Maintaining Reasoning Consistency in Compositional Visual Question Answering - Supplementary Material

Chenchen Jing[1], Yunde Jia[1], Yuwei Wu[1]*, Xinyu Liu[1], Qi Wu[2]
[1]Beijing Laboratory of Intelligent Information Technology,
School of Computer Science, Beijing Institute of Technology, China
[2]Australian Centre for Robotic Vision, University of Adelaide, Australia
{chenchen.jing,jiayunde,wuyuwei,liuxinyu18}@bit.edu.cn, qi.wu01@adelaide.edu.au

1. Overview

In this document, we provide:
1. detailed analyses of the GQA-Sub dataset (Sec. 2),
2. more qualitative results of the dialog-like reasoning method (Sec. 3).

2. Dataset analysis

The GQA-Sub dataset is constructed by decomposing the compositional questions in the GQA dataset [1] into sub-questions, in order to enable the quantitative evaluation of the reasoning consistency. Several examples of the sub-questions in the GQA-Sub dataset are shown in Fig. 1. For each example, we show an input image on the left side and list a compositional question and the sub-questions on the right side. The first row shows three compositional questions that have only one sub-question. The second row and the third row show examples with two and three sub-questions, respectively. As shown in the figure, the generated sub-questions are diverse and reasonable. In the following, we use these examples to introduce the dataset statistics and illustrate the details of the dataset curation.

2.1. Dataset statistics

The GQA-Sub dataset contains a train-sub split and a validation-sub split. The train-sub split contains $351,272$ sub-questions decomposed from the $943,000$ compositional questions in the train split of the GQA. The validation-sub split contains $45,043$ sub-questions decomposed from the $132,062$ compositional questions in the validation split of the GQA. Tab. 1 shows the numbers of compositional questions that have k sub-questions in the train split and validation split of the GQA. The reason why most questions have no sub-questions is that there are many sub-questions with similar concepts and answers. So we balance the generated sub-questions by removing most of these sub-questions to avoid dataset biases. For the same reason, a compositional question with fewer sub-questions does not necessarily have fewer visual concepts compared with a compositional question with more sub-questions. For example, the question in Fig. 1 (c) has more visual concepts than the question in Fig. 1 (d).

Train				Validation			
$k=0$	$k=1$	$k=2$	$k\geq 3$	$k=0$	$k=1$	$k=2$	$k\geq 3$
690,949	166,767	71,859	13,425	98,176	24,135	8,392	1,359

Table 1. The numbers of compositional questions that have k sub-questions in the train split and validation split of the GQA.

It is shown from the figure that the GQA-Sub dataset contains different types of sub-questions. Specifically, there are 47 detailed types of sub-questions in the dataset. These question types are listed in Tab. 2. From the semantic perspective, these question types can be divided into four categories: attribute-related (attr), category-related (cat), object-related (obj), and relationship-related (rel). From the structural perspective, these question types can be divided into three categories: choose, query, and verify. Note that, all these question types can be found in the GQA dataset. We only generate sub-questions that belong to question types that appear in the GQA dataset, to guarantee that the generated sub-questions are in-distribution samples for reasoning models trained on the GQA dataset.

2.2. Details of the dataset construction

Language graph generation and traversing. We construct and traverse a language graph, which represents the known visual concepts of a compositional question, to decompose the question into sub-questions. In particular, we use a kind of directed edges in the language graph to denote the referential relationship of the question. In language graph traversing, we start from the node whose in-degree

*corresponding author

Q: What is the person on the motorcycle wearing? (**GT**: helmet)
SQ: Do you see a person on the motorcycle? (**GT**: yes)

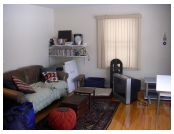
(a)

Q: What is the color of the buses in the middle of the picture? (**GT**: green)
SQ: Are any buses observable? (**GT**: yes)

(b)

Q: Do you see a couch near the wall that is made of brick? (**GT**: yes)
SQ: Is the wall brick or hardwood? (**GT**: brick)

(c)



