Maintaining Reasoning Consistency in Compositional Visual Question Answering

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

Abstract

A compositional question refers to a question that contains multiple visual concepts (e.g., objects, attributes, and relationships) and requires compositional reasoning to answer. Existing VQA models can answer a compositional question well, but cannot work well in terms of reasoning consistency in answering the compositional question and its sub-questions. For example, a compositional question for an image is: "Are there any elephants to the right of the white bird?" and one of its sub-questions is " Is any bird visible in the scene?". The models may answer "yes" to the compositional question, but "no" to the sub-question. This paper presents a dialog-like reasoning method for maintaining reasoning consistency in answering a compositional question and its sub-questions. Our method integrates the reasoning processes for the sub-questions into the reasoning process for the compositional question like a dialog task, and uses a consistency constraint to penalize inconsistent answer predictions. In order to enable quantitative evaluation of reasoning consistency, we construct a GQA-Sub dataset based on the well-organized GQA dataset. Experimental results on the GQA dataset and the GQA-Sub dataset demonstrate the effectiveness of our method.

1. Introduction

Compositional visual question answering (VQA) [3, 9, 27] is the task of providing an answer to a compositional question about an image based on the content of the image. A compositional question refers to a question that contains multiple visual concepts (*e.g.*, objects, attributes, and relationships). The task requires a comprehensive understanding of the multi-modal inputs and compositional relational reasoning based on the understanding.

Figure 1. Qualitative examples showing the inconsistency of existing compositional VQA models. For each example, we show an image on the left side. A compositional question (Q) and its sub-questions (Sub-Q) with ground-truth answers and predicted answers of two reasoning models (*i.e.*, the LCGN [9] and the MMN [4]) are shown on the right side.

Existing reasoning models for compositional VQA can answer a compositional question well, but cannot work well in terms of reasoning consistency in answering the compositional question and its sub-questions. A compositional question usually contains various known visual concepts in the image. To answer the question, reasoning models are supposed to infer unknown concepts based on the known concepts. For a compositional question " Are there any elephants to the right of the white bird?" (shown in Fig. 1 (a)), the known concepts are "white" and "bird", while the unknown concepts are "elephants" and "to the right of". Obviously, the compositional question reveals answers to sub-questions about these known concepts, such as " Is any bird visible in the scene?", and "What the white animal is called?". The question requires a stronger reasoning abil-

---

*corresponding author

ity to answer than its sub-questions. However, although a model accurately answers the compositional question, it may fail to provide correct answers for its sub-questions as shown in Fig. 1. The inconsistency indicates errors in the reasoning process for the compositional question.

This paper proposes a dialog-like reasoning method that integrates the reasoning processes for sub-questions into the reasoning process for a compositional question to maintain the reasoning consistency in compositional VQA. The method represents an input image as a structured visual graph and performs iterative language-guided graph convolution to learn contextual visual representations for compositional reasoning. The number of required iterations for each question is determined in advance. In the reasoning process of a compositional question, if the current number of iterations equals the required number of iterations of a sub-question, we use an answer classifier to answer the corresponding sub-question using current visual representations. A consistency constraint is further introduced to penalize inconsistent answer predictions. By answering sub-questions with the guidance of the compositional question, the method is supposed to capture the correlations of the question and its sub-questions for better consistency. The answering process for the sub-questions can also be regarded as the intermediate supervision of the reasoning process for the compositional question.

To enable the quantitative evaluation of reasoning consistency in compositional VQA, we build a GQA-Sub dataset based on the GQA dataset [11], a well-organized large-scale dataset for compositional VQA. We automatically decompose compositional questions of the GQA into sub-questions, and carefully balance the obtained sub-questions to avoid biases. Given a compositional question and its sub-questions, reasoning consistency can be quantitatively measured by evaluating whether predictions for these questions are contradictory. Experimental results on the GQA and the GQA-Sub demonstrate the effectiveness of our method. The data and code are publicly available at *https://github.com/jingchenchen/ReasoningConsistency-VQA*.

The contributions of this paper are two-fold:

1. We propose a dialog-like reasoning method that integrates reasoning processes for the sub-questions into the reasoning process for a compositional question to ensure reasoning consistency in compositional VQA.

2. We present a GQA-Sub dataset to evaluate the reasoning consistency of compositional VQA models.

2. Related Work

Consistency in VQA Consistency in VQA is defined as being able to answer questions posed from different semantic perspectives about a certain visual fact without any con-
tradiction [19]. Specifically, a visual fact is defined as a triplet in a scene graph of an image. Recent work has proposed various datasets to measure the consistency of VQA models. Ribeiro *et al.* [20] first studied the inconsistency of VQA models and argue that we need to consider the relationship between predictions to measure true understanding. They automatically generate implications for questions in the VQA v1 dataset [3] to evaluate the consistency. The implications are questions about the same fact as the original question. Given a question-answer pair as "What room is this? bathroom", they generate implications such as "Is this a bathroom?" and "Is there a bathroom in the picture?". Ray *et al.* [19] generate entailed questions for questions in the VQA v2 dataset [5]. The entailed questions are also about the same fact as the original question. Selvaraju *et al.* [22] distinguish reasoning questions and perception questions and build datasets based on the VQA v2 to test the consistency of models for reasoning questions and perception questions. Yuan *et al.* [28] transform counting question of the VQA v2 by performing object-oriented partition, reordering, or reversion, to test the perception ability of VQA models. Shah *et al.* [23] and Whitehead *et al.* [26] rephrase questions in the VQA v2 to evaluate the VQA models' robustness for linguistic variations. Previous work generates questions about one visual fact in the original question to measure the consistency of VQA models. By contrast, we focus on generating sub-questions about known visual facts in the compositional questions to test whether the VQA models are really capable of compositional reasoning.

