DAS-VQA: Dual Awareness Systems for Improving Generalization in Visual Question Answering

Trang Nguyen Naoaki Okazaki
Tokyo Institute of Technology,
Department of Computer Science, School of Computing
trang.nguyen@nlp.c.titech.ac.jp okazaki@c.titech.ac.jp

Abstract

Multimodal reasoning is a crucial factor of generalizability in the Visual Question Answering (VQA) task. Recent studies deal with generalization by refining unimodal models but pay less attention to the combination of multimodalities. We propose Dual Awareness Systems VQA (DAS-VQA), a novel framework to improve the Out-of-Distribution generalization by enhancing multimodal reasoning. DAS-VQA consists of two components: (1) the input processing component to identify the purpose of a question, and (2) the selection component for a proper strategy to give an answer. The experimental results show that DAS-VQA improves the generalization on real-life images and medical datasets without extra human annotations.

1 Introduction

The Visual Question Answering (VQA) task is to provide an answer to the questions regarding an image. Recent studies (Niu et al., 2021; Wen et al., 2021; Niu and Zhang, 2021) raised an issue that the models tend to answer a question while ignoring the input image, which hinders the Out-of-Distribution (OOD) generalization. For more details, OOD generalization is the ability to perform well in OOD cases rather than just recalling the co-occurrence of data during the independent and identically distributed (IID) training (Zhang et al., 2021; Kawaguchi et al., 2022). The reasons come from the biases in human knowledge that cause some correlations between the question and answer distribution (Niu et al., 2021; Wen et al., 2021). For instance, the question *Is this ...?* is usually supposed to be a Yes/No question (*e.g.* Is this a cat?) instead of a multiple choice question (*e.g.* Is this a cat or dog?). Therefore, there is a chance to give a correct answer just by replying *Yes* or *No*.

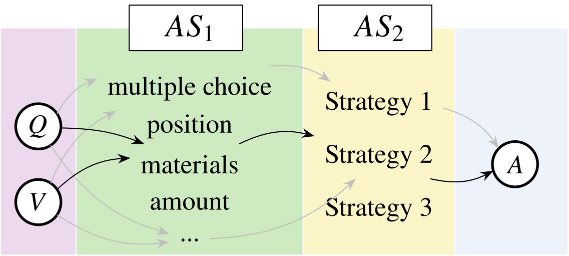


Figure 1: DAS-VQA overview: AS_1 (green area) implicitly recognizes question's purposes, and AS_2 (yellow area) selects a proper strategy to give the answer.

Some promising approaches to solve this issue are (1) reducing the linguistic correlation (Niu et al., 2021; Wen et al., 2021), (2) strengthening the visual processing (Yang et al., 2020), and (3) balancing the answer distribution by generating new image-question pairs (Gokhale et al., 2020). However, these approaches focus only on either language or the visual modality without considering the causality among input properties such as question types, objects, or backgrounds.

In fact, the VQA task requires not only the joint of visual and language processing but also the multimodal reasoning between them (Wang et al., 2022b; Nguyen et al., 2022). In addition, it is impossible to cover all combinations of the visual and language properties in a single dataset (Cao et al., 2021) due to the diversity of data in real life (Gokhale et al., 2020). Therefore, the reasoning across modularities should be considered as a core aspect to solve the bias issue and improve OOD generalization (Cao et al., 2021).

In this work, we propose Dual Awareness Systems VQA (DAS-VQA) as a novel framework to improve the OOD generalization by emphasizing multimodal reasoning, inspired by the causality theory (Daniel et al., 2015; Wang et al., 2022a). Assuming that question types lead to different strategies to answer the question, DAS-VQA consists of two components called the Awareness System (AS): (1) the first one is placed in the inputs processing part, which is responsible for implicitly recognizing the question's purpose; and (2) the second one takes the processed inputs and the question purpose into account to select one from a list of strategies to predict the answer.

Our contributions are summarized as follows: (1) We propose DAS-VQA as a novel framework to improve OOD generalization by enhancing the multimodal reasoning that is compatible with a diversity of multimodal tasks and domains; and (2) DAS-VQA is the first work for VQA task that represents the multimodal reasoning as the interaction of double mediators in a causal graph, which corresponds to the two layers of cognition.

