Noname manuscript No.
(will be inserted by the editor)

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

Radhika Dua · Sai Srinivas Kancheti · Vineeth N Balasubramanian

Received: date / Accepted: date

Abstract 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. In this work, we present a completely generative formulation where a multi-word answer is *generated* for a visual query. To take this a step forward, 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. We propose an end-to-end architecture to solve this task and describe how to evaluate it. We show that our model generates strong answers and rationales through qualitative and quantitative evaluation, as well as through a human Turing Test.

Keywords Visual Commonsense Reasoning · Visual Question Answering · Model with Generated Explanations

1 Introduction

Visual Question Answering(VQA) (Thomason et al. 2018; Lu et al. 2019; Storks et al. 2019; Jang et al. 2017; Lei et al. 2018) is a vision-language task that has seen a lot of attention in recent years. In general, the VQA task consists of either open-ended or multiple choice answers to a question about the image. There are an increasing number of models devoted to obtaining the best possible performance on benchmark VQA datasets, which intend to measure visual understanding based on visual questions. Most existing works perform VQA by using an attention mechanism and combining features from two modalities for predicting answers. However, answers in existing VQA datasets and models are largely one-word answers (average length is 1.1 words), which gives existing models the freedom to treat answer generation as a classification task. For the open-ended VQA task, the top-K answers are chosen, and models perform classification over this vocabulary.

However, many questions which require commonsense reasoning cannot be answered in a single word. A textual answer for a sufficiently complicated question may need to be a sentence. For example, a question of the type "What will happen...." usually cannot be answered completely using a single word. Fig 2 shows examples of such questions where multi-word answers are required (the answers and rationales in this figure are generated by our model in this work). Current VQA systems are not well-suited for questions of this type. To reduce this gap, more recently, the Visual Commonsense Reasoning (VCR) task (Zellers et al. 2018; Lu et al. 2019; Dua et al. 2019; Zheng et al. 2019; Talmor et al. 2018; Lin et al. 2019a) was proposed, which requires a greater level of visual understanding and an ability to reason about the world. More interestingly, the VCR dataset features multi-word answers, with an average answer length of 7.55 words. However, VCR is still a classification task, where the correct answer is chosen from a set of four answers. Models which solve classification tasks simply need to pick an answer in the case of VQA, or an answer and a rationale for

Radhika Dua
Indian Institute of Technology, Hyderabad, India
E-mail: radhikadua1997@gmail.com

Sai Srinivas Kancheti
Indian Institute of Technology, Hyderabad, India
E-mail: sai.srinivas.kancheti@gmail.com

Vineeth N Balasubramanian
Indian Institute of Technology, Hyderabad, India
E-mail: vineethnb@iith.ac.in

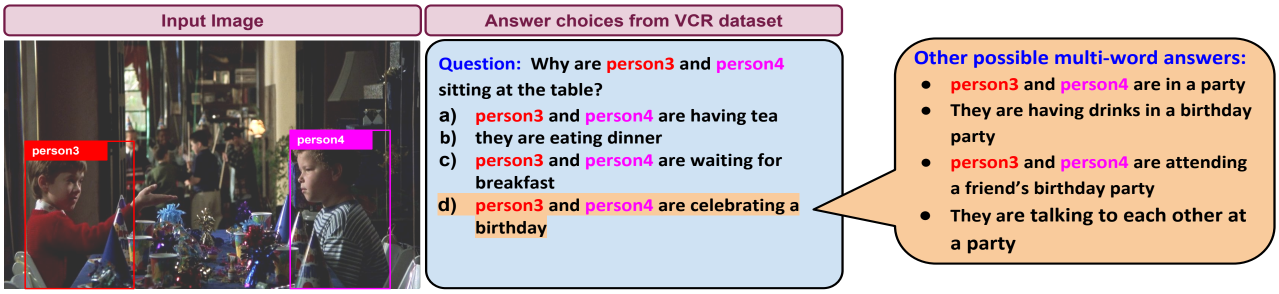


Fig. 1 An example from the VCR dataset (Zellers et al. 2018) shows that there can be many correct multi-word answers to a question, which makes classification setting restrictive. The highlighted option is the correct option present in the VCR dataset, the rest are examples of plausible correct answers.

VCR. However, when multi-word answers are required for a visual question, options are not sufficient, since the same 'correct' answer can be paraphrased in a multitude of ways, each having the same semantic meaning but differing in grammar. Fig 1 shows an image from the VCR dataset, where the first highlighted answer is the correct one among a set of four options provided in the dataset. The remaining three answers in the figure are included by us here (not in the dataset) as other plausible correct answers. Existing VQA models are fundamentally limited by picking a right option, rather to answer in a more natural manner. Moreover, since the number of possible 'correct' options in multi-word answer settings can be large (as evidenced by Fig 1), we propose that for richer answers, one would need to move away from the traditional classification setting, and instead let our model *generate the answer* to a given question. We hence propose a new task which takes a generative approach to multi-word VQA in this work.

Humans when answering questions often use a rationale to justify the answer. In certain cases, humans answer directly from memory (perhaps through associations) and then provide a post-hoc rationale, which could help improve the answer too - thus suggesting an interplay between an answer and its rationale. Following this cue, we also propose to generate a rationale along with the answer which serves two purposes: (i) it helps justify the generated answer to end-users; and (ii) it helps generate a better answer. Going beyond contemporary efforts in VQA, we hence propose, for the first time to the best of our knowledge, an approach that automatically generates both multi-word answers and an accompanying rationale, that also serves as a textual justification for the answer. We term this task `Visual Question Answering and Reasoning`(`ViQAR`) and propose an end-to-end methodology to address this task. This task is especially important in critical AI tasks such as VQA in the medical domain, where simply answering questions about the medical images is not sufficient. Instead, generating a rationale to justify the generated answer will help patients/doctors trust the models outcome and also understand the exact cause for the disease.

In addition to formalizing this new task, we provide a simple yet reasonably effective model consisting of four sequentially arranged recurrent networks to address this challenge. The model can be seen as having two parts: a generation module (GM), which comprises of the first two sequential recurrent networks, and a refinement module (RM), which comprises of the final two sequential recurrent networks. The GM first generates an answer, using which it generates a rationale that explains the answer. The RM generates a *refined* answer based on the rationale generated by GM. The refined answer is further used to generate a refined rationale. Our overall model design is motivated by the way humans think about answers to questions, wherein the answer and rationale are often mutually dependent on each other (one could motivate the other and also refine the other). We seek to model this dependency by first generating an answer-rationale pair and then using them as priors to regenerate a refined answer and rationale. We train our model on the VCR dataset, which contains open-ended visual questions along with answers and rationales. Considering this is a generative task, we evaluate our methodology by comparing our generated answer/rationale with the ground truth answer/rationale on correctness and goodness of the generated content using generative language metrics, as well as by human Turing Tests.

Our main contributions in this work can be summarized as follows: (i) We propose a new task `ViQAR` that seeks to open up a new dimension of Visual Question Answering tasks, by moving to a completely generative paradigm; (ii) We propose a simple and effective model based on generation and iterative refinement for `ViQAR`(which could serve as a baseline to the community); (iii) Considering generative models in general can be difficult to evaluate, we provide a discussion on how to evaluate such models, as well as study a comprehensive list of evaluation metrics for this task; (iv) We conduct a suite of experiments which show promise

Fig. 2 Given an image and a question about the image, we **generate** a natural language answer and reason that explains why the answer was generated. The images shown above are examples of outputs that our proposed model generates. These examples also illustrate the kind of visual questions for which a single-word answer is insufficient. Contemporary VQA models handle even such kinds of questions only in a classification setting, which is limiting.

of the proposed model for this task, and also perform ablation studies of various choices and components to study the effectiveness of the proposed methodology on `ViQAR`. We believe that this work could lead to further efforts on common-sense answer and rationale generation in vision tasks in the near future. To the best of our knowledge, this is the first such effort of automatically generating a multi-word answer and rationale to a visual question, instead of picking answers from a predefined list.

2 Related Work

In this section, we review earlier efforts from multiple perspectives that may be related to this work: Visual Question Answering, Visual Commonsense Reasoning and Image Captioning in general.

Visual Question Answering (VQA): VQA (Antol et al. 2015; Goyal et al. 2016; Jabri et al. 2016; Selvaraju et al. 2020) refers to the task of answering questions related to an image. VQA and its variants have been the subject of much research work recently. A lot of recent work has focused on varieties of attention-based models, which aim to 'look' at the relevant regions of the image in order to answer the question (Anderson et al. 2017; Lu et al. 2016b; Yu et al. 2017a; Xu and Saenko 2015; Yi et al. 2018; Xu and Saenko 2015; Shih et al. 2015; Chen et al. 2015; Yang et al. 2015). Other recent work has focused on better multimodal fusion methods (Kim et al. 2018; 2016; Fukui et al. 2016; Yu et al. 2017b), the incorporation of relations (Norcliffe-Brown et al. 2018; Li et al. 2019; Santoro et al. 2017), the use of multi-step reasoning (Cadène et al. 2019; Gan et al. 2019; Hudson and Manning 2018), and neural module networks for compositional reasoning (Andreas et al. 2016; Johnson et al. 2017; Chen et al. 2019; Hu et al. 2017). Visual Dialog (Das et al. 2018; Zheng et al. 2019) extends VQA but requires an agent to hold a meaningful conversation with humans in natural language based on visual questions.

