Cross Modality Bias in Visual Question Answering: A Causal View with Possible Worlds VQA

Ali Vosoughi[†*], Shijian Deng[‡*], Songyang Zhang[§], Yapeng Tian[‡], Chenliang Xu[§], Jiebo Luo[§,†], *Fellow, IEEE,*

[§]Department of Computer Science, University of Rochester, Rochester, NY 14620
[‡]Department of Computer Science, University of Texas Dallas, Dallas, TX 12345
[†]Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY 14620

***Abstract*—To increase the generalization capability of VQA systems, many recent studies have tried to de-bias spurious language or vision associations that shortcut the question or image to the answer. Despite these efforts, the literature fails to address the confounding effect of vision and language simultaneously. As a result, when they reduce bias learned from one modality, they usually increase bias from the other. In this paper, we first model a confounding effect that causes language and vision bias simultaneously, then propose a counterfactual inference to remove the influence of this effect. The model trained in this strategy can concurrently and efficiently reduce vision and language bias. To the best of our knowledge, this is the first work to reduce biases resulting from confounding effects of vision and language in VQA, leveraging causal explain-away relations. We accompany our method with an explain-away strategy, pushing the accuracy of the questions with numerical answers results compared to existing methods that have been an open problem. The proposed method outperforms the state-of-the-art methods in VQA-CP v2 datasets. Our codes are available at `https://github.com/ali-vosoughi/PW-VQA`.**

***Index Terms*—Visual Question Answering (VQA), bias reduction, language-vision interactions, confounding effects, causal inference.**

I. Introduction

VISUAL Question Answering (VQA) systems, positioned at the confluence of computer vision and natural language processing, are designed to provide natural language responses to queries based on both an image and a question. The effectiveness of these systems is contingent upon their ability to synergize visual and linguistic information, yielding accurate and robust answers pertinent to the images in question. However, despite advancements in integrating language-vision modalities, a prevailing issue in many VQA models is their propensity to shortcut directly from vision or language inputs to answers without fully leveraging the interplay between these two modalities [1], [2], [3]. This tendency, known as vision or language bias, results in an overdependence on one of these modalities and has been extensively investigated in recent research [4], [5], [2], [6], [7], [8].

To address these challenges, VQA models often resort to forming spurious correlations. They may either base their answers solely on the visual modality or directly link questions

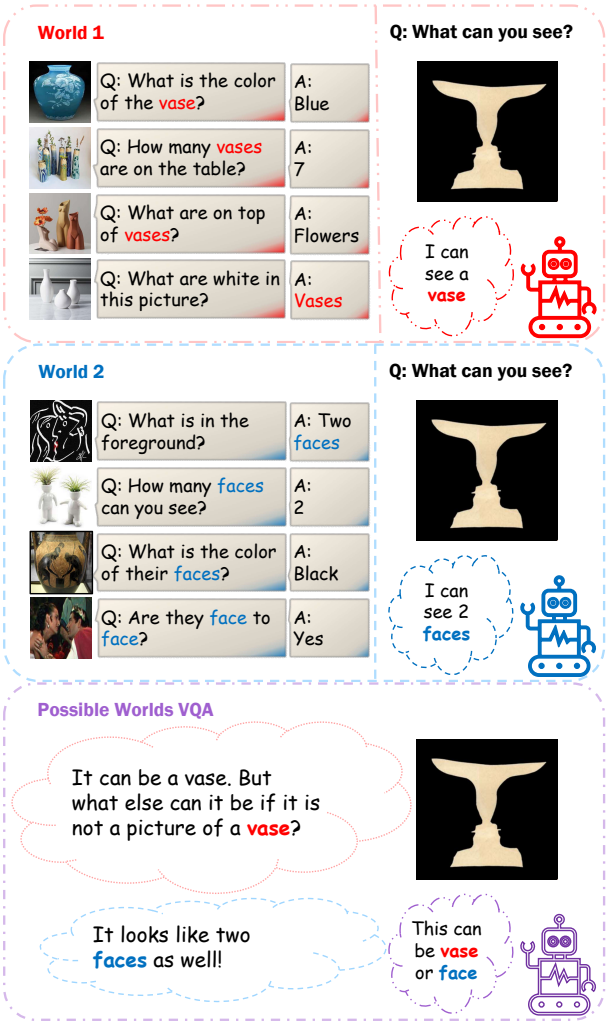


Fig. 1: The visual cognition of annotators, influenced by linguistic factors [9], affects the formulation of questions and selection of corresponding answers. Rubin's vase [10] exemplifies how memory and experience shape an annotator's perception.

* Equal contribution.

to answers, thereby circumventing the necessity for comprehensive multimodal reasoning. A striking instance of this can be found in a VQA dataset where the answer "tennis" to sports-related questions is correct 60% of the time, irrespective of the visual context [11], [7]. This highlights a critical flaw in these models: their tendency to forge direct, albeit superficial, connections between questions and answers, sidestepping an in-depth multimodal interrogation. A significant development is the VQA-CP dataset [12], created in response to these identified shortcuts in earlier VQA datasets [11]. The VQA-CP dataset aims to evaluate the generalizability of VQA models under shifting prior conditions. If a model relies on shortcuts, it is expected to falter when faced with altered distributions between training and test sets.

Approaches to tackle the VQA-CP challenge can be categorized into four primary methodologies. The first category comprises methods targeting language-only biases, modifying the language module, or utilizing a form of language prior to suppress or control language-centric shortcuts. Examples include isolating question-only branches or incorporating a language prior that can be subtracted or masked within the model [5], [2], [6]. The second category includes methods that focus on mitigating vision-only bias. These methods aim to reduce visual bias or enhance visual attention by incorporating human feedback as new visual input for training, thereby increasing the model's focus on visual data or reducing contextual biases that directly link vision to the answer [13], [14], [15]. However, both these categories often overlook the potential influence of the other modality on the bias.

Another approach involves synthesizing new data to balance and augment the training distribution, a strategy initially introduced by the VQA-CP dataset. Most methods in this category employ generative models to synthesize and augment visual and linguistic data, striving for balanced distributions [16], [17], [18], [19], [20]. Nevertheless, these methods sometimes neglect the core intent of the VQA-CP dataset: to challenge models with imbalanced training and test distributions, thereby assessing their generalization capabilities.

Lastly, we introduce a category of methods that aim to concurrently mitigate both language and vision biases, considering the dual modalities of vision and language for robust multimodal inference. This approach recognizes the intertwined nature of these biases and seeks a more holistic solution that addresses the complexities of VQA systems.

Causal inference, resonating with contemporary research addressing biases in VQA, has catalyzed diverse studies across multiple computer vision domains. These include visual explanations, scene graph generation, image recognition, and various learning paradigms like zero-shot, few-shot, and incremental learning, along with representation learning, semantic segmentation, and vision-language tasks [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [16], [31], [32], [33], [34]. Counterfactual learning, in particular, has emerged as a significant influence in recent VQA research [16], [31], [17].

This shift towards causal reasoning underscores its importance in developing more sophisticated and unbiased VQA systems. Exploring counterfactual scenarios is especially promising for reducing biases and enhancing the reliability of VQA models.

Drawing inspiration from the inherent biases in VQA datasets, we propose that these biases may be rooted in the design of the models themselves. For instance, the personal biases of data collectors, such as a preference for a specific sport like tennis, could lead to a disproportionate representation of related data. This, coupled with their experience and cultural background, influences how they annotate the data [9], [10], [35], [36]. Visual perception, a complex process influenced by memory and cultural factors, can affect an annotator's decision-making process, as illustrated by the classic example of Rubin's vase (1916), shown in Figure 1. This raises a question: Could machines also develop different perceptions based on how they are trained? The resulting biases in preference and perception potentially confound VQA datasets and affect their generalizability.

We further elucidate the phenomenon of collider bias within the paradigm of vision-language interplay. The term *collider bias* refers to the situation where the existence of one variable makes another variable appear insignificant, a concept deeply investigated in the field of causal inference literature [37]. To illustrate, consider the scenario where a person is identified as a celebrity based on their wealth or physical attractiveness. Once the person's wealth is known, the importance of their physical attractiveness in determining their celebrity status is reduced. In the realm of VQA, a strong correlation between a question and its answer can lessen the importance of visual data in formulating the answer, thereby introducing a bias.

In response to the limitations of current models like CF-VQA, which predominantly address language bias while often neglecting visual information, we introduce a novel system, Possible Worlds VQA (PW-VQA). Our approach stands in contrast to existing methods by concurrently addressing spurious correlations in both vision and language. PW-VQA employs a causal method to mitigate the confounding effects inherent in these two modalities. This model not only tackles the bias issue but also recognizes that biases often stem from datasets collected with inherent prejudices. By excluding biases originating from cultural influences in data collection, PW-VQA strives to remove biases in both vision and language, thereby enriching the multimodal capabilities of the system. Empirical evidence demonstrates that after training to remove these biases, PW-VQA exhibits significantly reduced bias in either language or vision modalities during testing. Notably, PW-VQA has shown remarkable performance improvements, especially in answering numerical questions, a challenge that has stymied previous models.

Our contributions are as follows. 1) We propose a causal graph separating the problem into two sub-graphs of anticausal learning and an explain-away network. We simultaneously model the visual and linguistic biases through the explain-away network to distinguish between bad and good language and vision biases. We model the experience bias of the annotator as an unobserved confounder that influences the choice of question and answer pairs. 2) We propose a counterfactual approach to reduce these bad biases while keeping the good ones. To the best of our knowledge, our work is the first to propose a causal method to simultaneously alleviate language

and vision biases. 3) We double the accuracy of the numerical questions, which has been an open question recently [7].

II. Related Work

In this section, we cover existing literature on VQA, focusing on the various approaches developed to enhance multimodal VQA models and the diverse strategies employed to identify and mitigate biases.

A. Linguistic Bias in VQA

Linguistic bias in VQA, where language inputs in questions lead to answers via shortcut learning, has garnered significant attention. Key approaches in this area involve data augmentation strategies [38], [39], [40], [16], [17], [18], [19]. Wen et al. [41] introduced a novel technique, DDG, designed to diminish biases in VQA models by generating and employing both positive and negative image-question pairs for training. Other methods include masking language inputs to prevent reliance on textual shortcuts [5], [2], [6] and fortifying the vision component in VQA systems [13], [14]. Additionally, a test-time adaptation method presented in [42] actively identifies and mitigates biases in the test data.

An emerging trend in combating language bias involves direct modifications at the model level, as opposed to data-centric augmentations. This approach, which aims to structurally reduce bias, is extensively discussed in works by Niu et al. [7], Goyal et al. [40], Agrawal et al. [12], and Nguyen et al. [43]. These model-level interventions offer a promising direction for creating inherently unbiased VQA systems.

B. Visual Bias in VQA

The interaction between different modes, like vision and language, or within visual modes across different frames or resolutions, has always been a significant issue [44], [45], [46]. However, exploration of visual bias in VQA systems represents a significant paradigm shift in the field, especially given the historical predominance of language as the primary source of bias. Recent studies have increasingly recognized the critical role of visual biases, which have been somewhat underrepresented in past research [8]. These biases emerge from VQA systems forming simplistic correlations directly from visual contextual cues to the answers, bypassing the need for deeper analytical processing [47]. This trend underscores the necessity of addressing biases that arise from the learning of specific visual elements, such as color and pixel patterns, or the overall context of the image. Moreover, there is a growing emphasis on the importance of accurately focusing on relevant parts of an image, a key area where biases can manifest [47].