2 Preliminaries

2.1 The causal-effect in general

A basic form of a causal relation is defined as $X \rightarrow Y$ with the treatment X and outcome Y. To explore the reasons of X *causes* Y, Pearl and Mackenzie (2018) mentioned *mediator* to dissect the effect into direct and indirect effects.

2.1.1 Mediators in causality

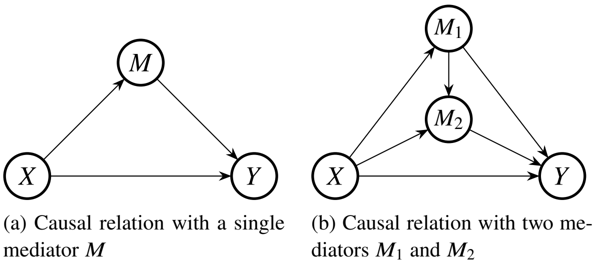
A single mediator M forms an indirect path $X \rightarrow M \rightarrow Y$ as described in Figure 2a. On the other hand, considering two ordered mediators M_1 and M_2 in Figure 2b, we have three indirect paths:

- Case 1: $X \rightarrow M_1 \rightarrow Y$: only through M_1
- Case 2: $X \rightarrow M_2 \rightarrow Y$: only through M_2
- Case 3: $X \rightarrow M_1 \rightarrow M_2 \rightarrow Y$: through both M_1, M_2

Following Daniel et al. (2015), we denote Y_i-jkl is the outcome conditioned by $i, j, k, l \in \{0, 1\}$. i equals to 1 when the treatment X is given and 0 otherwise when a virtual value is defined for X. j, k, and l equal to 1 when Y is affected by the indirect path described as *Case 1*, *Case 2*, and *Case 3*, respectively, and 0 otherwise. For instance, Y_1-001 represents the outcome that is affected by a given X and two factors: $X \rightarrow Y$ and $X \rightarrow M_1 \rightarrow M_2 \rightarrow Y$.

2.1.2 Total Effect and related terms

The term Total Effect (TE) described in Equation 1 compares the effect of X on Y through any indirect paths. The general form of Natural Direct Effect (NDE) is defined in Equation 2[1] to analyze a particular indirect case. In addition, the TE can be decomposed into the sum of NDE and Total Indirect Effect (TIE) as in Equation 3.



(a) Causal relation with a single mediator M

(b) Causal relation with two mediators M_1 and M_2

Figure 2: General forms of the causal graphs

$$\mathrm{TE} = \mathrm{Y}_1\text{-}111 - \mathrm{Y}_0\text{-}000 \quad (1)$$

$$\mathrm{NDE}\text{-}jkl = \mathrm{Y}_1\text{-}jkl - \mathrm{Y}_0\text{-}jkl \quad (2)$$

$$\mathrm{TIE} = \mathrm{TE} - \mathrm{NDE} \quad (3)$$

2.2 The causal-effect view in VQA

Consider the causal graph for the VQA task depicted in Figure 3a in which the inputs V and Q cause an answer A, and a mediator K represents the commonsense knowledge space. We have: **Direct paths**: $Q \rightarrow A$, $V \rightarrow A$ and **Indirect path**: $(V, Q) \rightarrow K \rightarrow A$. Notice that the direct paths represent the effects only from the question or image, while the indirect path represents the relation of the question, image, and the knowledge space to cause an answer. Therefore, the causality approach for the VQA task aims to enhance the indirect effect and eliminate the direct effects.

3 Proposed method: DAS-VQA

3.1 Dual Awareness Systems

The DAS-VQA, illustrated in Figure 1, provides an architecture backbone to promote the effects of multimodal reasoning between the input image, question, and commonsense knowledge to yield an answer in the VQA task. The key idea of DAS-VQA is to construct two awareness systems (AS) to have (1) distinct *approaches* in understanding the multimodal input and (2) various *strategies* in giving the answer by assuming that different questions' purposes lead to diverse ways to answer a question. Subsequently, DAS-VQA learns to operate a large number of multimodal

[1] We have eight cases of NDE corresponding to the increase of the binary string from 000 to 111 (Daniel et al., 2015)

reasoning flows, which is interpreted as choosing an appropriate pair of *approach* and *strategy* for a particular multimodal input. Therefore, DAS-VQA is able to deal with the diversity of multimodal combinations, which improves the OOD generalization.