Q: What are the blue bags sitting on? (**GT**: floor)
SQ1: Do you see a bag that is white? (**GT**: no)
SQ2: Does the bag look blue or tan? (**GT**: blue)

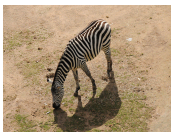
(d)

Q: Are there any shelves to the left of the drawer that is to the right of the TV stand? (**GT**: no)
SQ1: Is there a nightstand? (**GT**: no)
SQ2: What type of furniture is to the right of the TV stand? (**GT**: drawer)

(e)

Q: Does the sheep have the same color as the car? (**GT**: no)
SQ1: What is the vehicle? (**GT**: car)
SQ2: What is the name of the animal? (**GT**: sheep)

(f)



Q: Is the zebra that is eating striped and white? (**GT**: no)
SQ1: What is the name of the animal? (**GT**: zebra)
SQ2: What type of animal is eating? (**GT**: zebra)
SQ3: Is the zebra eating or playing? (**GT**: eating)

(g)

Q: Is the sofa to the left or to the right of the table that is brown? (**GT**: left)
SQ1: Do you see a nightstand? (**GT**: no)
SQ2: What the brown furniture is called? (**GT**: table)
SQ3: Is any sofa observable in this photograph? (**GT**: yes)

(h)

Q: Is the woman to the right of the elephant carrying a bag? (**GT**: no)
SQ1: What is the name of the animal? (**GT**: elephant)
SQ2: Is there a woman that is to the right of the elephant? (**GT**: yes)
SQ3: Is the woman to the right or to the left of the elephant? (**GT**: right)

(i)

Figure 1. Examples of the sub-questions in the GQA-Sub dataset. For each example, we show an input image on the left side and list a compositional question and the sub-questions on the right side. The first row shows three examples that have only one sub-question. The second row and the third row show examples with two and three sub-questions, respectively.

is zero. For example, for a compositional question "Is the woman to the right of the elephant carrying a bag?" (Fig. 1 (i)), the "woman" is referred to by the "elephant" and thus there is an edge from the "elephant" to the "person". Thus the first sub-question is about the "elephant" and the following sub-questions involve the "woman". In particular, for a question that contains two independent objects such as "Does the sheep have the same color as the car?" (Fig. 1 (f)), the language graph has two independent nodes. In this case, we randomly select a node to generate a sub-question and then select another node.

Decoys. We need to generate decoys for two kinds of questions: (1) questions that are about verifying and with an answer "no" such as "Do you see a bag that is white?" for blue bags (the first sub-question in Fig. 1 (d)), (2) questions about choosing such as "Is the zebra eating or playing" (the third sub-question in Fig. 1 (g)). To obtain high-quality decoys, we search the questions with decoys in the GQA dataset and obtain a set of candidate decoys for each concept. In the sub-question generation process, we exploit the scene graph of the corresponding image to select a reasonable decoy, for a specific compositional question. For example, in Fig. 1 (h), the known visual concept in the compositional question is "table", to generate a sub-question with a decoy, we select "nightstand" as the decoy because it is semantically similar to "table" and do not appear in the image. We guarantee the decoys (*e.g.*, objects, attributes, or relations) are reasonable by using the scene graph. For example, for a black dog, we only generate a sub-question as "Is the dog white" with the answer "no" when there is no white dog in the image.

Sampling. We perform three times sampling to balance the generated sub-questions. The first time of sampling and the third time of sampling are from the local-group level like [1] to balance the answer distribution of each group to avoid biases. For each question, we generate a local label that characterizes the semantics of the question. For example, for "What color is the dog?", its local label is "dog-color". By contrast, the global label in [1] for the question should be "color". Then we partition the questions into groups according to their local labels. In each group g, we count the number of corresponding questions $\mathrm{Num}(a, g)$ for each answer a and sort these answers in descending order. At last, we remove some questions to guarantee $\mathrm{Num}(a_{1st}, g) \leq \gamma * \mathrm{Num}(a_{2nd}, g)$, where a_{1st} denotes the answer with most questions in the group and a_{2nd} denotes the answer with the second most questions. γ is a factor that controls the smoothness of the answer distribution and we set it as 1.2.

