

# Learning to Reason: End-to-End Module Networks for Visual Question Answering

(Supplementary Material)

Ronghang Hu<sup>1</sup>   Jacob Andreas<sup>1</sup>   Marcus Rohrbach<sup>1,2</sup>   Trevor Darrell<sup>1</sup>   Kate Saenko<sup>3</sup>

<sup>1</sup>University of California, Berkeley   <sup>2</sup>Facebook AI Research   <sup>3</sup>Boston University

{ronghang, jda, trevor, rohrbach}@eecs.berkeley.edu, saenko@bu.edu

## 1. Details on the CLEVR dataset

In this section, we show more experimental results and detailed visualization of outputs from the proposed N2NMN model on the CLEVR dataset.

### 1.1. Accuracy vs question size

Figure 1 shows the accuracy of our model on questions of different length. It can be seen that on long questions our model can still achieve relatively high accuracy.

### 1.2. Visualized examples

In Figure 2, we visualize the outputs from our model in detail, showing the predicted layouts, textual attention weights and outputs from each module.

These visualizations show that the model indeed learns intuitive and modular reasoning behaviors. For example, consider the top-left question on the next page, *What size is the thing that is on the left side of the tiny purple matte cube and behind the cyan block?* To answer this question, the model first uses `find[0]` (the number in the bracket is the index of this module in the decoded layout sequence) to produce spatial attention for *the tiny purple matte cube*, and shifts that attention to the *left* using `relocate[1]`. It also uses `find[2]` to locate the *cyan block* and then look *behind* it using `relocate[3]`. Finally, it takes the intersection of the above two areas, which contains just a single object (the small purple shiny sphere), and applies the `describe[5]` module to determine the object’s *size*. This clearly shows that the model is actually learning grounded language parsing: it maps some words to specific objects and their properties, other words to spatial relationships, etc., and learns to compose these concepts together to compute the answer.

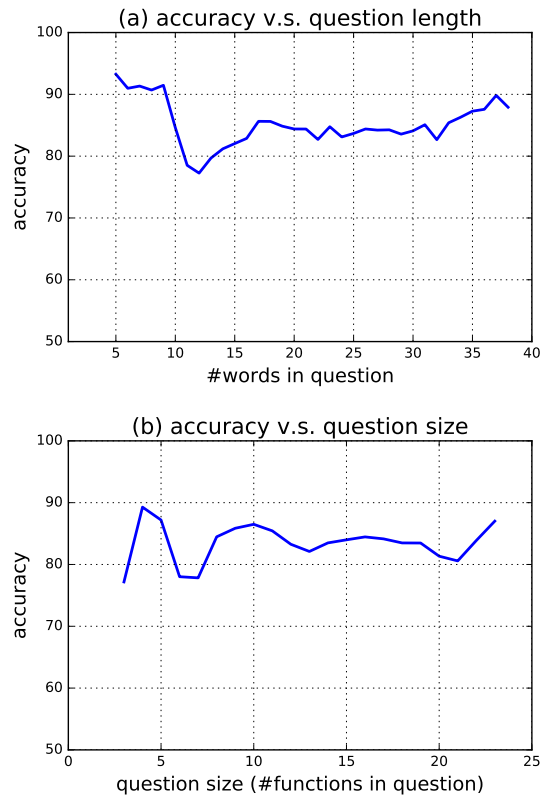
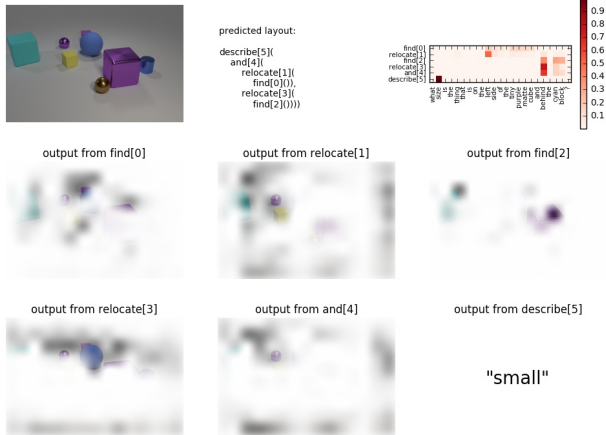


Figure 1: (a) The accuracy of our model v.s. question length (the number of words in the question) on the CLEVR validation set. Our model can still achieve relatively high accuracy on long questions with 30 or more words. (b) The accuracy of our model v.s. question size. On the CLEVR dataset, the size of a question is the number of functions in the functional program used to synthesize the question.