The efforts closest to ours are those that provide justifications along with answers (Li et al. 2018b; Hendricks et al. 2016; Li et al. 2018a; Park et al. 2018; Wu et al. 2019b; Park et al. 2018; Rajani and Mooney 2017), each of which however also answers a question as a classification task (and not in a generative manner) as described below. (Li et al. 2018b) create the VQA-E dataset that has an explanation along with the answer to the question. (Wu et al. 2019b) provide relevant captions to aid in solving VQA, which can be thought of as weak justifications. More recent efforts (Park et al. 2018; Patro et al. 2020) attempt to provide visual and textual explanations to justify the predicted answers. Datasets have also been proposed for VQA in the recent past to test visual understanding (Zhu et al. 2015; Goyal et al. 2016; Johnson et al. 2016); for e.g., the Visual7W dataset (Zhu et al. 2015) contains a richer class of questions about an image with textual and visual answers. However, all these aforementioned efforts continue to focus on answering a question as a classification task (often in one word, such as Yes/No), followed by simple explanations. We however, in this work, focus on *generating* multi-word answers with a corresponding multi-word rationale, which has not been done before.

Visual Commonsense Reasoning (VCR): VCR (Zellers et al. 2018) is a recently introduced vision-language dataset which involves choosing a correct answer (among four provided options) for a given question about the image, and then choosing a rationale (among four provided options) that justifies the answer. The task associated with the dataset aims to test for visual commonsense understanding and provides images, questions and answers of a higher complexity than other datasets such as CLEVR (Johnson et al. 2016). The dataset has attracted a few methods over the last year (Zellers et al. 2018; Lu et al. 2019; Dua et al. 2019; Zheng et al. 2019; Talmor et al. 2018; Lin et al. 2019a;b; Wu et al. 2019a; Ayyubi et al. 2019; Brad 2019; Yu et al. 2019; Wu et al. 2019a), each of which however follow the dataset's task and treat this as a classification problem. None of these efforts attempt to answer and reason using generated sentences.

Image Captioning and Visual Dialog: One could also consider the task of image captioning (Xu et al. 2015; You et al. 2016; Lu et al. 2016a; Anderson et al. 2017; Rennie et al. 2016), where natural language captions are generated to describe an image, as being close to our objective. However, image captioning is more a global description of an image than question-answering problems that may be tasked with answering a question about understanding of a local region in the image.

In contrast to all the aforementioned efforts, our work, `ViQAR`, focuses on automatic complete *generation* of the answer, and of a rationale, given a visual query. This is a challenging task, since the generated answers must be correct (with respect to the question asked), be complete, be natural, and also be justified with a well-formed rationale. We now describe the task, and our methodology for addressing this task.

3 ViQAR: Task Description

Let $\mathcal{V}$ be a given vocabulary of size $|\mathcal{V}|$ and $\mathbf{A} = (a_1, a_2, ...a_{l_a}) \in \mathcal{V}^{l_a}$, $\mathbf{R} = (r_1, r_2, ...r_{l_r}) \in \mathcal{V}^{l_r}$ represent answer sequences of length l_a and rationale sequences of length l_r respectively. Let $\mathbf{I} \in \mathbb{R}^D$ represent the image representation, and $\mathbf{Q} \in \mathbb{R}^B$ be the feature representation of a given question. We also allow the use of an image caption, if available, in this framework given by a feature representation $\mathbf{C} \in \mathbb{R}^B$. Our task is to compute a function $\mathcal{F} : \mathbb{R}^D \times \mathbb{R}^B \times \mathbb{R}^B \rightarrow \mathcal{V}^{l_a} \times \mathcal{V}^{l_r}$ that maps the input image, question and caption features to a large space of generated answers $\mathbf{A}$ and rationales $\mathbf{R}$, as given below:

$$\mathcal{F}(\mathbf{I}, \mathbf{Q}, \mathbf{C}) = (\mathbf{A}, \mathbf{R}) \tag{1}$$

Note that the formalization of this task is different from other tasks in this domain, such as Visual Question Answering (Agrawal et al. 2015) and Visual Commonsense Reasoning (Zellers et al. 2018). The VQA task can be formulated as learning a function $\mathcal{G} : \mathbb{R}^D \times \mathbb{R}^B \rightarrow C$, where C is a discrete, finite set of choices (classification setting). Similarly, the Visual Commonsense Reasoning task provided in (Zellers et al. 2018) aims to learn a function $\mathcal{H} : \mathbb{R}^D \times \mathbb{R}^B \rightarrow C_1 \times C_2$, where C_1 is the set of possible answers, and C_2 is the set of possible reasons. The generative task, proposed here in `ViQAR`, is harder to solve when compared to VQA and VCR. One can divide `ViQAR` into two sub-tasks:

- **Answer Generation:** Given an image, its caption, and a complex question about the image, a multi-word natural language answer is generated: $(\mathbf{I}, \mathbf{Q}, \mathbf{C}) \rightarrow \mathbf{A}$
- **Rationale Generation:** Given an image, its caption, a complex question about the image, and an answer to the question, a rationale to justify the answer is generated: $(\mathbf{I}, \mathbf{Q}, \mathbf{C}, \mathbf{A}) \rightarrow \mathbf{R}$

We also study variants of the above sub-tasks (such as when captions are not available) in our experiments. Our experiments suggest that the availability of captions helps performance for the proposed task, expectedly. We now present a methodology built using known basic components to study and show that the proposed, seemingly challenging, new task can be solved with existing architectures. In particular, our methodology is based on the understanding that the answer and rationale can help each other, and hence needs an iterative refinement procedure to handle such a multi-word multi-output task. We consider the simplicity of the proposed solution as an aspect of our solution by design, more than a limitation, and hope that the proposed architecture will serve as a baseline for future efforts on this task.

4 Proposed Methodology

We present an end-to-end, attention-based, encoder-decoder architecture for answer and rationale generation which is based on an iterative refinement procedure. The refinement in our architecture is motivated by the observation that answers and rationales can influence one another mutually. Thus, knowing the answer helps in generation of a rationale, which in turn can help in the generation of a more refined answer. The encoder part of the architecture generates the features from the image, question and caption. These features are used by the decoder to generate the answer and rationale for a question.

Feature Extraction: We use spatial image features as proposed in (Anderson et al. 2017), which are termed bottom-up image features. We consider a fixed number of regions for each image, and extract a set of k features, V, as defined below:

$$V = \{\mathbf{v_1}, \mathbf{v_2}, ..., \mathbf{v_k}\} \qquad \text{where } \mathbf{v_i} \in \mathbb{R}^D \tag{2}$$

We use BERT (Devlin et al. 2019) representations to obtain fixed-size (B) embeddings for the question and caption, $\mathbf{Q} \in \mathbb{R}^B$ and $\mathbf{C} \in \mathbb{R}^B$ respectively. The question and caption are projected into a common feature space $\mathbf{T} \in \mathbb{R}^L$ given by:

$$\mathbf{T} = g(W_t^T(tanh(W_q^T\mathbf{Q}) \oplus tanh(W_c^T\mathbf{C}))) \tag{3}$$

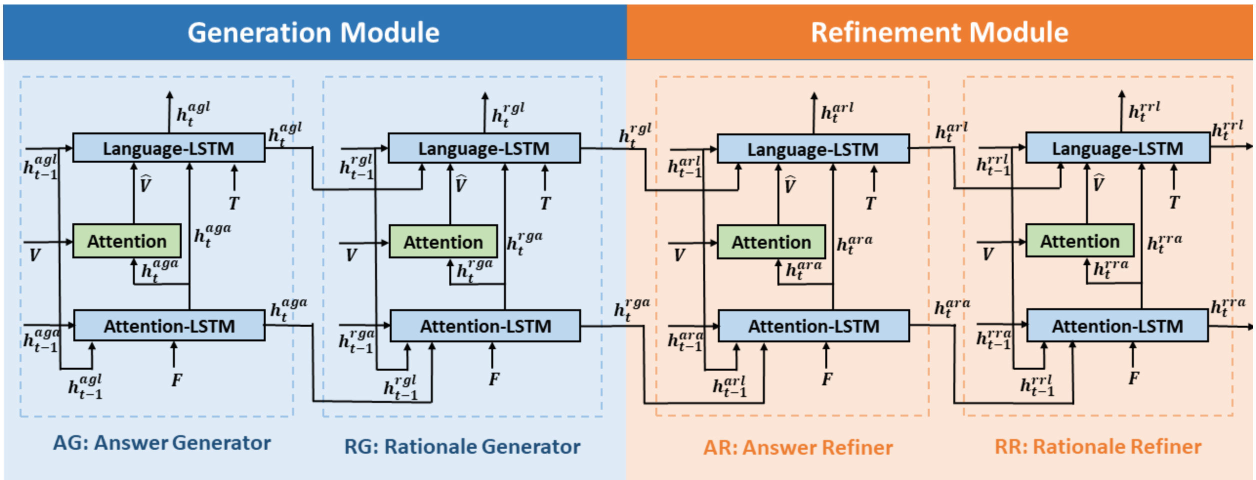


Fig. 3 The decoder of our proposed architecture: Given an image and a question on the image, the model must generate an answer to the question and a rationale to justify why the answer is correct.

where g is a non-linear function, $\oplus$ indicates concatenation and $W_t \in \mathbb{R}^{L\times L}$, $W_q \in \mathbb{R}^{B\times L}$ and $W_c \in \mathbb{R}^{B\times L}$ are learnable weight matrices of the layers (we use two linear layers in our implementation in this work).

