COCA: COllaborative CAusal Regularization for Audio-Visual Question Answering

Mingrui Lao[1], Nan Pu[1]*, Yu Liu[2], Kai He[1], Erwin M. Bakker[1], Michael S. Lew[1]

[1]LIACS Media Lab, Leiden University
[2]International School of Information Science & Engineering, Dalian University of Technology
{m.lao, n.pu, k.he, e.m.bakker, m.s.k.lew}@liacs.leidenuniv.nl, liuyu8824@dlut.edu.cn

Abstract

Audio-Visual Question Answering (AVQA) is a sophisticated QA task, which aims at answering textual questions over given video-audio pairs with comprehensive multimodal reasoning. Through detailed causal-graph analyses and careful inspections of their learning processes, we reveal that AVQA models are not only prone to over-exploit prevalent language bias, but also suffer from additional joint-modal biases caused by the shortcut relations between textual-auditory/visual co-occurrences and dominated answers. In this paper, we propose a COllabrative CAusal (COCA) Regularization to remedy this more challenging issue of data biases. Specifically, a novel Bias-centered Causal Regularization (BCR) is proposed to alleviate specific shortcut biases by intervening bias-irrelevant causal effects, and further introspect the predictions of AVQA models in counterfactual and factual scenarios. Based on the fact that the dominated bias impairing model robustness for different samples tends to be different, we introduce a Multi-shortcut Collaborative Debiasing (MCD) to measure how each sample suffers from different biases, and dynamically adjust their debiasing concentration to different shortcut correlations. Extensive experiments demonstrate the effectiveness as well as backbone-agnostic ability of our COCA strategy, and it achieves state-of-the-art performance on the large-scale MUSIC-AVQA dataset.

Introduction

AVQA is an emerging yet sophisticated QA task stemming from Visual-QA (Antol et al. 2015), Video-QA (Zhu et al. 2017) and Audio-QA (Fayek and Johnson 2020). Beyond these conventional QA tasks that concern reasoning from single- or cross-modality data, AVQA requires models to comprehensively understand multimodal data and perform spatio-temporal reasoning over audio-visual scenes (Li et al. 2022a). The additional core challenge in AVQA is that reasoning the answer necessarily needs to jointly understand the audio and the video as depicted in Fig. 1(a), which is easily hindered by data biases. Although there are many studies that focus on the data bias in VQA tasks, there has no research work on addressing the bias issue for AVQA. To reveal this problem and discover its potential over-dependences on different modalities, we conduct a group of exploratory experiments based on the state-of-the-art model (Li et al. 2022a). Concretely, we remove one or two modalities in the training process on MUSIC-AVQA dataset (Li et al. 2022a). Fig. 1(b) shows that there is no significant accuracy drop when deleting ether visual or audio modality. It is also surprising that only exploiting question can achieve a decent accuracy of 55% for open-ended questions. Undoubtedly, excluding textual information would remarkably impair the performance on audio-visual questions.

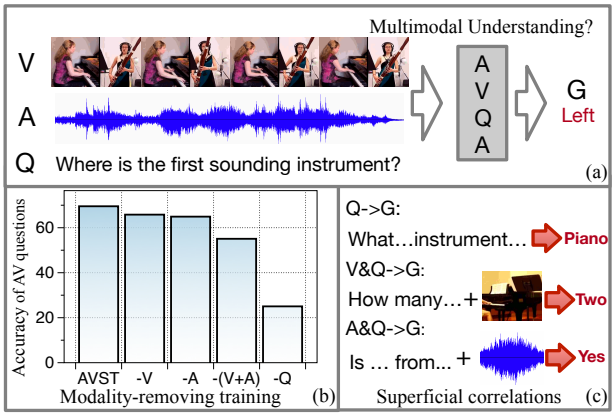

Figure 1: (a) AVQA requires a comprehensive multimodal reasoning to answer the textual question over video-audio information. (b) Exploratory experiments by removing one or two modalities at one time for AVQA training. (c) Illustration of the examples from three types of shortcut biases.

In this paper, by thoroughly analyzing the AVQA model with a causal-effect retrospection (Yao et al. 2021) and carefully inspecting its learning process, we find that the biases hindering robustness are mainly caused by three shortcut correlations: 1) Q→G: directly reasoning from question patterns and words to statistically frequent ground-truth answers G. 2) V&Q→G: over-exploiting the co-occurrence of questions and some specific visual objects to deduce related answers. 3) A&Q→G: only focusing on question patterns accompanied with certain audio waves to predict corresponding answers. Some illustrated examples for the aforementioned superficial correlations are in Fig. 1 (c). Compared with VQA or AQA tasks with uni-modal bias (Agrawal et al. 2018), the debiasing strategies for

*Corresponding Author.

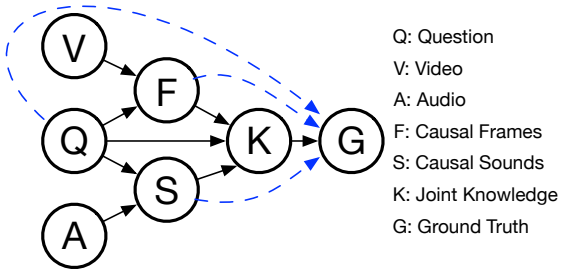

Figure 2: Causal Graph of AVQA, where dashed arrows in blue indicate the potential shortcut biases.

AVQA should collaboratively alleviate both uni-modal and joint-modal biases from different shortcut correlations.

