CLEVR-Ref+: Diagnosing Visual Reasoning with Referring Expressions

Runtao Liu, **Chenxi Liu**, Yutong Bai, Alan Yuille

02/12/2019

What is referring expression

giraffe on the right

What is wrong with referring expression

- Dataset bias
  - [1] showed that even if the referring expression is discarded and prediction is made solely from the image, accuracy is much higher than random

[1] Cirik, Volkan, Louis-Philippe Morency, and Taylor Berg-Kirkpatrick. "Visual Referring Expression Recognition: What Do Systems Actually Learn?." In NAACL, 2018

What is wrong with referring expression

- Dataset bias
  - [1] showed that even if the referring expression is discarded and prediction is made solely from the image, accuracy is much higher than random
- Cannot inspect the entire visual reasoning process
  - Getting the final result right does not imply true understanding
  - Getting all the intermediate results right is a much stronger evidence

[1] Cirik, Volkan, Louis-Philippe Morency, and Taylor Berg-Kirkpatrick. "Visual Referring Expression Recognition: What Do Systems Actually Learn?." In NAACL, 2018

What is wrong with referring expression

- Dataset bias
  - [1] showed that even if the referring expression is discarded and prediction is made solely from the image, accuracy is much higher than random
- **Solution: Synthetic Dataset**

- Cannot inspect the entire visual reasoning process
  - Getting the final result right does not imply true understanding
  - Getting all the intermediate results right is a much stronger evidence

[1] Cirik, Volkan, Louis-Philippe Morency, and Taylor Berg-Kirkpatrick. "Visual Referring Expression Recognition: What Do Systems Actually Learn?." In NAACL, 2018

What is wrong with referring expression

- Dataset bias
  - [1] showed that even if the referring expression is discarded and prediction is made solely from the image, accuracy is much higher than random
- **Solution: Synthetic Dataset**

- Cannot inspect the entire visual reasoning process
  - Getting the final result right does not imply true understanding
  - Getting all the intermediate results right is a much stronger evidence
- **Solution: Modular Approach**

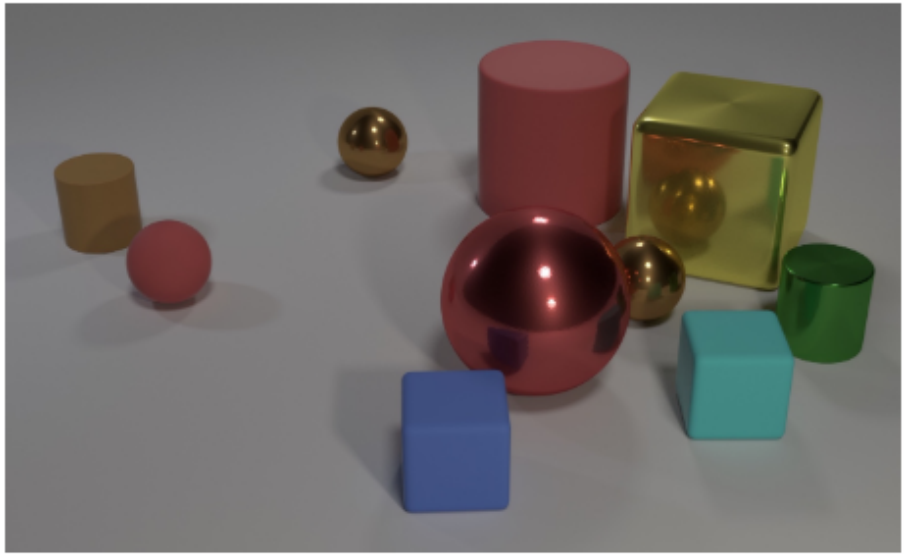
[1] Cirik, Volkan, Louis-Philippe Morency, and Taylor Berg-Kirkpatrick. "Visual Referring Expression Recognition: What Do Systems Actually Learn?." In NAACL, 2018

The CLEVR-Ref+ Dataset

CLEVR

- Simple, synthetic scene with full knowledge
- Synthetic sentences generated by templates
- Designed for visual question answering

Questions in CLEVR test various aspects of visual reasoning including **attribute identification**, **counting**, **comparison**, **spatial relationships**, and **logical operations**.



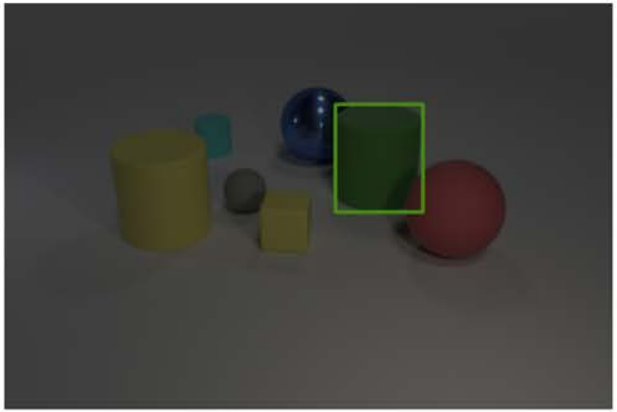
Q: Are there an **equal number** of **large things** and **metal spheres**?
Q: **What size** is the **cylinder** **that is left of** the **brown metal** thing **that is left of** the **big sphere**?
Q: There is a **sphere** with the **same size as** the **metal cube**; is it **made of the same material as** the **small red sphere**?
Q: **How many** objects are **either** **small cylinders** or **red** things?

Johnson, Justin, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning." In CVPR, 2017.

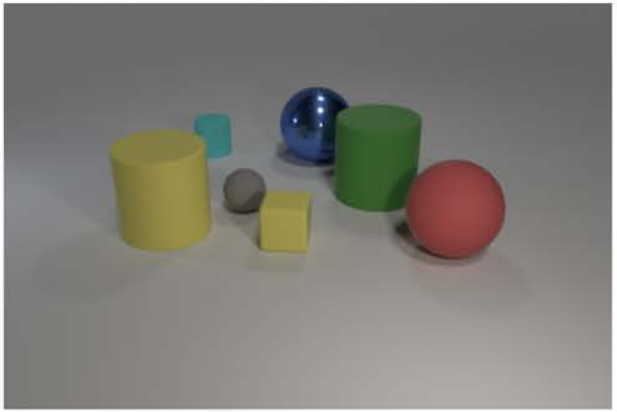
Step 1: Question -> Referring Expression

