Core-to-Global Reasoning for Compositional Visual Question Answering

Hao Zhou[1], **Tingjin Luo**[2*], **Zhangqi Jiang**[2]

[1]Department of Operational Research and Planning, Naval University of Engineering, Wuhan, Hubei, China
[2]College of Science, National University of Defense Technology, Changsha, Hunan, China
{zhouhao3075, tingjinluo}@hotmail.com, jiangzq@nudt.edu.cn

Abstract

Compositional visual question answering (Compositional VQA) needs to provide an answer to a compositional question, which requires the model to have advanced capabilities of multi-modal semantic understanding and logical reasoning. However, current VQA models mainly concentrate on enriching the visual representations of images and neglect the redundancy in the enriched information to bring some negative impacts. To enhance the value and availability of semantic features, we propose a novel core-to-global reasoning (CTGR) model for compositional VQA. The model first extracts both global features and core features from image and question through a feature embedding module. Then, to enhance the value of semantic features, we propose an information filtering module to align visual features and text features at the core semantic level and to filter out the redundancy carried by image and question features at the global semantic level, which can further strengthen cross-modal correlations. Besides, we design a novel core-to-global reasoning mechanism for multimodal fusion, which integrates content features from core learning and context features from global features for accurate answer predictions. Finally, extensive experimental results on GQA, GQA-sub, VQA2.0 and Visual7W demonstrate the effectiveness and superiority of CTGR.

Introduction

Compositional Visual Question Answering (Compositional VQA) refers to answering a compositional question based on image content, which contains various visual concepts such as objects, attributes, and relationships (Antol et al. 2015). As shown in Fig.1, the key to answering the compositional question is not only to understand the color of the curtain as white, but also to understand the spatial relationship between the chair and the curtain (Zerroug et al. 2022; Shen, Inoue, and Shinoda 2024). Compared to general VQA (Wu et al. 2017; Schwenk et al. 2022), compositional VQA requires models that can deeply mine multiple visual concepts and generate accurate answers for a challenging question. Due to the advantage of spatial and semantic understanding, the compositional VQA can be used in various fields, such as intelligent assistants, smart homes, and automatic driving (Barra et al. 2021; Andreas et al. 2016).

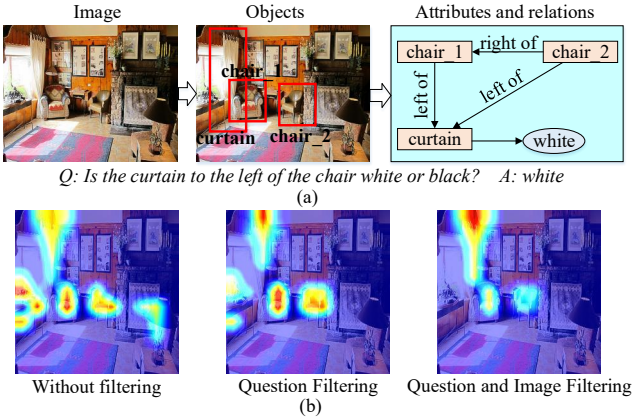


Figure 1: (a) An example of compositional VQA. (b) The visualization of the model's attention with different semantic values by filtering image and question features.

In literature, compositional VQA methods fall into two categories(Jing et al. 2022b): modular-based and holistic-based models. Modular-based methods (Chen et al. 2021; Shi, Zhang, and Li 2019) divide the VQA model into distinct modules according to the input question, and construct a dynamic neural modular network (NMN) with the semantic symbols for reasoning. For example, Andreas *et al.* (Andreas et al. 2016) decomposed the input question into linguistic substructures, such as objects, relations, attributes, etc., and then dynamically instantiated modular networks for reasoning based on these substructures. Most of them focus on the generalization and explainability of visual reasoning based on NMN (Xue, Qian, and Xu 2023). Holistic-based models (Jing et al. 2022a, 2020; Zhang et al. 2021) use a single model, such as graph neural network (GNN), for different inputs to learn multiple semantic features to achieve inference. For example, Hu *et al.* (Hu et al. 2019) constructed a GNN with objects in the scene as nodes and learned contextual features for inference by message passing.

However, most VQA models focus on enriching image representations without checking whether all enriched features are necessary to reason about the answer or not (Nguyen et al. 2022). Except for the core predicate semantics (e.g., object, relations, and attributes in images and questions), the enriched features still contain a large amount of

*Corresponding author

redundant information, which weakens the semantic value of image features and exacerbates the difficulty of multimodal semantic alignment. Figure 1(b) shows the visualization of the model's attention with different semantic values by filtering image and text features. We can find that even with enriched representations of image and question, the model is hard to accurately locate the visual regions related with the answer due to the interference of redundant information, therefore increasing inference difficulty and uncertainty (as shown in the left of Fig.1(b)). By removing redundant information from text or image features, the model more easily establishes the relations between core words and visual regions, improving the accuracy and reliability of model prediction (as shown in the middle and right of Fig.1(b)). To refine the semantic features of images, Nguyen *et al.* (Nguyen et al. 2022) proposed a framework based on the Faster RCNN network to extract some key semantic features like objects, relationships, and attributes in images, and constructed a coarse-to-fine reasoning, effectively improving the compositional VQA performance. However, the semantic features they extracted still contain some noises, which limits the model's reasoning ability. Actually, both questions and images contain considerable redundant information in compositional VQA, which becomes one of the causes of the semantic gaps in multimodal understanding (Xue, Qian, and Xu 2023; Wang et al. 2023).

