Supplementary Material for Variational Causal Inference Network for Explanatory Visual Question Answering

Dizhan Xue[1,2] Shengsheng Qian[1,2] Changsheng Xu[1,2,3]
[1]State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]Peng Cheng Laboratory
xuedizhan17@mails.ucas.ac.cn, {shengsheng.qian, csxu}@nlpr.ia.ac.cn

A. Computation Details

To facilitate a wide range of readers, we introduce some computation details in this section.

A.1. Evidence Lower Bound (ELBO)

We use evidence lower bound (ELBO) [13] in Equation 1. To deduce the ELBO, we have

$$\begin{aligned}
&\log p(A'|M) \\
=&\log \int_F p(A', F|M) dF \\
=&\log \int_F p(A', F|M) \frac{q(F|M)}{q(F|M)} dF \\
=&\log E_{q(F|M)}\Big[\frac{p(A', F|M)}{q(F|M)}\Big] \\
\geq& E_{q(F|M)}\Big[\log \frac{p(A', F|M)}{q(F|M)}\Big]
\end{aligned} \tag{17}$$

by using Jensen's Inequality. Then, we have

$$\begin{aligned}
&\log p(A'|M) \\
\geq& E_{q(F|M)}\Big[\log \frac{p(A', F|M)}{q(F|M)}\Big] \\
=& E_{q(F|M)}\Big[\log \frac{p(A'|M, F)p(F, M)}{q(F|M)}\Big] \\
=& E_{q(F|M)}[\log p(A'|M, F) + \log p(F|M) - \log q(F|M)]
\end{aligned} \tag{18}$$

A.2. KL Divergence between Gaussian Distributions

We use KL divergence between two Gaussian distributions in Equation 3. Denote $q(F) = q(F|E') = \mathcal{N}(\boldsymbol{\mu}_{E'}, \boldsymbol{\Sigma}_{E'})$ and $p(F) = p(F|E^*) = \mathcal{N}(\boldsymbol{\mu}_{E^*}, \boldsymbol{\Sigma}_{E^*})$ for simplicity, the KL divergence can be deduced as follows:

$$\begin{aligned}
&KL(q(F) \parallel p(F)) \\
=&\int_F q(F) \log \frac{q(F)}{p(F)} dF \\
=&\int_F \frac{1}{2} q(F) \Big[\log \frac{|\boldsymbol{\Sigma}_{E^*}|}{|\boldsymbol{\Sigma}_{E'}|} - (F - \boldsymbol{\mu}_{E'})^T \boldsymbol{\Sigma}_{E'}^{-1} (F - \boldsymbol{\mu}_{E'}) \\
&+ (F - \boldsymbol{\Sigma}_{E^*})^T \boldsymbol{\Sigma}_{E^*}^{-1} (F - \boldsymbol{\Sigma}_{E^*})\Big] dF \\
=&\frac{1}{2}\Big[\log \frac{|\boldsymbol{\Sigma}_{E^*}|}{|\boldsymbol{\Sigma}_{\boldsymbol{E'}}|} - tr\big\{E_q[(F - \boldsymbol{\Sigma}_{E'})(F - \boldsymbol{\Sigma}_{E'})^T]\boldsymbol{\Sigma}_{E'}^{-1}\big\} \\
&+ E_q[(F - \boldsymbol{\Sigma}_{E^*})^T \boldsymbol{\Sigma}_{E^*}^{-1} (F - \boldsymbol{\Sigma}_{E^*})]\Big] \\
=&\frac{1}{2}\Big[\log \frac{|\boldsymbol{\Sigma}_{E^*}|}{|\boldsymbol{\Sigma}_{\boldsymbol{E'}}|} - tr\{\boldsymbol{I}_{d_f}\} + tr\{\boldsymbol{\Sigma}_{E^*}^{-1}\boldsymbol{\Sigma}_{E'}\} + \Delta\boldsymbol{\mu}^T \boldsymbol{\Sigma}_{E^*}^{-1} \Delta\boldsymbol{\mu}\Big] \\
=&\frac{1}{2}\Big[\log \frac{|\boldsymbol{\Sigma}_{E^*}|}{|\boldsymbol{\Sigma}_{\boldsymbol{E'}}|} - d_f + tr\{\boldsymbol{\Sigma}_{E^*}^{-1}\boldsymbol{\Sigma}_{E'}\} + \Delta\boldsymbol{\mu}^T \boldsymbol{\Sigma}_{E^*}^{-1} \Delta\boldsymbol{\mu}\Big],
\end{aligned} \tag{19}$$

where d_f is the dimension of F and $\boldsymbol{I}_{d_f}$ is a identity matrix of size d_f.

