Machine Reasoning: A Vision Perspective

Karan Samel

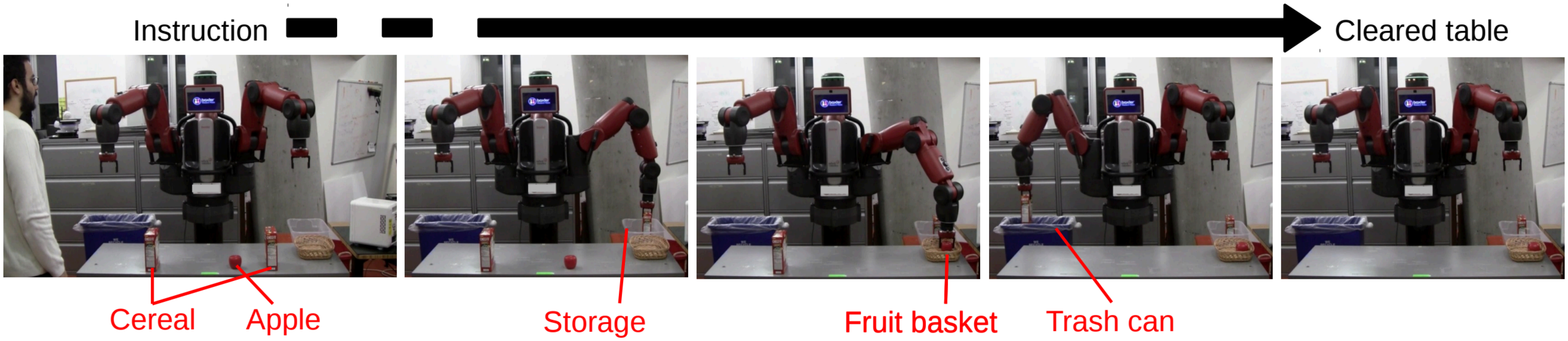
ksamel@gatech.edu

Georgia Institute of Technology – 7/2/20

Machine Reasoning

There are many works that optimize ML models by themselves.

How are these models leveraged in a larger system?



(a) Human: "The box on the left is empty. Clear the table."

Nyga, et al. Grounding Robot Plans from Natural Language Instructions with Incomplete World Knowledge. CoRL'18

Visual Reasoning

Part I: Visual Question Answering

VQA Motivation

Testbed in visual question answering (VQA). Why is it hard?

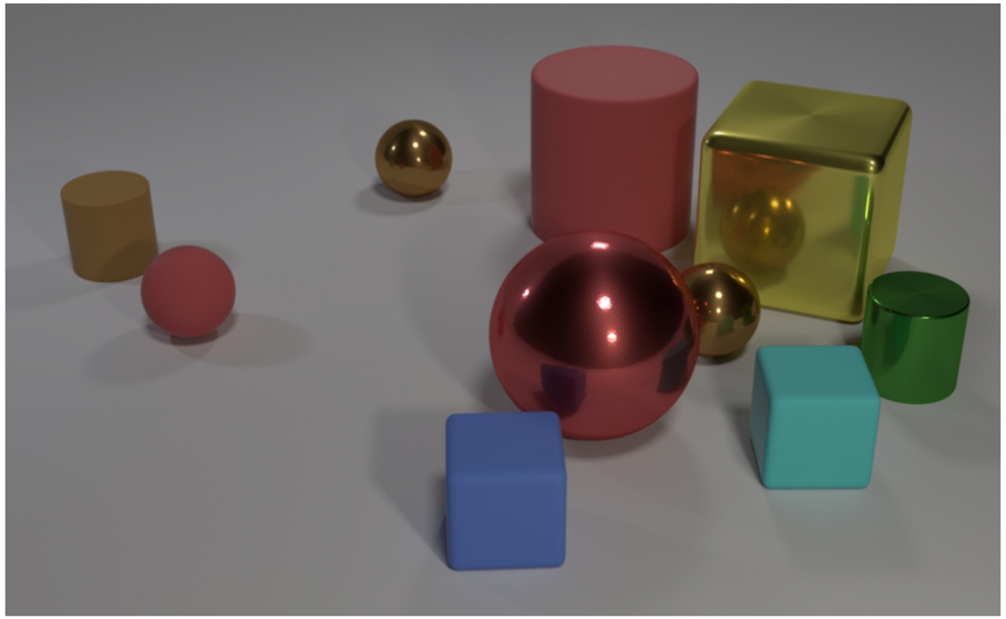
- **There are not a fixed number of output labels.**

- **Questions are compositional in nature, thus many possible inputs.**

Can we leverage the structure of scenes and questions to reduce the data requirement?

VQA Datasets

Given an image and question, how do we arrive to the answer?



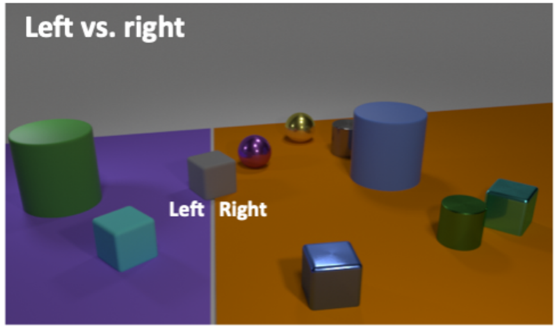
Q: Are there an equal number of large things and metal spheres?
Q: What size is the cylinder that is left of the brown metal thing that is left of the big sphere? **Q:** There is a sphere with the same size as the metal cube; is it made of the same material as the small red sphere?
Q: How many objects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

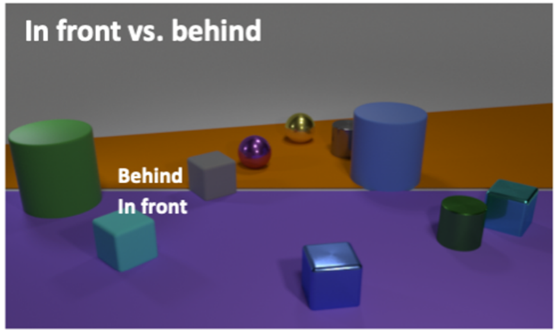
Figure 1: Examples from the new GQA dataset for visual reasoning and compositional question answering:
Is the ***bowl*** *to the right of the* ***green apple****?*
What type of ***fruit*** *in the image is* ***round****?*
What color is the ***fruit*** *on the right side, red or* ***green****?*
Is there any ***milk*** *in the* ***bowl*** *to the left of the* ***apple****?*

Johnson, et al. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. CVPR'17
Hudson, Manning. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. CVPR'19

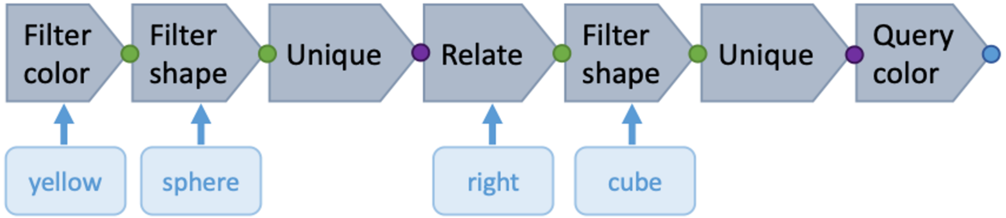
VQA Datasets

Auxiliary labels such as scene graphs and functional programs provided.

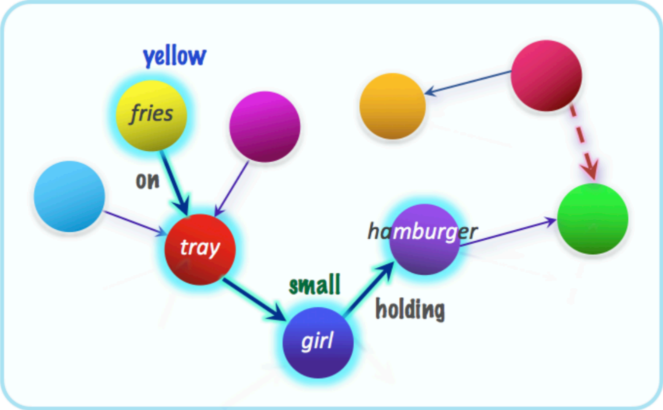




Sample chain-structured question:



What color is the cube to the right of the yellow sphere?

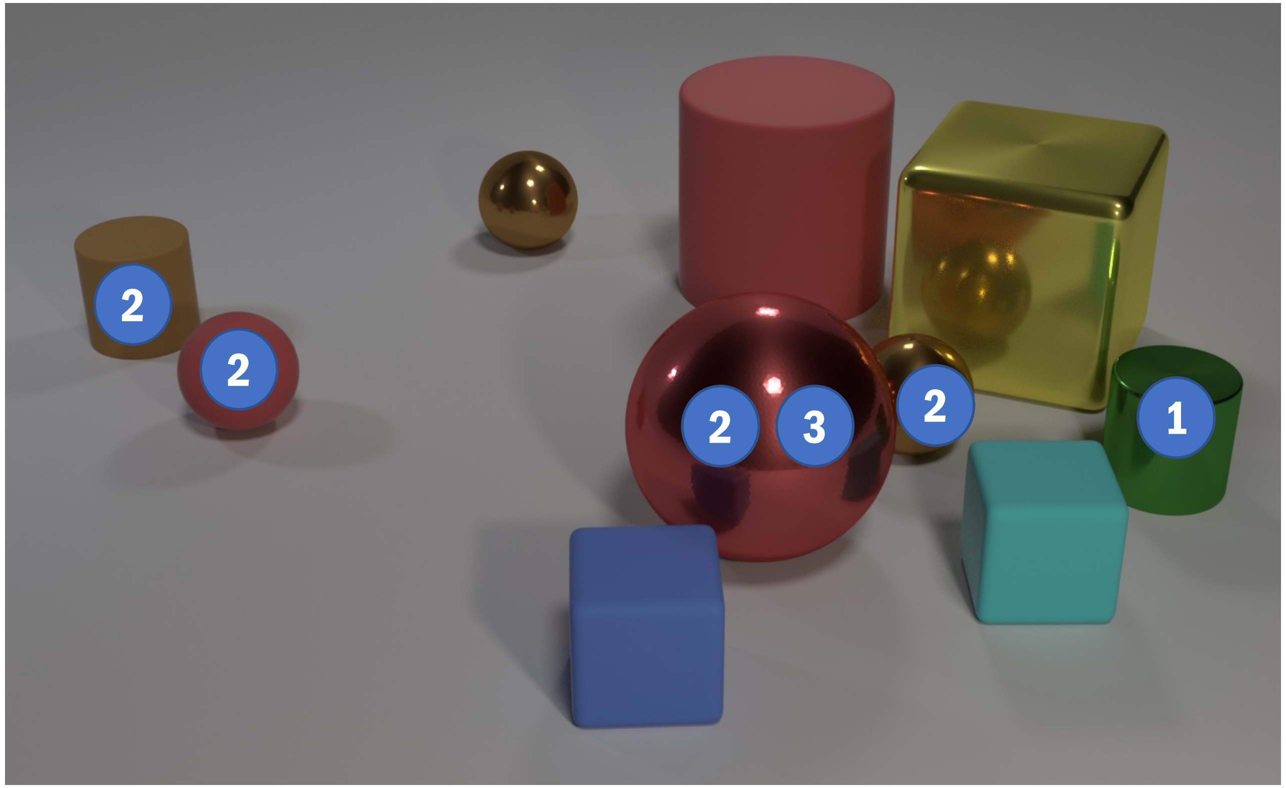


What color is the food on the red object left of the small girl that is holding a hamburger, yellow or brown?

