Towards Generalizable Visual Reasoning

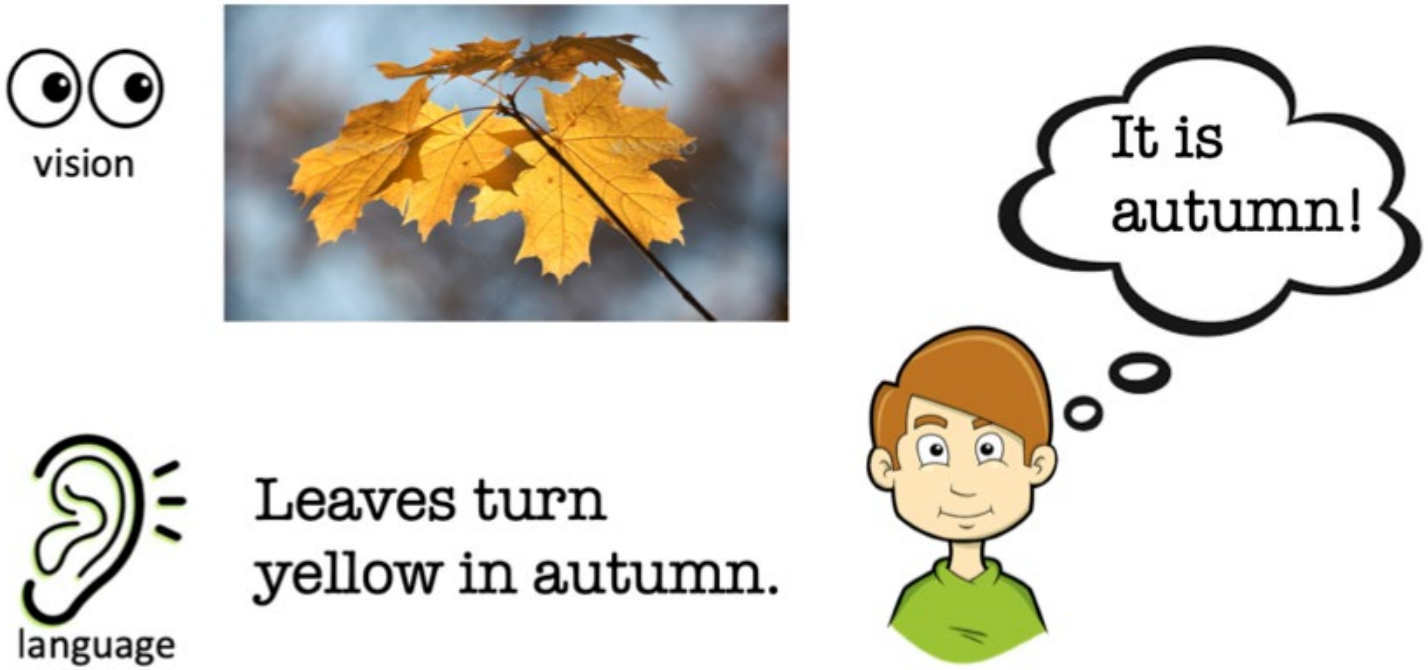
Zhuowan Li

The Department of Computer Science

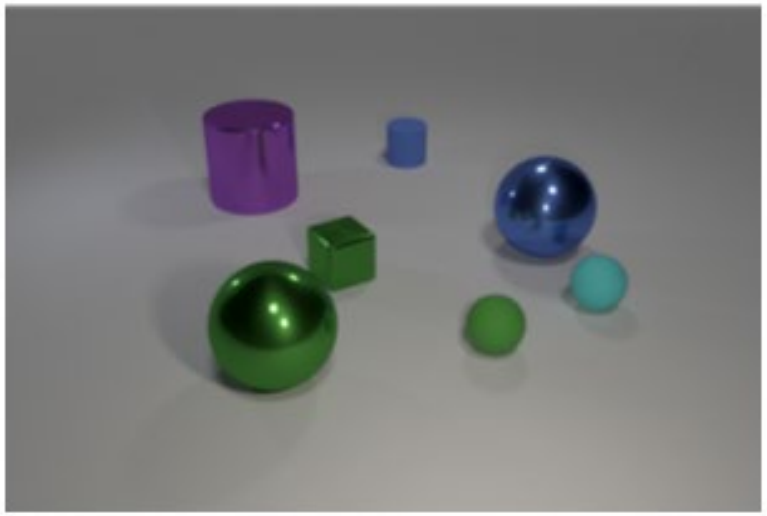
Johns Hopkins University

Advisors: Alan Yuille, Benjamin Van Durme

Humans reason with vision and language



Visual Question Answering is a challenging task

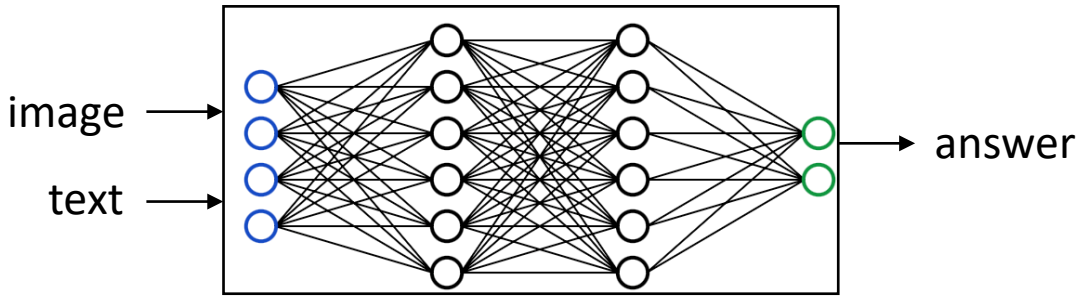


Q: The cylinder that is the same size as the metallic sphere is what color?
A: Purple

