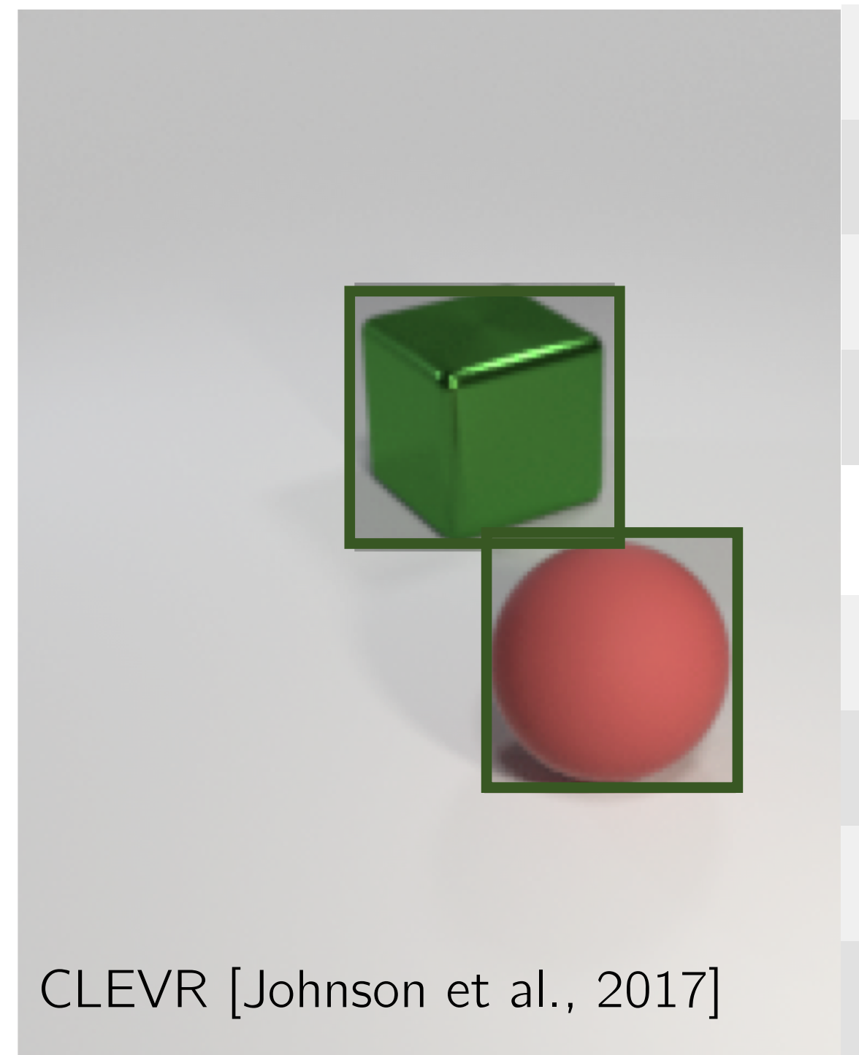
The Neuro-Symbolic Concept Learner

Interpreting Scenes, Words and Sentences from Natural Supervision

Jiayuan Mao[1,2] Chuang Gan[3] Pushmeet Kohli[4] Joshua B. Tenenbaum[1] Jiajun Wu[1]
1 Massachusetts Institute of Technology 2 IIIS, Tsinghua University 3 MIT-IBM Watson AI Lab 4 DeepMind

Concept Learning in Visual Reasoning


CLEVR [Johnson et al., 2017]

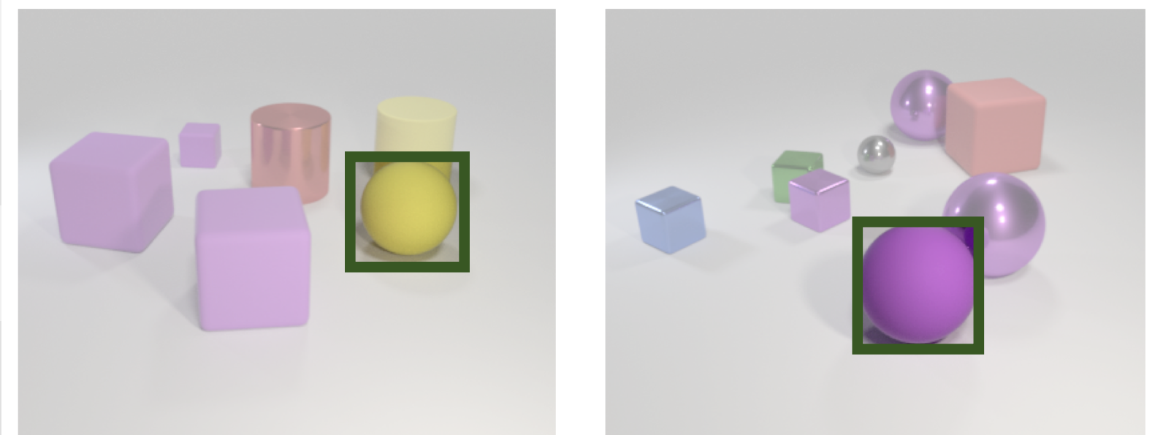
Color	Green
Shape	Cube
Material	Metal
……	……