To tackle the above challenges, we propose a novel Core-To-Global Reasoning (CTGR) framework for compositional VQA. Our CTGR mines core semantic features to enrich image and question representations, aligns and filters semantic features to reduce redundancy, and achieves core-to-global reasoning for efficient multimodal semantic fusion. Specifically, to obtain core semantic features, we first extract question predicates through a stop-word filter and extract object, attribute, and relationship features for images through an improved scene graph generation (SGG) framework. Then, to enhance the semantic value of core features, we design an information filtering module for feature refinement. The visual core features will be aligned with the text core features via a semantic alignment mechanism. And the unnecessary information will be filtered out from global features of questions and images. Besides, we propose a novel core-to-global learning to extract content and context features for multimodal information fusion and prediction. Finally, our model achieves 0.80%, 2.44%, 0.80% and 0.40% performance improvement on the GQA, GQA-sub, VQA2.0 and Visual7W datasets, respectively. Our main contributions can be summarized as follows:

- We introduce an effective framework to extract hierarchical semantics from question and image, including global and core features, which provide rich representations for model learning.
- We design an information filtering mechanism to enhance the semantic value of features, which promotes core semantic alignment between different modalities and removes redundant information from global features.
- We propose a core-to-global learning approach to fuse multimodal information, which learns content information from the core level and context information from the global level for answer predictions.

Related Works

VQA. It is crucial for VQA to extract meaningful features from images and texts for multi-modal fusion. Grid features(Jiang et al. 2020), BUTD features(Anderson et al. 2018), and object features (Nguyen et al. 2019) are usually used to extract image representation in VQA models. BERT (Devlin et al. 2018) and Glove (Pennington, Socher, and Manning 2014) are used to convert words or sentences in questions into vector features. Recently, Chen *et al.* (Chen et al. 2020) takes the large-scale pre-training models for image-text representation. For multimodal fusion, the attention mechanism is widely applied in various VQA models to obtain joint embeddings of images and questions. For examples, Kim *et al.* (Kim, Jun, and Zhang 2018) proposed the bilinear attention mechanism to fully utilize the visual-linguistic information and learn their interactions. Besides, some VQA models are dedicated to exploring the relations between visual regions and words in questions, such as through message passing, pairwise relationship modeling (Cadene et al. 2019), adversarial learning (Li et al. 2021).

Compositional VQA. Compositional VQA needs to generate an answer for a given compositional question based on the image content(Chen et al. 2021; Andreas et al. 2016). The compositional VQA models can be divided into two categories: holistic-based (Hu et al. 2019; Jing et al. 2022a, 2020; Zhang et al. 2021; Yang et al. 2020; Yu et al. 2019) and modular-based models (Andreas et al. 2016; Chen et al. 2021; Shi, Zhang, and Li 2019). The holistic-based models tend to build a unified multimodal fusion framework, often employing the graph neural network or attention mechanism for training and inference. Hudson *et al.* (Hudson and Manning 2019a) proposed an NSM model, which performed sequential inference on a probabilistic scene graph and achieved multi-hop reasoning for predictions. The modular-based models consist of various modules, which mainly reflect the reasoning structure implicit in the question. Andreas *et al.* (Andreas et al. 2016) decomposed the questions into language substructures and dynamically generated modular networks for inference. On this basis, Chen *et al.* (Chen et al. 2021) proposed an IoU-based Kullback-Leible divergence model to provide additional supervisory information for keeping the reasoning on track. Besides, the consistency in compositional VQA attracted the attention of some researchers(Ribeiro, Guestrin, and Singh 2019; Ray et al. 2019). Jing *et al.* (Jing et al. 2022b) generated a series of sub-questions for each question on GQA dataset to test the model's reasoning ability for compositional questions.

Method

Model Overview

Figure 2 illustrates an overview of the proposed core-to-global reasoning framework for compositional VQA. The framework takes a question and an image as inputs, and consists of three modules: feature embedding module, information filtering module, and multimodal fusion and inference

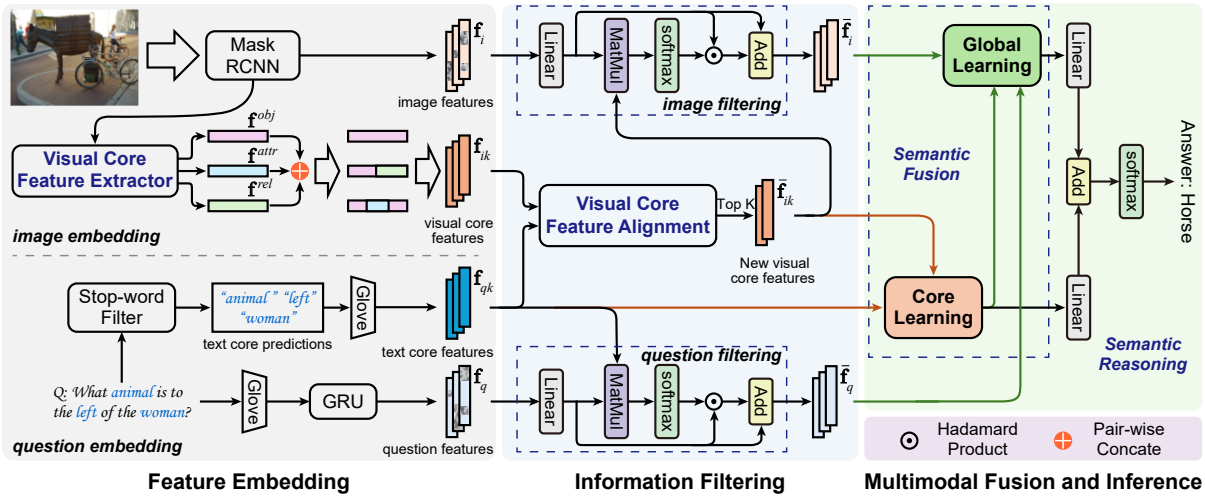


Figure 2: The framework of CTGR model. It consists of three modules: feature embedding module, information fultering module, and multimodal fusion and inference module.

module. In the feature embedding module, image features and visual core features are extracted through an improved SGG framework, while question features and text core features are derived using a question embedding framework. The core features include object, attribute, and relationship features in both question and image. The information filtering module is adopted to enhance the semantic value of visual and text representations, including visual core features alignment, image features filtering, and question features filtering. In the multimodal fusion and inference module, semantics at different levels are fused through core learning and global learning with hierarchical bilinear attention, and answers will be predicted via the core-to-global reasoning.