Let the mean of the extracted spatial image features (as in Eqn 2) be denoted by $\bar{\mathbf{V}} \in \mathbb{R}^D$. These are concatenated with the projected question and caption features to obtain $\mathbf{F}$:

$$\mathbf{F} = \bar{\mathbf{V}} \oplus \mathbf{T} \tag{4}$$

We use $\mathbf{F}$ as the common input feature vector to all the LSTMs in our architecture.

Architecture: Fig. 3 shows our end-to-end architecture to address `ViQAR`. As stated earlier, our architecture has two modules: *generation* (GM) and *refinement* (RM). The GM consists of two sequential, stacked LSTMs, henceforth referred to as answer generator (AG) and rationale generator (RG) respectively. The RM seeks to refine the generated answer as well as rationale, and is an important part of the proposed solution as seen in our experimental results. It also consists of two sequential, stacked LSTMs, which we denote as answer refiner (AR) and rationale refiner (RR).

Each sub-module (presented inside dashed lines in the figure) is a complete LSTM. Given an image, question, and caption, the AG sub-module unrolls for l_a time steps to generate an answer. The hidden state of Language and Attention LSTMs after l_a time steps is a representation of the generated answer. Using the representation of the generated answer from AG, RG sub-module unrolls for l_r time steps to generate a rationale and obtain its representation. Then the AR sub-module uses the features from RG to generate a refined answer. Lastly, the RR sub-module uses the answer features from AR to generate a refined rationale. Thus, a refined answer is generated after $l_a + l_r$ time steps and a refined rationale is generated after l_a further time steps. The complete architecture runs in $2l_a + 2l_r$ time steps.

The LSTMs: The two layers of each stacked LSTM (Hochreiter and Schmidhuber 1997) are referred to as the Attention-LSTM ($\mathcal{L}_a$) and Language-LSTM ($\mathcal{L}_l$) respectively. We denote h_t^a and x_t^a as the hidden state and input of the Attention-LSTM at time step t respectively. Analogously, h_t^l and x_t^l denote the hidden state and input of the Language-LSTM at time t. Since the four LSTMs are identical in operation, we describe the attention and sequence generation modules of one of the sequential LSTMs below in detail.

Spatial Visual Attention: We use a soft, spatial-attention model, similar to (Anderson et al. 2017) and (Lu et al. 2016a), to compute attended image features $\hat{\mathbf{V}}$. Given the combined input features $\mathbf{F}$ and previous hidden states h_{t-1}^a, h_{t-1}^l, the current hidden state of the Attention-LSTM is given by:

$$x_t^a \equiv h^p \oplus h_{t-1}^l \oplus \mathbf{F} \oplus \pi_t \tag{5}$$

$$h_t^a = \mathcal{L}_a(x_t^a, h_{t-1}^a) \tag{6}$$

where $\pi_t = W_e^T \mathbf{1}_t$ is the embedding of the input word, $W_e \in \mathbb{R}^{|\mathcal{V}|\times E}$ is the weight of the embedding layer, and $\mathbf{1}_t$ is the one-hot representation of the input at time t. h^p is the hidden representation of the previous LSTM (answer or rationale, depending on the current LSTM).

The hidden state h_t^a and visual features V are used by the attention module (implemented as a two-layered MLP in this work) to compute the normalized set of attention weights $\boldsymbol{\alpha}_t = \{\alpha_{1t}, \alpha_{2t}, ..., \alpha_{kt}\}$ (where α_{it} is the normalized weight of image feature $\mathbf{v_i}$) as below:

$$y_{i,t} = W_{ay}^T(\tanh(W_{av}^T \mathbf{v}_i + W_{ah}^T h_t^a)) \tag{7}$$

$$\boldsymbol{\alpha}_t = \mathrm{softmax}(y_{1t}, y_{23}, ..., y_{kt}) \tag{8}$$

In the above equations, $W_{ay} \in \mathbb{R}^{A\times 1}$, $W_{av} \in \mathbb{R}^{D\times A}$ and $W_{ah} \in \mathbb{R}^{H\times A}$ are weights learned by the attention MLP, H is the hidden size of the LSTM and A is the hidden size of the attention MLP.

The attended image feature vector $\mathbf{\hat{V}_t} = \sum_{\mathbf{i=1}}^{\mathbf{k}} \alpha_{\mathbf{it}} \mathbf{v_i}$ is the weighted sum of all visual features.

Sequence Generation: The attended image features $\hat{\mathbf{V}}_t$, together with $\mathbf{T}$ and h_t^a, are inputs to the language-LSTM at time t. We then have:

$$x_t^l \equiv h^p \oplus \hat{\mathbf{V}}_t \oplus h_t^a \oplus \mathbf{T} \tag{9}$$

$$h_t^l = \mathcal{L}_l(x_t^l, h_{t-1}^l) \tag{10}$$

$$y_t = W_{lh}^T h_t^l + b_{lh} \tag{11}$$

$$p_t = softmax(y_t) \tag{12}$$

where h^p is the hidden state of the previous LSTM, h_t^l is the output of the Language-LSTM, p_t is the conditional probability over words in $\mathcal{V}$ at time t. The word at time step t is generated by a single-layered MLP with learnable parameters: $W_{lh} \in \mathbb{R}^{H \times |\mathcal{V}|}$, $b_{lh} \in \mathbb{R}^{|\mathcal{V}| \times 1}$. The attention MLP parameters W_{ay}, W_{av} and W_{ah}, and embedding layer's parameters W_e are shared across all four LSTMs. (We reiterate that although the architecture is based on well-known components, the aforementioned design decisions were obtained after significant study.)

Loss Function: For a better understanding of our approach, Figure 4 presents a high-level illustration of our proposed generation-refinement model.

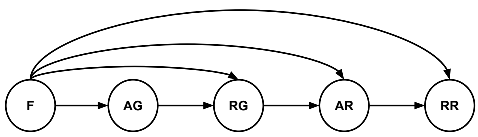


Fig. 4 High-level illustration of our proposed Generation-Refinement model

Let $A_1 = (a_{11}, a_{12}, ..., a_{1l_a})$, $R_1 = (r_{11}, r_{12}, ..., r_{1l_r})$, $A_2 = (a_{21}, a_{22}, ..., a_{2l_a})$ and $R_2 = (r_{21}, r_{22}, ..., r_{2l_r})$ be the generated answer, generated rationale, refined answer and refined rationale sequences respectively, where a_{ij} and r_{ij} are discrete random variables taking values from the common vocabulary $\mathcal{V}$. Given the common input F, our objective is to maximize the likelihood $P(A_1, R_1, A_2, R_2|F)$ given by:

$$\begin{aligned} P(A_1, R_1, A_2, R_2|F) &= P(A_1, R1|F)P(A_2, R_2|F, A_1, R_1) \\ &= P(A_1|F)P(R_1|F, A_1) \\ &\quad P(A_2|F, A_1, R_1)P(R_2|F, A_1, R_1, A_2) \end{aligned} \tag{13}$$

In our model design, each term in the RHS of Eqn 13 is computed by a distinct LSTM. Hence, minimizing the sum of losses of the four LSTMs becomes equivalent to maximizing the joint likelihood. Our overall loss is the sum of four cross-entropy losses, one for each LSTM, as given below:

$$\mathcal{L} = -\left(\sum_{t=1}^{l_a} \log p_t^{\theta_1} + \sum_{t=1}^{l_r} \log p_t^{\theta_2} + \sum_{t=1}^{l_a} \log p_t^{\theta_3} + \sum_{t=1}^{l_r} \log p_t^{\theta_4} \right) \tag{14}$$

where θ_i represents the i^{th} sub-module LSTM, p_t is the conditional probability of the t^{th} word in the input sequence as calculated by the corresponding LSTM, l_a indicates the ground-truth answer length, and l_r the ground truth rationale length. Other loss formulations, such as a weighted average of the cross entropy terms did not perform better than a simple sum. We tried weights from $0.0, 0.25, 0.5, 0.75, 1.0$ for the loss terms. More implementation details are provided in Section 5.