Hudson and Manning [11] devise a consistency metric by generating entailed questions for compositional questions in the GQA [11]. In Sec. 4, we analyze the differences between entailed questions of the GQA and our sub-questions in detail.

Compositional VQA. Existing methods for compositional VQA can be mainly divided into two categories: modular and holistic. Modular methods [2, 4, 8, 24] assemble various modules according to an input question and execute the modules for reasoning. Holistic methods [9, 10, 12, 13, 16, 29] use a single model for different inputs to achieve reasoning. Our work aims to ensure the reasoning consistency for compositional VQA and is orthogonal to previous work.

3. Method

3.1. Overview

The compositional VQA task is to provide an answer A for a natural language compositional question Q about an image I. The image I can be represented by a set of objects $\mathcal{O}_v = \{o_i\}_{i=1}^M$, where M is the number of objects in the image. A compositional question Q contains multiple concepts such as objects, relations, and attributes.
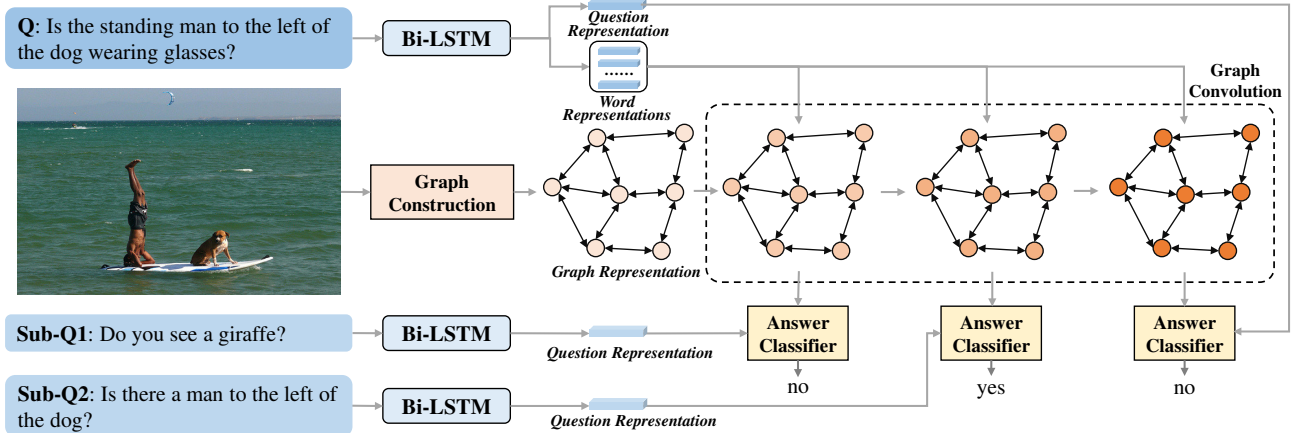
Figure 2. The reasoning process of the proposed method for an input compositional question, which has two sub-questions. We represent an input image as a visual graph and obtain its representation via graph construction. For simplicity, not all edges in the graph are shown. A Bi-LSTM is used to obtain question representations of all questions and word representations of the compositional question. According to the word representations, we perform iterative graph convolution to gradually learn contextual visual representations. The numbers of required iterations of graph convolution for the question, the first sub-question, and the second sub-question are 3, 1, 2, respectively. After each iteration, an answer classifier is used to answer the corresponding question based on its representation and current graph representations.

We decompose the question Q into a set of sub-questions $\mathcal{Q} = \{Q_i\}_{i=1}^N$, where N is the number of sub-questions. We use $\mathcal{Q}^f = \{Q\} \cup \mathcal{Q}$ to denote a set of questions consisting of the question Q and its sub-questions $\mathcal{Q}$.

As shown in Fig. 2, we represent an input image as a visual graph to characterize its relational layout. For each question $q \in \mathcal{Q}^f$, the proposed method performs T_q iterations of graph convolution guided by the question q to learn contextual visual representations to answer the question. Specifically, the number of iterations T_q is determined by performing Part-of-speech (POS) tagging for the question via the Spacy tool [7] and computing the number of nouns in it. An answer classifier $\mathcal{F}$ is used to fuse the question and the graph to predict the answer. In particular, in the reasoning process of the compositional question Q, after the t-th iteration, where $t \in \{1, 2, ..., T_Q\}$, we use the classifier $\mathcal{F}$ to answer a sub-question $Q_i \in \mathcal{Q}$, if and only if $t = T_{Q_i}$, according to the current graph.

3.2. Feature Encoding

For an input image, we build a visual graph $G = \{V, E\}$, where $V = \{v_i\}_{i=1}^M$ is a set of nodes and $E = \{e_{ij}\}_{i,j=1}^M$ denotes the relationships among objects. An undirected edge e_{ij} connects each pair of nodes v_i and v_j. Thus G is a fully-connected graph. For each node v_i, we extract an appearance feature $\boldsymbol{l}_i \in \mathbb{R}^{d_l \times 1}$ via the bottom-up attention model [1]. We encode the location and size information via a spatial feature $\boldsymbol{b}_i \in \mathbb{R}^{d_b \times 1}$. Then we obtain the node embedding $\boldsymbol{v}_i = \boldsymbol{W}_v[\boldsymbol{l}_i; \boldsymbol{b}_i]$ by concatenating the two features and projecting the concatenated feature to a common space $\mathbb{R}^{d \times 1}$, where $\boldsymbol{W}_v \in \mathbb{R}^{d \times (d_l + d_b)}$ is a learnable matrix.

