Multimodal Rationales for Explainable Visual Question Answering Supplementary Material

Kun Li[1], George Vosselman[1], Michael Ying Yang[2]
[1]University of Twente, [2]University of Bath
[1]{k.li, george.vosselman}@utwente.nl, [2]myy35@bath.ac.uk

In this supplementary document, we provide detailed explanations of the synthesized dataset in Sec. A. Additional ablation studies in terms of the input for the detector and the metric are provided in Sec. B. Moreover, we explain the societal impact in Sec. C and provide a discussion part in Sec. D.

A. Synthesized Dataset

For the process, briefly, we complement the visual explanations by incorporating the annotations from COCO [10]. We begin by aligning potential annotations with samples from the VQA-E [8] dataset based on shared objects. Figure 1 illustrates the overall process involved in synthesizing the multimodal EVQA dataset. We further analyze the synthesized multimodal EVQA dataset (VQA-E-Syn) and compare it with available datasets in Table 1. The distinctions from previous EVQA datasets are highlighted in the last three columns (*i.e.,* the quantity and format of multimodal rationales). For instance, in comparison to the VQA-X dataset [13], our VQA-E-Syn dataset provides more comprehensive and precise visual rationales. We believe this synthesized dataset will foster future research in multimodal EVQA, which will be made publicly available.

B. Additional Ablation Study

We further report two groups of ablation studies.

B.1. Ablation of Linguistic Features for Vision Predictor

To assess the impact of linguistic features on the vision predictor Grounding-DINO, we performed experiments on VQA-E-Syn using features derived from either questions ortextual rationales as input. We observe that the variant using features from input questions struggled (41.93% AP), as these do not provide comprehensive and clear referring information like visual grounding [7]. In contrast, features derived from textual rationales significantly enhanced the detection model, leading to 47.45% AP (a 5.52% improvement) in the MRVQA-E model. In addition, Fig. 2 shows

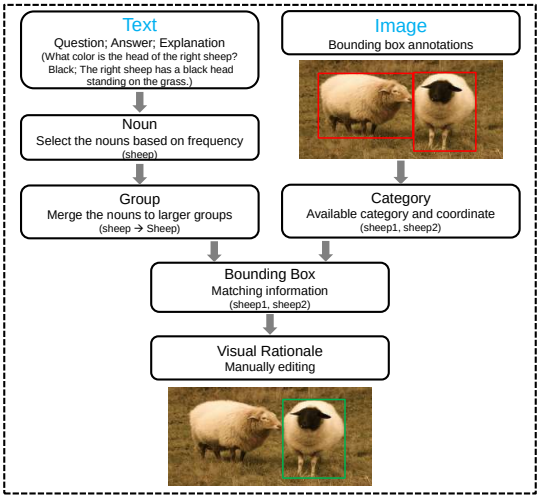

Figure 1. The overall process of the multimodal EVQA dataset synthesis. The text in the brackets shows an example.

a qualitative comparison between the variant and our approach using both MRVQA-E and MRVQA-C models. The results demonstrate that our proposed approach leveraged the generation of textual rationales to significantly enhance performance for multimodal EVQA.

B.2. Ablation of the New Metric

To verify whether the introduced metric vtS effectively measures the quality of the generated multimodal rationales, we performed a series of analyses in this part. Table 2 reports the results. We began by evaluating the textual similarity score (TS) derived from the GTE [9] model. The results in the first two columns show that the quality of textual rationales, as represented by TS, aligned well with the SPICE metric. Additionally, we explored various combinations of visual and textual aspects in the multimodal rationale evaluation. Specifically, we considered three options: (1) averaging AP and TS by taking their mean; (2) multiplying AP and TS; (3) dividing the result from (2) by the result from (1), which We refer to as vtS. We observe that

Table 1. Statistics for the comparison between the synthesized EVQA dataset and prior datasets. Bounding box and scene graph are abbreviated as B-box and S-G, respectively.

