Multimodal Rationales for Explainable Visual Question Answering

Kun Li[1], George Vosselman[1], Michael Ying Yang[2]
[1]University of Twente, [2]University of Bath

Background

Explainable Visual Question Answering (EVQA): aims to link the predicted answer to the given image-question pair by providing explanations in different formats . **Motivation**: either single-track or compositional EVQA methods cannot support human-friendly explanations without additional information.

Contributions

- A new EVQA model that provides multimodal explanations for answer predictions through textual and visual rationale generation.
- A text alignment loss function.
- A new evaluation metric vtS from a visual perspective.
- Achieved state-of-the-art performance.

Formulation:

$$\{A, TR, VR\} = P(I, Q; \theta)$$

Loss:

$$L_{tr} = -\sum_{i}^{M}\sum_{t}^{l} \log p_\theta(w_t^i | f_{llm}^i, w_1^i, \dots, w_{t-1}^i)$$

$$L_{ta} = 1 - \sum_{i}^{M} \cos(f_{llm}^{\prime i}, f_{cross}^{\prime i})$$

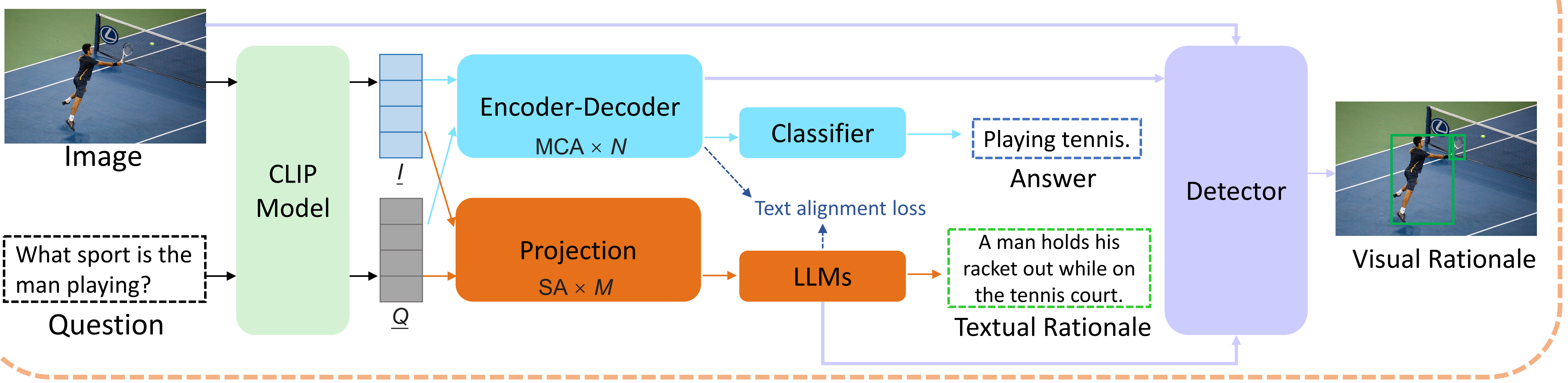
$$L = L_{ans} + L_{tr} + L_{vr} + \lambda L_{ta}$$

vtS score:

$$TS = \frac{1 + \cos(\mathrm{GTE}(T_{pred}), \mathrm{GTE}(T_{gt}))}{2}$$

$$vtS = \frac{2 \times TS \times AP}{TS + AP}$$

Method



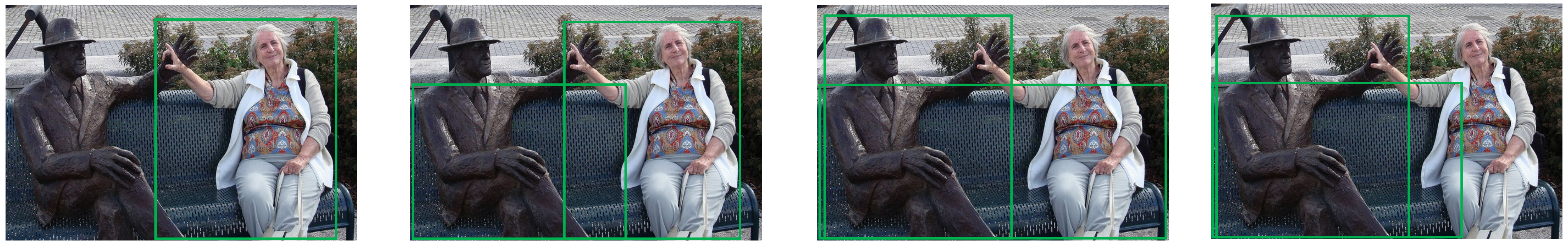
Quantitative Results

Table 1. Comparison results on the VQA-E-Syn.