Cho et al. [48] put forward a groundbreaking method utilizing Generative Adversarial Networks (GANs). Their approach is centered on learning the bias distributions in a target VQA model's data, with the aim of using this insight to train the model for greater resilience against such biases. The study by Liu et al. [49] tackled compositional generalization by creating a framework that enhances VQA performance through dense interactions within and between modalities. Jing et al. [50] suggested a dialog-like reasoning that merges reasoning for sub-questions into the main process, using consistency rules to ensure logical answer predictions [50]. Li et al. [51] investigated the impact of primitives for compositional generalization in vision-and-language tasks. They introduced a self-supervised learning framework that provides vision-and-language methods with semantic consistency and stability, proving its effectiveness in tasks like video grounding over time and visual question answering [51]. Additionally, Zhang et al. [52] introduced a causality-based multimodal interaction enhancement strategy tailored for multiple-choice VQA scenarios. This method specifically aims to mitigate the vision-answer bias, a critical and often overlooked component of visual bias in VQA systems.

C. Memory Bias in VQA

Memory bias in VQA underscores the impact of human culture, experience, and belief systems on the perception of images and language. The way annotators, influenced by their individual backgrounds, perceive images and interpret questions can lead to significant biases in dataset formulation [7], [12]. Such biases imply that identical images might evoke diverse questions and answers, further complicated by the intricacies of visual memory bias. Recent studies like those by Liu et al. [35] and Zhang et al. [36] have delved into this issue, highlighting that visual memory bias can be shaped by various factors, including the language, location, time, and cultural context of annotators.

A notable aspect of memory bias is its geographical and cultural slant, predominantly favoring North American and Western perspectives [35]. Current initiatives are focused on diversifying datasets to encompass a broader range of cultures and languages, thereby countering this skewness. Memory bias is not only about the diversity of content but also how the same visual stimulus can be interpreted differently. This is exemplified in Figure 1, showcasing a visual illusion where a single image can be perceived in multiple ways, such as seeing a vase or two faces. These kinds of visual illusions and memory biases introduce considerable variability in the annotation process of VQA datasets, leading to a range of questions and answers for the same image.

D. Causality in VQA

Causality-inspired methods in VQA have increasingly employed counterfactual and interventional techniques, focusing on vision-language tasks. These methods have shown remarkable efficacy in both data generation and model-based interventions, substantially enhancing the capabilities of existing VQA models [7], [1], [21], [53], [22], [23]. A significant part of these efforts involves the synthesis of data pairs to equilibrate training datasets across various models [16], [31], [17].

In the forefront of this domain, Zhang et al. [52] introduced a causality-based multimodal interaction enhancement method specifically tailored for multiple-choice VQA scenarios, aiming to reduce vision-answer biases. CopVQA [43] represents another stride in the field, enhancing VQA generalization by

delineating causal reasoning pathways between multimodal inputs and question answering. The application of counterfactual techniques to minimize language biases in VQA is particularly noteworthy [7], [2], [1].

Our approach diverges from the prevailing trend in the literature, which largely concentrates on de-biasing through data augmentation or reducing language priors. We adopt a novel perspective, exploring language-vision bias through the lens of causality. This approach advocates for a more robust utilization of multimodal information to facilitate unbiased inference, even under conditions where training data is inherently biased. By doing so, we aim to contribute a unique dimension to the realm of causality in VQA, addressing the intricate interplay between language and vision biases more holistically.

III. Motivation and Background

Our method is motivated by Counterfactual VQA, CF-VQA [7], which was motivated by Reducing Unimodal Biases for VQA, RUBi [2]. We review these two methods and their evolution in III-A and III-B and then discuss their limitations in III-C.

A. Reducing Unimodal Biases for VQA

The undirected graph of a RUBi is shown in Figure 2a, with $\{V, Q, K, A, M\}$ as set of nodes, V: image, Q: question, K: multimodal knowledge, A: answer out of a set of answers $\mathcal{A} = \{a\}$, M: question mask. $\mathcal{F}_Q$ is an encoder for questions, and $\mathcal{F}_V$ is for images. Consequently, a multimodal function $\mathcal{F}_{VQ}$ is used to obtain $k = \mathcal{F}_{VQ}(v, q)$. An auxiliary neural network nn_q is trained to classify answers based on only $\{q, a\}$ pairs. Then, the classification head is discarded at inference to obtain the masks $m = \sigma(nn_q(\mathcal{F}_Q(q)))$, where σ is the *sigmoid* function. The masks are then applied to the multimodal classification $k \odot m$ to reduce the language bias.

B. Counterfactual VQA (CF-VQA)

CF-VQA uses counterfactual thinking and causal inference to improve RUBi, by only adding one learnable parameter. The causal graph of CF-VQA is shown in Figure 2b. The graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ is a Directed Acyclic Graph (DAG), where $\mathcal{V} = \{V, Q, K, A\}$ with a set of causal edges such that if $Q \rightarrow K$, then Q is a direct cause of K. Moreover, Q is an indirect cause of A through the *mediator* K, as $Q \rightarrow K \rightarrow A$. The causal edge assumption states that every parent is a direct cause of all its children. The answer a can be defined in a multi-class classifier using logits (score) Z. Therefore, for h as a fusion function, for question q, image v, and multimodal knowledge k, these scores for question-only, multimodal fused and vision-only are:

$$Z_q = \mathcal{F}_Q(q), \quad Z_v = \mathcal{F}_V(v), \quad Z_k = \mathcal{F}_{VQ}(v, q), \\ Z_{q,v,k} = h(Z_q, Z_v, Z_k), \tag{1}$$

Denoting answer score $Z_{q,v,k}$ as:

$$Z_{q,v,k} = Z(Q = q, V = v, K = k), \tag{2}$$

the total effect (TE) of $V = v$ and $Q = q$ on $A = a$, according to [7], is defined as:

$$TE = Z_{q,v,k} - Z_{q^*,v^*,k^*}, \tag{3}$$

where Z_{q^*,v^*,k^*} is answer logits Z for counterfactual question q^*, counterfactual image v^*, and counterfactual multimodal knowledge k^*. The total effect can be decomposed into natural direct effect (NDE) and total indirect effect (TIE):

$$TE = TIE + NDE. \tag{4}$$

NDE for the question-only branch is $Q \rightarrow A$ by comparing Z_{q,v^*,k^*} and Z_{q^*,v^*,k^*}:

$$NDE = Z_{q,v^*,k^*} - Z_{q^*,v^*,k^*}. \tag{5}$$

Finally, using (3), (4), and (5), TIE will be:

$$TIE = Z_{q,v,k} - Z_{q,v^*,k^*}, \tag{6}$$

as shown in Figure 2c. Consequently, the logits $Z_{q,v,k}$ is parametrized as $\mathcal{F}_Q$: $Q \rightarrow A$, and $\mathcal{F}_{VQ}$: $(V, Q) \rightarrow K \rightarrow A$. The question-only and vision-only logits Z_q and Z_v will be as:

$$Z_b = \begin{cases} z_b = \mathcal{F}_B(b) & \text{if } B = b \\ z_b^* = c & \text{if } b = \varnothing \end{cases}, \tag{7}$$

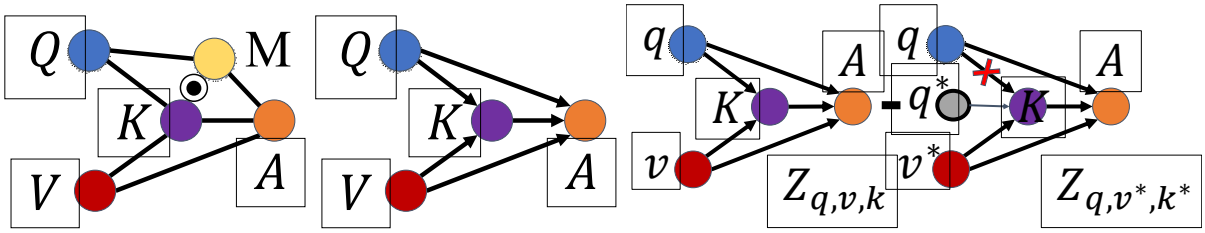
where $B \in \{Q, V\}$, and c as a constant, learnable feature, as described in [7], and z_b^* is a counterfactual realization of Z_b. Furthermore, multimodal knowledge's logit Z_k is defined as:

$$Z_k = \begin{cases} z_k = \mathcal{F}_{VQ}(v, q) & \text{if } V = v \text{ and } Q = q \\ z_k^* = c & \text{if } V = \varnothing \text{ or } Q = \varnothing \end{cases}. \tag{8}$$

C. Limitations of Prior Methods

Addressing Visual and Language Biases in VQA: While the predominance of language biases in VQA systems has been extensively studied, visual bias remains relatively underexplored. Recent investigations reveal that VQA systems may bypass essential visual analysis, relying instead on contextual cues such as color or spatial information [47]. In more comprehensive studies, such as those conducted on UrbanCars and ImageNet-W datasets, the phenomenon of multi-shortcut learning was observed, where models depend on various spurious visual cues, overshadowing primary visual concepts [54]. A critical challenge identified is the inadvertent amplification of one type of bias when attempting to mitigate another [54]. Our work introduces a counterfactual explain-away framework, aiming to concurrently alleviate both language and vision-related biases, thereby enhancing the robustness of VQA models.

Impact of Memory on Perception in VQA: The influence of memory on visual perception is a critical aspect that shapes the interpretation of images in VQA systems [9]. This concept is exemplified through Rubin's face illusion [10], as illustrated in Figure 1. Rubin's theory on memory bias suggests that an individual's past experiences significantly shape their perception of visual information [10], [55]. Studies further indicate that factors like language, geographical context, and temporal aspects can profoundly affect image interpretation [9], [36], [35]. Our approach proposes to consider the experience of the annotator as an unobserved confounder, thereby addressing the issue of experience bias. This method aims to provide a more nuanced and unbiased understanding of visual content in VQA systems.



(a) RUBi (b) CF-VQA graph (c) Total indirect effect

Fig. 2: VQA graphs related to RUBi and CF-VQA are shown. a) In RUBi, question Q and image V are fused through multimodal knowledge K to obtain an answer A, while question-only mask M is applied on K; b) causal graph of CF-VQA is shown, where $Q \rightarrow A$ and $V \rightarrow A$ are vision and language shortcuts, all V, Q, and K are factual; c) output of VQA with counterfactual question $Q = q^*$ and vision $V = v^*$ is subtracted from a regular VQA with factual $V = v$ and $Q = q$.

IV. Possible Worlds VQA (PW-VQA)

In this section, we explain the proposed method in five subsections. First, we simultaneously model the language and vision bias using a causal view. Then we model experience bias as unobserved confounders of the VQA systems. Third, a counterfactual method is proposed in the subsequent subsection to solve these problems. Fourth, we propose a novel strategy to fuse multimodal vision and language information in VQA systems. Finally, we detail the architecture of the proposed method.

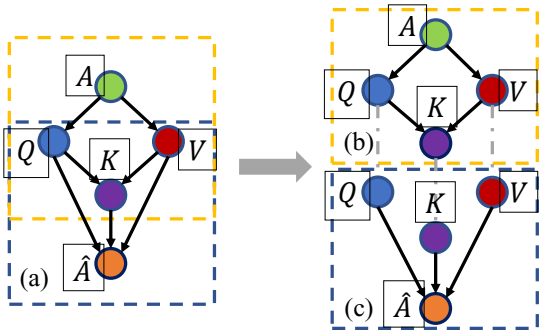


Fig. 3: The proposed causal graph reformulates the VQA problem by stating that a) the answer A is a cause of the question Q, and vision V, and the final estimated answer $\hat{A}$ is achieved by fusing V and Q information. b) The anticausal subgraph consists of the ground-truth answer A that is a cause of the V and Q, which leads to multimodal knowledge K. c) The collider $Q \rightarrow \hat{A} \leftarrow V$ is an explain-away network that models the language-vision bias.

Assume that a multimodal knowledge K contains fused information of question Q and vision V used in a VQA system. We propose the causal graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ with the set of nodes $\mathcal{V} = \{Q, V, K, A, \hat{A}\}$, which is shown in Figure 3a to model VQA systems.