In this paper, we propose a novel COllaborative CAusal (COCA) Regularization to cooperatively overcome the complex biases and improve model robustness from two aspects:

Bias-centered Causal Regularization (BCR). BCR aims at mitigating the bias caused by a specific shortcut correlation through regularization from *counterfactual* and *factual* views. The former promotes AVQA models to introspect the bias about the question "what will the answer G will be, if other modality information is inconsistent with the bias-centered modality". For the most intuitive Q→G shortcut bias, BCR conducts a counterfactual intervention on bias-irrelevant video and audio features, and meanwhile maintains the question-centered shortcut relation to synthesize counterfactual samples. Thus, the shortcut bias Q→G could be alleviated by misleading its counterfactual answer prediction. The latter mainly considers the type of questions "what will answer G will be, if other information is intervened, but is still causally related to shortcut-based modality". To this end, BCR replaces original bias-unrelated modality information with the generated informative factual alternatives. As the causal correlation between factual inputs and answer G is uninterrupted, BCR encourages the factual answer prediction to be consistent with the prediction from original inputs. Benefited from factual regularization, BCR could effectively avoid AVQA model from over-correcting in counterfactual scenarios, and further strengthen the generalization ability of multimodal representations.

Multi-shortcut Collaborative Debiasing (MCD). Although BCR could reduce the data bias from a specific shortcut, how to cooperatively leverage BCR for multi-shortcut debiasing is still a challenge to improve model robustness. It is mainly due to the fact that the dominated shortcut bias resulted in model vulnerability for different AVQA samples may be different. To this end, our MCD strategy presents a novel Information Entropy driven Metric I to measure how different shortcut biases impair model robustness for different samples. Then, through the comparisons among Is from different shortcut biases, MCD assigns instance-aware loss weights to not only balance the contributions of BCR for three shortcut bias, but also adjust the interdependence between counterfactual and factual regularization. Benefiting from MCD, our COCA is capable of following the principle of "suit the remedy to the case", and dynamically focuses on handling with the dominated bias for different samples.

The contributions of this paper are summarized as following: 1)To our best knowledge, this work is the first attempt to analyze the potential biases in AVQA task from the perspective of causal graph, and further reveal the multi-shortcut biases problem in this task. 2) We propose a backbone-agnostic COCA approach to cooperatively alleviate different biases with an instance-aware manner, and enhance the multimodal reasoning capacity of AVQA models. Extensive experiments verify the effectiveness of COCA, and show state-of-the-art performance on the MUSIC-AVQA dataset.

A Causal View with Bias Revelation on AVQA

In this section, we exhibit the causal graph of AVQA from the perspective of causal theory (Glymour, Pearl, and Jewell 2016), to indepthly disclose its reasoning process. Furthermore, through comprehensively analysing the causal graph, we derive three important shortcut bias, which heavily degrade the model's robustness and generalization ability.

Causal Graph of AVQA

Given the multimodal inputs and the ground-truth answer G, the causal graph is illustrated in Fig. 2, where nodes and links refer to variables and causal-effect relations.

- $V \rightarrow F \leftarrow Q$: The causal video frames F are inferred from the jointly reasoning between question Q and raw video V, which normally filters the question-related frames with important visual objects through visual attention mechanism.
- $A \rightarrow S \leftarrow Q$: The causal audio sounds S are determined by the question Q and raw audio A to discover the crucial auditory cues from the question-related time slice in the whole redundant audio.
- $K \rightarrow G$: Joint knowledge K is entirely conditional upon the comprehensive understanding of the causal effect $Q \rightarrow K$, $F \rightarrow G$ and $S \rightarrow G$ from three modalities through multimodal fusion. Hence, this causal-effect relation denotes the unbiased reasoning by projecting the joint knowledge into the final answer prediction.

Potential Shortcut Biases in AVQA

By deducing the causal graph, we find that the question-related variables can be spuriously correlated with answers, which results in uni-modal and joint-modal shortcut biases.

- $Q \dashrightarrow G$: implies the uni-modal language bias from Q to G, which is prone to caused by statistical regularities between answer occurrences and question patterns. Practically, if the language bias is severe, $Q \rightarrow K$ tends to dominant the contribution for joint knowledge K integration, whereas the cause-and-effect relations $F \rightarrow G$ and $S \rightarrow G$ would be ignored.
- $F \dashrightarrow G \& S \dashrightarrow G$: indicate the joint-modal biases caused by the superficial relations between textual-visual or textual-audio co-occurrences and ground-truth answers. For example, even though the training samples with question "where...first sound...?" are unbiased towards answers "right" and "left", the joint-modality bias can also discover a frequent combination of "first sound" and

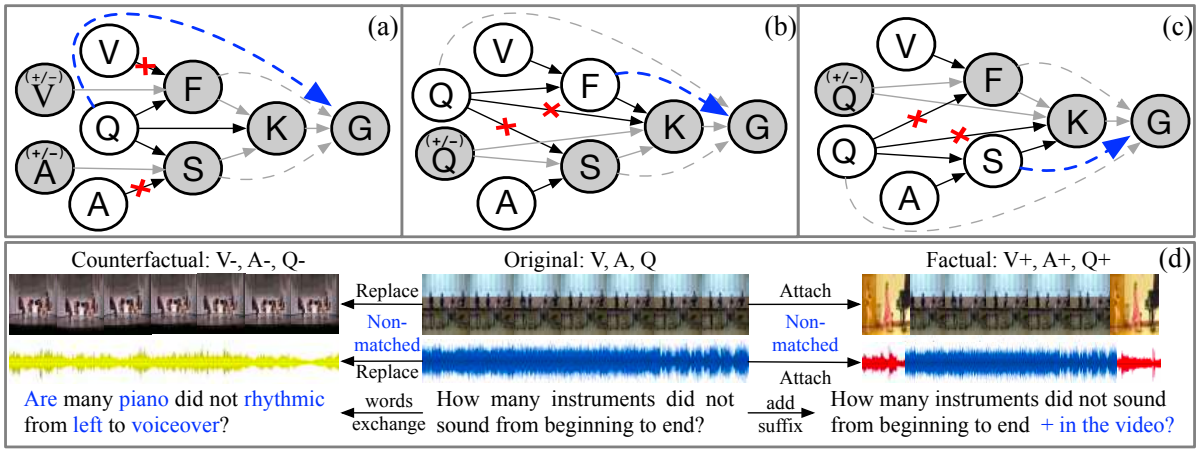

Figure 3: (a) $Q \dashrightarrow G$ centered factual/counterfactual (+/-) scenarios under causal intervention. (b) $F \dashrightarrow G$ centered +/- scenarios under causal intervention. (c) $S \dashrightarrow G$ centered +/- scenarios under causal intervention. (d) Illustrations of counterfactual (-) and factual (+) inputs synthesizing for video, audio and question required in BCR. In (a), (b) and (c), the shallowed nodes and links designate intervened variables and cause-and-effect relationships. The links with red cross means the relations were interrupted.

the scenes of "piano" to infer its statistically related answer "left" without considering audio information. In practice, if the textual-audio bias is severe, the causal video frame F determined by Q and V would significantly weaken the causal reasoning from $Q \rightarrow K$ and $S \rightarrow K$, and impair unbiased learning. Analogously, we can derive the other joint-modal shortcut bias, $S \dashrightarrow G$.

