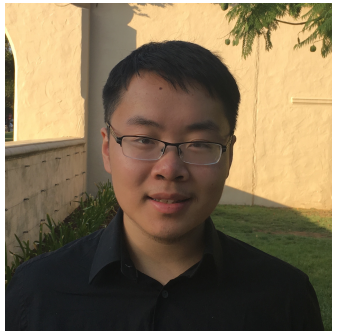
Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding

http://nsvqa.csail.mit.edu

NeurIPS 2018



Kexin Yi[1*]

Jiajun Wu[2*]

Chuang Gan[3]

Antonio Torralba[2]

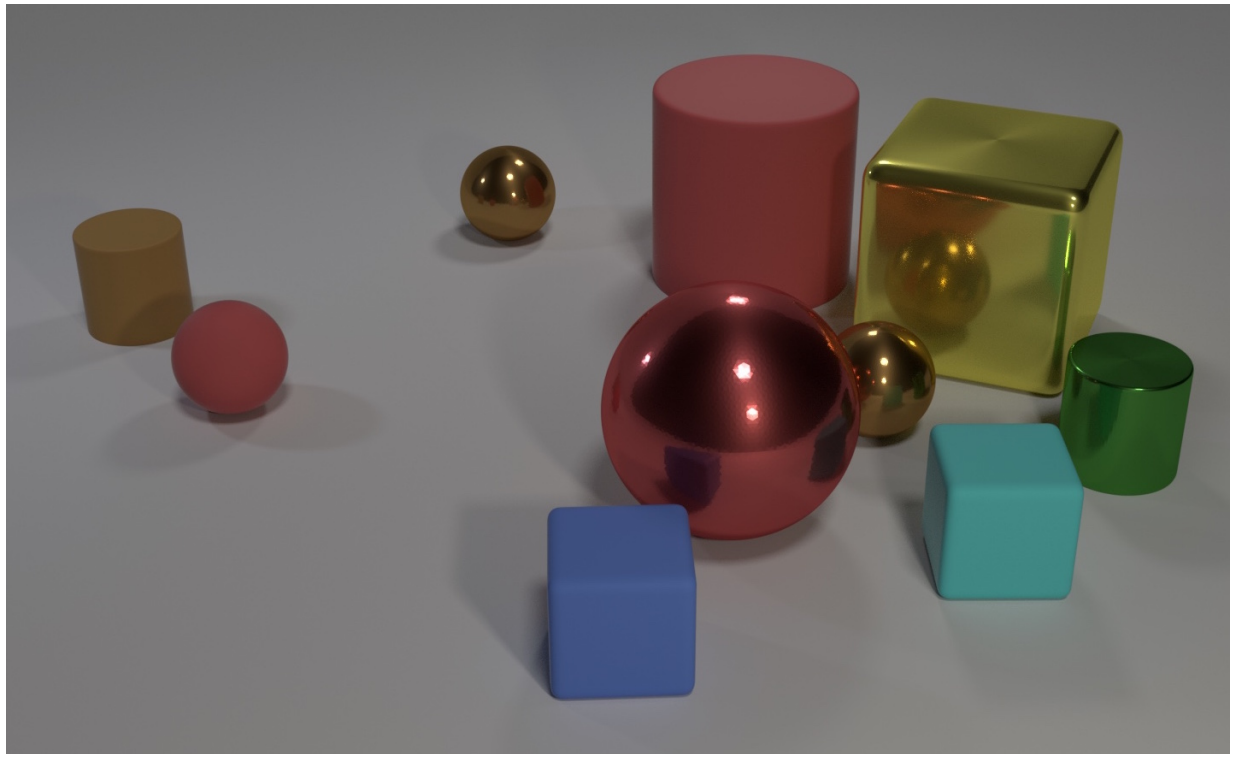
Pushmeet Kohli[4]

Joshua B. Tenenbaum[2]

[1]Harvard University [2]MIT CSAIL [3]MIT-IBM Watson AI Lab [4]DeepMind

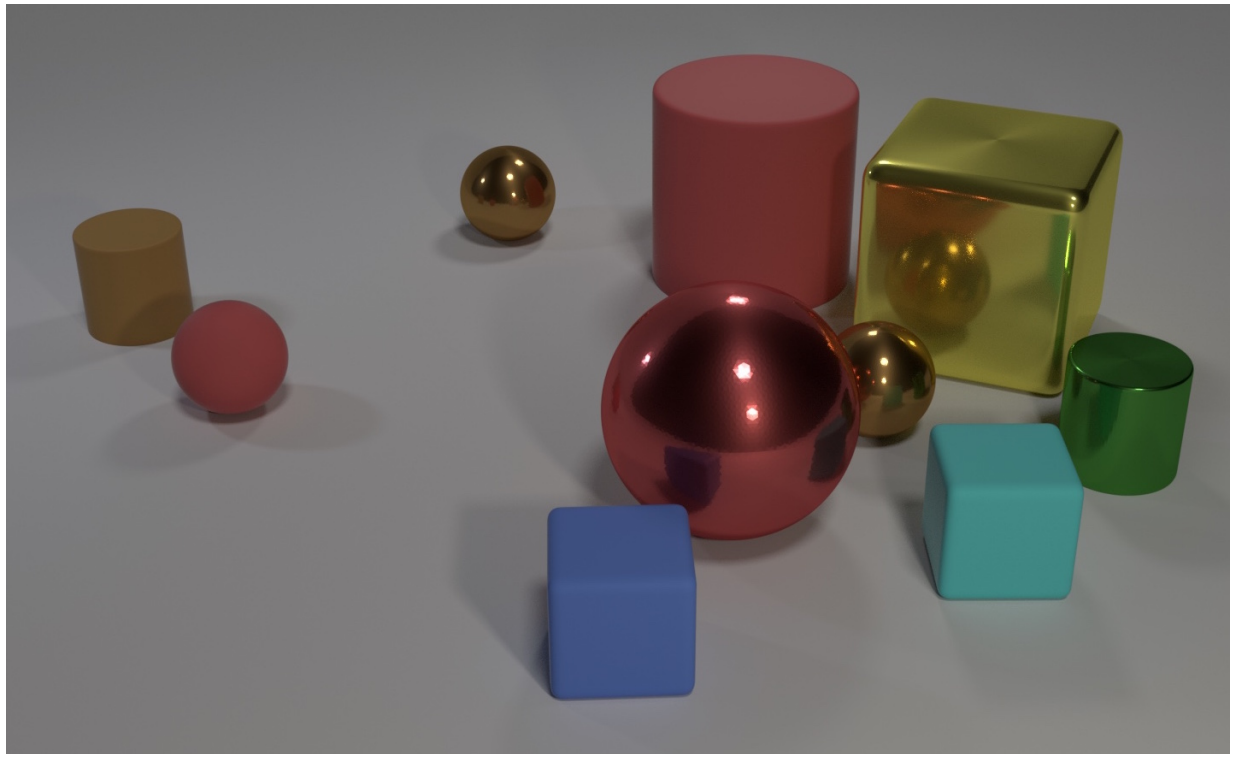
(* equal contributions)

Task: Visual Reasoning



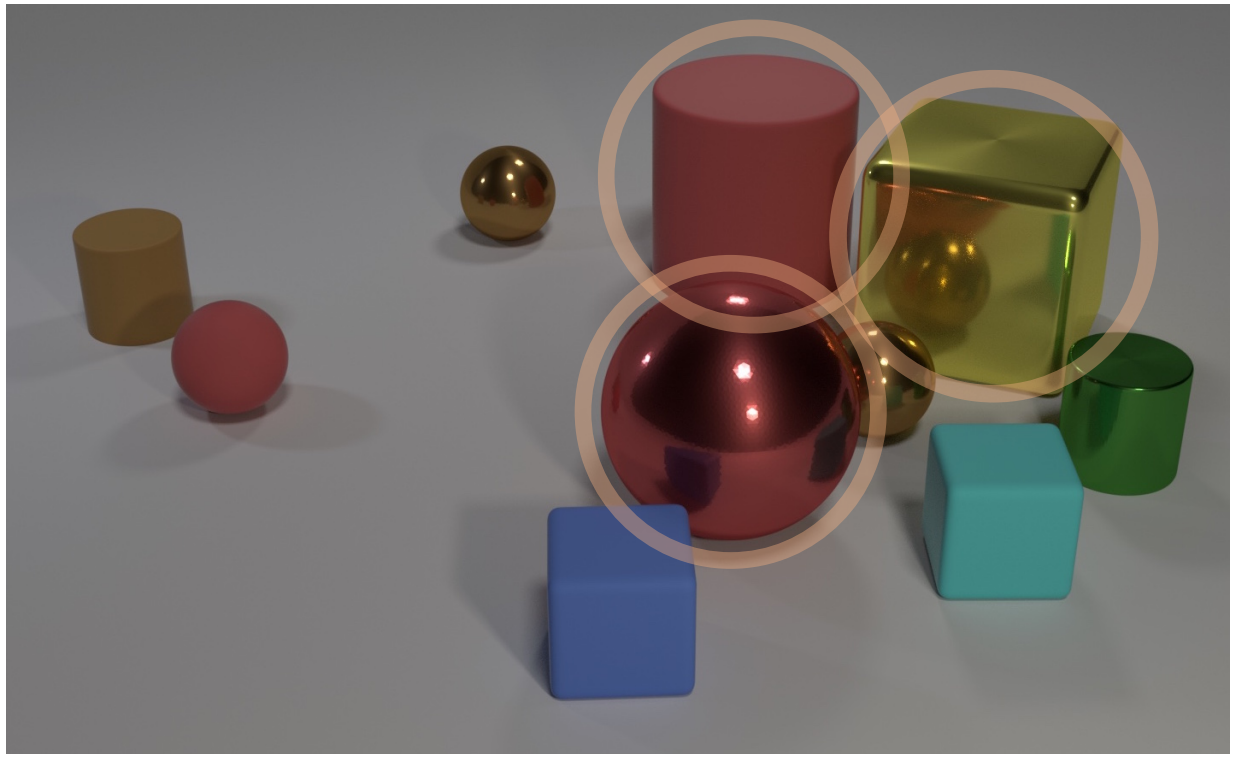
Question: *Are there an equal number of large things and metal spheres?*