Select: hamburger ⟶ **Relate:** girl,holding ⟶ **Filter** size: small ⟶ **Relate:** object, left ⟶ **Filter** color: red ⟶ **Relate:** food,on ⟶ **Choose** color: yellow | brown

Program structure for VQA data

CLEVR



"What color is the large round object is to the left of the green cylinder?"

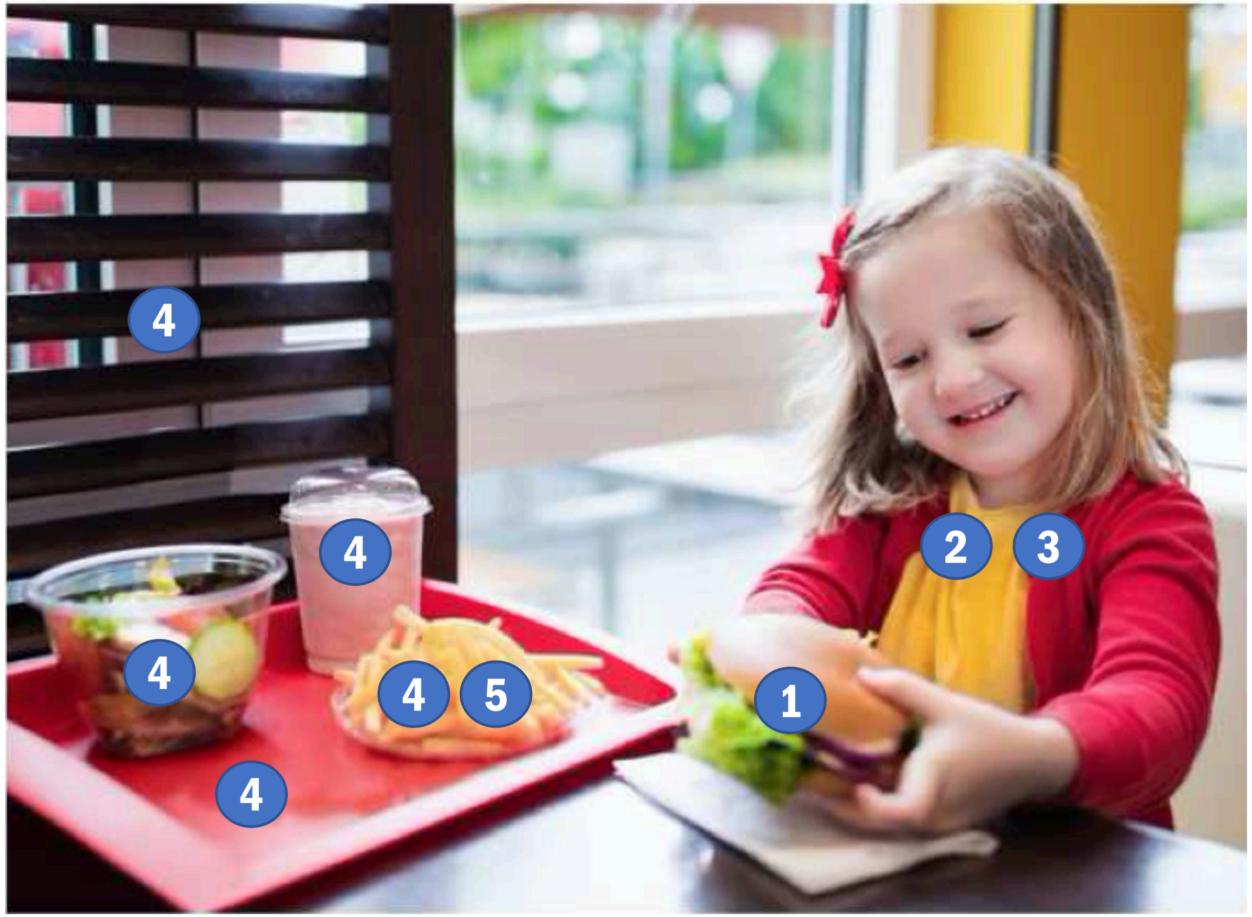
Program	Selection
Filter ([green, cylinder])	1
Relate (1 , left)	2
Filter (2 , [round, large])	3

A: Red

Johnson, et al. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. CVPR'17

Program structure for VQA data

GQA



"What is the yellow food to the left of the small girl that is holding the hamburger?"

Program	Selection
Filter (hamburger)	1
Relate (1 , holding)	2
Filter (2 , [small, girl])	3
Relate (3 , left)	4
Filter (4 , yellow])	5

A: Fries

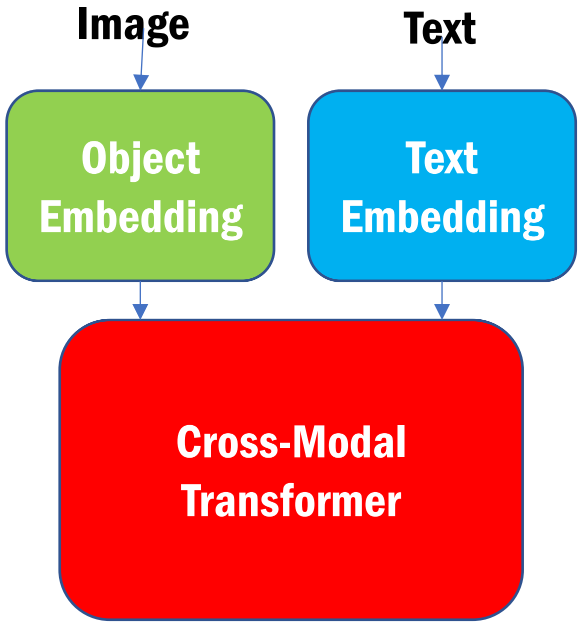
Hudson, Manning. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. CVPR'19

Modeling Approaches

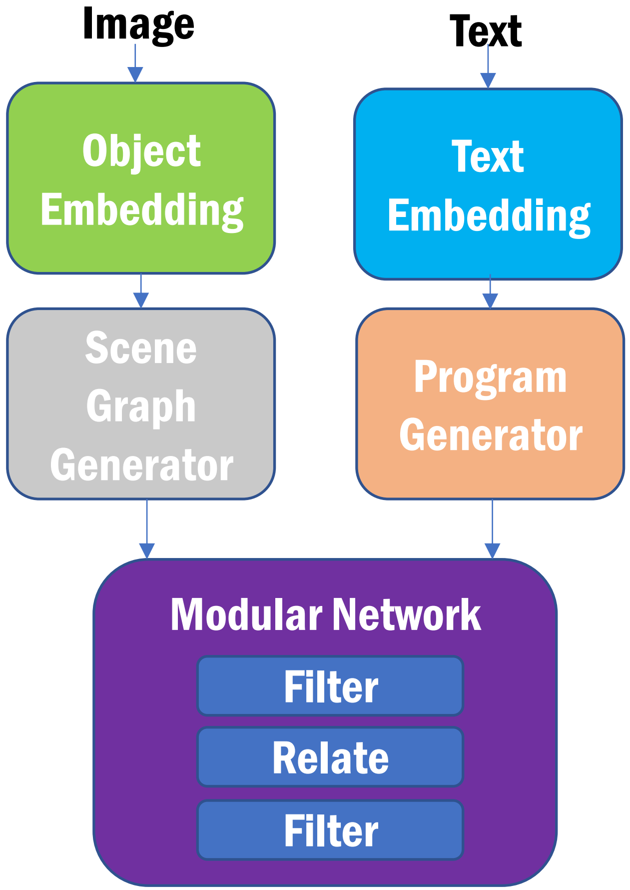
Tradeoffs:

Deep Networks		Modular Networks
✓	1) Implementation	✗
✗	2) Interpretability	✓
↑	3) Label requirements	↓

Deep Networks



Modular Networks



Modeling Approaches - Transformers

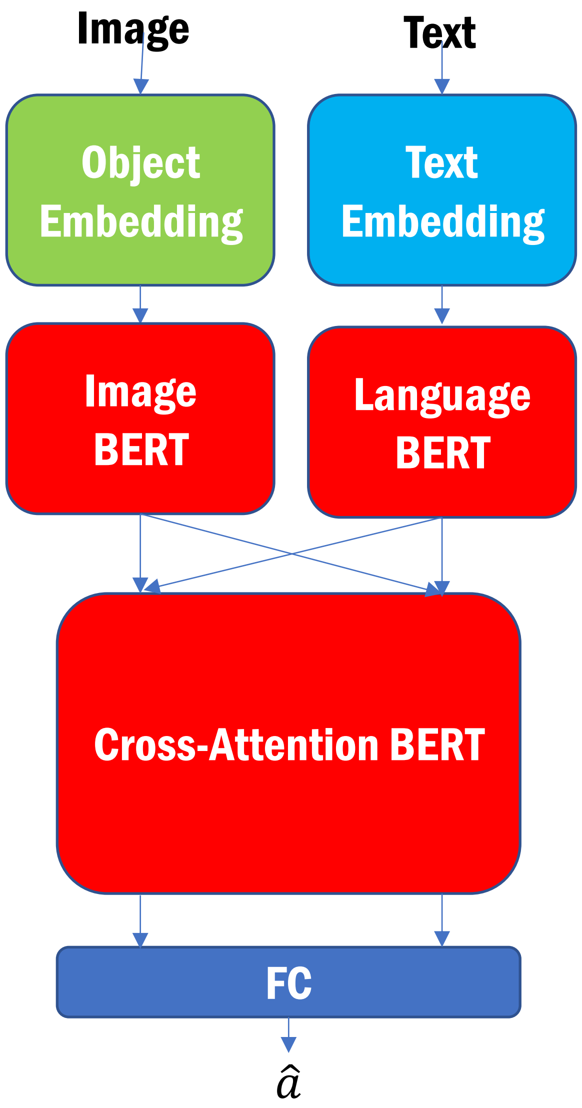
LXMERT

Object Embedding through R-CNN and ResNet
Text Embedding initialized from scratch