Remark. Compared with the causal analysis (Niu et al. 2021) of VQA tasks, the derived joint-modal shortcut biases additional exist in AVQA tasks because of involving more modalities for reasoning.

Methodology

Given inputs $v \in V$, $a \in A$ and $q \in Q$, an AVQA model aims to learn a function F to generate a distribution over the ground-truth answer space $\mathcal{G}$, which is formulated as:

$$P\left(\mathcal{G} \mid v, a, q\right) = F\left(v, a, q\right). \tag{1}$$

To reach competitive performance, conventional AVQA models (Li et al. 2022a; Yun et al. 2021) prefer to parse the audio-question as well as visual-question pairs through attention mechanism $h(\cdot)$ before multimodal fusion $f(\cdot)$. Thus, we can revisit the formulation of AVQA model as:

$$F\left(v, a, q\right) = f\left(h_{qv}(v, q), h_{qa}(a, q), q\right), \tag{2}$$

where $h_{qa}(\cdot)$ and $h_{qv}(\cdot)$ imply the question-guide audio and visual attention for causal sounds and frames extraction.

Subsequently, we elaborate our proposed COllaborative CAusal (COCA) Regularization, which consists of two crucial modules: 1) Bias-centered Causal Regularization (BCR) and 2) Multi-shortcut Collabrative Debiasing (MCD).

Bias-Centered Causal Regularization

Removing the bias caused by a specific shortcut is a fundamental yet challenging procedure for AVQA models. To address the issue, BCR introduces bias-centered **counterfactual** and **factual** scenarios by causal interventions on bias-irrelevant inputs.

Uni-modal Shortcut Bias. As depicted in Fig. 3(a), we assume that the language bias $Q \dashrightarrow G$ is due to the fact that, Q dominates the joint knowledge K aggregation through $Q \rightarrow K$, thereby limiting the contributions from paths $F \rightarrow K$ and $S \rightarrow K$. In order to uncover and highlight the spurious relation, we establish $Q \dashrightarrow G$ centered counterfactual and factual scenarios by intervening the bias-irrelevant video V and audio A with counterfactual $(-)$ and factual $(+)$ alternatives as shown in Fig. 3(a).

Specifically, in counterfactual world, we change them by an entirely different and also un-matched video-audio pair (V^- and A^-) in Fig. 3(d). Then, we exploit a counterfactual regularization loss to enforces the unreasonable prediction from (V^-,A^- ,Q) fails to answer the question by yielding uniform prediction over answer candidates:

$$\mathcal{L}_{cf}^{qs} = KL\left(F\left(v^-, a^-, q\right) \| \mathcal{X}\right), \tag{3}$$

where $\mathcal{L}_{cf}^{qs}$ denotes the counterfactual loss for $Q \dashrightarrow G$ shortcut debiasing, and $\mathcal{X}$ indicates the uniform distribution over the ground-truth answer candidates. As the interference thoroughly shears the causal relations $F \rightarrow K$ as well as $S \rightarrow K$, $\mathcal{L}_{cr}^{qs}$ can forthrightly mitigate the uni-modal shortcut bias by significantly increasing its predictive uncertainty.

In factual scenarios, we generate the factual V^+ and Q^+ by attaching extra frames and voices into the front and back sides of V and Q as shown in the right column of Fig. 3(d). Though the raw representation of V and Q are disturbed, their original causal relations to joint knowledge K are exhaustively maintained. Then, we employ the factual regularization loss to encourage the reasonable prediction from (V^+, A^+, Q), thereby holding the semantic consistency:

$$\mathcal{L}_{f}^{qs} = KL\left(F\left(v^+, a^+, q\right) \| F\left(v, a, q\right)\right), \tag{4}$$

where $\mathcal{L}_{f}^{qs}$ implies the loss function to overcome the uni-modal shortcut bias from the factual view. Accompanied

by the counterfactual regularization, the factual regularization prohibits AVQA model from over-correcting problem in counterfactual scenarios, triggered by overwhelmingly increasing the predictive uncertainty for any interventions upon V and A, without considerations of whether the causal relations are changed. Meanwhile, it boosts the diversity of feature representations and effectively strengths the semantic generalization of AVQA models.

Joint-Modal Shortcut Bias. The joint-modal shortcut biases $F \dashrightarrow G$ and $S \dashrightarrow G$ are owing to the over-dependence on the causal effect $F \rightarrow K$ and $S \rightarrow K$. To overcome the bias $F \dashrightarrow G$ in Fig. 3(b), we select to conduct causal interventions on Q in the paths $Q \rightarrow K$ and $A \rightarrow S \leftarrow Q$ to disturb shortcut-irrelevant cause-and-effect correlations $Q \rightarrow K$ and $S \rightarrow K$. Meanwhile, the input Q in $V \rightarrow F \leftarrow Q$ is unchanged to maintain the causality from F to K, so as to establish bias $F \dashrightarrow G$ centered counterfactual and factual scenarios. In practice, the causal relations of $Q \rightarrow K$ and $V \rightarrow F \leftarrow Q$ are implemented by multimodal fusion $f(\cdot)$ and question-guide visual attention $h_{qv}(\cdot)$ in Eq. (2), respectively. In the counterfactual scenarios, we present to synthesize Q^- by randomly exchanging 50% of the words between raw question and other sampled question sentence to damage its semantic information as shwon in Fig. 3(d). Then, the counterfactual loss $\mathcal{L}_{cf}^{fs}$ for reducing $F \dashrightarrow G$ bias is:

$$\mathcal{L}_{cf}^{fs} = KL\left(f\left(h_{qv}(v,q), h_{qa}(a,q^-), q^-\right) \| \mathcal{X}\right). \quad (5)$$