Feature Embedding Module

The feature embedding module includes image embedding and question embedding. Specifically, the image embedding is to obtain image features and visual core features. Image features are extracted by an object detector with the Mask RCNN (He et al. 2017). To obtain more refined visual core features, we utilize the SGG framework proposed by Tang *et al.* (Tang et al. 2019) to extract object and relational features. Given an image I, the model can extract image RoI features $\mathbf{f}^{roi}$ and object proposal boxes B through Mask RCNN.

$$\mathbf{f}^{roi}, B = RCNN(I), \tag{1}$$

Similar to BUTD, we consider image RoI features as image features, i.e., $\mathbf{f}_i = \mathbf{f}^{roi}$. In the SGG framework, we construct a TreeLSTM network by prim's algorithm (Prim 1957) and adopt BiTreeLSTM (Tang et al. 2019; Zhou et al. 2023) to extract object features and relations features.

$$\mathbf{f}^{obj} = BiTreeLSTM(\mathbf{f}^{roi}), \tag{2}$$

$$\mathbf{f}^{rel} = BiTreeLSTM(\mathbf{f}^{obj}). \tag{3}$$

To extract object attribute features, we incorporate an additional attribute branch into the SGG framework. Taking the refined RoI features $\bar{\mathbf{f}}^{roi}$ from BiTreeLSTM as inputs, attribute branch learns object attribute features and predicts attribute categories via cross entropy loss. The object attribute features can be expressed as:

$$\mathbf{f}^{attr} = MLP(\bar{\mathbf{f}}^{roi}, B). \tag{4}$$

We take three learnable linear projection functions to project the pairs of object, object-attribute, and object-relations-object into the same dimension. And the object, attribute, and relationship features in the image will be combined into a matrix and form the visual core features $\mathbf{f_{ik}}$.

The question embedding is to obtain question features and text core features. Specifically, given a question consisting of n words $Q = \{q_1, q_2, ..., q_n\}$, word embedding is applied to map them to feature vectors $x = \{x_1, x_2, ..., x_n\}$. The embedding vectors x_i are then sequentially fed into the GRU unit (Cho et al. 2014) to obtain question features $\mathbf{f_q}$. Following CFR(Nguyen et al. 2022), text core predicates are extracted through a stop-word filter, which consists of two lists. The first list is to remove meaningless words based on stop-words list in the NLTK. The second list includes rare words whose frequency is less than 10 from all questions. The text core predicates can be obtained via the stop word filter. We applied 300-dim Glove word embedding to extract text core features $\mathbf{f}_{qk}$ from text core predicates.

Information Filtering Module

Enriched image and question features often contain a lot of redundant information in VQA task. There are usually fewer objects and regions in images associated with text questions, while irrelevant information increases the difficulty of model locating effective regions for prediction. Therefore, we design an information filtering module to reduce unnecessary information and enhance the semantic value of images and question features. The information filtering module includes visual core features alignment, image features filtering, and question features filtering.

Visual core features alignment. The detailed structure of visual core features alignment is shown in Fig.3. We take the

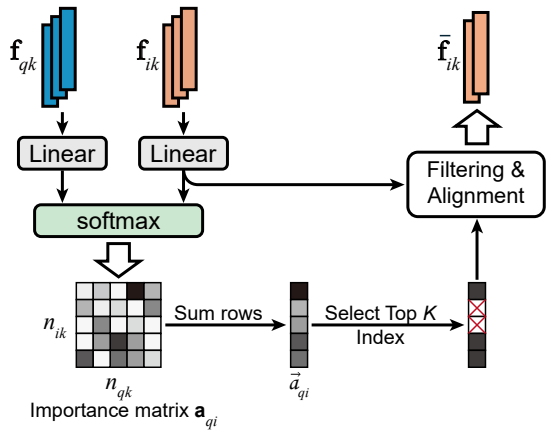


Figure 3: The structure of visual core features alignment.

text core features as the supervisory information to calculate the importance of all semantic pairs in visual core features, and then select the semantic pairs that are closely related to the questions as new visual core features. Specifically, given visual core features $\mathbf{f}_{ik} \in \mathbb{R}^{\mathrm{n}_{ik} \times \mathrm{d}_{ik}}$ and text core features $\mathbf{f}_{qk} \in \mathbb{R}^{\mathrm{n}_{qk} \times \mathrm{d}_{qk}}$, where n_{ik} and n_{qk} represent the number of feature instances, and d_{ik} and d_{qk} represent the dimension of feature instances. The importance of semantic pairs related to text core features can be computed as:

$$\mathbf{a}_{qi}^{T} = softmax(\tau_{qk}(\mathbf{f}_{qk})\tau_{ik}(\mathbf{f}_{ik})^{T}), \tag{5}$$

where $\mathbf{a}_{qi}^{T} \in \mathbb{R}^{\mathrm{n}_{qk} \times \mathrm{n}_{ik}}$, $\tau_{qk}()$ and $\tau_{ik}()$ are learnable linear projection functions. They can project $\mathbf{f}_{ik} \in \mathbb{R}^{\mathrm{n}_{ik} \times \mathrm{d}_{ik}}$ and $\mathbf{f}_{qk} \in \mathbb{R}^{\mathrm{n}_{qk} \times \mathrm{d}_{qk}}$ into $\mathbf{f'}_{ik} \in \mathbb{R}^{\mathrm{n}_{ik} \times \mathrm{d}_{\tau}}$ and $\mathbf{f'}_{qk} \in \mathbb{R}^{\mathrm{n}_{qk} \times \mathrm{d}_{\tau}}$. The importance vector $\vec{a}_{qi}$ can be obtained by summing the importance matrix $\mathbf{a}_{qi}$ by rows, whose value represents the importance of each semantic pairs in the visual core features.