Method	BELU-4	METEOR	ROUGE	CIDEr	SPICE	vtS	Number	Yes/No	Other	OA
PJ-X	8.78	16.94	35.65	89.31	15.32	49.19	-	-	-	-
VQA-E	8.93	17.02	35.96	90.84	16.83	51.88	54.26	63.98	66.45	64.67
DMRFNet	13.34	19.44	40.76	95.68	20.41	56.06	59.08	72.28	73.84	72.15
REX	14.71	21.35	41.86	100.75	21.08	56.38	59.73	72.91	74.18	72.84
OFA	16.38	22.09	42.74	103.25	22.65	56.58	60.48	73.26	75.37	73.49
VCIN	15.77	22.38	43.10	104.63	22.07	57.89	61.59	73.47	73.10	72.43
Ours-E	**16.97**	**23.02**	**44.28**	**107.04**	**23.68**	**59.16**	**65.11**	**76.40**	**75.87**	**75.01**

Qualitative Results

Q: Is the man on the bench alive?



TR: A woman is sitting outside. **A:** Yes.
DMRFNet

TR: The person pointing on the bench. **A:** No.
VCIN

TR: A statue is sitting on the bench. **A:** No.
Ours-E

TR: A statue is sitting on the bench. **A:** No.
GT

Table 2. Comparison results on the VQA-X.

Method	B	M	R	C	S	OA
VQA-E	20.7	18.6	44.2	75.6	15.6	70.2
VCIN	25.9	21.6	48.5	93.7	**19.4**	77.7
Ours-C	**26.6**	**22.0**	**48.9**	**98.4**	19.2	**78.8**

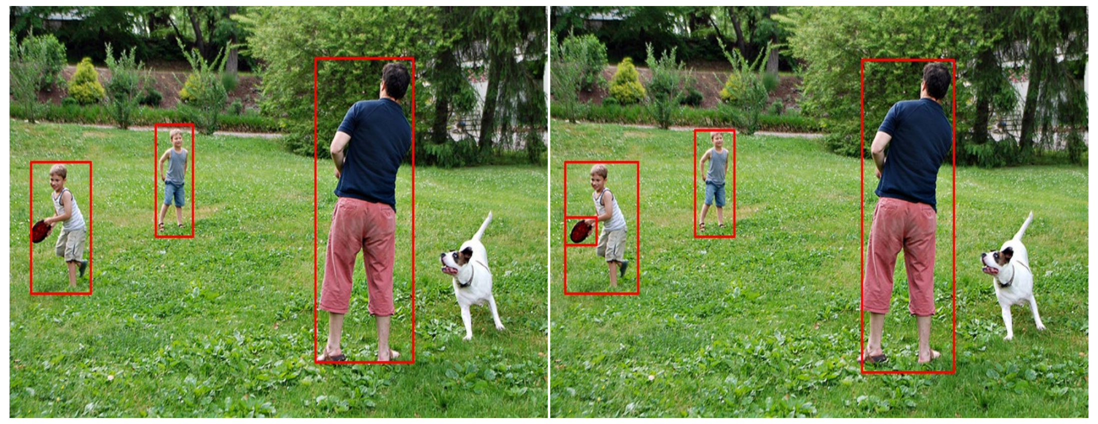
Table 3. Comparison results on the GQA-REX.

Method	B	M	R	C	S	OA
VQA-E	42.6	34.5	73.6	358	40.4	57.2
VCIN	58.6	41.6	81.5	519	54.6	60.6
Ours-C	**58.8**	**42.0**	**83.1**	**528**	**55.0**	**61.3**

Q: What are these people doing?

A: Playing frisbee.

TR: A father plays frisbee with his two sons.



Ours + Q   Ours + TR

Conclusion

We propose a multimodal EVQA method that predicts not only answers but also textual and visual rationales to support the answers. A text alignment loss function and an evaluation metric are introduced to inspire further research in the EVQA field.

k.li@utwente.nl