Color	Red
Shape	Sphere
Material	Rubber
……	……

Visual Question Answering
Q: What's the shape of the red object?
A: Sphere.

Image Captioning
There are a red sphere and a green cube.

Instance Retrieval: rubber sphere.

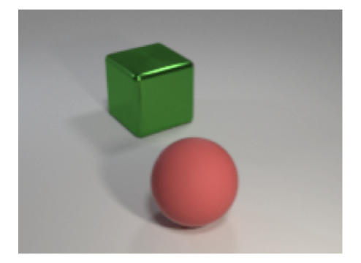


Overview of Visual Reasoning Methods

Models	Visual Features	Semantics	Extra Labels		Inference
			# Prog.	Attr.	
FiLM (Perez et al., 2018)	Convolutional	Implicit	0	No	Feature Manipulation
IEP (Johnson et al., 2017b)	Convolutional	Explicit	700K	No	Feature Manipulation
MAC (Hudson & Manning, 2018)	Attentional	Implicit	0	No	Feature Manipulation
Stack-NMN (Hu et al., 2018)	Attentional	Implicit	0	No	Attention Manipulation
TbD (Mascharka et al., 2018)	Attentional	Explicit	700K	No	Attention Manipulation
NS-VQA (Yi et al., 2018)	Object-Based	Explicit	0.2K	Yes	Symbolic Execution
NS-CL	Object-Based	Explicit	0	No	Symbolic Execution

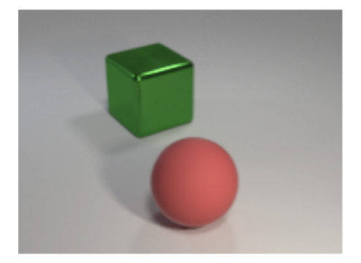
Curriculum Learning

□ Lesson1: Object-based questions.



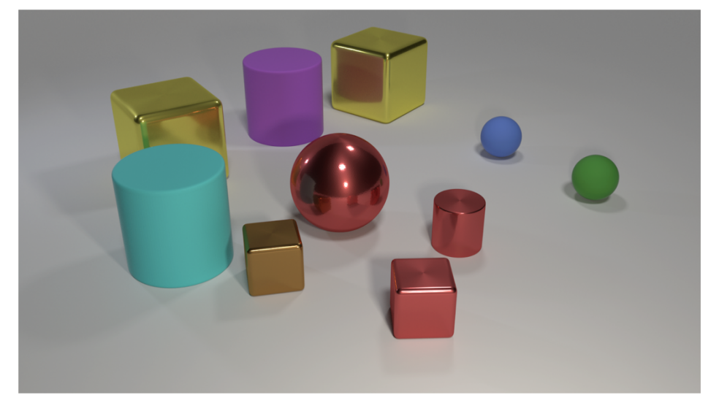
Q: What is the shape of the red object?
A: Sphere.

□ Lesson2: Relational questions.