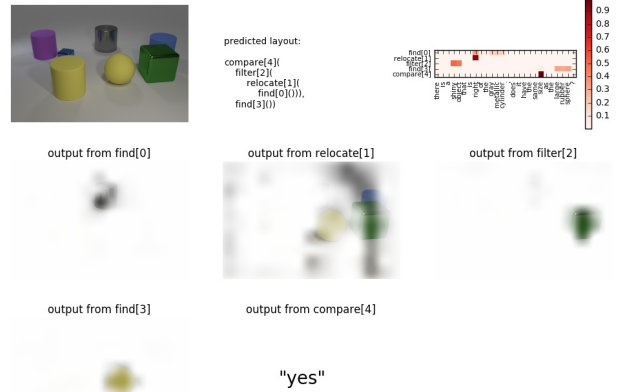
question: what size is the thing that is on the left side of the tiny purple matte cube and behind the cyan block ?

ground-truth answer: "small" predicted answer: "small"



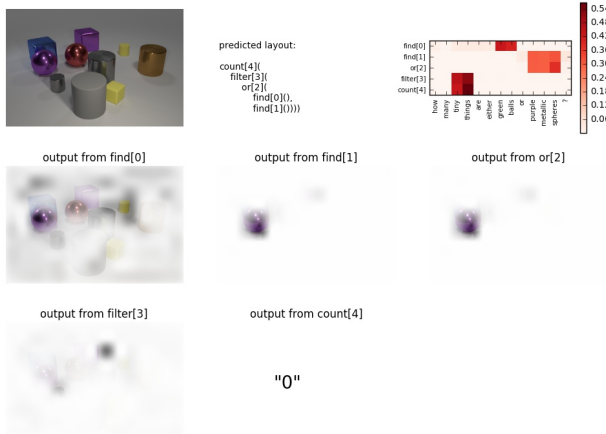
question: there is a shiny object that is right of the gray metallic cylinder ; does it have the same size as the large rubber sphere ?

ground-truth answer: "yes" predicted answer: "yes"



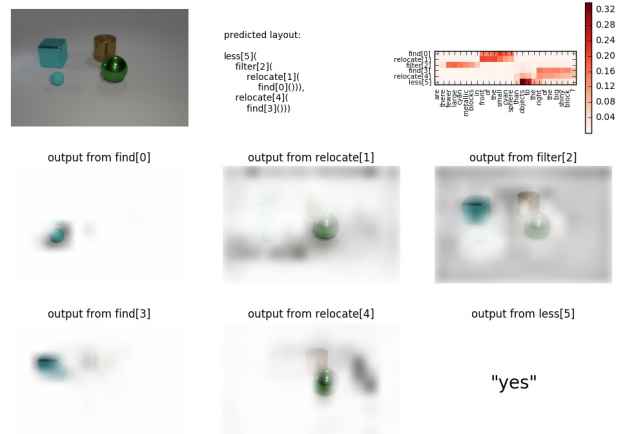
question: how many tiny things are either green balls or purple metallic spheres ?

ground-truth answer: "0" predicted answer: "0"



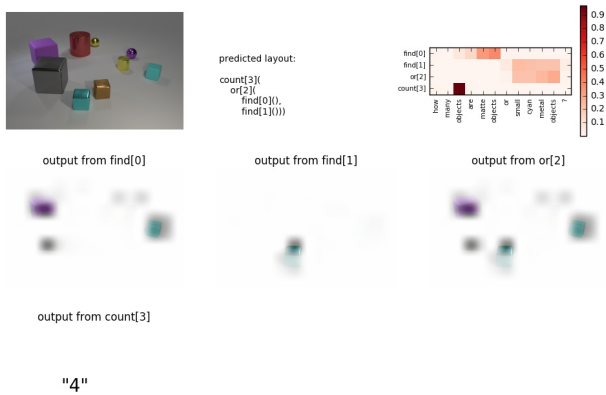
question: are there fewer large cyan metallic blocks in front of the small cyan sphere than objects to the right of the big shiny block ?

ground-truth answer: "yes" predicted answer: "yes"