By sorting $\vec{a}_{qi}$ by descending, we select Top K semantic pairs to form a new visual core features $\bar{\mathbf{f}}_{ik} \in \mathbb{R}^{\mathrm{n}_K \times \mathrm{d}_{\tau}}$ and filter out unimportant instances. The value of K is:

$$K = \lambda \times \mathrm{mean}(d_{qk}), \tag{6}$$

where λ is a hyper-parameter and $\mathrm{mean}(d_{qk})$ denotes the average of text core predicates extracted from all questions.

Features filtering of image and question. To reduce the redundancy and enhance the semantic value of image and question features, we take the visual core features and text core features as the supervision to refine image features and question features, respectively. For visual core features $\bar{\mathbf{f}}_{ik}$ and image features $\mathbf{f}_i$, the weighting map ψ_i is computed as:

$$\psi_i = \mathrm{softmax}(\sum_{p=0}^{K} \tau_i(\mathbf{f}_i)(\bar{\mathbf{f}}_{ik})_p^T), \tag{7}$$

where $\tau_i()$ is learnable linear projection functions, which projects $\mathbf{f}_i \in \mathbb{R}^{\mathrm{n}_i \times \mathrm{d}_i}$ into $\mathbf{f'}_i \in \mathbb{R}^{\mathrm{n}_i \times \mathrm{d}_{\tau}}$. The refined image features can be obtained by:

$$\bar{\mathbf{f}}_i = (\psi_i \cdot \mathbf{1}^T) \odot \mathbf{f'}_i + \mathbf{f'}_i, \tag{8}$$

where $\mathbf{1} \in \mathbb{R}^{d_{\tau}}$ is a channel-scaled vector, $\odot$ denotes the Hardamard product. Similarly, taken the text core features $\mathbf{f'}_{qk}$ as supervision information, we can compute the weighting map ψ_q and refined question features:

$$\psi_q = softmax(\sum_{p=0}^{n_{ik}} \tau_q(\mathbf{f}_q)(\mathbf{f'}_{qk})_p^T), \tag{9}$$

$$\bar{\mathbf{f}}_q = (\psi_q \cdot \mathbf{1}^T) \odot \mathbf{f'}_q + \mathbf{f'}_q, \tag{10}$$

where $\tau_q()$ is learnable linear projection functions, which projects $\mathbf{f}_q \in \mathbb{R}^{\mathrm{n}_q \times \mathrm{d}_q}$ into $\mathbf{f'}_i \in \mathbb{R}^{\mathrm{n}_q \times \mathrm{d}_{\tau}}$.

Multimodal Fusion and Inference Module

Effective multimodal fusion can deepen the model's understanding of both image and question information, allowing it to discern cross-modal semantic correlations and improve the accuracy of VQA. In this section, we propose a core-to-global learning for multimodal information fusion and prediction. In core-level learning, visual core features and text core features are fused. The model can learn the interaction between them, which provides essential and content features for answer prediction. In global-level learning, the model will fuse the visual features and question features, which promotes the model's understanding of multimodal semantics and provide context features for accurate answer predictions. We have designed a hierarchical bilinear attention mechanism for core-level learning and global-level learning.

In core-level learning, the bilinear attention mechanism (Kim, Jun, and Zhang 2018) is used to fuse visual core features and text core features. It outputs a joint representation vector $\bar{c}_{lk}$ with d_{lk} dimensions. For m-th element in the joint representation $\bar{c}_{lk}$, the value $\bar{c}_{lk}^{m}$ can be computed by:

$$\bar{c}_{lk}^{m} = (\mathbf{f'}_{qk} W_{qk})_i^T A^{lk} (\bar{\mathbf{f}}_{ik} W_{ik})_m, \tag{11}$$

where $W_{qk} \in \mathbb{R}^{\mathrm{d}_{\tau} \times \mathrm{d}_{lk}}$ and $W_{ik} \in \mathbb{R}^{\mathrm{d}_{\tau} \times \mathrm{d}_{lk}}$ are learnable factor matrices, $A^{lk} \in \mathbb{R}^{\mathrm{n}_{qk} \times \mathrm{n}_K}$ is the bilinear attention distribution map and it can be computed by:

$$A^{lk} = softmax((\mathbf{f'}_{qk} W'_{qk})((\bar{\mathbf{f}}_{ik} W'_{ik})^T), \tag{12}$$

where $W'_{qc} \in \mathbb{R}^{\mathrm{d}_{\tau} \times \mathrm{d}_{lk}}$ and $W'_{ik} \in \mathbb{R}^{\mathrm{d}_{\tau} \times \mathrm{d}_{lk}}$ are learnable parameter matrices.