The second time of sampling is from the visual concept level to guarantee that we only generate one sub-question for each known concept of a compositional question. For each sub-question, we first compute a balance score as $\mathrm{Num}(a_{1st}, g') - \mathrm{Num}(a', g')$, where a' and g' are the answer and the group of the sub-question, respectively. If there are more than one sub-questions for a single visual concept, only the sub-question with the highest balance score remains.

Another solution to balance the dataset is to first perform sampling from the visual concept level and then perform sampling from the local-group level. However, the remained sub-questions are much fewer than performing three times of sampling.

3. Qualitative results

In this section, we provide additional qualitative examples of the proposed method. For each example, we show visual attention maps, numbers of required iterations, and predicted answers of our method for a compositional question and its sub-questions about an input image.

Fig. 2 shows the reasoning processes of our method for two compositional questions that have only one sub-question. The compositional question in Fig. 2 (a) is relatively simple. Our method attends to the cow and the vehicle in the image and provides correct answers to the input questions. In Fig. 2 (b), the compositional question contains multiple objects but has only one sub-question. The sub-question involves two objects and also requires relational reasoning to answer. Our method attends to the critical objects and answers these questions consistently.

Semantic	Structural	Detailed
attr	choose	activityChoose
		chooseAttr
		materialChoose
		stateChoose
		typeChoose
	query	activity
		directOf
		directWhich
		how
		material
		state
	verify	materialVerify
		materialVerifyC
		typeVerify
		typeVerifyC
		verifyAttr
		verifyAttrC
		verifyAttrCThis
		verifyAttrThis
cat	choose	categoryThatChoose
		categoryThisChoose
		objThisChoose
	query	activityWho
		category
		categoryAttr
		categoryThat
		categoryThis
obj	verify	exist
		existAttr
		existAttrC
		existC
		existMaterial
		existMaterialC
		existThat
		existThatC
rel	choose	catcategoryRelOChoose
		relChooser
	query	catcategoryRelO
		catcategoryRelS
		relO
		relS
	verify	existRelS
		existRelSC
		existRelSRC
		relVerify
		relVerifyCo
		relVerifyCr

Table 2. The question types of the GQA-Sub Dataset.

Fig. 3 depicts two compositional questions that have two sub-questions. These compositional questions contain two relatively simple sub-questions. Our method accurately locates the corresponding objects in the image for each sub-question. The visual attention map for the compositional question is relatively smooth, especially in Fig. 3 (b). The possible reason is that after several iterations of graph convolution, the node representations contain too much contextual information and these representations may be similar.

Fig. 4 depicts two complex compositional questions that have three sub-questions and requires strong relational reasoning ability to answer. We observe that throughout the reasoning process, our method accurately answers the simpler sub-questions. Specifically, for sub-questions that only contain one object, our method usually focuses on the critical object with a high attention value. For sub-questions about the relations of two objects, our method can still attend to corresponding objects. Thus we are more convinced that the model relies on compositional reasoning to predict the answers to the original compositional questions rather than dataset biases.

Fig. 5 depicts two typical failure cases of our method. In both cases, the proposed method fails to maintain reasoning consistency. It provides the correct answer to the compositional question but doesn't answer the sub-question accurately. In Fig. 5 (a), the method attends to the cucumber and the bread accurately but makes a wrong prediction for the sub-question. Considering the spatial relational is easier to determine, a possible reason for the wrong prediction is that the but in the image is hard to recognize for the answer classifier. In this case, a more advanced answer prediction module may be beneficial. In Fig. 5 (b), the method fails to attend to the pair of glasses in the two sub-questions and thus answers both sub-questions wrongly. The main reason is that the pair of glasses is too small to attend to. A possible solution is to use a stronger object detector in visual graph construction.

These qualitative results clearly demonstrate the effectiveness of dialog-like reasoning. On the one hand, the answering of sub-questions makes the overall reasoning process more understandable. On the other hand, it provides some hints to improve our method.