Likewise, the counterfactual loss $\mathcal{L}_{cf}^{ss}$ for alleviating the bias $S \dashrightarrow G$ is achieved by replacing Q with Q^- in question-guide visual attention $h_{qv}(\cdot)$ and multimodal fusion $f(\cdot)$, to break the causal effect from $Q \rightarrow G$ and $F \rightarrow G$ as depicted in Fig. 3(c). We define the loss function as:

$$\mathcal{L}_{cf}^{ss} = KL\left(f\left(h_{qv}(v,q^-), h_{qa}(a,q), q^-\right) \| \mathcal{X}\right). \quad (6)$$

For factual regularization, as shown in Fig. 3(d) we synthesize factual Q^+ by adding extra suffixes (e.g., "in the video" or "were heard") to original Q, thereby enriching its textual representation with no semantic change. The factual regularization loss $\mathcal{L}_{f}^{fs}$ for $F \dashrightarrow G$ is formatted as:

$$\mathcal{L}_{f}^{fs} = KL\left(f\left(h_{qv}(v,q), h_{qa}(a,q^+), q^+\right) \| F\left(v,a,q\right)\right). \quad (7)$$

The factual loss function L_{f}^{ss} for $S \dashrightarrow G$ could be defined similarly based on Eq. (7).

Multi-Shortcut Collabrative Debiasing

For different AVQA samples, the dominated shortcut biases they suffered from are presumably different. Consequently, the unbiased AVQA models are required to selectively remove biases for different samples, and encourage these debiasing objectives to complement mutually. To achieve this goal, we present a novel Multi-shortcut Collaborative Debiasing (MCD) strategy through **quantifying** the degrees of suffered biases for each sample, and further **weighting** different causal regularization loss functions.

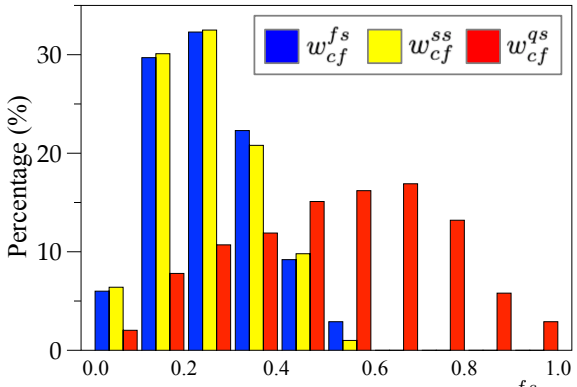

Figure 4: The distributions of different weights w_{cf}^{fs}, w_{cf}^{ss}, w_{cf}^{qs} on the MUSIC-AVQA training split.

For quantifying bias, we introduce an information entropy driven metric I^B to measure the how much the AVQA sample suffers from a specific shortcut bias B. Specifically, it is defined by the comparison between the maximum information entropy H_{max} over answer distribution and the entropy $H(\cdot)$ of B-centered counterfactual prediction P_{cf}^{B}:

$$\begin{aligned} I^B &= H_{max} - H(P_{cf}^{B}) \\ &= -\sum_{i}^{N} \frac{1}{N} log \frac{1}{N} + \sum_{i}^{N} P_{cf}(i)^B log P_{cf}^{B}(i), \end{aligned} \quad (8)$$

where N designates the number of answer candidates in ground-truth, and i is the i_{th} answer label. For instance, if an AVQA sample severely suffers from joint-modal bias $F \dashrightarrow G$, its counterfactual prediction P_{cf}^{fs} is prone to be biased toward its statistically-related answers with lower uncertainty. Thus, the gap of information entropy between the uniform distribution $\mathcal{X}$ (maximum entropy) and its counterfactual prediction P_{cf}^{fs} would obviously larger than those from other shortcut biases.

After acquiring the bias index I^B for each shortcut, we assign different weights for their counterfactual and factual loss functions. Explicitly, based on the example of $F \dashrightarrow G$, its counterfactual w_{cf}^{fs} and factual w_{f}^{fs} weights are:

$$w_{cf}^{fs} = \frac{I^{fs}}{I^{fs} + I^{ss} + I^{qs}}, \quad w_{f}^{fs} = \frac{I^{ss} + I^{qs}}{I^{fs} + I^{ss} + I^{qs}}. \quad (9)$$

The weights for the other two shortcuts can be determined in similar manners. On the one hand, for the counterfactual regularization for different shortcuts, we prompt AVQA model to focus on more-biased shortcut debiasing. On the other hand, for a specific shortcut bias, we encourage to assign larger weight to its factual regularization loss when the debiasing concentration for counterfactual scenario decreases, thereby enhancing its generalization and avoiding over-correcting in debiasing process.

Fig. 4 visualizes the weight distributions of counterfactual regularization loss for different shortcuts, which explicitly reveals how they affect unbiased multimodal learning for

different samples. It can be clearly observed that, for most AVQA samples, the uni-modal language biases are the dominated bias to be over-exploited, where the average value of W_{cf}^{qs} is around 0.5 on training dataset. On the contrary, the influences of joint-modal bias for most samples are relatively not severe, as the values of W_{cf}^{ss} and W_{cf}^{fs} are concentrated on the interval from 0.1 to 0.4.

Ultimately, we train the parameters in the whole AVQA classification network to jointly minimize the loss terms:

$$\mathcal{L}_{all} = \mathcal{L}_{task} + \alpha \sum_{B \in \{fs, ss, qs\}} (w_{cf}^{B} \mathcal{L}_{cf}^{B} + w_{f}^{B} \mathcal{L}_{f}^{B}), \quad (10)$$

where α is the trade-off factor to adjust the contributions between the task and the COCA objectives. It is worth noting that, COCA is model-agnostic debiasing strategy, which can be incorporated into various AVQA architectures.