Q: Is the green cube left to the red sphere?
A: Yes

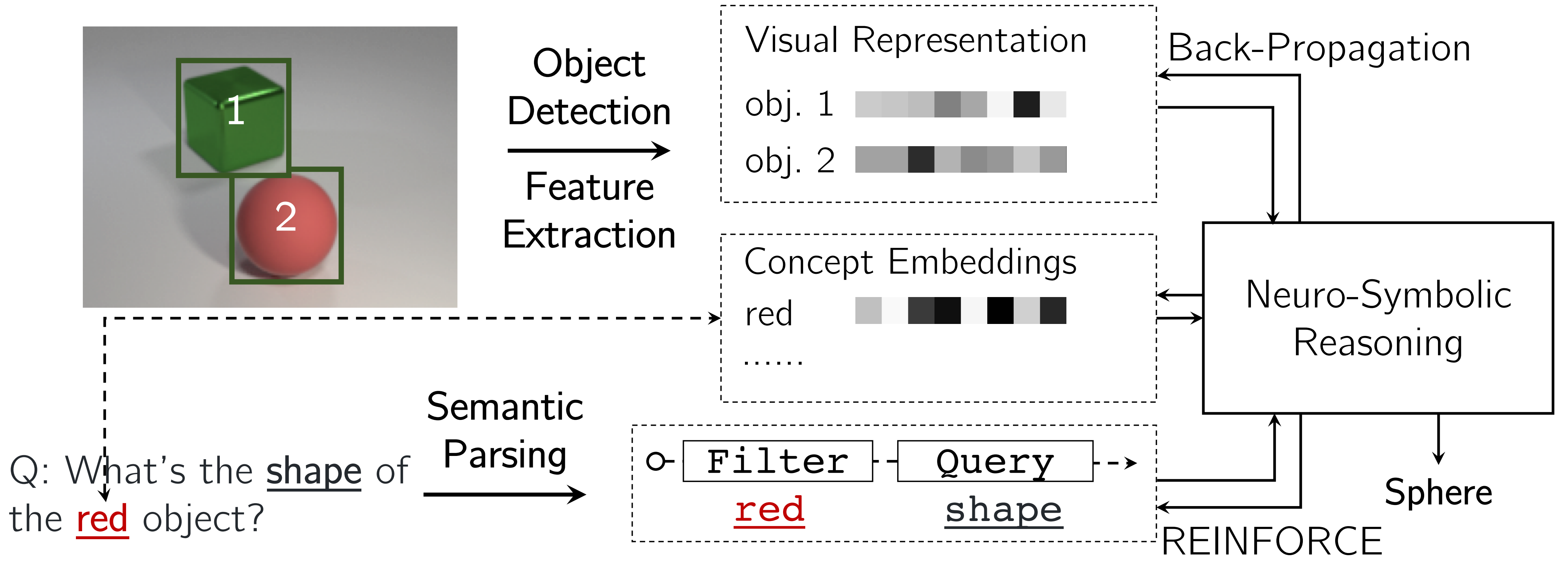
□ Lesson3: complex scenes, complex questions



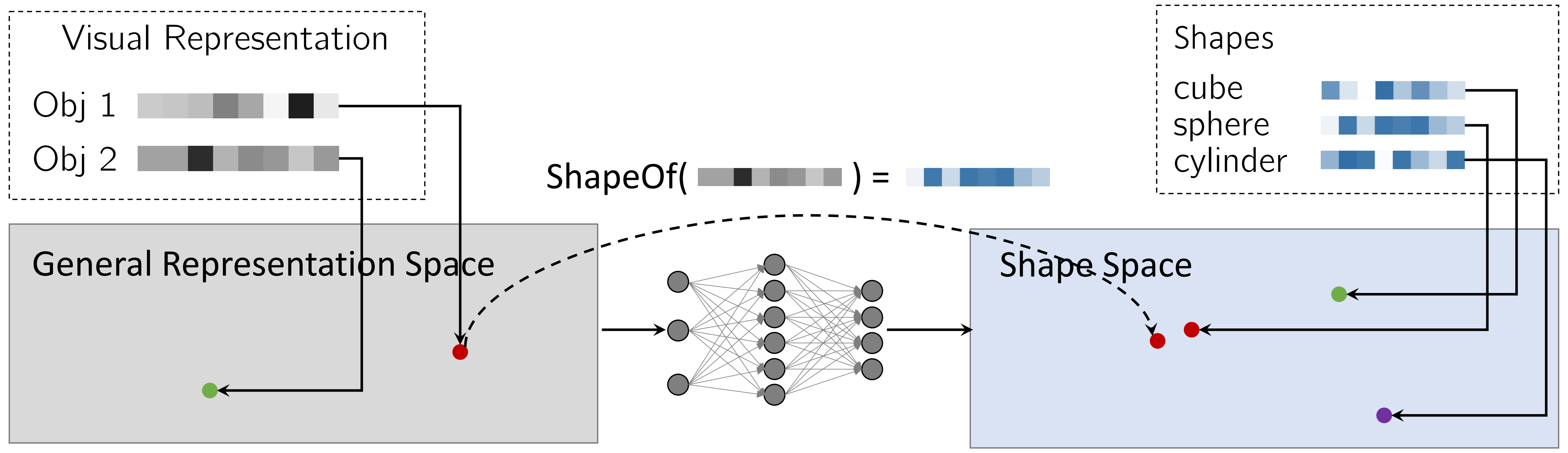
Q: Does the big matte object behind the big sphere have the same color as the cylinder left of the small brown cube?
A: No.

