# Beyond VQA: Generating Multi-word Answer and Rationale to Visual Questions

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

In this supplementary section, we provide additional supporting information including:

- Results on VCR classification task (extension of Section 5 of main paper)
- Additional qualitative results of our model (extension of Section 5 of main paper)
- Qualitative results on impact of the Refinement module in our model, and a study on the effect of adding refinement module to VQA-E
- Qualitative results on transferring our model to the VQA task (extension of results in Section 5 of main paper)
- A discussion on existing objective evaluation metrics for this task, and the need to go beyond (extension of section 5 of our paper)

## 1. VCR classification task

We evaluate the performance of our model on the classification task. For every question, there are four answer choices and four rationale choices provided in the VCR dataset. We compute the similarity scores between each of the options and our generated answer/ rationale, and choose the option with the highest similarity score. Accuracy percentage for answer classification, rationale classification and overall answer+rationale classification (denoted as Overall) are reported in Table 1. Only samples that correctly predict *both* answers and rationales are considered for overall answer+rationale classification accuracy. The results show the difficulty of the `ViQAR` task, expounding the need for opening up this problem to the community.

## 2. Additional qualitative results

In addition to the qualitative results presented in Section 5 of the main paper, Figure 1 presents more qualitative results from our proposed model on the VCR dataset for the `ViQAR` task. We observe that our model is capable of generating answer-rationale pairs to complex subjective questions of the type: Explanation (why, how come), Activity (doing, looking, event), Temporal (happened, before, after, etc), Mental (feeling, thinking, love, upset), Scene (where, time) and Hypothetical sentences (if, would, could). For completeness of understanding, we also show a few more examples on which our model fails to generate a good answer-rationale pair in Figure 2. As stated earlier in Section 5, even on these results, we observe that our model does generate both answers and rationales that are grammatically correct and complete. Improving the semantic correctness of the generations will be an important direction of future work.

## 3. Impact of refinement module

Figure 3 provides a few examples to qualitatively compare the model with and without the refinement module, in continuation to the discussion in Section 6. We observe that the model without the refinement module fails to generate answers and rationale for complex image-question pairs. However, our proposed Generation-Refinement model is capable of generating a meaningful answer with a supporting explanation. Hence the addition of the refinement module to our model is useful to generate answer-rationale pairs to complex questions. We also performed a study on the effect of adding refinement to VQA models, particularly to VQA-E [Li *et al.*2018], which can be considered close to our work since it provides explanations as a classification problem (it classifies the explanation among a list of options, unlike our model which generates the explanation). To add refinement, we pass the last hidden state of the LSTM that generates the explanation along with joint representation to another classification module. However, we did not observe improvement in classification accuracy when the refinement module is added for such a model. This may be attributed to the fact that the VQA-E dataset consists largely of one-word answers to visual questions. We infer that the interplay of answer and rationale, which is important to generate a better answer and provide justification, is more useful in multi-word answer settings which is the focus of this work.

| Metrics | Q+I+C | | | Q+I | | | Q+C | | |
|---|---|---|---|---|---|---|---|---|---|
| | Answer | Rationale | Overall | Answer | Rationale | Overall | Answer | Rationale | Overall |
| Infersent | 34.90 | 31.78 | 11.91 | 34.73 | 31.47 | 11.68 | 30.50 | 27.99 | 9.17 |
| USE | 34.56 | 30.81 | 11.13 | 34.7 | 30.57 | 11.17 | 30.15 | 27.57 | 8.56 |

Table 1: Quantitative results on the VCR dataset. Accuracy percentage for answer classification, rationale classification and overall answer-rationale classification is reported.


Figure 1: *(Best viewed in color)* Qualitative results for `ViQAR` task from our Generation Refinement architecture. Blue box = question about image; Green = Ground truth; Red = Generated results from our proposed architecture. (Object regions shown on image is for reader's understanding and are not given as input to model.)