Experiments

We first conduct experiments to verify the superiority of COCA compared with the state-of-the-art AQA, VideoQA and AVQA methods. Then, we introduce the in-depth ablation studies to validate the effectiveness of each component in our COCA method. Next, we explore the model-agnostic ability of COCA on different baseline models. Afterward, we visualize and analysis the qualitative results.

Dataset. We evaluate our method on the large-scale MUSIC-AVQA dataset (Li et al. 2022a), which consists of more than 40K question-answer pairs covering comprehensive question types over textual, visual and audio modalities. To our best knowledge, MUSIC-AVQA is the only officially released AVQA dataset, which includes audio-visual questions requiring joint understanding over auditory and visual information for answer prediction.

Implementation Details. We build a simple and effective model with audio-video temporal attention as the baseline. For video features, we employ pretrained ResNet18 (He et al. 2016) to extract features of 10 sampled frames for each video; for audio, we exploit VGGish (Hershey et al. 2017) to obtain 128-D feature vector from 16kHz sound; for textual features, we set the dimension of word embedding is 512, and fix the max length of words is 14. The initial learning rate is 1e-4, which would be decayed by multiplying 0.1 for every 8 epochs. The mini-batch and maximun epoch number are 64 and 24. The optimal trade-off factor we select is $\alpha = 0.75$, which is also validated in Fig. 5.

Compared Methods. To verify the effectiveness of our COCA, we first compare COCA with 6 state-of-the-art methods: 1) ConvLSTM (Fayek and Johnson 2020) is established by the combination of widely-used ConvLSTM AQA model with visual attention; 2) PSAC (Li et al. 2019) is an advanced video-QA method with a positional self-attention with co-attention; 3) HME (Fan et al. 2019) introduces a heterogeneous memory to enhance accuracy; 4) AVSD (Schwartz, Schwing, and Hazan 2019) is a state-of-the-art method for video dialog task; 5) LAViT (Yun et al. 2021) is an advanced Pano-AVQA approach based on multiple transformer auto-encoders; 6) AVST (Li et al. 2022a) is the most competitive state-of-the-art AVQA method, which introduce a spatio-temporal grounded audio-visual network with a two-stage training strategy. Next, to verify the model-agnostic ability of COCA, we incorporate COCA into two additional model architectures: 1) VQAT-based: As current existed AVQA model is limited, we modify the well-known VideoQA model VQAT (Yang et al. 2021a) with audio attention to form a VQAT-based AVQA model; 2) ConvLSTM-based: Analogously, we add a widely-used visual attention module upon the ConvLSTM (Fayek and Johnson 2020).

State-Of-The-Art Comparisons

From the performance between AQA/VideoQA models and AVQA approaches on Audio/Visual Questions, apart from the audio comparative question types, most AVQA models achieve superior performance, which indicates that additionally considering corresponding auditory/visual information can facilitates the question answering over video/audio input. For the comparison within AVQA methods, baseline and AVST approaches outperform AVSD and LAVit models by establishing more effective temporal visual-auditory attention mechanisms to find crucial cues for question answering. Our COCA strategy established upon 'baseline' model significantly enhances the overall performance with 1.67% accuracy boost over A-V questions, and even slightly outperforms the state-of-the-art AVST model. Moreover, through the integration with AVST, our method achieves state-of-the-art performance of 72.33% overall accuracy on MUSIC-AVQA test set, and occupies all the first places for average results on audio, visual and audio-visual questions, which further demonstrates the effectiveness of COCA.

Ablation Studies

As depicted in Tab. 2, we conduct extensive ablation studies to validate the contributions for different components in COCA on baseline model. We find that exploiting counterfactual regularization for each shortcut bias increases the average accuracy, among which L_{cf}^{qs} remarkably enhances the performance by around 0.9% upon baseline. It can be explained by the fact that, even though AVQA model suffers from multiple biases, most samples are still dominated by the uni-modal language bias, which is consistent to the weight distribution in Fig. 4. Based on counterfactual loss, their factual regularization slightly improves the performance by strengthening the generalization of shortcut-irrelevant multimodal inputs. Then, from the last four rows in Tab. 2, compared with evenly considering the contributions from three shortcut regularization, our MCD strategy effectively facilitates their debiasing synergy by dynamically assigning different weights to focus on dominated bias for different AVQA samples. Detailed comparisons for between MCD and average weighting are illustrated in Fig. 5. We can see that MCD is consistently superior to the average weighting under various setting of trade-off factor α, and reach its best accuracy when $\alpha = 0.75$ on both causal and counterfactual regularization testing. There results further validate the effectiveness and stability of MCD.

Task	Method	Audio Question			Visual Question			Audio-Visual Question Answering						All
		CNT	Comp	Avg.	CNT	LOC	Avg.	Exist	LOC	CNT	Comp	Temp	Avg.	Avg.
AQA	ConvLSTM	74.07	**68.89**	72.15	67.47	54.56	60.94	82.91	50.81	63.03	60.27	51.58	62.24	63.65
VQA	PSAC	75.64	66.06	72.09	68.64	69.79	69.22	77.59	55.02	63.42	61.17	59.47	63.52	66.54
	HME	74.76	63.56	70.61	67.97	69.46	68.76	80.30	53.18	63.19	62.69	59.83	64.05	66.45
AVQA	AVSD	72.41	61.90	68.52	67.39	74.19	70.83	81.61	58.79	63.89	61.52	61.41	65.49	67.44
	LAViT	74.36	64.56	70.73	69.39	75.65	72.56	81.21	59.33	64.91	64.22	63.23	66.64	68.93
	Baseline	74.53	65.99	71.38	69.76	74.69	72.25	81.38	61.20	67.04	61.20	63.99	67.80	69.61
	+COCA	79.35	66.50	74.61	72.35	76.08	74.24	**83.50**	64.02	**70.99**	63.40	64.48	69.47	71.64
	AVST	78.18	67.05	74.06	71.56	**76.38**	74.00	81.81	64.51	70.80	**66.01**	63.23	69.54	71.52
	+COCA	**79.94**	67.68	**75.42**	**75.10**	75.43	**75.23**	**83.50**	**66.63**	69.72	64.12	**65.57**	**69.96**	**72.33**

Table 1: State-of-the-art Comparisons on MUSIC-AVQA dataset. Best and second best numbers are in bold and underlined. CNT, comp, Avg, loc, Exist and Temp imply question types counting, comparative, average, location, existential and temporal.