Embeds both the text and image features through the BERT encoder

Constructs a joint representation of the image and language BERT features

Pretrained on VQA and image training splits, FC fine tuned on each task



Tan, Bansal. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. EMNLP'19

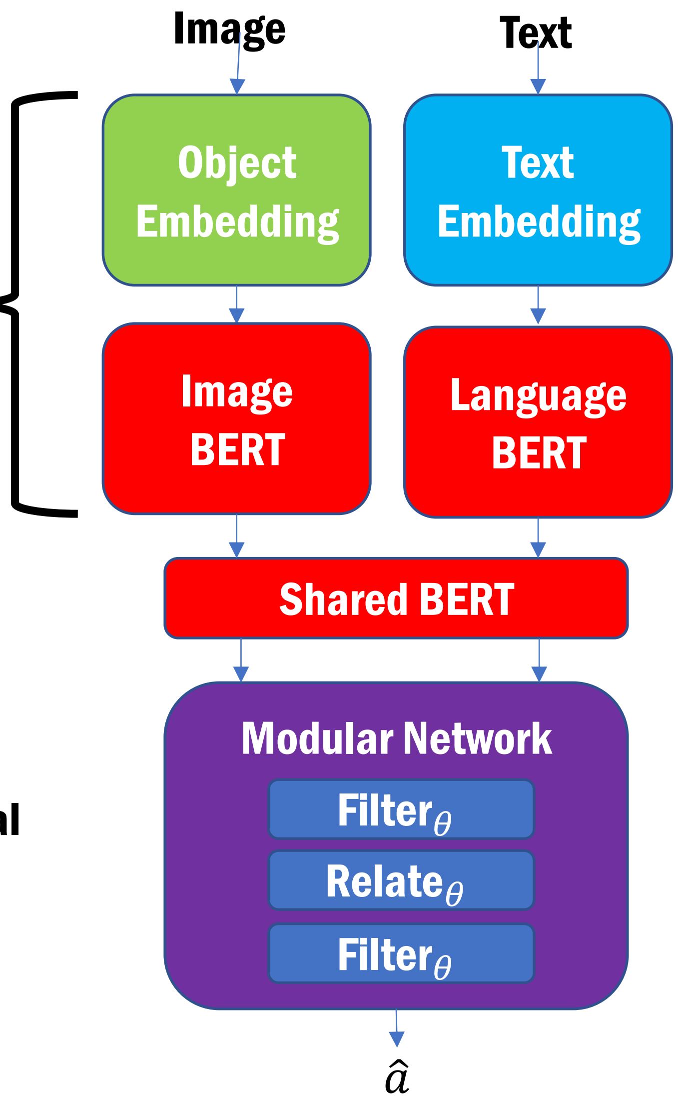
Modeling Approaches - Transformers

SVDR

Same as LXMERT

Disjoint Language and Image encoder for pretraining

Modular network functions as cross-modal transformers



Liu, et al. Static-Dynamic Reasoning via Functional Programs. (WIP)

Modeling Approaches – Graph Traversal

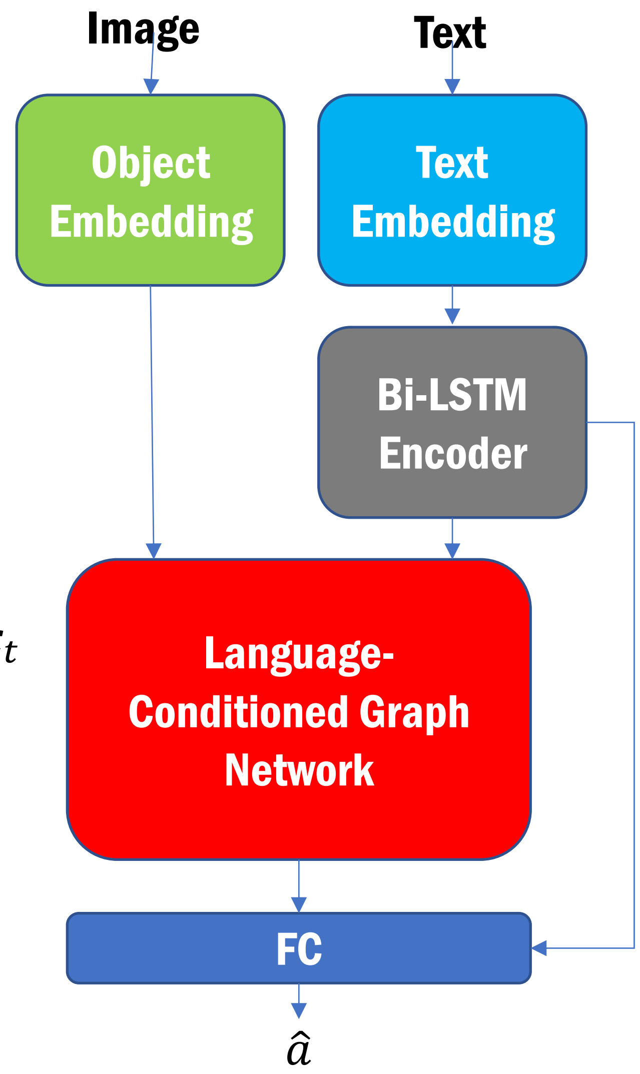
Language Conditioned Graph Networks (LCGN)

Use LSTM to encode question q
Text commands $c_t \mid q$ are encoded for T timesteps

Message passing on the object embeddings x_i occurring T times:

w_{ji}^t computed from $x_i^{t-1}, x_j^{t-1}, c_t$
m_{ji}^t computed from w_{ji}^t, x_j^{t-1}, c_t
$x_i^t = FC\ ([x_i^{t-1}; \sum_j m_{ji}^t])$

Use the final representations for $FC\left([\sum_i x_i^T; q]\right) = \hat{a}$



Hu, et al. Language-Conditioned Graph Networks for Relational Reasoning . ICCV'19

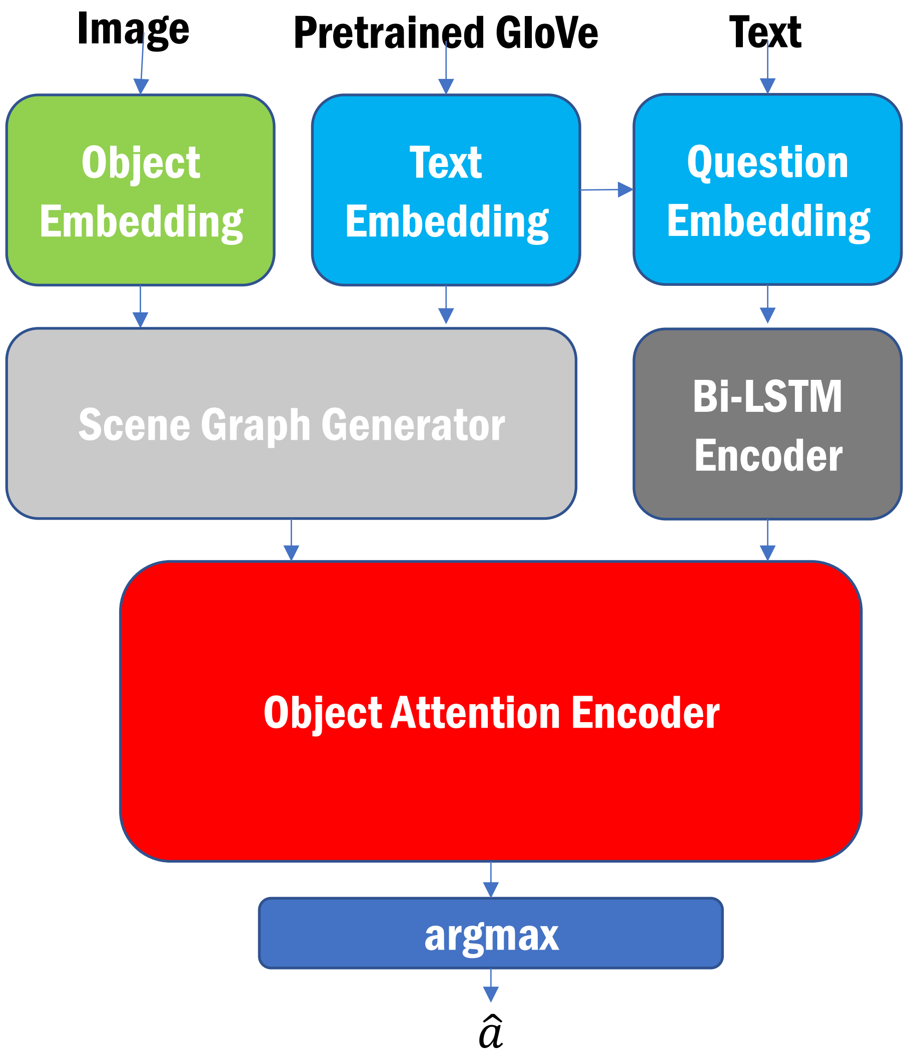
Modeling Approaches – Graph Traversal

Neural State Machine (NSM)

A scene graph of object attribute and relations is computed. These representations are shared with the text embedding.

Successively compute the probability of object x_i as the traversal object for question step c_t :

Take the final object with the highest attention to answer the query.



Hudson, Manning. Learning by Abstraction: The Neural State Machine. NeurIPS'19

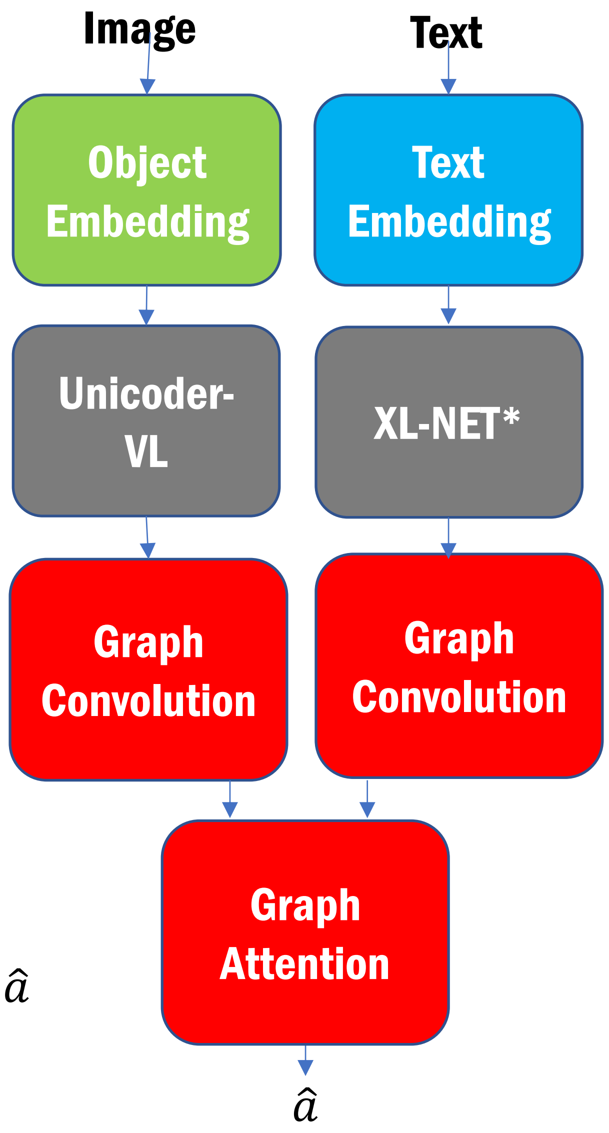
Modeling Approaches – Graph Traversal

DREAM

Encode images with a visual language transformers.
Encode text with XL-NET and minimize the graph distance of the embeddings as well.