Task: Visual Reasoning

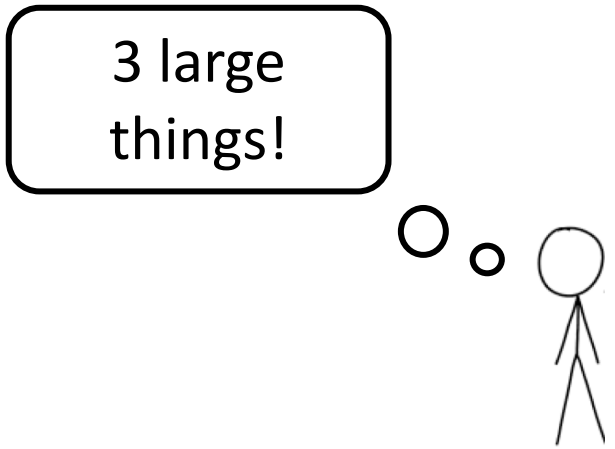


Question: *Are there an equal number of large things and metal spheres?*

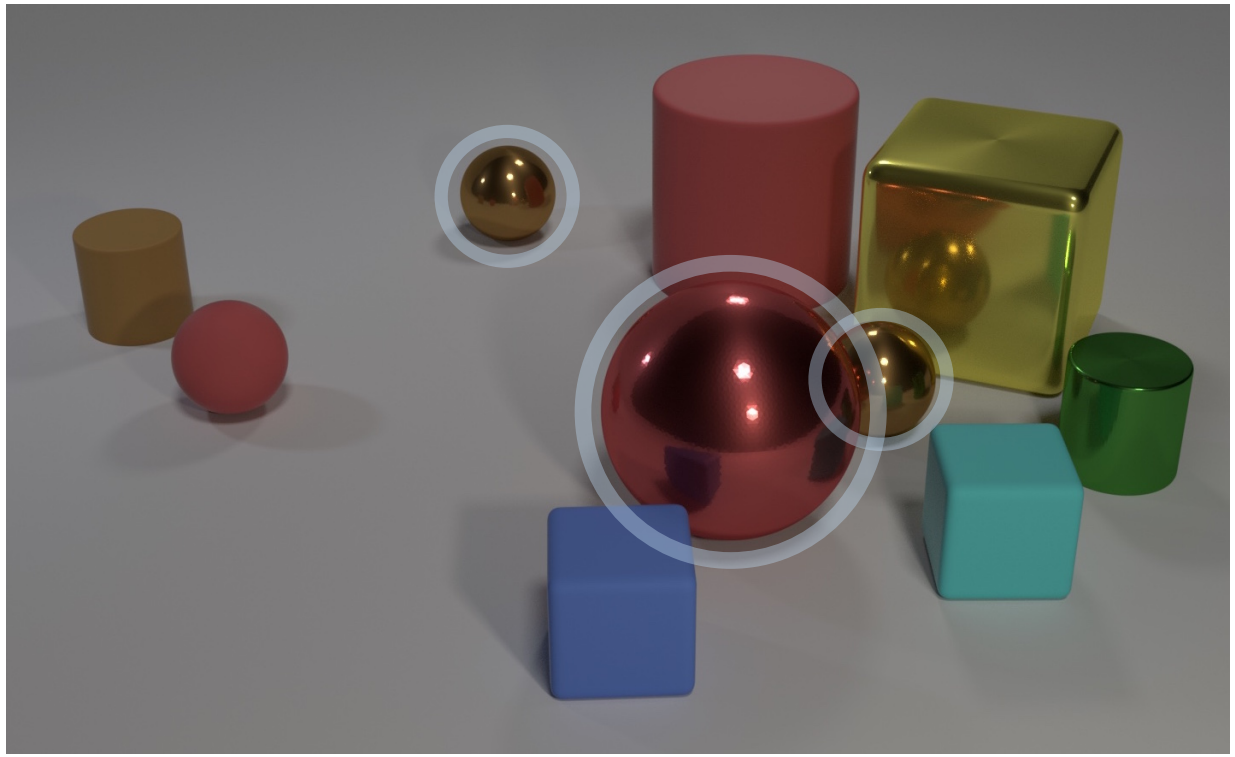
Task: Visual Reasoning



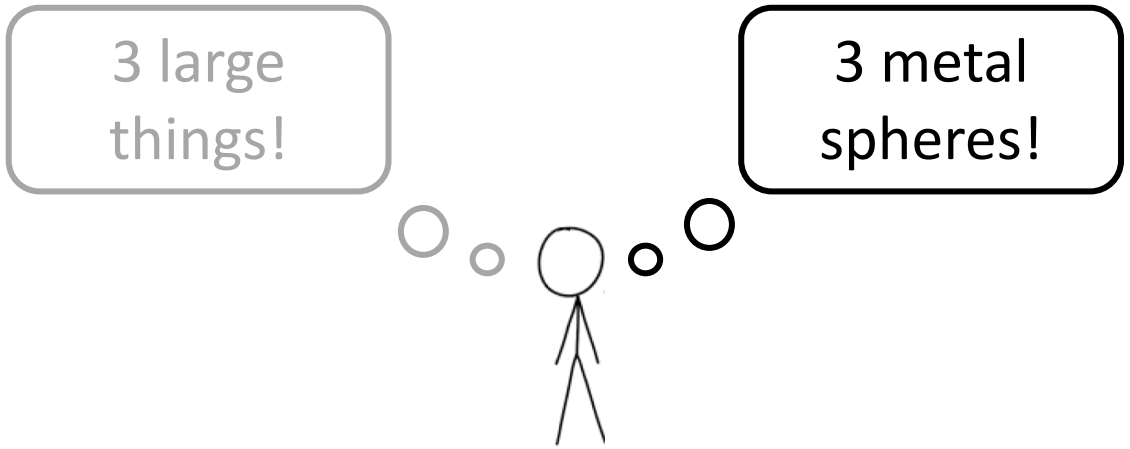
Question: *Are there an equal number of* ***large things*** *and metal spheres?*



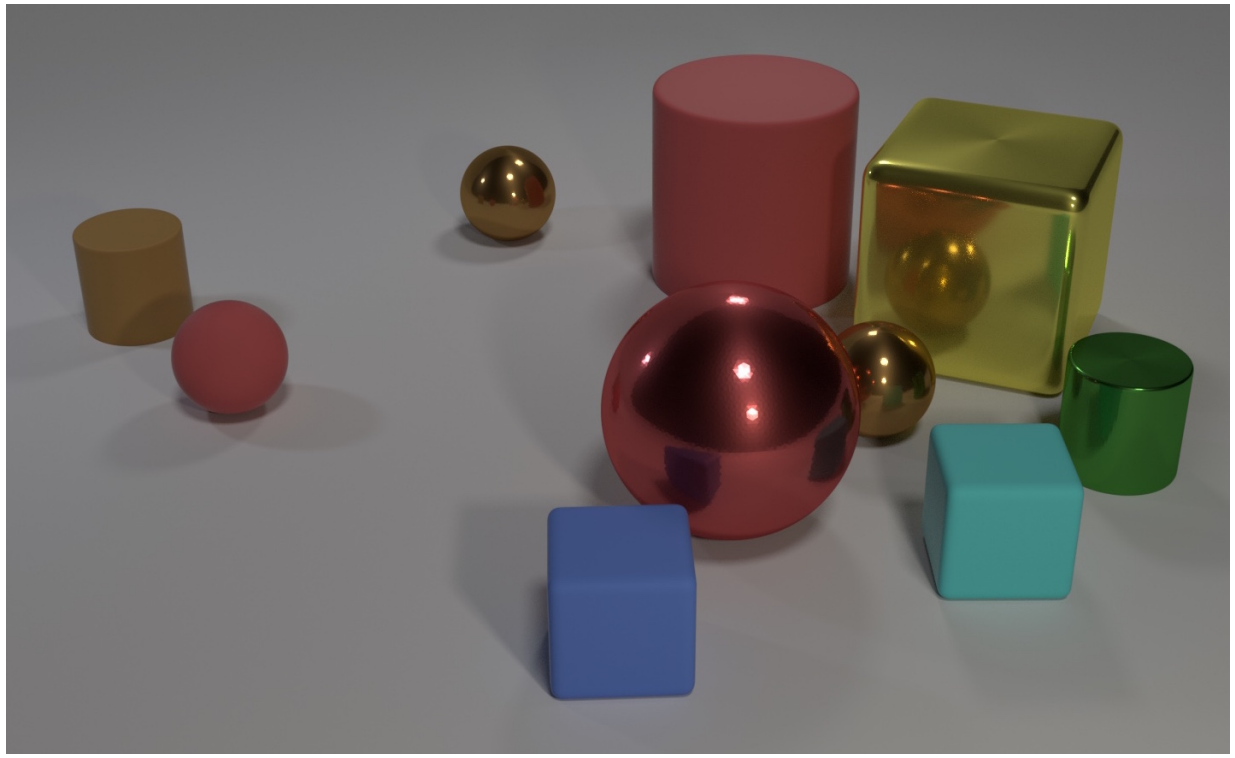
Task: Visual Reasoning



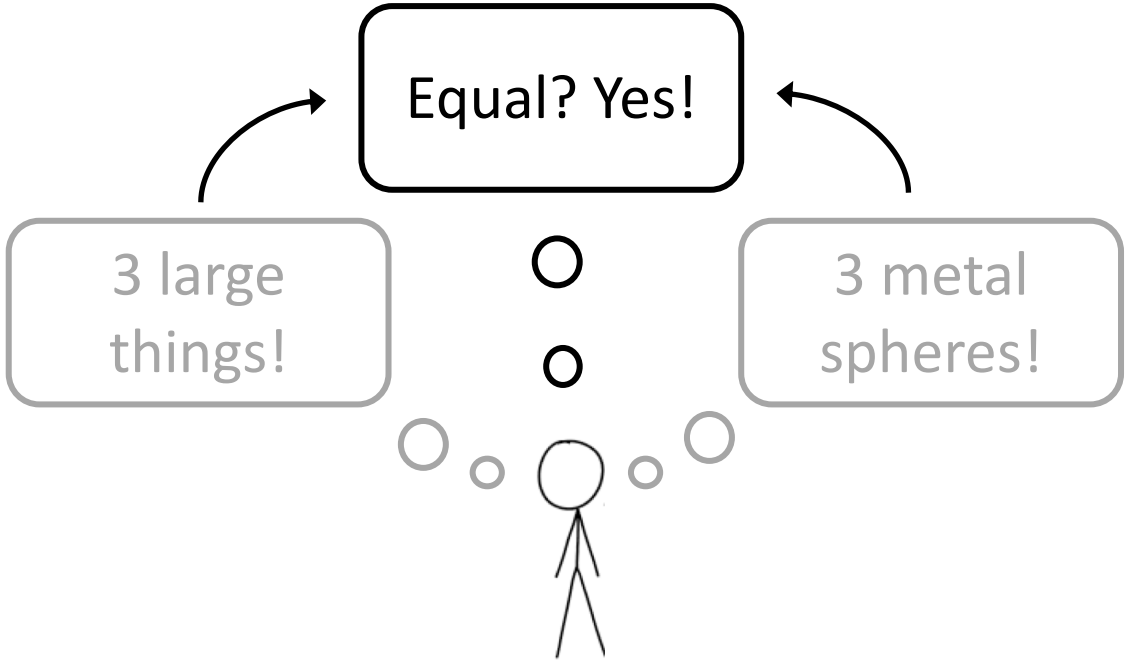
Question: *Are there an equal number of large things and* ***metal spheres****?*