Method	Component							Avg.
	L_{cf}^{qs}	L_{f}^{qs}	L_{cf}^{fs}	L_{f}^{fs}	L_{cf}^{ss}	L_{f}^{ss}	C	
Baseline								69.61
$Q \dashrightarrow G$	✓							70.50
	✓	✓						70.86
$F \dashrightarrow G$			✓					69.91
			✓	✓				70.12
$S \dashrightarrow G$					✓			70.04
					✓	✓		70.37
Multiple Shortcuts	✓		✓		✓			70.81
	✓		✓		✓		✓	71.21
	✓	✓	✓	✓	✓	✓		71.03
	✓	✓	✓	✓	✓	✓	✓	71.64

Table 2: Ablation study for different components in COCA based on the average accuracy on MUSIC-AVQA test set. C indicates our collaborative debiasing strategy

Model-Agnostic Evaluation Intuitively, COCA overcomes multiple shortcut bias by intervening multimodal inputs and conduct debiasing regularization for factual/counterfactual output logits. Therefore, COCA should be orthogonal to model architecture design, and incorporated into various AVQA models. To testify its backbone-agnostic ability, we integrate COCA into multiple different AVQA models in Tab. 3. Established upon Baseline, VQAT-based and ConvLSTM-based backbones, COCA can consistently improve the performance over all question types. Furthermore, blending COCA with the state-of-the-art AVST model could still enhance the performance reached at 72.33% with 0.81% accuracy boost. It also exhibits the effectiveness of COCA on advanced AVQA model. Also, we conclude the reason for the improvement gap between Baseline and AVST is that, AVST employs superior model architectures, such as visual-spatial attention, self-supervised matching loss with two-stage training. These advanced strategies facilitate the multimodal interactions, and indirectly prevents AVQA model from over-exploiting partial information.

Qualitative Results For the first sample in Fig. 6, though both baseline and our method select the correct answer for the given question, their reasoning behaviors may be different from their attention weights. Baseline model tends to evenly focus on temporal video frames and audio waves,

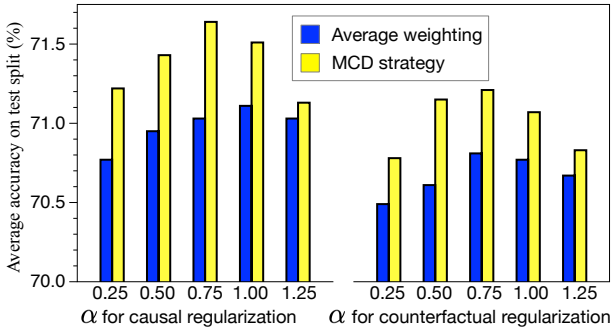

Figure 5: The performance of average weighting and our MCD to integrate multiple shortcut debiasing under various settings of trade-off factor α.

and we deduce that the correct answer may be determined by language bias in training set, which also supported by its higher weight W_{cf}^{qs} for uni-modal debiasing. On the contrary, COCA is capable to concentrate more on the question-related video frames and audio clips, and increase its explainability over visual-auditory information. For the second sample, we can see that the uncertainty for temporal audio-visual attention distributions in COCA are obviously less than those in baseline model, so as to unbiasedly obtain the correct answer 'Piano'. The illustrated superiority of temporal attention for COCA also support its remarkable performance on audio-visual temporal question type in Tab. 1.

Related Works

Audio-Visual Question Answering. In the last few years, several question answering (QA) tasks have realized impressive progress in different modalities, including text question answering (Rajpurkar et al. 2016), visual question answering (Antol et al. 2015; Jang et al. 2017; Yu et al. 2015; Lee, Cheon, and Han 2021; Ye and Kovashka 2021; Tanaka, Nishida, and Yoshida 2021; Lao et al. 2021b), audio question answering (Fayek and Johnson 2020), and video question answering (Zhu et al. 2017). Beyond these QA tasks that concern reasoning from single- or cross-modality data, Audio-Visual Question Answering (AVQA) requires models to comprehensively understand three modalities and perform spatio-temporal reasoning over audio-visual scenes (Li et al.