Message pass image and text features

Distribution attention weights between objects to retrieve $\hat{a}$



Zhong, et al. Reasoning Over Semantic-Level Graph for Fact Checking. arXiv

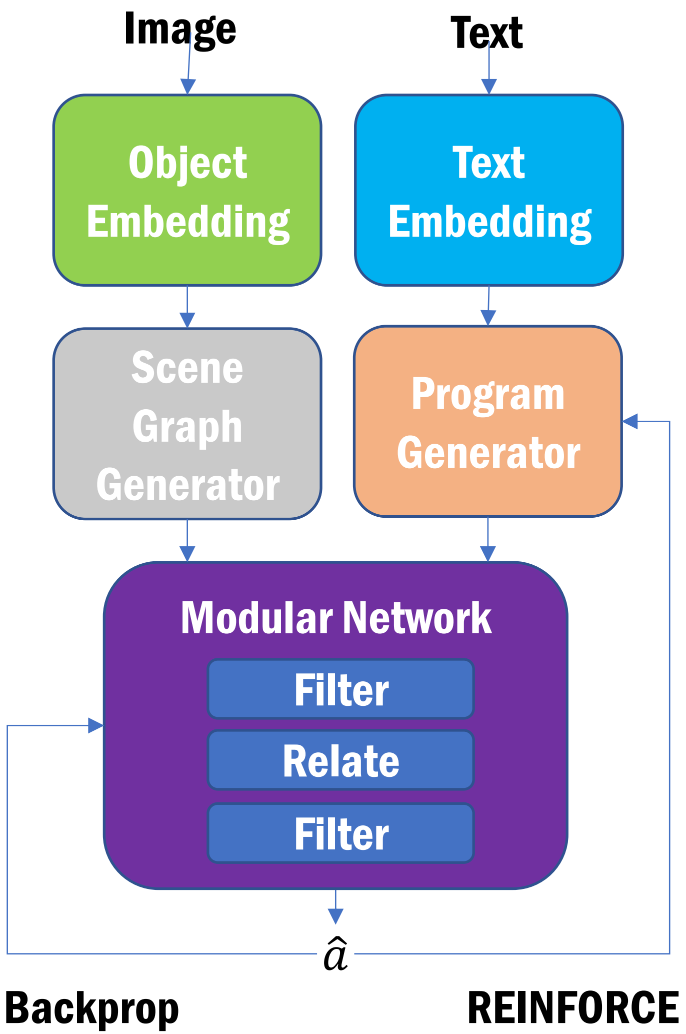
Modeling Approaches – Neuro-Symbolic

Leverages the scene graph and the program generator to softly traverse the graph to $\hat{a}$

Computes embedding concepts for the program tokens to match the corresponding object attributes

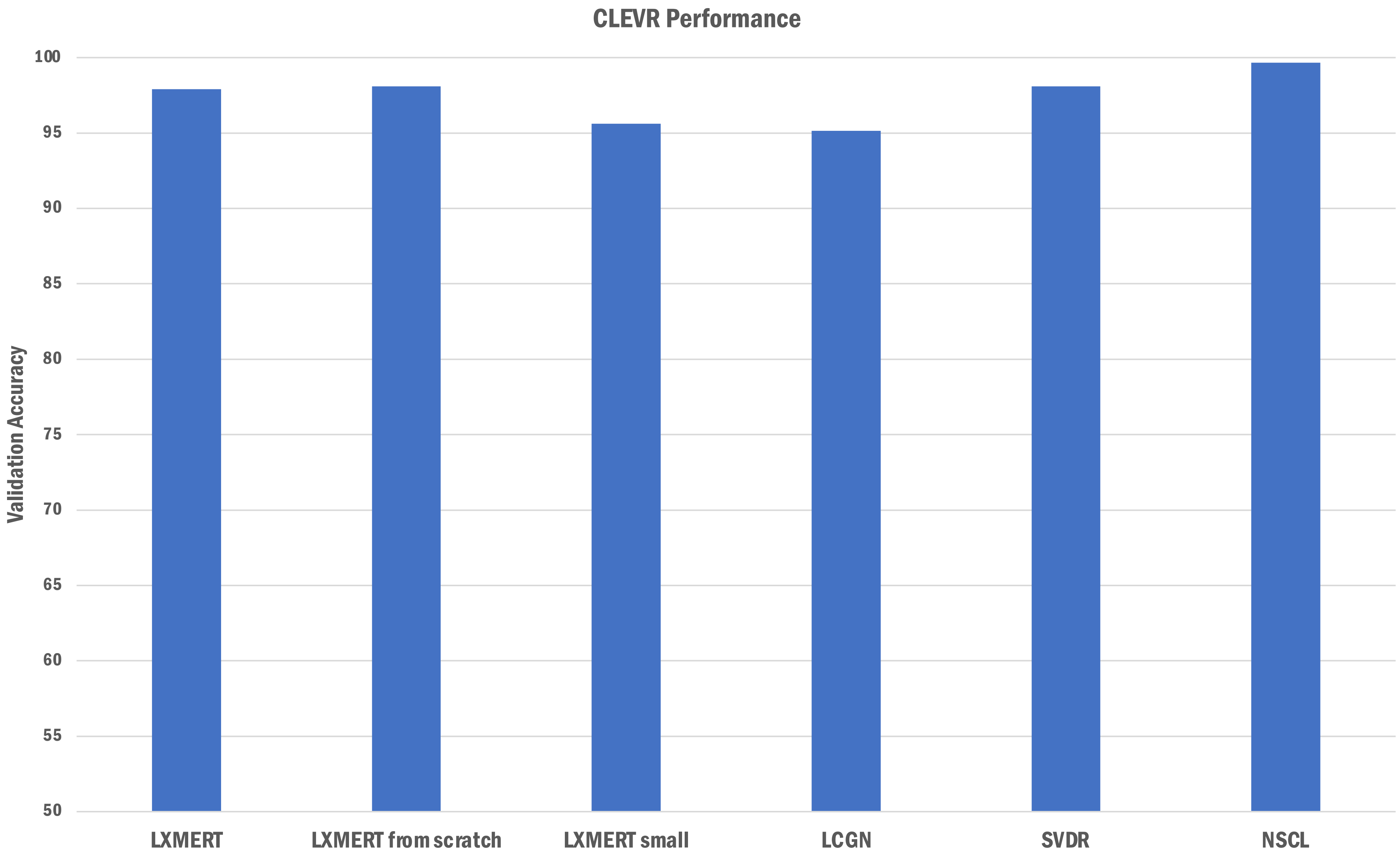
Ex: Learn concepts $sphere$ and W_{shape} such that $\langle W_{shape} \times ResNet($$), sphere\rangle \approx 1$

Modular functions are hand coded and differentiable

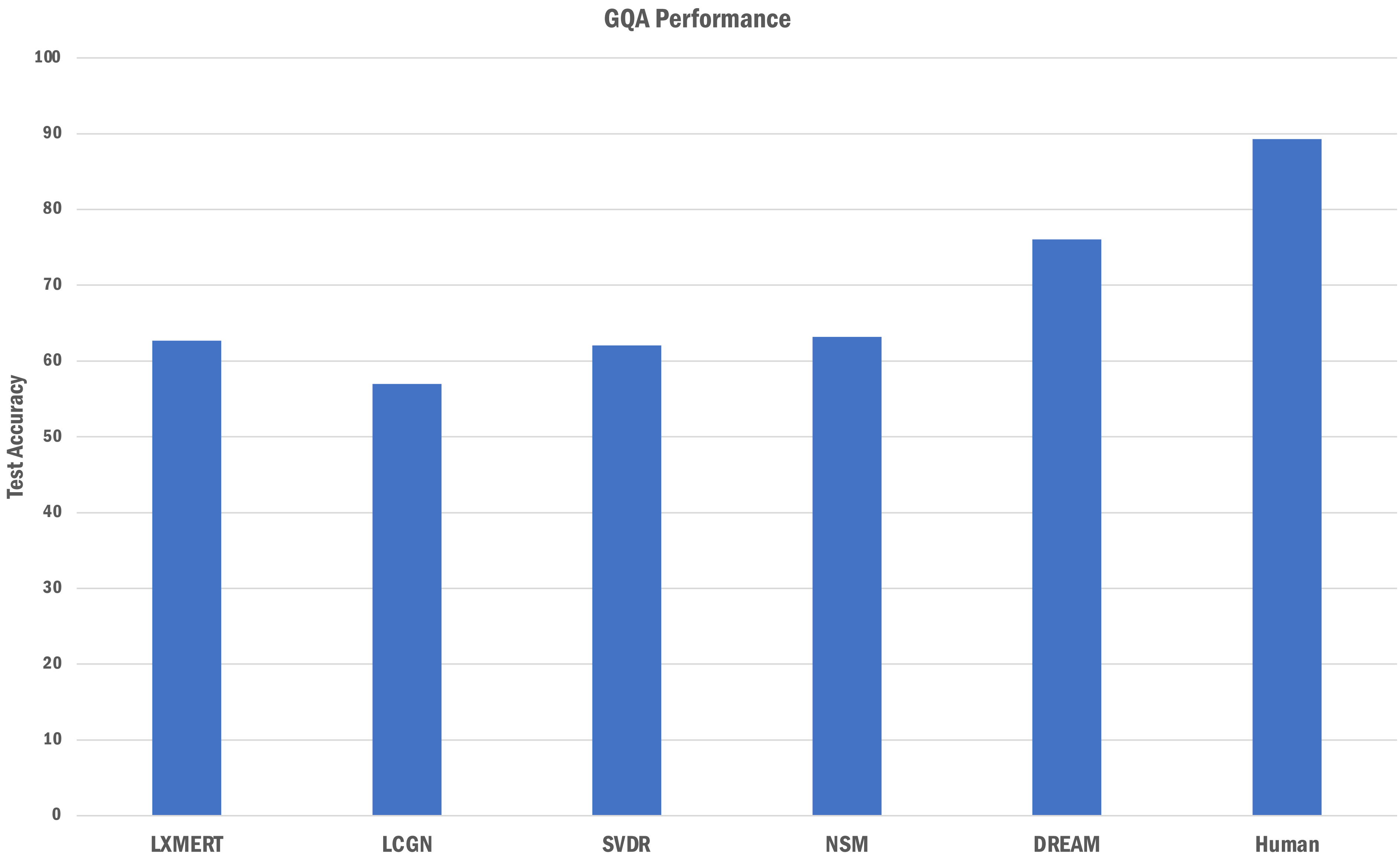


Mao, et al. The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences from Natural Supervision. ICLR'19

