REX: Reasoning-aware and Grounded Explanation (Supplementary Materials)

Shi Chen Qi Zhao
Department of Computer Science and Engineering,
University of Minnesota
{chen4595, qzhao}@umn.edu

This supplementary materials provide additional experimental results and details of the proposed framework. Specifically,

- We carry out an ablation study on the explanation generation.
- We elaborate the detailed templates for constructing partial explanations with atomic operations (Section 2).
- We present the statistics of the proposed GQA-REX dataset, and provide qualitative examples of our reasoning-aware and grounded explanations (Section 3).
- We provide the implementation details of our proposed explanation generation method (Section 4).

1. Ablation Study of Explanation Generation

Our proposed explanation generation model bridges key components across the visual-textual modalities, and simultaneously models multi-modal explanations with an explicit grounding module. In the main paper, we show its advantages over existing explanation generation methods that independently generate explanations of different modalities. To further demonstrate the effectiveness of our model, we conduct two ablation experiments: (1) replacing visual grounding in our explanations with corresponding object names, and (2) optimizing the visual (*i.e.,* attention maps) and textual explanations separately. The first experiment shows a drop of 1.28% VQA accuracy, and the second one reports 65.53% vs 66.16% (our full model) accuracy, 32.4% vs 67.95% (our full model) explanation grounding score, both highlighting the significance of simultaneously modeling multi-modality

2. Templates for Constructing Explanations

Our functional program proposed in the main paper progressively traverses the reasoning process and uses predefined templates to construct partial explanation at each reasoning step. In this section, we present the details of the templates. Our templates are designed based on the semantic meaning of each atomic operation, and take into account both information extracted in the current reasoning step and that passed from previous steps. As shown in Table 1, we define three general functions shared across different templates: [OBJ] selects a specific type of visually grounded objects, [ATTR] finds desired attributes specified in the atomic operation, and [DEP] collects partial explanations from dependent nodes in the previous steps. Other functions are more specific to a single atomic operation: [CHECK_EXISTENCE] examines if a certain type of objects exist in the scene; [QUERY_ATTR] queries the value of a specific type of attributes; [VERIFY_ATTR] examines if the selected objects have certain attributes; [FIND_COMMON] finds the commons attributes shared by both groups of objects; [COMPARE_ATTR] compares two groups of objects based on a specific type of attributes; [RELATION] finds the desired relationships between two groups of objects; [LOGICAL AND/OR] denotes logical operations.

Operation	Template
Select	[OBJ]
Exist	There [CHECK_EXISTENCE] [DEP]
Filter	[ATTR] [OBJ]
Query	[DEP] is/are [QUERY_ATTR]
Verify	[DEP] is/are [VERIFY_ATTR]
Common	both [DEP 1] and [DEP 2] are [FIND_COMMON]
Same	[DEP 1][DEP 2] are [ATTR] / [DEP1] is [ATTR1] and [DEP 2] is [ATTR 2]
Different	[DEP1] is [ATTR1] and [DEP 2] is [ATTR 2] / [DEP 1][DEP 2] are [ATTR]
Compare	[DEP 1] is [COMPARE_ATTR] than [DEP 2]
Relate	[DEP 1] [RELATION] [DEP 2]
And/Or	[DEP 1] [LOGICAL AND/OR] [DEP 2]

Table 1. Templates for constructing partial explanation with atomic operations.

With the aforementioned templates, we sequentially update the explanation by selectively attending to different regions of interest, investigating the desired attributes, and ac-

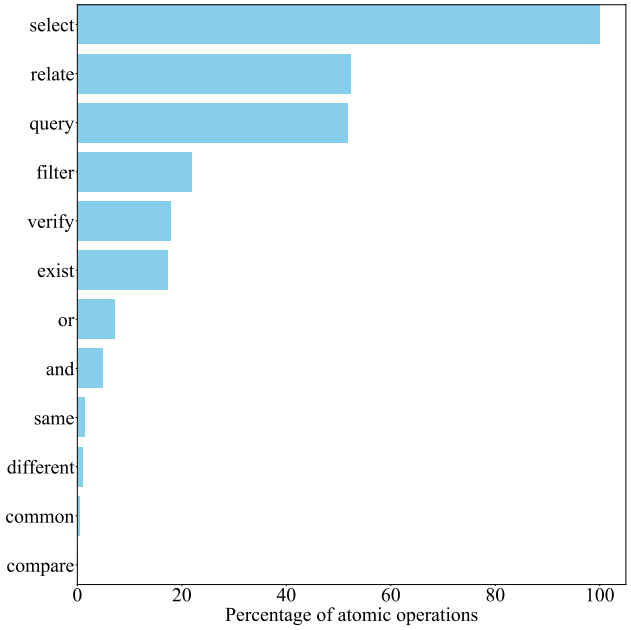


Figure 1. Percentage of questions with different types of reasoning operations.

cumulating information along the reasoning-process. The templates not only enable the collection of our new GQA-REX dataset, but also provide a general paradigm for automatically constructing explanations based on the reasoning process.

