# Supplementary Material of "CRIC: A VQA Dataset for Compositional Reasoning on Vision and Commonsense"

Difei Gao, *Student Member, IEEE*, Ruiping Wang, *Senior Member, IEEE*,
Shiguang Shan, *Fellow, IEEE*, and Xilin Chen, *Fellow, IEEE*

## OVERVIEW

In the supplementary material, we provide more details of CRIC dataset, baselines and experiments:
1 - Function Definition of CRIC
2 - Details of NMN-CS
3 - Some QA samples in CRIC
4 - More Experimental Results.

## 1 FUNCTION DEFINITION OF CRIC

In this section, we introduce the 10 basic functions that our dataset aims to evaluate. These functions operate on some values that are indicated in a question or generated by some neural modules and output an object list or a concept.

**Inputs of functions.** These basic functions have two types of inputs. The first one is the text input that is indicated in a question:

- `object`: An object name, e.g., *dog, double decker*.
- `attribute`: An attribute name, e.g., *blue, open*.
- `predicate`: A predicate name, e.g., *on, holding*.
- `type`: A category name that indicates recognizing one type of concepts, e.g., *color, object, animal*.
- `KG_Query`: A commonsense query indicates a knowledge item, where some elements are replaced with BLANK or VISION for specific functions, e.g., *<BLANK, Can, climb the VISION>*, which means finding objects that can climb a specified object (e.g., a tree) in an image.
- `hypernym`: A category name that indicates finding objects belonging to one category, e.g., *animal, furniture*.

Another type of input is a vector generated by some other modules:

- *objects*: A set of objects (could contain zero, one or multiple objects) in an image.

**Outputs of functions.** Our functions have two types of outputs:

- *objects*: A set of objects in a given image.
- *concept*: A concept that could be the name of a visual concept (object, attribute, scene, etc.) or a boolean value (indicates yes or no).

**Basic Functions.** In this part, we introduce 7 visual basic functions (the other 3 commonsense functions have been illustrated in our main paper in Sec.3 **Function Definition**).

- Find: Given a set of objects, filter the objects by the object name or the attribute name or both two, e.g., find "cat", find "black", find "black cat".
- Relate: Return all objects in the image that have the specified relation `predicate` to the input objects, where input objects are the "subject", output objects are the "object". For example, find all objects that the man ("subject") is holding ("predicate").
- Relate Reverse: Return all objects in the image that have the specified relation `predicate` to the input objects, where input objects are the "object", output objects are the "subject". For example, find all objects that are on ("predicate") the table ("object").
- Recognition: Recognize the concept in the *objects* among one `type` of concepts, e.g., recognize the color in one image region.
- And: Return the intersection of two sets of objects.
- Verify: Given a set of objects, output *yes* if the set is non-empty and *no* if it is empty.
- Initial: Output the set of all objects in the image.

TABLE 1
The details of visual modules in NMN-CS. Each function takes text inputs $\mathbf{t}$ indicated in the question (i.e., the attended question feature), some attention inputs, such as $\mathbf{a}$, $\mathbf{a_1}$, and $\mathbf{a_2}$, generated by some other modules, and features of all objects $\mathbf{v}$ as inputs, then achieves corresponding function and outputs an attention map $\mathbf{y}_a$ (shorted as "att") or a probability vector $\mathbf{y}_c$ over all candidate answers. The operator $\odot$ is element-wise multiplication, $\mathrm{sum}$ is summing the results over spatial dimensions, and $\mathbf{e}_N$ is an N dimensional (the number of objects in the image) vector where all elements are 1.

| Module | text inputs | attention inputs | output | Implementation details |
|---|---|---|---|---|
| Find | $\mathbf{t}$ | $\mathbf{a}$ | att | $\mathbf{y}_a = \mathrm{sigmoid}(\mathrm{FC}(\mathrm{FC}(\mathbf{a} \odot \mathbf{v}) \odot \mathrm{FC}(\mathbf{t}))$ |
| Relate/Relate_Reverse | $\mathbf{t}$ | $\mathbf{a}$ | att | $\mathbf{y}_a = \mathrm{sigmoid}(\mathrm{FC}(\mathrm{FC}(\mathbf{v}) \odot \mathrm{FC}(\mathrm{sum}(\mathbf{a} \odot \mathbf{v})) \odot \mathrm{FC}(\mathbf{t})))$ |
| Recognition | $\mathbf{t}$ | $\mathbf{a}$ | concept | $\mathbf{y}_c = \mathrm{sigmoid}(\mathrm{FC}(\mathrm{FC}(\mathrm{sum}(\mathbf{a} \odot \mathbf{v})) \odot \mathrm{FC}(\mathbf{t})))$ |
| And | (none) | $\mathbf{a}_1$, $\mathbf{a}_2$ | att | $\mathbf{y}_a = \mathrm{sigmoid}(\mathrm{FC}(\mathrm{FC}(\mathbf{a}_1) \odot \mathrm{FC}(\mathbf{a}_2)))$ |
| Verify | (none) | $\mathbf{a}$ | concept | $\mathbf{y}_c = \mathrm{softmax}(\mathrm{FC}(\mathbf{a}))$ |
| Initial | (none) | (none) | att | $\mathbf{y}_a = \mathbf{e}_N$ |