Technically, the causal view of DAS-VQA, as presented in Figure 3c, contains the given values of treatment as (v, q) that causes an answer A, controlled by the two ASs denoted as mediators AS_1 and AS_2. Subsequently, any paths that do not go through AS_1 (*e.g.* $q \rightarrow A$, $q \rightarrow AS_2 \rightarrow A$) are considered as *unimodal* paths since they do not produce any multimodal understanding from the input pair. Likewise, any paths that do not go through AS_2 (*e.g.* $(v, q) \rightarrow AS_1 \rightarrow A$) are considered as *monolithic* paths since they do not involve separated strategies to give the answer. Finally, DAS-VQA emphasizes multimodal reasoning by eliminating paths without a completed reasoning flow, including *unimodal* and *monolithic* paths. Notice that the question types and strategies are implicitly distinguished by the model during training, not explicitly defined by humans.

3.2 Implementation details

Following the notation in Section 2.1.1, we denote Z_I-$S_1S_2S_{12} \in \mathbb{R}^N$ is the predicted answer controlled by $I, S_1, S_2, S_{12} \in \{0, 1\}$ (N is the vocabulary size). Specifically, I is 1 when the input pair is given; S_1 is 1 when the path corresponding to *Case 1* exists and 0 otherwise. S_2 and S_{12} are defined similarly with *Case 2* and *Case 3*.

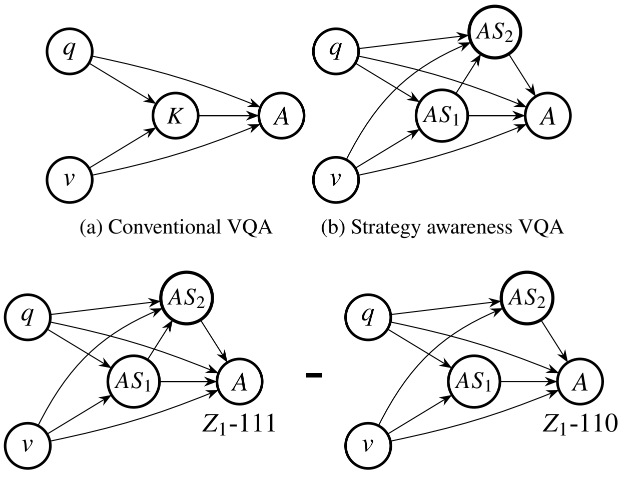
Let x denote the multimodal representation of the input pair computed by AS_1. Next, define the list of answers $z \in \mathbb{R}^N$ computed by Neural Network (NN) models: $z_v = W_v(v)$ and $z_q = W_q(q)$ as direct effects; $z_{v^*} = W_{v^*}(v)$ and $z_{q^*} = W_{q^*}(q)$ as indirect effects only through AS_2; $z_{MF} = W_{MF}(x)$ as an indirect effect only through AS_1 in which W_{MF} is designed as $\underline{m}$onolithic $\underline{f}$unciton and $z_{AS} = W_{AS}(x)$ as indirect effect through both of ASs.

$$Z_1\text{-}111 = z_q + z_v + z_{q^*} + z_{v^*} + z_{MF} + z_{AS} \quad (4)$$

$$Z_1\text{-}110 = z_q + z_v + z_{q^*} + z_{v^*} + z_{MF} \quad (5)$$

Training AS_1 and the NNs above are trained by *Cross Entropy Loss* on Z_1-111, which is described in Equation 4, with the target as the correct answer in the training data.

Inference Following Equation 3, the predicted answer probability is defined as the subtraction of Z_1-111 as TE and Z_1-110 (Equation 5) as NDE, with the meaning of eliminating effects from uncompleted reasoning paths.



(a) Conventional VQA (b) Strategy awareness VQA

(c) DAS-VQA

Figure 3: General forms of the causal graphs

For more details about the model design of this study, we refer the readers to Appendix B.

4 Experiments and Results

Our experiments validate two hypotheses: (1) the awareness of the question's purpose is beneficial to improving the VQA performance, and (2) the awareness of multimodal reasoning helps enhance the OOD generalization.