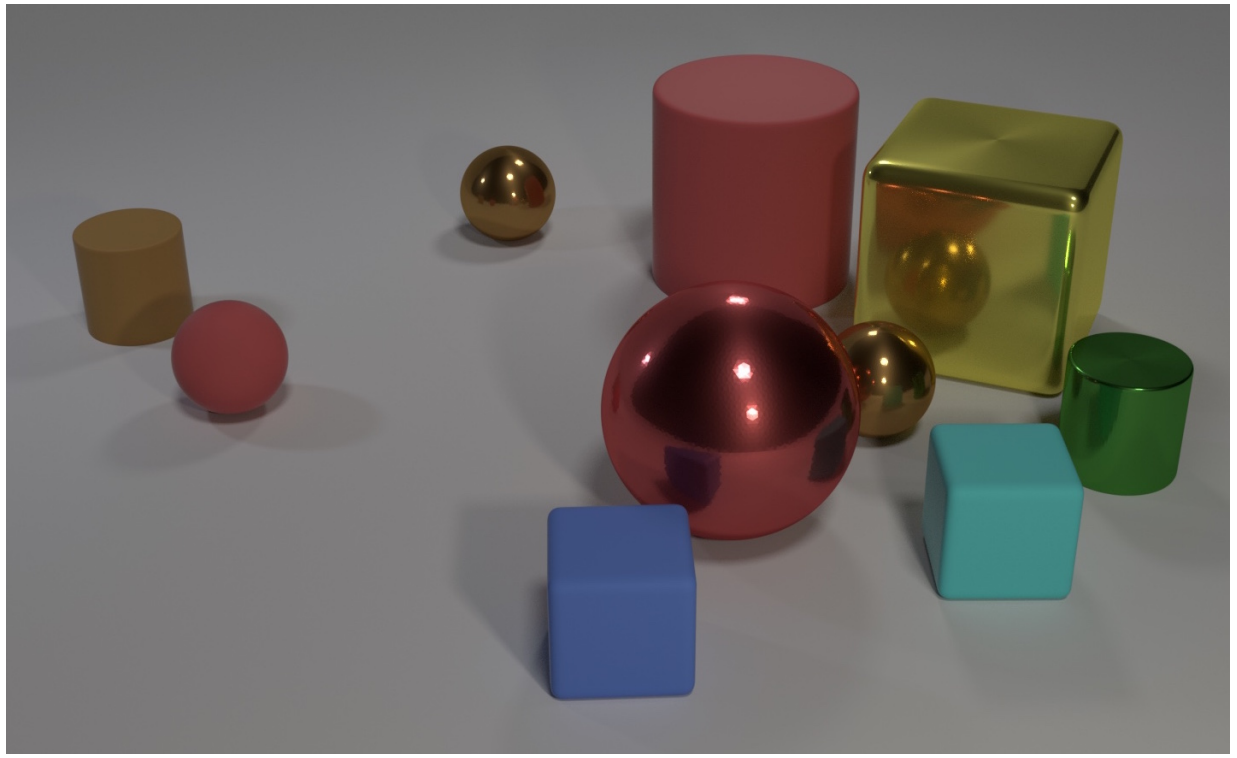
Task: Visual Reasoning



Question: *Are there an* ***equal number*** *of large things and metal spheres?*



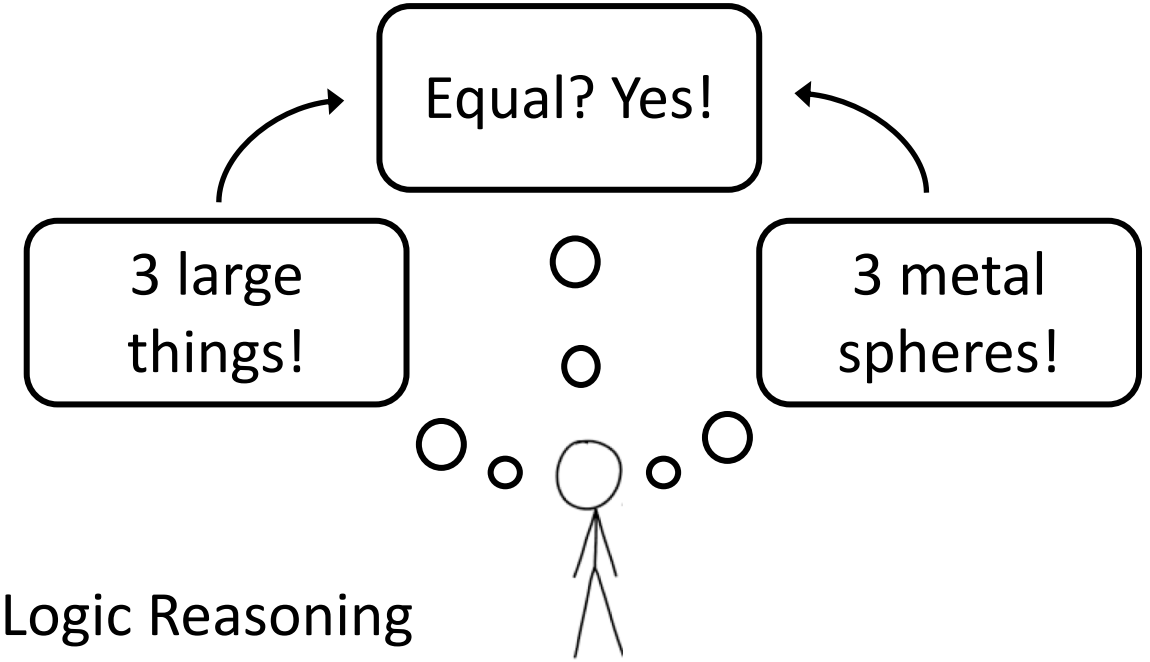
Task: Visual Reasoning



Visual Perception

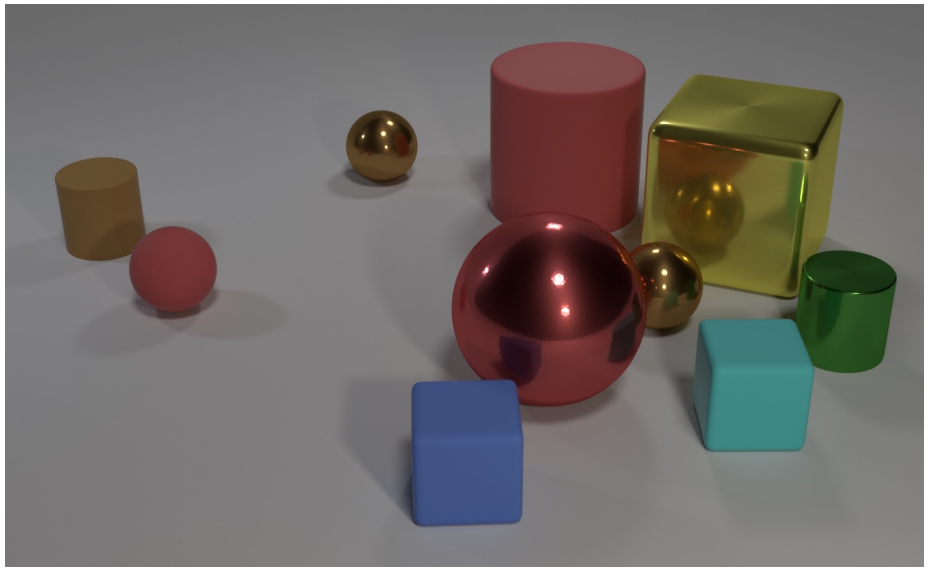
Question Understanding

Question: *Are there an equal number of large things and metal spheres?*



Logic Reasoning

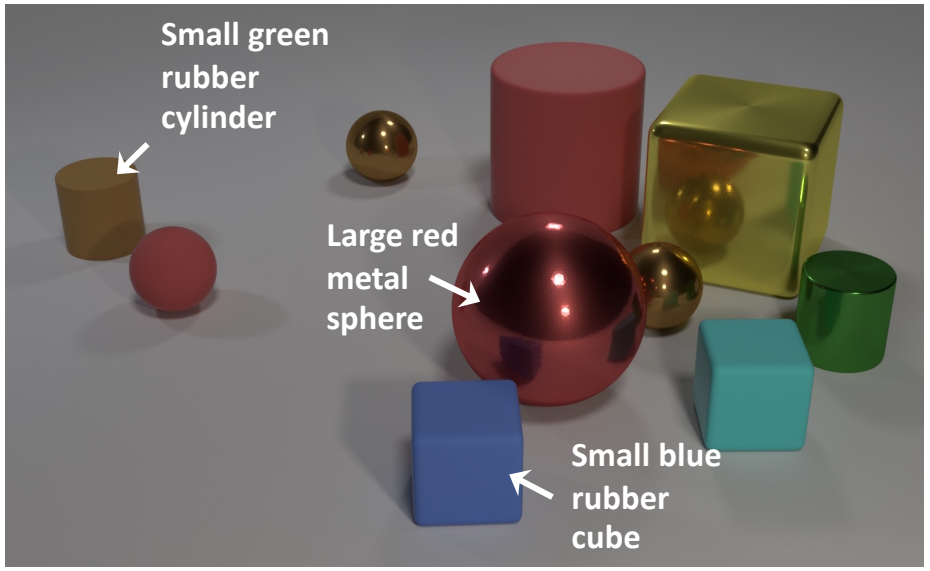
CLEVR Dataset



[Johnson et al. CVPR 2017]