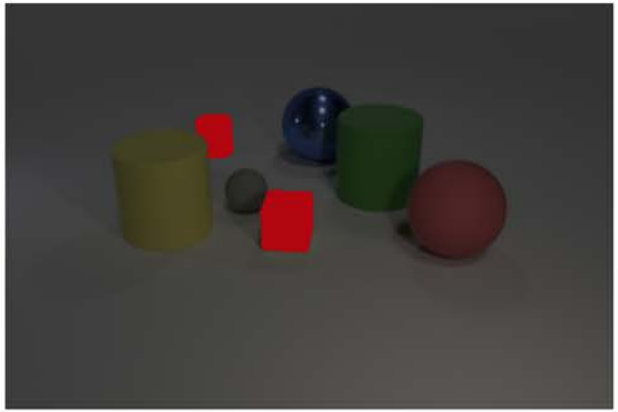
Category	Question (CLEVR)	Referring Expression (CLEVR-Ref+)
Basic	How many cyan cubes are there?	The cyan cubes.
Spatial Relation	Are there any green cylinders to the left of the brown sphere?	The green cylinders to the left of the brown sphere.
AND Logic	How many green spheres are both in front of the red cylinder and left to the yellow cube?	The green spheres that are both in front of the red cylinder and left to the yellow cube.
OR Logic	Are there any cylinders that are either purple metal objects or small red matte things?	Cylinders that are either purple metal objects or small red matte things.
Same Relation	Are there any other things that have the same size as the red sphere?	The things/objects that have the same size as the red sphere.
Compare Integer	Are there more brown shiny objects behind the large rubber cylinder than gray blocks?	-
Comparison	Does the small ball have the same color as the small cylinder in front of the big sphere?	-

Step 2: Answer -> Bounding Box/Segmentation Mask



The big thing(s) that are behind the second one of the big thing(s) from front and to the right of the first one of the large sphere(s) from left





Any other things that are the same size as the fifth one of the thing(s) from right

Step 3: Module Additions

- We listed the frequent words in RefCOCO+ [1] and manually categorized them
- We found that ordinal and visible are not covered by referring expression templates, so we added them

Category	Example words	Frequency
object	shirt,head,chair,hat,pizza	63.66%
human	man,woman,guy,girl,person	42.54%
color	white,black,blue,red,green	38.76%
spatial	back,next,behind,near,up	23.86%
animal	zebra,elephant,horse,bear	15.36%
attribute	big,striped,small,plaid,long	10.55%
action	standing,holding,looking	10.34%
ordinal	**closest,furthest,first,third**	**5.797%**
compare	smaller,tallest,shorter,older	5.247%
visible	**fully visible,barely seen**	**4.639%**

[1] Yu, Licheng, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. "Modeling context in referring expressions." In ECCV, 2016.

Step 4: Changes to Generation Procedure

- Better balance between templates
- Remove referring expressions that are too peculiar and rare
- Better prevention of the referring expression from being degenerate
- Refer to at least one object at the end

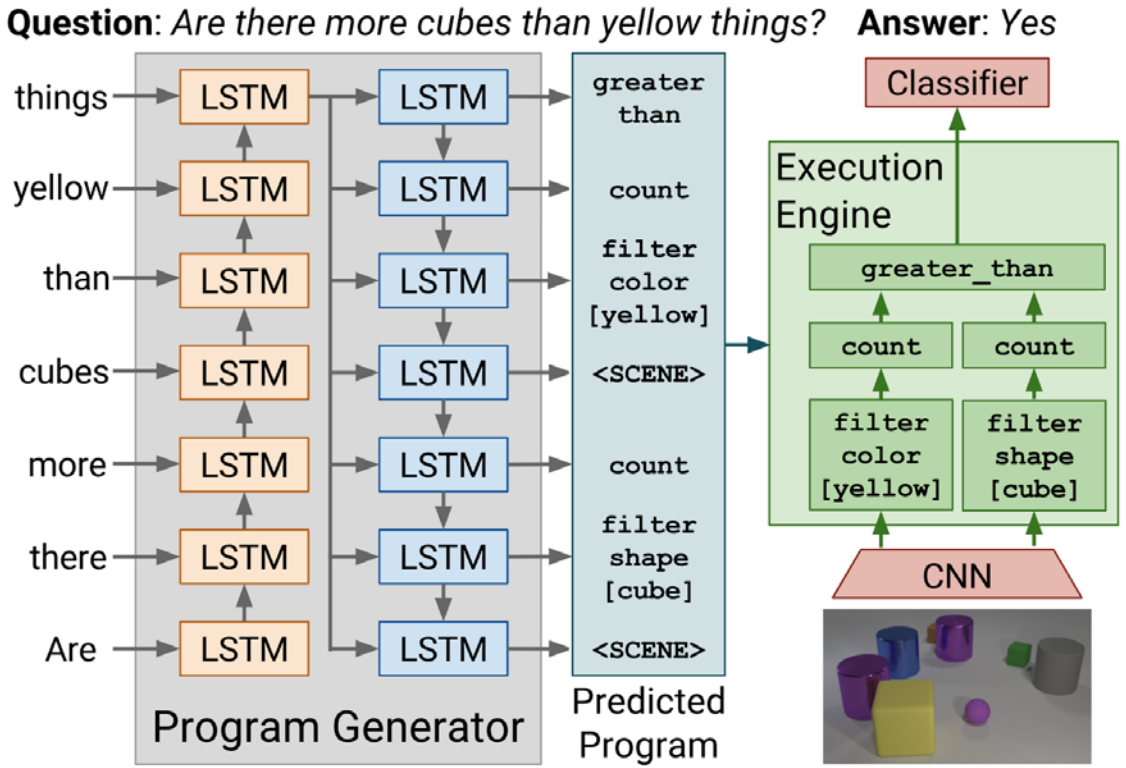
Summary

- Same scenes as CLEVR
  - 70K images in train
  - 15K images in val
  - 15K images in test
- Every image is associated with 10 referring expressions

The IEP-Ref Model: Unifying Segmentation and Diagnosis

IEP

- Stands for “Inferring and Executing Programs”
- Modular approach
- Designed for visual question answering

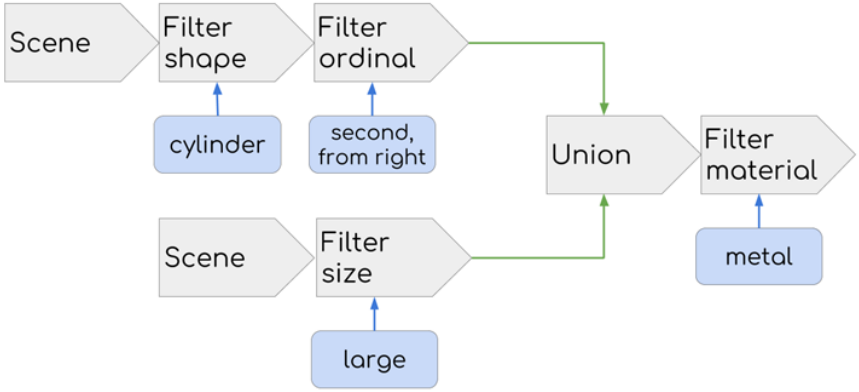


Johnson, Justin, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. "Inferring and executing programs for visual reasoning." In ICCV, 2017.