Q: What does the little boy in front of the table hold?
A: Toothbrush

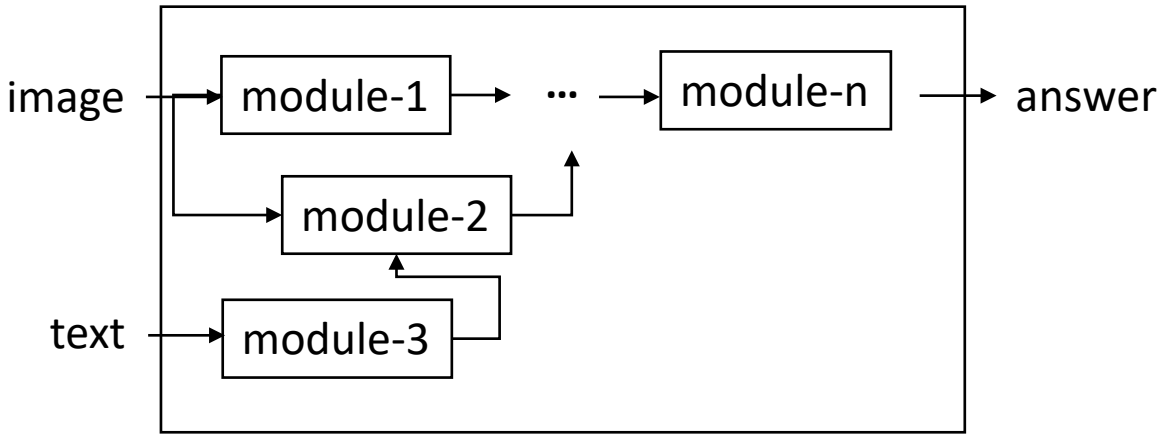
A paradox in VQA

Standard end-to-end models



- Perform well on IID setting
- Not robust to distribution shifts

Neural modular methods



- More robust
- Not perform well on standard real datasets

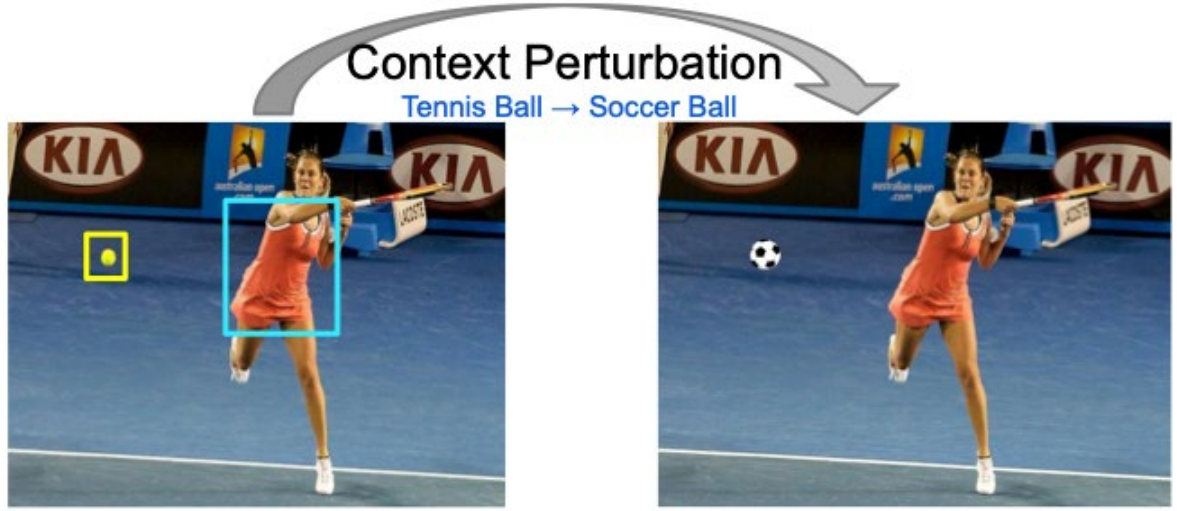
We want to understand this better

Standard VQA models are not robust

Text shortcuts

- “what color..” → white
- “is there..” → yes
- “how many..” → 2
- “What color is the banana” → yellow

Visual contexts

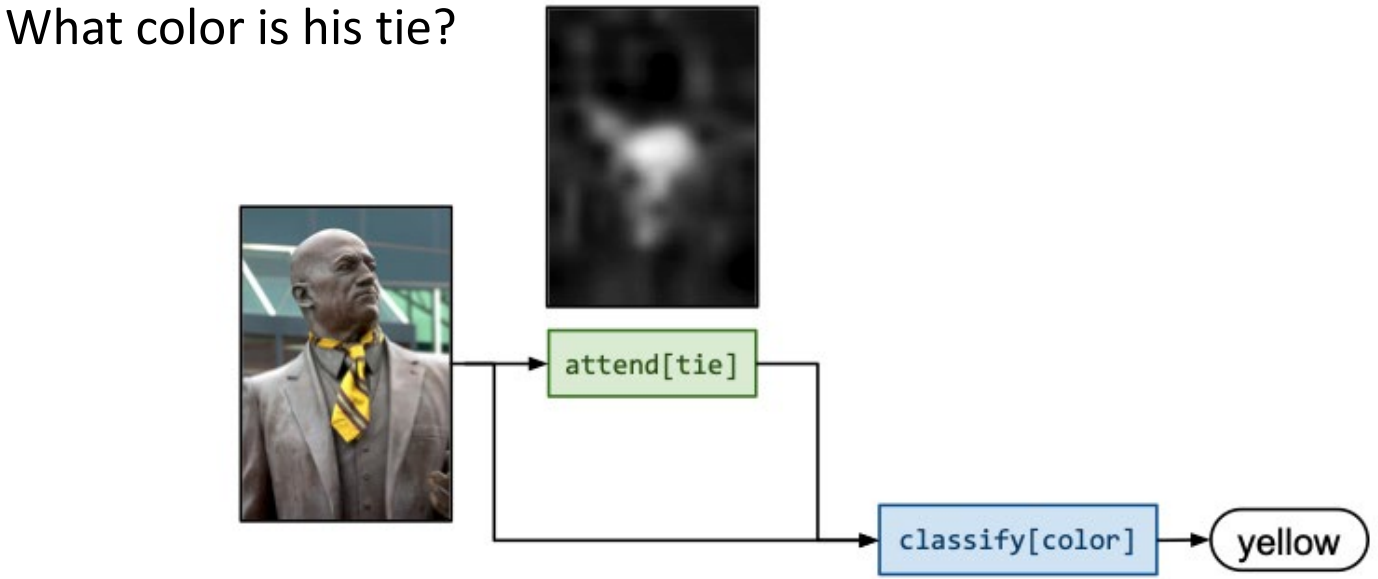


Question: What color is the woman’s dress?
Ground-truth answer: Orange.
Model prediction: Orange. ✓ **Model prediction:** White. ✗

Swapmix: Diagnosing and regularizing the over-reliance on visual context in visual question answering.
Vipul Gupta, **Zhuowan Li**, Adam Kortylewski, Chenyu Zhang, Yingwei Li, Alan Yuille.
In CVPR 2021.