3. The GQA-REX Dataset

Aiming to provide an explanation benchmark that encodes the reasoning process and grounding across the visual-textual modalities, we propose a new GQA-REX dataset that consists of 1,040,830 reasoning-aware and visually grounding explanations. In this section, we present the statistics of the our dataset, including the distribution of atomic operations and the distribution of visually grounding objects. We also provide qualitative examples of our defined explanations.

As shown in Figure 1, our dataset covers the explanations for a wide range of visual questions: All of the questions require attending to specific regions of interest (*i.e.,* the *select* operation) to derive the answers, which highlights the need to explain decisions with visual grounding. A large proportion of questions involve recognizing certain attributes (*i.e., relate*, *query*, *filter*, and *verify*), which is one of the fundamental skills for understanding the visual world. Some questions require examining the existence of certain types of objects or performing logical operations, which correspond to the considerable amount of yes/no questions. There are also relatively difficult questions that ask models to investigate all attributes of two groups of objects (*i.e.,* *same*, *different*, *common*, and *compare*).

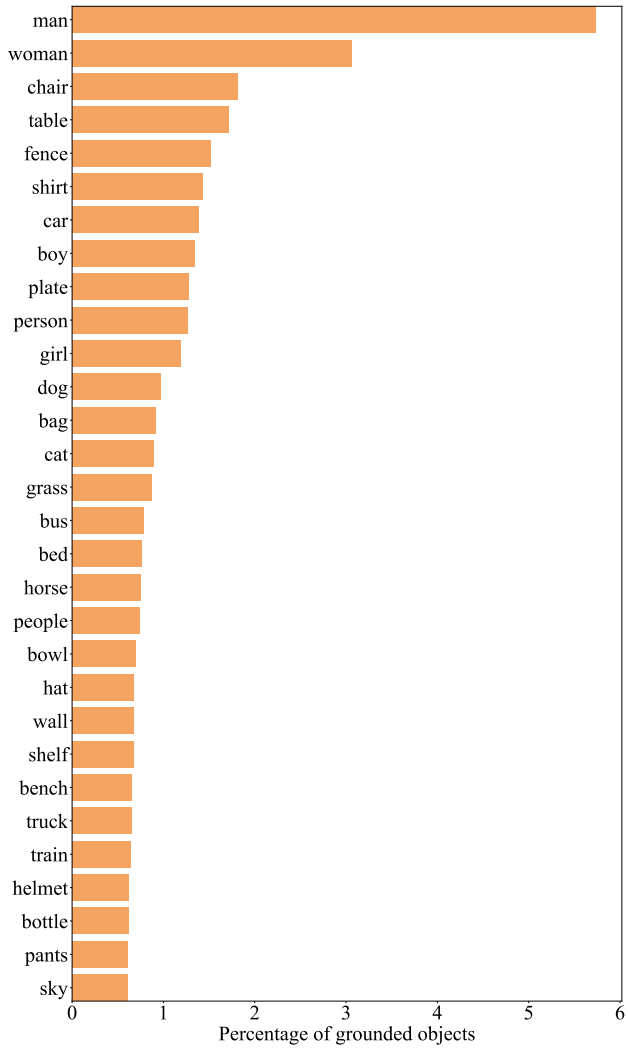


Figure 2. Top-20 object categories for visual grounding and their percentages in our dataset.

To explain the decision making for various questions with multi-modal evidence, we link a diverse collection of objects with their corresponding regions of interest. As shown in Figure 2, unlike [4] that focuses on grounding a single type of objects (*i.e.,* humans), our dataset takes into account 1,660 unique types of object categories and provides more fine-grained categorization (*e.g.,* human characters are categorized based on their genders, *i.e.,* woman and man, and ages, *i.e.,* boy and man). The visual grounding plays an essential role in explaining how different components in the visual-textual modalities contribute to the decision-making, and enabling the development of computational models with improved multi-modal understanding of the reasoning process (*e.g.,* VisualBert-REX in the main paper).

In Figure 3, we visualize examples of our explanations for different types of questions, *e.g.*, questions examining the existence of certain object in the 1^{st} row, questions relating to the attributes of multiple objects in the 2^{nd} and the 3^{rd} rows, and questions investigating different types of relationships in the 4^{th} row. They demonstrate the effectiveness of our defined explanations on elaborating the rationales behind the answers, and validate the usefulness of our functional program in automatically constructing the explanations.