Inspired by the anticausal learning [56], [57], we model the answer A as a cause of both the images V and question Q. Unlike previous works [7], [2], [1], we distinguish between the ground-truth answer A for the training of the VQA model and the estimated answer $\hat{A}$ when the model is used in practice (test). Therefore, as shown in Figure 3b, Q and V have a causal effect on K and are also a child of the answer A.

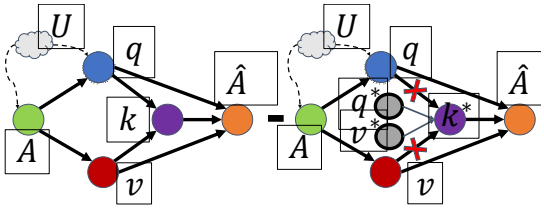


Fig. 4: The multimodal knowledge $K = k^*$ is counterfactual, while Q and V are facts ($Q = q, V = v, K = k^*$), then, the natural indirect effect (NDE) is subtracted from the total effect (TE) to obtain total indirect effect (TIE). The values $V = v$ and $Q = q$ are fact, and $V = v^*$ and $Q = q^*$, which leads to $K = k^*$ are counterfactuals.

A. Collider Confounder in Vision and Language

The relationship $Q \rightarrow A$ creates a spurious correlation between the question Q directly to the answer A. Therefore, the $V \rightarrow K \rightarrow A$ information is ignored. Contrarily the VQA models may shortcut visual information to answer $V \rightarrow K \rightarrow A$ rather than multimodal knowledge [47]. By looking at the subgraph shown in Figure 3c, the explain-away network, or collider bias network simultaneously can model vision and language bias. The relationship $Q \rightarrow \hat{A} \leftarrow V$ is a collider, a primitive graph structure, *aka* explain-away network. Consequently, having a strong connection $Q \rightarrow \hat{A}$ removes the dependency of the $\hat{A}$ on V. Noteworthy that there are useful information and harmful biases in both vision and language. Our explain-away method aims to remove biases but keep good information. Therefore we introduce the collider of $Q \rightarrow \hat{A} \leftarrow V$ as a source of vision-language bias in VQA models.

B. Memory Bias as an Unobserved Confounder in VQA Systems

In the context of VQA systems, our investigation highlights a novel source of confounding bias, predominantly stemming from the annotator's experience during dataset preparation, as illustrated in the proposed causal graph $\mathcal{G}$. This form of bias, termed 'memory bias', is exemplified by the visual illusion challenge depicted in Figure 1. Specifically, this bias emerges when annotators select questions Q and formulate answers A based on a visual input V, with their judgments being

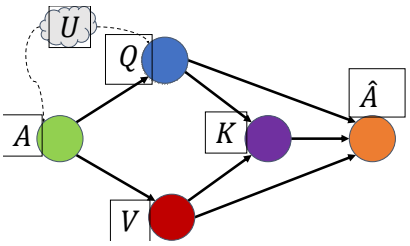


Fig. 5: The causal graph of the VQA where the question Q and the answer A are influenced by unobserved confounder U.

inherently influenced by personal experiences and preferences. This introduces an unobserved bias U, which critically impacts the outcome.

The proposed causal graph for VQA models, incorporating this unobserved confounder, is detailed in Figure 5. To dissect the influence of this bias, we analyze various causal paths leading to the predicted answer $\hat{A}$. These paths include $U \rightarrow A \rightarrow Q \rightarrow K \rightarrow \hat{A}$ and $U \rightarrow A \rightarrow V \rightarrow \hat{A}$, which directly link the unobserved bias through annotator's answers to the visual and question-related elements of the VQA system. Additionally, the path $U \rightarrow A \rightarrow V \rightarrow K \rightarrow \hat{A}$ underscores the compound effect of bias when considering both visual and knowledge-based elements. While paths originating from $U \rightarrow Q$ are also present, our primary focus is on those impacting $K \rightarrow \hat{A}$, as they are most relevant to the predictive mechanisms of VQA models.

C. Explain-Away Fusion Strategy (EA)

We propose the following Explain-Away (EA) fusion function as follows. For parametrization, we use similar notations as [7]. Therefore, the score $Z_{q,v,k}$ which is the feature space of the fusion K, is parametrized as $\mathcal{F}_Q$: $Q \rightarrow \hat{A}$, and $\mathcal{F}_{VQ}$: $(V, Q) \rightarrow K \rightarrow \hat{A}$.

Based on Z_q, Z_v, and Z_k, we define the fusion function as follows:

$$\text{(EA)} \qquad h(Z_q, Z_v, Z_k) = \frac{1}{\alpha + 1} \log(Z_{\text{EA}}), \tag{9}$$

where Z_{EA} is defined as:

$$\begin{aligned} Z_{\text{EA}} =& \sigma(Z_q)^{\alpha} \sigma(Z_v)^{\alpha+1} \sigma(Z_k)^{\alpha+1} \\ &+ \sigma(Z_q)^{\alpha+1} \sigma(Z_v)^{\alpha} \sigma(Z_k)^{\alpha+1} \\ &+ \sigma(Z_q)^{\alpha+1} \sigma(Z_v)^{\alpha+1} \sigma(Z_k)^{\alpha}, \end{aligned} \tag{10}$$

with $\alpha \geq 0$ being a tunable parameter determined through empirical experimentation. The choice of α influences the balance between the contributions of each feature space, thus affecting the model's ability to mitigate biases effectively. Our empirical analysis, discussed later, provides insights into the selection of an optimal value for α that ensures a balanced representation of both visual and linguistic aspects in the VQA process.

D. Unobserved Confounding Bias Reduction

Since the model relies on the fused information K of V and Q, and as shown in Figure 4, the confounding bias of vision-language can be removed by maximizing the total indirect effect (TIE) by subtracting natural direct effect (NDE) of this confounding influence from its total effect (TE) [58]:

$$\begin{aligned} TIE &= TE - NDE \\ &= h(Z_q, Z_v, Z_k) - h(Z_q, Z_v, Z_{k^*}), \end{aligned} \tag{11}$$

where K^* is a counterfactual of K, as described in (8). As the influence of the unobserved confounding bias is subtracted in (11), it will block the influence of the explain-way explain-away of vision-language and experience biases altogether. By blocking the two paths $V \rightarrow K$ and $Q \rightarrow K$, all influences from unobserved confounding bias are blocked.

E. Architecture of the PW-VQA

Here we discuss the PW-VQA framework's architecture, shown in Figure 7. The framework consists of two branches: regular VQA, which can be of any baseline method, and the counterfactual version of the same network and parameters, shown as a biased branch in the figure. Four different losses are simultaneously used during training to formulate the causal relationship between each modality. In addition, a constant c is jointly trained here to capture the natural indirect effect of vision-language confounding biased injected during the annotation process. Finally, in the inference stage, PW-VQA uses the logits of regular VQA subtracted by the biased VQA and gets a debiased answer. The letters a in this figure denote the answer.

Training: For the training of the network, we use vision-only branch $\mathcal{L}_{VA}(v, a)$, question-only branch $\mathcal{L}_{QA}(q, a)$, and multimodal fusion branch $\mathcal{L}_{VQA}(v, q, a)$. As illustrated in Figure 5, given a triplet (v, q, a) where a is the ground-truth answer of image-question pair (v, q), the branches are optimized by minimizing the cross-entropy losses over the scores $Z_{q,v,k}$, Z_q and Z_v: [7]:

$$\mathcal{L}_{cls} = \mathcal{L}_{VQA}(v, q, a) + \mathcal{L}_{QA}(q, a) + \mathcal{L}_{VA}(v, a), \tag{12}$$

where $\mathcal{L}_{VQA}$, $\mathcal{L}_{QA}$ and $\mathcal{L}_{VA}$ are over $Z_{q,v,k}$, Z_q and Z_v. A learnable parameter c in Eq. (7)-(8), which controls the sharpness of the distribution of Z_{q,v^*,k^*} is also included, as the sharpness of NDE should be similar to that of TE [68], [7]. An improper c would lead to the domination of TIE in Eq. (11) by either TE or NDE. Thus, we use KL-divergence to estimate c:

$$\mathcal{L}_{kl} = \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} -p(a|q, v, k) \log p(a|q, v^*, k^*), \tag{13}$$

where $p(a|q, v, k) = \text{softmax}(Z_{q,v,k})$ and $p(a|q, v^*, k^*) = \text{softmax}(Z_{q,v^*,k^*})$. Only c is updated when minimizing $\mathcal{L}_{kl}$. The final loss is the combination of $\mathcal{L}_{cls}$ and $\mathcal{L}_{kl}$:

$$\mathcal{L}_{final} = \sum_{(v,q,a) \in \mathcal{D}} \mathcal{L}_{cls} + \mathcal{L}_{kl} \tag{14}$$

A key question arises regarding the balancing of these two loss components in expression (14). To explore this, we conduct a series of ablation studies, examining the impact of different weightings on the overall performance and learning dynamics of the model. This investigation aims to determine the optimal balance that maximizes the model's effectiveness in handling the complexities of VQA tasks.

TABLE I: The table lists the accuracy values for the most recent studies, especially on both VQA-CP v2 and VQA v2 datasets. We show the best-performing method with bold and the second-best-performing method with an underline. We use a dash for the papers that miss reporting performance values on datasets.

Test set		VQA-CP v2 test				VQA v2 test			
Methods	Base	All	Y/N	Num.	Other	All	Y/N	Num.	Other
GVQA [12]	-	31.30	57.99	13.68	22.14	42.24	72.03	31.17	34.65
SAN [59]	-	24.96	38.35	11.14	21.74	52.41	70.06	39.28	47.84
UpDn [60]	-	39.74	42.27	11.93	46.05	63.48	81.18	42.14	**55.66**
BLIP-1 [61]		34.42	54.97	14.48	35.10	54.59	70.70	40.00	46.21
BLIP-2 [62]	-	34.93	52.21	15.15	36.10	-	-	-	-
S-MRL [2]	-	38.46	42.85	12.81	43.20	63.10	-	-	-
Methods based on modifying language module or using language prior:									
DLR [4]	UpDn	48.87	70.99	18.72	45.57	57.96	76.82	39.33	48.54
VGQE [63]	UpDn	48.75	-	-	-	**64.04**	-	-	-
VGQE [63]	S-MRL	50.11	66.35	27.08	46.77	63.18	-	-	-
AdvReg. [5]	UpDn	41.17	65.49	15.48	35.48	62.75	79.84	42.35	55.16
RUBi [2]	UpDn	44.23	67.05	17.48	39.61	-	-	-	-
RUBi [2]	S-MRL	47.11	68.65	20.28	43.18	61.16	-	-	-
LM [6]	UpDn	48.78	72.78	14.61	45.58	63.26	81.16	42.22	55.22
LM+H [6]	UpDn	52.01	72.58	31.12	46.97	56.35	65.06	37.63	54.69
CF-VQA (SUM) [7]	UpDn	53.55	**91.15**	13.03	44.97	63.54	**82.51**	**43.96**	54.30
CF-VQA (SUM) [7]	S-MRL	55.05	90.61	21.50	45.61	60.94	81.13	43.86	50.11
GGE-DQ-tog [64]	UpDn	57.32	87.04	27.75	**49.59**	59.11	73.27	39.99	54.39
Methods based on reducing visual bias or enhancing visual attention/grounding:									
AttAlign [13]	UpDn	39.37	43.02	11.89	45.00	63.24	80.99	42.55	55.22
HINT [13]	UpDn	46.73	67.27	10.61	45.88	63.38	81.18	42.99	55.56
SCR [14]	UpDn	49.45	72.36	10.93	48.02	62.20	78.80	41.60	54.50
Methods mitigating both language and vision:									
LMH+Fisher [8]	UpDn	54.55	74.03	49.16	45.82	-	-	-	-
PW-VQA (ours)	UpDn	59.06	88.26	52.89	45.45	62.63	81.80	43.90	53.01
PW-VQA (ours)	S-MRL	**60.26**	88.09	**59.13**	45.99	61.25	80.32	43.17	51.53
PW-VQA (ours)	BLIP-1	49.53	84.36	45.38	33.24	45.56	61.48	27.39	38.42
PW-VQA (ours)	BLIP-2	45.84	85.17	19.16	32.73	-	-	-	-
Methods that synthesize data to augment and balance training splits:									
CVL [17]	UpDn	42.12	45.72	12.45	48.34	-	-	-	-
Unshuffling [18]	UpDn	42.39	47.72	14.43	47.24	68.08	78.32	42.16	52.81
RandImg [65]	UpDn	55.37	83.89	41.60	44.20	57.24	76.53	33.87	48.57
SSL [19]	UpDn	57.59	86.53	29.87	50.03	63.73	-	-	-
CSS [16]	UpDn	58.95	84.37	49.42	48.21	59.91	73.25	39.77	55.11
CSS+CL [66]	UpDn	59.18	86.99	49.89	47.16	57.29	67.27	38.40	54.71
Mutant [67]	UpDn	61.72	88.90	49.68	50.78	62.56	82.07	42.52	53.28
LMH+ECD [20]	UpDn	59.92	83.23	52.59	49.71	57.38	69.06	35.74	54.25