Model Results on CLEVR



Model Results on GQA



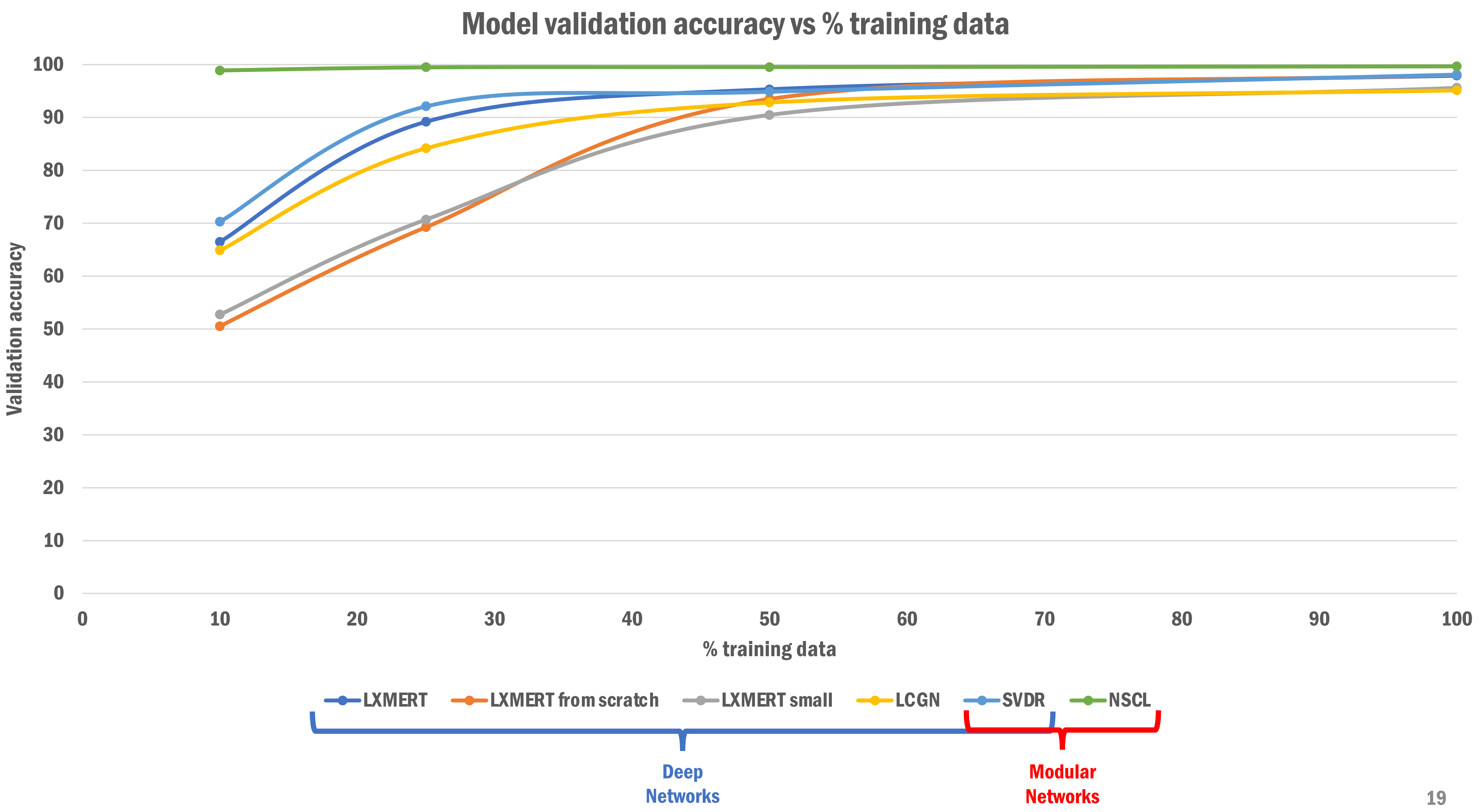
Objective: Do More with Less

- Datasets contain large number of data points.

- Data is cleanly annotated and rich.

- How can we perform well on less data?

Model Results on CLEVR

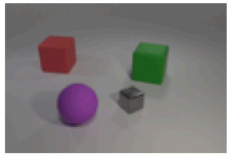


Model Results on CLEVR – NSCL Strategies

Curriculum Learning

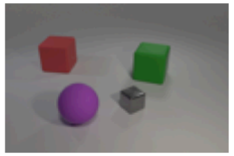
Initialized with DSL and executor.

Lesson1: Object-based questions.



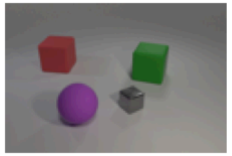
Q: What is the shape of the red object?
A: Cube.

Lesson2: Relational questions.



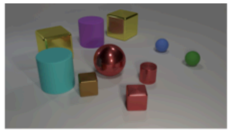
Q: How many cubes are behind the sphere?
A: 3

Lesson3: More complex questions.



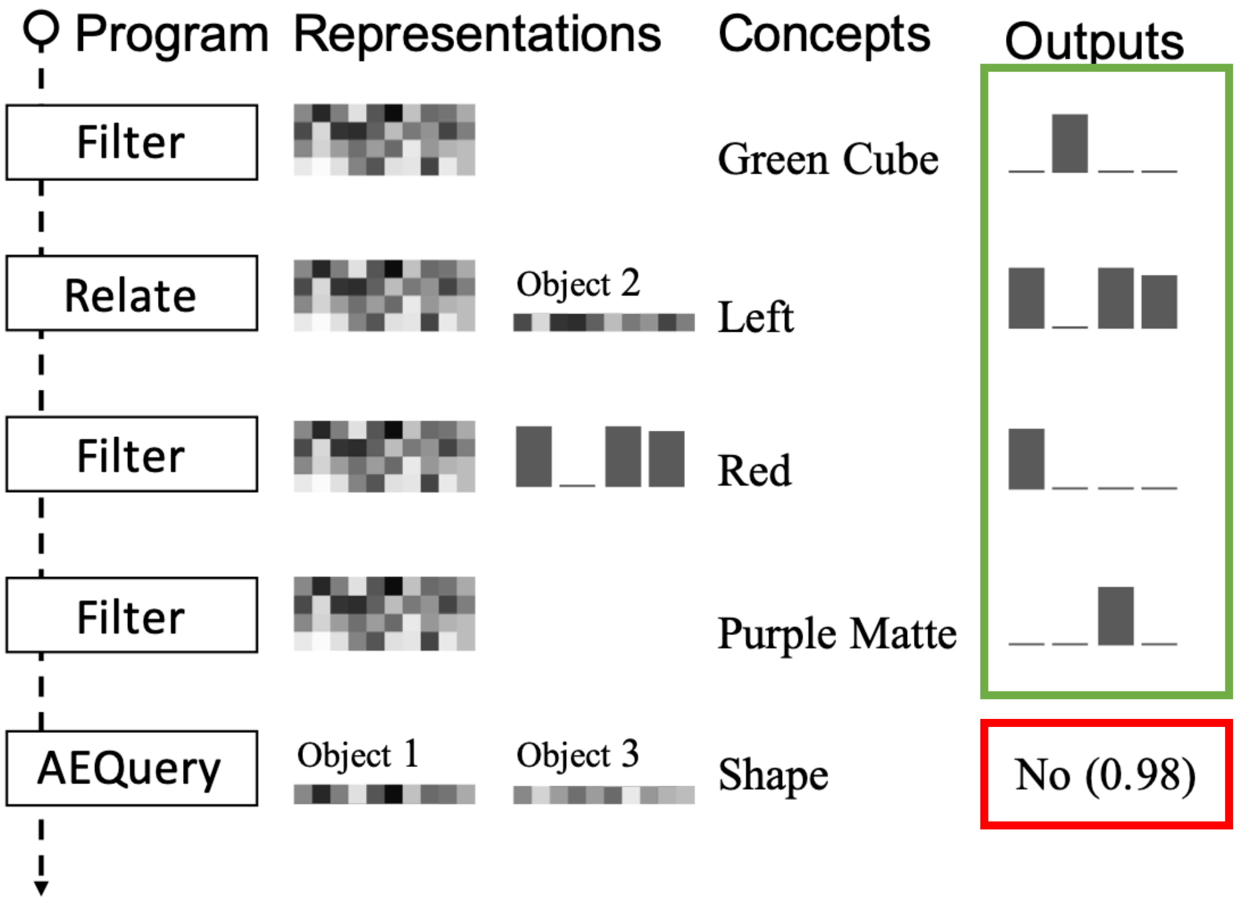
Q: Does the red object left of the green cube have the same shape as the purple matte thing?
A: No

Deploy: complex scenes, complex questions



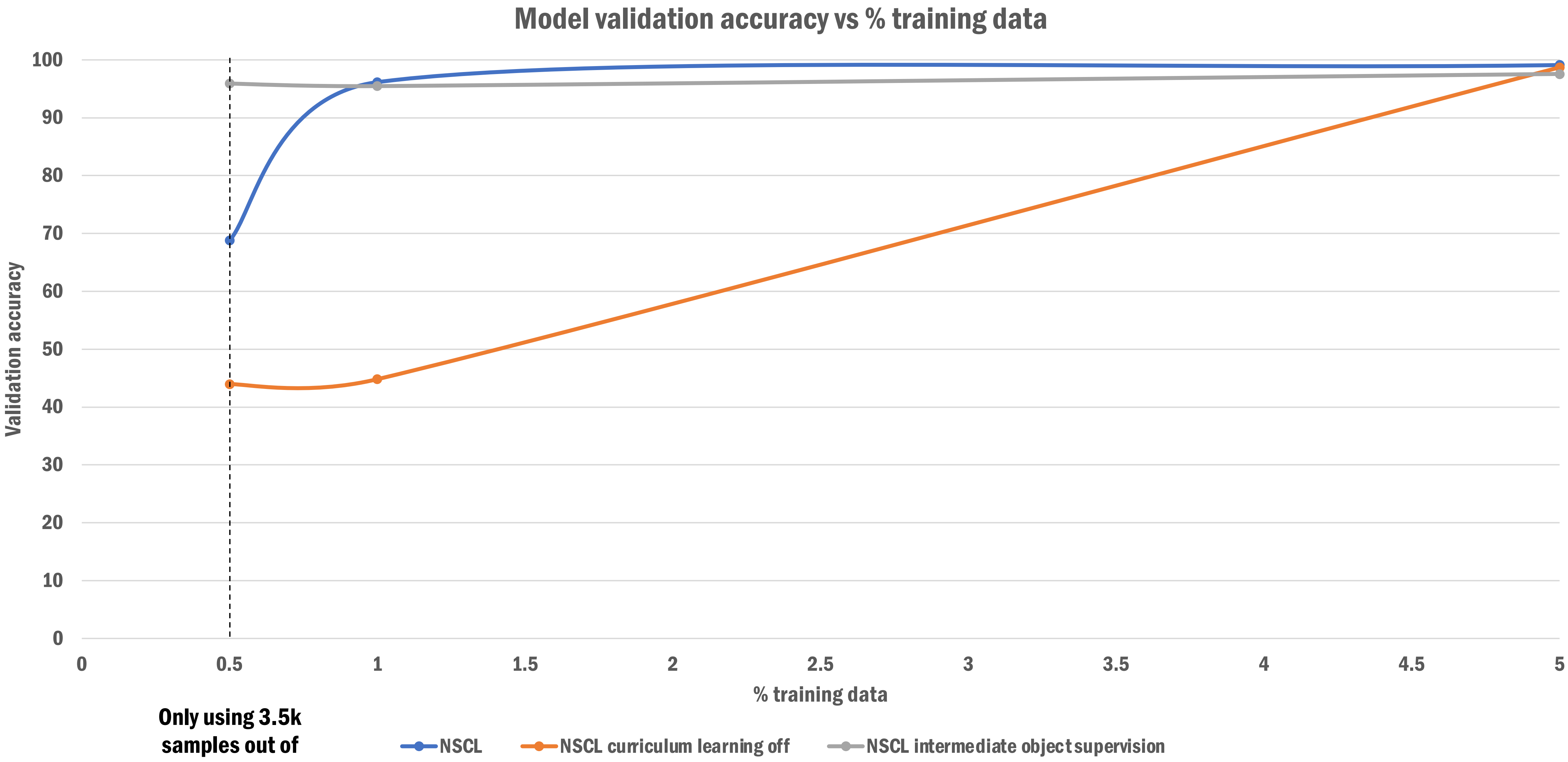
Q: Does the matte thing behind the big sphere have the same color as the cylinder left of the small matte cube?
A: No.