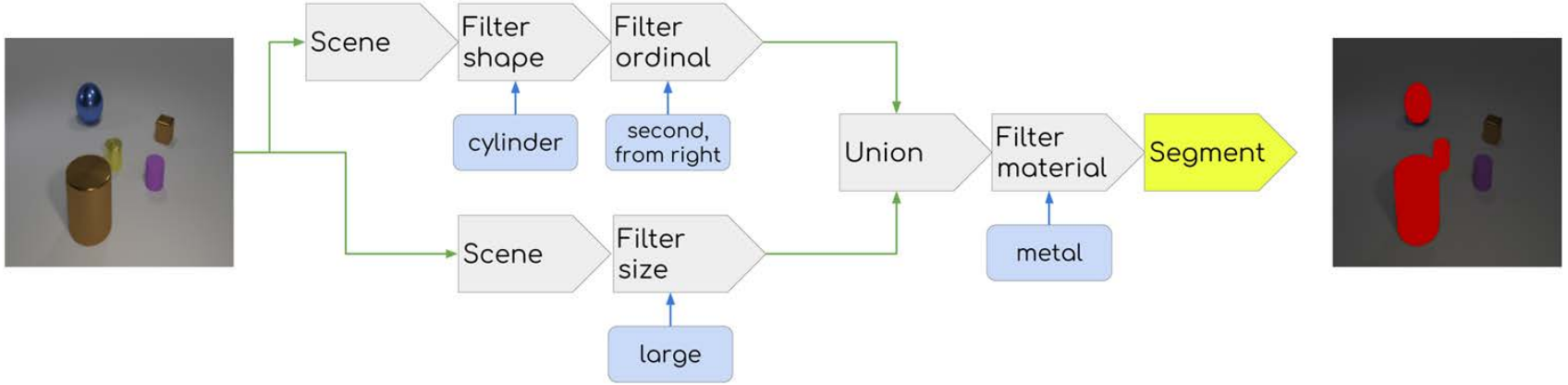
Step 1: Parsing

The metallic things that are the second one of the cylinder(s) from right or large object(s)

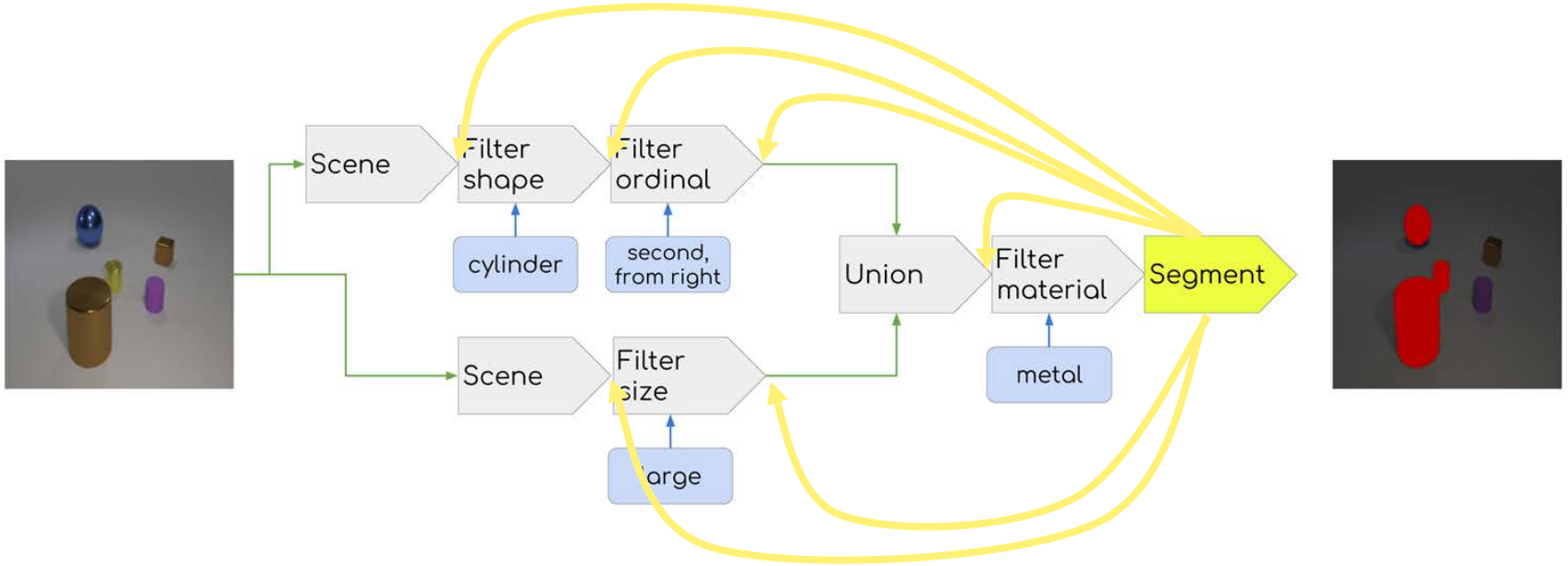
LSTM Parser



Step 2: Segmentation



Step 3: Diagnosis



Diagnosing Visual Reasoning: REF vs VQA

- The “logic” part is very similar. Both have to understand spatial relationship, logical operations etc.

Diagnosing Visual Reasoning: REF vs VQA

- The “logic” part is very similar. Both have to understand spatial relationship, logical operations etc.
- Recall that we are interested in visualizing intermediate steps. The output space of **visualization is basically the same as segmentation** (as in the REF setting), as opposed to textual answers (as in the VQA setting).
- In other words, we would have very little insight by plugging the “Exist” module (VQA setting) at intermediate steps.