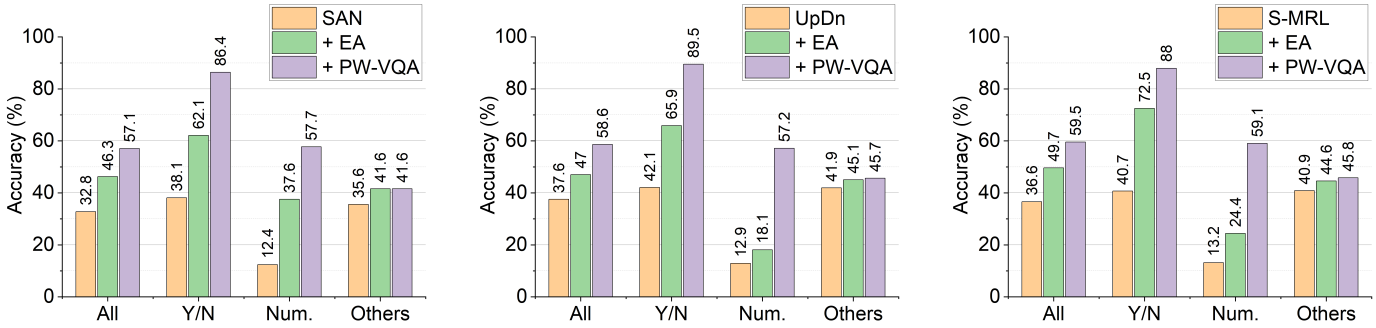


Fig. 6: The plots show the backbones using our proposed causal framework (PW-VQA) and fusion strategy (EA). The results are consistently improving for all three different backbones, namely, SAN [59], S-MRL [2], and UpDn [60].

Inference. For the inference, we use the debiased causal effect for inference, which is implemented as:

$$\begin{aligned} TIE = TE - NDE &= Z_{q,v,k} - Z_{q,v^*,k^*} \\ &= h(z_q, z_v, z_k) - h(z_q, z_v^*, z_k^*). \end{aligned} \tag{15}$$

V. Experiments

A. Experimental Setup

In our study, we employed a robust experimental framework to validate the efficacy of our VQA model. The experiments were conducted using the VQA-CP v2 dataset [12], and VQA dataset [11], which comprises approximately 438K training and 220K test questions, all in English. This dataset was chosen for its comprehensive coverage of question-answer pairs, providing a suitable benchmark for our model.

Model Backbones: Our VQA model was applied on three core backbones: Stacked Attention Network (SAN)[59], Bottom-up and Top-down Attention (UpDn)[60], Simplified MUREL (S-MRL)[3], [2], BLIP-1[61], and BLIP-2 [62] net-

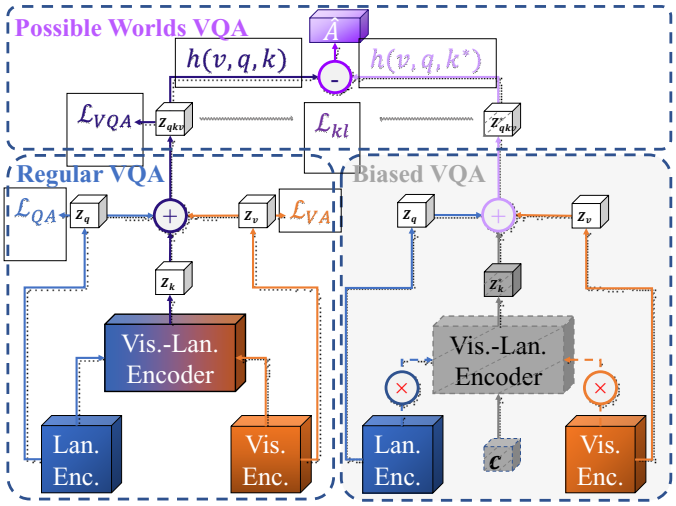


Fig. 7: Network architecture of the proposed PW-VQA framework.

works. We further explore generative model of BLIP-1, pre-trained with web data, excels in zero-shot performance for video language tasks, while BLIP-2 leverages a lightweight Q-Former for multimodal representation alignment. For implementation of both BLIP-1 and BLIP-2 models, we use their feature extractor for image and text, and use an attention layer to obtain multimodal features. Finally, we concatenate these features with question and image features extracted by BLIP encoders. We then use this joint feature as a logit for the classification layer.

Hardware Specifications: The experiments utilized NVIDIA GeForce GTX 1080 Ti, RTX A6000, A100, and RTX 4090 GPUs based on availability on our local servers. This choice of hardware demonstrates the model's adaptability to different computational environments. It is noteworthy that the model is operable even on a singular GeForce GTX 1080 GPU.

Training and Validation: Training durations varied based on the backbone used: approximately 8 hours for SAN and an average of 3 hours for both S-MRL and UpDn. The training for BLIP models takes longer and is in the range of 12 hours. Validation on the test split of the dataset was completed in about 10 minutes. The model's performance was evaluated using accuracy as the primary metric.

Hyperparameter Optimization: A manual search was conducted to identify optimal hyperparameters. Our empirical findings indicated that the model does not converge for $\alpha < 1$, leading to the establishment of $\alpha \geq 1$ as a lower bound. The value of $\alpha = 1.5$ was identified as optimal, resulting in the best performance across multiple trials.

Training Convergence: The model consistently converged to stable results within 22 epochs, negating the need for additional training beyond this point. We utilized a batch size of 256 for all experiments. Through this rigorous experimental setup, we aimed to ensure the robustness and reliability of our proposed VQA model, adequately addressing the challenges posed by biases in both vision and language modalities.

B. Quantitative Results

To compare our method with the available literature on the benchmark datasets, we list the performance values in Table. I. Then, to compare reasonably with the existing methods, we divide them into four categories: 1) Methods like DLR [4], VGQE [63]), AdvReg [5], RUBi [2], LM [6], LM+H [6], CF-VQA[7], GGE-DQ-tog [64] modify language modules or use language before suppress, control, or mask language shortcuts. However, these methods only consider spurious language correlations and neglect vision in their schema. 2) Some approaches, such as AttAlign [13], HINT [13], SCR [14] mitigate visual biases by loosening contextual ties to the answer or improving visual grounding and attention via human feedback, de-coupling shortcuts that couple vision to answer. 3) Other approaches like LMH+Fisher [8] mitigate both language and vision bias together, attempting to balance two modalities of vision and language for robust multimodal inference. Our proposed method here is in this class. 4) Methods such as CVL [17], Unshuffling [18], RandImg [65], SSL [19], Mutant [67], CSS [16], CSS+CL [66], LMH+ECD [20] synthesize samples and augment data to balance training and test sets. Some of these methods also have higher accuracy than counterparts, for instance Mutant in I; however, since these methods are based on balancing distributions and violate the main idea of the VQA-CP v2 dataset, it is not fair to compare them with our method. We include them in our results for inclusiveness.

As listed in Table. I, our method outperforms most of the competing methods on the benchmark datasets, especially in numerical questions, which was introduced as an open problem recently [7]. Moreover, the results indicate that our method improves the accuracy of all the S-MRL, UpDn, BLIP-1, and BLIP-2 backbones, demonstrating that they are generalizable to all of these architectures. Noteworthy to mention that there are higher accuracy values for methods that augment data. In contrast, these methods are not comparable to ours as they do not obey the main idea of the VQA-CP v2 dataset, conducting unbiased inference under biased training. Simulation results of our proposed EA fusion strategy and the PW-VQA are shown in Figure 6. Both the EA fusion strategy and PW-VQA framework increase the accuracy of all question types. Particularly, the accuracy of numerical results with SAN as backbone increases from 12.4 to 37.6 by adding EA fusion and further increases to 57.7 by adding the PW-VQA framework. Furthermore, the improvements are consistent for all backbones, including SAN, S-MRL, and UpDn. More improvements can be achieved using large pre-trained language-vision open-ended generative models. In our experiments, we employed the generative BLIP decoder [61] and CLIP encoders [69] as an exemplary model, which we will discuss later.

C. Qualitative Results

In our qualitative analysis, we conducted simulations on the VQA-CP v2 dataset to compare the performance of our method against CF-VQA and traditional VQA systems. Selected examples from this dataset are illustrated in Figure 8.

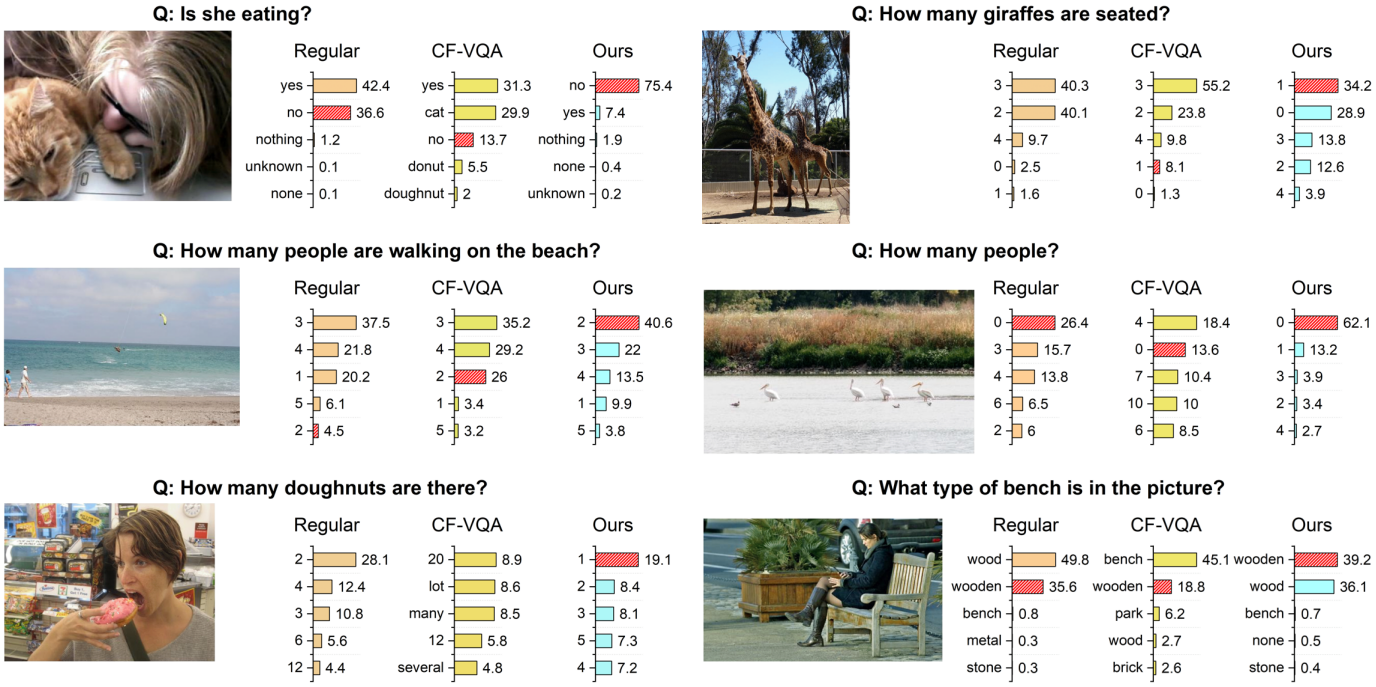


Fig. 8: Bar plot comparison of answer correctness probabilities in the VQA-CP v2 test split: Regular VQA, CF-VQA [7], and our method. Each bar represents the probability (out of 100%) that a given method's answer is correct, with the red bars, characterized by a sparse pattern, indicating ground-truth values.