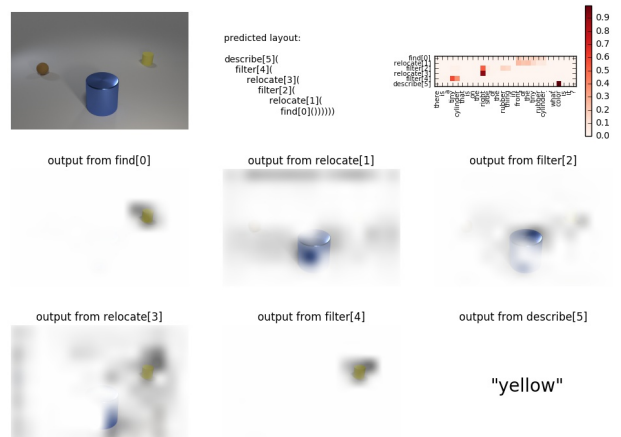
question: how many objects are matte objects or small cyan metal objects ?

ground-truth answer: "3" predicted answer: "4"

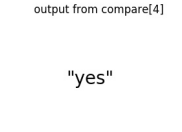
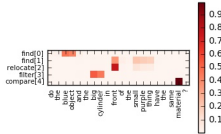


question: there is a tiny cylinder that is on the right side of the rubber thing in front of the tiny rubber cylinder ; what color is it ?

ground-truth answer: "yellow" predicted answer: "yellow"

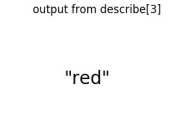
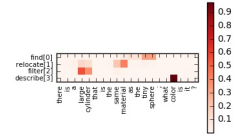


ground-truth answer: "yes"    predicted answer: "yes"



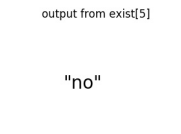
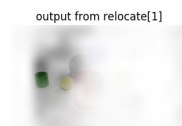
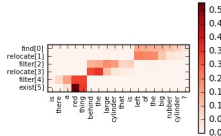
"yes"

ground-truth answer: "red"      predicted answer: "red"



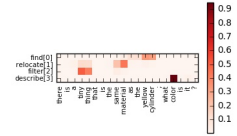
```
"red"
```

ground-truth answer: "no"      predicted answer: "no"



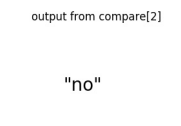
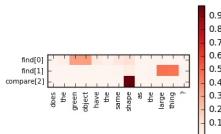
"no"

ground-truth answer: "blue"    predicted answer: "blue"



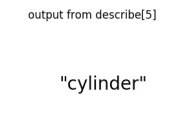
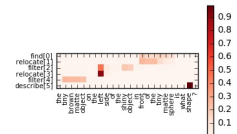
"blue"

ground-truth answer: "no"      predicted answer: "no"



"no"

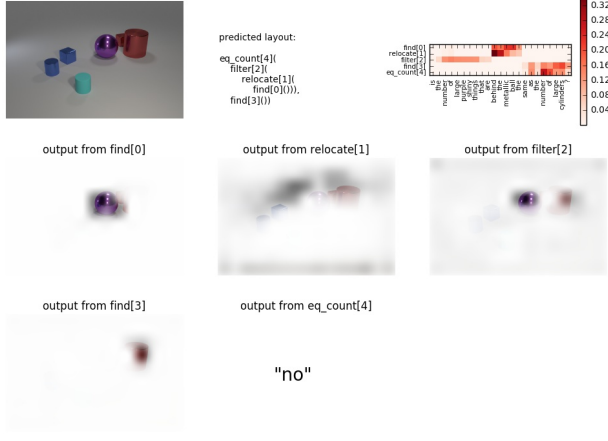
ground-truth answer: "cylinder"      predicted answer: "cylinder"



"cylinder"

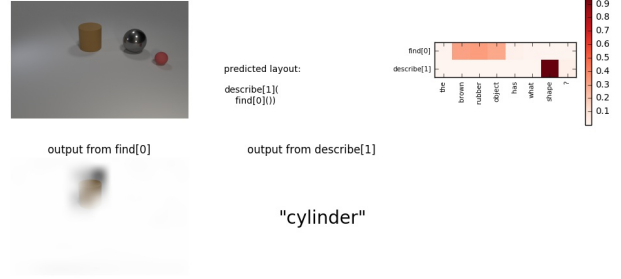
question: is the number of large purple shiny things that are behind the metallic ball the same as the number of large cylinders ?

ground-truth answer: "no" predicted answer: "no"