4.1 Experiment setup

4.1.1 Datasets

To examine the first hypothesis, we conduct an experiment on four datasets in two domains: (1) real-life: VQA-CPv2 (Agrawal et al., 2017) and VQAv2 and (2) medical: PathVQA (He et al., 2021) and VQA-RAD (Lau et al., 2018). To examine the second hypothesis, we observe results on the VQA-CPv2 dataset since this is an OOD dataset with significant differences in answer distribution per question category between the training and test sets.

4.1.2 Baselines

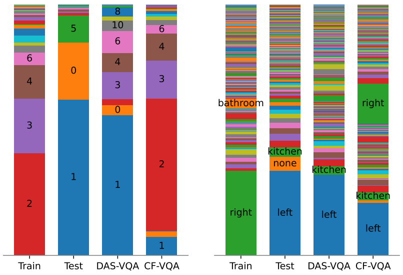
In VQA-CPv2 and VQAv2, we compare DAS-VQA and CFVQA (Niu et al., 2021), which use a similar approach, i.e., the causal view of the VQA task to overcome the language priors. CFVQA attempts to eliminate the effect of the question-only branch, which is distinct from our method that promotes multimodal reasoning. We compare DAS-

Table 1: Comparison on VQA-CPv2 and VQAv2. The bolded values indicate the best results.

Test set	VQA-CPv2 Test				VQAv2 Val			
Method	Overall	Y/N	Num.	Other	Overall	Y/N	Num.	Other
CF-VQA	$55.1^{\pm0.1}$	$90.6^{\pm0.3}$	$21.5^{\pm0.9}$	$45.6^{\pm0.2}$	$60.9^{\pm0.2}$	$81.1^{\pm0.2}$	**43.8**$^{\pm0.4}$	$50.1^{\pm0.1}$
DAS-VQA	**57.8**$^{\pm0.3}$	**91.1**$^{\pm0.3}$	**41.6**$^{\pm0.2}$	**46.4**$^{\pm0.1}$	**62.2**$^{\pm0.6}$	**81.4**$^{\pm0.3}$	**43.8**$^{\pm0.4}$	**52.4**$^{\pm0.2}$

Table 2: Overall score comparison on PathVQA and VQA-RAD. The bolded values indicate the best results.

Test set	PathVQA	VQA-RAD
MMQ-VQA	$48.8^{\pm0.12}$	$68.2^{\pm0.2}$
DAS-VQA	**50.9**$^{\pm0.3}$	**70.0**$^{\pm0.3}$



(a) How many ... ?

(b) Which ... ?

Figure 4: Answers distributions on VQA-CPv2

VQA to MMQ-VQA (Binh D. Nguyen, 2019) on PathVQA and VQA-RAD as the first work that reports results on PathVQA and the state-of-the-art on VQA-RAD. We report the mean and standard error over 5 seeds for all methods.

4.2 Experiment results

4.2.1 Quantitative Results

The IID results on VQAv2, PathVQA, and VQA-RAD are presented in Tables 1 and 2. Overall, DAS-VQA outperforms the baselines in all answer categories. Taking a deeper look at the OOD result on VQA-CPv2, DAS-VQA robustly outperforms the baseline by a large margin, especially +20.1 point in the *Number* type.

4.2.2 Qualitative Results

Debiased answer distribution As illustrated in Figure 4, DAS-VQA exhibits the OOD generalizability when overcoming the biased answers in training set on multiple question categories, whereas the baseline shows the biased distribution of the memorized answers.

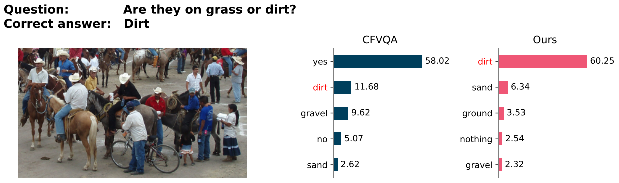


Figure 5: Sample of debiased case. DAS-VQA recognize the multiple choice question and give a correct answer

Debiased sample Figure 5 provides a comparison of debiasing results from the baseline and DAS-VQA. DAS-VQA answers the question correctly by recognizing the question purpose (the multiple choice question) accurately, although the baseline is trapped in the Yes/No question type and gives an incorrect answer. For further discussion on qualitative results, we refer the readers to Appendix D.