5 Experiments and Results

In this section, we describe the dataset used for this work, implementation details, as well as present the results of the proposed method and its variants.

Dataset: Considering there has been no dataset explicitly built for this new task, we study the performance of the proposed method on the recent VCR (Zellers et al. 2018) dataset, which has all components needed for our approach. We train our proposed architecture on VCR, which contains ground truth answers and ground truth rationales that allow us to compare our generated answers and rationales against. We also show in Section 6 on how a model trained on the VCR dataset can be used to give a rationale for images from Visual7W (Zhu et al. 2015), a VQA dataset with no ground-truth rationale.

VCR is a large-scale dataset that consists of 290k triplets of questions, answers, and rationales over 110k unique movie scene images. For our method, we also use the captions provided by the authors of VCR as input to our method (we later perform an ablation study without the captions). At inference, captions are generated using (Vinyals et al. 2014) (trained on provided captions) and use them as input to our model. Since the test set of the dataset is not provided, we randomly split the train set into train-train-split (202,923 samples) and train-val-split (10,000 samples) while using the validation set as our test data (26,534 samples). All our baselines are also compared on the same setting for fair comparison.

Dataset	**Avg. A length**	**Avg. Q length**	**Avg. R length**	**Complexity**
VCR	7.55	6.61	16.2	High
VQA-E	1.11	6.1	11.1	Low
VQA-X	1.12	6.13	8.56	Low

Table 1 Statistical comparison of VCR with VQA-E, and VQA-X datasets. VCR dataset is highly complex as it is made up of complex subjective questions.

VQA-E (Li et al. 2018b) and VQA-X (Park et al. 2018) are competing datasets that contains explanations along with question-answer pairs. Table 1 shows the high-level analysis of the three datasets. Since VQA-E and VQA-X are derived from VQA-2, many of the questions can be answered in one word (a yes/no answer or a number). In contrast, VCR asks open-ended questions and has longer answers. Since our task aims to generate rich answers, the VCR dataset provides a richer context for this work. CLEVR (Johnson et al. 2016) is another VQA dataset that measures the logical reasoning capabilities by asking the question that can be answered when a certain sequential reasoning is followed. This dataset however does not contain reasons/rationales on which we can train. Also, we do not perform a direct evaluation on CLEVR because our model is trained on real-world natural images while CLEVR is a synthetic shapes dataset.

In order to study our method further, we also study the transfer of our learned model to another challenging dataset, Visual7W (Zhu et al. 2015), by generating an answer/ rationale pair for visual questions in Visual7W (please see Section 6 more more details). Visual7W is a large-scale visual question answering (VQA) dataset, which has multi-modal answers i.e visual 'pointing' answers and textual 'telling' answers.

Implementation Details: We use spatial image features generated from (Anderson et al. 2017) as our image input. Fixed-size BERT representations of questions and captions are used. Hidden size of all LSTMs is set to 1024 and hidden size of the attention MLP is set to 512. We trained using the ADAM optimizer with a decaying learning rate starting from $4e^{-4}$, using a batch size of 64. Dropout is used as a regularizer. Our code is publicly available at radhikadua123.github.io/ViQAR.

Evaluation Metrics: We use multiple objective evaluation metrics to evaluate the goodness of answers and rationales generated by our model. Since our task is generative, evaluation is done by comparing our generated sentences with ground-truth sentences to assess their semantic correctness as well as structural soundness. To this end, we use a combination of multiple existing evaluation metrics. Word overlap-based metrics such as METEOR (Lavie and Agarwal 2007), CIDEr (Vedantam et al. 2014) and ROUGE (Lin 2004) quantify the structural closeness of the generated sentences to the ground-truth. While such metrics give a sense of the structural correctness of the generated sentences, they may however be insufficient for evaluating generation tasks, since there could be many valid generations which are correct, but not share the same words as a single ground truth answer. Hence, in order to measure how close the generation is to the ground-truth in meaning, we additionally use embedding-based metrics (which calculate the cosine similarity between sentence embeddings for generated and ground-truth sentences) including SkipThought cosine similarity (Kiros et al. 2015), Vector Extrema cosine similarity (Forgues and Pineau 2014), Universal sentence encoder (Cer et al. 2018), Infersent (Conneau et al. 2017) and BERTScore (Zhang et al. 2019). We use a comprehensive suite of all the aforementioned metrics to study the performance of our model. Subjective studies are presented later in this section.

Classification Accuracy: We also 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 2. 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.

Results: *Qualitative Results:* Fig 5 shows examples of images and questions where the proposed model generates a meaningful answer with a supporting rationale. Qualitative results indicate that our model is capable of generating answer-rationale pairs to complex subjective questions starting with 'Why', 'What', 'How', etc. Given the question, "What is person8 doing right now?", the generated rationale: "person8 is standing in front of a microphone" actively supports the generated answer: "person8 is giving a speech", justifying the choice of a two-stage generation-refinement architecture. For completeness of understanding, we also present a few examples on which our model fails to generate the semantically correct answer in Fig. S2. However, even on these results, we observe that our model does generate both answers and rationales that are grammatically correct and complete (where rationale supports the generated answer). Improving the semantic correctness of the generations will be an important direction of future work. More qualitative results are presented in the supplementary owing to space constraints.

Quantitative Results: Quantitative results on the suite of evaluation metrics stated earlier are shown in Table 3. Since this is a new task, there are no known methods to compare against. We hence compare our model against a basic two-stage LSTM that generates the answer and reason independently as a baseline (called *Baseline* in Table 3), and a VQA-based method (Anderson et al. 2017) that extracts multi-modal features to

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

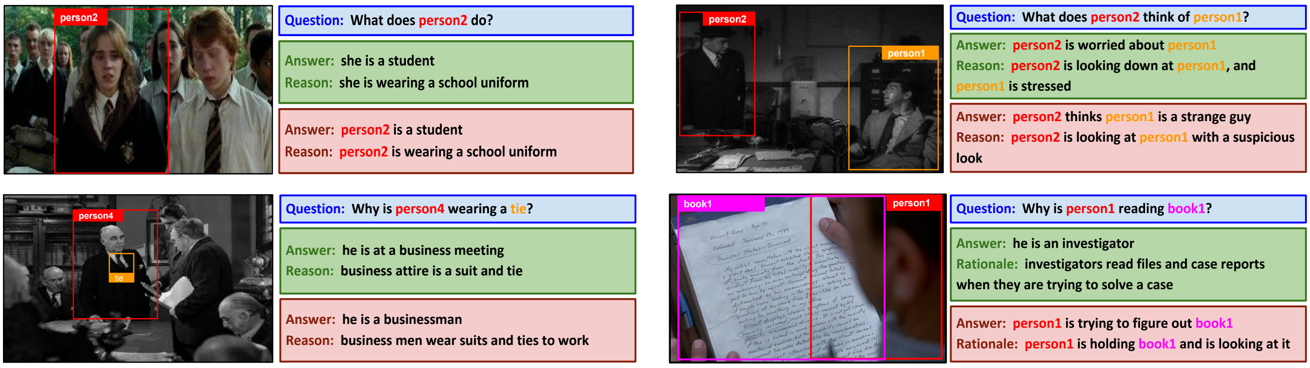


Fig. 5 *(Best viewed in color)* Sample qualitative results for `ViQAR` from our proposed Generation-Refinement architecture. Blue box = question about image; Green = Ground truth answer and rationale; Red = Answer and rationale generated by our architecture. *(Object regions shown on image is for reader's understanding and are not given as input to model.)*