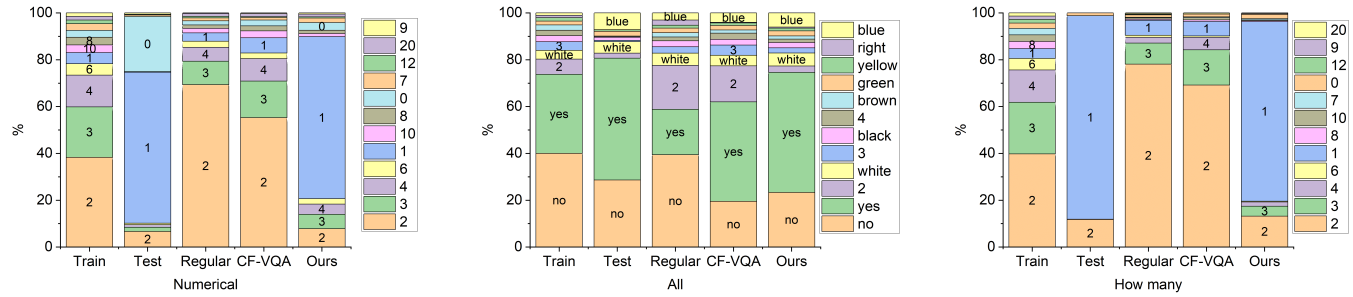


Fig. 9: The distributions of the train, test sets, the previous methods, namely regular VQA, CF-VQA [7], and our proposed method are shown. Note that there is a subtle difference between "how many" questions versus questions with "numerical" answers, which is related to the difference between recognizing "numerical" rather than counting "how many".

These examples demonstrate our method's reduced susceptibility to biases inherent in language and vision. For instance, when presented with an image of a cat and a woman and the question "Is she eating?", our method accurately denies the action, whereas traditional VQA incorrectly affirms it. Notably, CF-VQA displays a pronounced bias towards the more visually salient object (the cat) in the image, erroneously responding with high confidence to the unrelated answer "cat".

Furthermore, both traditional VQA and CF-VQA fail to effectively utilize key information from questions, leading to biased responses influenced by prominent visual elements like foreground animals.

A critical factor in our method's enhanced performance is the integration of the Explain-Away (EA) fusion mechanism. EA fusion is pivotal in our approach, ensuring a more balanced consideration of the image, the textual content of the question, and the multimodal knowledge K. This balanced approach, distinct from existing methods, allows for a harmonious integration of visual, textual, and contextual information.

Through the EA fusion, our model processes and integrates visual cues from images, textual semantics, and contextual understanding from shared knowledge K. This comprehensive approach ensures that each component receives equal attention, avoiding overreliance on any single information source. The significant impact of EA fusion is evident in our ablation studies, which show a notable enhancement in the model's performance with its inclusion, particularly in the accuracy of numerical responses.

Finally, Figure 9 presents distributions of numerical answers across training, testing, traditional VQA, CF-VQA, and our

model. This visualization highlights that while CF-VQA inherits significant biases for numerical queries from the training dataset, our model mitigates these biases. Our approach, through de-confounding causal inference, achieves an answer distribution that more closely mirrors the test dataset, thereby reducing both language and vision biases.

D. Ablation Studies

1) Evaluating Zero-Shot Performance with CLIP-BLIP Model: This study employs the Contrastive Language-Image Pretraining (CLIP) model, a system trained on extensive image-text pair datasets using a contrastive loss mechanism [69]. We integrate CLIP as a baseline in our ablation study, focusing on large pretrained vision-language models. CLIP serves as both a text and image encoder in our framework, operating without fine-tuning. We process (v, q, a) tuples—comprising images (v), questions (q), and answers (a)—by tokenizing and encoding the text components (questions and answers) with CLIP, followed by normalization. Similarly, image features are embedded via CLIP's image encoder and normalized. The model architecture further incorporates Multilayer Perceptron (MLP) layers to generate vision-only and language-only logits, which are instrumental in our PW-VQA method. Fusion of image and language features is executed through concatenation, followed by MLP-based processing for vision-language logits derivation.

For the inference phase, we utilize the Bootstrapping Language-Image Pretraining for Unified Vision-Language Understanding and Generation (BLIP) model [61]. BLIP's generative capabilities, tailored for open-ended visual question answering, complement our CLIP-based approach. To ensure the integrity of question-only logits, we introduce zero tensors as image inputs within the BLIP framework, thereby preventing unintended image feature influence. Subsequent processing involves normalizing and integrating CLIP-generated features into our PW-VQA system.

Table. II details the performance metrics of the BLIP-CLIP model alongside other evaluated models. Notably, BLIP-CLIP demonstrates superior accuracy; however, it's important to contextualize this finding, considering BLIP and CLIP's extensive training on vast, high-quality image-text pair datasets. Figure 10 illustrates the implemented CLIP-BLIP model architecture. In our analysis, listed in Table. II, our PW-VQA method demonstrated comparable performance to the CLIP-BLIP model. This observation prompts a critical discussion about the metrics and criteria used for comparison. Given that the BLIP model has been extensively trained on a vast dataset comprising over 400 million image-text pairs, its scalability and robustness in various scenarios differ significantly from our other backbones that we used for PW-VQA approach. This disparity in training data volume and diversity raises questions about the direct comparability of these models in the context of open-ended visual question answering.

Further, the intensive training regime of BLIP could potentially mask subtleties in model performance, particularly in nuanced or less represented scenarios within its training dataset. Therefore, a more in-depth analysis is necessary

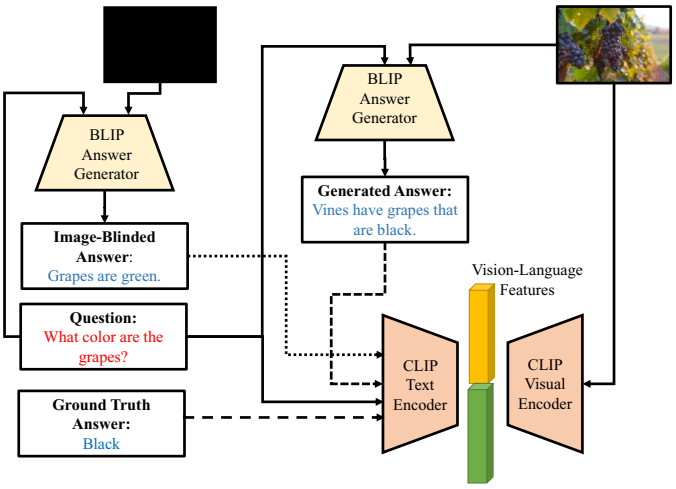


Fig. 10: The architecture of CLIP-BLIP network as a baseline for PW-VQA. We combine generative pretrained BLIP [61] with encoding of the CLIP [69] as a transformer-based large vision-language model.

to understand the interaction and efficacy of the PW-VQA method when applied to the pretrained CLIP-BLIP framework. Future research should focus on conducting comprehensive studies, possibly involving diverse and challenging datasets, to rigorously evaluate the effectiveness of PW-VQA in enhancing the generative capabilities of models like CLIP-BLIP. Such studies should aim to dissect the models' performance in various scenarios, especially where training data may not have provided sufficient representation or depth. This approach will enable a more nuanced understanding of the strengths and limitations of applying causal methods like PW-VQA to large-scale, pretrained models in the domain of visual question answering.

2) Stabilizing Logarithmic Computations in Fusion Equations: During our examination, we identified that the inclusion of a small constant, denoted as ϵ, is crucial for maintaining consistency in the logarithmic operations integral to our fusion equations. This stabilization technique addresses potential computational instability, particularly in scenarios involving values approaching zero. To empirically determine the optimal value of ϵ, a range of constants were evaluated, and their effects were systematically analyzed. Our experimental results, summarized in Table. III, illustrate the impact of varying ϵ on the model's performance during both training and testing phases. Additionally, Figure 11 visually represents these findings, providing clear insights into the stabilizing effect of different ϵ values.

3) Categorized Improvements on SAN, UpDn, and SMRL Baselines: The plots in Figure 12 show performance metrics for different methods as baseline and percent of improvements compared to baseline on each class of question types when using our proposed method, PW-VQA. In all of these simulations, $\alpha = 1.4$ are set. As seen in Figure 12, PW-VQA consistently improves the existing method, confirming the generalizability of the method to several existing methods.

TABLE II: The table lists the accuracy values for different backbones based on VQA-CP v2 and VQA v2 datasets. We use different backbones, UpDn, S-MRL, and CLIP-BLIP, to show the effect of the backbone on the accuracy. We show the best-performing method with bold and the second-best-performing method with an underline.

Test set		VQA-CP v2 test				VQA v2 test			
Methods	Base	All	Y/N	Num.	Other	All	Y/N	Num.	Other
PW-VQA (ours)	UpDn	59.06	<u>88.26</u>	52.89	45.45	<u>62.63</u>	<u>81.80</u>	<u>43.90</u>	<u>53.01</u>
PW-VQA (ours)	S-MRL	<u>60.26</u>	88.09	<u>59.13</u>	<u>45.99</u>	61.25	80.32	43.17	51.53
PW-VQA (ours)	CLIP-BLIP	**76.57**	**97.23**	**69.39**	**67.72**	**78.17**	**97.27**	**62.26**	**67.85**

TABLE III: Ablation study for the impact of varying ϵ values on model performance in the VQA-CP v2 test set using the S-MRL network as the backbone. This table illustrates the effect of different ϵ values on the accuracy and stability of the model, offering insights into the optimal setting for ϵ in logarithmic computations within the fusion equations.

ϵ	All	Yes/No	Number	Other
1.00×10^{-12}	58.6	87.81	58.59	45.82
5.00×10^{-12}	59.51	88.51	59.47	45.77
1.00×10^{-11}	59.22	87.78	58.66	45.79
5.00×10^{-11}	59.71	88.18	58.31	45.85
1.00×10^{-10}	59.6	88.03	59.32	45.19
5.00×10^{-10}	59.31	87.07	59.44	45.02
1.00×10^{-09}	59.44	87.71	59.6	44.77
5.00×10^{-09}	58.13	86.94	57.78	43.41

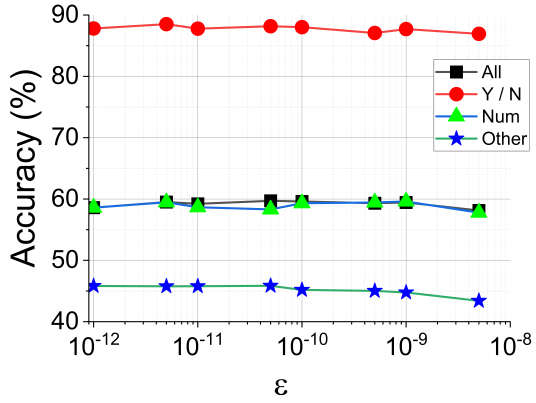


Fig. 11: Ablation of different values of epsilon on VQA-CP v2 test set. Variations of ϵ has a slight effect on improving the results, though the reason may be related to computational stability. These results are related to PW-VQA with $\alpha = 1.5$ and $\epsilon = \{10^{-12}, 5 \times 10^{-12}, \ldots, 5 \times 10^{-9}\}$.