CLEVR Dataset

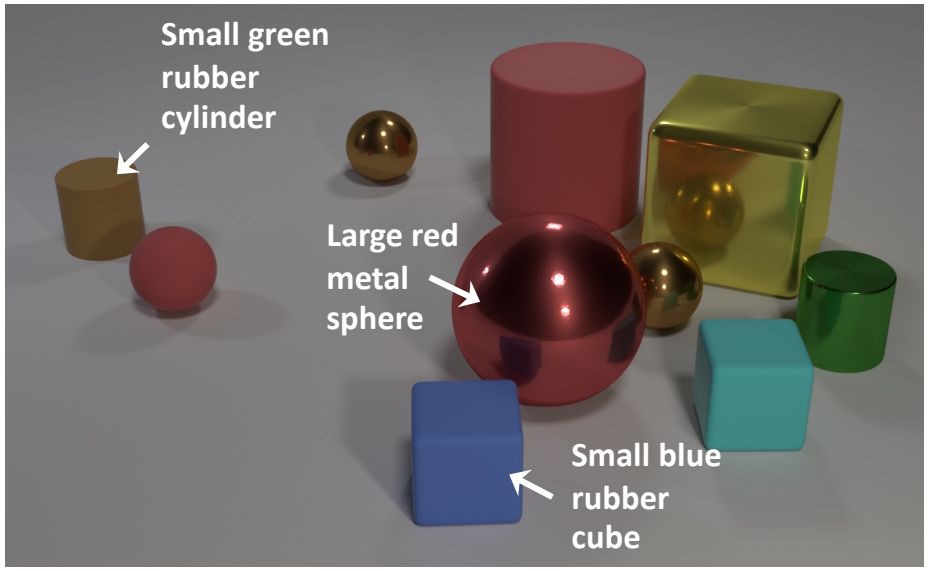
- Synthetic images of shapes with compositional attributes



[Johnson et al. CVPR 2017]

CLEVR Dataset

- Synthetic images of shapes with compositional attributes
- Machine generated questions paired with programs



Question: *Are there an equal number of large things and metal spheres?*

Program: equal_number(count(filter_size(Scene, Large)), count(filter_material(filter_shape(Scene, Sphere), Metal)))

Answer: *Yes*

[Johnson et al. CVPR 2017]

Neural-Symbolic Visual Question Answering (NS-VQA)

Mask R-CNN

CNN

ID	Size	Shape	Material	Color	x	y	z
1	Small	Cube	Metal	Purple	-0.45	-1.10	0.35
2	Large	Cube	Metal	Green	3.83	-0.04	0.70
3	Large	Cube	Metal	Green	-3.20	0.63	0.70
4	Small	Cylinder	Rubber	Purple	0.75	1.31	0.35
5	Large	Cube	Metal	Green	1.58	-1.60	0.70

I. *Neural* Scene Parsing

II. *Neural* Question Parsing

How many cubes that are behind the cylinder are large?

LSTM Encoder

LSTM → `1. filter_shape(scene, cylinder)`

LSTM → `2. relate(behind)`

LSTM → `3. filter_shape(scene, cube)`

LSTM → `4. filter_size(scene, large)`

LSTM → `5. count(scene)`

III. *Symbolic* Program Execution

```
1. filter_cylinder
2. relate_behind
```

ID	Size	Shape	…
1	Small	Cube	…
2	Large	Cube	…
3	Large	Cube	…
5	Large	Cube	…

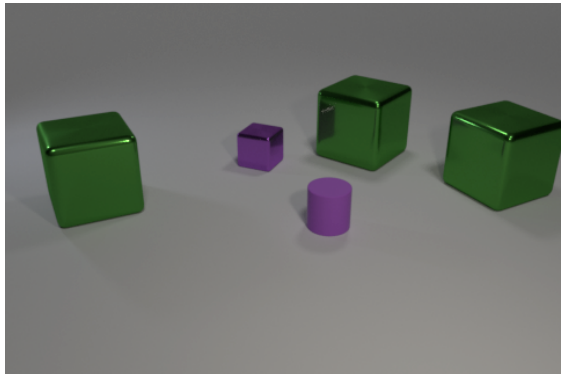
```
3. filter_cube
4. filter_large
```

ID	Size	…
2	Large	…
3	Large	…
5	Large	…

```
5. count
```

Answer: 3

Neural-Symbolic Visual Question Answering (NS-VQA)



I. *Neural* Scene Parsing

II. *Neural* Question Parsing

How many cubes that are behind the cylinder are large? →

LSTM Encoder

LSTM → `1. filter_shape(scene, cylinder)`

LSTM → `2. relate(behind)`

LSTM → `3. filter_shape(scene, cube)`

LSTM → `4. filter_size(scene, large)`

LSTM → `5. count(scene)`

III. *Symbolic* Program Execution

`1. filter_cylinder`
`2. relate_behind`

ID	Size	Shape	…
1	Small	Cube	…
2	Large	Cube	…
3	Large	Cube	…
5	Large	Cube	…

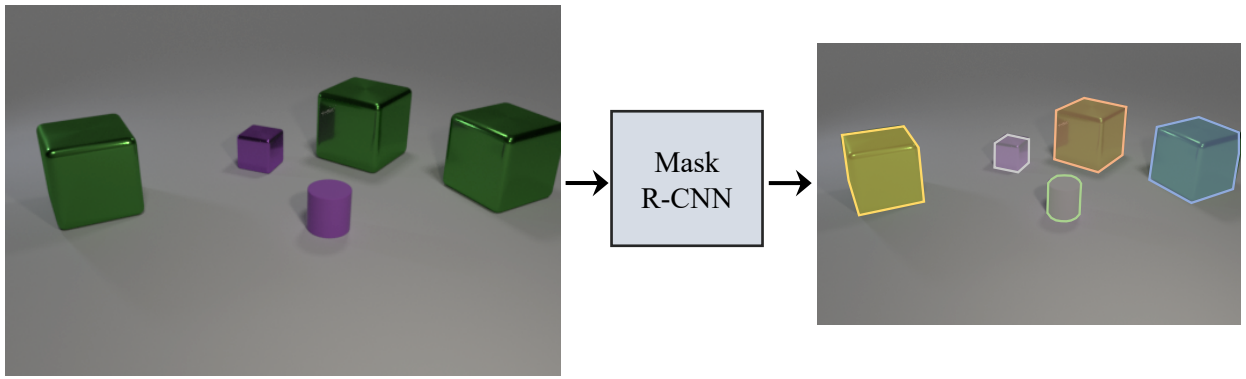
`3. filter_cube`
`4. filter_large`

ID	Size	…
2	Large	…
3	Large	…
5	Large	…

`5. count`

Answer: 3

Neural-Symbolic Visual Question Answering (NS-VQA)



I. *Neural* Scene Parsing

II. *Neural* Question Parsing

How many cubes that are behind the cylinder are large? → LSTM Encoder

LSTM → `1. filter_shape(scene, cylinder)`
LSTM → `2. relate(behind)`
LSTM → `3. filter_shape(scene, cube)`
LSTM → `4. filter_size(scene, large)`
LSTM → `5. count(scene)`

III. *Symbolic* Program Execution

`1. filter_cylinder`
`2. relate_behind`

ID	Size	Shape	…
1	Small	Cube	…
2	Large	Cube	…
3	Large	Cube	…
5	Large	Cube	…

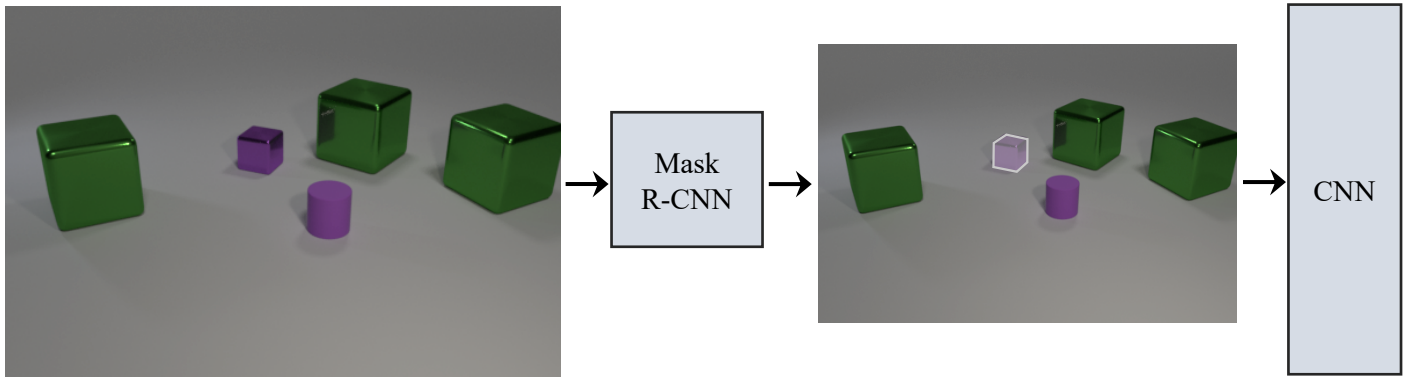
`3. filter_cube`
`4. filter_large`

ID	Size	…
2	Large	…
3	Large	…
5	Large	…

`5. count`

Answer: 3

Neural-Symbolic Visual Question Answering (NS-VQA)



ID	Size	Shape	Material	Color	x	y	z
1	Small	Cube	Metal	Purple	-0.45	-1.10	0.35

I. *Neural* Scene Parsing

II. *Neural* Question Parsing

How many cubes that are behind the cylinder are large? → LSTM Encoder

LSTM → `1. filter_shape(scene, cylinder)`
LSTM → `2. relate(behind)`
LSTM → `3. filter_shape(scene, cube)`
LSTM → `4. filter_size(scene, large)`
LSTM → `5. count(scene)`