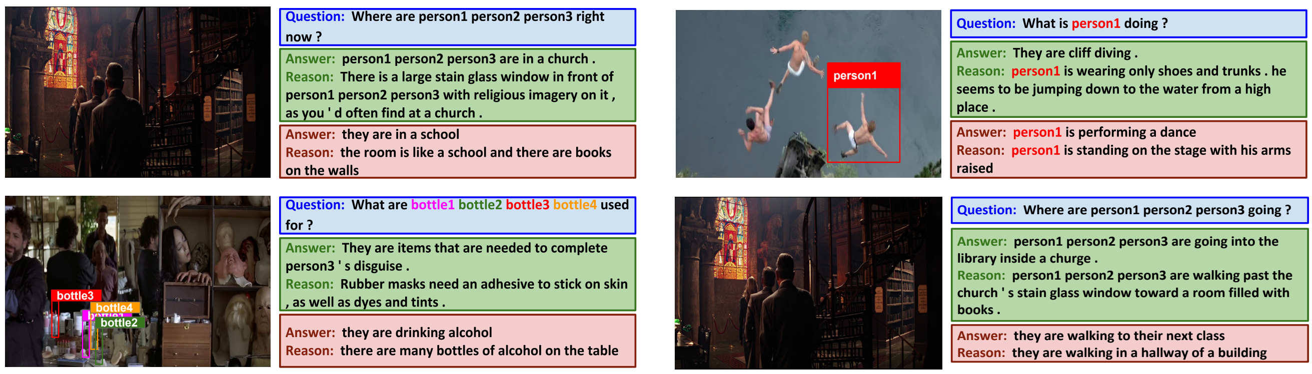


Fig. 6 *(Best viewed in color)* Challenging inputs for which our model fails to generate the semantically correct answer, but the generated answer and rationale still show grammatical correctness and completeness, showing the promise of the overall idea. Blue box = question about the image; Green = Ground truth; Red = Results generated by our architecture. *(Object regions shown on image are for reader's understanding and are not given as input to model.)*

generate answers and rationales independently (called *VQA- Baseline* in Table 3). (Comparison with other standard VQA models is not relevant in this setting, since we perform a generative task, unlike existing VQA models.) We show results on three variants of our proposed Generation-Refinement model: Q+I+C (when question, image and caption are given as inputs), Q+I (question and image alone as inputs), and Q+C (question and caption alone as inputs). Evidently, our Q+I+C performed the most consistently across all the evaluation metrics. (The qualitative results in Fig 5 were obtained using the (Q+I+C) setting.) Importantly, our model outperforms both baselines, including the VQA-based one, on every single evaluation metric, showing the utility of the proposed architecture.

Human Turing Test: In addition to the study of the objective evaluation metrics, we also performed a human Turing test on the generated answers and rationales. 30 human evaluators were presented each with 50 randomly sampled image-question pairs, each containing an answer to the question and its rationale. The test aims to measure how humans score the generated sentences w.r.t. ground truth sentences. Sixteen of the fifty questions had ground truth answers and rationales, while the rest were generated by our proposed model. For each sample, the evaluators had to give a rating of 1 to 5 for five different criteria, with 1 being very poor and 5 being very good. The results are presented in Table 4. Despite the higher number of generated answer-rationales judged by human users, the answers and rationales produced by our method were deemed to be fairly correct grammatically. The evaluators also agreed that our answers were relevant to the

Metrics	VQA-Baseline (Anderson et al. 2017)	Baseline	Q+I+C	Q+I	Q+C
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

Table 3 Quantitative evaluation on VCR dataset; CS = cosine similarity; we compare against a basic two-stage LSTM model and a VQA model (Anderson et al. 2017) as baselines; remaining columns are proposed model variants

question and the generated rationales are acceptably relevant to the generated answer.

6 Discussions and Analysis

Ablation Studies on Refinement Module: We evaluated the performance of the following variations of our proposed generation-refinement architecture M: (i) $M - RM$: where the refinement module is removed; and (ii) $M + RM$: where a second refinement module is added, i.e. the model has one generation module and two refinement modules (to see if further refinement of answer and rationale helps). Table 5 shows the quantitative results.

Metrics	#Refine Modules 0	1	2
Univ Sent Encoder	0.453	**0.455**	0.430
Infersent	0.434	**0.438**	0.421
Embedding Avg Cosine similarity	0.85	0.846	0.840
Vector Extrema Cosine Similarity	0.482	**0.493**	0.462
Greedy Matching Score	0.659	**0.672**	0.639
METEOR	0.101	**0.116**	0.090
Skipthought Cosine Similarity	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

Table 5 Comparison of proposed Generation-Refinement Architecture with variations in num of Refinement modules

We observe that our proposed model, which has one refinement module has the best results. Adding additional refinement modules causes the performance to go down. We hypothesize that the additional parameters (in a second Refinement module) in the model makes it harder for the network to learn. Removal of the refinement module also causes performance to drop, supporting our claim on the usefulness for a Refinement module too. We also studied the classification accuracy in these variations, and observed that 1-refinement model (original version of our method) with 11.9% accuracy outperforms 0-refinement model (11.64%) and 2-refinement model (10.94%) for the same reasons. Fig. 7 provides a few qualitative results with and without the refinement module, supporting our claim. More results are presented in the supplementary.

Transfer to Other Datasets: We also studied whether the proposed model, trained on the VCR dataset, can provide answers and rationales to visual questions in other VQA datasets (which do not have ground truth rationale provided). To this end, we tested our trained model on the widely used Visual7W Zhu et al. (2015) dataset without any additional training. Fig 8 presents qualitative results for `ViQAR` task on the Visual7W dataset. We also perform a Turing test on the generated answers and rationales to evaluate the model's performance on Visual7W. Thirty human evaluators were presented each with twenty five hand-picked image-question pairs, each of which contains a generated answer to the question and its rationale. The results, presented in Table 6 show 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. More results are presented in the supplementary.

More results, including qualitative examples, for various settings are included in the Supplementary Section. Since `ViQAR` is a completely generative task, objective evaluation is a challenge, as in any other generative methods. For comprehensive evaluation, we use a suite of objective metrics typically used in related vision-language tasks. We perform a detailed analysis in the Supplementary material and show that even our successful results (qualitatively speaking) may have low scores on objective evaluation metrics at times, since generated sentences may not match a ground truth sentence word-by-word. We hope that opening up this dimension of generated explanations will only motivate a better metric in the near future.

7 Conclusion

In this paper, we propose `ViQAR`, a novel task for generating a multi-word answer and a rationale given an image and a question. Our work aims to go beyond classical VQA by moving to a completely generative

Criteria	Generated Mean ± std	Ground-truth Mean ± std
How well-formed and grammatically correct is the answer?	4.15 ± 1.05	4.40 ± 0.87
How well-formed and grammatically correct is the rationale?	3.53 ± 1.26	4.26 ± 0.92
How relevant is the answer to the image-question pair?	3.60 ± 1.32	4.08 ± 1.03
How well does the rationale explain the answer with respect to the image-question pair?	3.04 ± 1.36	4.05 ± 1.10
Irrespective of the image-question pair, how well does the rationale explain the answer ?	3.46 ± 1.35	4.13 ± 1.09

Table 4 Results of human Turing test performed with 30 people who rated samples consisting of question, corresponding answer and rationale on 5 criteria. For each criterion, a rating of 1 to 5 was given. Results show mean score and standard deviation for each criterion for generated and ground truth samples.

Fig. 7 Qualitative results for our model with (in green, last row) and without (in red, penultimate row) Refinement module

Fig. 8 Qualitative results on Visual7W dataset (note that there is no rationale provided in this dataset, and all above rationales were generated by our model)

Criteria	Generated Mean ± std
How well-formed and grammatically correct is the answer?	3.98 ± 1.08
How well-formed and grammatically correct is the rationale?	3.80 ± 1.04
How relevant is the answer to the image-question pair?	4.11 ± 1.17
How well does the rationale explain the answer with respect to the image-question pair?	3.83 ± 1.23
Irrespective of the image-question pair, how well does the rationale explain the answer ?	3.83 ± 1.28

Table 6 Results of the Turing test on Visual7W dataset performed with 30 people who had to rate samples consisting of a question and its corresponding answer and rationales on five criteria. For each criterion, a rating of 1 to 5 was given. The table gives the mean score and standard deviation for each criterion for the generated samples.

paradigm. To solve `ViQAR`, we present an end-to-end generation-refinement architecture which is based on the observation that answers and rationales are dependent on one another. We showed the promise of our model on the VCR dataset both qualitatively and quantitatively, and our human Turing test showed results comparable to the ground truth. We also showed that this model can be transferred to tasks without ground truth rationale. We hope that our work will open up a broader discussion around generative answers in VQA and other deep neural network models in general.

Acknowledgements We acknowledge MHRD and DST, Govt of India, as well as Honeywell India for partial support of this project through the UAY program, IITH/005. We also thank Japan International Cooperation Agency and IIT Hyderabad for the generous compute support.