4) Comprehensive Statistical Analysis of Performance Metrics: This ablation study presents an in-depth statistical analysis of the performance metrics, specifically focusing on the variance and mean values of the accuracy results of our proposed VQA method. These statistical measures are crucial for understanding the consistency and reliability of the model under various conditions. The variance of the accuracy results, listed in Table. IV, provides insights into the model's stability across different datasets and question types.

5) Role of Explain-Away Strategy in Accuracy and Numerical Question Performance: Table V presents a detailed ablation study focusing on the impact of different backbone architectures in conjunction with our enhanced causal graph approach. This study is critical for understanding the efficacy of the proposed counterfactual mechanism in mitigating vision-language fusion collider bias, a notable contributor to inaccuracies in VQA systems.

In this study, we deliberately exclude the Explain-Away (EA) fusion strategy to isolate and evaluate the intrinsic capabilities of our causal counterfactual approach. By doing so, we aim to provide a clearer insight into the fundamental performance improvements attributed solely to the causal graph dynamics. Our results demonstrate a marked improvement in handling vision-language associations, particularly evident in the enhanced accuracy of responses to numerically oriented questions. This improvement underscores the effectiveness of our causal graph in disentangling complex vision-language interdependencies.

6) Effect of α Variations on Performance: In our ablation study, we examined the influence of varying the α parameter within the PW-VQA model. This parameter, integral to our model's architecture, was tested across a range from 1 to 2 to understand its effect on performance. The results of this investigation, as detailed in Table. VI, indicate that an α value of 1.4 yields optimal performance across a majority of the tested backbones. This suggests a potential sweet spot for α in balancing the model's underlying mechanisms, with significant implications for tuning the PW-VQA model for different applications. The table further provides a comparative analysis of performance across various backbones, highlighting the robustness and versatility of the model at this specific α value.

7) Analyzing the Impact of Weight Parameters in the Loss Function: Our ablation study methodically examines the influence of weight parameters w_{kl} and w_{cls} within the final loss function. These parameters control the contribution of the Kullback-Leibler (KL) divergence loss ($\mathcal{L}_{kl}$) and the classification loss ($\mathcal{L}_{cls}$). We adjust these weights to explore their effect on model performance, focusing on the balance between the two loss components. Following expression (14) the final loss function is reformulated as:

$$\mathcal{L}_{final} = \sum (v, q, a) \in \mathcal{D} w_{cls} \cdot \mathcal{L}_{cls} + w_{kl} \cdot \mathcal{L}_{kl} \qquad (16)$$

This modified equation enables a detailed investigation into the significance of each loss term in enhancing model performance. Our findings, summarized in Table. VII, indicate that while minor variations in weights do not significantly impact results, a critical threshold exists. Notably, the model's performance deteriorates when $w_{cls} = 1 - w_{kl} = 0.0$, underscoring the necessity of a balanced trade-off between $\mathcal{L}_{kl}$ and $\mathcal{L}_{cls}$ for optimal outcomes.

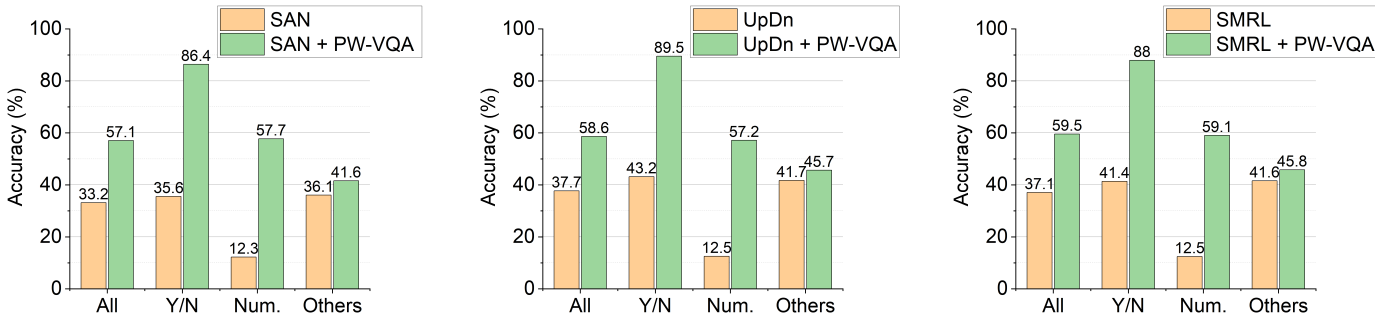


Fig. 12: The plots here show the performance metrics in percent for different backbones using our proposed method, PW-VQA. All $\alpha = 1.4$ are set for all these simulations. Digits on the bars are rounded up to one digit. As shown here, the proposed method consistently improves the performance of all the backbone.

TABLE IV: Variances of the accuracy performance of our method based on five performing simulations with different random seeds.

Dataset	Base	Overall	Y/N	Num.	Other
VQA v2	S-MRL	$60.76^{\pm 0.10}$	$79.60^{\pm 0.47}$	$42.75^{\pm 0.07}$	$51.21^{\pm 0.03}$
VQA v2	UpDn	$62.62^{\pm 0.01}$	$81.80^{\pm 0.06}$	$43.59^{\pm 0.05}$	$53.09^{\pm 0.01}$
VQA-CP v2	S-MRL	$59.71^{\pm 0.14}$	$87.98^{\pm 0.24}$	$59.41^{\pm 0.61}$	$45.51^{\pm 0.09}$
VQA-CP v2	UpDn	$58.70^{\pm 0.20}$	$89.19^{\pm 0.23}$	$58.85^{\pm 0.56}$	$45.17^{\pm 0.05}$

TABLE V: Ablation study on different backbones, namely SAN, S-MRL, UpDn, and BLIP-1 as listed here. As listed, our proposed method is improving the results when used with all of the backbones here, and also improves as we use the fusion and causal graph that is proposed. The fusion function is with $\alpha = 1.4$ as the free parameter and based on empirical study.

	All	Y/N	Num.	Other
SAN [59]	32.77	38.12	12.38	35.56
SAN+EA	46.25	62.13	37.58	40.31
+PW-VQA (EA)	57.06	86.4	57.73	41.57

	All	Y/N	Num.	Other
UpDn [60]	37.55	42.11	12.88	41.93
UpDn (EA)	47.02	65.89	18.11	45.06
+PW-VQA (EA)	58.64	89.51	57.15	45.68

	All	Y/N	Num.	Other
BLIP-1 [61]	33.60	54.26	14.80	34.22
BLIP-1 (EA)	42.37	73.19	17.03	34.09
+PW-VQA (EA)	46.21	84.65	18.42	33.98

	All	Y/N	Num.	Other
S-MRL [2]	36.59	40.71	13.17	40.85
S-MRL (EA)	49.65	72.48	24.42	44.6
+PW-VQA (EA)	59.54	87.95	59.05	45.83

TABLE VI: Ablation of PW-VQA α values on the final result for values ranging from 1 to 2. As shown the value of $\alpha = 1.4$ works well for most of the backbones.

	All	Y / N	Num	Other
SAN	33.18	38.57	12.25	36.1
α =1	56.23	86.2	57.72	40.3
α =1.1	56.27	87.45	58.5	39.54
α =1.2	56.96	86.84	58.07	41.57
α =1.3	52.9	76.87	57.67	39.95
α =1.4	57.06	86.4	57.73	41.57
α =1.5	42.75	51.24	52.24	39.38
α =1.6	55.1	85.64	58.31	38.39
α =1.7	56.2	86.41	58.21	41.15
α =1.8	53.85	87.36	54.65	39.3
α =1.9	43.41	73.47	57.52	34.49
α =2	53.39	86.22	52.64	39.26

	All	Y / N	Num	Other
UpDn	37.69	43.17	12.53	41.72
α =1	57.75	89.09	53.25	45.05
α =1.1	58.45	89.8	55.5	45.86
α =1.2	57.55	89.22	57.24	43.08
α =1.3	57.64	89.24	54.1	45.67
α =1.4	58.64	89.51	57.15	45.68
α =1.5	59.13	89.34	57.71	45.38
α =1.6	58.59	88.18	58.08	45.2
α =1.7	56.96	89.07	45.16	44.78
α =1.8	57.09	89.34	54.69	44.83
α =1.9	58.91	88.51	59.66	44.14
α =2	58.67	88.16	59.84	43.95

	All	Y / N	Num	Other
S-MRL	37.09	41.39	12.46	41.6
α =1	59.47	88.57	58.55	45.44
α =1.1	59.49	89.1	58.95	45.45
α =1.2	59.17	87.76	59.53	45.54
α =1.3	59.24	87.86	59.04	45.71
α =1.4	59.54	87.95	59.05	45.83
α =1.5	59.71	88.18	58.31	45.85
α =1.6	59.44	88.02	58.83	45.43
α =1.7	59.42	87.79	59.27	45.24
α =1.8	59.26	87.5	59.56	44.8
α =1.9	58.82	87.11	59.29	44.31
α =2	58.4	86.51	57.55	43.99

	All	Y / N	Num	Other
BLIP-1	34.42	54.97	14.48	35.10
α =1	45.32	84.65	17.15	32.52
α =1.1	45.32	84.84	17.66	32.49
α =1.2	45.41	84.84	17.47	32.51
α =1.3	45.25	84.68	17.04	32.61
α =1.4	46.21	84.65	18.42	33.98
α =1.5	47.00	84.70	24.92	33.98
α =1.6	47.92	84.94	30.59	33.79
α =1.7	48.81	84.94	37.3	33.85
α =1.8	49.22	84.64	41.87	33.76
α =1.9	49.53	84.36	45.38	33.24
α =2	48.32	83.78	46.50	31.14

TABLE VII: Ablation study results are listed for the impact of adjusting loss term weights in the final loss function (Expression (14)) with parameter settings of $\alpha = 1.5$ and $w_{cls} = 1 - w_{kl}$. This table delineates the performance variations observed when systematically altering the balance between classification loss ($\mathcal{L}_{cls}$) and KL divergence loss ($\mathcal{L}_{kl}$), providing insights into the optimal weighting strategy for enhanced model efficacy.

	All	Y / N	Num	Other
S-MRL	37.09	41.39	12.46	41.60
$w_{kl} = 0.0$	46.49	83.39	40.00	30.98
$w_{kl} = 0.1$	49.30	86.55	57.01	30.11
$w_{kl} = 0.2$	44.54	84.22	32.18	27.00
$w_{kl} = 0.3$	36.00	65.13	33.73	22.93
$w_{kl} = 0.4$	39.96	81.83	45.24	22.18
$w_{kl} = 0.5$	42.82	88.38	27.13	26.99
$w_{kl} = 0.6$	48.73	85.22	57.77	29.71
$w_{kl} = 0.7$	48.48	87.77	27.67	33.88
$w_{kl} = 0.8$	44.74	84.96	29.11	28.72
$w_{kl} = 0.9$	45.72	85.89	33.05	29.49
$w_{kl} = 1.0$	0.03	0.00	0.01	0.06

VI. Conclusions

VQA systems suffer from leveraging information only from one modality, especially the language modality from the given question. Many methods have been proposed to address this kind of problem. However, the previous method didn't consider that biases that come from each modality are highly confounded through the annotation process. VQA systems that ignore this effect cannot avoid increasing the bias learned from one modality while trying to reduce bias from another modality. We formulate the Explain-Away effect that causes the bias of both vision and language modalities with a novel causal framework for VQA systems. This framework can be implemented on the different VQA backbones and improve their generalizability significantly. The proposed framework successfully helps VQA systems reduce language bias without increasing vision bias. Experiment results show that our proposed method achieved state-of-the-art performance on debias oriented dataset VQA-CP especially doubled the accuracy

on numerical questions from the previous best model.

