language bias improves the performance of the existing model in general, and particularly boosts CF-VQA with S-MRL baseline to the third place of VQA-CP v2.0 benchmark with the score of 55.05%. Notably, almost all the existing models on VQA v2.0 would have a performance drop (e.g., as large as 23.74% [8]) while testing on the VQA-CP v2.0 dataset. In our experiment, CSS-VQA only has 1.76% performance drop compared to CF-VQA that has 5.89% performance drop, which implies CSS-VQA has a better generalization ability and is more invariant toward distribution shift.

On the other hand, for VC R-CNN, applying VC feature on traditional bottom-up features truly gain excellent performance on VQA task. Moreover, right now the framework just concatenates VC features with other traditional features directly, there might be better ways to utilize those features like further processing of the bottom-up features of objects themselves, or the processing of VC feature before concatenation.

As for ViLBERT and DeVLBert, DeVLBert obtains a performance boost over the ViLBERT on the VQA task, thus we can see intervention-based architectures can truly outperform the traditional association-based one. Moreover, considering framework of VC R-CNN directly combining commonsense features (intervention-base features) with traditional features (association-based features), they have exactly the same ideology since DeVLBert combines deconfounding features with the knowledge of the downstream task. The competitive results between DeVLBert and VC R-CNN shows cross-modal pretraining and deconfounding are essential for cross-modal downstream tasks.

## 8 Conclusion & Future Work

To conclude, we've shown the causal thinking from the three levels of Pearl's Ladder of Causality can be adopted to help interactive AIs to answer a open-world question based on what they've seen. None of them uses existing causal models such as SCM directly because the raw data (e.g., image, language contexts) is non-trivial to be converted into structural data, but solely applying the concepts from causality bring pleasant performance boost to many existing models. Notably, instead of developing a completely new model or system regardless of its difficulties, we noticed that most causality papers tend to develop a plug-and-play component, which may also be a popular trend in the coming years.

## References


[1] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering, 2018.

[2] A K Akobeng. Understanding randomised controlled trials. *Archives of Disease in Childhood*, 90(8):840–844, 2005.

[3] 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, 2018.

[4] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments, 2018.

[5] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015.

[6] Stanislaw Antol, C. Lawrence Zitnick, and Devi Parikh. Zero-shot learning via visual abstraction. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 401–416, Cham, 2014. Springer International Publishing.

[7] Remi Cadene, Corentin Dancette, Hedi Ben-younes, Matthieu Cord, and Devi Parikh. Rubi: Reducing unimodal biases in visual question answering, 2020.

[8] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering, 2020.

[9] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases, 2019.

[10] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, Jose M. F. Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[11] 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, 2017.

[12] Ruocheng Guo, Lu Cheng, Jundong Li, P. Richard Hahn, and Huan Liu. A Survey of Learning Causality with Data. *ACM Computing Surveys (CSUR)*, 53(4):1–37, 2020.

[13] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257, 1991.

[14] Fredrik D. Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference, 2018.

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.

[16] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, 2019.

[17] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias, 2021.

[18] Nick Pawlowski, Daniel C. Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference, 2020.

[19] Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3(none):96 – 146, 2009.

[20] Judea Pearl. The do-calculus revisited, 2012.

[21] Judea Pearl and Dana Mackenzie. *The Book of Why*. Basic Books, New York, 2018.

[22] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. Toward causal representation learning. *Proceedings of the IEEE*, pages 1–23, 2021.

[23] Scott Thiebes, Sebastian Lins, and Ali Sunyaev. Trustworthy artificial intelligence. *Electronic Markets*, pages 1–18, 2020.

[24] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense r-cnn, 2020.

[25] Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. A survey on causal inference, 2020.

[26] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering, 2019.

[27] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning, 2019.

[28] Shengyu Zhang, Tan Jiang, Tan Wang, Kun Kuang, Zhou Zhao, Jianke Zhu, Jin Yu, Hongxia Yang, and Fei Wu. Devlbert. *Proceedings of the 28th ACM International Conference on Multimedia*, Oct 2020.

[29] C. Lawrence Zitnick and Devi Parikh. Bringing semantics into focus using visual abstraction. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '13, page 3009–3016, USA, 2013. IEEE Computer Society.

## Appendix A: Code

We've made our home folders (luo00042, zou00080) publicly accessible under CSCI 5525 group in MSI. The code from each model is from the following GitHub repositories:

1. VC-RCNN (https://github.com/Wangt-CN/VC-R-CNN)
2. CSS-VQA (https://github.com/yanxinzju/CSS-VQA)
3. CF-VQA (https://github.com/yuleiniu/cfvqa)
4. DeVLBert (https://github.com/shengyuzhang/DeVLBert)
5. ViLBert (https://github.com/jiasenlu/vilbert_beta)

The datasets can be downloaded either from the official website of VQA challenges (https://visualqa.org/) or from the download script located at "/home/csci5525/luo00042/cfvqa/cfvqa/datasets/scripts/"

For CF-VQA and CSS-VQA, the slurm scripts for training and testing are located at "/home/csci5525/luo00042/scripts/" and their corresponding outputs are saved at "/home/csci5525/luo00042/cfvqa/" and "/home/csci5525/luo00042/CSS-VQA/" with the name of format "{$name_of_run}_output.txt".