$[\,\cdot\,;\,\cdot\,]$ denotes the concatenation operation of two vectors. We do not obtain representations for edges in G because the representations are not involved in graph convolution.

For a question $q \in \mathcal{Q}^f$ containing K words $\{w_k\}_{k=1}^K$, we use a Bi-LSTM [21] to encode the question and project it into the common space to obtain the sentence-level representation $\boldsymbol{q} \in \mathbb{R}^{d \times 1}$. The word representation $\boldsymbol{w}_k \in \mathbb{R}^{d \times 1}$ of the word w_k is obtained by concatenating corresponding forward and backward hidden vectors and projecting it into the common space.

3.3. Architecture

The proposed method uses the language-guided graph convolution to learn context-aware visual representations for each question $q \in \mathcal{Q}^f$, to achieve reasoning. For each node v_i, we first build a local representation $\boldsymbol{v}_i^{loc} = \boldsymbol{v}_i$ and initialize a contextual representation $\boldsymbol{v}_{i,0}^{ctx} \in \mathbb{R}^{d \times 1}$ from a learned parameter as [9].

At the t-th iteration of graph convolution, the two representations are fused to obtain a joint representation $\tilde{\boldsymbol{v}}_{i,t} \in \mathbb{R}^{3d \times 1}$ as

$$\tilde{\boldsymbol{v}}_{i,t} = \left[\boldsymbol{v}_i^{loc}; \boldsymbol{v}_{i,t-1}^{ctx}; \left(\boldsymbol{W}_1 \boldsymbol{v}_i^{loc}\right) \circ \left(\boldsymbol{W}_2 \boldsymbol{v}_{i,t-1}^{ctx}\right)\right], \quad (1)$$

where $\circ$ denotes element-wise multiplication. $\boldsymbol{W}_1 \in \mathbb{R}^{d \times d}$ and $\boldsymbol{W}_2 \in \mathbb{R}^{d \times d}$ are learnable matrices. We use a language attention to generate a command vector $\boldsymbol{c}_t \in \mathbb{R}^{d \times 1}$ as

$$\begin{aligned} \boldsymbol{c}_t &= \sum_{k=1}^K \alpha_{t,k}^L \boldsymbol{w}_k, \\ \alpha_{t,k}^L &= \text{Softmax}_k \left(\boldsymbol{W}_3 \left(\boldsymbol{w}_k \circ \left(\boldsymbol{W}_4^t \left(\boldsymbol{W}_5 \boldsymbol{q}\right)\right)\right)\right), \end{aligned} \quad (2)$$

where $\boldsymbol{W}_3 \in \mathbb{R}^{1 \times d}$, $\boldsymbol{W}_4^t \in \mathbb{R}^{d \times d}$, and $\boldsymbol{W}_5 \in \mathbb{R}^{d \times d}$ are learnable matrices.

Then we update the contextual representation $\boldsymbol{v}_{i,t}^{ctx}$ for each node v_i by using the joint representations $\{\tilde{\boldsymbol{v}}_{i,t}\}_{i=1}^M$ of both the connected nodes and the node itself, based on the the command vector $\boldsymbol{c}_t$. The $\boldsymbol{v}_{i,t}^{ctx}$ is computed by

$$
\begin{aligned}
\boldsymbol{v}_{i,t}^{ctx} &= \boldsymbol{W}_6 \left[\boldsymbol{v}_{i,t-1}^{ctx}; \hat{\boldsymbol{v}}_{i,t}^{ctx} \right], \\
\hat{\boldsymbol{v}}_{i,t}^{ctx} &= \sum_{j=1}^M \beta_{i,j}^t \circ (\boldsymbol{W}_7 \tilde{\boldsymbol{v}}_{j,t}), \\
\beta_{i,j}^t &= \mathrm{Softmax}_j \left(((\boldsymbol{W}_8 \tilde{\boldsymbol{v}}_{i,t}) \circ (\boldsymbol{W}_9 \boldsymbol{c}_t)) (\boldsymbol{W}_{10} \tilde{\boldsymbol{v}}_{j,t})^T \right),
\end{aligned} \tag{3}
$$

where $\boldsymbol{W}_6 \in \mathbb{R}^{d \times 2d}$, $\boldsymbol{W}_7 \in \mathbb{R}^{d \times 3d}$, $\boldsymbol{W}_8 \in \mathbb{R}^{d \times 3d}$, $\boldsymbol{W}_9 \in \mathbb{R}^{d \times d}$, and $\boldsymbol{W}_{10} \in \mathbb{R}^{d \times 3d}$ are learnable matrices.

After the graph convolution, an answer classifier is used for answer prediction. The answer classifier fuses local and contextual representations of nodes to obtain output representations, uses the top-down attention [1] to aggregate output representations, and uses the aggregated representations to obtain the predicted answer distribution $\boldsymbol{y}_{pred}^q$ as

$$
\begin{aligned}
\boldsymbol{y}_{pred}^q &= \boldsymbol{W}_{ans} \left(\boldsymbol{W}_{11} \left[\sum_{i=1}^M \alpha_i^V \boldsymbol{v}_{i,T_q}^{out}; \boldsymbol{q} \right] \right), \\
\alpha_i^V &= \mathrm{Softmax}_i \left(\boldsymbol{W}_{12} \left(\boldsymbol{v}_{i,T_q}^{out} \circ (\boldsymbol{W}_{13} \boldsymbol{q}) \right) \right), \\
\boldsymbol{v}_{i,T_q}^{out} &= \boldsymbol{W}_{14} \left[\boldsymbol{v}_i^{loc}; \boldsymbol{v}_{i,T_q}^{ctx} \right],
\end{aligned} \tag{4}
$$

where $\boldsymbol{W}_{ans} \in \mathbb{R}^{N^{ans} \times d}$, $\boldsymbol{W}_{11} \in \mathbb{R}^{d \times 2d}$, $\boldsymbol{W}_{12} \in \mathbb{R}^{1 \times d}$, $\boldsymbol{W}_{13} \in \mathbb{R}^{d \times d}$, and $\boldsymbol{W}_{14} \in \mathbb{R}^{d \times 2d}$ are learnable matrices. N^{ans} is the number of answer categories. For simplicity, we use a function $\mathcal{F}$ to represent the classifier. Then we have $\boldsymbol{y}_{pred}^q = \mathcal{F}(\boldsymbol{q}, \boldsymbol{V}_{T_q}^q)$, where $\boldsymbol{V}_{T_q}^q$ denotes node representations after T_q iterations of graph convolution guided by a question q.