ACKNOWLEDGMENTS

This work was supported by NSF 1909912 and 2202124, NNSA NA0004078, NIH R01EY034562, DARPA HR00112220003, and the Center of Excellence in Data Science, an Empire State Development-designated Center of Excellence. Jiebo Luo's research was partially funded by the NSF under Award No. 2238208. Shijian Deng and Yapeng Tian's work was supported by a research award from Cisco. The content of the information does not necessarily reflect the position of the Government, and no official endorsement should be inferred.

LIMITATIONS

Our method, while demonstrating proficiency in various scenarios, exhibits limitations, particularly in contexts requiring extensive background knowledge and reasoning. This challenge is not unique to our approach but is a common shortfall across current VQA systems, including CF-VQA [7] and regular VQA models, as more examples illustrated in Figure 13. In this figure, we compare the performance of different methods on the VQA-CP v2 test split. The ground-truth answers are indicated by red bars, contrasting with other bars representing the prediction probabilities.

Dependency on Background Knowledge: A critical observation is the models' struggle with questions necessitating historical or contextual information. For example, accurately responding to a question like "What year was this picture taken?" demands knowledge of specific time periods associated with visual elements in the image, such as sneaker and bicycle designs. Such inquiries require the model to infer a time range (e.g., 1960-1980) by engaging in visual reasoning that considers both minor and significant details. Similarly, questions like "What bridge is this?" also necessitate background information that is not readily available in the training data.

Reasoning Limitations: In cases where reasoning is essential, our method, although successful in matching the expected ground-truth answer in the dataset, reveals an inherent limitation. For instance, in answering "How many bears are in the picture?", our model identifies the correct answer but lacks the capability for the kind of reasoning a human might employ. It operates as a multimodal system, leveraging vision and language to classify answers, but does not engage in the deeper reasoning processes that such questions may require.

Future Directions: These limitations highlight the need for future VQA systems to incorporate mechanisms for contextual reasoning and background knowledge integration. While our method marks progress in multimodal learning, the quest for a truly comprehensive VQA system continues.

ETHICS STATEMENT

As fundamental components in various AI applications, including visual dialog and question-answering systems, Visual Question Answering (VQA) systems bear significant ethical responsibilities. The potential for these systems to inadvertently propagate or amplify unethical content, such as racial or gender biases, warrants careful consideration, especially when deployed at scale.

Potential Risks: One of the primary ethical concerns revolves around bias in VQA systems. Biases in training data can lead to biased outputs, perpetuating stereotypes or unfair representations. This risk is particularly acute in systems that interact with diverse user populations and in contexts with significant social implications.

Benefits and Social Impact: On the positive side, VQA systems hold tremendous potential for societal benefits, notably in assisting individuals with disabilities or visual impairments. By enabling visual queries through natural language interfaces, these systems can significantly enhance accessibility and independence for many users.

Mitigation Strategies: To address these ethical challenges, it is imperative to incorporate robust measures during the development and deployment of VQA systems. This includes careful curation of training datasets to minimize bias, ongoing monitoring for unintended discriminatory patterns, and transparency in algorithmic decision-making processes. Furthermore, involving diverse stakeholder groups in the development process can provide valuable insights into potential ethical pitfalls and user-centric solutions.

REFERENCES

[1] Yulei Niu and Hanwang Zhang, "Introspective Distillation for Robust Question Answering," *Adv. Neural Inform. Process. Syst.*, Vol. 34, pp. 16292–16304, 2021.

[2] Rémi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, *et al.*, "RUBi: Reducing Unimodal Biases for Visual Question Answering," *Adv. Neural Inform. Process. Syst.*, Vol. 32, pp. 841–852, 2019.

[3] Remi Cadene, Hedi Ben-Younes, Matthieu Cord, and Nicolas Thome, "Murel: Multimodal Relational Reasoning for Visual Question Answering," *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 1989–1998, 2019.

[4] Chenchen Jing, Yuwei Wu, Xiaoxun Zhang, Yunde Jia, and Qi Wu, "Overcoming Language Priors in VQA via Decomposed Linguistic Representations," *AAAI*, pp. 11181–11188, 2020.

[5] Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee, "Overcoming Language Priors in Visual Question Answering with Adversarial Regularization," *Adv. Neural Inform. Process. Syst.*, pp. 1541–1551, 2018.

[6] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer, "Don't Take the Easy Way Out: Ensemble Based Methods for Avoiding Known Dataset Biases," *Joint Conf. on Empirical Methods in Natural Language Processing and the Int. Joint Conf. on Natural Language Processing*, pp. 4060–4073, 2019.

[7] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen, "Counterfactual VQA: A Cause-Effect Look at Language Bias," *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 12700–12710, 2021.

[8] Itai Gat, Idan Schwartz, Alexander Schwing, and Tamir Hazan, "Removing Bias in Multi-modal Classifiers: Regularization by Maximizing Functional Entropies," *Adv. Neural Inform. Process. Syst.*, Vol. 33, pp. 3197–3208, 2020.

[9] Gary Lupyan, Rasha Abdel Rahman, Lera Boroditsky, and Andy Clark, "Effects of Language on Visual Perception," *Trends in Cognitive Sciences*, Vol. 24, No. 11, pp. 930–944, 2020.

[10] Edgar Rubin, "Synsoplevede Figurer," 1915.

[11] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra, "VQA: Visual Question Answering," *International Journal of Computer Vision*, Vol. 123, No. 1, pp. 4–31, 2017.

[12] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi, "Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering," *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 4971–4980, 2018.

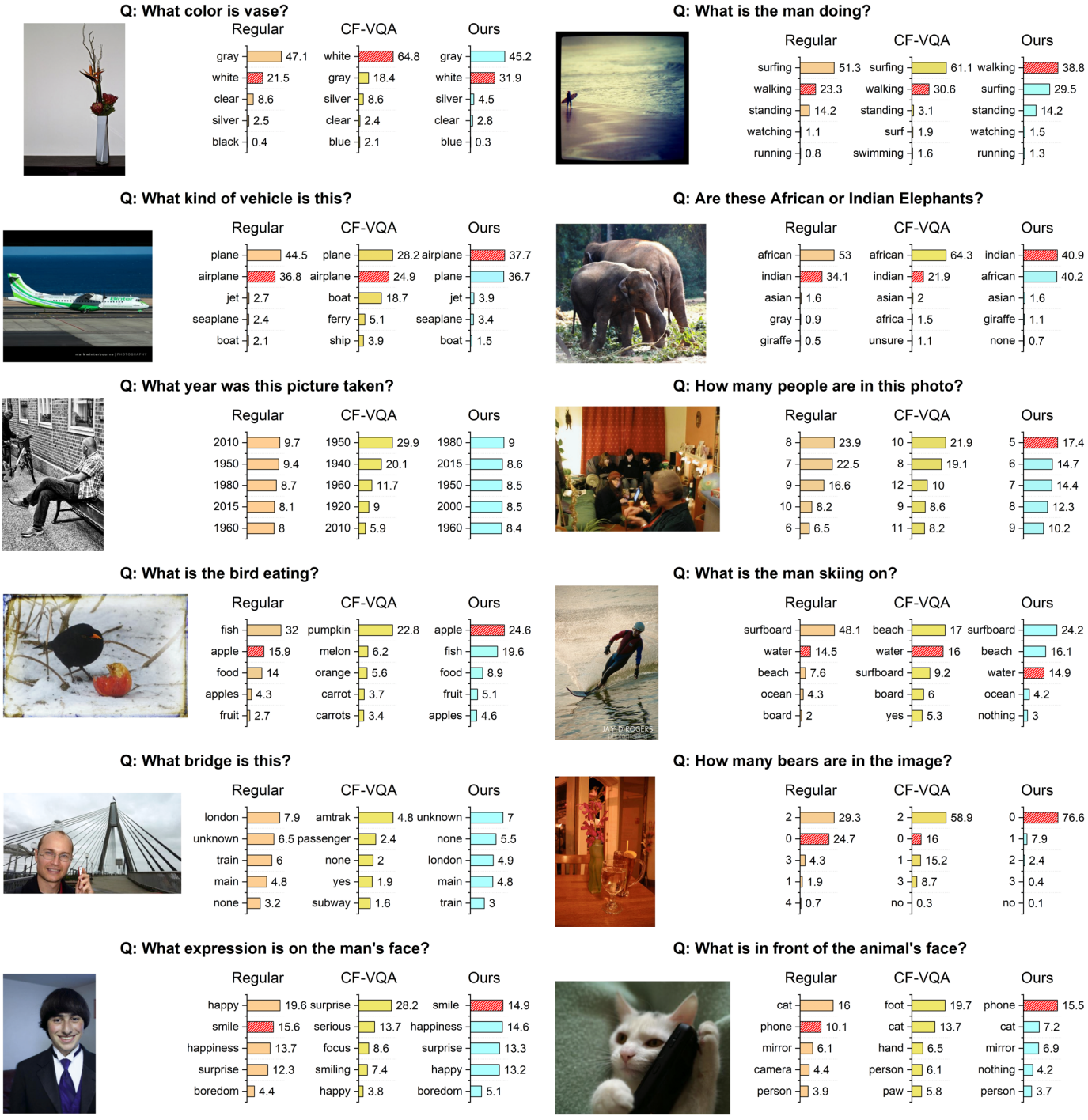


Fig. 13: Qualitative comparison of VQA-CP v2 test split, our method vs. CF-VQA [7] and regular VQA are shown in these images. Red bars denote the ground-truth one, while the other bars denote the prediction probability corresponding to their value. A limitation of VQA models is shown with two examples where PW-VQA fails to answer correctly; however, CF-VQA and regular VQA also fail.

[13] 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," *Int. Conf. Comput. Vis.*

[14] Jialin Wu and Raymond Mooney, "Self-Critical Reasoning for Robust Visual Question Answering," *Adv. Neural Inform. Process. Syst.*, Vol. 32, 2019.

[15] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra, "Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions?," *Computer Vision and Image Understanding*, Vol. 163, pp. 90–100, 2017.

[16] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang, "Counterfactual Samples Synthesizing for Robust Visual Question Answering," *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 10800–10809, 2020.

[17] Ehsan Abbasnejad, Damien Teney, Amin Parvaneh, Javen Shi, and Anton van den Hengel, "Counterfactual Vision and Language Learning," *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 10044–10054, 2020.

[18] Damien Teney, Ehsan Abbasnejad, and Anton van den Hengel, "Unshuffling Data for Improved Generalization in Visual Question Answering," *Int. Conf. Comput. Vis.*, pp. 1417–1427, 2021.

[19] Xi Zhu, Zhendong Mao, Chunxiao Liu, Peng Zhang, Bin Wang, and Yongdong Zhang, "Overcoming Language Priors with Self-Supervised Learning for Visual Question Answering," *IJCAI*, pp. 1083–1089, 2021.

[20] Camila Kolling, Martin More, Nathan Gavenski, Eduardo Pooch, Otávio Parraga, and Rodrigo C Barros, "Efficient Counterfactual Debiasing for Visual Question Answering," *Winter Conf. Applications Comput. Vision*, pp. 3001–3010, 2022.

[21] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee, "Counterfactual Visual Explanations," Proceedings of Machine Learning Research (PMLR), pp. 2376–2384, 2019.

[22] Pei Wang and Nuno Vasconcelos, "Scout: Self-Aware Discriminant Counterfactual Explanations," *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 8981–8990, 2020.