An alternative: Neural Modular Methods

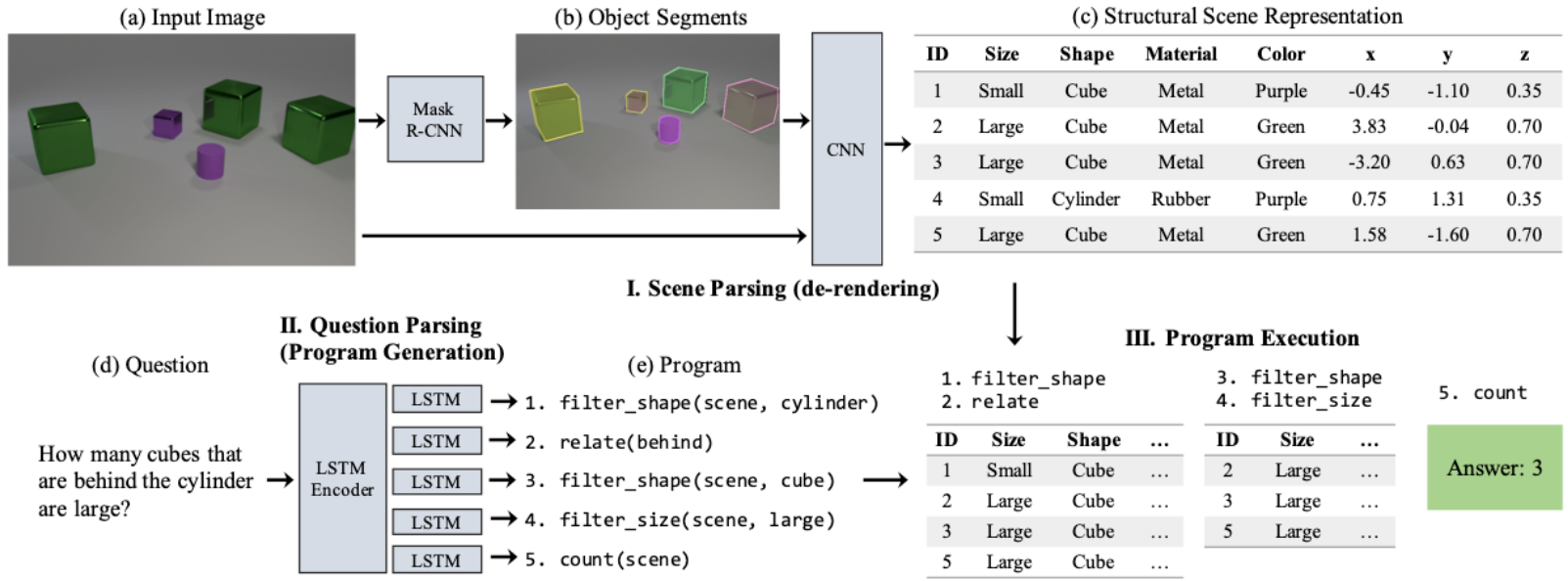
- Parse a question into a series of operations
- Each operation is implemented as a separate neural module

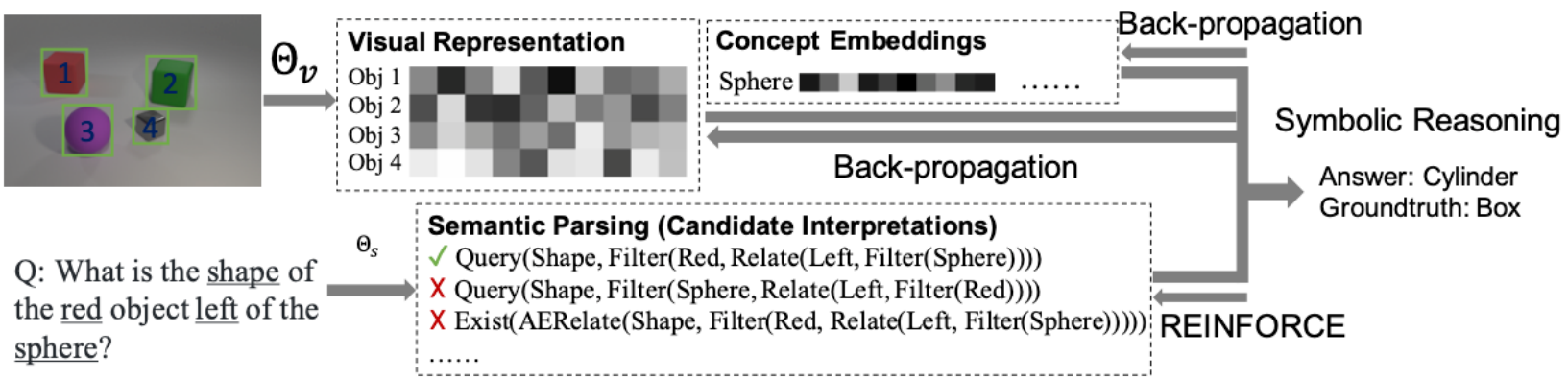


Andreas, Jacob, et al. "Neural module networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

An alternative: Neural Modular Methods

Neural Symbolic VQA





Neural Symbolic Concept Learner

Yi, Kexin, et al. "Neural-symbolic vqa: Disentangling reasoning from vision and language understanding." *NeurIPS* 2018.