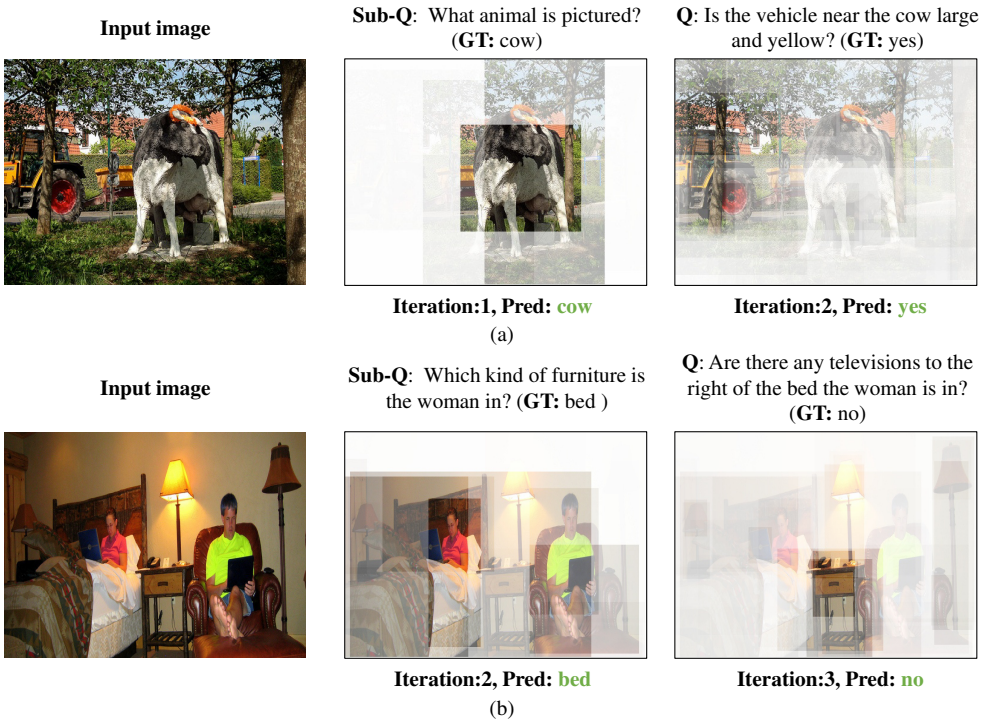


Figure 2. Qualitative examples about compositional questions with only one sub-question. For each compositional question and its sub-questions about an input image, we provide the visual attention maps, the number of required iterations, and the predicted answers of our method.