Dataset	Image	#Image	#Q&A	#TR	#VR	VR Format
VQA-HAT [3]	COCO	20K	59K	-	62K	Region
VQA-X [13]	COCO	28K	32K	41K	6K	Region
VQA-E [8]	COCO	108K	269K	269K	-	-
VCR [16]	Movies	99K	264K	264K	-	-
Vizwiz-G [1]	Phone	9K	9K	-	9K	Boundary
GQA-REX [2]	GQA	82K	1040K	1040K	82K	S-G
VQA-E-Syn	COCO	20,367	33,726	33,726	93,377	B-box

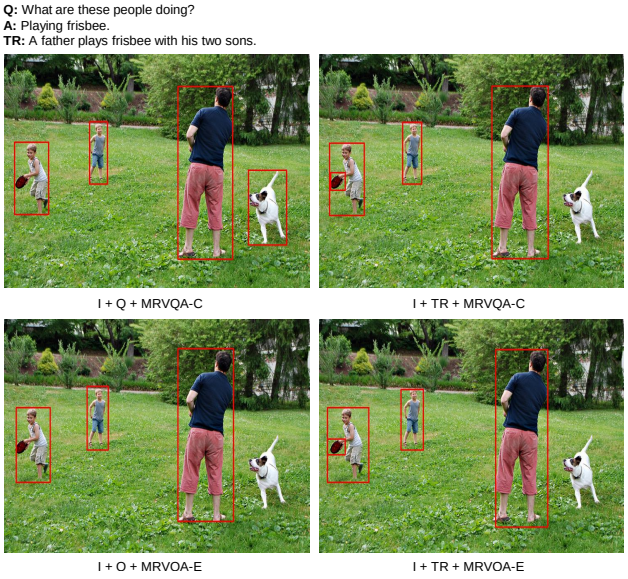

Figure 2. Examples of generated visual rationales (bounding boxes in red) with different linguistic inputs for our MRVQA-C and MRVQA-E models.

options (1) and (2) failed to accurately reflect model performance when there is a significant discrepancy between the two components (*e.g.,* comparing VCIN and MRVQA-C, and considering that visual predictions would deteriorate further if the model's performance were to decline). In contrast, option (3) mitigated the impact of large discrepancies, offering a more balanced assessment (*e.g.,* 0.3% with (3) compared to more than 0.7% with (1)). Consequently, we used the combination approach outlined in (3) as the new evaluation metric vtS for our multimodal EVQA task.

Furthermore, we tried to assess the introduced metric and compared it with manual evaluation results. We randomly selected 200 samples from VQA-E-Syn and employed three annotators to evaluate the quality of predictions across four aspects like VQA-E (including *Fluent*, *Correct*, *Relevant*, *Complementary*), using a grading system that ranges from 1 to 5, where 1 represents 'very poor' and 5 signifies 'very good'. Table 3 reports the results. The quality of generated rationales is good for human preference compared to DMRFNet. The vtS scores are also consistent with the human evaluation results.

Table 2. Ablation results of the multimodal evaluation metric on VQA-E-Syn.

Method	S	TS	AP	(1)	(2)	(3)
PJ-X [13]	15.32	68.32	38.43	53.38	26.26	49.19
VQA-E [8]	16.83	71.67	40.65	56.16	29.13	51.88
FME [14]	18.77	73.16	42.57	57.87	31.14	53.82
DMRFNet [17]	20.41	75.06	44.74	59.90	33.58	56.06
VCIN [15]	22.07	78.13	45.97	62.05	35.92	57.89
MRVQA-C	22.19	**79.57**	45.86	62.72	36.49	58.19
MRVQA-E	**23.68**	78.56	**47.45**	**63.00**	**37.28**	**59.16**

Table 3. Comparisons between human evaluation and the proposed vtS metric.

Method	Fluent	Correct	Relevant	Complementary	vtS (%)
DMRFNet	3.29	3.33	3.16	3.09	55.63
MRVQA-E	3.87	3.50	3.42	3.47	60.27
GT	4.64	4.44	4.04	3.93	80.82

C. Societal Impact

This paper in explanatory visual question answering presents an approach that holds significant societal implications. By advancing the capabilities of machines to understand and respond to human queries based on visual content, the system has the potential to enhance the human-computer interactions, leading to more intuitive and efficient communication. The societal implications of the answering system extend beyond technological advancements, offering solutions that can positively impact education, healthcare, and accessibility on a broader scale.