References

Agrawal A, Lu J, Antol S, Mitchell M, Zitnick CL, Parikh D, Batra D (2015) Vqa: Visual question answering. ICCV

Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, Zhang L (2017) Bottom-up and top-down attention for image captioning and visual question answering. CVPR

Andreas J, Rohrbach M, Darrell T, Klein D (2016) Neural module networks. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp 39–48

Antol S, Agrawal A, Lu J, Mitchell M, Batra D, Zitnick CL, Parikh D (2015) Vqa: Visual question answering. In: ICCV

Ayyubi HA, Tanjim MM, Kriegman DJ (2019) Enforcing reasoning in visual commonsense reasoning. ArXiv abs/1910.11124

Brad F (2019) Scene graph contextualization in visual commonsense reasoning. In: ICCV 2019

Cadène R, Ben-younes H, Cord M, Thome N (2019) Murel: Multimodal relational reasoning for visual question answering. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp 1989–1998

Cer D, Yang Y, yi Kong S, Hua N, Limtiaco N, John RS, Constant N, Guajardo-Cespedes M, Yuan S, Tar C, Sung YH, Strope B, Kurzweil R (2018) Universal sentence encoder. ArXiv

Chen K, Wang J, Chen L, Gao H, Xu W, Nevatia R (2015) ABC-CNN: an attention based convolutional neural network for visual question answering. CoRR abs/1511.05960, URL `http://arxiv.org/abs/1511.05960`, `1511.05960`

Chen W, Gan Z, Li L, Cheng Y, Wang WWJ, Liu J (2019) Meta module network for compositional visual reasoning. ArXiv abs/1910.03230

Conneau A, Kiela D, Schwenk H, Barrault L, Bordes A (2017) Supervised learning of universal sentence representations from natural language inference data. CoRR

Das A, Kottur S, Gupta K, Singh A, Yadav D, Lee S, Moura JMF, Parikh D, Batra D (2018) Visual dialog. IEEE TPAMI 41

Devlin J, Chang MW, Lee K, Toutanova K (2019) Bert: Pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT

Dua D, Wang Y, Dasigi P, Stanovsky G, Singh S, Gardner M (2019) Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In: NAACL-HLT

Forgues G, Pineau J (2014) Bootstrapping dialog systems with word embeddings

Fukui A, Park DH, Yang D, Rohrbach A, Darrell T, Rohrbach M (2016) Multimodal compact bilinear pooling for visual question answering and visual grounding. In: EMNLP

Gan Z, Cheng Y, Kholy AE, Li L, Liu J, Gao J (2019) Multi-step reasoning via recurrent dual attention for visual dialog. In: ACL

Goyal Y, Khot T, Summers-Stay D, Batra D, Parikh D (2016) Making the v in vqa matter: Elevating the role of image understanding in visual question answering. CVPR

Hendricks LA, Akata Z, Rohrbach M, Donahue J, Schiele B, Darrell T (2016) Generating visual explanations. In: ECCV

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Computation 9:1735–1780

Hu R, Andreas J, Rohrbach M, Darrell T, Saenko K (2017) Learning to reason: End-to-end module networks for visual question answering. 2017 IEEE International Conference on Computer Vision (ICCV) pp 804–813

Hudson DA, Manning CD (2018) Compositional attention networks for machine reasoning. ArXiv abs/1803.03067

Jabri A, Joulin A, van der Maaten L (2016) Revisiting visual question answering baselines. In: ECCV

Jang Y, Song Y, Yu Y, Kim Y, Kim G (2017) Tgif-qa: Toward spatio-temporal reasoning in visual question answering. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp 1359–1367

Johnson J, Hariharan B, van der Maaten L, Fei-Fei L, Zitnick CL, Girshick RB (2016) Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. CVPR

Johnson J, Hariharan B, van der Maaten L, Hoffman J, Fei-Fei L, Zitnick CL, Girshick RB (2017) Inferring and executing programs for visual reasoning supplementary material

Kim J, On KW, Lim W, Kim J, Ha J, Zhang B (2016) Hadamard product for low-rank bilinear pooling. CoRR abs/1610.04325, URL `http://arxiv.`

org/abs/1610.04325, 1610.04325
Kim JH, Jun J, Zhang BT (2018) Bilinear attention networks. ArXiv abs/1805.07932
Kiros R, Zhu Y, Salakhutdinov R, Zemel RS, Urtasun R, Torralba A, Fidler S (2015) Skip-thought vectors. In: NIPS
Lavie A, Agarwal A (2007) Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In: WMT@ACL
Lei J, Yu L, Bansal M, Berg TL (2018) Tvqa: Localized, compositional video question answering. In: EMNLP
Li L, Gan Z, Cheng Y, Liu J (2019) Relation-aware graph attention network for visual question answering. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) pp 10312–10321
Li Q, Fu J, Yu D, Mei T, Luo J (2018a) Tell-and-answer: Towards explainable visual question answering using attributes and captions. In: EMNLP
Li Q, Tao Q, Joty SR, Cai J, Luo J (2018b) Vqa-e: Explaining, elaborating, and enhancing your answers for visual questions. In: ECCV
Lin BY, Chen X, Chen J, Ren X (2019a) Kagnet: Knowledge-aware graph networks for commonsense reasoning. ArXiv
Lin CY (2004) Rouge: A package for automatic evaluation of summaries. In: ACL
Lin J, Jain U, Schwing AG (2019b) Tab-vcr: Tags and attributes based vcr baselines. In: NeurIPS
Lu J, Xiong C, Parikh D, Socher R (2016a) Knowing when to look: Adaptive attention via a visual sentinel for image captioning. CVPR
Lu J, Yang J, Batra D, Parikh D (2016b) Hierarchical question-image co-attention for visual question answering. In: NIPS
Lu J, Batra D, Parikh D, Lee S (2019) Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. ArXiv
Norcliffe-Brown W, Vafeias E, Parisot S (2018) Learning conditioned graph structures for interpretable visual question answering. In: NeurIPS
Park DH, Hendricks LA, Akata Z, Rohrbach A, Schiele B, Darrell T, Rohrbach M (2018) Multimodal explanations: Justifying decisions and pointing to the evidence. CVPR
Patro BN, Pate S, Namboodiri VP (2020) Robust explanations for visual question answering. ArXiv abs/2001.08730
Rajani NF, Mooney RJ (2017) Ensembling visual explanations for vqa
Rennie SJ, Marcheret E, Mroueh Y, Ross J, Goel V (2016) Self-critical sequence training for image captioning. CVPR
Santoro A, Raposo D, Barrett DGT, Malinowski M, Pascanu R, Battaglia PW, Lillicrap TP (2017) A simple neural network module for relational reasoning. In: NIPS
Selvaraju RR, Tendulkar P, Parikh D, Horvitz E, Ribeiro MT, Nushi B, Kamar E (2020) Squinting at vqa models: Interrogating vqa models with subquestions. ArXiv abs/2001.06927
Shih KJ, Singh S, Hoiem D (2015) Where to look: Focus regions for visual question answering. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp 4613–4621
Storks S, Gao Q, Chai JY (2019) Commonsense reasoning for natural language understanding: A survey of benchmarks, resources, and approaches. ArXiv
Talmor A, Herzig J, Lourie N, Berant J (2018) Commonsenseqa: A question answering challenge targeting commonsense knowledge. In: NAACL-HLT
Thomason J, Gordon D, Bisk Y (2018) Shifting the baseline: Single modality performance on visual navigation and qa. In: NAACL-HLT
Vedantam R, Zitnick CL, Parikh D (2014) Cider: Consensus-based image description evaluation. CVPR
Vinyals O, Toshev A, Bengio S, Erhan D (2014) Show and tell: A neural image caption generator. CVPR
Wu A, Zhu L, Han Y, Yang Y (2019a) Connective cognition network for directional visual commonsense reasoning. In: NeurIPS
Wu J, Hu Z, Mooney RJ (2019b) Generating question relevant captions to aid visual question answering. In: ACL
Xu H, Saenko K (2015) Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In: ECCV
Xu K, Ba J, Kiros R, Cho K, Courville AC, Salakhutdinov R, Zemel RS, Bengio Y (2015) Show, attend and tell: Neural image caption generation with visual attention. In: ICML
Yang Z, He X, Gao J, Deng L, Smola AJ (2015) Stacked attention networks for image question answering. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp 21–29
Yi K, Wu J, Gan C, Torralba A, Kohli P, Tenenbaum JB (2018) Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. In: NeurIPS
You Q, Jin H, Wang Z, Fang C, Luo J (2016) Image captioning with semantic attention. CVPR
Yu D, Fu J, Mei T, Rui Y (2017a) Multi-level attention networks for visual question answering. CVPR
Yu W, Zhou J, Yu W, Liang X, Xiao N (2019) Heterogeneous graph learning for visual commonsense rea-

soning. In: NeurIPS

Yu Z, Yu J, Fan J, Tao D (2017b) Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. 2017 IEEE International Conference on Computer Vision (ICCV) pp 1839–1848

Zellers R, Bisk Y, Farhadi A, Choi Y (2018) From recognition to cognition: Visual commonsense reasoning. In: CVPR

Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y (2019) Bertscore: Evaluating text generation with bert. ArXiv abs/1904.09675

Zheng Z, Wang W, Qi S, Zhu SC (2019) Reasoning visual dialogs with structural and partial observations. In: CVPR

Zhu Y, Groth O, Bernstein MS, Fei-Fei L (2015) Visual7w: Grounded question answering in images. CVPR

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

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

- Additional qualitative results of our model (extension of Section 5 of main paper)
- More results on transferring our model to the VQA task (extension of results in Section 6 of main paper)
- More results on impact of the Refinement module in our model, and a study on the effect of adding refinement module to VQA-E
- A discussion on existing objective evaluation metrics for this task, and the need to go beyond (extension of section 6 of our paper)
- A study, using our attention maps, on the interpretability of our results

S1 Additional Qualitative Results

In addition to the qualitative results presented in Sec 5 of the main paper, Fig S1 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 examples on which our model fails to generate a good answer-rationale pair in Fig S2. 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.