III. *Symbolic* Program Execution

`1. filter_cylinder`
`2. relate_behind`

ID	Size	Shape	…
1	Small	Cube	…
2	Large	Cube	…
3	Large	Cube	…
5	Large	Cube	…

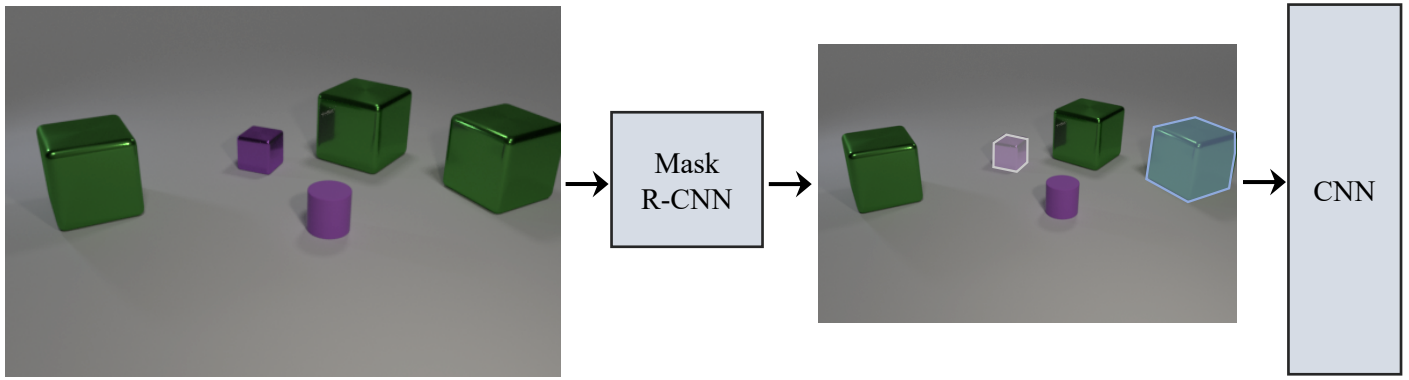
`3. filter_cube`
`4. filter_large`

ID	Size	…
2	Large	…
3	Large	…
5	Large	…

`5. count`

Answer: 3

Neural-Symbolic Visual Question Answering (NS-VQA)



ID	Size	Shape	Material	Color	x	y	z
1	Small	Cube	Metal	Purple	-0.45	-1.10	0.35
2	Large	Cube	Metal	Green	3.83	-0.04	0.70

I. *Neural* Scene Parsing

II. *Neural* Question Parsing

How many cubes that are behind the cylinder are large?

III. *Symbolic* Program Execution

1. filter_cylinder
2. relate_behind

ID	Size	Shape	…
1	Small	Cube	…
2	Large	Cube	…
3	Large	Cube	…
5	Large	Cube	…

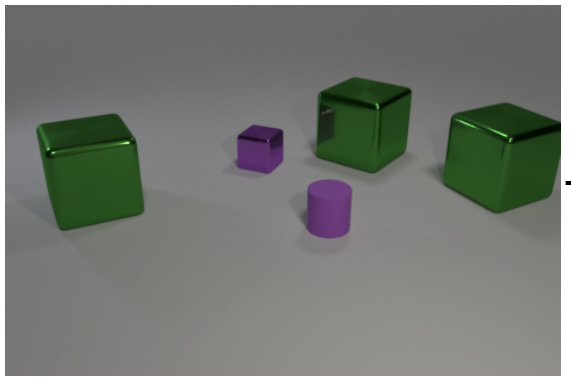
3. filter_cube
4. filter_large

ID	Size	…
2	Large	…
3	Large	…
5	Large	…

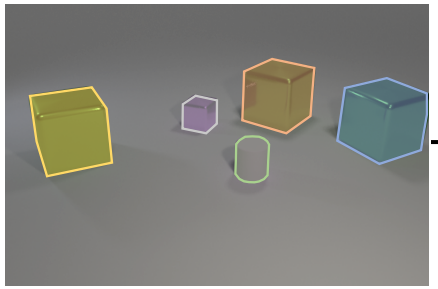
5. count

Answer: 3

Neural-Symbolic Visual Question Answering (NS-VQA)



Mask R-CNN



CNN

ID	Size	Shape	Material	Color	x	y	z
1	Small	Cube	Metal	Purple	-0.45	-1.10	0.35
2	Large	Cube	Metal	Green	3.83	-0.04	0.70
3	Large	Cube	Metal	Green	-3.20	0.63	0.70
4	Small	Cylinder	Rubber	Purple	0.75	1.31	0.35
5	Large	Cube	Metal	Green	1.58	-1.60	0.70

I. *Neural* Scene Parsing

II. *Neural* Question Parsing

How many cubes that are behind the cylinder are large?

LSTM Encoder

LSTM → `1. filter_shape(scene, cylinder)`

LSTM → `2. relate(behind)`

LSTM → `3. filter_shape(scene, cube)`

LSTM → `4. filter_size(scene, large)`

LSTM → `5. count(scene)`

III. *Symbolic* Program Execution

`1. filter_cylinder`
`2. relate_behind`

ID	Size	Shape	…
1	Small	Cube	…
2	Large	Cube	…
3	Large	Cube	…
5	Large	Cube	…

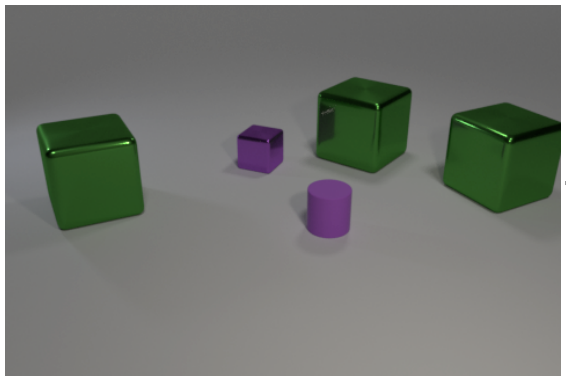
`3. filter_cube`
`4. filter_large`

ID	Size	…
2	Large	…
3	Large	…
5	Large	…

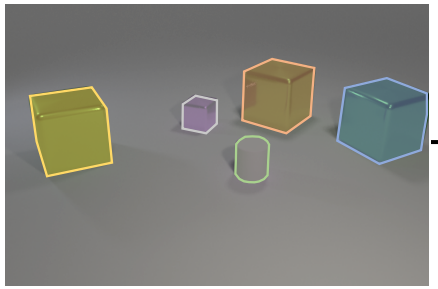
`5. count`

Answer: 3

Neural-Symbolic Visual Question Answering (NS-VQA)



Mask R-CNN



CNN

ID	Size	Shape	Material	Color	x	y	z
1	Small	Cube	Metal	Purple	-0.45	-1.10	0.35
2	Large	Cube	Metal	Green	3.83	-0.04	0.70
3	Large	Cube	Metal	Green	-3.20	0.63	0.70
4	Small	Cylinder	Rubber	Purple	0.75	1.31	0.35
5	Large	Cube	Metal	Green	1.58	-1.60	0.70

I. *Neural* Scene Parsing

II. *Neural* Question Parsing

How many cubes that are behind the cylinder are large?

III. *Symbolic* Program Execution

```
1. filter_cylinder
2. relate_behind
```

ID	Size	Shape	…
1	Small	Cube	…
2	Large	Cube	…
3	Large	Cube	…
5	Large	Cube	…

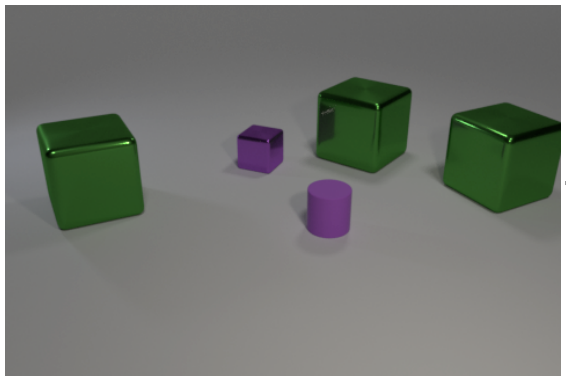
```
3. filter_cube
4. filter_large
```

ID	Size	…
2	Large	…
3	Large	…
5	Large	…

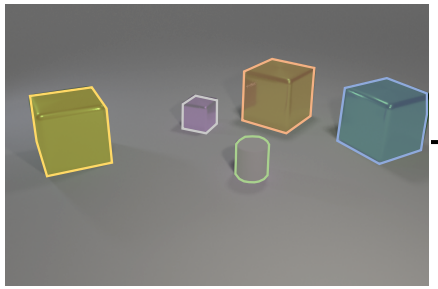
```
5. count
```

Answer: 3

Neural-Symbolic Visual Question Answering (NS-VQA)



Mask R-CNN



CNN

ID	Size	Shape	Material	Color	x	y	z
1	Small	Cube	Metal	Purple	-0.45	-1.10	0.35
2	Large	Cube	Metal	Green	3.83	-0.04	0.70
3	Large	Cube	Metal	Green	-3.20	0.63	0.70
4	Small	Cylinder	Rubber	Purple	0.75	1.31	0.35
5	Large	Cube	Metal	Green	1.58	-1.60	0.70

I. *Neural* Scene Parsing

II. *Neural* Question Parsing

How many cubes that are behind the cylinder are large?

LSTM Encoder

LSTM

LSTM

LSTM

LSTM

LSTM

III. *Symbolic* Program Execution

```
1. filter_cylinder
2. relate_behind
```

ID	Size	Shape	…
1	Small	Cube	…
2	Large	Cube	…
3	Large	Cube	…
5	Large	Cube	…

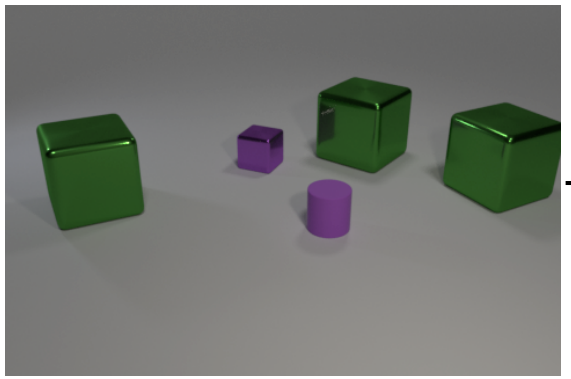
```
3. filter_cube
4. filter_large
```

ID	Size	…
2	Large	…
3	Large	…
5	Large	…

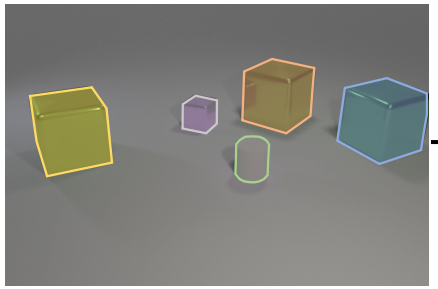
```
5. count
```

Answer: 3

Neural-Symbolic Visual Question Answering (NS-VQA)



Mask R-CNN



CNN

ID	Size	Shape	Material	Color	x	y	z
1	Small	Cube	Metal	Purple	-0.45	-1.10	0.35
2	Large	Cube	Metal	Green	3.83	-0.04	0.70
3	Large	Cube	Metal	Green	-3.20	0.63	0.70
4	Small	Cylinder	Rubber	Purple	0.75	1.31	0.35
5	Large	Cube	Metal	Green	1.58	-1.60	0.70

I. *Neural* Scene Parsing

II. *Neural* Question Parsing

How many cubes that are behind the cylinder are large?