A.3. Normalized Weighted Geometric Mean (NWGM)

We use Normalized Weighted Geometric Mean (NWGM) [2] in Equation 14. For simplicity, we denote $\boldsymbol{x} = [F|\boldsymbol{T}_0]$ and $f(\boldsymbol{x}) = LN(\boldsymbol{x})\boldsymbol{W}_a$ which is a linear transformation. Then, from the results in [2, 11] that $E_F[softmax(f(\boldsymbol{x}))] \approx NWGM[f(\boldsymbol{x})]$, we have

$$\begin{aligned}
&E_F\{p(A|M, F)\} = E_F[softmax(f(\boldsymbol{x}))] \\
\approx& NWGM[f(\boldsymbol{x})] = softmax(E_F\{f(\boldsymbol{x})\}) \\
=& softmax(f(E_F\{\boldsymbol{x}\})) = softmax(f([E\{F\}|\boldsymbol{T}_0])) \\
=& softmax(f([\boldsymbol{\mu}_{E^*}|\boldsymbol{T}_0])),
\end{aligned} \tag{20}$$

where we use $E_F\{f(\boldsymbol{x})\} = f(E_F\{\boldsymbol{x}\})$ since f is a linear transformation.

B. Correlation of Objectives Proposed in Section 4.1 and Section 4.2

We describe the optimization objectives in a causal perspective in Section 4.1 and deduce objectives by variational inference in Section 4.2. In this section, we indicate the objectives proposed in Section 4.1 and Section 4.2 are consistent.

The objectives proposed in Section 4.1 are as follows:

$$\begin{aligned}\mathcal{O}_1 &= \arg\min -P(E=E'|M)P(A=A'|M,E=E')\\ \mathcal{O}_2 &= \arg\min KL(P(F|E')\|P(F|E^*)).\end{aligned} \tag{21}$$

The objectives proposed in Section 4.2 are as follows:

$$\begin{aligned}\mathcal{L}_{ans} &= \arg\min -E_{q(F|E')}[\log p(A'|M,F)]\\ &\quad + KL(q(F|E') \parallel p(F|E^*))\\ \mathcal{L}_{exp} &= \arg\min -\log p(E'|M).\end{aligned} \tag{22}$$

Therefore, we have

$$\begin{aligned}&\mathcal{L}_{ans}+\mathcal{L}_{exp}\\ &= \arg\min -E_{q(F|E')}[\log p(A'|M,F)]\\ &\quad + KL(q(F|E') \parallel p(F|E^*)) - \log p(E'|M)\\ &= \arg\min -\log P(A'|M,E') + KL(P(F|E')\|P(F|E^*))\\ &\quad - \log P(E'|M)\\ &= \arg\min -\log P(E'|M) - \log P(A'|M,E')\\ &\quad + KL(P(F|E')\|P(F|E^*))\\ &= \log \mathcal{O}_1 + \mathcal{O}_2.\end{aligned} \tag{23}$$

Therefore, the objectives deduced in Section 4.2 are consistent with the objectives proposed in Section 4.1.

C. Comparisons of Experimental Details

As VQA is a sub-task in our work but we follow the experimental protocol of EVQA [3], it is important to clarify the differences in experimental details compared to prior VQA studies conducted on the GQA dataset, which is extended to GQA-REX dataset in EVQA.

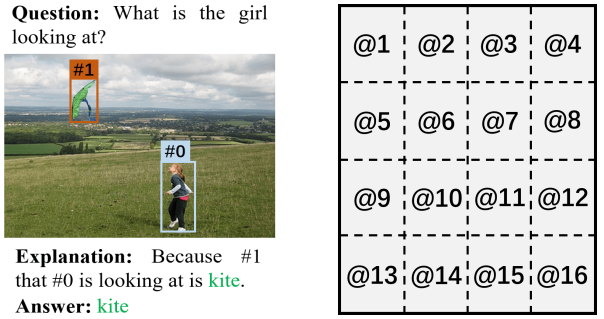
C.1. Difference in Data Split.

In training, we use the balanced GQA-REX training set (of 912,934 samples), which is slightly smaller than the balanced GQA training set (of 943,000) as samples without explanations are removed. In validation, we use the balanced GQA-REX validation set (of 127,900 samples), which is also slightly smaller than the balanced GQA validation set (of 132,062) as samples without explanations are removed. In test, we use the standard GQA-REX test set (of 11,183,447 samples), which is identical to the standard GQA test set.

However, prior VQA studies may not use the balanced GQA training set for training. For instance, LXMERT [10], which is used as the backbone in our model, is finetuned on the standard (or named *all*) training set (of 14,305,356 samples) plus the standard validation set (of 2,011,853 samples) of GQA. Therefore, it is unfair to directly compare the answering accuracy of LXMERT and ours though our test accuracy is higher than LXMERT's. Similarly, state-of-the-art VQA methods are typically trained on the standard GQA training set or standard GQA training + validation set, such as CFR [8], DPT [7], VinVL [16], TRRNet [14], and OSCAR [5].