Experimental Results: Accuracy

Models

- In addition to IEP-Ref (seg), we also evaluated three existing SOTA referring expression models on CLEVR-Ref+, to see their strengths and weaknesses:
  - Speaker-Listener-Reinforcer (det) [1]
  - MAttNet (det) [2]
  - Recurrent Multimodal Interaction (seg) [3]

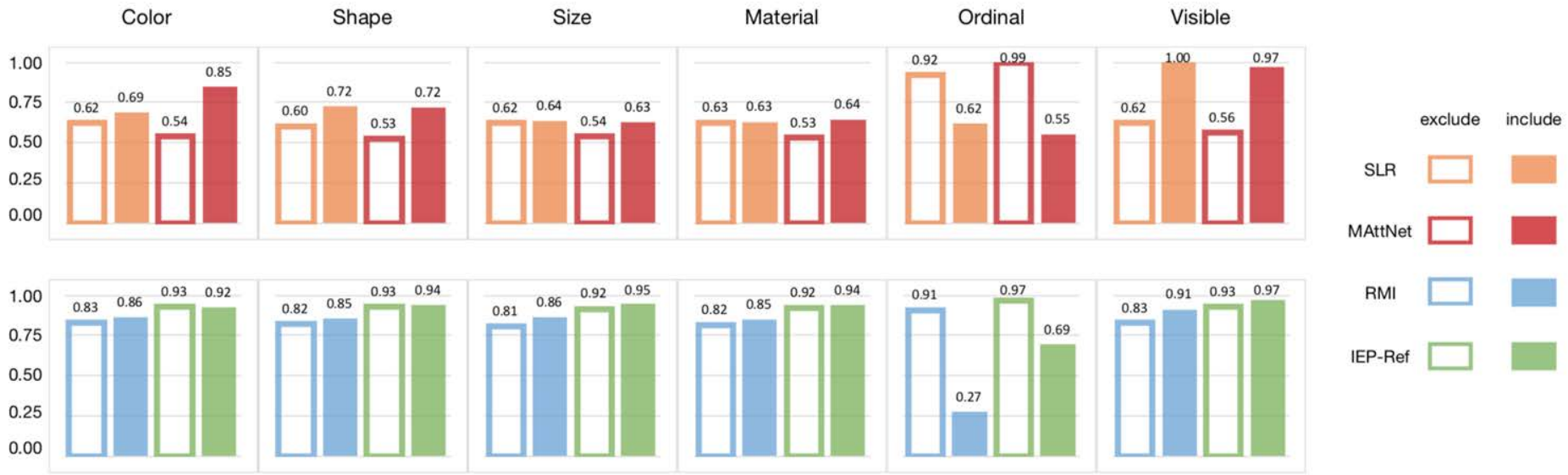
[1] Yu, Licheng, Hao Tan, Mohit Bansal, and Tamara L. Berg. "A joint speakerlistener-reinforcer model for referring expressions." In CVPR, 2017.
[2] Yu, Licheng, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L. Berg. "Mattnet: Modular attention network for referring expression comprehension." In CVPR, 2018.
[3] Liu, Chenxi, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, and Alan L. Yuille. "Recurrent Multimodal Interaction for Referring Image Segmentation." In ICCV, 2017

Overall Evaluation

	Basic	Spatial Relation			Logic				
	0-Relate	1-Relate	2-Relate	3-Relate	AND	OR	Same	Accuracy	IoU
SLR [35]	0.627	0.569	0.570	0.584	0.594	0.701	0.444	0.577	-
MAttNet [33]	0.566	0.623	0.634	0.624	0.723	0.737	0.454	0.609	-
RMI [21]	0.822	0.713	0.736	0.715	0.585	0.679	0.251	-	0.561
IEP-Ref (GT)	0.928	0.895	0.908	0.908	0.879	0.881	0.647	-	0.816
IEP-Ref (700K prog.)	0.920	0.884	0.902	0.898	0.860	0.869	0.636	-	0.806
IEP-Ref (18K prog.)	0.907	0.858	0.874	0.862	0.829	0.847	0.605	-	0.782
IEP-Ref (9K prog.)	0.910	0.858	0.847	0.811	0.778	0.791	0.626	-	0.760

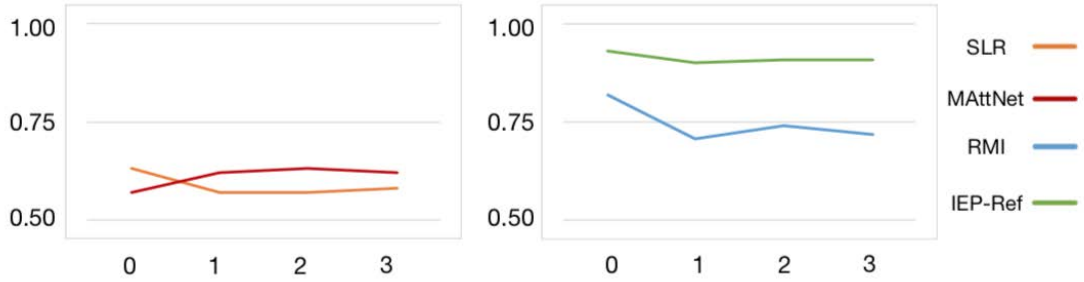
Basic Referring Ability

- Easy: color, shape, visibility
- Hard: ordinality



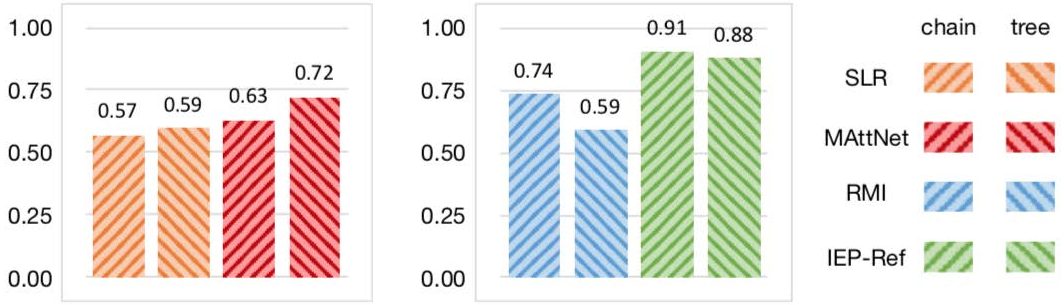
Spatial Reasoning Ability

- Seems that once the model has grasped spatial reasoning, there is little trouble in successfully applying it multiple times.



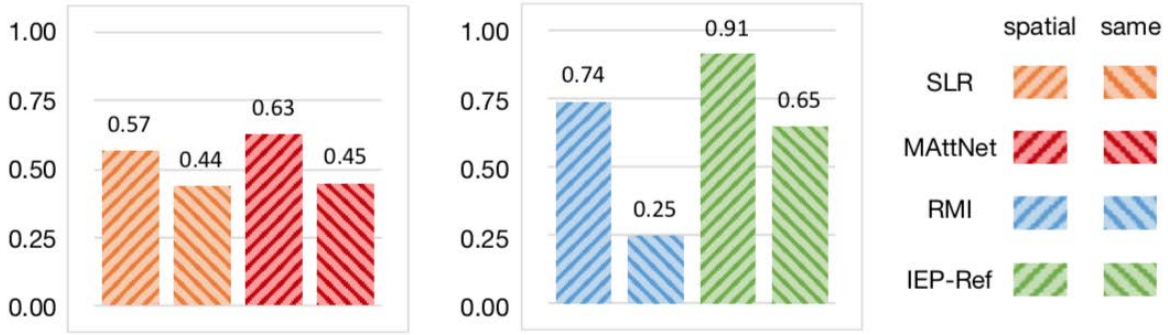
Different Reasoning Topologies

- Trees are generally harder, though not consistent.