End-to-end vs Intermediate Supervision



Mao, et al. The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences from Natural Supervision. ICLR’19

Model Results on CLEVR – Less Labels Results

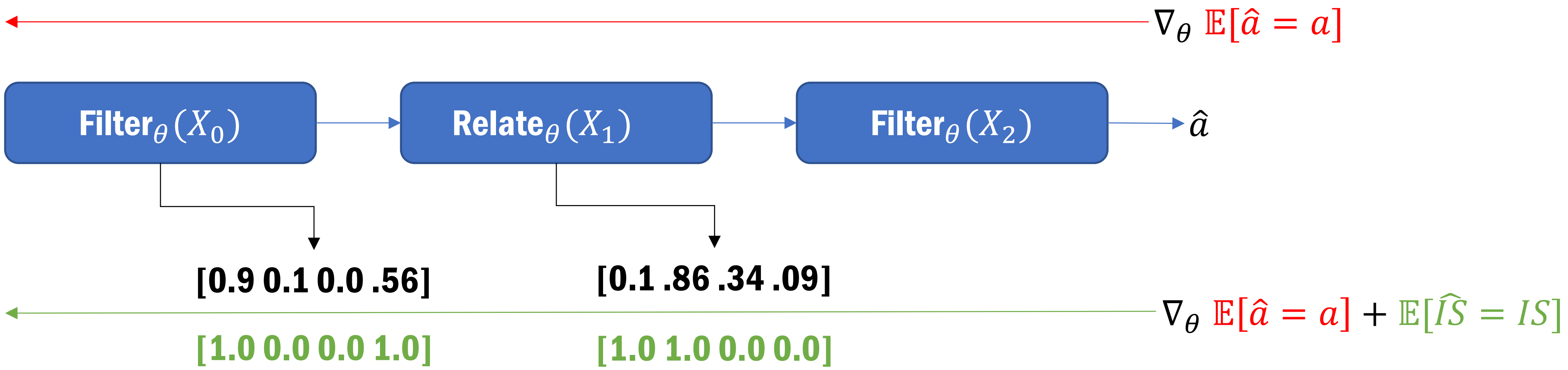


Only using 3.5k samples out of 700k (0.5%)!

Research Directions

Modular Network functions are hand coded and must be differentiable.

Using neural networks instead are simple to instantiate, but difficult to train end-to-end.

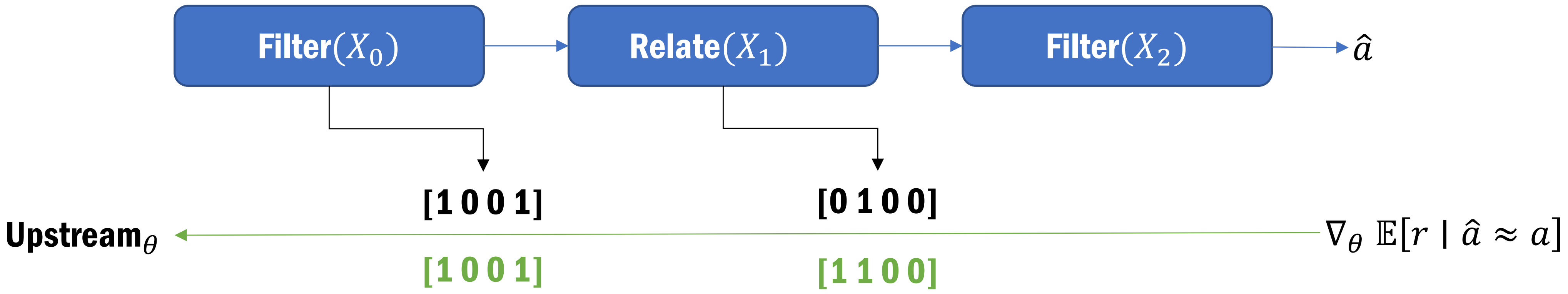


Can we leverage our intermediate supervision (IS) results efficiently?

Evaluate function complexity versus label requirements.

Research Directions

Similarly, can we use generic functions that *do not* have to be differentiable?



We need to leverage reinforcement; how should we define our reward?

Labeling by Abduction

- What if we don't have intermediate supervision?

- We have to enumerate the possible intermediate labels till the answer is satisfied

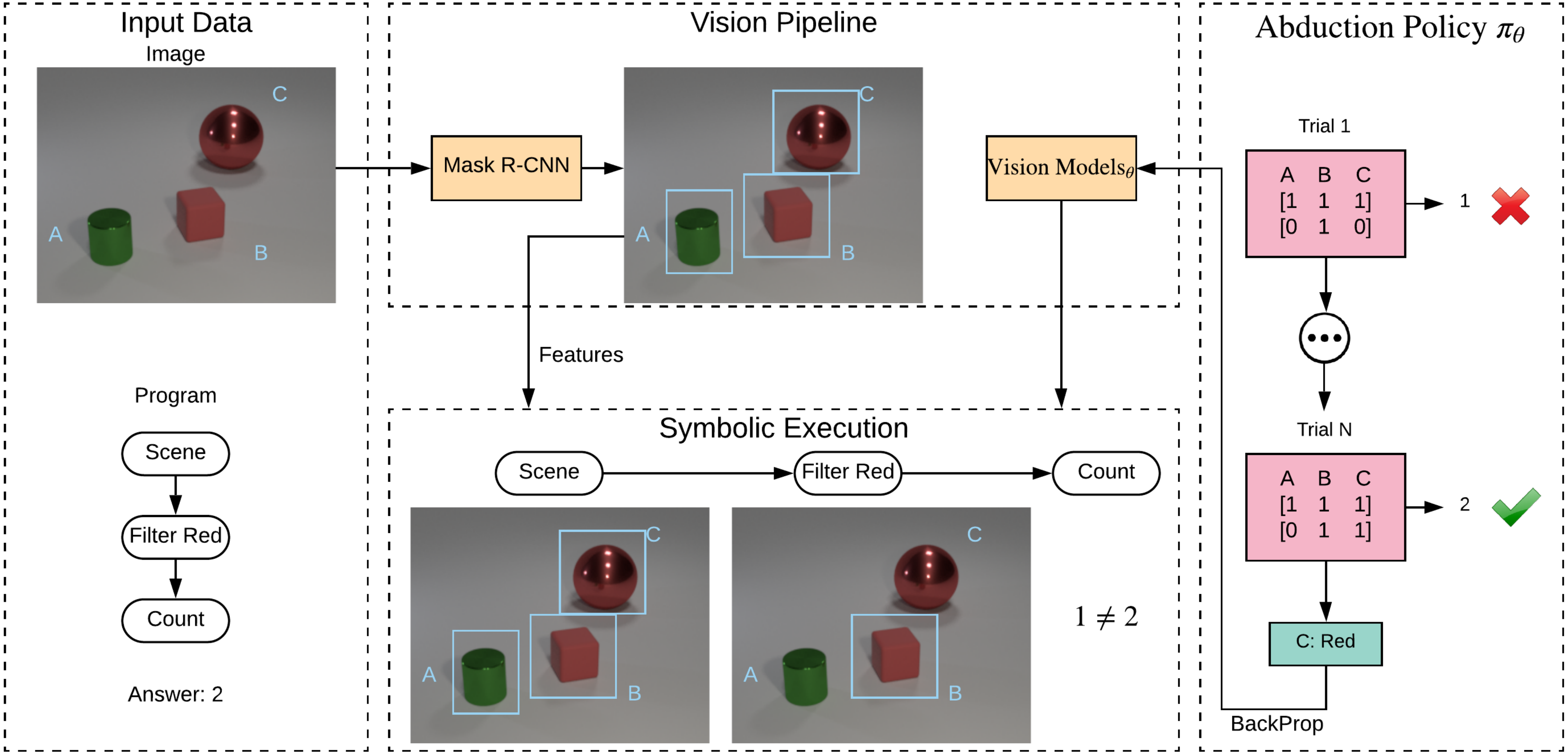
- Given our current predictions, what is the simplest change needed? Explore abductive reasoning!

$$Background \cup\ Hypothesis \vDash Observation$$

$$Modules \cup Pedictions \vDash Answer$$

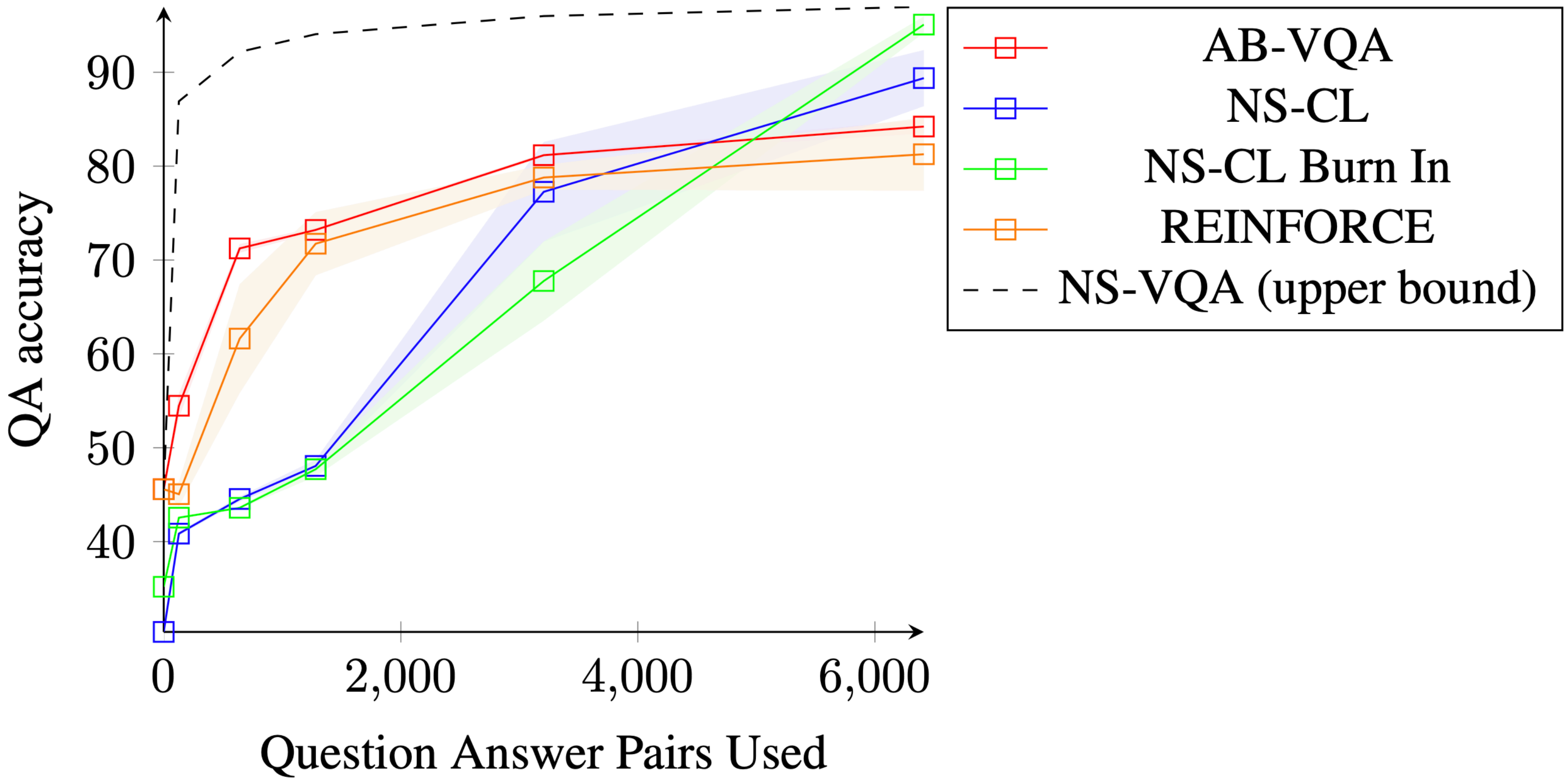
Labeling by Abduction

What are the minimal changes needed to “correct” the predicted intermediate labels