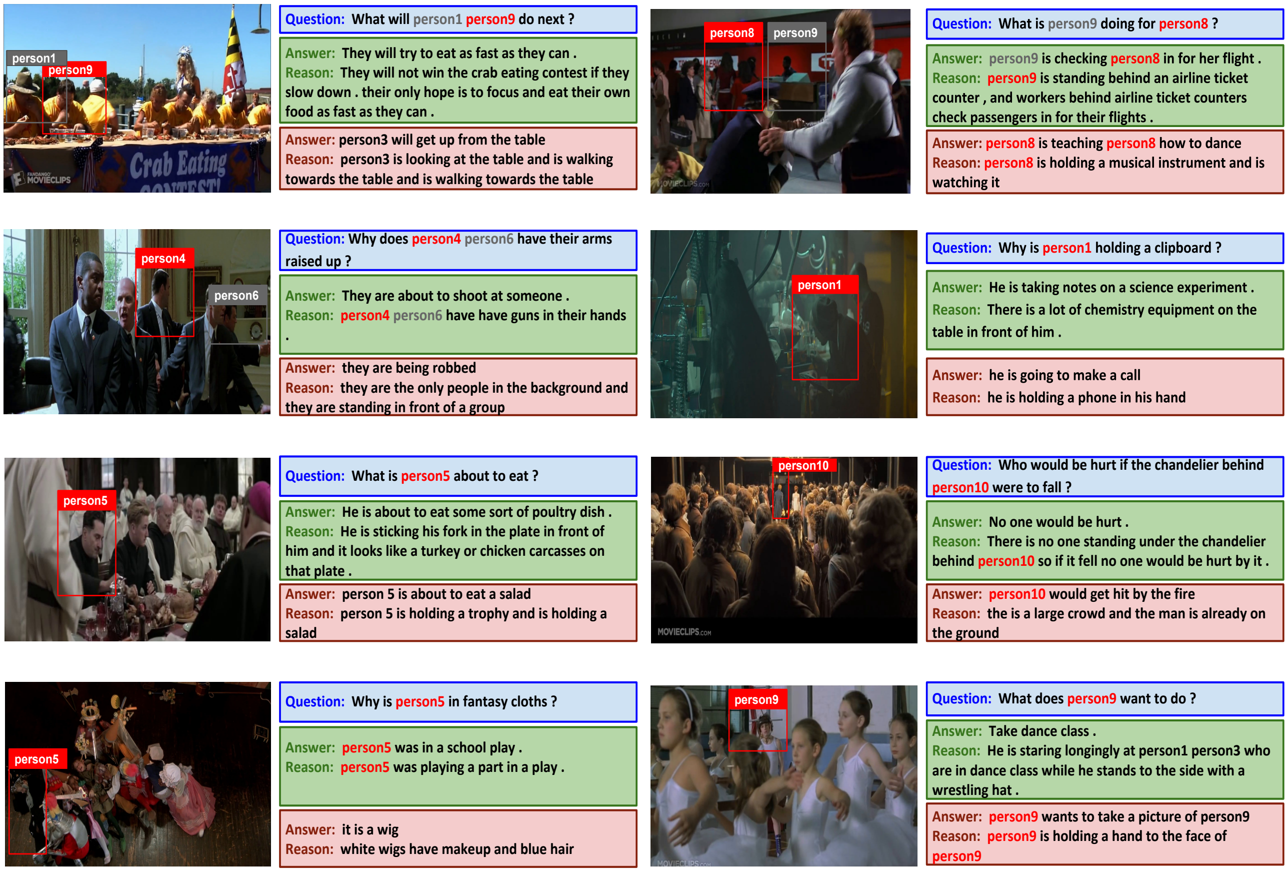



Figure 2: *(Best viewed in color)* Challenging examples for which our model fails to generate the semantically correct answer and rationale. Blue box = question about image; Green = Ground truth; Red = Generated results from our proposed architecture. (Object regions shown on image are for reader's understanding and are not given as input to model.)