LSTM Encoder

LSTM → `1. filter_shape(scene, cylinder)`

LSTM → `2. relate(behind)`

LSTM → `3. filter_shape(scene, cube)`

LSTM → `4. filter_size(scene, large)`

LSTM → `5. count(scene)`

III. *Symbolic* Program Execution

`1. filter_cylinder`
`2. relate_behind`

ID	Size	Shape	…
1	Small	Cube	…
2	Large	Cube	…
3	Large	Cube	…
5	Large	Cube	…

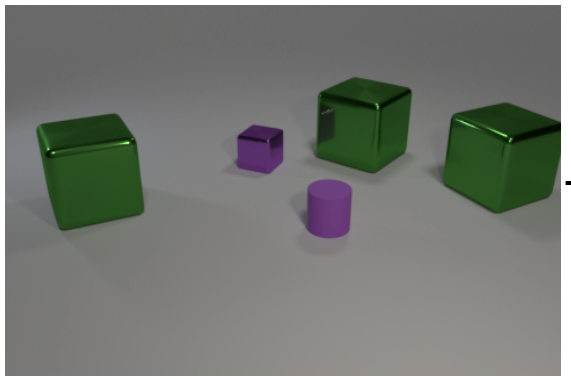
`3. filter_cube`
`4. filter_large`

ID	Size	…
2	Large	…
3	Large	…
5	Large	…

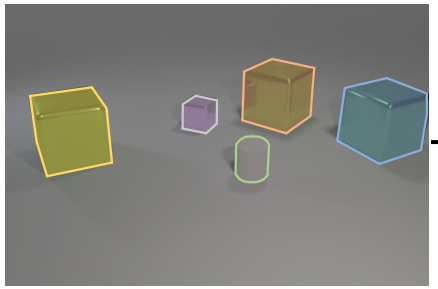
`5. count`

Answer: 3

Neural-Symbolic Visual Question Answering (NS-VQA)



Mask R-CNN



CNN

ID	Size	Shape	Material	Color	x	y	z
1	Small	Cube	Metal	Purple	-0.45	-1.10	0.35
2	Large	Cube	Metal	Green	3.83	-0.04	0.70
3	Large	Cube	Metal	Green	-3.20	0.63	0.70
4	Small	Cylinder	Rubber	Purple	0.75	1.31	0.35
5	Large	Cube	Metal	Green	1.58	-1.60	0.70

I. *Neural* Scene Parsing

II. *Neural* Question Parsing

How many cubes that are behind the cylinder are large?

LSTM Encoder

LSTM → `1. filter_shape(scene, cylinder)`

LSTM → `2. relate(behind)`

LSTM → `3. filter_shape(scene, cube)`

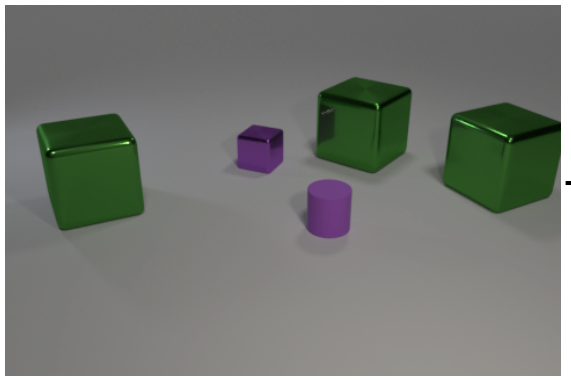
LSTM → `4. filter_size(scene, large)`

LSTM → `5. count(scene)`

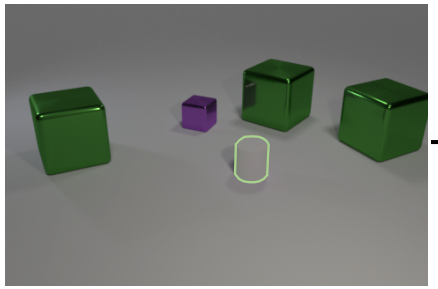
III. *Symbolic* Program Execution

```
1. filter_cylinder
2. relate_behind
3. filter_cube
4. filter_large
5. count
```

Neural-Symbolic Visual Question Answering (NS-VQA)



Mask R-CNN



CNN

ID	Size	Shape	Material	Color	x	y	z
1	Small	Cube	Metal	Purple	-0.45	-1.10	0.35
2	Large	Cube	Metal	Green	3.83	-0.04	0.70
3	Large	Cube	Metal	Green	-3.20	0.63	0.70
4	Small	Cylinder	Rubber	Purple	0.75	1.31	0.35
5	Large	Cube	Metal	Green	1.58	-1.60	0.70

I. *Neural* Scene Parsing

II. *Neural* Question Parsing

How many cubes that are behind the cylinder are large?

LSTM Encoder

LSTM → `1. filter_shape(scene, cylinder)`

LSTM → `2. relate(behind)`

LSTM → `3. filter_shape(scene, cube)`

LSTM → `4. filter_size(scene, large)`

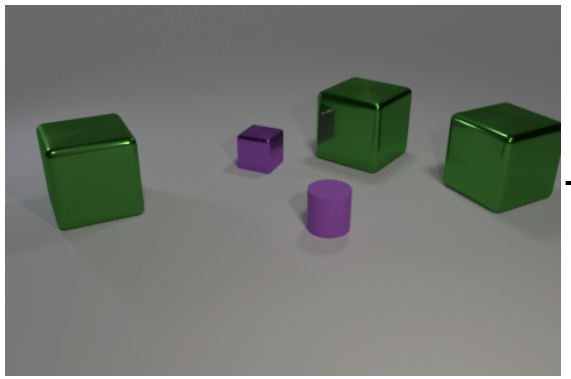
LSTM → `5. count(scene)`

III. *Symbolic* Program Execution

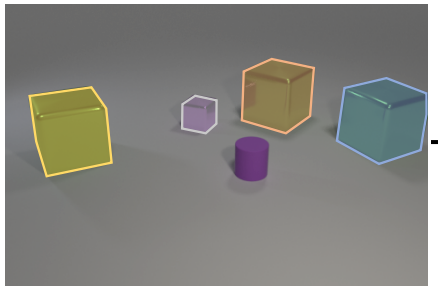
```
1. filter_cylinder
2. relate_behind
3. filter_cube
4. filter_large
5. count
```

ID	Size	Shape	Material	Color
4	Small	Cylinder	Rubber	Purple

Neural-Symbolic Visual Question Answering (NS-VQA)



Mask R-CNN



CNN

ID	Size	Shape	Material	Color	x	y	z
1	Small	Cube	Metal	Purple	-0.45	-1.10	0.35
2	Large	Cube	Metal	Green	3.83	-0.04	0.70
3	Large	Cube	Metal	Green	-3.20	0.63	0.70
4	Small	Cylinder	Rubber	Purple	0.75	1.31	0.35
5	Large	Cube	Metal	Green	1.58	-1.60	0.70

I. *Neural* Scene Parsing

II. *Neural* Question Parsing

How many cubes that are behind the cylinder are large?

LSTM Encoder

LSTM → `1. filter_shape(scene, cylinder)`

LSTM → `2. relate(behind)`

LSTM → `3. filter_shape(scene, cube)`

LSTM → `4. filter_size(scene, large)`

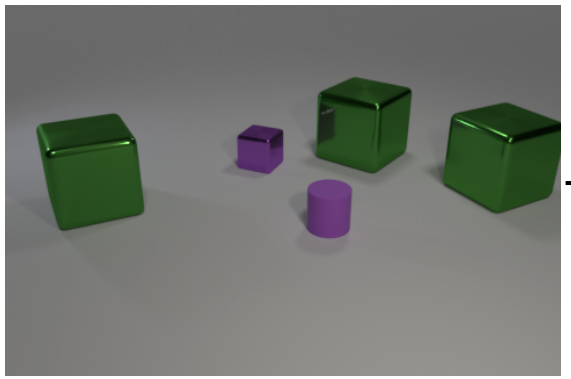
LSTM → `5. count(scene)`

III. *Symbolic* Program Execution

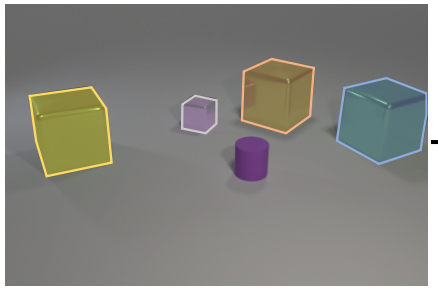
`1. filter_cylinder`
`2. relate_behind`
`3. filter_cube`
`4. filter_large`
`5. count`

ID	Size	Shape	Material	Color
1	Small	Cube	Metal	Purple
2	Large	Cube	Metal	Green
3	Large	Cube	Metal	Green
5	Large	Cube	Metal	Green

Neural-Symbolic Visual Question Answering (NS-VQA)



Mask R-CNN



CNN

ID	Size	Shape	Material	Color	x	y	z
1	Small	Cube	Metal	Purple	-0.45	-1.10	0.35
2	Large	Cube	Metal	Green	3.83	-0.04	0.70
3	Large	Cube	Metal	Green	-3.20	0.63	0.70
4	Small	Cylinder	Rubber	Purple	0.75	1.31	0.35
5	Large	Cube	Metal	Green	1.58	-1.60	0.70

I. *Neural* Scene Parsing

II. *Neural* Question Parsing

How many cubes that are behind the cylinder are large?