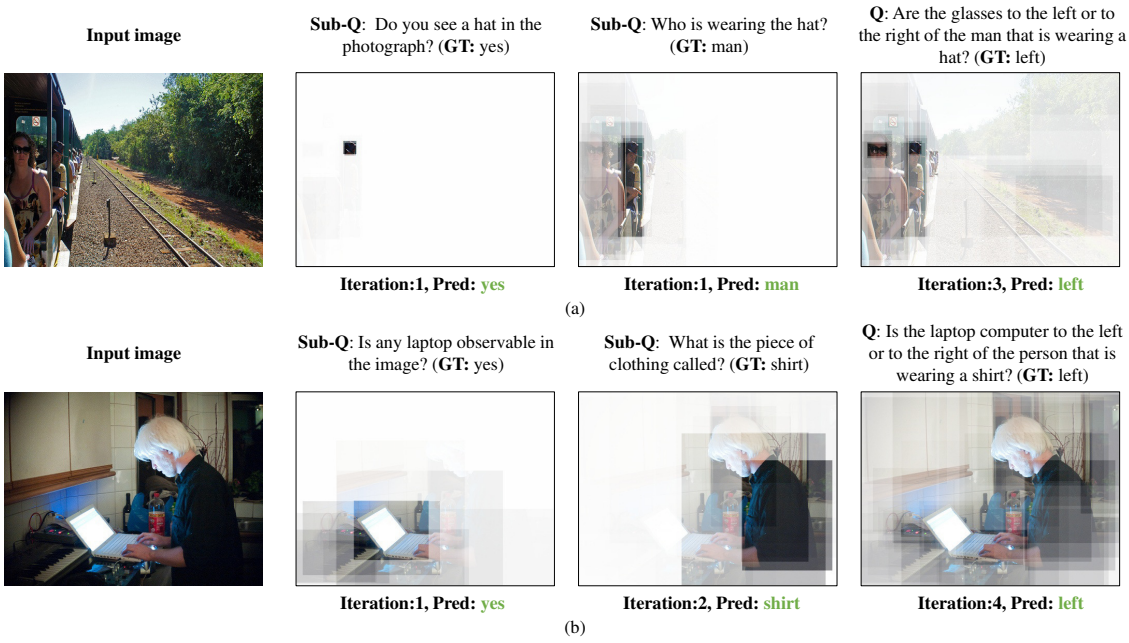


Figure 3. Qualitative examples about compositional questions with two sub-questions. For each compositional question and its sub-questions about an input image, we provide the visual attention maps, the number of required iterations, and the predicted answers of our method.

Figure 4. Qualitative examples about compositional questions with three sub-questions. For each compositional question and its sub-questions about an input image, we provide the visual attention maps, the number of required iterations, and the predicted answers of our method.

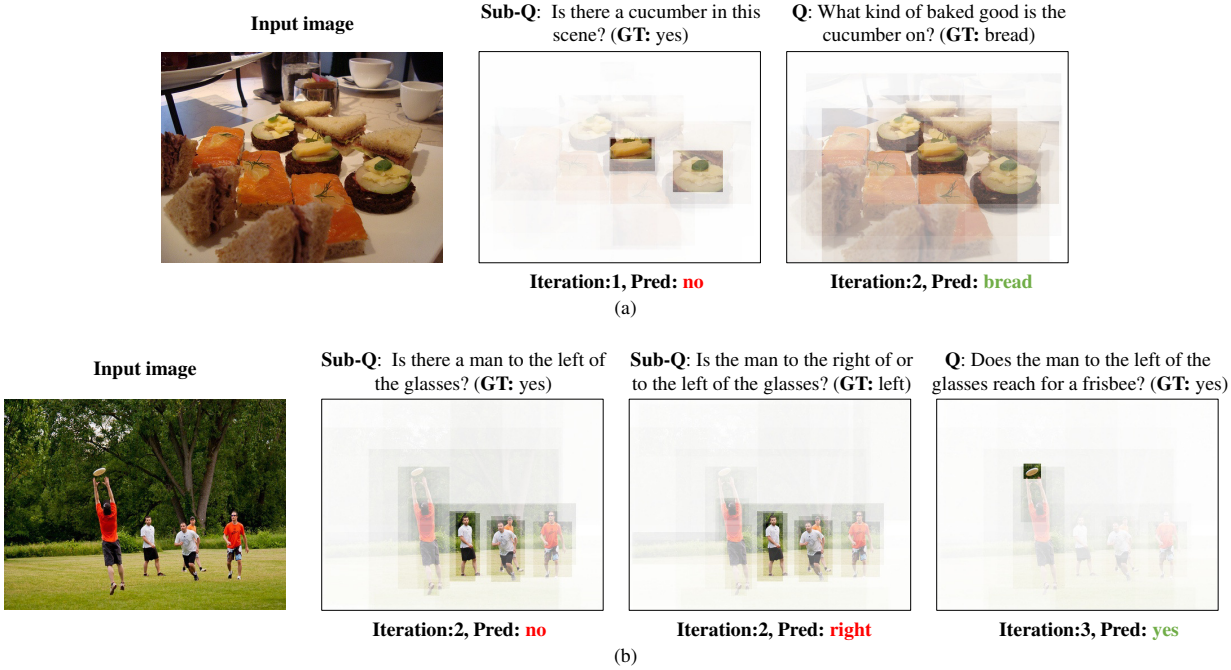


Figure 5. Failure cases of the proposed method. In each case, we provide the visual attention maps, the number of required iterations, and the predicted answers of our method, for a compositional question and its sub-questions about an input image.

References

[1] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6700–6709, 2019. 1, 2