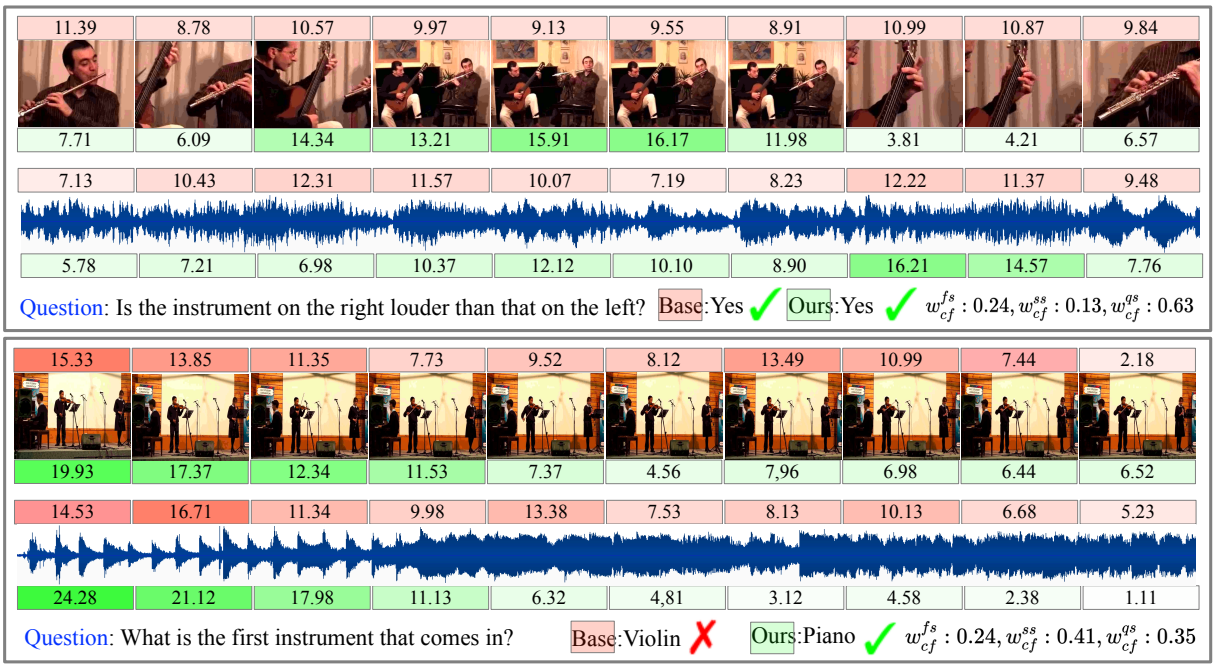

Figure 6: Two AVQA samples for qualitative analysis of our COCA based on the baseline model. The numbers above and below the videos and audio waves are the temporal visual and auditory attention weights for question answering, among which numbers in red and green refer to the weights generated from baseline and COCA approach respectively.

Method	V Ques	A Ques	A-V Ques	Average
Baseline	72.25	71.38	67.80	69.61
+COCA	74.24	74.61	69.47	71.64
AVST	74.00	74.06	69.54	71.52
+COCA	75.23	75.42	69.96	72.33
VQAT-based	71.59	71.14	67.66	69.32
+COCA	75.43	73.25	68.78	70.81
ConvLSTM-based	70.54	72.57	66.42	68.37
+COCA	72.87	73.12	67.15	69.74

Table 3: Performance for visual, auditory, visual-auditory and all questions types on MUSIC-AVQA test set.

2022a). Particularly, for audio-visual questions in AVQA task, without considering either visual or auditory information would fail to deduce correct answers reasonably. MUSIC-AVQA (Li et al. 2022a) and Pano-AVQA (Yun et al. 2021) are proposed to explore such high-level reasoning tasks. Pano-AVQA (Yun et al. 2021) designs a robust transformer-based multimodality encoder for addressing cross-modality inference, while its dataset has not been released yet. MUSIC-AVQA (Li et al. 2022a) proposes a spatio-temporal grounding approach for long-term audio-visual scenes. *Unlike those AVQA approaches that develop advanced network architectures, our method attends mainly on reducing multi-shortcut data biases of AVQA models.*

Causal Inference for Debiasing. Causal inference is the process of determining the independent, actual effect of a particular phenomenon (Pearl 2009), which has been explored for years in psychology, politics and epidemiology (Keele 2015; Richiardi, Bellocco, and Zugna 2013). Recently, some works (Agarwal, Shetty, and Fritz 2020; Cadene et al. 2019; Chen et al. 2020; Niu et al. 2021; Li et al. 2022b; Yang et al. 2021b; Pan et al. 2022) introduced causal inference into VQA tasks. Most of recent solutions to reduce the bias in VQA can be grouped into three categories: strengthening visual grounding (Selvaraju et al. 2019; Wu and Mooney 2019), weakening language prior (Ramakrishnan, Agrawal, and Lee 2018; Abbasnejad et al. 2020; Lao et al. 2022; Chen et al. 2020; Lao et al. 2021a), and implicit/explicit data argumentation (Abbasnejad et al. 2020; Zhu et al. 2020). *Different from the aforementioned methods that focus only on uni-modality shortcut biases, our work is the first to study a multi-variable causal analysis for the potential joint-modality biases in the AVQA task, and correspondingly proposes a collaborative debiasing method MCD that is tailor-made for multi-shortcut biases.*

Conclusion

In this paper, we have introduced a model-agnostic COllaborative CAusal (COCA) Regularization to overcome unimodal and joint-modal biases in AVQA models. Specifically, a novel Bias-centered Causal Regularization was proposed to alleviate shortcut bias by causal intervention to introspect from counterfactual/factual scenarios. We also presented a Multi-Shortcut Collaborative Debiasing strategy to facilitate the cooperation among different Causal Regularization. We validated the effectiveness of COCA through extensive comparative and ablative studies. Moving forward, we are going to extend COCA for other textual-visual-auditory tasks which require comprehensive reasoning.

Acknowledgments

This work was supported mainly by the LIACS Media Lab at Leiden University, and partially by the China Scholarship Council and the NSF of China under grant 62102061.

References

Abbasnejad, E.; Teney, D.; Parvaneh, A.; Shi, J.; and Hengel, A. v. d. 2020. Counterfactual vision and language learning. In *CVPR*, 10044–10054.

Agarwal, V.; Shetty, R.; and Fritz, M. 2020. Towards Causal VQA: Revealing and Reducing Spurious Correlations by Invariant and Covariant Semantic Editing. In *CVPR*, 9690–9698.

Agrawal, A.; Batra, D.; Parikh, D.; and Kembhavi, A. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *CVPR*, 4971–4980.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. Vqa: Visual question answering. In *ICCV*, 2425–2433.

Cadene, R.; Dancette, C.; Cord, M.; Parikh, D.; et al. 2019. Rubi: Reducing unimodal biases for visual question answering. *NeurIPS*, 841–852.

Chen, L.; Yan, X.; Xiao, J.; Zhang, H.; Pu, S.; and Zhuang, Y. 2020. Counterfactual samples synthesizing for robust visual question answering. In *CVPR*, 10800–10809.