For a compositional question Q, after the t-th iteration, if t equals T_{Q_i}, the number of required iterations for a sub-question Q_i, we use the answer classifier to predict the answer for the sub-question Q_i as $\hat{\boldsymbol{y}}_{pred}^{Q_i} = \mathcal{F}(\boldsymbol{Q}_i, \boldsymbol{V}_{T_{Q_i}}^Q)$, where $\hat{\boldsymbol{y}}_{pred}^{Q_i}$ is the predicted answer distribution for a sub-question Q_i in the reasoning process for the compositional question Q.

3.4. Optimization

We use the cross-entropy loss to supervise the answering process of the proposed method. For each questions $q \in \mathcal{Q}^f$, the VQA loss is computed by $\mathcal{L}_{vqa} = -\boldsymbol{y}_{gt}^q \log(\boldsymbol{y}_{pred}^q)$, where $\boldsymbol{y}_{gt}^q$ is an one-hot label for question q. Considering we answer sub-questions $\mathcal{Q}$ in the reasoning process for the corresponding compositional question Q, we have an extra VQA loss for sub-questions. Specifically, for a sub-question Q_i, the loss is given by $\mathcal{L}_{vqa}^{sub} = -\boldsymbol{y}_{gt}^{Q_i} \log(\hat{\boldsymbol{y}}_{pred}^{Q_i})$.

We introduce a consistency constraint to encourage the model to answer the compositional question and the sub-questions consistently. We believe that sub-questions $\mathcal{Q}$ should be answered more confidently and more accurately than the compositional question Q because the answering of the sub-question is the basis of answering the compositional question. We use $\hat{a}_{pred}^{Q_i}$ to denote the predicted probability for the ground-truth answer of a sub-question Q_i in the reasoning process of the compositional question Q, and a_{pred}^Q is the predicted probability for the ground-truth answer of a Q. The consistency constraint enforces the $\hat{a}_{pred}^{Q_i}$ to be higher than a_{pred}^Q as

$$
\mathcal{L}_{cons} = \max(\log(a_{pred}^Q) - \log(\hat{a}_{pred}^{Q_i}), 0). \tag{5}
$$

The consistency constraint is also used to guarantee the compatibility of a_{pred}^Q and $a_{pred}^{Q_i}$.

The overall objective of the proposed method is given by

$$
\mathcal{L} = \mathcal{L}_{vqa} + \lambda_{sub} \mathcal{L}_{vqa}^{sub} + \lambda_{cons} \mathcal{L}_{cons}, \tag{6}
$$

where λ_{sub} and λ_{cons} are two hyper-parameters that balance the loss terms. Considering that at the beginning of the training stage, the predictions of the model may not be accurate enough for consistency constraint to be effective, we first set λ_{cons} in Eq. (6) as 0 and train the model for several epochs, and then train it with the full objective.

3.5. Implementation Details

For the visual input, the dimensions of the appearance features d_l, the spatial features d_b and the common space d are set as 2048, 96, and 512, respectively. We keep the top 48 bounding boxes ranked by confidence score as [4] with the positional information of each bounding box in the form of [top-left-x, top-left-y, bottom-right-x, bottom-right-y], normalized by the image width and height. For the language input, we first use the pre-trained GloVe [18] to initialize the 300-d embedding of words and then input them to the Bi-LSTM. The number of training epochs is set as 25 and we first train the model for 5 epochs without the consistency constraint. The hyper-parameter λ_{sub} and λ_{cons} are set as 0.05 and 0.01, respectively.

4. GQA-Sub dataset

In this section, we introduce the GQA-Sub dataset, which enables the quantitative evaluation of reasoning consistency for compositional VQA models. The GQA-Sub dataset is constructed based on the GQA dataset [11], a large-scale dataset for real-world visual reasoning and compositional question answering. The GQA contains four splits: a train split for training, a validation split and a test-dev split for validation, and a test split for online testing. We only generate sub-questions for questions of the train split and the validation split of GQA because the ground-truth scene graphs of the two splits are available.

Question: What is the name of the animal that is watching the black television above the DVD player? (**GT**: cat)

Scene Graph

(b)

cat → orange   screen

watching   black   to the left of

television → above → DVD player

Language Graph

(c)

black ④

③   ②   ①

television → above → DVD player

Sub-questions:

(d)
1. Is there a DVD player? (**GT**: yes)