In global-level learning, the model takes both question features and joint representation vector $\bar{c}_{lk}$ as supervision to fuse image features and question features. The supervision information at global level is: $\bar{\mathbf{f}}_{si} = (\bar{c}_{lk} || \bar{\mathbf{f}}_q W_q)$, where $||$ denotes the concatenation. Similarly, the bilinear attention mechanism is used to fuse image features and question features at global level. It outputs a joint representation vector $\bar{c}_{gc}$ with d_{gc} dimensions. For j-th element in joint representation $\bar{c}_{gc}$, the value $\bar{c}_{gc}^{j}$ can be computed by:

$$\bar{c}_{gc}^{j} = (\bar{\mathbf{f}}_{si} W_{si})_j^T A^{gc} (\bar{\mathbf{f}}_i W_i)_j, \tag{13}$$

where $W_{si} \in \mathbb{R}^{\mathrm{d}_{lk} \times \mathrm{d}_{gc}}$ and $W_i \in r^{d_{\tau} \times \mathrm{d}_{gc}}$ are learnable factor matrices, $A^{gc} \in \mathbb{R}^{(\mathrm{n}_q+1) \times \mathrm{n}_i}$ is the bilinear attention distribution map and it can be computed by:

$$A^{gc} = \mathrm{softmax}((\bar{\mathbf{f}}_{si} W'_{si})((\bar{\mathbf{f}}_i W'_i)^T), \tag{14}$$

where $W'_{si} \in \mathbb{R}^{\mathrm{d}_{lk} \times d_{gc}}$ and $W'_i \in \mathbb{R}^{\mathrm{d}_{\tau} \times \mathrm{d}_{gc}}$ are learnable parameter matrices.

Model	Val acc (%)	Test acc (%)
BUTD	/	49.7
BAN	61.5	55.2
CTI	61.7	54.9
MCAN	/	57.4
RCVQA	/	59.6
MMN	/	60.4
LXMERT	59.8	60.0
OSCAR	/	61.6
CFR	73.6	72.1
CTGR	**75.0**	**72.9**

Table 1: Results on GQA validation and test set.

In semantic reasoning, the model combines $\bar{c}_{lk}$ and $\bar{c}_{gc}$, and selects appropriate information with the core-to-global reasoning for answer prediction. Specifically, taken $\bar{c}_{lk}$ and $\bar{c}_{gc}$ as inputs, the model will output the probability distribution via softmax classifier. The candidate answer with the highest probability will be selected as the prediction:

$$\text{Ans} = \max(\text{softmax}(\tau_{lk}(\bar{c}_{lk}) + \tau_{gc}(\bar{c}_{gc}) + b_{an})), \quad (15)$$

where τ_{lk} and τ_{gc} are learnable linear projection functions. They can convert the dimensions of vectors $\bar{c}_{lk}$ and $\bar{c}_{gc}$ to d_A, respectively, and A is the number of candidate answers.

Finally, through the end-to-end training process, the learned model can learn adaptive weights to remove the noise information in images and question features, align multimodal core semantics, and realize deep fusion of multimodal semantic features via core-to-global reasoning.

Experiments

Experimental Setup

Datasets. We evaluate our CTGR model on four benchmarks: GQA (Hudson and Manning 2019b), GQA-sub (Jing et al. 2022b), VQA2.0 (Goyal et al. 2017), and Visual7W(Zhu et al. 2016), and follow the same split in each dataset for training and testing. GQA is a real-world dataset for visual reasoning and compositional VQA. It contains 113000 images and 22 million novel questions, with a vocabulary of 3097 words in questions and 1878 candidate answers. GQA-sub is a subset of GQA dataset that can quantitatively evaluate the inference consistency in the composition VQA. VQA2.0 contains 204,721 images of real world, with 332,793 questions and 29,332 answers. Visual7W is a subset of the visual genome dataset, containing 47300 images and 327929 QA pairs. The questions are composed of what/here/how/he/how/why with 4 candidate answers.

Implementation details. We trained and tested our model on an NVIDIA RTX 3090TI GPU. The model is trained with a batch size of 32 and an initial learning rate of 0.001 using Adam optimizer. Similar to CFR(Nguyen et al. 2022), we use the Glove to extract embedding vectors for questions. To extract image semantic features, we retrained both the attribute branch and relationship branch on SGG framework with the weight parameters proposed by Tang (Tang et al. 2019). The parameters d_{lk} and d_{gc} are empirically set to 768.

Metrics. We adopt the standard accuracy metric to evaluate our model. For Visual7W dataset with multiple choices, the accuracy is directly used to evaluate model performance. For GQA, GQA-sub, and VQA2.0 dataset, the accuracy is defined as a prediction that is judged to be correct if it matches the answers provided by at least three annotators.

Model	ACC	ACC(sub)	RC(1)	RC(2)	RC(3)
Language-only	43.86	41.16	31.61	18.81	7.31
Visual-only	56.63	53.97	46.63	28.16	16.26
MAC	62.08	62.63	56.1	41.67	33.96
LCGN	64.16	63.74	57.37	44.32	35.09
MMN	65.05	64.46	58.79	43.98	33.96
DIS	69.71	70.31	64.98	53.55	48.86
CTGR	**69.75**	**72.75**	**65.58**	**54.78**	**50.77**

Table 2: Results on GQA-sub validation and test sets.

Experimental Results

Experimental results on GQA. We compare our model with some recent VQA models, including: BUTD(Anderson et al. 2018), BAN(Kim, Jun, and Zhang 2018), CTI(Do et al. 2019), RCVQA(Jing et al. 2022b), MCAN(Yu et al. 2019), MMN(Chen et al. 2021), LXMERT(Tan and Bansal 2019), OSCAR(Li et al. 2020), and CFR (Nguyen et al. 2022). The results on GQA validation set and test set are shown in Tab.1.

From Tab.1, compared to BAN model, the accuracy of CTGR on validation and test sets has improved by 13.5% and 17.7%, respectively. These significant improvements indicate that when facing complex compositional questions, our model can better understand questions and images by extracting objects, relationships, and attributes features, and accurately establish associations between text core predicates and image core regions by multimodal semantic alignment. Compared with the CFR model, our model gets 1.4% and 0.8% improvements on validation set and test set, respectively. Meanwhile, CTGR model also achieves the best performance among all models, which indicates that the core-to-global reasoning promotes the model's understanding of images and questions, and enables the model to accurately locate and generate candidate answers.