5 Related Work

Causality approaches in VQA Niu et al. (2021) creates the question-only, (*i.e.* $Q \rightarrow A$) to basically capture the linguistic biases. They utilize the counterfactual training objective similarly to Equation 3 to subtract the bias from the conventional answer distribution to obtain a debiased one. In contrast, DAS-VQA extends the biases not only from the linguistic but also from the monolithic strategy of the model for different question purposes.

Multimodal reasoning in VQA Wang et al. (2022b) create multimodal reasoning by combining the two knowledge graphs of visual-level with extracted objects and concept-level with multimodal information. In contrast to DAS-VQA , we define multimodal reasoning as the choices of paths in the causal graph to understand the multimodal and select a proper strategy to predict the answer.

6 Conclusion

In this study, we proposed DAS-VQA as a novel framework that improves the OOD generalization in the VQA task by promoting multimodal reasoning without any additional human annotations or labels. The experiment results demonstrate that DAS-VQA outperforms the baselines in both IID and OOD cases with real-life images and the medical domain datasets.

7 Acknowledgement

These research results were partly obtained from the commissioned research (No. 225) by National Institute of Information and Communications Technology (NICT), Japan.

References

Agrawal, A., Batra, D., Parikh, D., and Kembhavi, A. (2017). Don't just assume; look and answer: Overcoming priors for visual question answering. In **CVPR**.

Binh D. Nguyen, Thanh-Toan Do, B. X. N. T. D. E. T. Q. D. T. (2019). Overcoming data limitation in medical visual question answering. In **MICCAI**.

Cao, Q., Wan, W., Wang, K., Liang, X., and Lin, L. (2021). Linguistically routing capsule network for out-of-distribution visual question answering. In **2021 IEEE/CVF International Conference on Computer Vision (ICCV)**, pages 1594–1603.

Daniel, R. M., De Stavola, B. L., Cousens, S. N., and Vansteelandt, S. (2015). Causal mediation analysis with multiple mediators. **Biometrics**, 71(1):1–14.

Gokhale, T., Banerjee, P., Baral, C., and Yang, Y. (2020). MUTANT: A training paradigm for out-of-distribution generalization in visual question answering. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pages 878–892, Online. Association for Computational Linguistics.

He, X., Cai, Z., Wei, W., Zhang, Y., Mou, L., Xing, E., and Xie, P. (2021). Towards visual question answering on pathology images. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)**, pages 708–718, Online. Association for Computational Linguistics.

Kawaguchi, K., Bengio, Y., and Kaelbling, L. (2022). Generalization in deep learning. In **Mathematical Aspects of Deep Learning**, pages 112–148. Cambridge University Press.

Lau, J. J., Gayen, S., Demner, D. L., and Abacha, A. B. (2018). Visual question answering in radiology (vqa-rad). In **Open Science Framework**.

Nguyen, B. X., Do, T., Tran, H., Tjiputra, E., Tran, Q. D., and Nguyen, A. (2022). Coarse-to-fine reasoning for visual question answering. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**, pages 4558–4566.

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**, pages 12695–12705.

Niu, Y. and Zhang, H. (2021). Introspective distillation for robust question answering. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, **Advances in Neural Information Processing Systems**, volume 34, pages 16292–16304. Curran Associates, Inc.

Pearl, J. and Mackenzie, D. (2018). **The Book of Why: The New Science of Cause and Effect**. Basic Books, Inc., USA, 1st edition.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. (2021). Toward causal representation learning. **Proceedings of the IEEE**, 109(5):612–634.

Wang, W., Lin, X., Feng, F., He, X., Lin, M., and Chua, T.-S. (2022a). Causal representation learning for out-of-distribution recommendation. In **Proceedings of the ACM Web Conference 2022**, WWW '22, page 3562–3571, New York, NY, USA. Association for Computing Machinery.

Wang, Y., Yasunaga, M., Ren, H., Wada, S., and Leskovec, J. (2022b). Vqa-gnn: Reasoning with multimodal semantic graph for visual question answering.

Wen, Z., Xu, G., Tan, M., Wu, Q., and Wu, Q. (2021). Debiased visual question answering from feature and sample perspectives. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, **Advances in Neural Information Processing Systems**.