2. Is the black television above or under the DVD player? (**GT**: above)

3. Is there a black television above the DVD player? (**GT**: yes)

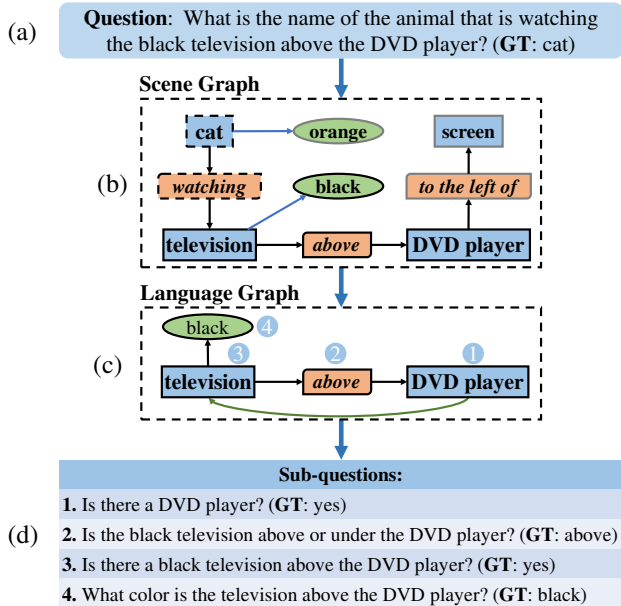**4.** What color is the television above the DVD player? (**GT**: black)

Figure 3. Illustration of the sub-question generation process. (a) shows a compositional question. (b) is the scene graph of an image about the question. In the graph, we use black solid borders to denote known concepts in the question, black dotted borders to denote unknown concepts, and gray borders to denote unmentioned concepts. For simplicity, not all unmentioned concepts are shown. Black edges denote visual relations between objects and blue edges link objects and their attributes. (c) shows the language graph composed of the known concepts. The green edge denotes the referential relationship. (d) shows the sub-questions generated by traversing the language graph.

4.1. Sub-question Generation Pipeline

In the following, we illustrate how we obtain sub-questions by decomposing compositional questions of the GQA dataset. The ground-truth scene graphs for images and functional programs for questions are used for sub-question generation. The graphs and programs are provided by the GQA. As preparations, we manually annotate patterns for each type of question to generate sub-questions.

Language Graph Generation. To decompose a compositional question into multiple sub-questions, we first construct a language graph to represent the known visual concepts of the question. Generally, a "known concept" can be determined by only checking a question, while an "unknown concept" can not be determined until a corresponding image is observed. As shown in Fig. 3 (a), for a compositional question "What is the name of the animal that is watching the black television above the DVD player" with the answer "cat", the known concepts in the question are "DVD player", "above", "television", and "black" while the unknown concepts are "cat" and "watching". We obtain the known visual concepts of the question based on its func-

tional program and confirm these concepts indeed exist in the scene graph as shown in Fig. 3 (b). A language graph as shown in Fig. 3 (c) is further constructed to represent these concepts. In particular, we add a kind of directed edges to denote the referential relationship in the language graph. The "television" is determined by the "DVD player" and "above", which means there may be more than one black television but only one is above the DVD player. Thus we add an edge from the "DVD player" to the "television".

Language Graph Traversing. Then we traverse the language graph to generate sub-questions, as shown in Fig. 3 (c) and (d). We sequentially select visual concepts from the root node, whose in-degree is zero, to other nodes alongside the edges representing referential relationships. Specifically, we generate three kinds of sub-questions about the three kinds of visual concepts: existential questions about objects, relational questions about relations, and questions about an attribute. For a given visual concept, various types of questions can be generated. For each question type, we randomly select a pattern and generate a sub-question.

Decoys. We generate decoys for two kinds of questions: (1) questions that are about verifying and with an answer "no" such as "Is the dog white" for a black dog, (2) questions about choosing such as "Is the dog brown or black". We exploit the questions of the GQA dataset and obtain a set of high-quality decoys for each concept such as "white" and "brown" for "black".

Balancing. Finally, we balance the generated sub-questions to avoid language biases. We perform three times of sampling for these generated sub-questions. Please refer to the supplementary material for details of the sampling.

4.2. Dataset Analysis

After all the stages, we obtain $351, 272$ and $45, 043$ sub-questions for the train split and the validation split of the GQA dataset, respectively. These sub-questions form two splits: a train-sub split and a validation-sub split for the proposed GQA-Sub dataset. As shown in Fig. 4 (a), the generated sub-questions are informative and expressive. Besides, the types of our sub-questions are diverse.

We compare our sub-questions and the entailed questions, which are used to evaluate the consistency metric in the GQA dataset, in Fig. 4 (b). It is shown that the entailed questions are different from our sub-questions. Firstly, some entailed questions are paraphrases of the original question. Thus the questions do not require different strengths of reasoning ability to answer. Secondly, some entailed questions are redundant and similar. The number of entailed questions for compositional questions varies significantly. While the first question only has one entailed question, the second question has 27 entailed questions. We do not show all entailed questions for simplicity. Among the 27 entailed questions, most of them are similar or even

Q: What is the food on top of the metal counter? (**GT**: pink)
Sub-Q1: Do you see a counter that is wood in the photograph? (**GT**: no)
Sub-Q2: How does the counter look, metal or cardboard? (**GT**: metal)

Q: What is the black bike leaning on? (**GT**: tree)
Sub-Q1: Is there a bike that is black? (**GT**: yes)
Sub-Q2: What color is the bike? (**GT**: black)

Entailed-Q: What is the bike leaning on? (**GT**: tree)

Q: The dog to the right of the woman sits on what? (**GT**: blanket)
Sub-Q1: What is this, a woman or a man? (**GT**: woman)
Sub-Q2: What kind of animal is to the right of the woman? (**GT**: dog)
Sub-Q3: Is the dog to the left or to the right of the woman? (**GT**: right)

(a)

Q: Do both the vehicle to the right of the car and the van to the left of the spectators have white color? (**GT**: yes)
Sub-Q1: Is there a vehicle that is to the right of the car? (**GT**: yes)
Sub-Q2: Is the vehicle to the right or to the left of the car? (**GT**: right)
Sub-Q3: Which kind of vehicle is to the left of the spectators? (**GT**: van)