Experimental results on GQA-sub. The GQA sub dataset can be used to evaluate the performance on combined VQA and the inference consistency. We compare our model with Language-only, Visual-only, MAC(Hudson and Manning 2018), LCGN(Hu et al. 2019), MMN(Chen et al. 2021), and DIS (Liu et al. 2024). And the results on GQA-sub dataset are shown in Tab.2.

The metrics in Tab.2 reflect the accuracy of model performance (ACC) and reasoning consistency (RC(K)). ACC and ACC (sub) represent the accuracy of validation set on GQA and GQA-sub datasets, respectively. RC(K) represents the consistency score of reasoning with K sub-problems. As shown in Tab.2, CTGR achieves best performance in both accuracy and reasoning consistency. Specifically, compared to MMN model, CTGR gets 6.79%, 10.80%, and 16.81% performance improvements on RC(1), RC(2), and RC(3), respectively. Compared with the SOTA DIS model, CTGR improves the performance of ACC (sub) from 70.31% to 72.75%, and achieves improvements in consistency scores of 0.60%, 1.23%, and 1.91%, respectively. These improve-

Model	Val acc (%)	Test acc (%)
BAN	66.0	70.0
DFAF	66.2	70.2
CTI	66.0	70.1
MCAN	67.2	70.6
ViLT	/	70.9
LXMERT	/	72.4
CFR	69.7	72.5
CTGR	**71.0**	**73.3**

Table 3: Results on VQA 2.0 validation and test sets.

Model	Val acc (%)	Test acc (%)
BAN	65.7	67.5
fPMC	/	66.0
STL	67.5	68.2
CTI	67.0	69.3
CFR	69.8	71.9
CTGR	**70.5**	**72.3**

Table 4: Results on Visual7W validation and test sets.

ments primarily stem from CTGR model's ability to comprehensively understand the core semantics of both images and questions. By employing the core-to-global reasoning, our model can deepen its understanding of questions and images, accurately recommend relevant candidate regions, thereby strengthening its capacity for consistent reasoning.

Results on VQA2.0 and Visual7W. For a fair comparison, we compare our model with some recent VQA models, including: BAN(Kim, Jun, and Zhang 2018), DFAF(Gao et al. 2019), CTI(Do et al. 2019), MCAN(Yu et al. 2019), LXMERT(Tan and Bansal 2019), CFR(Nguyen et al. 2022), ViLT (Kim, Son, and Kim 2021). The results on VQA2.0 dataset are shown in Tab.3.

From Tab.3, we can find that CTGR model gets the best performance on VQA2.0 dataset. Compared to BAN model, CTGR model achieves 5.0% and 3.3% improvements on the VQA 2.0 validation set and test set, respectively. The semantic features extracted in CTGR model, such as objects, relationships, and attribute features, are more refined than with CFR model. And compared with it, the performance of our model has been improved by 1.3% and 0.8% on the validation set and test set, respectively. This result also demonstrates that extracting multiple core semantic features from images can effectively enhance the visual reasoning ability. Besides, we compare our model with several models trained on Visual7W dataset, including: BAN(Kim, Jun, and Zhang 2018), fPMC(Hu, Chao, and Sha 2018), STL(Wang et al. 2018), CTI(Do et al. 2019), and CFR(Nguyen et al. 2022). The results on Visual7W dataset are shown in Tab.4. The table reveals that compared to the latest CTI and CFR (Nguyen et al. 2022) models, CTGR model has shown respective performance improvements of 3.0% and 0.4% on the test set, and also achieved the best performance on Visual7W.

Ablation Study

We conduct the following ablation studies with different variants and parameters to verify the effectiveness of CTGR.

Feature embedding	Information filtering	Multimodal fusion	Val acc
Q+I(BUTD)	/	/	63.3
Coarse2Fine features	/	/	66.7
✓	/	/	68.0
✓	Text filtering	/	68.9
✓	Image filtering	/	70.2
✓	✓	/	70.8
✓	✓	Bilinear attention	72.7
✓	✓	Hierarchical attention	74.1
✓	✓	✓	75.0

Table 5: Ablation studies of proposed modules.

λ	2	4	5	6	7	9
Acc	70.5	72.8	73.9	75.0	74.8	74.0

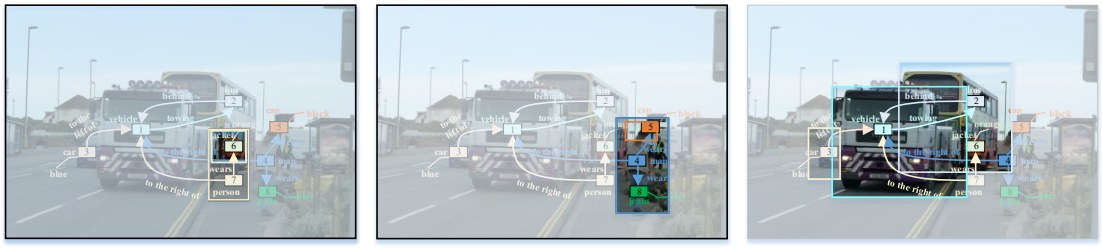
Table 6: The effect of the hyper-parameter λ.

Effectiveness of the proposed modules. To fully explore the roles of the various modules proposed in CTGR model, we design four ablation experiments to verify the influence of three modules and their variants on the model performance. Tab.5 shows the results of these experiments on GQA validation set.

The contribution of each module. In Tab.5, we study the contribution of each module to model performance. When equipped the feature embedding module, the performance of the model increased by 4.7%. This improvement demonstrates that core features can enhance the model's understanding of images and questions, thereby improving the model's performance.When the information filtering module is embedded in the model, its performance improves from 68.0% to 70.8%. This result validates that by filtering redundant information and aligning multimodal semantics in images and questions, the model can pay more attention to useful information directly related to the answers. When equipped with the multimodal fusion and inference module, the model can fuse multimodal semantic features and learn both content features and context features for answers prediction via core-to-global reasoning, its performance is improved by 4.2%. More importantly, the performance improves consistently when three modules are used together.