C.2. Difference in Visual Objects.

We follow REX [3], TRRNet [14], and LXMERT [10] to extract 36 objects in every image by pretrained Faster R-CNN. The file of extracted features is identical to that used by REX and LXMERT. However, some VQA methods also use pretrained Faster R-CNN but extract 50 objects in every image, such as CFR [8], DPT [7], VinVL [16], and OSCAR [5]. It is not surprising to find using more objects can usually improve the answering accuracy.



(a) An example in GQA-REX dataset (b) Image segmentation in GQA-REX

Figure 1. Task definition in GQA-REX dataset.

D. Consistency Metric

In this section, we introduce motivation and computation details of our proposed **Consistency (Con.)** metric on GQA-REX dataset. As an example shown in Figure 1 (a), we can find that the explanation is consistent with the answer because it directly contains the answer "kite". Motivated by this finding, we have analyzed the whole GQA-REX dataset to dig out whether the answer always occurs in the explanation. We have finally found that except for two special kinds of questions, the answer or related direction token always occurs in the explanation. The first kind is asking for yes or no, such as "(qid = 07452748) **Question**: Is the cheese to the left of the food on the plate? **Explanation**: Because #9 is to the left of #1. **Answer**: yes". The second kind is asking for the common attribute of two objects, such as "(qid = 00226745) **Ques-**

Explanation	Because #21 is parrot.	Because middle #14 is blue.	Because #33 is located at @5.	Because #35 is located at @8.	Because #1 to the right of #34 is table.	Because #26 is to the left of #0.	Because #25 hanging on #29 is lamp.
Answer	parrot	blue	left	right	table	left	chandelier
Con.	1	1	1	1	1	1	0

Explanation	Because #8 to the right of #33 is pepper.	Because black #23 is located at @12.	Because yellow #30 is located at @1.	Because #3 in front of #8 is woman.	Because #10 that #1 is behind is cow.	Because there is #24 is brown and old.	Because both #34 and #16 are wood.
Answer	cucumber	left	right	girl	horse	yes	material
Con.	0	0	0	0	0	excluded	excluded

Figure 2. Examples of computing our proposed Consistency (Con.) metric. Consistent tokens in explanations with corresponding answers are colored green. The answers ($\in B$) of excluded types of samples are colored red.

tion: What do the box and the taxi have in common? **Explanation**: Because both #5 and #33 are white. **Answer**: color". Fortunately, only answers to the above two kinds of questions are in $B = \{yes, no, color, material, shape\}$. Therefore, we exclude these two kinds of questions while computing Con. by checking the ground truth answers of questions. Moreover, in GQA-REX, every image is segmented into 16 regions as shown in Figure 1 (b). Therefore, an answer of "left" or "right" can relate to tokens in $L = \{@1, @2, @5, @6, @9, @10, @13, @14\}$ or $R = \{@3, @4, @7, @8, @11, @12, @15, @16\}$ in the explanation. For example, "(qid = 07159849) **Question**: On which side of the image is the black backpack? **Explanation**: Because black #32 is located at @1. **Answer**: left".

Motivated by the above findings, we calculate Con. of a single sample as follows:

$$Con. = \begin{cases} 1 & , A = \text{“}left\text{”} \wedge (A \in E \vee \exists t(t \in L \wedge t \in E)) \\ 1 & , A = \text{“}right\text{”} \wedge (A \in E \vee \exists t(t \in R \wedge t \in E)) \\ 1 & , A \in E, \\ 0 & , A \notin E, \end{cases} \tag{24}$$

where A and E are predicted answer and explanation, and we only compute Con. of samples of which the ground truth answers are not in B. To verify whether Con. can precisely reflect the consistency correlation, we computed Con. of the ground truth annotations on GQA-REX, which is 99.85%. We have carefully checked the samples of which Con.=0 and found it is due to annotation mistakes in the dataset. Therefore, the ground truth supports the rationality of Con. metric. In experiments, we report the average Con. score on GQA-REX validation set.

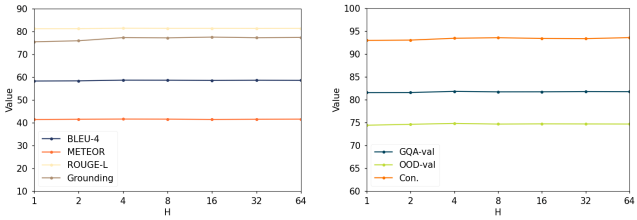
In Figure 2, we show various examples of computing Con. on real results predicted by REX and VCIN.