For ViLBERT and DeVLBERT, the slurm scripts for training and testing are located at "/home/csci5525/zou00080/scripts/".

## Appendix B: Visualization

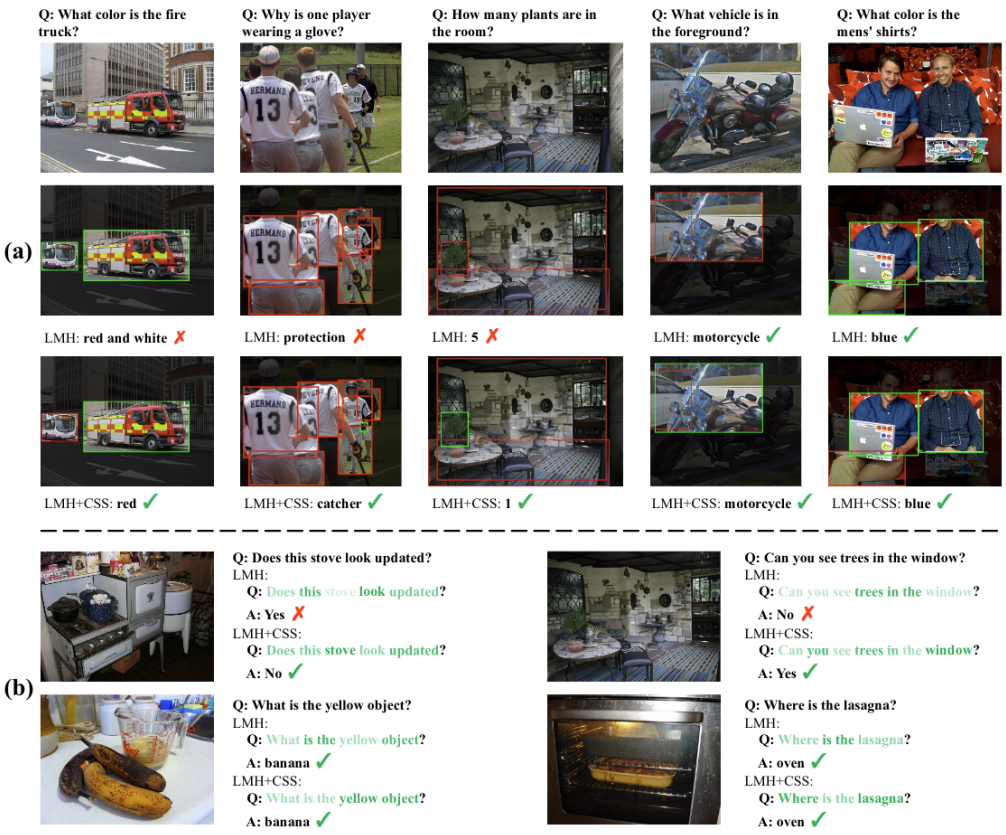


Figure 9: A visualization of visual-explainable ability (top) and question-sensitive ability (bottom).

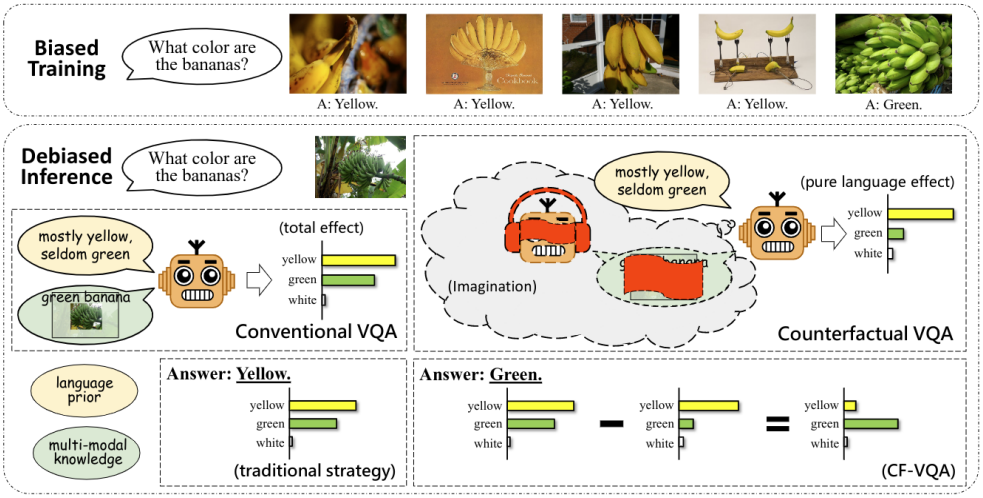


Figure 10: A straightforward visualization of how CF-VQA subtracts the causal effect caused by language bias from the total causal effect to extract the true causal inference, compared to conventional VQA.