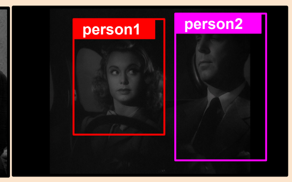
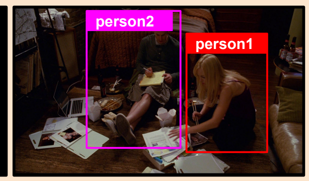
## 4. Qualitative results on transfer to VQA task

As stated in Section 6 of the main paper, we also studied whether the proposed model, trained on the VCR dataset, can provide answers and rationales to visual questions in standard VQA datasets (which do not have ground truth rationale provided). Figure 4 presents additional qualitative results for `ViQAR` task on the Visual7W dataset. We observe that our algorithm generalizes reasonably well to the other VQA dataset and generates answers and rationales relevant to the image-question pair, without any explicit training for this dataset. This adds a promising dimension to this work.

## 5. On objective evaluation metrics for generative tasks: A Discussion

Since `ViQAR` is a completely generative task, objective evaluation is a challenge, as in any other generative methods. Hence, for comprehensive evaluation, we use a suite of several well-known objective evaluation metrics to measure the performance of our method quantitatively. There are various reasons why our approach may seem to give relatively lower scores than typical results for these scores on other language processing tasks. Such evaluation metrics measure the similarity between the generated and ground-truth sentences. For our task, there may be multiple correct answers and rationales, and each of them can be expressed in numerous ways. Figure 5 shows a few examples of images and questions along with their corresponding ground-truth, generated answer-rationale pair, and corresponding evaluation metric scores. We observe that generated answers and rationales are relevant to the image-question pair but may be different from the ground-truth answer-rationale pair. Hence, the evaluation metrics reported here have low scores even when the results are actually qualitatively good (as evidenced in the Human Turing test results in Section 5 of the main paper). Thus, in this task, textual similarity to the ground truth may not be the only sign of the quality and may even indicate that the network is overfitting. We hence

| | | | | |
|---|---|---|---|---|
| Image |  |  |  |  |
| Question | What is person3 ' s profession ? | What will person2 do next? | Where is person1 going ? | Why is everyone in the room looking in a particular direction ? |
| Generation Module | Answer: person4 is a teacher Reason: person4 is standing in front of the classroom | Answer: car will take a drink from the cup Reason: car is holding a glass in his hand | Answer: she is going to meet her boyfriend Reason: she is dressed up and looks happy she is going to go to a date | Answer: they are watching something on the ground Reason: they are looking in the same direction and are looking at the same thing |
| Generation - Refinement Module | Answer: person3 is a doctor Reason: person3 is wearing a white coat and standing in front of a desk | Answer: person2 will order a drink Reason: person2 is at a bar | Answer: person1 is going to the kitchen to get some food Reason: person1 is walking towards a dining table with a plate of food in her hand | Answer: they are looking at something interesting Reason: they are looking in the same direction with a look of surprise on their faces |

| | | | | |
|---|---|---|---|---|
| Image |  |  |  |  |
| Question | What job is person8 doing right now ? | Where was person2 previously ? | Where could person1 person2 be driving from ? | What are person1 person2 drinking ? |
| Generation Module | Answer: person9 is working as a lawyer Reason: bartenders stand behind the bar at a table | Answer: he was at the bar Reason: he is holding a beer with a beer | Answer: they are riding a plan Reason: the car is in a vehicle and the vehicle is very narrow | Answer: they are drinking coffee Reason: they are holding a coffee cup and there is a coffee bottle on the table |
| Generation - Refinement Module | Answer: person8 is a waiter Reason: he is standing behind the bar and is wearing a uniform | Answer: he was outside Reason: he is wearing a coat and a hat | Answer: they might be driving to a hotel Reason: they are both wearing suits and ties and are in a car | Answer: they are drinking wine Reason: there is a bottle of wine in front of them |