The Neuro-Symbolic Concept Learner

Principle 1: **Explicit** visual grounding of concepts with **neuro-symbolic** reasoning.
Principle 2: **Joint** learning of concepts and language with developmental **curriculum**.

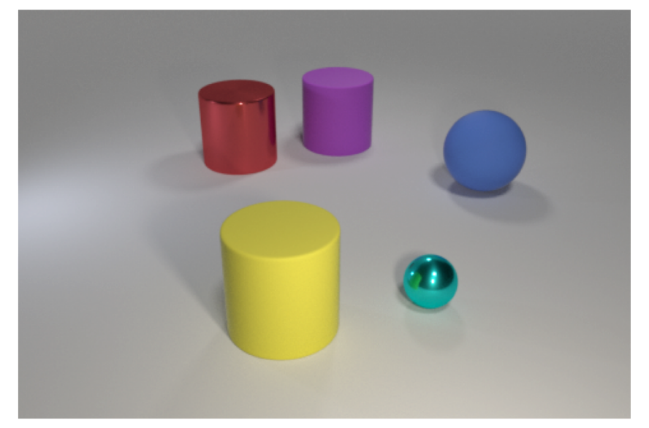


Visual-Semantic Embeddings for Shape Query



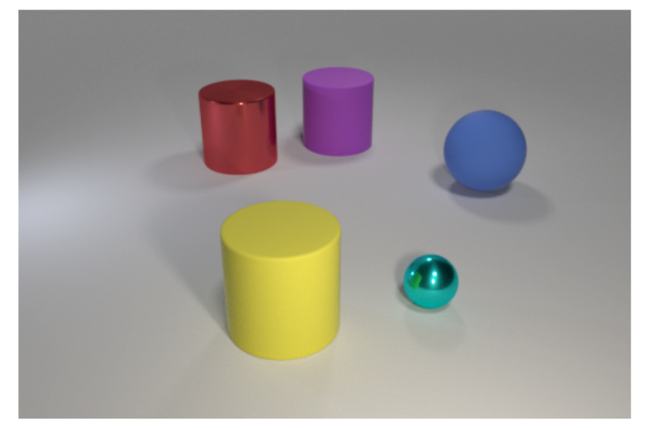
Combinatorial Generalization

A: #objects ≤ 6, depth ≤ 4



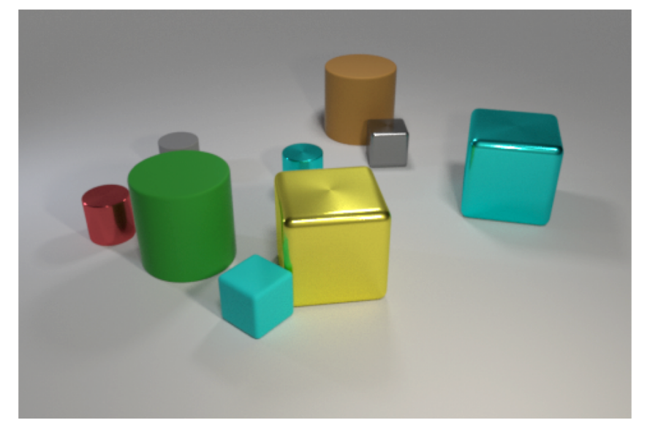
Q: What's the shape of the big yellow thing?

B: #objects ≤ 6, depth > 4



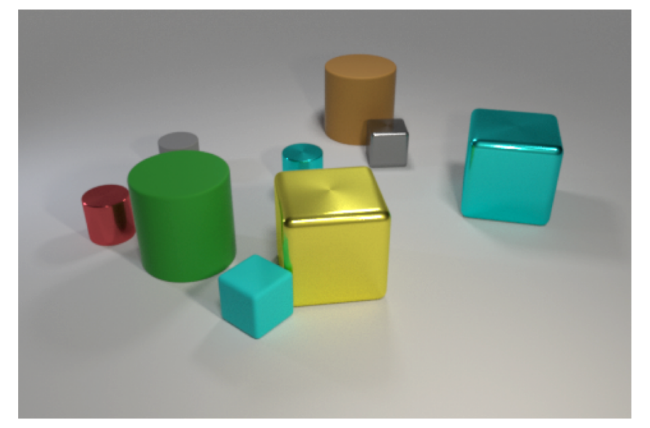
Q: What size is the cylinder that is left of the cyan thing that is in front of the big sphere?