LSTM Encoder

LSTM → `1. filter_shape(scene, cylinder)`

LSTM → `2. relate(behind)`

LSTM → `3. filter_shape(scene, cube)`

LSTM → `4. filter_size(scene, large)`

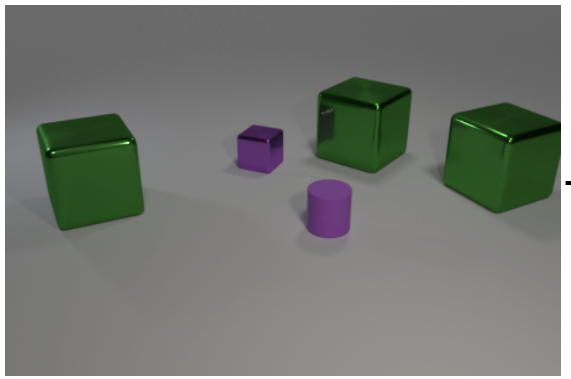
LSTM → `5. count(scene)`

III. *Symbolic* Program Execution

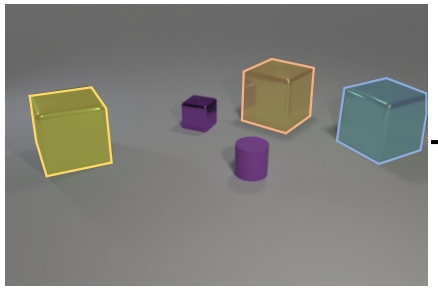
`1. filter_cylinder`
`2. relate_behind`
`3. filter_cube`
`4. filter_large`
`5. count`

ID	Size	Shape	Material	Color
1	Small	Cube	Metal	Purple
2	Large	Cube	Metal	Green
3	Large	Cube	Metal	Green
5	Large	Cube	Metal	Green

Neural-Symbolic Visual Question Answering (NS-VQA)



Mask R-CNN



CNN

ID	Size	Shape	Material	Color	x	y	z
1	Small	Cube	Metal	Purple	-0.45	-1.10	0.35
2	Large	Cube	Metal	Green	3.83	-0.04	0.70
3	Large	Cube	Metal	Green	-3.20	0.63	0.70
4	Small	Cylinder	Rubber	Purple	0.75	1.31	0.35
5	Large	Cube	Metal	Green	1.58	-1.60	0.70

I. *Neural* Scene Parsing

II. *Neural* Question Parsing

How many cubes that are behind the cylinder are large?

LSTM Encoder

LSTM → `1. filter_shape(scene, cylinder)`

LSTM → `2. relate(behind)`

LSTM → `3. filter_shape(scene, cube)`

LSTM → `4. filter_size(scene, large)`

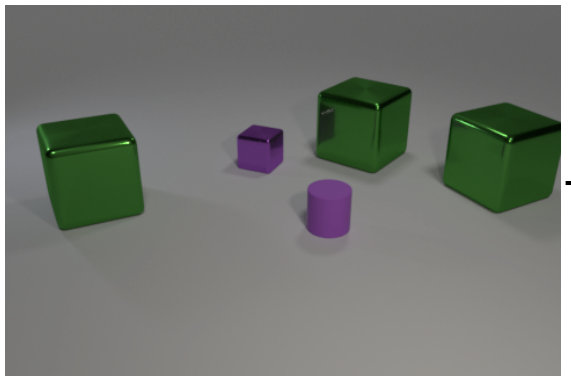
LSTM → `5. count(scene)`

III. *Symbolic* Program Execution

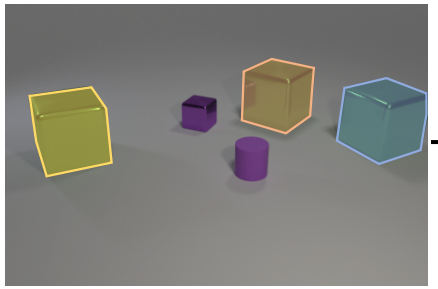
`1. filter_cylinder`
`2. relate_behind`
`3. filter_cube`
`4. filter_large`
`5. count`

ID	Size	Shape	Material	Color
2	Large	Cube	Metal	Green
3	Large	Cube	Metal	Green
5	Large	Cube	Metal	Green

Neural-Symbolic Visual Question Answering (NS-VQA)



Mask R-CNN



CNN

ID	Size	Shape	Material	Color	x	y	z
1	Small	Cube	Metal	Purple	-0.45	-1.10	0.35
2	Large	Cube	Metal	Green	3.83	-0.04	0.70
3	Large	Cube	Metal	Green	-3.20	0.63	0.70
4	Small	Cylinder	Rubber	Purple	0.75	1.31	0.35
5	Large	Cube	Metal	Green	1.58	-1.60	0.70

I. *Neural* Scene Parsing

II. *Neural* Question Parsing

How many cubes that are behind the cylinder are large?

LSTM Encoder

LSTM → `1. filter_shape(scene, cylinder)`

LSTM → `2. relate(behind)`

LSTM → `3. filter_shape(scene, cube)`

LSTM → `4. filter_size(scene, large)`

LSTM → `5. count(scene)`

III. *Symbolic* Program Execution

```
1. filter_cylinder
2. relate_behind
3. filter_cube
4. filter_large
5. count
```

ID	Size	Shape	Material	Color
2	Large	Cube	Metal	Green
3	Large	Cube	Metal	Green
5	Large	Cube	Metal	Green

Answer: 3

Advantage 1: High Accuracy

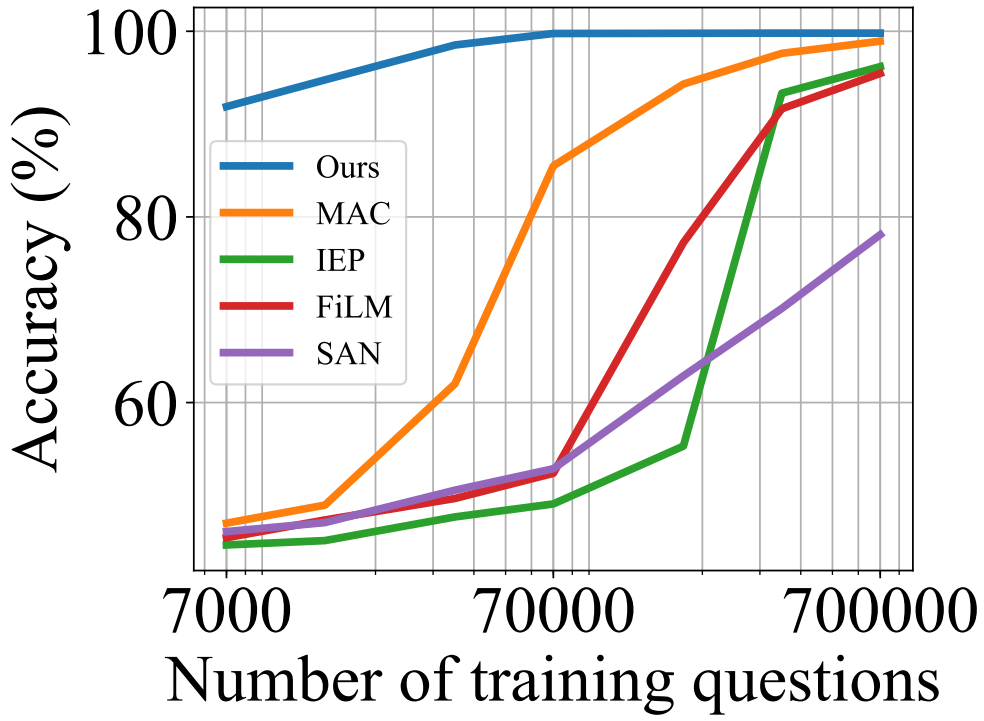
- Symbolic reasoning is robust to longer logic traces
- Our model outperforms current state-of-the-art methods on CLEVR

Method	Accuracy (%)
Human	92.6
RN	95.5
IEP	96.9
FiLM	97.6
MAC	98.9
TbD	99.1
NS-VQA (Ours)	**99.8**

[Johnson et al. ICCV 2017, Santoro et al. NIPS 2017, Perez et al. AAAI 2018, Hudson et al. ICLR 2018, Mascharka et al. CVPR 2018]