Study on variants of feature embedding module. To study the impact of feature embedding module, the model takes different features and visual modalities as inputs, including image+question features, Coarse2Fine features extracted by CFR, and the features extracted by CTGR model. The results are shown in the upper row of Tab.5. Both CFR and CTGR extracted the core features of questions and images, but the text core features and visual core features extracted by our model are more refined than those obtained by CFR, and CTGR model gets 1.3% improvements. This result indicates that fine semantic features can help the model to deepen the accurate understanding of the image and text.

Study on variants of information filtering module. As shown in the middle row of Tab.5, we design three variants of information filtering module to study its effectiveness: (1) question features filtering only, (2) image features filtering and refining, and (3) question+image features filtering and



Is the **jacket orange** or **blue**?
Answer: orange
(a)

Who is **wearing** the **cap**?
Answer: man
(b)

Are there **buses** **to the** **left** **of the** **vehicle** **on the left of the picture?**
Answer: no
(c)

Figure 4: The visualization results on the GQA dataset.

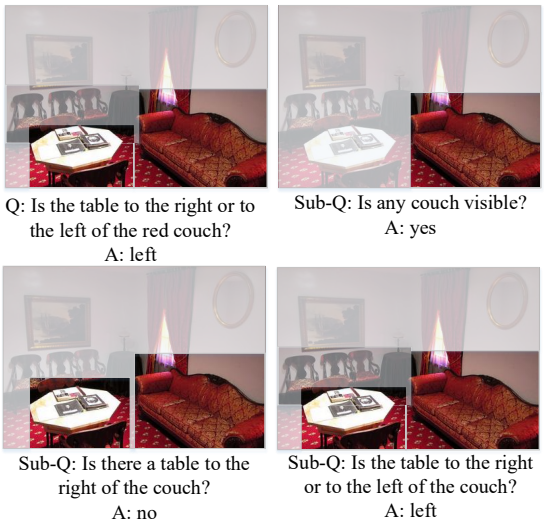


Figure 5: The visualization results on the GQA-sub dataset.

refining. By filtering question feature, the model's performance is improved by 0.9%. By introducing image features filtering and refining, the accuracy increases from 68.0% to 70.2%. When using the proposed information filtering module, the model gets 2.8% improvements and achieves the highest accuracy among the three variants. This result indicates that filtering unnecessary information in the text and image features can enhance the semantic value of the features, thereby mitigating the negative impact of redundancy.

Study on variants of multimodal fusion and inference module. We mainly construct bilinear attention, hierarchical attention, and hierarchical bilinear attention for multimodal features fusion. As shown in the bottom row of Tab.5, hierarchical bilinear attention combines the advantages of bilinear and hierarchical attention, and achieves 2.3% and 0.9% performance improvements, respectively.

The effect of the hyperparameter λ. In this section, we further investigate the influence of different hyperparameter λ. We set λ=2, 4, 5, 6, 7, 9, and the corresponding results are shown in Tab. 6. From Tab. 6, we find that when $\lambda = 6$, our model can get the best performance. When $\lambda > 6$, the improvement of model performance is trivial, and it will expand trainable parameters. Therefore, we recommend λ=6 to determine the value of Top K to filter visual core features.

Visualization

Fig.4 shows the visualization results on the GQA dataset. The larger the grayscale value of words or the higher the transparency of image regions the more important they are in Fig.4. The result demonstrates the effectiveness of CTGR model in inferring the correct answer. As shown in Fig.4(c), a compositional question includes several core words, such as "buses", "vehicle", and "left". The model first aligns the image region of "vehicle" and "bus" with the core word "vehicle" and "buses". And then by integrating the text core word features of "left" with the visual core features (such as "bus-behind-vehicle"), the model can correctly predict the answer as "No". Fig.5 shows the visualization results on the GQA-sub dataset. For the question, our model first focuses on the right region of the picture to answer the first sub-question. Then, it pays attention to the relationship between "couch" and "table" when answering the second and third sub questions. Finally, our model can infer the correct answer as "left" for original question based on the color attribute of "couch" and the relationship between "couch" and "table". In general, our method answers the compositional questions and their sub-questions consistently, which might improve the interpretability of the model.

Conclusion

In this paper, we have introduced a new core-to-global reasoning model for compositional VQA. After extracting global features and core features of image and question, our model achieves the alignment for visual core features and text core features, and filters out the redundant information for global features to enhance the value of semantic features via a feature filtering module. Furthermore, to fuse multimodal semantics, we propose a new reasoning module to learn the content and context features for answer predictions in a core-to-global manner. CTGR model has been validated on four datasets and achieves better performance in both accuracy and consistency inference. The ablation studies also demonstrate the effectiveness of the proposed module. In future work, research on multimodal semantic alignment is worth further exploration.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (No. 62302516 and 62376281).

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and vqa. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Andreas, J.; Rohrbach, M.; Darrell, T.; and Klein, D. 2016. Neural Module Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 39–48.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, 2425–2433.

Barra, S.; Bisogni, C.; De Marsico, M.; and Ricciardi, S. 2021. Visual question answering: Which investigated applications? *Pattern Recognition Letters*, 151: 325–331.

Cadene, R.; Ben-Younes, H.; Cord, M.; and Thome, N. 2019. Murel: Multimodal relational reasoning for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1989–1998.

Chen, W.; Gan, Z.; Li, L.; Cheng, Y.; Wang, W.; and Liu, J. 2021. Meta module network for compositional visual reasoning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 655–664.

Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, 104–120. Springer.

Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Do, T.; Do, T.-T.; Tran, H.; Tjiputra, E.; and Tran, Q. D. 2019. Compact trilinear interaction for visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 392–401.

Gao, P.; Jiang, Z.; You, H.; Lu, P.; Hoi, S. C.; Wang, X.; and Li, H. 2019. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6639–6648.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6904–6913.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, 2961–2969.

Hu, H.; Chao, W.-L.; and Sha, F. 2018. Learning answer embeddings for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5428–5436.

Hu, R.; Rohrbach, A.; Darrell, T.; and Saenko, K. 2019. Language-conditioned graph networks for relational reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10294–10303.

Hudson, D.; and Manning, C. D. 2019a. Learning by abstraction: The neural state machine. *Advances in Neural Information Processing Systems*, 32.

Hudson, D. A.; and Manning, C. D. 2018. Compositional attention networks for machine reasoning. *arXiv preprint arXiv:1803.03067*.

Hudson, D. A.; and Manning, C. D. 2019b. 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*, 6700–6709.

Jiang, H.; Misra, I.; Rohrbach, M.; Learned-Miller, E.; and Chen, X. 2020. In defense of grid features for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10267–10276.

Jing, C.; Jia, Y.; Wu, Y.; Li, C.; and Wu, Q. 2022a. Learning the dynamics of visual relational reasoning via reinforced path routing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 1122–1130.

Jing, C.; Jia, Y.; Wu, Y.; Liu, X.; and Wu, Q. 2022b. Maintaining reasoning consistency in compositional visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5099–5108.

Jing, C.; Wu, Y.; Zhang, X.; Jia, Y.; and Wu, Q. 2020. Overcoming language priors in vqa via decomposed linguistic representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 11181–11188.

Kim, J.-H.; Jun, J.; and Zhang, B.-T. 2018. Bilinear attention networks. *Advances in Neural Information Processing Systems*, 31.

Kim, W.; Son, B.; and Kim, I. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, 5583–5594. PMLR.

Li, L.; Lei, J.; Gan, Z.; and Liu, J. 2021. Adversarial vqa: A new benchmark for evaluating the robustness of vqa models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2042–2051.

Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference of Computer Vision*, 121–137. Springer.

Liu, Y.; Peng, D.; Wei, W.; Fu, Y.; Xie, W.; and Chen, D. 2024. Detection-Based Intermediate Supervision for Visual Question Answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 14061–14068.

Nguyen, A.; Tran, Q. D.; Do, T.-T.; Reid, I.; Caldwell, D. G.; and Tsagarakis, N. G. 2019. Object captioning and retrieval with natural language. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 1–10.

Nguyen, B. X.; Do, T.; Tran, H.; Tjiputra, E.; Tran, Q. D.; and Nguyen, A. 2022. Coarse-to-fine reasoning for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4558–4566.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Prim, R. C. 1957. Shortest connection networks and some generalizations. *The Bell System Technical Journal*, 36(6): 1389–1401.

Ray, A.; Sikka, K.; Divakaran, A.; Lee, S.; and Burachas, G. 2019. Sunny and dark outside?! improving answer consistency in vqa through entailed question generation. *arXiv preprint arXiv:1909.04696*.

Ribeiro, M. T.; Guestrin, C.; and Singh, S. 2019. Are red roses red? evaluating consistency of question-answering models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6174–6184.

Schwenk, D.; Khandelwal, A.; Clark, C.; Marino, K.; and Mottaghi, R. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, 146–162. Springer.

Shen, R.; Inoue, N.; and Shinoda, K. 2024. Pyramid Coder: Hierarchical Code Generator for Compositional Visual Question Answering. *arXiv preprint arXiv:2407.20563*.

Shi, J.; Zhang, H.; and Li, J. 2019. Explainable and explicit visual reasoning over scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8376–8384.

Tan, H.; and Bansal, M. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Tang, K.; Zhang, H.; Wu, B.; Luo, W.; and Liu, W. 2019. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6619–6628.

Wang, Y.; Yasunaga, M.; Ren, H.; Wada, S.; and Leskovec, J. 2023. Vqa-gnn: Reasoning with multimodal knowledge via graph neural networks for visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21582–21592.

Wang, Z.; Liu, X.; Wang, L.; Qiao, Y.; Xie, X.; and Fowlkes, C. 2018. Structured triplet learning with pos-tag guided attention for visual question answering. In *IEEE Winter Conference on Applications of Computer Vision*, 1888–1896. IEEE.

Wu, Q.; Teney, D.; Wang, P.; Shen, C.; Dick, A.; and Van Den Hengel, A. 2017. Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding*, 163: 21–40.

Xue, D.; Qian, S.; and Xu, C. 2023. Variational causal inference network for explanatory visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2515–2525.

Yang, X.; Lin, G.; Lv, F.; and Liu, F. 2020. Trrnet: Tiered relation reasoning for compositional visual question answering. In *European Conference of Computer Vision*, 414–430. Springer.

Yu, Z.; Yu, J.; Cui, Y.; Tao, D.; and Tian, Q. 2019. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6281–6290.

Zerroug, A.; Vaishnav, M.; Colin, J.; Musslick, S.; and Serre, T. 2022. A benchmark for compositional visual reasoning. *Advances in Neural Information Processing Systems*, 35: 29776–29788.

Zhang, P.; Li, X.; Hu, X.; Yang, J.; Zhang, L.; Wang, L.; Choi, Y.; and Gao, J. 2021. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5579–5588.

Zhou, H.; Zhang, J.; Luo, T.; Yang, Y.; and Lei, J. 2023. Debiased Scene Graph Generation for Dual Imbalance Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 4274–4288.

Zhu, Y.; Groth, O.; Bernstein, M.; and Fei-Fei, L. 2016. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4995–5004.