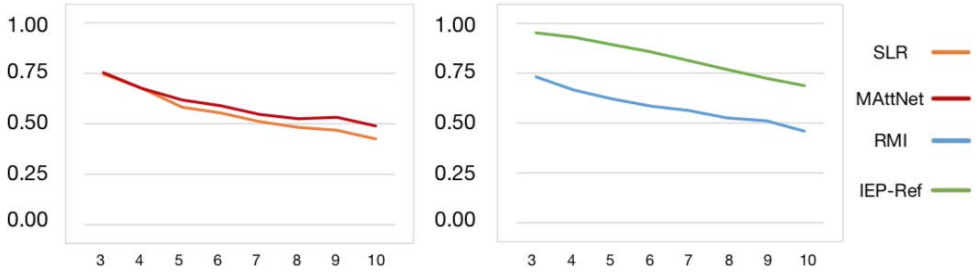
Different Relation Types

- “Same” is harder than “Spatial”
- Presumably because “Same” requires global context



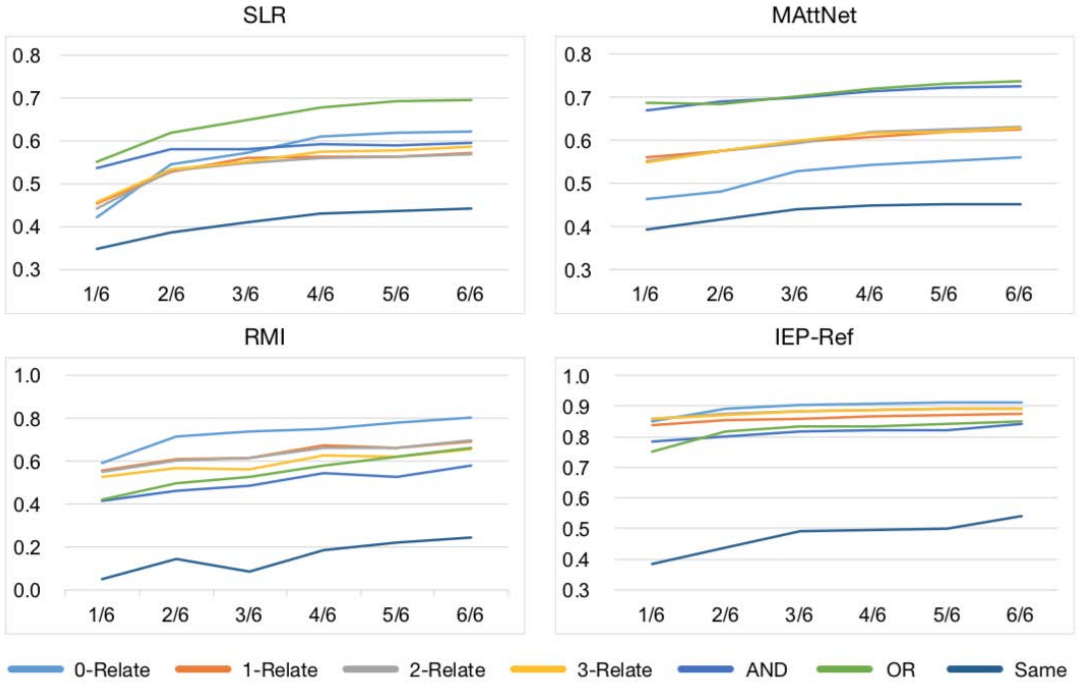
Number of Objects in a Scene

- More objects -> Harder



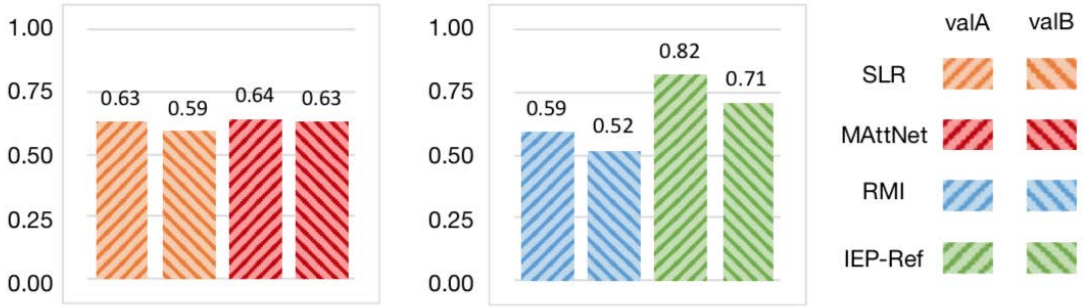
Schedule of Acquiring Reasoning Abilities

