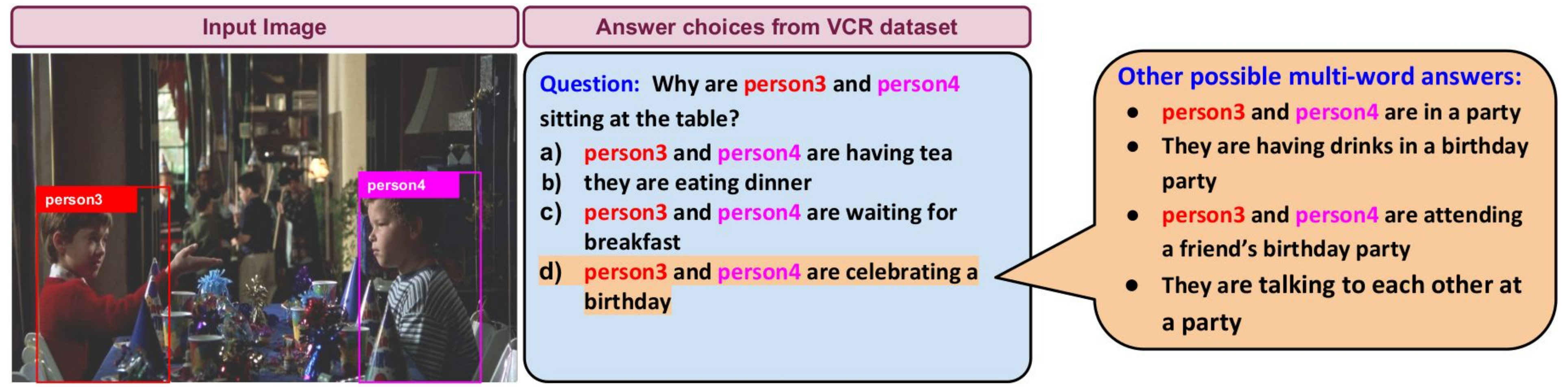
Beyond VQA: Generating Multi-word Answers and Rationales to Visual Questions

Radhika Dua*, Sai Srinivas Kancheti*, Vineeth N Balasubramanian
{radhika, cs21resch01004, vineethnb}@iith.ac.in
Indian Institute of Technology Hyderabad, India