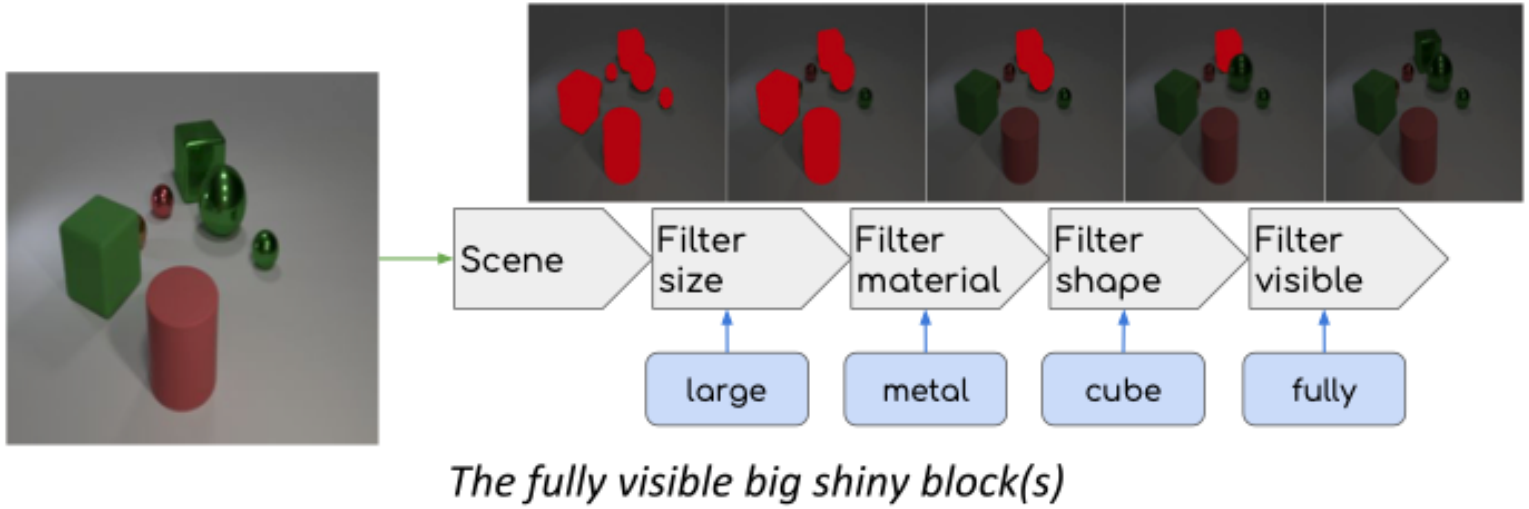
Mao, Jiayuan, et al. “The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision.” ICLR 2019.

Neural modular methods

Pros:

- Interpretable

CLEVR-Ref+ (by CCVL)
CVPR 2019



- Data-efficient
- Robust
- SoTA performance on CLEVR
- …

Liu, Runtao, et al. "Clevr-ref+: Diagnosing visual reasoning with referring expressions." *CVPR*. 2019.

Neural modular methods

Pros:

- Interpretable
- Data-efficient
- Robust (Need to be verified)
- SoTA performance on CLEVR
- …

Cons:

- Need explicit reasoning programs (partly addressed by NSCL)
- Low performance on real images

Neural modular methods

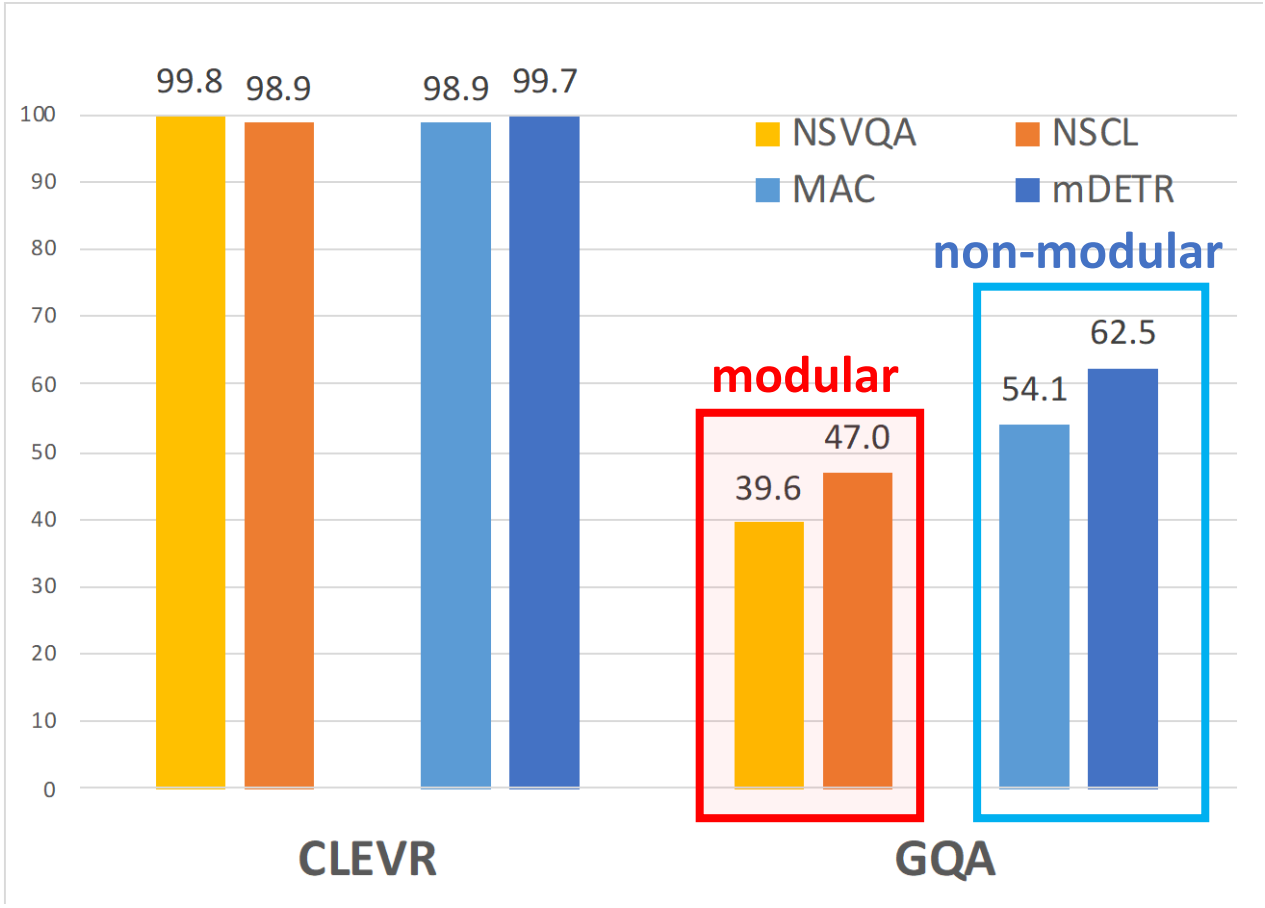
Pros:

- Interpretable
- Data-efficient
- Robust (Need to be verified)
- SoTA performance on CLEVR
- …

Cons:

- Need explicit reasoning programs (partly addressed by NSCL)
- **Low performance on real images**

Modular methods suffer on real data



Outline

- Why do symbolic methods suffer on real images? How to improve them? [ICCV 2021]
- Super-CLEVR: How to study domain robustness more systematically? [CVPR 2023 Highlight]
- Extension of Super-CLEVR with Part, Pose, Occlusion [Ongoing]

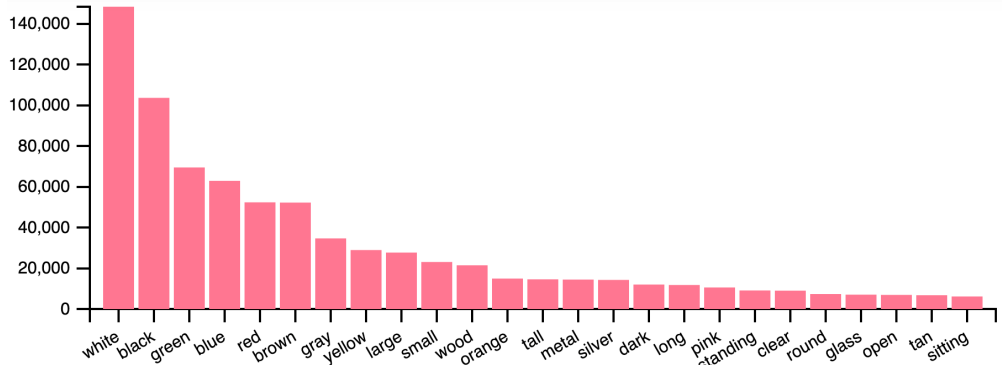
Why do Symbolic Methods suffer on Real Images? How to improve them?

Calibrating Concepts and Operations: Towards Symbolic Reasoning on Real Images.
Zhuowan Li, Elias Stengel-Eskin, Yixiao Zhang, Cihang Xie, Quan Hung Tran, Benjamin Van Durme, Alan Yuille
In ICCV 2021.

Overview: two reasons, and solutions

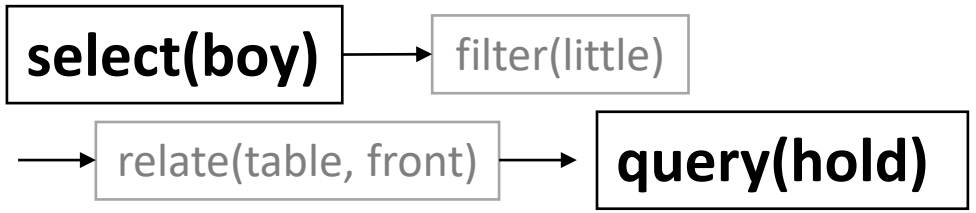
Reason-1:

Long-tail distribution



Reason-2:

Unequal importance of reasoning steps



What does the little boy in front of the table hold?

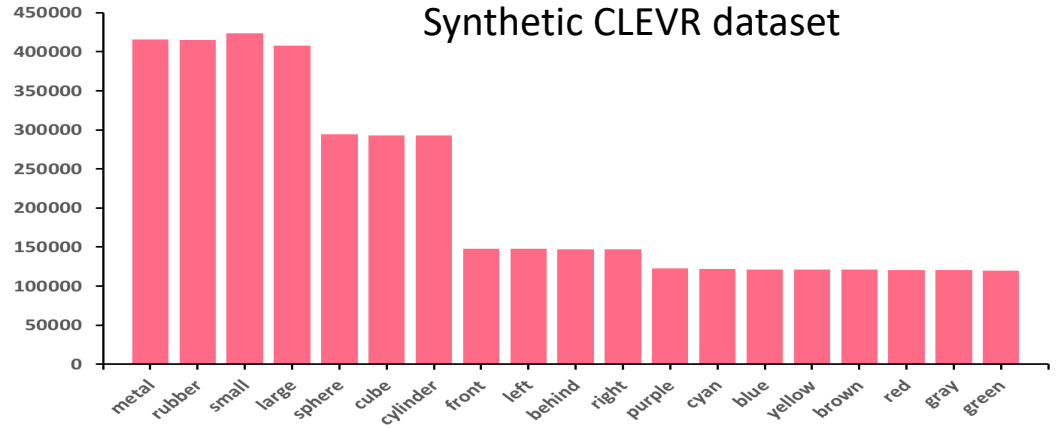
Solution-1:

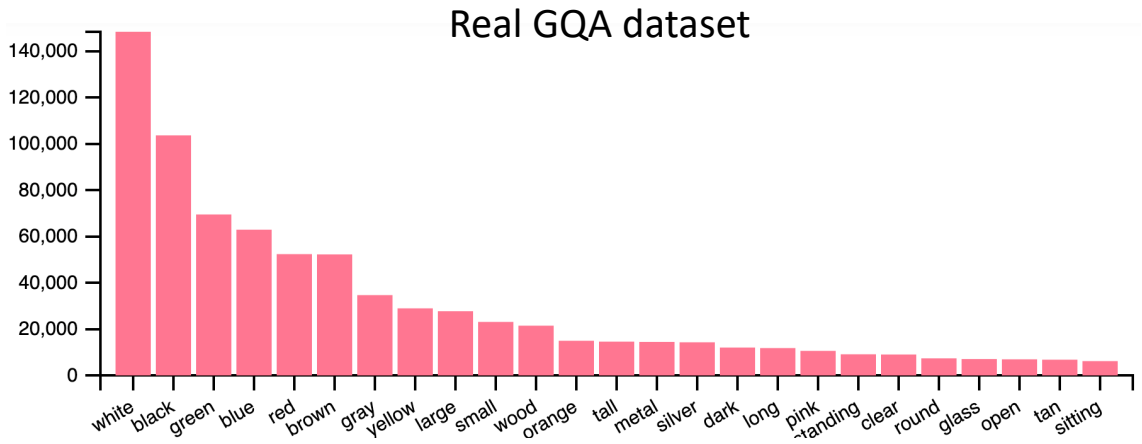
Calibrating concepts

Solution-2:

Calibrating operations

1. Real data suffers from long-tailed distribution



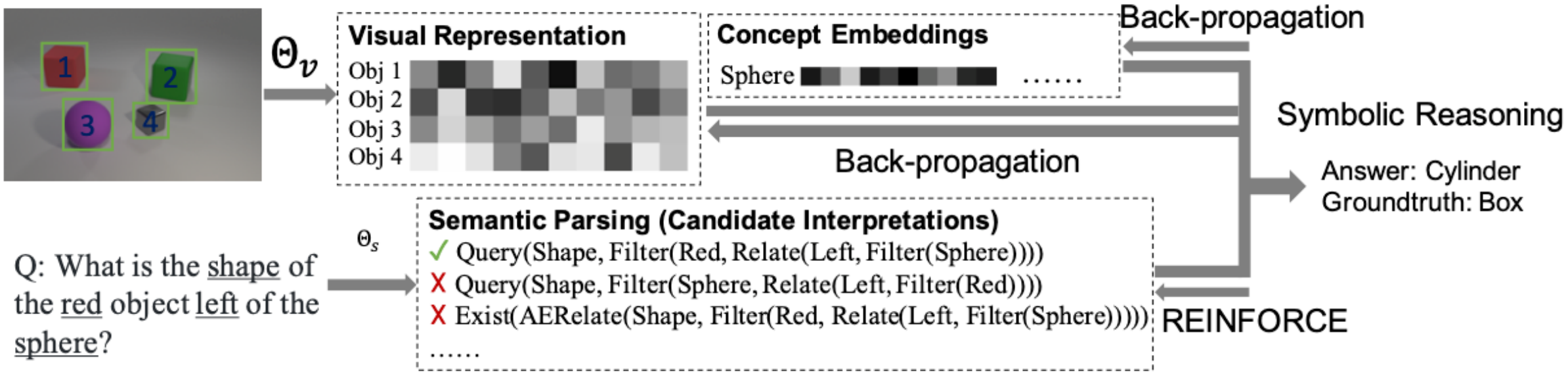


white, black, green, …

tan, sitting, open, …

How do long tails affect concept learning?

Recall NSCL:



Normalized Concept embeddings hinder the learning the concept distributions

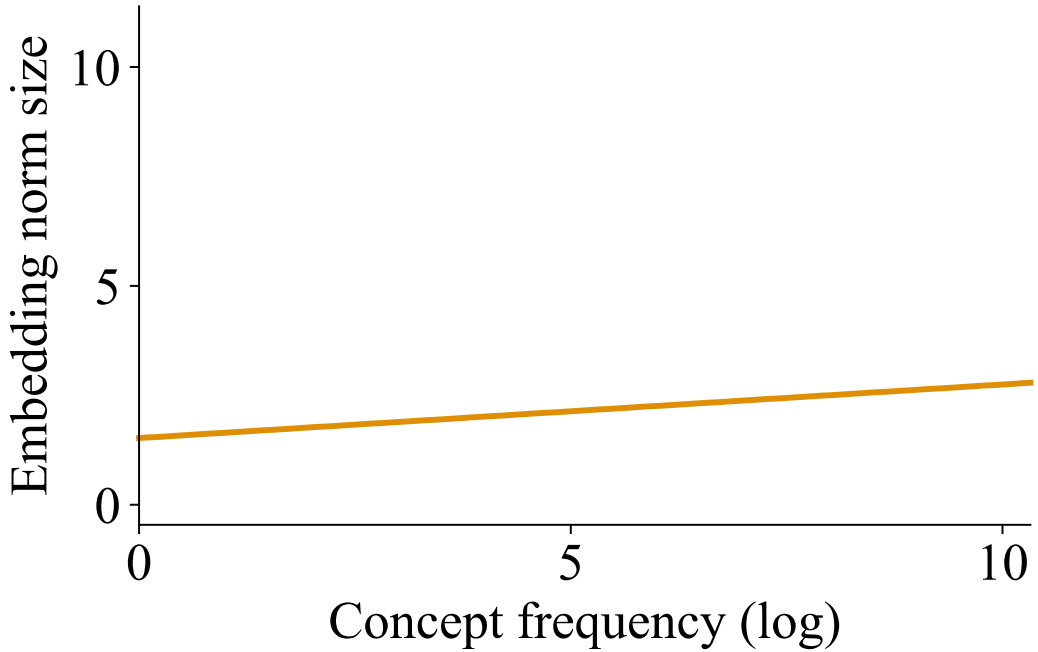
Mao, Jiayuan, et al. "The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision." ICLR 2019.

How does this distribution affect concept learning?

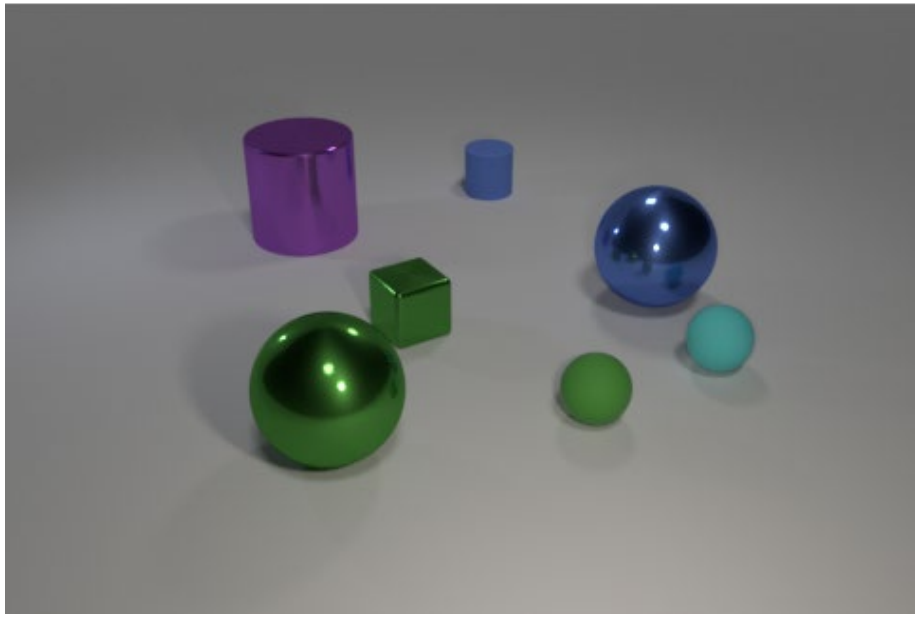
- By default, concept embeddings are normalized.
- Simply removing the normalization on concept embedding yields substantially better performance (+3.4%)
- **The magnitudes of the embeddings matter!**