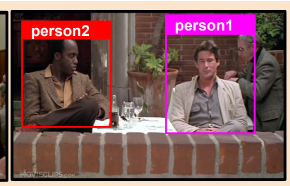
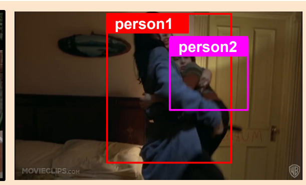
| | | | | |
|---|---|---|---|---|
| Image |  |  |  |  |
| Question | Who does dog2 belong to ? | What is person14 doing ? | Why does person1 seem annoyed ? | What is person1 doing to person2 ? |
| Generation Module | Answer: pottedplant 1 belongs to person 2 Reason: pottedplant 1 is on the ground next to person2 | Answer: he is giving a speech Reason: he is standing in front of a microphone and everyone is looking at him | Answer: he is trying to get his food to do Reason: he is sitting at a table in a restaurant | Answer: person2 is trying to hit bed1 Reason: person2 is holding a knife and bed1 is trying to choke him |
| Generation - Refinement Module | Answer: dog2 belongs to person1 Reason: person1 is holding dog2 and people usually hold their own dogs | Answer: person14 is standing in court for a trial Reason: person14 is sitting at the head of the courtroom | Answer: he is not pleased with what person2 is saying Reason: he is looking at person2 with a scowl on his face | Answer: person1 is hugging person2 Reason: person1 is holding person2 up against her face |

Figure 3: *(Best viewed in color)* Qualitative results for the model with and without Refinement module. Blue box = question about image; Green = Results from model with Refinement module; Red = Results from model without Refinement module. (Object regions shown on image are for reader's understanding and are not given as input to model.)