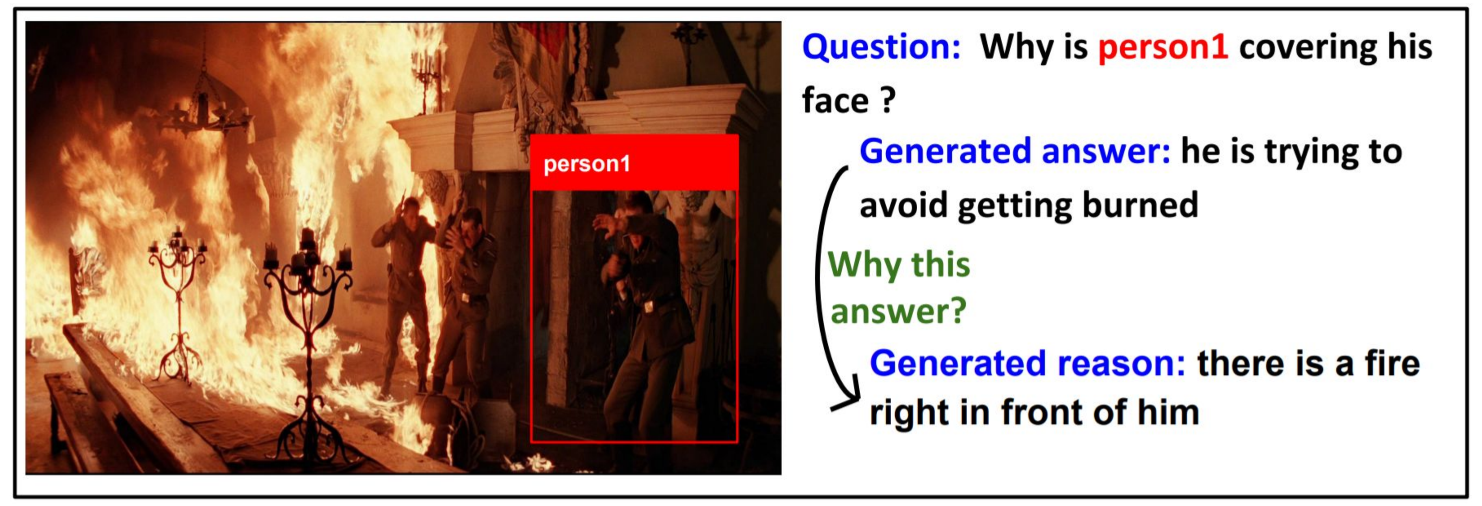
Motivation

- Visual Question Answering is a multi-modal task that aims to measure high-level visual understanding.
- Contemporary VQA models are restrictive in the sense that answers are obtained via classification over a limited vocabulary (in the case of open-ended VQA), or via classification over a set of multiple-choice-type answers.
- There can be many correct multi-word answers to a question which makes this classification setting restrictive.



Task Description

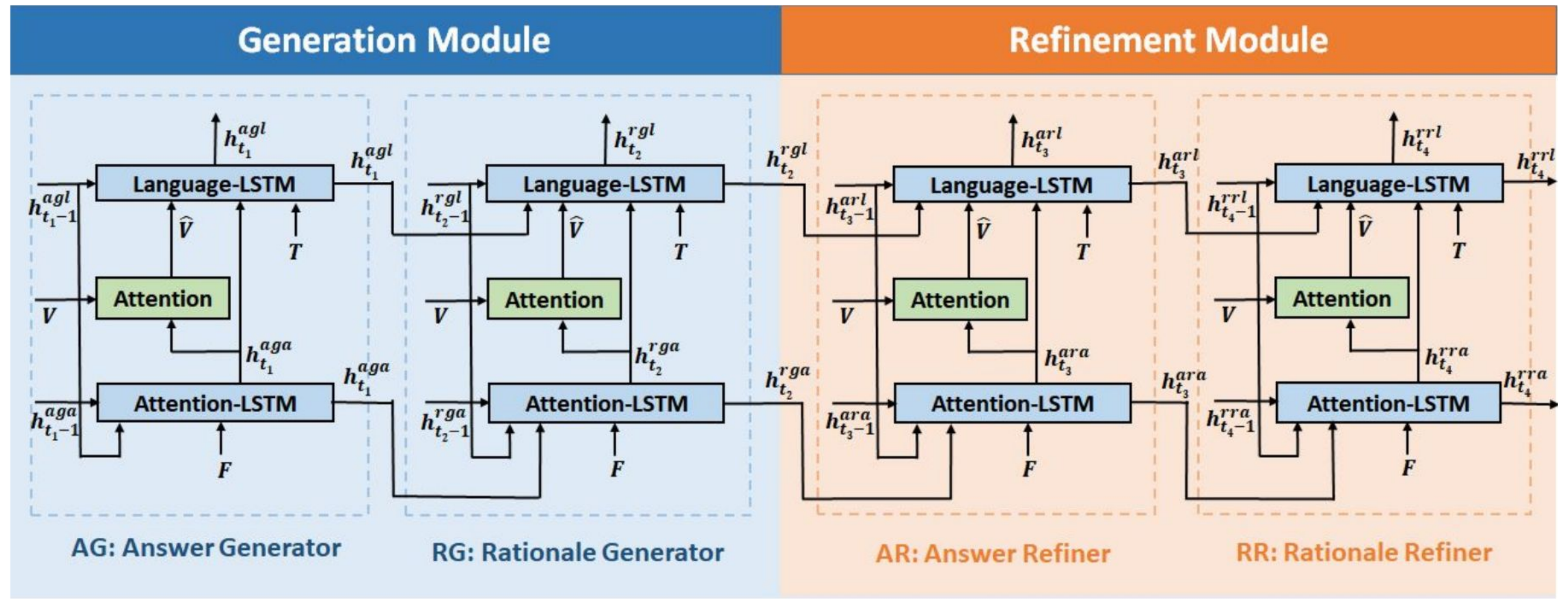
- We introduce a new task: **ViQAR** (Visual Question Answering and Reasoning), wherein a model must generate the complete answer and a rationale that seeks to justify the generated answer.