C: #objects > 6, depth ≤ 4



Q: What's the shape of the big yellow thing?

D: #objects > 6, depth > 4



Q: What size is the cylinder that is left of the cyan thing that is in front of the gray cube?

Model	Test			
	Split A	Split B	Split C	Split D
MAC	97.3	N/A	92.9	N/A
IEP	96.1	92.1	91.5	90.9
TbD	98.8	94.5	94.3	91.9
NS-CL	**98.9**	**98.9**	**98.7**	**98.8**

Training Set: Split A Only.

Results on the CLEVR Dataset

70k images, 700k questions, 19 concepts [Johnson et al., 2017]

Model	Prog. Anno.	Overall	Count	Cmp. Num.	Exist	Query Attr.	Cmp. Attr.
Human	N/A	92.6	86.7	86.4	96.6	95.0	96.0
NMN	700K	72.1	52.5	72.7	79.3	79.0	78.0
N2NMN	700K	88.8	68.5	84.9	85.7	90.0	88.8
IEP	700K	96.9	92.7	98.7	97.1	98.1	98.9
DDRprog	700K	98.3	96.5	98.4	98.8	99.1	99.0
TbD	700K	99.1	97.6	99.4	99.2	99.5	99.6
RN	0	95.5	90.1	93.6	97.8	97.1	97.9
FiLM	0	97.6	94.5	93.8	99.2	99.2	99.0
MAC	0	98.9	97.2	99.4	99.5	99.3	99.5
NS-CL (10% data)	0	98.9	98.2	99.0	98.8	99.3	99.1
NS-CL (full data)	0	**99.6**	**99.3**	**99.6**	**99.7**	**99.8**	**99.6**

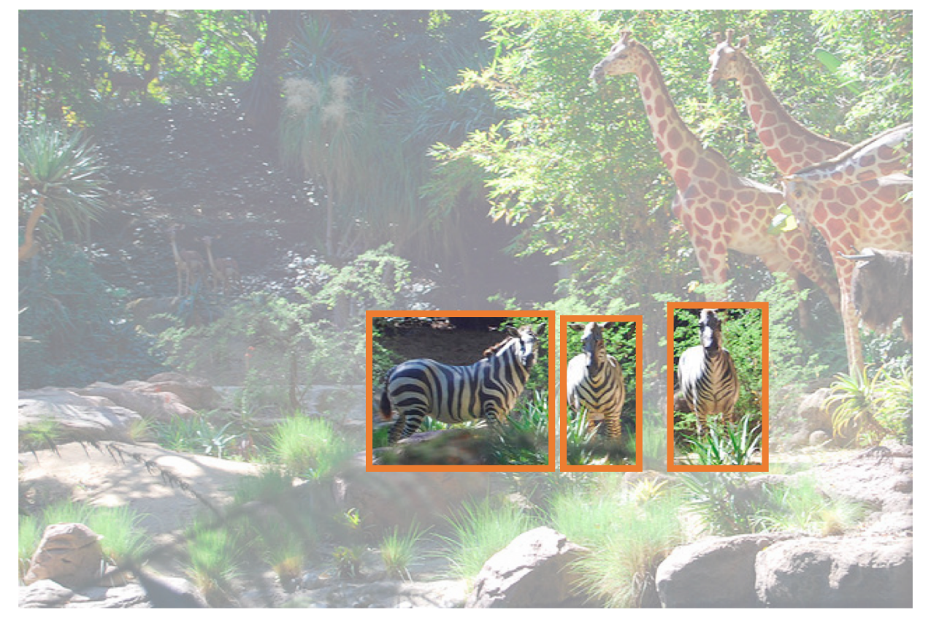
Results on the VQS Dataset

30k images, 90k questions, 9k concepts [Gan et al. 2017]

Q: What **color** is the **fire hydrant**?

Filter fire_hydrant → Query color

A: Yellow



Q: How many **zebras** are there?

Filter zebra → Count

A: 3

Concepts for Instance Retrieval

Horse

Person On a Skateboard