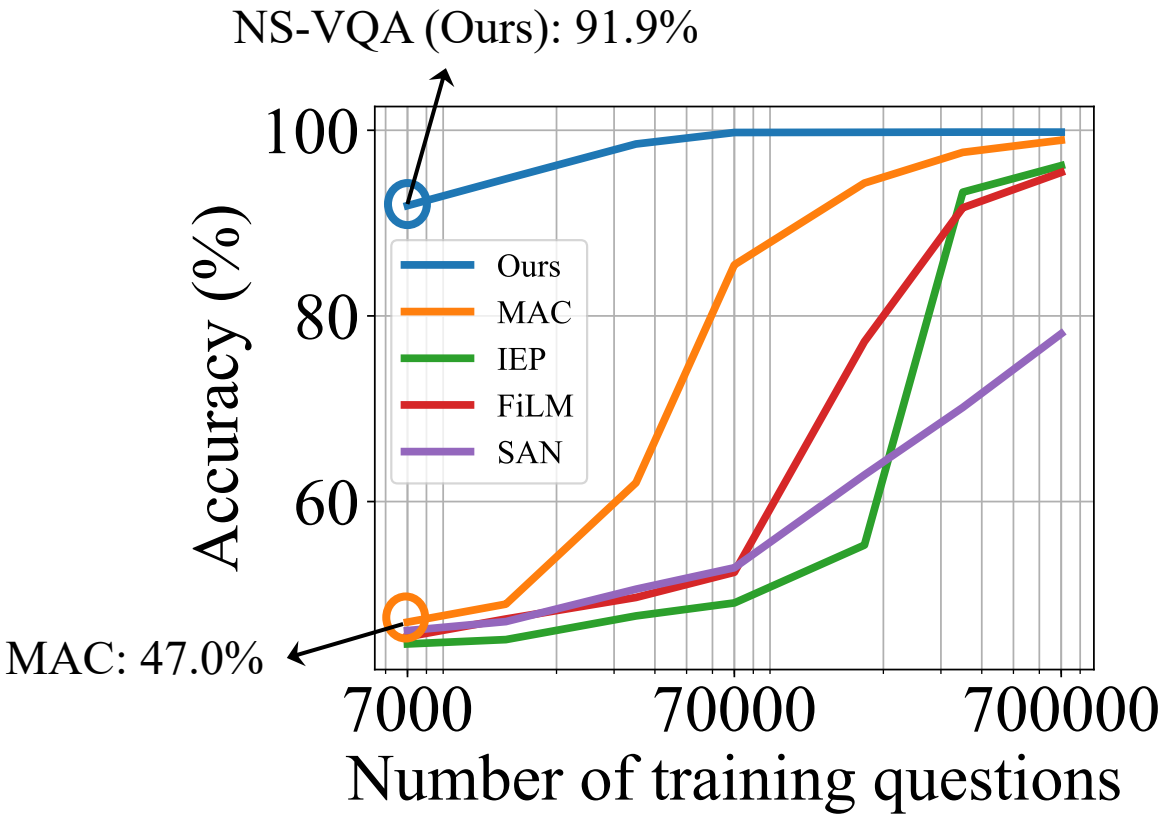
Advantage 2: Data Efficiency

- Our disentangled model requires fewer questions for training



Advantage 2: Data Efficiency

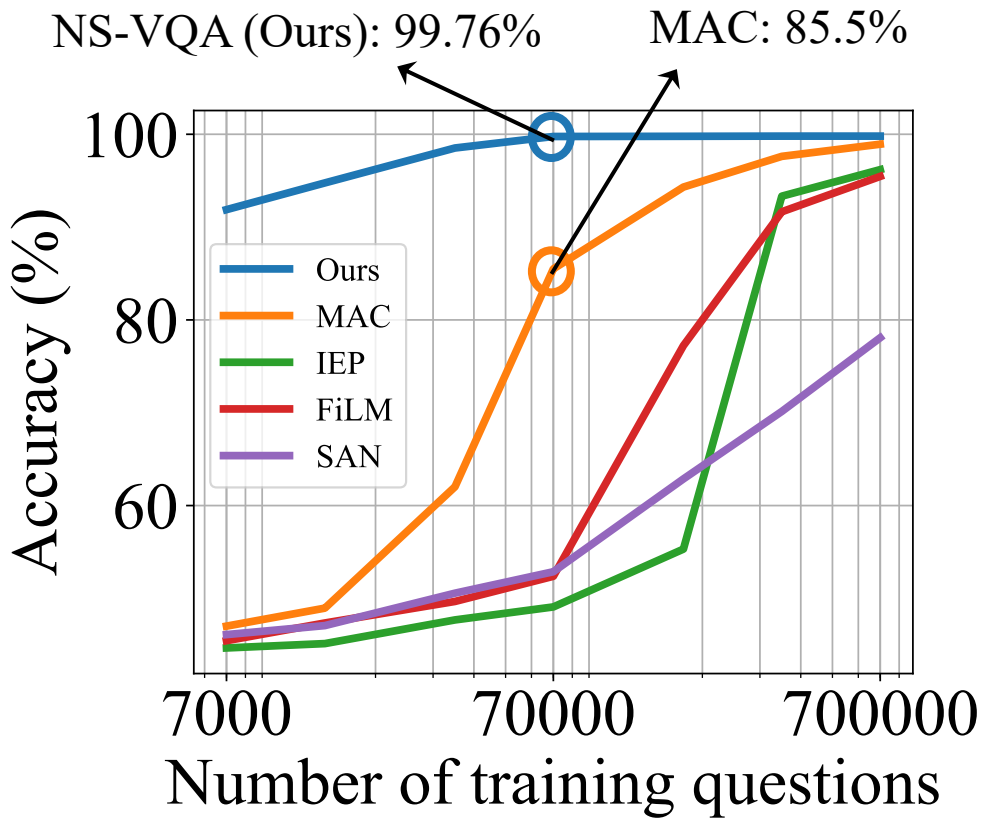
- Our disentangled model requires fewer questions for training
- 91% accuracy when trained on 1% questions (44% higher than strongest baseline)



[Hudson et al. ICLR 2018]

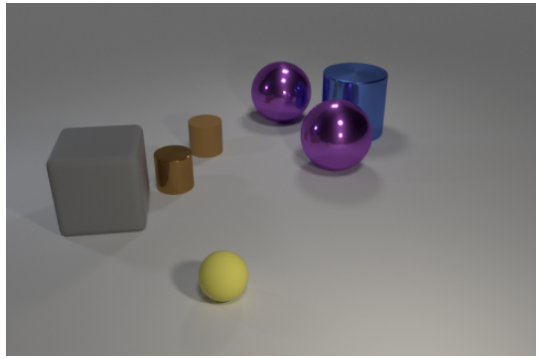
Advantage 2: Data Efficiency

- Our disentangled model requires fewer questions for training
- 91% accuracy when trained on 1% questions (44% higher than strongest baseline)
- 99.7% accuracy when trained on 10% questions (14% higher than strongest baseline)



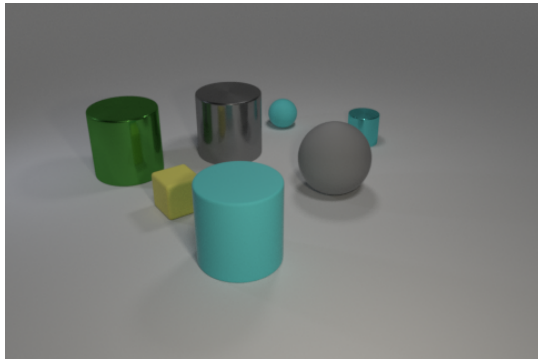
[Hudson et al. ICLR 2018]

Advantage 3: Transparency and Interpretability



Question: What number of cylinders are gray objects or tiny brown matte objects?

Ours	IEP
scene	filter_small
filter_small	filter_brown
filter_brown	filter_large
filter_rubber	filter_cyan
scene	...(25 modules)
filter_gray	filter_metal
union	union
filter_cylinder	filter_cylinder
count	count
Answer: 1	**Answer: 2**

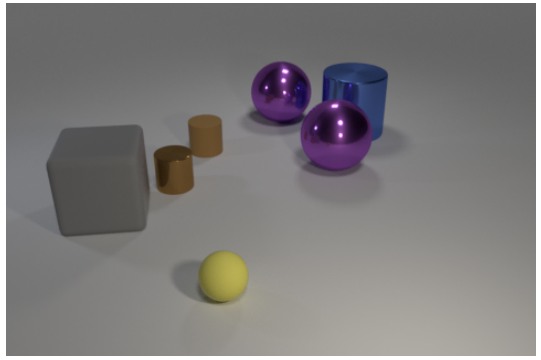


Question: Are there more yellow matte things that are right of the gray ball than cyan metallic objects?

Ours	IEP
scene	filter_small
filter_cyan	filter_cyan
filter_metal	union
count	filter_brown
...(4 modules)	...(25 modules)
scene	filter_small
filter_yellow	filter_yellow
filter_rubber	filter_rubber
count	count
greater_than	greater_than
Answer: no	**Answer: no**

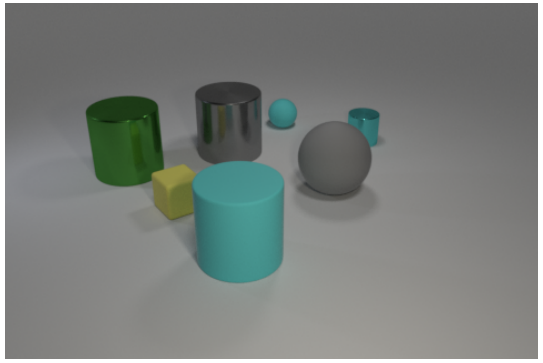
[Johnson et al. ICCV 2017]

Advantage 3: Transparency and Interpretability



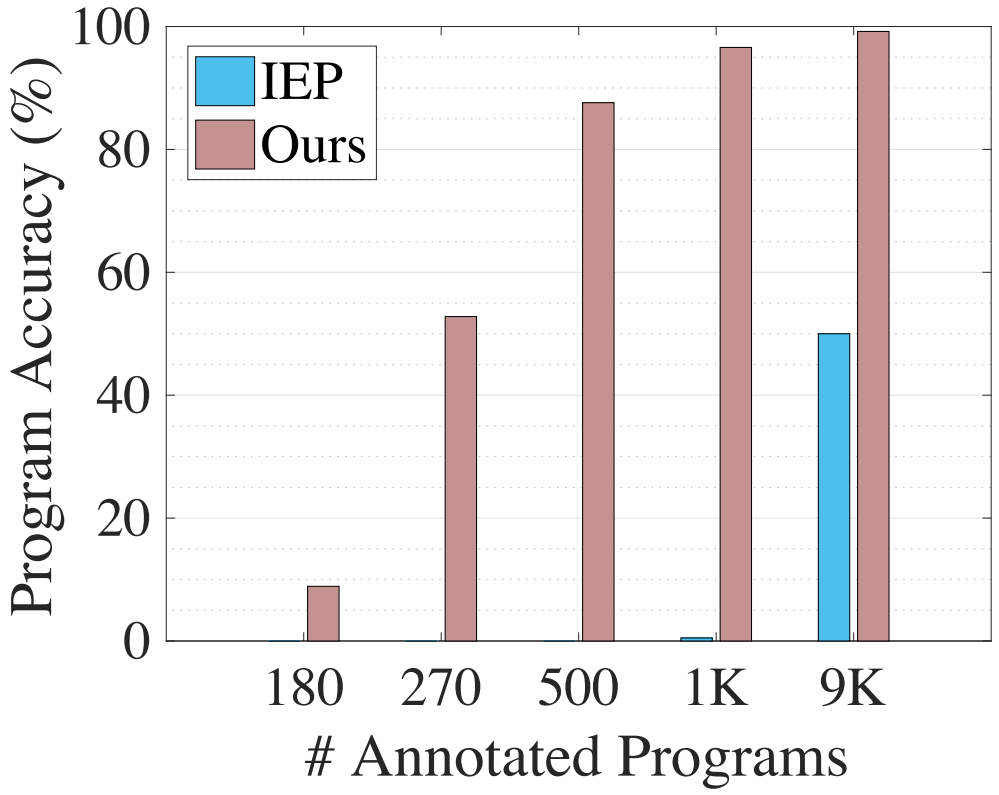
Question: What number of cylinders are gray objects or tiny brown matte objects?

Ours	IEP
scene	filter_small
filter_small	filter_brown
filter_brown	filter_large
filter_rubber	filter_cyan
scene	...(25 modules)
filter_gray	filter_metal
union	union
filter_cylinder	filter_cylinder
count	count
Answer: 1	**Answer: 2**



Question: Are there more yellow matte things that are right of the gray ball than cyan metallic objects?

Ours	IEP
scene	filter_small
filter_cyan	filter_cyan
filter_metal	union
count	filter_brown
...(4 modules)	...(25 modules)
scene	filter_small
filter_yellow	filter_yellow
filter_rubber	filter_rubber
count	count
greater_than	greater_than
Answer: no	**Answer: no**



[Johnson et al. ICCV 2017]

Summary

- Neural-Symbolic VQA (NS-VQA)
  - Disentangled visual reasoning
  - *Neural* scene and question parsing
  - *Symbolic* program execution

- Advantages
  - High accuracy (99.8% on CLEVR)
  - Data efficiency (99.7% with 10% training data)
  - Interpretability and transparency

Method	Accuracy (%)
Human	92.6
RN	95.5
IEP	96.9
FiLM	97.6
MAC	98.9
TbD	99.1
NS-VQA (Ours)	**99.8**