Yang, X., Lin, G., Lv, F., and Liu, F. (2020). Trrnet: Tiered relation reasoning for compositional visual question answering. In **ECCV**.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. volume 64, page 107–115, New York, NY, USA. Association for Computing Machinery.

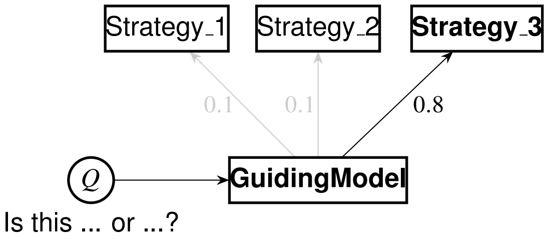


Figure 6: The strategy awareness VQA from the view of ICM

A The flexibility of DAS-VQA

The flexibility of DAS-VQA is described by two points: the design of the two awareness systems and the choices of NDE. First, technically, these two ASs can be designed as any deep learning architecture that serves the desired responsibility. Second, as mentioned in Section 2.1.1, we have eight forms of NDE, then, adapting to different training objectives of arbitrary multimodal tasks, different NDE is executed. The following are eight equations of the prediction probability Z_I-$S_1S_2S_{12}$ preparing for NDE computations.

$$\begin{aligned}
Z_1\text{-}000 &= z_q + z_v \\
Z_1\text{-}001 &= z_q + z_v + z_{AS} \\
Z_1\text{-}010 &= z_q + z_v + (z_{v^*} + z_{q^*}) \\
Z_1\text{-}011 &= z_q + z_v + (z_{v^*} + z_{q^*}) + z_{AS} \\
Z_1\text{-}100 &= z_q + z_v + z_{MF} \\
Z_1\text{-}101 &= z_q + z_v + z_{MF} + z_{AS} \\
Z_1\text{-}110 &= z_q + z_v + z_{MF} + (z_{v^*} + z_{q^*}) \\
Z_1\text{-}111 &= z_q + z_v + z_{MF} + (z_{v^*} + z_{q^*}) + z_{AS}
\end{aligned} \tag{6}$$

B Implementation details

B.1 Designs of awareness systems

In this study, we utilize the independent causal mechanisms (ICM) structure (Schölkopf et al., 2021) for both components. The main idea of the ICM is to decompose the original design into smaller parts that are learned independently. We manipulate one causal mechanism for each question's purpose and each strategy. In this work, all of the guiding models and mechanisms are designed as multilayer perceptron (MLP).

The general flow of ICM is illustrated in Figure 6, which contains a list of strategy models and a guiding model.

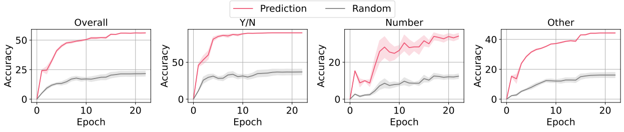


Figure 7: The causal impact of strategy selection. The *Prediction* values indicate results from the selected strategy by AS_2, while the *Random* values represent a random selection of strategies, collected from VQA-CPv2.

The key objective of ICM is to train the strategy models independently, with means the weight-updating process of this model does not affect other models' weight. The role of the guiding model is to select k strategies to be activated by a k-*hot* Gumbel max. Afterward, only the selected strategy would contribute to the output of ICM.

B.2 Monolithic models design

DAS-VQA utilizes the following Monolithic models: W_v, W_q, W_{MF}. In this work, we use MLP for these models.

C Evaluation method

Our evaluation methods follow the previous works (Niu et al., 2021; Binh D. Nguyen, 2019), Overall, we evaluate the predicted answer by word-by-word accuracy.

In addition, for VQA-CPv2 and VQAv2, we also report the accuracy of *Yes/No* and *Numbers* related questions, the rest of the question types denoted as *Other* type. While for PathVQA and VQA-RAD, we report the overall score only.

D Further discussion on qualitative results

The Causal Impact of Strategy Selection The causal impact of strategy selection is depicted in Figure 7. DAS-VQA proves the ability to learn a list of strategies independently and gain a significant distance in roles of each strategy by the drop of performance when just randomly selecting a strategy.