Unnormalizing concept weights

Positive correlation between concept frequency and embedding norm size



2. Reasoning steps are of unequal importance



Q: The cylinder that is the same size as the metallic sphere is what color?
A: Purple
Prog: select(sphere) → filter(metallic) → samesize() → filter(cylinder) → querycolor()

Q: What does the little boy in front of the table hold?
A: Toothbrush
Prog: select(boy) → filter(little) → relate_s(table, front) → query_rel_o(hold)

Intuition: what if we adjusting operation weights?

Question: Is there a bag in this image that is not black?

Groundtruth: No

(1) Select(bag) scores:

[−7.0, −6.0, **2.1**, −9.9]

(2) Filter(not black) scores:

[0.8, −0.7, **−1.7**, 2.1]

Merge: (1) + (2):

[−6.2, −5.3, **0.4**, −7.8] → Exist? → ✗ Answer: Yes

With weight: (1) + 2*(2)

[−5.4, −4.6, **−1.3**, −5.7] → Exist? → ✓ Answer: No

Method: Calibrating Concepts and Operations

- Calibrating concepts
  - Explicitly learnable norm size for each concept:

$$\mathbf{c}_{concept} = w^{type}_{concept}\mathbf{c}_{concept}$$

- Calibrating operations
  - Train a LSTM weight predicter to predict weights for each module using context
  - Merge operation results with learned weights:

$$\mathbf{d}_i = \sum_{j \in \mathcal{D}(p_i)} w_j \mathbf{d}_j$$

Calibrating operations:
predict weights for each operation

Answer: horse

Scene Representation

horse
spectator
...
hurdle

Faster-RCNN

Executor

Query ← d
Select → $* w_3$
Relate → $* w_2$
Select → $* w_1$

Operation Weight Predictor

LSTM ← $(p_1, p_2, \ldots, p_m)$

Program

Query(name)
Relate_s(left)
Select[name] (animal)
Select[name](spectator)

Question: How is the animal to the left of the spectators called?

Semantic Repr.
Object embeddings
gate
MLP
MLP
M_{attr}
d
Positional Repr.
Learnable concept weight
$w *$
Concept embedding

Calibrating Concepts:
Learn magnitudes for each concept in each module

Both the concept and operation calibrations help!

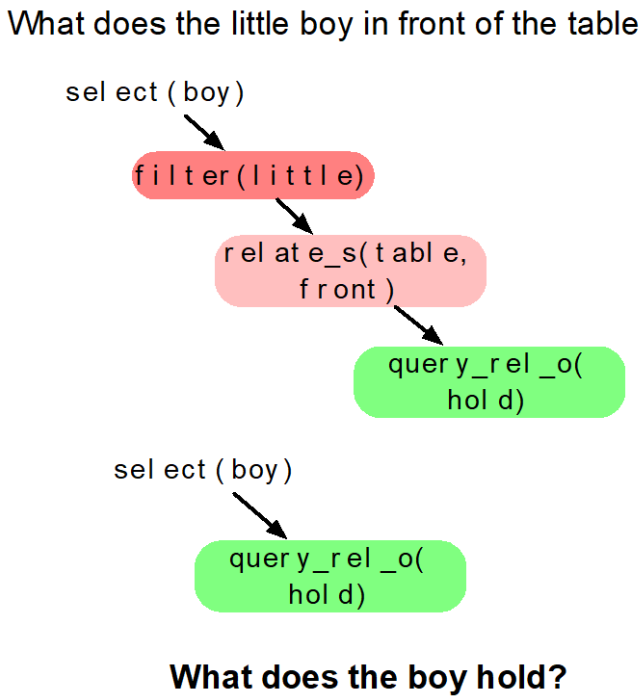
Model Level	Concept	Operation	Accuracy
1 (Baseline)	Normalized	Average	47.01
2	Normalized	Calibrated	51.03
3	Unnormalized	Calibrated	54.65
4 (Ours)	**Calibrated**	**Calibrated**	**56.13**

Our method helps bridge the gap

	Model	Accuracy
non-symbolic	LXMERT (Tan and Bansal, 2019)	60.33
	NSM (Hudson and Manning, 2019)	63.17
	MNM (Chen et al., 2021)	60.83
symbolic	∇-FOL (Amizdeh et al., 2020)	54.76
	CCO (Ours, 2021)	**56.38**

Analysis of operation weights

- Prune low-weight modules progressively from the question
- The proposed perturb split can be used to analyze model behaviors