## 2 Details of NMN-CS

In Sec.4 of the body, we illustrate the commonsense-related modules of NMN-CS. In Table 1, we show the details of visual modules in NMN-CS. A visual module is a function $y = f(\mathbf{a}_1, ..., \mathbf{a}_n, \mathbf{v}, \mathbf{t})$ that takes $n$ ($n \in \{0, 1, 2\}$ in our model, and $n = 0$ indicates that no $\mathbf{a}_i$ is inputted into the function) tensors $(\mathbf{a}_1, ...., \mathbf{a}_n)$ generated from other neural modules, image features $\mathbf{v}$ and text input feature $\mathbf{t}$ extracted from the question (an attended question features generated by program prediction module) as inputs, and outputs a tensor $\mathbf{y}$ which is either an attention map $\mathbf{a}$ over image regions or the probability $\mathbf{c}$ over all possible answers.

## 3 Some QA samples in CRIC

In Fig. 1, we show more QA samples in the CRIC dataset and some questions from GQA and VQA v2, which share the same images with the CRIC to present the differences between the datasets.

## 4 More Experimental Results

### 4.1 Prediction Results of Representative Methods

In Fig. 2, we also show some prediction results of Memory-VQA+$l_{att}$, ViLBERT+$l_{att}$, and a KB-aware version of ViLBERT (ViLBERT+ERNIE+$l_{att}$. It can be seen that although the performance of ViLBERT+$l_{att}$ is much higher than Memory-VQA+$l_{att}$, there are indeed some questions where Memory-VQA+$l_{att}$ performs better than ViLBERT+$l_{att}$, e.g., Q1-Q3. These samples usually contain some confusing objects in images, e.g., multiple computers and monitors with different colors in Q2, or the appearance of the target objects are relatively rare, e.g., the carpet and bridge in Q1 and Q3 are relatively rare in the dataset. Without the explicit use of knowledge items, ViLBERT+$l_{att}$ is harder to precisely align these rare or confusing objects to the implicit knowledge depicted in parameters. In contrast, this issue can be alleviated to a certain extent when the ViLBERT+$l_{att}$ has access to knowledge items, i.e., ViLBERT+ERNIE+$l_{att}$ correctly predicts these answers. In addition, from Q4-Q6, we can observe the strength of ViLBERT in visual representation ability. It can precisely recognize some small objects (glasses in Q4, car in Q6), and fine-grained objects (olive oil in Q5).

### 4.2 Performance on revised version of CRIC

As described in main body Sec.3, we used GingerIt to provide modification suggestions of CRIC questions. We adopted suggestions other than modifying prepositions (this type of modification will change the meaning of the questions) and obtained a revised version of CRIC. We then evaluated several methods on this revised version dataset. The results are shown in Tab. 2. It can be seen that the results on the revised version are almost the same as the original version, which reflects that the linguistic quality does not have much impact on model evaluation.

TABLE 2
Model performances on the CRIC dataset before and after revision.

| Method | Final Score - Before Revision | Final Score - After Revision |
|---|---|---|
| MAC | 26.19 | 26.01 |
| Memory-VQA+$l_{att}$ | 38.87 | 38.22 |
| ViLBERT+$l_{att}$ | 53.76 | 53.59 |

**CRIC:**
**1.** Is there an object which can be used for holding cloth? no
**2.** Is the brown object that is near the couch a type of herbivore? no
**3.** What is on top of the furniture that is usually used for holding things? tv
4. What black object on the floor can be used for illuminating area? lamp
**VQA v2:**
**1.** What is the man currently doing in this picture?
**2.** What color is the dog?
**3.** What is the make of the laptop computer?

**CRIC:**
**1.** What object is a type of dessert? cookie
**2.** What size is the object that is on the plate and is made from flour? small
**3.** Is the object that is on the small plate a type of citrus fruit? no
**4.** What is the small sandwich on and can be used for holding food? plate

**VQA v2:**
**1.** How many plates of food?
**2.** What color are the plates?
**3.** Do each plate have carrots on them?

**CRIC:**
**1.** Which object that has the tail can carry people? horse
**2.** Can the animal that has the tail pull white object? yes
**3.** Is there a road that is cement and can be used for driving a car on? no
**4.** Can the vehicle which is white travel on road? no
**VQA v2:**
**1.** What kind of horses are these?
**2.** What is behind the horses?
**3.** Is this modern transportation?