E. Implementation Details

We adopt mini-batch Adam [4] optimizer to optimize VCIN. The mini-batch size is 128 and the initial learning rate is 1e-5 for all trainable parameters. The dimensions d_f, d_g, and d_o of hidden units are set as 768. We take the multi-head trick for all attention layers with the head number 4. We set the number L of Transformer layers as 2 and the maximum length T of explanations as 18. The MC sampling number H is set as 4. The whole model is implemented by Pytorch and trained on two Tesla V100 GPUs.

F. Additional Experiments

Due to the space limitation, we include more experiments in this section.



(a) (b)

Figure 3. Hyper-parameter Analysis of Monte Carlo sampling number H.

F.1. Hyper-parameter Analysis

In this section, we analyze the sensitivity of Monte Carlo sampling number H in Equation 15, which is set as 4 by default. We vary H in $\{1, 2, 4, 8, 16, 32, 64\}$ and report the performance of VCIN under the same training settings. As shown in Figure 3, the performance of VCIN slightly improves when H increases from 1 to 4, and becomes stable when H increases even larger. These results show our VCIN does not rely on large H, which significantly reduces the training costs. This is because samples per class in train-

ing ($\frac{912,934}{1,824} \approx 500$) is sufficient and several epochs of training with $H = 4$ can effectively eliminate the bias brought by Monte Carlo sampling.

Table 1. Performance comparisons among causal models on GQA-REX. The best results are highlighted in bold.

Model	BLEU-4	METEOR	CIDEr	Grounding	GQA-val	Con.
REX-LXMERT	54.79	39.51	466.01	70.79	78.19	84.90
REX-CF-LMT	54.61	39.63	468.68	70.66	79.29	84.97
VCIN	**58.65**	**41.57**	**519.23**	**77.33**	**81.80**	**93.44**

G. Comparisons to Existing Causal VQA

Since causal inference has been already studied for VQA, we clarify the difference between existing causal VQA methods and our VCIN in this section.

Existing causal VQA methods [1, 6, 12, 9, 15] focus on **eliminating** biased dependency (aka, spurious correlations) in learning to improve the accuracy of question answering. For instance, Agarwal et al. [1] propose automated semantic image manipulations to alleviate spurious correlations while learning. Niu et al. [9] propose a counterfactual inference framework to capture and mitigate language bias in VQA. Yang et al. [15] propose a causal attention mechanism to reduce the confounding bias which can mislead attention modules.

Differently, our proposed variational causal inference aims at **establishing** the causal correlation between the predicted answers and explanations to improve the answer-explanation consistency. Furthermore, as we have analyzed in Experiments, since our model can utilize the explanation information for answer prediction, the answering accuracy is also improved. Due to the huge difference of considered problems, our VCIN can also be integrated with the existing causal VQA methods, which can be left to future work.

In Table 1, we show the results of REX-CF-LMT that is built by applying CF-VQA [9] to REX-LXMERT. We can observe that while REX-CF-LMT improves answering accuracy (GQA-val), it cannot improve the quality of generated explanations or the answer-explanation consistency. Therefore, CF-VQA does not solve the problem of effective and credible reasoning for EVQA

H. Limitations

One of our major limitations is that we do not handle the bias existing in the dataset, which has been found in previous work [9, 15] and may affect the generalization of models.

References

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

[2] Pierre Baldi and Peter Sadowski. The dropout learning algorithm. *Artificial intelligence*, 210:78–122, 2014.

[3] Shi Chen and Qi Zhao. Rex: Reasoning-aware and grounded explanation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15586–15595, 2022.

[4] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[5] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020.

[6] Zujie Liang, Weitao Jiang, Haifeng Hu, and Jiaying Zhu. Learning to contrast the counterfactual samples for robust visual question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 3285–3292, 2020.

[7] Yuhang Liu, Wei Wei, Daowan Peng, and Feida Zhu. Declaration-based prompt tuning for visual question answering. *arXiv e-prints*, pages arXiv–2205, 2022.

[8] Binh X Nguyen, Tuong Do, Huy Tran, Erman Tjiputra, Quang D Tran, and Anh Nguyen. Coarse-to-fine reasoning for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4558–4566, 2022.

[9] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12700–12710, 2021.

[10] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, 2019.

[11] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10760–10770, 2020.

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

[13] Xitong Yang. Understanding the variational lower bound. *variational lower bound, ELBO, hard attention*, 22:1–4, 2017.

[14] Xiaofeng Yang, Guosheng Lin, Fengmao Lv, and Fayao Liu. Trrnet: Tiered relation reasoning for compositional visual question answering. In *European Conference on Computer Vision*, pages 414–430. Springer, 2020.

[15] Xu Yang, Hanwang Zhang, Guojun Qi, and Jianfei Cai. Causal attention for vision-language tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9847–9857, 2021.

[16] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021.