Entailed-Q1: Does the trailer have black color? (**GT**: yes)
Entailed-Q2: Does the trailer to the right of the car look black? (**GT**: no)
Entailed-Q3: Is that trailer white or blue? (**GT**: white)
Entailed-Q4: Is that trailer white or blue? (**GT**: white)
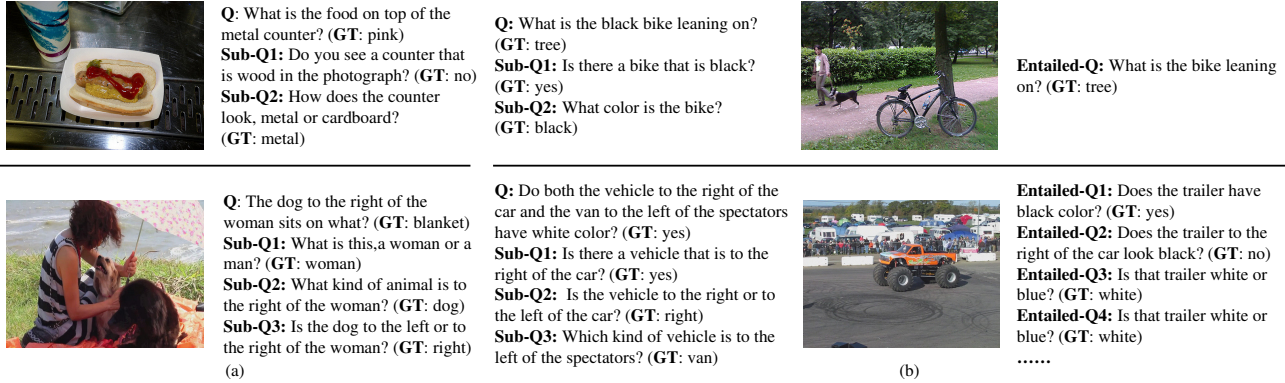**......**

(b)

Figure 4. (a) Examples of the GQA-Sub dataset. For each image on the left, we list a compositional question and the sub-questions on the right side. (b) Comparisons of the entailed questions in GQA and our sub-questions. For each example, we list a compositional question and its sub-questions on the left side, show an image in the middle, and list entailed questions of the compositional question on the right.

the same. Thus these entailed questions can not be used for evaluation of the reasoning consistency.

4.3. Reasoning Consistency Score

For a reasoning model to be consistent for questions and the sub-questions, the answers to all the questions should be correct. To measure the reasoning consistency of models, a simple solution is to compute the accuracy of models for answering sub-questions. But we argue that this score only measures the ability to answer sub-questions correctly rather than the ability to answer compositional questions and sub-questions consistently. An ideal consistency score should take into account both two kinds of questions.

We devise a reasoning consistency score $RC(k)$, which measures the ability of reasoning models to answer compositional questions that have at least k sub-questions consistently. The $RC(k)$ is computed by

$$RC(k) = \frac{\sum_{Q \subset \mathbb{Q}, N >= k} \text{Correct}^f \left(Q, \{Q_i\}_{i=1}^N \right)}{\sum_{Q \subset \mathbb{Q}, N >= k} \text{Correct}(Q)}, \quad (7)$$

where $\mathbb{Q}$ denotes the set of compositional questions. $\text{Correct}(\cdot)$ and $\text{Correct}_f(\cdot, \cdot)$ are two indicator functions. For a question q, we set $\text{Correct}(q)$ as 1 if it is correctly answered by a model and 0 otherwise. $\text{Correct}^f \left(Q, \{Q_i\}_{i=1}^N \right)$ indicates whether a compositional question Q and all its sub-questions $\{Q_i\}_{i=1}^N$ are correctly answered. We set it as 1 if Q and $\{Q_i\}_{i=1}^N$ are correctly answered and 0 otherwise. Obviously, the value of $RC(k)$ is in the range of $[0, 1]$. A model with higher reasoning consistency score is more consistent than a model with a lower score. This definition disentangles the accuracy and the reasoning consistency of compositional VQA. A model with lower accuracy may still have a higher reasoning consistency score if most correctly answered compositional questions are consistently correctly answered.

5. Experiments

We evaluate the proposed method on the GQA dataset [11] and our GQA-Sub dataset. In this section, we first analyze the reasoning consistency of the proposed method. Then we present ablation studies to investigate the effectiveness of several critical components in it. Further, we compare the accuracy of the proposed method with state-of-the-art compositional VQA methods. Finally, qualitative results of the proposed method are shown.

5.1. Reasoning Consistency Performance

We first evaluate three state-of-the-art compositional VQA models: MAC [10], LCGN [9], and MMN [4]. The MAC and LCGN are typical holistic methods for compositional VQA while the MMN is a modular method. The results of these methods and the proposed method are listed in Tab. 1, where "Acc" is the accuracy on the validation split of the GQA dataset, "Acc (sub)" is the accuracy on the validation-sub split of the GQA-Sub dataset, and "RC(k)" is the reasoning consistency scores with different k. For more fair comparisons, we trained the three methods in a

Methods	Acc	Acc (sub)	RC(1)	RC(2)	RC(3)
Language-only	43.86	41.16	31.60	18.81	7.31
Visual-only	56.63	53.97	46.63	28.16	16.26
MAC [10]	62.08	62.63	56.10	41.67	33.96
MAC + DA	61.04	71.42	65.78	54.09	43.14
LCGN [9]	64.16	63.74	57.37	44.32	35.09
LCGN + DA	64.14	73.46	68.93	58.94	50.05
MMN [4]	65.05	64.46	58.79	43.98	33.96
MMN + DA	65.65	74.60	69.59	57.98	48.17
Ours	**66.26**	**76.02**	**71.47**	**61.94**	**52.80**

Table 1. Results of our method and the state-of-the-art. The "DA" means the data augmentation.

data augmentation manner by using both the train split of the GQA and the train-sub of the GQA-Sub for training. We also evaluate the performance of a visual-only model and a question-only model. The former model only takes the local representations of nodes as input for answer prediction while the latter only uses the question representations. Since the original MMN uses the validation split of the GQA dataset, we retrain an MMN model by removing this split and remaining other settings to evaluate its consistency. We didn't evaluate the transformer-based method because these methods [14, 16, 25] use the validation set of the GQA dataset in the pre-training stage.