D. Discussion and Limitations

While our approach has shown advancements through extensive experiments across various datasets, there remain several areas for further exploration in future research. In the proposed model, we utilize the pre-trained CLIP model to represent input questions and images. However, the model may encounter problems when users' focus is on local details within images or when contextual information is limited. We show two failure cases in Fig. 3 to illustrate these problems. In the left example, our model predicted the wrong answer and failed to identify the relevant objects due to the challenge of distinguishing the target among numerous similar objects. This issue may stem from the CLIP model's strength in capturing global features from large-scale image-caption datasets, which might limit its ability to provide detailed local representations. In the example shown right, although our model accurately understood the cross-modal input and successfully located the relevant objects, the limited contextual information hampered the pre-

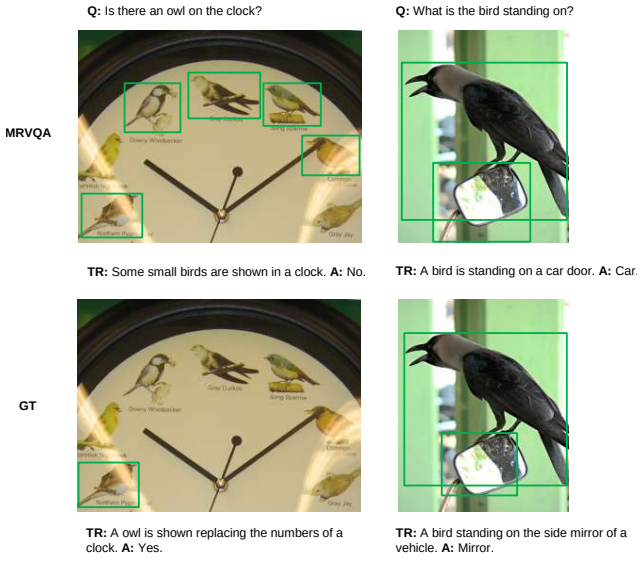

Figure 3. Two failure cases of our method.

cision of the linguistic prediction. The challenge arises because the zoomed-in local view complicates the recognition of the object (*e.g.,* mirror). This scenario emphasizes the importance of a multimodal EVQA model, which provides clear insights compared to a "black box" answering system. Therefore, a promising direction for future work is to explore the usage of models with more advanced feature representations and enhanced cross-modal recognition capabilities, such as VilBert[12] and BLIP [6]. Moreover, the zero-shot capabilities of recent powerful VLMs [4, 5, 11] offer promising potential for multimodal EVQA.

References

[1] Chongyan Chen, Samreen Anjum, and Danna Gurari. Grounding answers for visual questions asked by visually impaired people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19098–19107, 2022. 2

[2] 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. 2

[3] 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*, 163: 90–100, 2017. 2

[4] Chunyuan Li. Large multimodal models: Notes on cvpr 2023 tutorial. *arXiv preprint arXiv:2306.14895*, 2023. 3

[5] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024. 3

[6] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900, 2022. 3

[7] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 1

[8] Qing Li, Qingyi Tao, Shafiq Joty, Jianfei Cai, and Jiebo Luo. Vqa-e: Explaining, elaborating, and enhancing your answers for visual questions. In *Proceedings of the European Conference on Computer Vision*, pages 552–567, 2018. 1, 2

[9] Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*, 2023. 1

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755, 2014. 1

[11] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916, 2023. 3

[12] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[13] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8779–8788, 2018. 1, 2

[14] Jialin Wu and Raymond Mooney. Faithful multimodal explanation for visual question answering. In *Proceedings of the ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 103–112, 2019. 2

[15] Dizhan Xue, Shengsheng Qian, and Changsheng Xu. Variational causal inference network for explanatory visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2515–2525, 2023. 2

[16] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6720–6731, 2019. 2

[17] Weifeng Zhang, Jing Yu, Wenhong Zhao, and Chuan Ran. Dmrfnet: deep multimodal reasoning and fusion for visual question answering and explanation generation. *Information Fusion*, 72:70–79, 2021. 2