Fan, C.; Zhang, X.; Zhang, S.; Wang, W.; Zhang, C.; and Huang, H. 2019. Heterogeneous memory enhanced multimodal attention model for video question answering. In *CVPR*, 1999–2007.

Fayek, H. M.; and Johnson, J. 2020. Temporal reasoning via audio question answering. *IEEE-ACM T AUDIO SPE*, 28: 2283–2294.

Glymour, M.; Pearl, J.; and Jewell, N. P. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Hershey, S.; Chaudhuri, S.; Ellis, D. P.; Gemmeke, J. F.; Jansen, A.; Moore, R. C.; Plakal, M.; Platt, D.; Saurous, R. A.; Seybold, B.; et al. 2017. CNN architectures for large-scale audio classification. In *ICASSP*, 131–135.

Jang, Y.; Song, Y.; Yu, Y.; Kim, Y.; and Kim, G. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *CVPR*, 2758–2766.

Keele, L. 2015. The statistics of causal inference: A view from political methodology. *Political Analysis*, 23(3): 313–335.

Lao, M.; Guo, Y.; Chen, W.; Pu, N.; and Lew, M. S. 2022. VQA-BC: Robust Visual Question Answering Via Bidirectional Chaining. In *ICASSP*, 4833–4837.

Lao, M.; Guo, Y.; Liu, Y.; Chen, W.; Pu, N.; and Lew, M. S. 2021a. From Superficial to Deep: Language Bias Driven Curriculum Learning for Visual Question Answering. In *ACM MM*, 3370–3379.

Lao, M.; Guo, Y.; Pu, N.; Chen, W.; Liu, Y.; and Lew, M. S. 2021b. Multi-stage hybrid embedding fusion network for visual question answering. *Neurocomputing*, 423: 541–550.

Lee, D.; Cheon, Y.; and Han, W.-S. 2021. Regularizing attention networks for anomaly detection in visual question answering. In *AAAI*, volume 35, 1845–1853.

Li, G.; Wei, Y.; Tian, Y.; Xu, C.; Wen, J.-R.; and Hu, D. 2022a. Learning to Answer Questions in Dynamic Audio-Visual Scenarios. In *CVPR*, 19108–19118.

Li, X.; Song, J.; Gao, L.; Liu, X.; Huang, W.; He, X.; and Gan, C. 2019. Beyond rnns: Positional self-attention with co-attention for video question answering. In *AAAI*, volume 33, 8658–8665.

Li, Y.; Wang, X.; Xiao, J.; Ji, W.; and Chua, T.-S. 2022b. Invariant grounding for video question answering. In *CVPR*, 2928–2937.

Niu, Y.; Tang, K.; Zhang, H.; Lu, Z.; Hua, X.-S.; and Wen, J.-R. 2021. Counterfactual vqa: A cause-effect look at language bias. In *CVPR*, 12700–12710.

Pan, Y.; Li, Z.; Zhang, L.; and Tang, J. 2022. Causal Inference with Knowledge Distilling and Curriculum Learning for Unbiased VQA. *ACM T MULTIM COMPUT*, 18(3): 1–23.

Pearl, J. 2009. Causal inference in statistics: An overview. *Statistics surveys*, 3: 96–146.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. SQuAD: 100, 000+ Questions for Machine Comprehension of Text. In *EMNLP*, 2383–2392.

Ramakrishnan, S.; Agrawal, A.; and Lee, S. 2018. Overcoming language priors in visual question answering with adversarial regularization. In *NeurIPS*, volume 31, 1548–1558.

Richiardi, L.; Bellocco, R.; and Zugna, D. 2013. Mediation analysis in epidemiology: methods, interpretation and bias. *International journal of epidemiology*, 42(5): 1511–1519.

Schwartz, I.; Schwing, A. G.; and Hazan, T. 2019. A simple baseline for audio-visual scene-aware dialog. In *CVPR*, 12548–12558.

Selvaraju, R. R.; Lee, S.; Shen, Y.; Jin, H.; Ghosh, S.; Heck, L.; Batra, D.; and Parikh, D. 2019. Taking a hint: Leveraging explanations to make vision and language models more grounded. In *ICCV*, 2591–2600.

Tanaka, R.; Nishida, K.; and Yoshida, S. 2021. Visualmrc: Machine reading comprehension on document images. In *AAAI*, volume 35, 13878–13888.

Wu, J.; and Mooney, R. 2019. Self-critical reasoning for robust visual question answering. *NeurIPS*, 32: 8604–8614.

Yang, A.; Miech, A.; Sivic, J.; Laptev, I.; and Schmid, C. 2021a. Just Ask: Learning To Answer Questions From Millions of Narrated Videos. In *ICCV*, 1686–1697.

Yang, X.; Zhang, H.; Qi, G.; and Cai, J. 2021b. Causal Attention for Vision-Language Tasks. In *CVPR*, 9847–9857.

Yao, L.; Chu, Z.; Li, S.; Li, Y.; Gao, J.; and Zhang, A. 2021. A survey on causal inference. *TKDD*, 15(5): 1–46.

Ye, K.; and Kovashka, A. 2021. A case study of the shortcut effects in visual commonsense reasoning. In *AAAI*, volume 35, 3181–3189.

Yu, L.; Park, E.; Berg, A. C.; and Berg, T. L. 2015. Visual madlibs: Fill in the blank description generation and question answering. In *ICCV*, 2461–2469.

Yun, H.; Yu, Y.; Yang, W.; Lee, K.; and Kim, G. 2021. Pano-avqa: Grounded audio-visual question answering on 360deg videos. In *ICCV*, 2031–2041.

Zhu, L.; Xu, Z.; Yang, Y.; and Hauptmann, A. G. 2017. Uncovering the temporal context for video question answering. *IJCV*, 124(3): 409–421.

Zhu, X.; Mao, Z.; Liu, C.; Zhang, P.; Wang, B.; and Zhang, Y. 2020. Overcoming Language Priors with Self-supervised Learning for Visual Question Answering. In *IJCAI*, 1083–1089.