- The above example also illustrates the kind of visual question for which a **single-word answer is insufficient**.

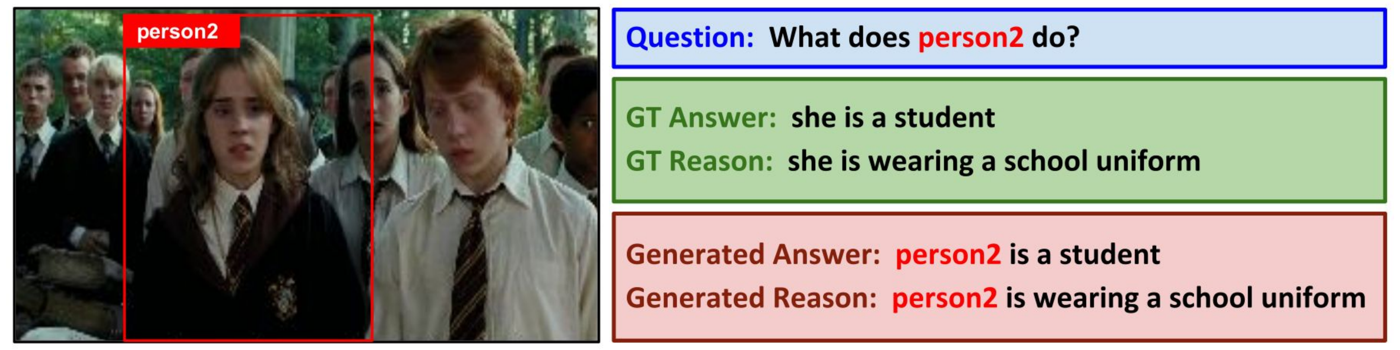
Methodology

- Humans often use a rationale to answer a question, and sometimes vice versa. This suggests a close interplay between answer and rationale. Inspired by this interplay, we propose a model for answer and rationale generation.
- Our architecture consists of the *Generation Module* and the *Refinement Module*, both comprising of two sequential, stacked LSTMs.
- Simplicity of our approach suggests the tractability of our task.

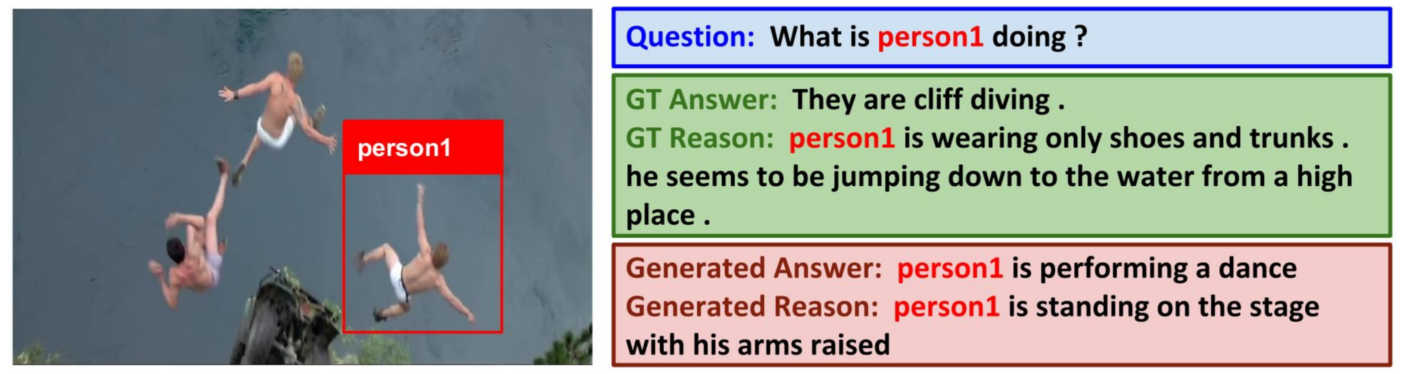


Qualitative Results

Example output from our proposed model:



Challenging input for which our model fails:



Quantitative Results

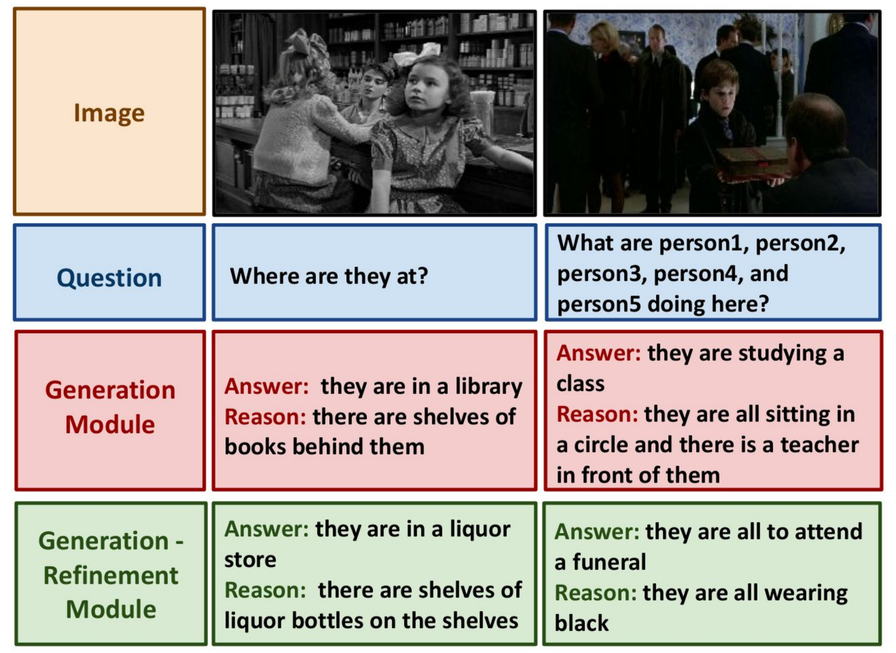
- We compare our proposed model and its variants against a basic two-stage LSTM model and a VQA model [1] as baselines [CS = cosine similarity].

Metrics	VQA-Baseline	Baseline	Q+I+C (Ours)	Q+I (Ours)	Q+C (Ours)
Univ Sent Encoder CS	0.419	0.410	**0.455**	0.454	0.440
Infersent CS	0.370	0.400	0.438	**0.442**	0.426
Embedding Avg CS	0.838	0.840	0.846	**0.853**	0.845
Vector Extrema CS	0.474	0.444	**0.493**	0.483	0.475
Greedy Matching Score	0.662	0.633	**0.672**	0.661	0.657
METEOR	0.107	0.095	**0.116**	0.104	0.103
Skipthought CS	0.430	0.359	**0.436**	0.387	0.385
RougeL	0.259	0.206	**0.262**	0.232	0.236
CIDEr	0.364	0.158	**0.455**	0.310	0.298
F-BERTScore	0.877	0.860	**0.879**	0.867	0.868

- We also perform a human Turing test on the generated answers and rationales, and find that our model generates grammatically correct and consistent answers and rationales.

Ablation Results

- We compare our proposed architecture with variations in number of refinement modules.



Metrics	#Refine Modules		
	0	1	2
Univ Sent Encoder CS	0.453	**0.455**	0.430
Infersent CS	0.434	**0.438**	0.421
Embedding Avg CS	0.850	0.846	0.840
Vector Extrema CS	0.482	**0.493**	0.462
Greedy Matching Score	0.659	**0.672**	0.639
METEOR	0.101	**0.116**	0.090
Skipthought CS	0.384	**0.436**	0.375
RougeL	0.234	**0.262**	0.198
CIDEr	0.314	**0.455**	0.197
F-BertScore	0.868	**0.879**	0.861

[1] Peter Anderson et al. "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering". In: *CVPR*. 2018.