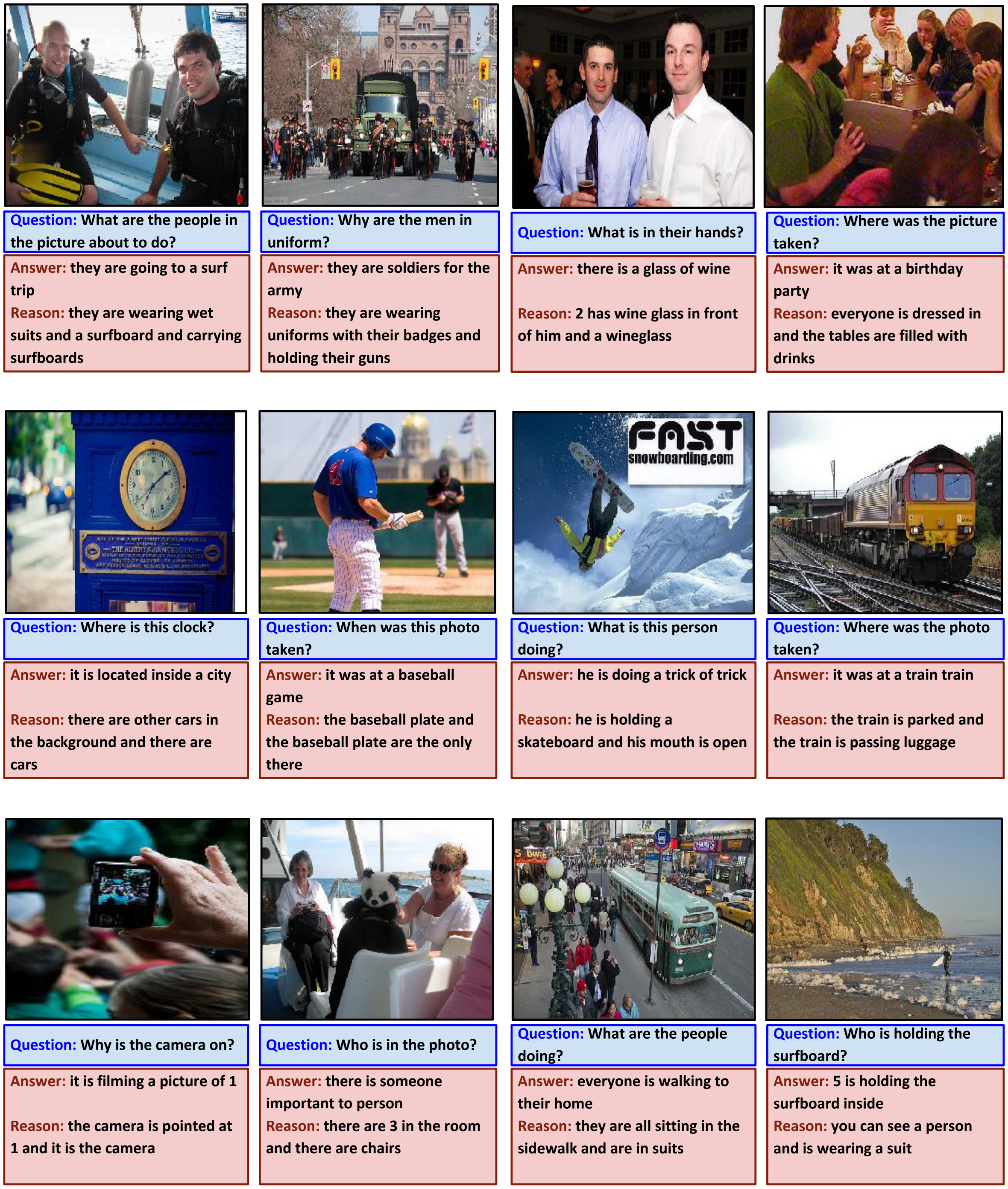



Figure 4: *(Best viewed in color)* Qualitative results on Visual7W dataset for `ViQAR` task from our proposed Generation-Refinement architecture. Blue box = Question about image; Red box = Generated results from our proposed architecture. (Note that there is Reason provided in the Visual7W dataset, and all the reasons in the figures are generated by our model.)

use Turing Tests (described in Section 5 of the main paper) to better estimate the performance of our model. An overall assessment that considers the different metrics used provides a more holistic view of the performance of our model.