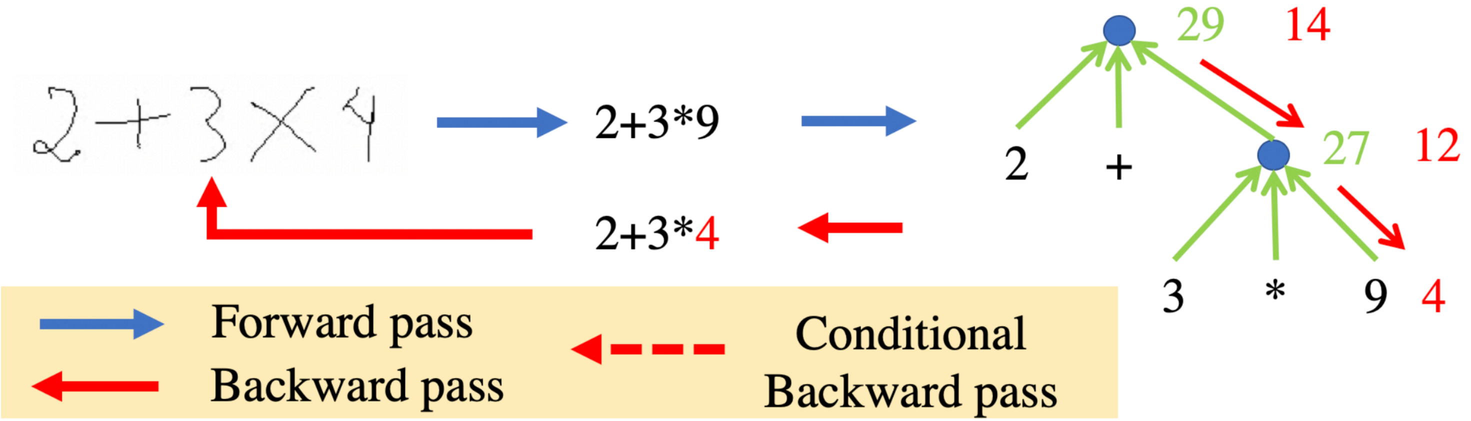
Results on CLEVR

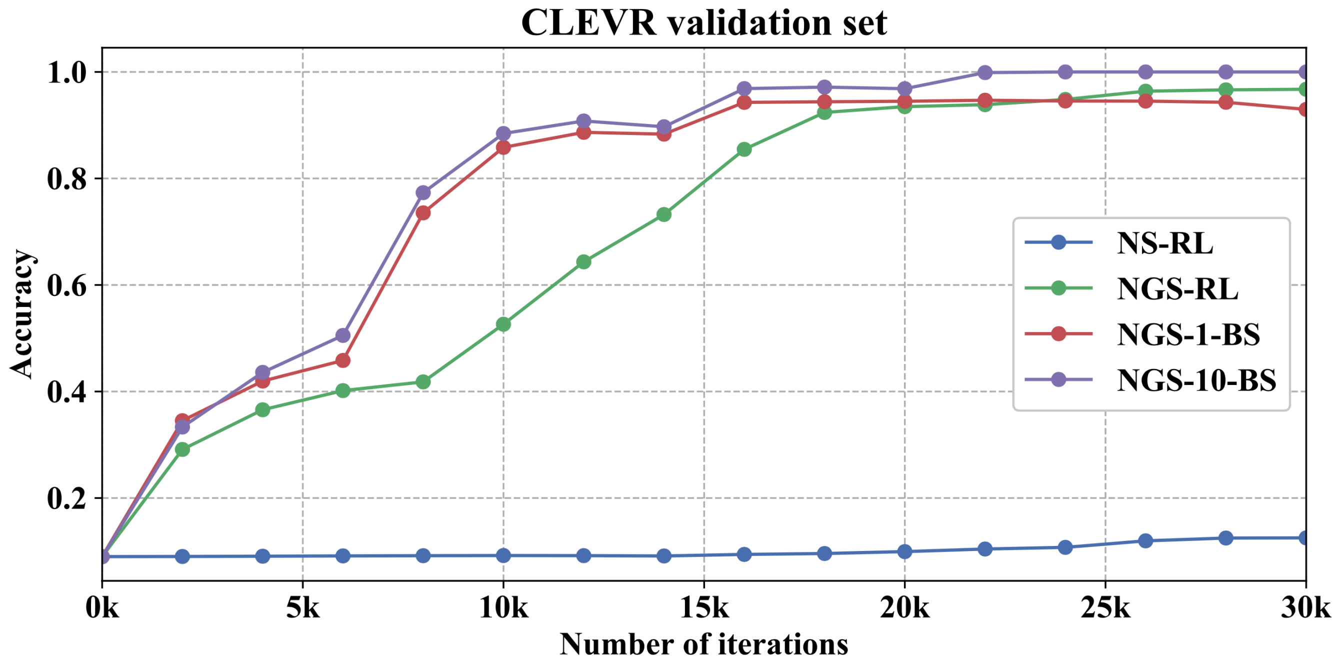
- Trained on 7 supervised image scene graphs (7/70k = 0.01%)
- Minimize edit distance between predicted and abduced labels



Labeling by Back Search

When encountering an incorrect answer work backwards to find corresponding labels.





Li, et al. Closed Loop Neural-Symbolic Learning via Integrating Neural Perception, Grammar Parsing, and Symbolic Reasoning. ICML'20

Visual Reasoning

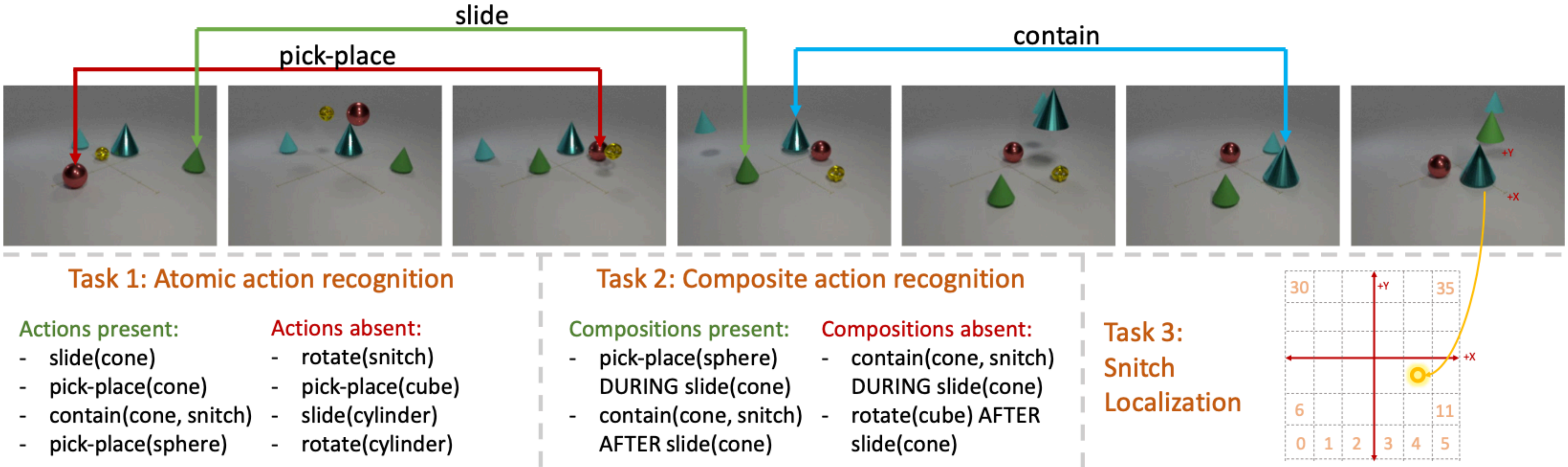
Part II: Video Understanding

Video Reasoning

- Introduce an extra temporal dimension to our VQA task.
- Many applications in video summarization and understanding.
- Methods involve a large end to end network but cannot reliably capture the state space over long periods of time.
- Can we model this state space discretely for reasoning tasks?