It is shown from Tab. 1 that the state-of-the-art reasoning models suffer from inconsistency. These methods are capable of answering compositional questions with at least one sub-questions. But among these corrected answered compositional questions, only 60% are consistently answered. As the k grows, the reasoning consistency score of all methods gradually declines. The language-only model and the visual-only model perform worst in all the cases, which is in line with our intuition. The proposed method outperforms the other methods in terms of reasoning consistency in all the cases. Note that our method also achieves superior performance compared with the state-of-the-art trained with data augmentation. There are two reasons for this. Firstly, the proposed method integrates the reasoning processes for the sub-questions into the reasoning process for compositional questions to encourage the model to exploit the correlations of the two kinds of questions. Secondly, the consistency constraint directly penalizes inconsistent answer predictions of models.

5.2. Ablation Studies

To investigate the effectiveness of several important components of our method, we train different variants of our model by ablating certain components. The results of those models are shown in Tab. 2.

We first evaluate the effectiveness of dialog-like reasoning. We train a graph reasoning model that only performs language-guided graph convolution. The model is named "Graph". The comparisons between the "Graph" and "Ours" demonstrate that dialog-like reasoning can sig-

nificantly improve reasoning consistency. We also observe that dialog-like reasoning is beneficial to the accuracy for the compositional questions. The possible reason is that the answering processes for the sub-questions can provide intermediate supervision for the answering of the compositional question. To evaluate the effectiveness of the consistency constraint, we remove the constraint of our method and obtain a model called "Ours (w/o CC)". The model is also inferior to our full model. To further investigate the effect of the order of sub-questions, we've trained a model by randomly setting the required iterations of graph convolution. The comparisons between the obtained model "Ours (rand)" and our full model show the positive effect of the order of sub-questions.

Then we compare our method with a graph reasoning model trained via data-augmentation and named "Graph + DA", since the dialog-like reasoning uses the sub-questions for training. We found that the simple data augmentation strategy can improve the accuracy for the sub-questions but doesn't benefit the accuracy for the compositional questions considerably. The comparison between the "Graph + DA" and "Ours" demonstrates that the proposed method indeed improves the accuracy and consistency. We observe our full model significantly outperforms the model for compositional questions with at least two questions.

5.3. Compositional Reasoning Performance

In this part, we compare the accuracy of the proposed method on the test split of the GQA dataset with the state-of-the-art. In the implementation, we follow [4, 25] and first train the model with the "all" split of questions of the GQA dataset and fine-tune the model with the balanced split of the GQA. The results are shown in Tab. 3. We obverse that the proposed method achieves competitive results compared with the state-of-the-art methods. The reason why we do not surpass all methods is that we mainly focus on maintaining reasoning consistency in compositional VQA. We

Methods	Acc	RC(1)	RC(2)	RC(3)
Graph	65.29	57.08	41.59	33.62
Graph + DA	65.58	70.25	61.70	51.62
Ours (w/o CC)	66.08	71.35	61.26	51.64
Ours (rand)	65.71	70.78	60.52	50.25
Ours	**66.26**	**71.47**	**61.94**	**52.80**

Table 2. Results of different variants of our model. The "DA", "DR", and "CC" means the data augmentation, the dialog-like reasoning and the consistency constraint, respectively.

Methods	Required Inputs	Acc
Bottom-Up [1]	V+L	49.74
MAC [10]	V+L	54.06
NMN [2]	V+L+Program	55.70
BAN [15]	V+L	57.10
GRN [6]	V+L	57.04
LCGN [9]	V+L	57.07
RPR [12]	V+L	59.43
LXMERT [25]	V+L	60.33
12-in-1 [16]	V+L	60.65
MMN [4]	V+L+Program	60.83
MDETR [14]	V+L	61.99
Ours	V+L	59.58

Table 3. Results of our method and the state-of-the-art on the test split of the GQA dataset.

Input image | **Sub-Q**: Is any backpack observable in this photo? (**GT**: yes) | **Sub-Q**: Is there a woman to the right of the backpack? (**GT**: yes) | **Sub-Q**: Is the woman to the right or to the left of the backpack? (**GT**: right) | **Q**: Is the woman to the right of the backpack holding a racket? (**GT**: no)

Iteration:1, Pred: yes — Iteration:2, Pred: yes — Iteration:2, Pred: right — Iteration:3, Pred: no

(a)

Input image | **Sub-Q**: What is the animal called in the picture? (**GT**: elephant) | **Sub-Q**: Do you see a man that is to the left of the woman? (**GT**: no) | **Sub-Q**: Is the man to the right or to the left of the woman? (**GT**: right) | **Q**: Is the elephant to the left or to the right of the man that is to the right of the woman? (**GT**: right)

Iteration:1, Pred: elephant — Iteration:2, Pred: no — Iteration:2, Pred: right — Iteration:3, Pred: right

(b)

Figure 5. Qualitative examples showing the reasoning consistency of our method. For a 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.

only use simple graph convolutions to learn contextual visual representations for compositional reasoning and didn't explore more sophisticated architecture or training strategies to further improve the question-answering ability. For example, transformer-based methods such as [14, 16, 25] use large corpus vision-and-language dataset in pre-training to learn multi-modality representations. The MMN [17] uses annotated structural scene graphs as the supervision for modules at the training stage.