- No obvious schedule; all modules are learned at same pace



Novel Compositions

- Small drop in performance; generalize well in general



Experimental Results:
IEP-Ref Interpretability

Visualization in IEP

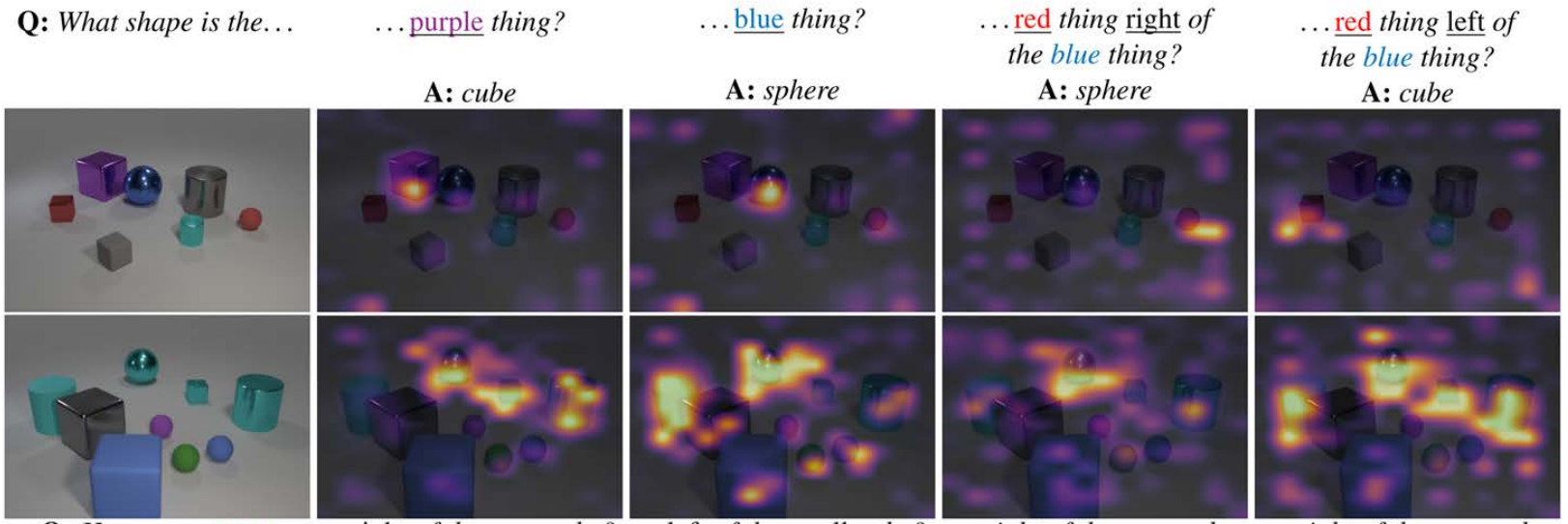
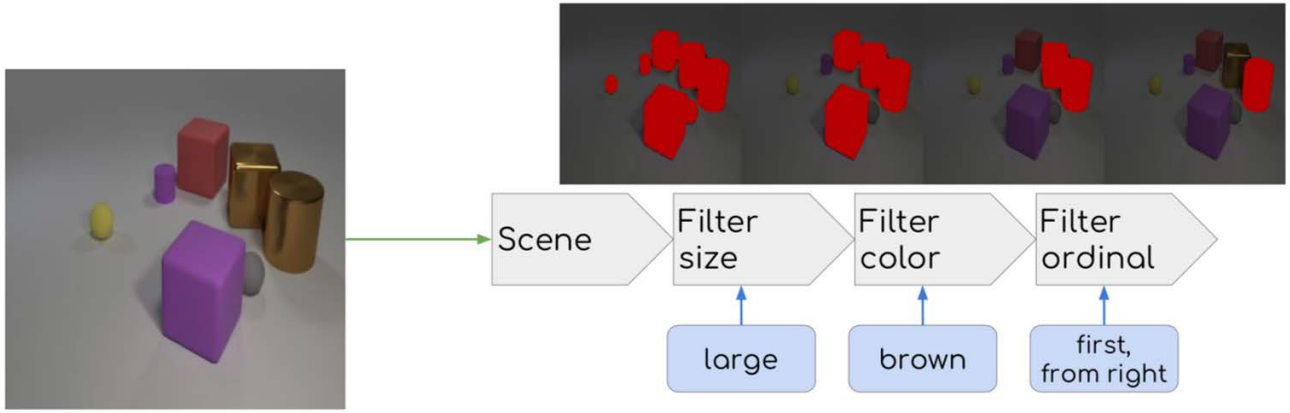


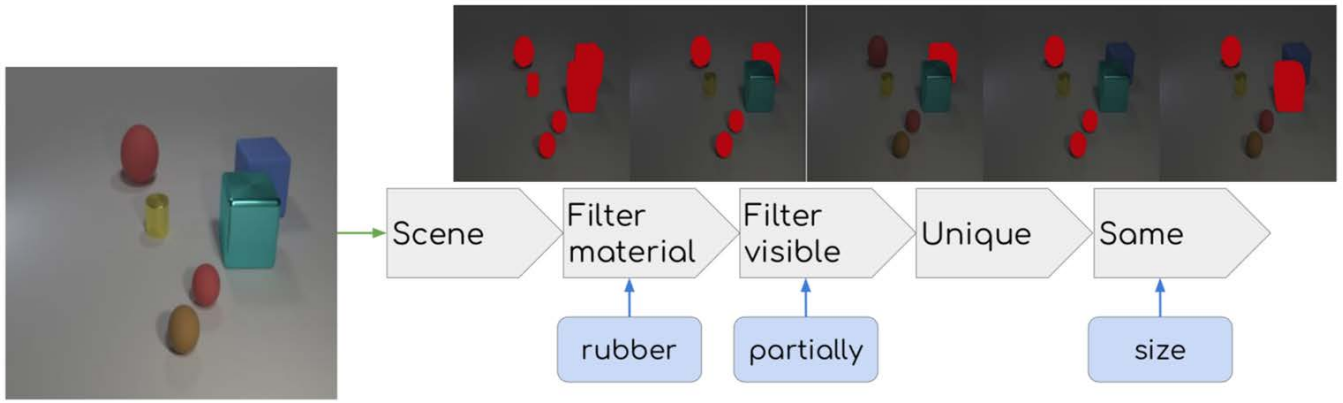
Figure 3. Visualizations of the norm of the gradient of the sum of the predicted answer scores with respect to the final feature map. From left to right, each question adds a module to the program; the new module is *underlined* in the question. The visualizations illustrate which objects the model attends to when performing the reasoning steps for question answering. Images are from the validation set.

Step-by-step inspection: chain structure



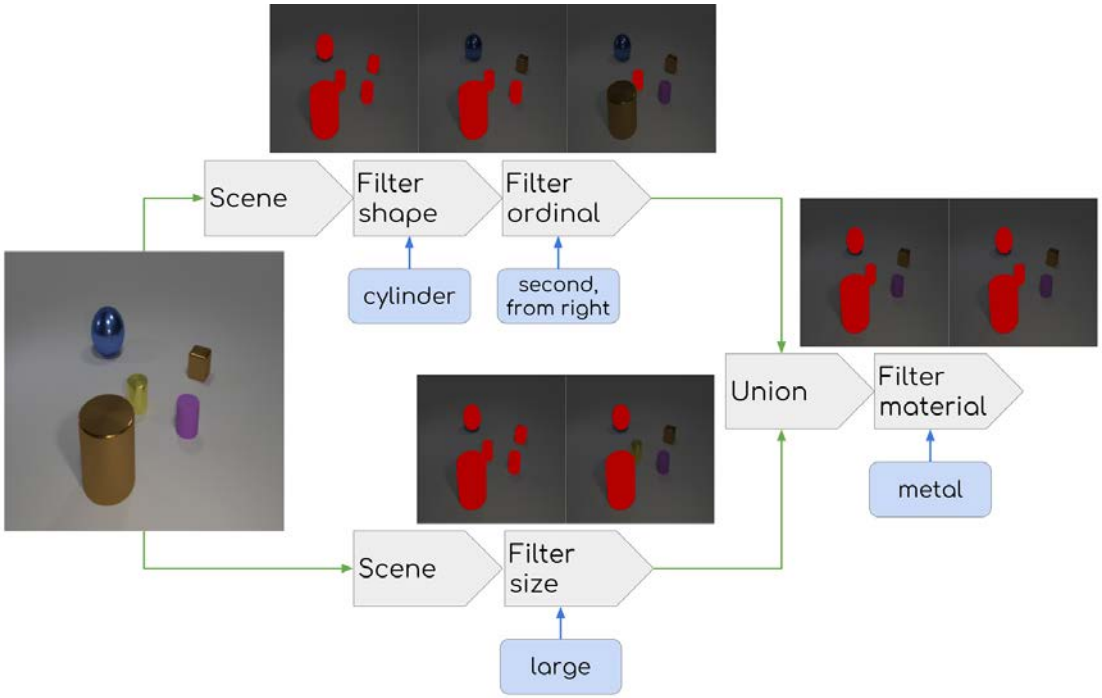
The first one of the big brown thing(s) from right