**CRIC:**
**1.** What red food on the plastic cutting board is usually used for eating? meat
**2.** What eating utensil which is made of metal does chopping require? knife
**3.** is the fruit that is on the counter sweet? yes
**4.** Can the red object that is on the plastic cutting board be used for eating? yes
**GQA:**
**1.** Which side is the red meat on?
**2.** How is the food to the left of the cutting board in this image called?
**3.** What color does the tray to the left of the apple have?

**CRIC:**
**1.** Is there a maroon object that is next to the phone and can be used for signing checks? no
**2.** What type of object is next to the black object that I can use for talking to someone? pen
**3.** Is there an electronic device that can be used for listening to music? no
**GQA:**
**1.** Which kind of furniture is white?
**2.** Are there any black phones?
**3.** Are there both bags and phones in the photo?

**CRIC:**
**1.** Which utensil in the picture does frying require? frying pan
**2.** What metal is the object that is on the stove and is a type of kitchen utensil made of? stainless steel
**3.** Is there a silver kitchenware that I can use for heating the object that is in the frying pan? yes

**GQA:**
**1.** What is this appliance called?
**2.** What kind of appliance is it?
**3.** Are there any eggs on the stainless steel pan?

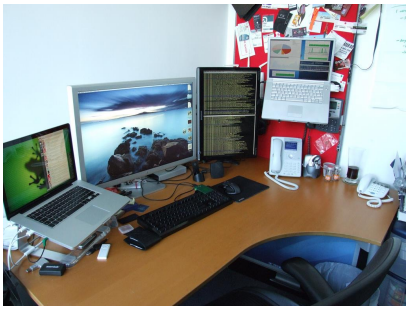
**1.** Which object is usually used for chilling the object behind the glass? refrigerator
**2.** What is in the home appliance that can be used for chilling drinks? bottle
**3.** Is there an object that is a type of soft drinks? yes
4. Is there an object that can be used for lying down? yes

**1.** Which large vehicle can use diesel fuel? truck
**2.** Can the large vehicle sail through sea? no
**3.** Which color is the vehicle that can use diesel fuel? blue

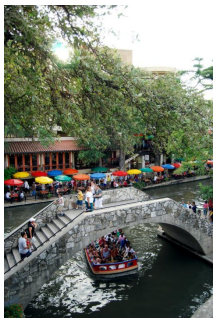
**1.** Is there a tableware that is bright blue and is usually used for keeping the furniture that is on the patio clean? no
**2.** What beverage on the table is liquid? wine
**3.** What is the shape of the object that is on the pole and can be used for lighting a room? oval
**4.** Is the object on the table a type of soft drinks? no

Fig. 1. The top one row displays the COCO images in Visual Genome and the corresponding questions from CRIC and VQA [1]. The middle and bottom rows display the Flickr images in Visual Genome and the corresponding questions from CRIC and GQA [2]. VQA v2 [1] mainly focuses on questions about querying visual facts with less compositionality. More recent GQA [2] mainly evaluates the complex compositional reasoning on visual facts. While CRIC contains compositional questions requiring reasoning on visual and commonsense.

**Question1:** Which color is the object that I can use for protecting feet from the floor
**Memory-VQA +$l_{att}$ :** brown
**ViLBERT+$l_{att}$:** red
**ViLBERT+ERNIE+$l_{att}$:** brown



**Question2:** Is there an object that is on top of the desk, is black and can display images
**Memory-VQA +$l_{att}$ :** yes
**ViLBERT+$l_{att}$:** no
**ViLBERT+ERNIE+$l_{att}$:** yes



**Question3:** Is there a cement place that is usually used for crossing a river
**Memory-VQA +$l_{att}$ :** yes
**ViLBERT+$l_{att}$:** no
**ViLBERT+ERNIE+$l_{att}$ :** yes

**Question4:** Who is wearing the accessory that is usually used for improving eyesight
**Memory-VQA +$l_{att}$ :** girl
**ViLBERT+$l_{att}$:** man
**ViLBERT+ERNIE+$l_{att}$:** man

**Question5:** Which thing depicted in the image is fluid?
**Memory-VQA +$l_{att}$ :** soup
**ViLBERT+$l_{att}$ :** olive oil
**ViLBERT+ERNIE+$l_{att}$:** olive oil

**Question6:** What color is the vehicle that is in the parking lot and can move quickly
**Memory-VQA +$l_{att}$ :** white
**ViLBERT+$l_{att}$:** black
**ViLBERT+ERNIE+$l_{att}$:** black

Fig. 2. Some example prediction results of Memory-VQA+$l_{att}$, ViLBERT+$l_{att}$, ViLBERT+ERNIE+$l_{att}$. The answer in green indicates that it is correct. Red indicates the answer is wrong.

## ACKNOWLEDGMENTS

## REFERENCES

[1] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "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 (CVPR)*, 2017, pp. 6904–6913.
[2] D. A. Hudson and C. D. Manning, "Gqa: a new dataset for compositional question answering over real-world images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 6700–6709.