S2 More 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 S3 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.

S3 Impact of Refinement Module

Figure S4 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. (2018b), 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.

S4 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 S5 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

Fig. S1 *(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.)

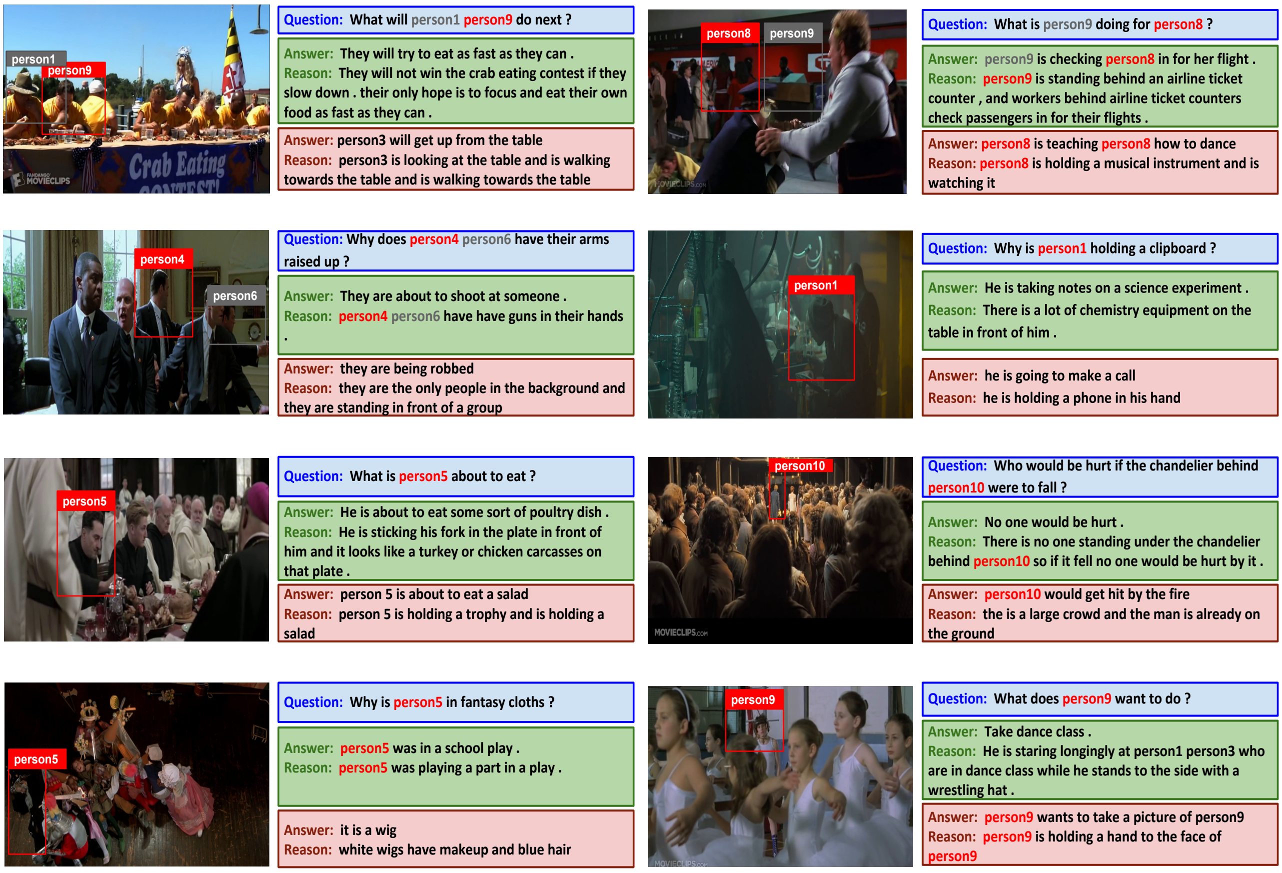


Fig. S2 *(Best viewed in color)* Challenging examples for which our model fails to generate the semantically correct answer and rationale. Blue box = question about image; Green = Ground truth; Red = Generated results from our proposed architecture. (Object regions shown on image are for reader's understanding and are not given as input to model.)

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 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.

S5 How Interpretable are our Results? A Study using Attention Maps

We visualized the attention weights for the model with and without the Refinement module. Addition of a Refinement module causes the model to attend to more relevant regions and leads to better answers and rationales. Figure S6 presents an example to illustrate the visualization of the attention weights used by generation model and generation-refinement model while generating each word of answer and rationale. The attention map visualizations also show that our model is visualizing the appropriate entities in the given image when generating the answer and rationale.

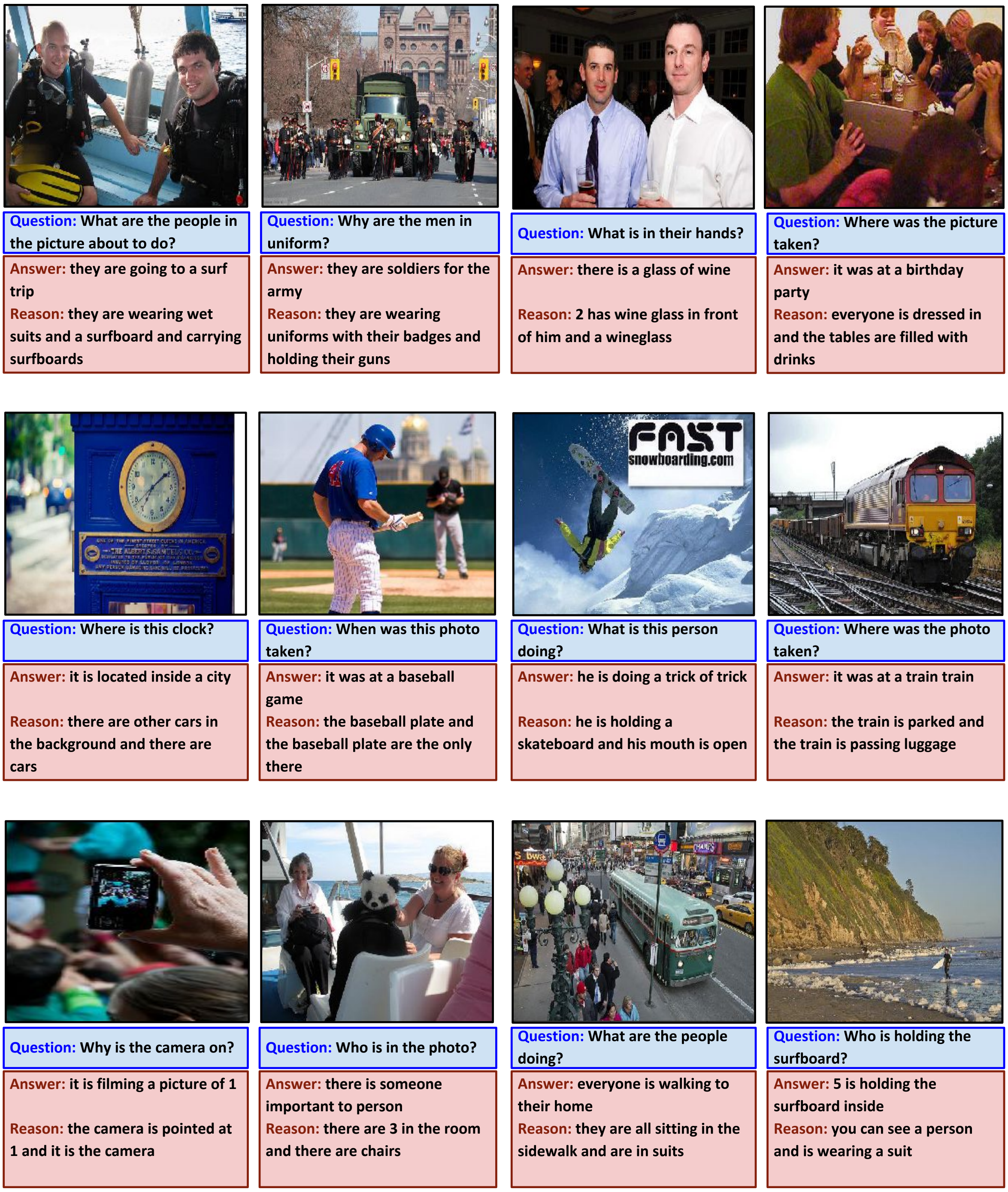


Fig. S3 *(Best viewed in color)* Qualitative results on Visual7W dataset for `ViQAR` task from our proposed Generation-Refinement architecture. Blue box = Question about image; Red box = Generated results from our proposed architecture. (Note that there is Reason provided in the Visual7W dataset, and all the reasons in the figures are generated by our model.)