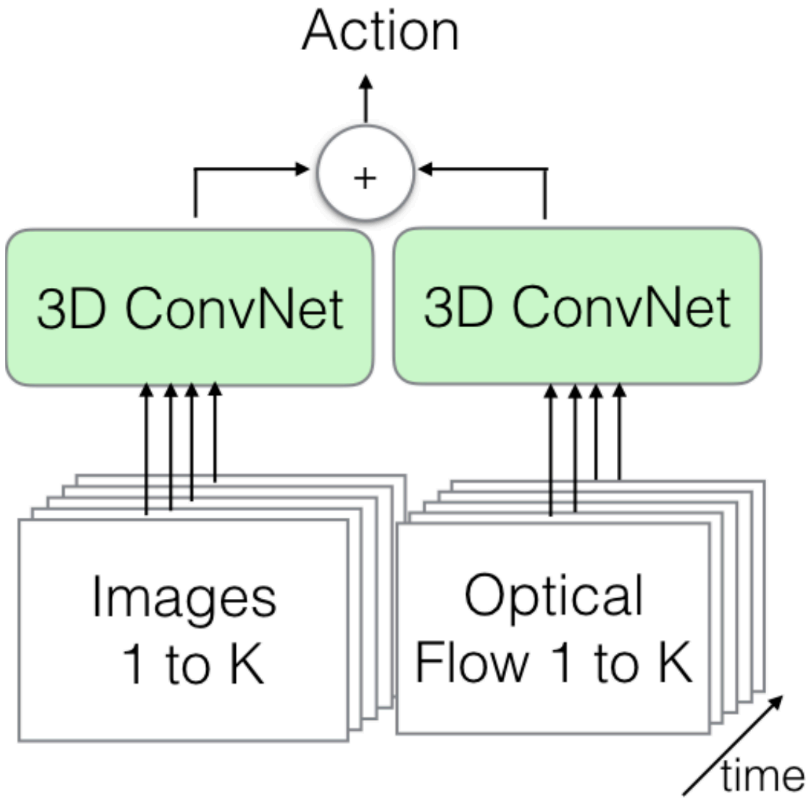
CATER Dataset

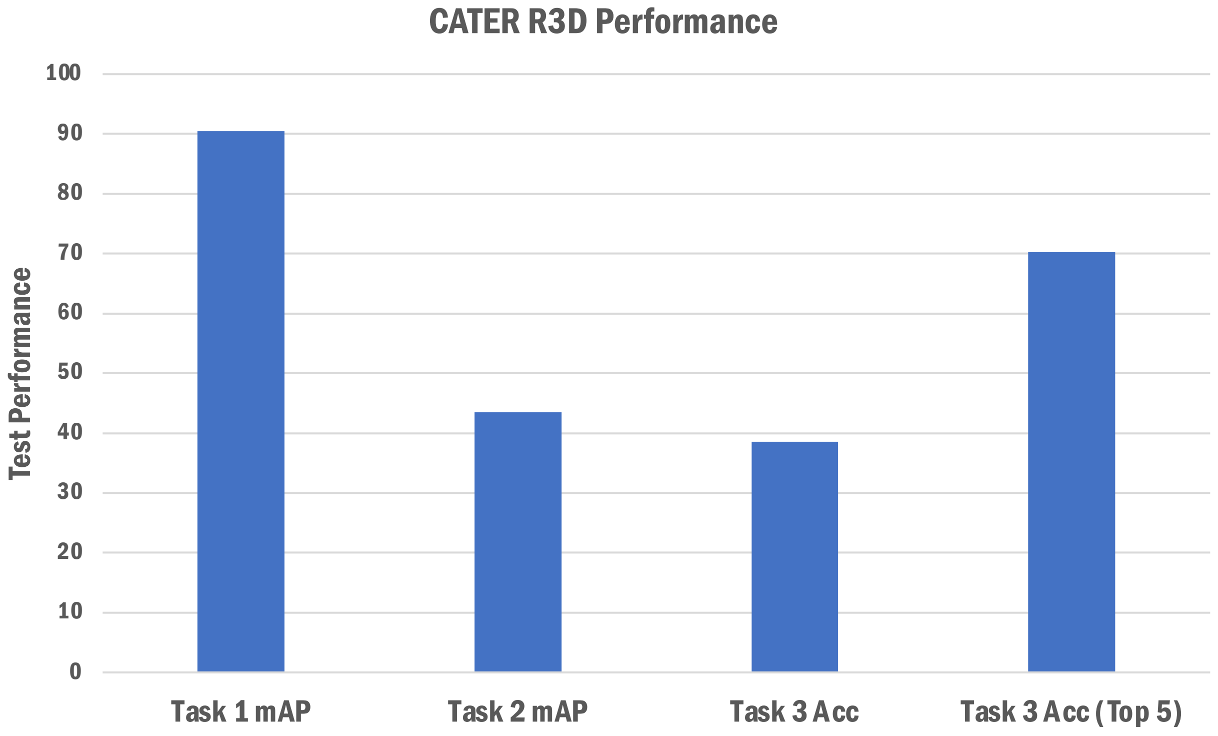
- **4 basic atomic action events, rotate, relocate, slide, contain.**
- **Multiple reasoning tasks.**



CATER Dataset: Baseline

Tested on variations of a 3D CNN model (R3D).





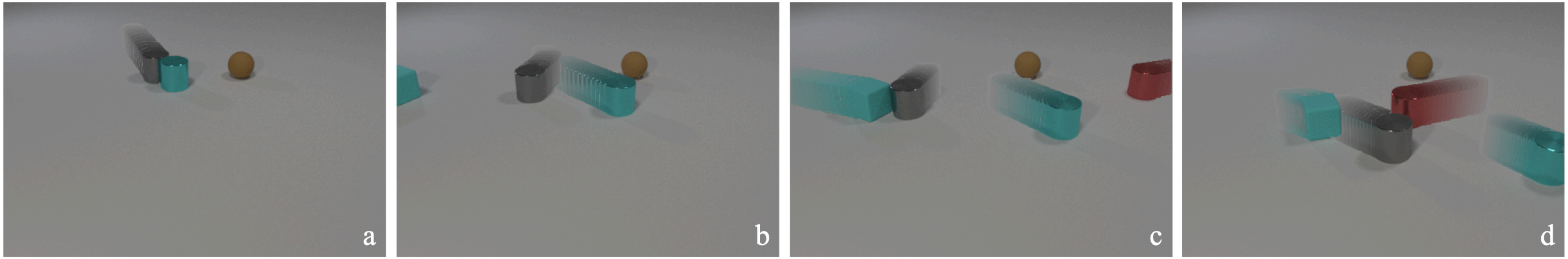
Carreria, Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. NeurIPS'19

CLEVRER Dataset

Physics dynamics of moving objects.

Introduces multiple question types:

(a) First collision (b) Cyan cube enters (c) Second collision (d) Video ends



I. Descriptive

Q: *What shape is the object that collides with the cyan cylinder?* **A:** *cylinder*

Q: *How many metal objects are moving when the video ends?* **A:** *3*

II. Explanatory

Q: *Which of the following is responsible for the gray cylinder's colliding with the cube?*

a) The presence of the sphere

b) The collision between the gray cylinder and the cyan cylinder **A:** *b)*

III. Predictive

Q: *Which event will happen next*

a) The cube collides with the red object

b) The cyan cylinder collides with the red object **A:** *a)*

IV. Counterfactual

Q: *Without the gray object, which event will not happen?*

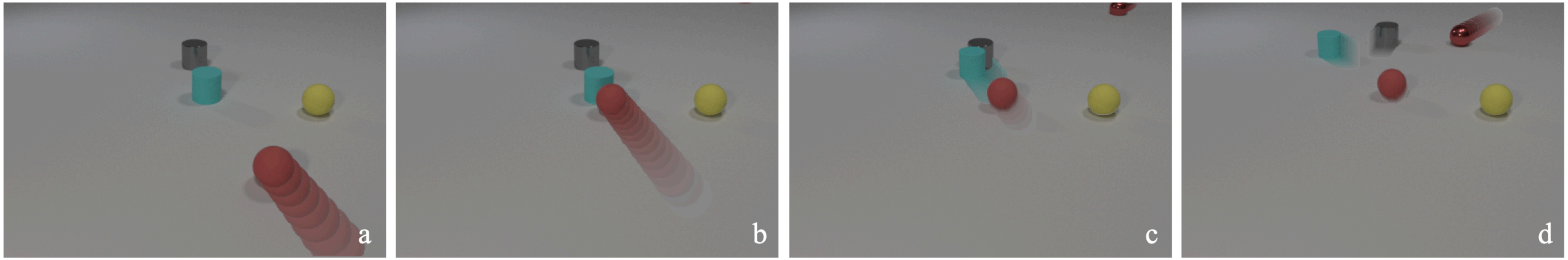
a) The cyan cylinder collides with the sphere

b) The red object and the sphere collide **A:** *a), b)*

Yi, et al. CLEVRER: Collision Events for Video Representation and Reasoning. ICLR'20

CLEVRER Dataset

Like CLEVR, the ground truth interactions and the program structure are provided.



Objects					
ID	1	2	3	4	5
Color	Cyan	Gray	Yellow	Red	Red
Material	Rubber	Metal	Rubber	Rubber	Metal
Shape	Cylinder	Cylinder	Sphere	Sphere	Sphere

Events					
Mode	Observation			Pred.	CF.
Frame	50	65	70	155	70
Type	Collision	Enter	Collision	Collision	Collision
Object ID	1, 4	5	1, 2	4, 5	2, 4

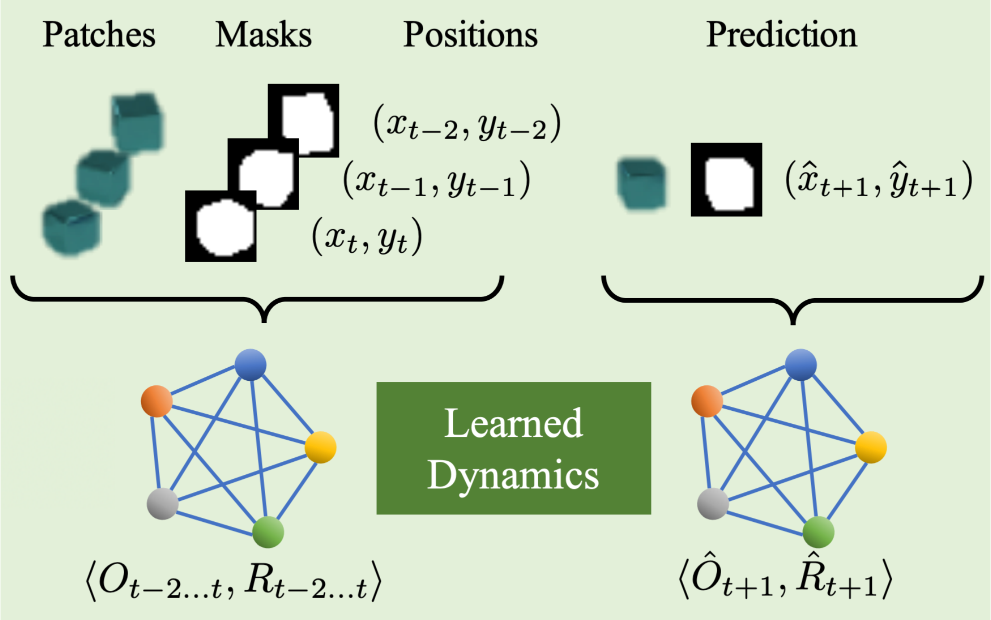
Question: *What shape is the first object to collide with the cyan object?*

Program: query_shape(get_col_partner(filter_order(filter_collision(Events, filter_color(Objects, Cyan)), First), filter_color(Objects, Cyan)))

Answer: *Sphere*

CLEVRER Dataset: Baselines

Baselines includes a physics propagation network (NS-DR) and a GQA model baseline (MAC).



Methods	Descriptive	Explanatory		Predictive		Counterfactual	
		per opt.	per ques.	per opt.	per ques.	per opt.	per ques.
NS-DR	88.1	87.6	79.6	82.9	68.7	74.1	42.2
NS-DR (NE)	85.8	85.9	74.3	75.4	54.1	76.1	42.0
MAC (V+)	86.4	70.5	22.3	59.7	42.9	63.5	25.1

Li, et al. Propagation Networks for Model-Based Control under Partial Observation. ICRA'19
Hudson, Manning. Compositional Attention Networks for Machine Reasoning. ICLR'18

Research Directions

- CATER contains a variety of event models while CLEVRER contains events and a flavor of time series forecasting.
- Current methods rely on modeling temporal interactions in a probabilistic setting
  - R3D: Encodes representation in a temporal recurrent model, requires more complex or longer videos to generalize well. This is a state space retrieval.
  - NS-DR: Models are similarly defined by recurrent networks and have a graphical flavor of the interactions between objects. This is state space inference.
- Can we leverage explicit state spaces and reason over the tasks using temporal logic?

Conclusion

- Rich applications in multi-modal vision and text reasoning.

- Most dataset challenges are label rich, unlike real world tasks.

- There is inherent structure in reasoning tasks, how do we leverage these to build label efficient models?

Machine Reasoning: A Vision Perspective

Karan Samel

ksamel@gatech.edu

Georgia Institute of Technology – 7/2/20