[23] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum, "CLEVRER: Collision Events for Video Representation and Reasoning," *Int. Conf. Learn. Represent.*, 2020.

[24] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang, "Unbiased Scene Graph Generation from Biased Training," *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 3716–3725, 2020.

[25] Zhongqi Yue, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua, "Interventional Few-Shot Learning," *Adv. Neural Inform. Process. Syst.*, Vol. 33, pp. 2734–2746, 2020.

[26] Zhongqi Yue, Tan Wang, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang, "Counterfactual Zero-Shot and Open-Set Visual Recognition," *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 15404–15414, 2021.

[27] Xinting Hu, Kaihua Tang, Chunyan Miao, Xian-Sheng Hua, and Hanwang Zhang, "Distilling Causal Effect of Data in Class-Incremental Learning," *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 3957–3966, 2021.

[28] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun, "Visual Commonsense R-CNN," *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 10760–10770, 2020.

[29] Shengyu Zhang, Tan Jiang, Tan Wang, Kun Kuang, Zhou Zhao, Jianke Zhu, Jin Yu, Hongxia Yang, and Fei Wu, "DeVLBert: Learning Deconfounded Visio-Linguistic Representations," *ACM Int. Conf. Multimedia*, pp. 4373–4382, 2020.

[30] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xian-Sheng Hua, and Qianru Sun, "Causal Intervention for Weakly-Supervised Semantic Segmentation," *Adv. Neural Inform. Process. Syst.*, Vol. 33, pp. 655–666, 2020.

[31] Damien Teney, Ehsan Abbasnedjad, and Anton van den Hengel, "Learning What Makes a Difference from Counterfactual Examples and Gradient Supervision," *European Conference on Computer Vision*. Springer, pp. 580–599, 2020.

[32] Xu Yang, Hanwang Zhang, and Jianfei Cai, "Deconfounded Image Captioning: A Causal Retrospect," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021.

[33] Tsu-Jui Fu, Xin Wang, Scott Grafton, Miguel Eckstein, and William Yang Wang, "Iterative Language-Based Image Editing via Self-Supervised Counterfactual Reasoning," *Conf. on Empirical Methods in Natural Language Processing*, pp. 4413–4422, 2020.

[34] Xu Yang, Hanwang Zhang, Guojun Qi, and Jianfei Cai, "Causal Attention for Vision-Language Tasks," *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 9847–9857, 2021.

[35] Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott, "Visually Grounded Reasoning across Languages and Cultures," *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 10467–10485, 2021.

[36] Michael Zhang and Eunsol Choi, "SituatedQA: Incorporating Extra-Linguistic Contexts into QA," *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7371–7387, 2021.

[37] Judea Pearl and Dana Mackenzie, *The Book of Why: The New Science of Cause and Effect*, Basic Books, 2018.

[38] Aishwarya Agrawal, Dhruv Batra, and Devi Parikh, "Analyzing the Behavior of Visual Question Answering Models," *Conf. on Empirical Methods in Natural Language Processing*, pp. 1955–1960, 2016.

[39] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh, "Yin and Yang: Balancing and Answering Binary Visual Questions," *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 5014–5022, 2016.

[40] 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," *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 6904–6913, 2017.

[41] Zhiquan Wen, Yaowei Wang, Mingkui Tan, Qingyao Wu, and Qi Wu, "Digging out discrimination information from generated samples for robust visual question answering," *Findings Assoc. for Comput. Linguistics: ACL*, pp. 6910–6928, 2023.

[42] Zhiquan Wen, Shuaicheng Niu, Ge Li, Qingyao Wu, Mingkui Tan, and Qi Wu, "Test-time model adaptation for visual question answering with debiased self-supervisions," *IEEE Trans. Multimedia*, 2023.

[43] Trang Nguyen and Naoaki Okazaki, "Causal Reasoning through Two Cognition Layers for Improving Generalization in Visual Question Answering," *Conf. on Empirical Methods in Natural Language Processing*, pp. 9221–9236, 2023.

[44] Zhong Ji, Junhua Hu, Deyin Liu, Lin Yuanbo Wu, and Ye Zhao, "Asymmetric Cross-Scale Alignment for Text-Based Person Search," *IEEE Trans. Multimedia*, 2022.

[45] Kaining Ying, Qing Zhong, Weian Mao, Zhenhua Wang, Hao Chen, Lin Yuanbo Wu, Yifan Liu, Chengxiang Fan, Yunzhi Zhuge, and Chunhua Shen, "CTVIS: Consistent Training for Online Video Instance Segmentation," *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 899–908, 2023.

[46] Lin Yuanbo Wu, Lingqiao Liu, Yang Wang, Zheng Zhang, Farid Boussaid, Mohammed Bennamoun, and Xianghua Xie, "Learning Resolution-Adaptive Representations for Cross-Resolution Person Re-Identification," *IEEE Trans. Image Process.*, 2023.

[47] 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," *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 5078–5088, 2022.

[48] Jae Won Cho, Dong-Jin Kim, Hyeonggon Ryu, and In So Kweon, "Generative Bias for Robust Visual Question Answering," *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 11681–11690, 2023.

[49] Fei Liu, Jing Liu, Zhiwei Fang, Richang Hong, and Hanqing Lu, "Visual Question Answering with Dense Inter- and Intra-Modality Interactions," *IEEE Trans. Multimedia*, Vol. 23, pp. 3518–3529, 2020.

[50] Chenchen Jing, Yunde Jia, Yuwei Wu, Xinyu Liu, and Qi Wu, "Maintaining Reasoning Consistency in Compositional Visual Question Answering," *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 5099–5108, 2022.

[51] Chuanhao Li, Zhen Li, Chenchen Jing, Yunde Jia, and Yuwei Wu, "Exploring the Effect of Primitives for Compositional Generalization in Vision-and-Language," *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 19092–19101, 2023.

[52] Xi Zhang, Feifei Zhang, and Changsheng Xu, "Reducing Vision-Answer biases for Multiple-choice VQA," *IEEE Trans. Image Process.*, 2023.

[53] Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata, "Grounding Visual Explanations," *Eur. Conf. Comput. Vis.*, pp. 264–279, 2018.

[54] Zhiheng Li, Ivan Evtimov, Albert Gordo, Caner Hazirbas, Tal Hassner, Cristian Canton Ferrer, Chenliang Xu, and Mark Ibrahim, "A Whac-a-Mole Dilemma: Shortcuts Come in Multiples Where Mitigating One Amplifies Others," *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 20071–20082, 2023.

[55] Edgar Rubin, *Visuell wahrgenommene Figuren: Studien in psychologischer Analyse*, Vol. 1, Gyldendalske boghandel, 1921.

[56] Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, Joris M Mooij, and Bernhard Schölkopf, "On Causal and Anticausal Learning," pp. 1255–1262, 2012.

[57] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz, "Invariant Risk Minimization," *arXiv preprint arXiv:1907.02893*, 2019.

[58] Judea Pearl, "Direct and Indirect Effects," *Probabilistic and Causal Inference*, 2001.

[59] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola, "Stacked Attention Networks for Image Question Answering," *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 21–29, 2016.

[60] 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," *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 6077–6086, 2018.

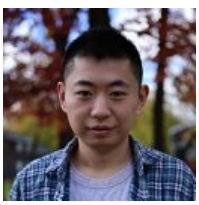
[61] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi, "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation," *International Conference on Machine Learning*, 2022.

[62] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *arXiv preprint arXiv:2301.12597*, 2023.
[63] Gouthaman Kv and Anurag Mittal, "Reducing Language Biases in Visual Question Answering with Visually-Grounded Question Encoder," *Eur. Conf. Comput. Vis.* Springer, pp. 18–34, 2020.
[64] Xinzhe Han, Shuhui Wang, Chi Su, Qingming Huang, and Qi Tian, "Greedy Gradient Ensemble for Robust Visual Question Answering," *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 1584–1593, 2021.
[65] Damien Teney, Ehsan Abbasnejad, Kushal Kafle, Robik Shrestha, Christopher Kanan, and Anton Van Den Hengel, "On the Value of Out-of-Distribution Testing: An Example of Goodhart's Law," *Adv. Neural Inform. Process. Syst.*, Vol. 33, pp. 407–417, 2020.
[66] Zujie Liang, Weitao Jiang, Haifeng Hu, and Jiaying Zhu, "Learning to Contrast the Counterfactual Samples for Robust Visual Question Answering," *Conf. on Empirical Methods in Natural Language Processing*, pp. 3285–3292, 2020.
[67] Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang, "MUTANT: A Training Paradigm for Out-of-Distribution Generalization in Visual Question Answering," *Conf. on Empirical Methods in Natural Language Processing*, pp. 878–892, 2020.
[68] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the Knowledge in a Neural Network," *arXiv preprint arXiv:1503.02531*, 2015.
[69] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, *et al.*, "Learning Transferable Visual Models from Natural Language Supervision," *International Conference on Machine Learning*. PMLR, pp. 8748–8763, 2021.

VII. Biography Section

Ali Vosoughi is presently pursuing his PhD at the University of Rochester. He earned his Bachelor's degree from Sharif University of Technology and his Master's degree from Bogazici University.

Shijian Deng is currently pursuing a Ph.D. at The University of Texas at Dallas, where his research intersects the innovative realms of causal learning and multimodal learning. Focused on harmonizing insights from computer vision, auditory processing, and natural language processing, he seeks to pioneer methods that enhance AI's interpretive capabilities across these varied domains.



Songyang Zhang is an applied scientist at Amazon. He received the Ph.D. degree from University of Rochester, in 2023. Before that, he got his master's degree from Zhejiang University and his bachelor's degree from Southeast University. His research is on the intersection of computer vision and natural language processing, including multi-modal vision-language understanding and generation.

Yapeng Tian (Member, IEEE) received the B.E. degree in electronic engineering from Xidian University, Xi'an, China, in 2013, the M.E. degree in electronic engineering from Tsinghua University, Beijing, China, in 2017, and the Ph.D. degree in computer science from the University of Rochester, Rochester, NY, USA, in 2022. He is currently an Assistant Professor with the Department of Computer Science, The University of Texas at Dallas, Richardson, TX, USA. His research interests include computer vision, computer audition, and machine learning.

Chenliang Xu is an Associate Professor in the Department of Computer Science at the University of Rochester. He received his Ph.D. in Computer Science from the University of Michigan in 2016, an M.S. in Computer Science from the University at Buffalo in 2012, and a B.S. in Information and Computing Science from Nanjing University of Aeronautics and Astronautics, China, in 2010. His research originates in computer vision and tackles interdisciplinary topics, including video understanding, audio-visual learning, vision and language, and methods for trustworthy AI.

Jiebo Luo (IEEE Fellow) is the Albert Arendt Hopeman Professor of Engineering and Professor of Computer Science at the University of Rochester which he joined in 2011 after a prolific career of fifteen years at Kodak Research Laboratories. He has authored over 600 technical papers and holds over 90 U.S. patents. His research interests include computer vision, NLP, machine learning, data mining, computational social science, and digital health. He has been involved in numerous technical conferences, including as a program co-chair of ACM Multimedia 2010, IEEE CVPR 2012, ACM ICMR 2016, and IEEE ICIP 2017, as well as a general co-chair of ACM Multimedia 2018 and IEEE ICME 2024. He has served on the editorial boards of the IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), IEEE Transactions on Multimedia (TMM), IEEE Transactions on Circuits and Systems for Video Technology (TCSVT), IEEE Transactions on Big Data (TBD), ACM Transactions on Intelligent Systems and Technology (TIST), Pattern Recognition, Knowledge and Information Systems (KAIS), Machine Vision and Applications, and Journal of Electronic Imaging. He was the Editor-in-Chief of the IEEE Transactions on Multimedia (2020-2022). Professor Luo is also a Fellow of NAI, ACM, AAAI, SPIE, and IAPR.