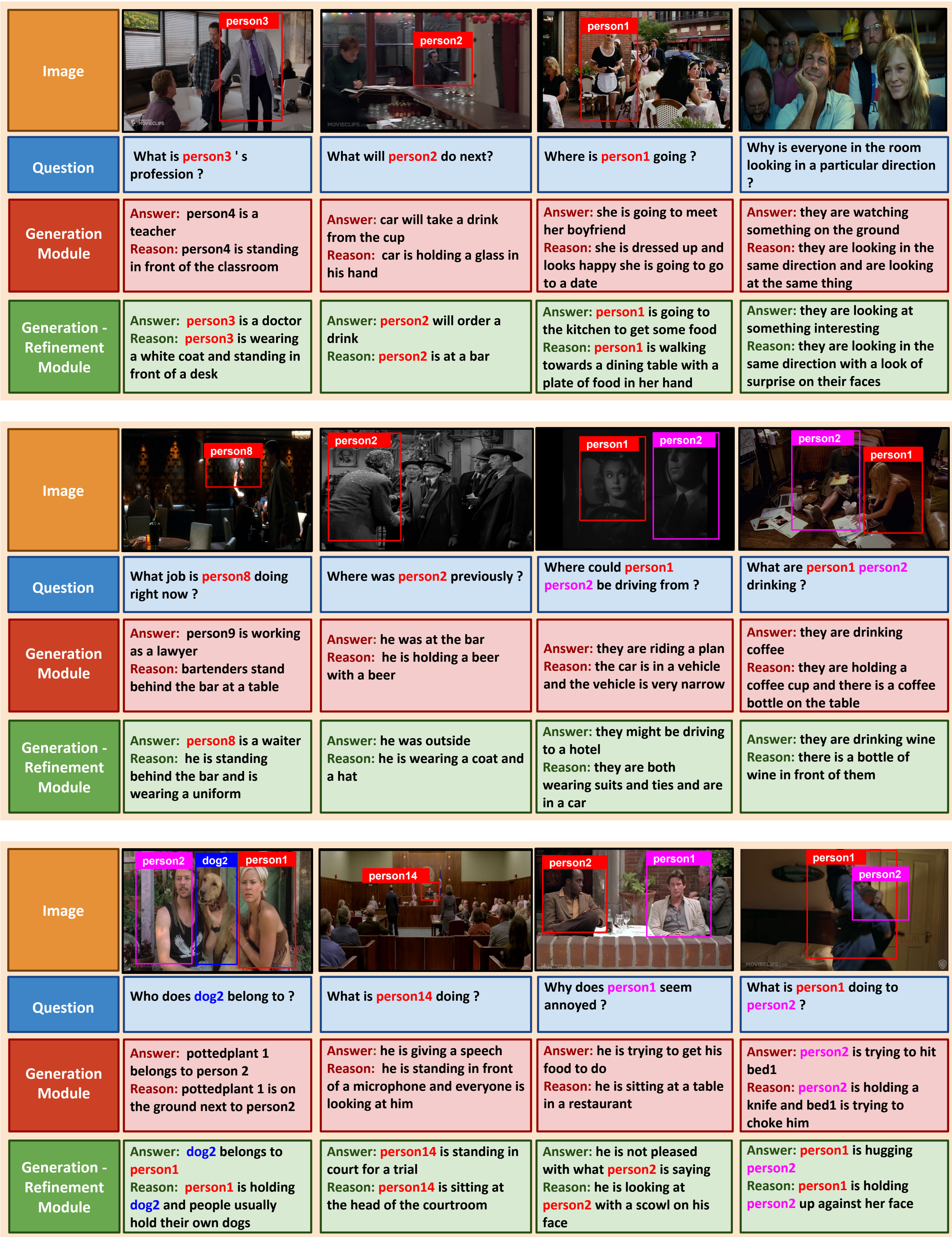


Fig. S4 *(Best viewed in color)* Qualitative results for the model with and without Refinement module. Blue box = question about image; Green = Results from model with Refinement module; Red = Results from model without Refinement module. (Object regions shown on image are for reader's understanding and are not given as input to model.)

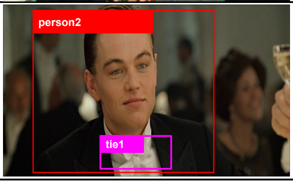
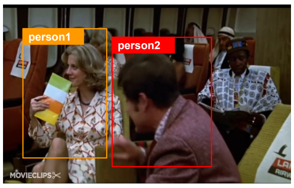
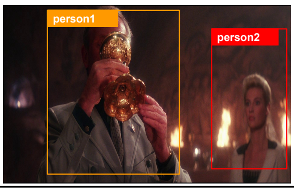
Question	Image	Groundtruth Answer and Reason	Generated Answer and Reason	Evaluation Metrics
What is person1 trying to do to person2 ?		Answer: person1 is trying to get person2 to tell his parents to vote for him . Reason: person1 looks like a politician and is talking to person2 who is a child . children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

Fig. S5 *(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.

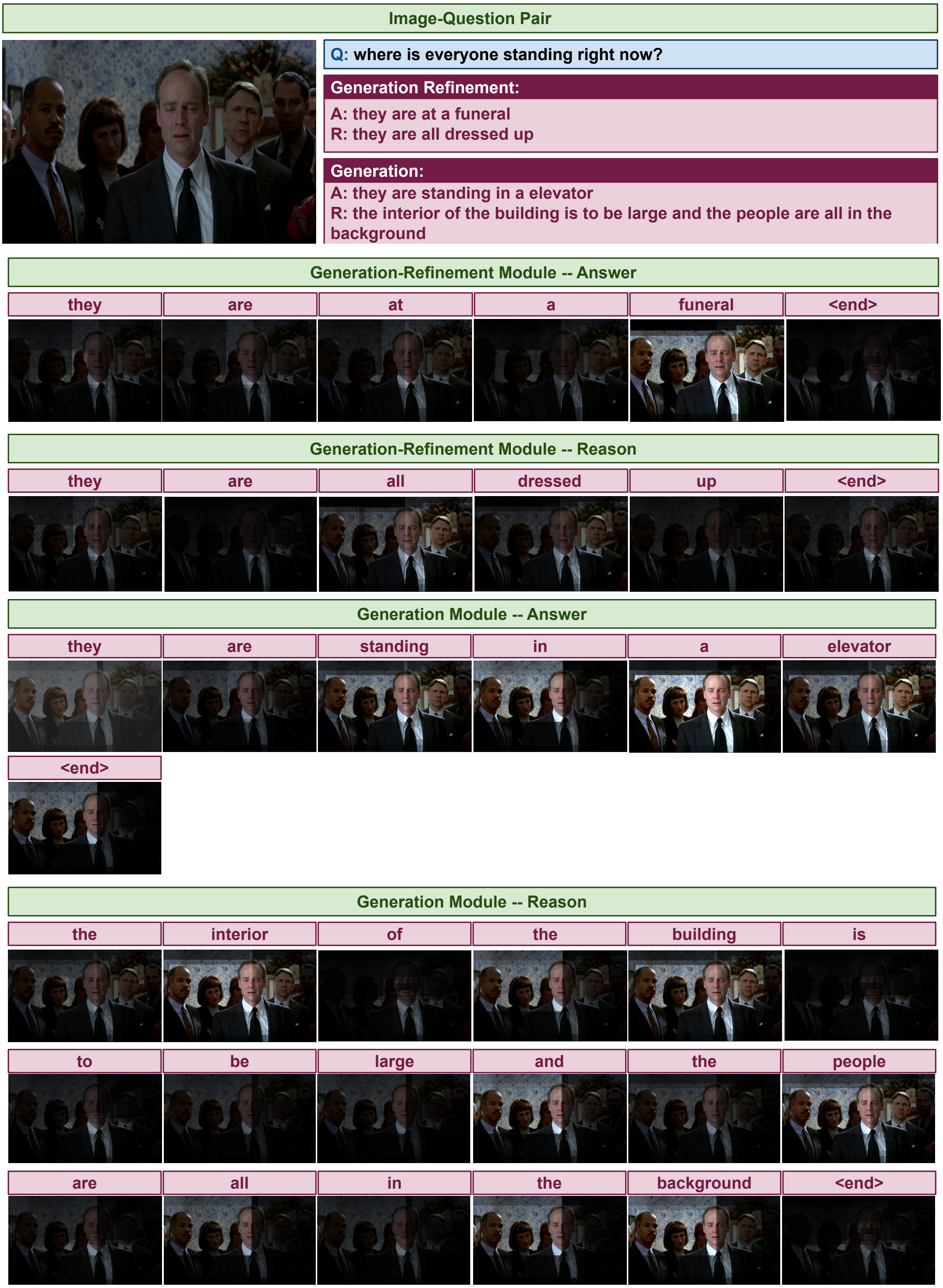


Fig. S6 *(Best viewed in color)* Visualization of attention weights that models use while generating each word.