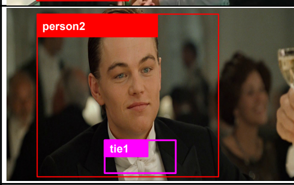
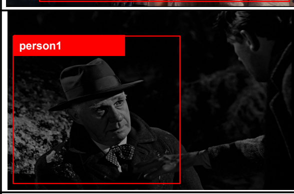
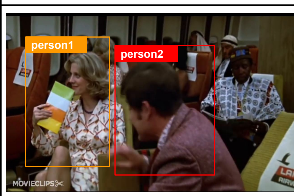
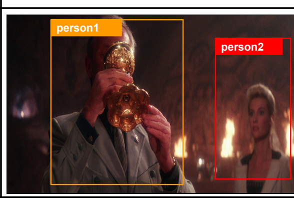
| Question | Image | Groundtruth Answer and Reason | Generated Answer and Reason | Evaluation Metrics |
|---|---|---|---|---|
| What is person1 trying to do to person2 ? |  | Answer: person1 is trying to get person2 to tell his parents to vote for him .<br><br>Reason: person1 looks like a politician and is talking to person2 who is a child . children can not vote . | Answer: person1 is trying to convince person2 to do something<br><br>Reason: person1 is leaning over and looking directly at person2 | EmbeddingAverageCS 0.86<br>infersent 0.54<br>USE 0.54<br>Vector Extrema CS 0.45<br>Greedy Matching Score 0.76<br>METEOR 0.16<br>Skipthought CS 0.22<br>RougeL 0.34 |
| Why is person2 wearing tie1 ? |  | Answer: person2 is at a formal event .<br><br>Reason: person2 was nominated for an award . | Answer: person2 is wearing a tie because he is at a wedding<br><br>Reason: people wear ties to formal events | EmbeddingAverageCS 0.7<br>infersent 0.44<br>USE 0.3<br>Vector Extrema CS 0.39<br>Greedy Matching Score 0.56<br>METEOR 0.11<br>Skipthought CS 0.19<br>RougeL 0.24 |
| Why is person1 laying down ? |  | Answer: person1 has a hangover .<br><br>Reason: person1 has a nauseous expression , and it ' s morning . | Answer: person1 is sick and asleep<br><br>Reason: person1 is laying down in bed with his eyes closed | EmbeddingAverageCS 0.63<br>infersent 0.42<br>USE 0.51<br>Vector Extrema CS 0.3<br>Greedy Matching Score 0.56<br>METEOR 0.18<br>Skipthought CS 0.22<br>RougeL 0.15 |
| Is person1 happy right now ? |  | Answer: No , person1 looks more confused than anything .<br><br>Reason: person1 is crinkling up his face and his eyes seem a bit intimidated by person2 . | Answer: no person1 is not happy<br><br>Reason: person1 has a very serious look on his face | EmbeddingAverageCS 0.84<br>infersent 0.64<br>USE 0.57<br>Vector Extrema CS 0.55<br>Greedy Matching Score 0.71<br>METEOR 0.16<br>Skipthought CS 0.25<br>RougeL 0.18 |
| Why are person1 person2 person3 , and person4 sitting at diningtable1 ? |  | Answer: They are in a work meeting .<br><br>Reason: They are dressed in professional outfits and are in an office . | Answer: they are working in a meeting<br><br>Reason: they are sitting at a table with papers in front of them | EmbeddingAverageCS 0.92<br>infersent 0.61<br>USE 0.62<br>Vector Extrema CS 0.66<br>Greedy Matching Score 0.7<br>METEOR 0.19<br>Skipthought CS 0.31<br>RougeL 0.39 |
| What does person3 have on his forehead ? |  | Answer: It ' s a cold compress .<br><br>Reason: The object is in the shape of a compress which his hands are holding to keep it on his forehead . | Answer: person3 has a bandage on his head<br><br>Reason: person3 has a cast on his forehead which is used for a bandage | EmbeddingAverageCS 0.71<br>infersent 0.36<br>USE 0.47<br>Vector Extrema CS 0.4<br>Greedy Matching Score 0.69<br>METEOR 0.09<br>Skipthought CS 0.22<br>RougeL 0.2 |
| 'Why is person2 turning toward person1 ? |  | Answer: person2 wants to ask person1 a question .<br><br>Reason: person2 is reaching toward person1 as if to ask her a leading question . this is how men sometimes appear when asking questions . | Answer: person2 is trying to talk to person1<br><br>Reason: person2 is looking at person1 s direction and is looking at person1 | EmbeddingAverageCS 0.84<br>infersent 0.62<br>USE 0.62<br>Vector Extrema CS 0.49<br>Greedy Matching Score 0.73<br>METEOR 0.15<br>Skipthought CS 0.23<br>RougeL 0.31 |
| Why is person2 looking at person1 in that way ? |  | Answer: She is shocked that he would drink out of such a valuable cup .<br><br>Reason: He is holding a cup made of gold up to his face . | Answer: he is wondering what person 1 is doing<br><br>Reason: person 2 is looking at person 1 with a look of disgust on his face | EmbeddingAverageCS 0.89<br>infersent 0.36<br>USE 0.36<br>Vector Extrema CS 0.42<br>Greedy Matching Score 0.65<br>METEOR 0.09<br>Skipthought CS 0.21<br>RougeL 0.23 |

Figure 5: *(Best viewed in color)* Sample quantitative and qualitative results that show that evaluation metrics can have low scores even when results are qualitatively good. Blue box = question about image; Green = Ground truth; Red = Generated results from our proposed architecture.

## References

[Li *et al.*2018] Qing Li, Qingyi Tao, Shafiq R. Joty, Jianfei Cai, and Jiebo Luo. Vqa-e: Explaining, elaborating, and enhancing your answers for visual questions. In *ECCV*, 2018.