Step-by-step inspection: chain structure



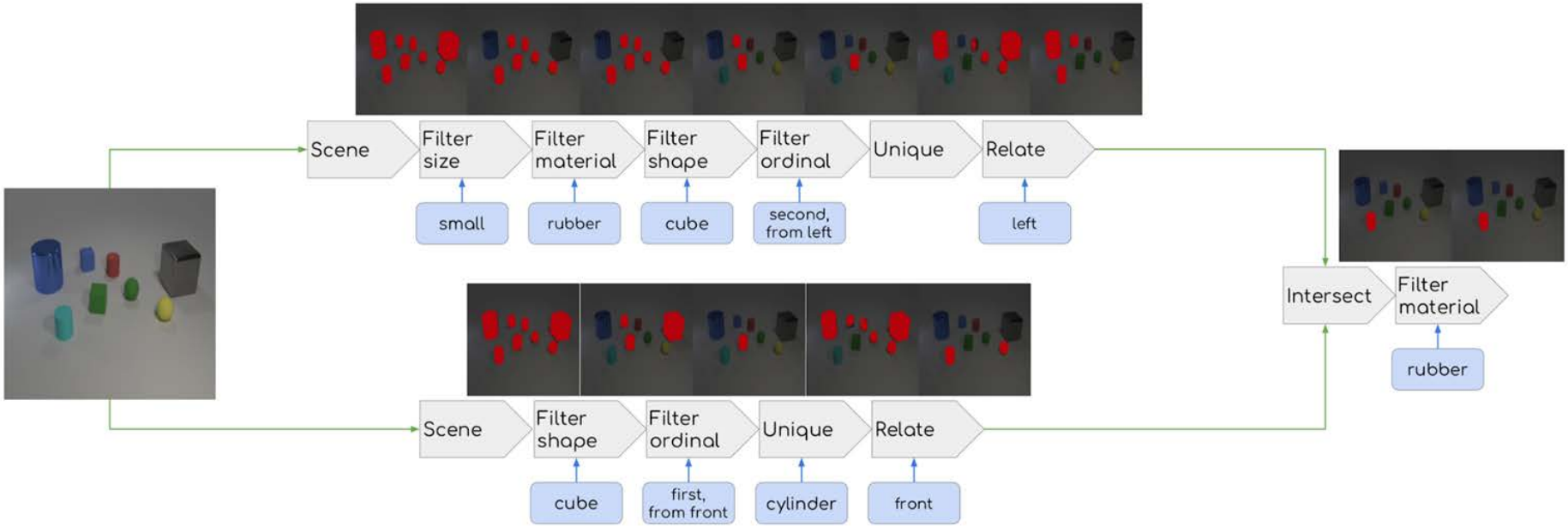
Any other things that are the same size as the partially visible rubber thing(s)

Step-by-step inspection: tree structure



The metallic things that are the second one of the cylinder(s) from right or large object(s)

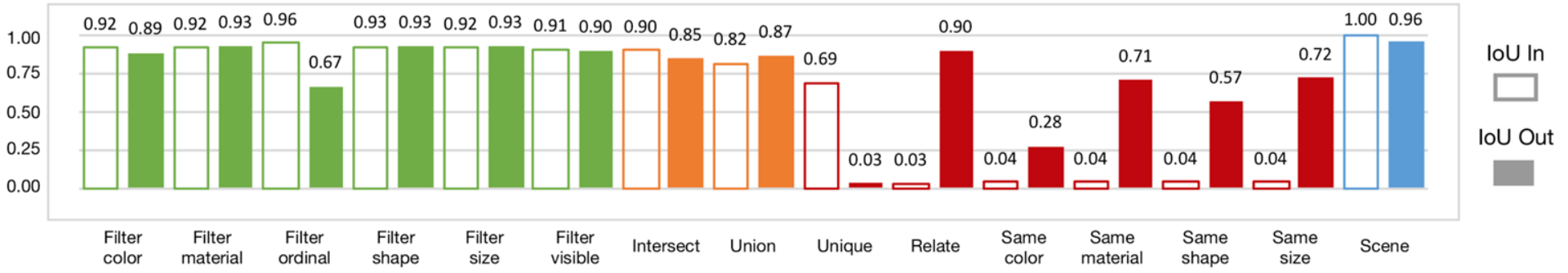
Step-by-step inspection: tree structure



The rubber object(s) that are in front of the first one of the cube(s) from front and left of the second one of the tiny rubber cube(s) from left

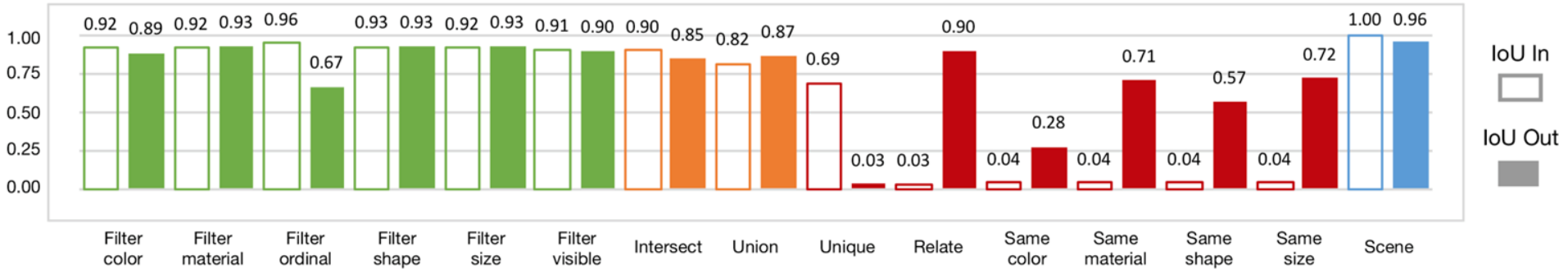
Step-by-step inspection: quantitative

- Every module is performing its intended job pretty well, except the red group.