4. Implementation of VisualBert-REX

In this section, we provide the implementation details of the proposed explanation generation method, *i.e.,* VisualBert-REX in the main paper. Similar to the VisualBert-EXP baseline, our method adopts the state-of-the-art VisualBert [2] as our visual reasoning backbone and the LSTM-based language generator from [3], and jointly predicts the answer and corresponding explanation. We concatenate word embeddings of the question and UpDown regional features [1], and use VisualBert to learn cross-modal features from them. Cross-modal features extracted at first token (*i.e.,* [CLS]) is utilized for predicting the answer and initializing the hidden state of the language generator. When sequentially generating each word in the explanation, we measure the similarity between the hidden state for the current step and the VisualBert features for all visual regions, and normalize the results to obtain the probabilities of grounding the current word in specific regions (*i.e.,* y_i^g in Equation 5 of the main paper). The grounding result is adaptively combined with the prediction determined based on the hidden state (*i.e.,* y_i^f in Equation 5 of the main paper), and the combined result is used to determine the next word in the explanation.

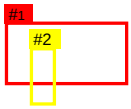
References

[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 *CVPR*, pages 6077–6086, 2018. 3

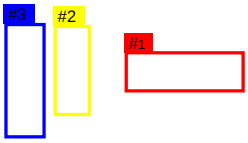
[2] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. What does BERT with vision look at? In *ACL*, pages 5265–5275, 2020. 3

[3] Jialin Wu and Raymond Mooney. Faithful multimodal explanation for visual question answering. In *ACL*, pages 103–112, 2019. 3

[4] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, pages 6713–6724, 2019. 2

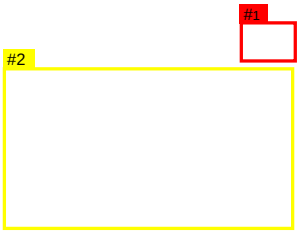


Question: Are both plates and forks in the picture?
Answer: yes
Explanation: Because there are # 1 and #2.



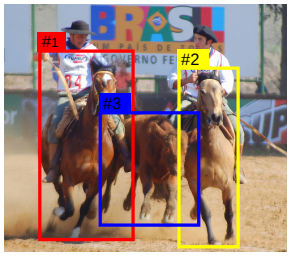
Question: Is there any surfboard to right of the man the people are standing by?
Answer: yes
Explanation: Because # 1 is to the right of #2 standing near #3.

Question: Do you see a chair to the left of pillow?
Answer: no
Explanation: Because there is no chair to the left of #1.



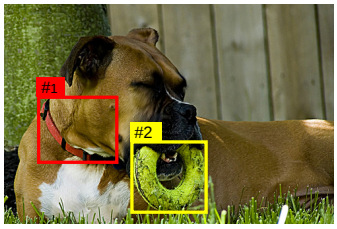
Question: What do the window and the bed have in common?
Answer: shape
Explanation: Because both #1 and #2 are round.

Question: What do the end table and the frame have in common?
Answer: material
Explanation: Because both #1 and #2 are wood.

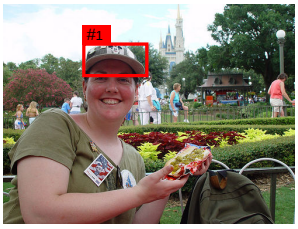


Question: Are these animals of different types?
Answer: yes
Explanation: Because #1 and #2 are horse, #3 is bull.

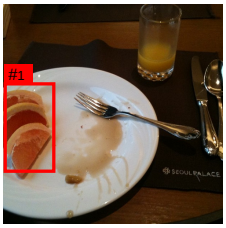
Question: Is the cap different in color than the sign?
Answer: no
Explanation: Because #1 and #2 are all black.



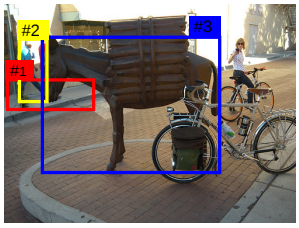
Question: Is the toy of the dog different in color than the collar?
Answer: yes
Explanation: Because #1 is red, #2 is green.



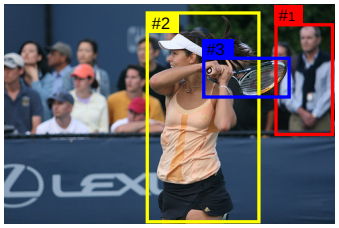
Question: What kind of clothing is not short sleeved?
Answer: cap
Explanation: Because #1 is cap and it is not short sleeved.



Question: Which side of the photo is the pink grapefruit on?
Answer: left
Explanation: Because #1 is located at (0, 0.4).



Question: The man that is to the left of the horse is walking where?
Answer: street
Explanation: Because #1 that #2 to the left of #3 is walking on is street.



Question: Are there any man to the right of the person that is holding a tennis racket?
Answer: yes
Explanation: Because #1 is to the right of #2 holding #3.

Figure 3. Qualitative examples of explanations in our GQA-REX dataset.