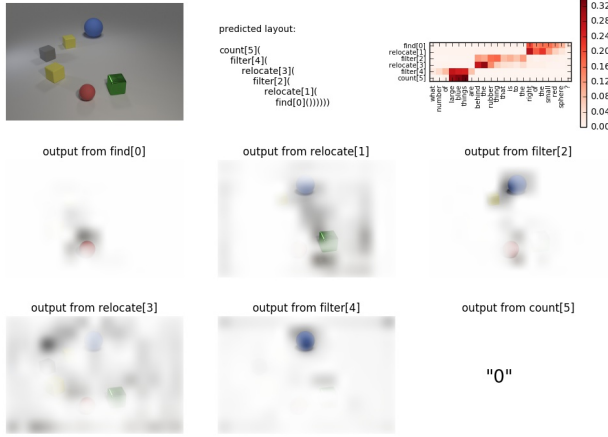
question: the brown rubber object has what shape ?

ground-truth answer: "cylinder" predicted answer: "cylinder"



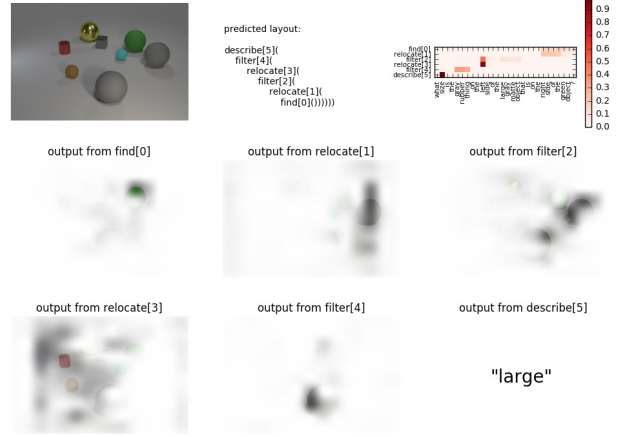
question: what number of large blue things are behind the rubber thing that is to the right of the small red sphere ?

ground-truth answer: "0" predicted answer: "0"



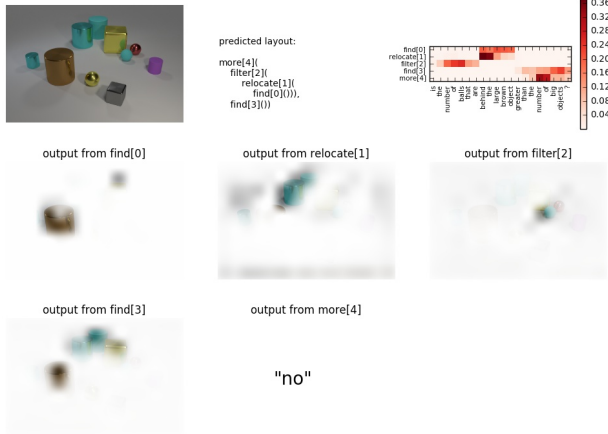
question: what size is the gray rubber thing on the left side of the large gray matte object that is on the right side of the green object ?

ground-truth answer: "large" predicted answer: "large"



question: is the number of balls that are behind the large brown object greater than the number of big objects ?

ground-truth answer: "no" predicted answer: "no"



question: is the material of the yellow thing the same as the cube behind the big brown rubber thing ?

ground-truth answer: "no" predicted answer: "no"

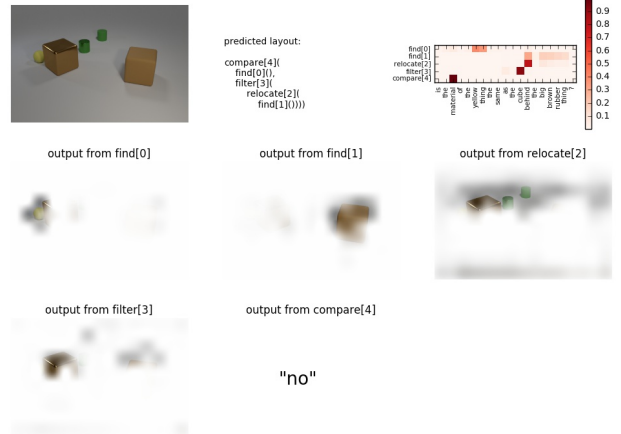


Figure 2: Detailed visualization of the predictions from our model on the CLEVR validation set. In each example, the first row shows the image, the predicted layout and the generated textual attention weights. Then, the output of each module is visualized as either an image attention map or an answer, depending on the output type of this module.