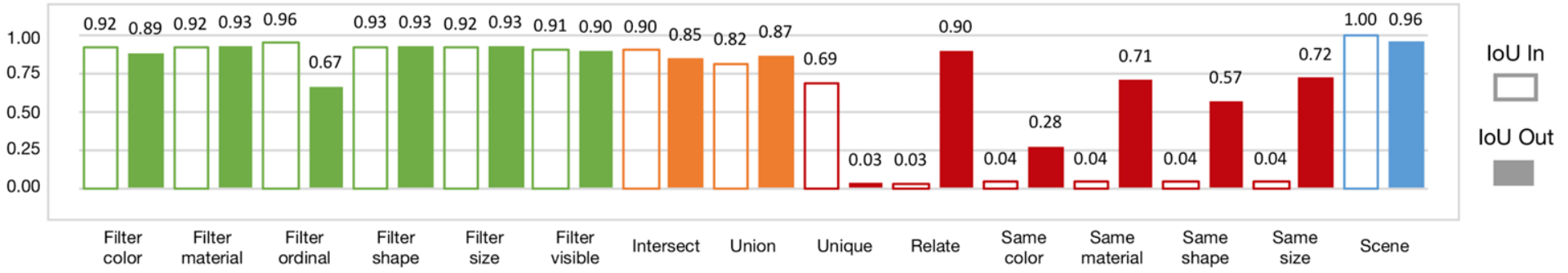
Step-by-step inspection: quantitative

- Every module is performing its intended job pretty well, except the red group.
- “Same” and “Relate” are the only modules that may come after “Unique”.
- IoU after “Unique” is very low; yet after one more module, the segmentation mask becomes normal again.



Step-by-step inspection: quantitative

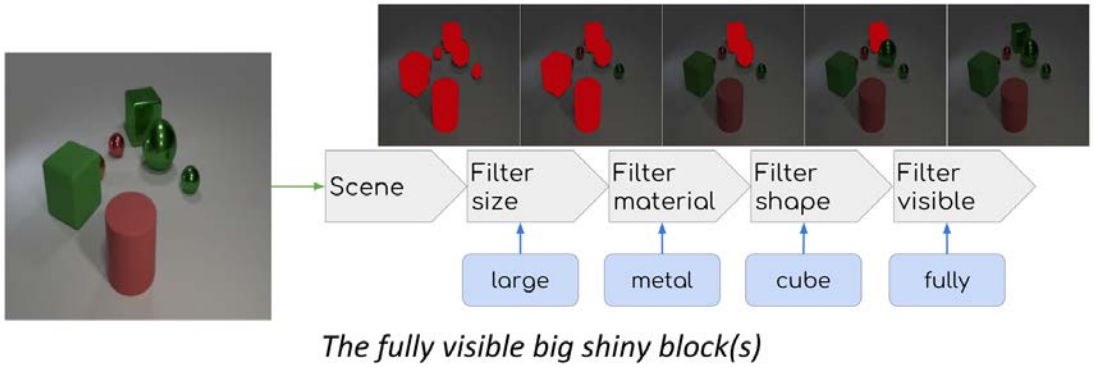
- Every module is performing its intended job pretty well, except the red group.
- “Same” and “Relate” are the only modules that may come after “Unique”.
- IoU after “Unique” is very low; yet after one more module, the segmentation mask becomes normal again.
- Perfect disentanglement for other modules, except learning some mechanism to treat “Unique” as the preprocessing step of “Same” and “Relate”.



False-premise referring expressions: quantitative

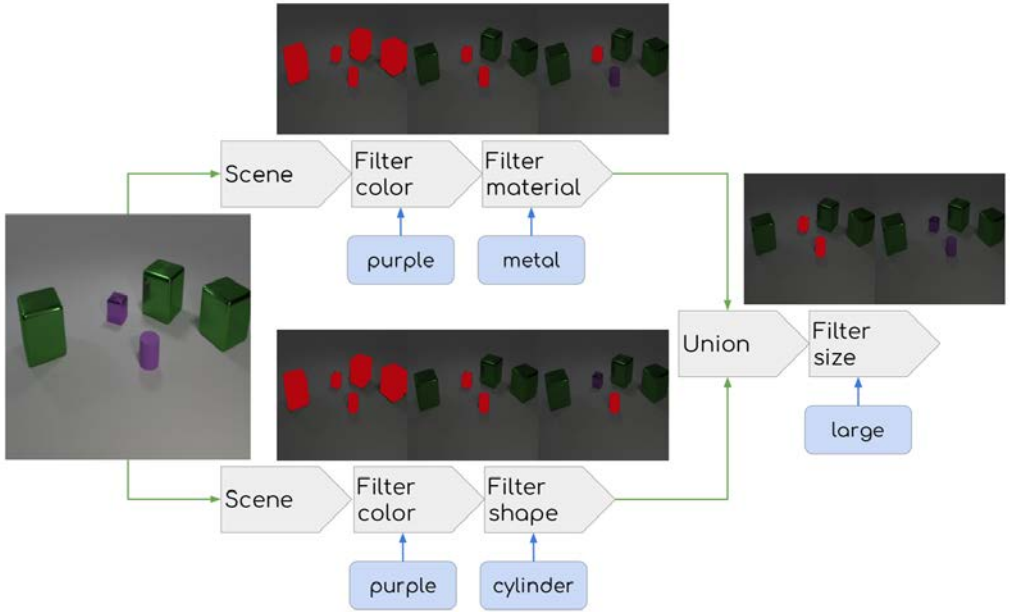
- We tested 10,000 referring expressions that refer to zero object at the end.
- Note that in every training example, at least one object is referred.
- IEP-Ref predicts:
  - 0 foreground pixel more than ¼ of the time
  - <= 8 foreground pixels more than ⅓ of the time

False-premise referring expressions: chain structure



The fully visible big shiny block(s)

False-premise referring expressions: tree structure



The big objects that are either purple metal object(s) or purple cylinder(s)

Conclusion

1. CLEVR-Ref+: Synthetic dataset
    - not-so-novel, but necessary for 3 and 4
2. IEP-Ref: Modular approach
    - not-so-novel, but necessary for 3 and 4
3. Detailed analysis of the strengths and weaknesses of existing models
    - useful; impossible on real image datasets
4. An easy technique to clearly reveal the entire visual reasoning process
    - novel; definitive and quantitative proof that neural modules are doing the intended jobs!

Everything has been released

- Paper:
  - https://arxiv.org/abs/1901.00850
- Dataset:
  - https://cs.jhu.edu/~cxliu/data/clevr_ref+_1.0.zip
  - https://cs.jhu.edu/~cxliu/data/clevr_ref+_cogent_1.0.zip
- Code:
  - https://github.com/ccvl/clevr-refplus-dataset-gen
  - https://github.com/ccvl/iep-ref

Thank you!