5.4. Quantitative Results

We provide qualitative examples in Fig. 5 to show the effectiveness of our method. For a compositional question and its sub-questions about an input image, we provide visual attention maps, numbers of required iterations, and predicted answers of our method. The compositional question in the first example is about whether the woman determined by a backpack holds a racket. We observe that our method focuses on the backpack to answer the first sub-question about it. Then our method focuses on the woman and the backpack simultaneously to answer the following two sub-questions. Finally, our method figures out there is no racket held by the woman and answers "no". Throughout the reasoning process, the method attends to critical object(s). As a result, our method answers the compositional questions and their sub-questions consistently. We believe that the attention maps and predicted answers for the sub-questions can serve as explanations for humans to be more convinced that the model relies on compositional reasoning to predict the answer rather than dataset biases. For more quantitative

results, please refer to the supplementary material.

6. Conclusion and Future Work

In this paper, we have presented a dialog-like reasoning method to maintain reasoning consistency in compositional VQA. We integrate the reasoning processes for the sub-questions into the reasoning process for a compositional question like a dialog task. Based on the GQA dataset, we constructed a GQA-Sub dataset, which enables the quantitative evaluation of reasoning consistency for compositional VQA models. Experimental results show that our method can answer a compositional question and its sub-questions consistently and accurately.

In our implementation, the number of required iterations for each question is determined by using an off-the-shelf POS tagging method. The tagging method is not specifically designed for our task and may introduce noises. In the future, we will devise a policy network for the reasoning model to determine whether the current contextual representations are sufficient to answer a specific question. Thus we can use reinforcement learning to enable the model to learn to integrate reasoning processes for the sub-questions into the reasoning process for a compositional question in an end-to-end manner for better performance.

Acknowledgments

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6077–6086, 2018. 3, 4, 7

[2] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 39–48, 2016. 2, 7

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2425–2433, 2015. 1, 2

[4] Wenhu Chen, Zhe Gan, Linjie Li, Yu Cheng, William Wang, and Jingjing Liu. Meta module network for compositional visual reasoning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 655–664, 2021. 1, 2, 4, 6, 7

[5] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6904–6913, 2017. 2

[6] Dalu Guo, Chang Xu, and Dacheng Tao. Graph reasoning networks for visual question answering. *arXiv preprint arXiv:1907.09815*, 2019. 7

[7] Matthew Honnibal and Ines Montani. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. 2017. 3

[8] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. Learning to reason: End-to-end module networks for visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 804–813, 2017. 2

[9] Ronghang Hu, Anna Rohrbach, Trevor Darrell, and Kate Saenko. Language-conditioned graph networks for relational reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10294–10303, 2019. 1, 2, 3, 6, 7

[10] Drew A Hudson and Christopher D Manning. Compositional attention networks for machine reasoning. In *International Conference on Learning Representations (ICLR)*, 2018. 2, 6, 7

[11] 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. 2, 4, 6

[12] Chenchen Jing, Yunde Jia, Yuwei Wu, Chuanhao Li, and Qi Wu. Learning the dynamics of visual relational reasoning via reinforced path routing. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2022. 2, 7

[13] Chenchen Jing, Yuwei Wu, Xiaoxun Zhang, Jia Yunde, and Qi Wu. Overcoming language priors in vqa via decomposed linguistic representations. In *Thirty-Forth AAAI Conference on Artificial Intelligence (AAAI)*, pages 11181–11188, 2020. 2

[14] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1780–1790, 2021. 7, 8

[15] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear Attention Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1571–1581, 2018. 7

[16] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10437–10446, 2020. 2, 7, 8

[17] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *International Conference on Learning Representations (ICLR)*, 2018. 8

[18] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 4

[19] Arijit Ray, Karan Sikka, Ajay Divakaran, Stefan Lee, and Giedrius Burachas. Sunny and dark outside?! improving answer consistency in vqa through entailed question generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5863–5868, 2019. 2

[20] Marco Tulio Ribeiro, Carlos Guestrin, and Sameer Singh. Are red roses red? evaluating consistency of question-answering models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6174–6184, 2019. 2

[21] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997. 3

[22] Ramprasaath R Selvaraju, Purva Tendulkar, Devi Parikh, Eric Horvitz, Marco Tulio Ribeiro, Besmira Nushi, and Ece Kamar. Squinting at vqa models: Introspecting vqa models with sub-questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10003–10011, 2020. 2

[23] Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. Cycle-consistency for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6649–6658, 2019. 2

[24] Jiaxin Shi, Hanwang Zhang, and Juanzi Li. Explainable and explicit visual reasoning over scene graphs. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[25] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5100–5111. Association for Computational Linguistics, 2019. 7, 8

[26] Spencer Whitehead, Hui Wu, Yi Ren Fung, Heng Ji, Rogerio Feris, and Kate Saenko. Learning from lexical perturbations for consistent visual question answering. *arXiv preprint arXiv:2011.13406*, 2020. 2

[27] Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding*, 163:21–40, 2017. 1

[28] Yuanyuan Yuan, Shuai Wang, Mingyue Jiang, and Tsong Yueh Chen. Perception matters: Detecting perception failures of vqa models using metamorphic testing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16908–16917, 2021. 2

[29] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5579–5588, 2021. 2