Analysis: Perturbed Test Split

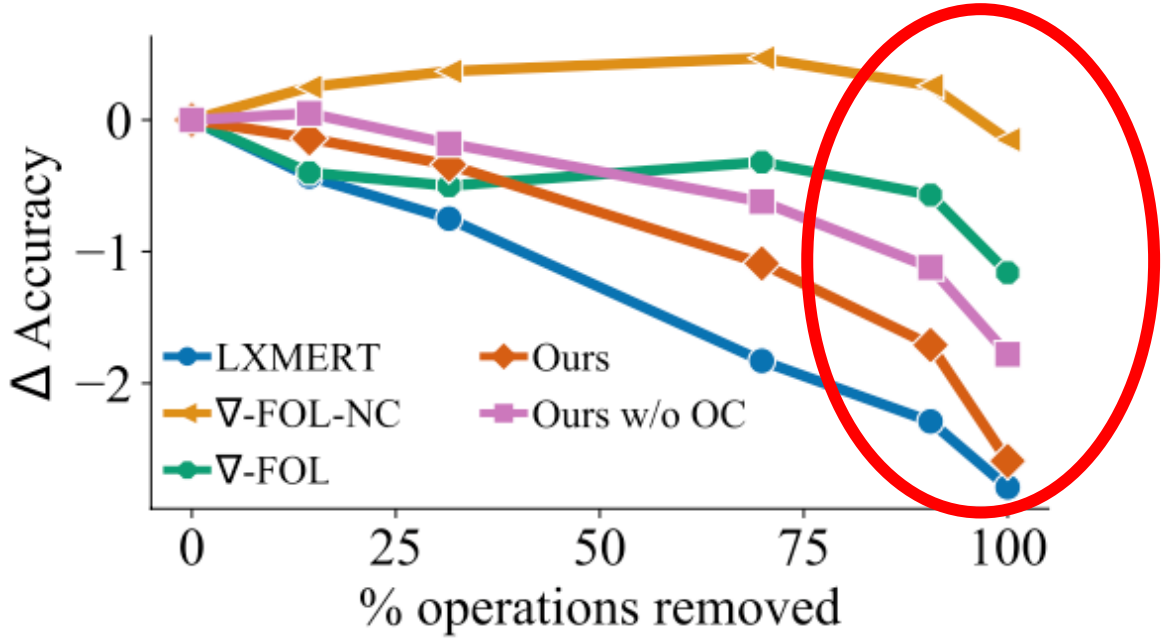


Figure 5: Accuracy drop of different models when the testing questions are progressively perturbed by removing reasoning operations with low weights.

The weight predictor assigns higher weights to more important operations.

Analysis: Perturbed Test Split

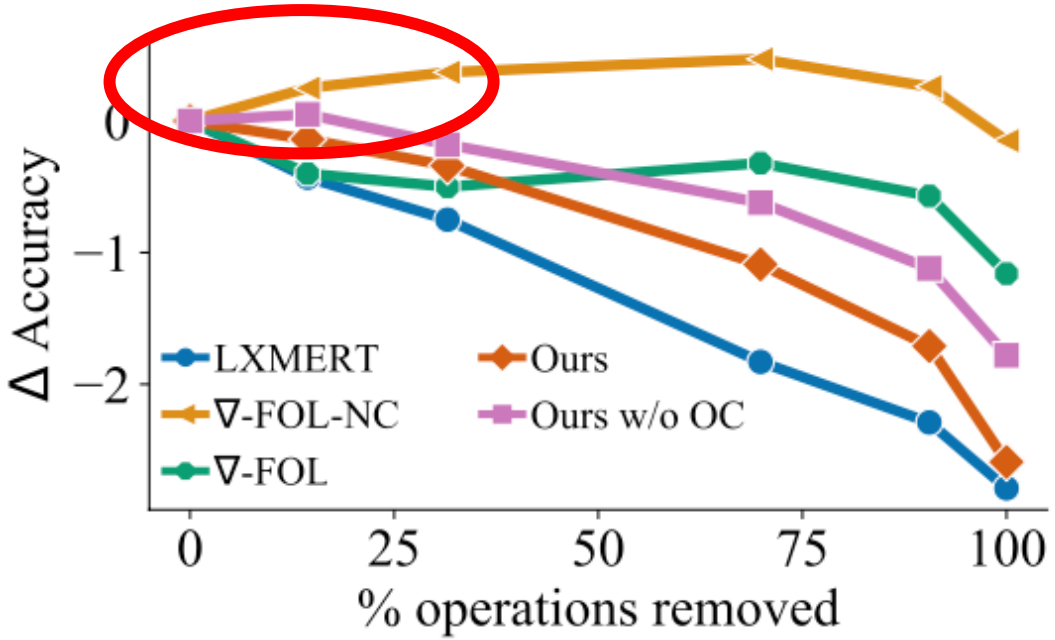


Figure 5: Accuracy drop of different models when the testing questions are progressively perturbed by removing reasoning operations with low weights.

Symbolic methods even increase performance when low-weighted operations are removed.

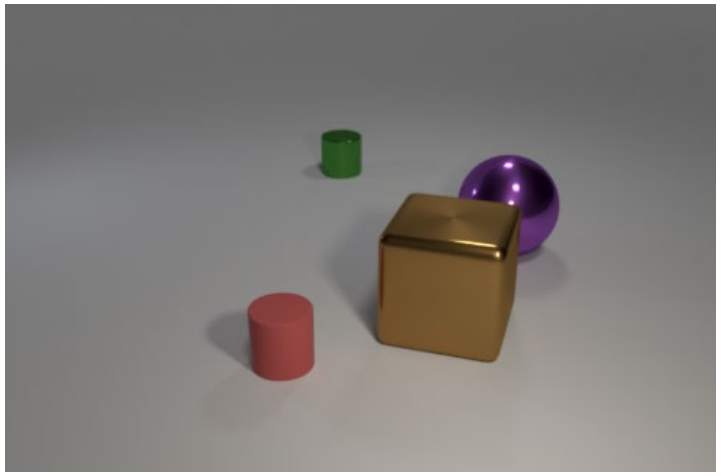
Summary of Neuromodular on Real Data

- The performance of Neuromodular methods is improved by the two methods described above. But this does not solve the problem. Why can’t Neuromodular or standard methods get 100% success?
- What else is going on?
- Conjectures:
- (1) The standard end-to-trained methods can exploit the biases of the datasets, but the Neuromodular approaches are less effective at this.
- (2) The Neuromodular methods use deep networks as their vision modules. They need better vision modules.
- (3) The training confounds the vision and the language by training them together.
- How to study this? Better controlled datasets. Out-of-distribution testing.

How to study the domain robustness more systematically?

Super-CLEVR: A Virtual Benchmark to Diagnose Domain Robustness in Visual Reasoning?
Zhuowan Li, Xingrui Wang, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin Van Durme, Alan Yuille.
CVPR 2023 Highlight.

Super-CLEVR dataset



CLEVR

Super-CLEVR

- Create the Super-CLRVR dataset using more realistic objects.
- This is more challenging
- It controls domain shifts to study robustness.

Super-CLEVR

- Super-CLEVR is more complex than CLVR. It contains classes of objects – vehicles – from the ShapeNet repository. These are rendered to generate semi-realistic images.
- Super-CLEVR is controllable. We can systematically vary factors like the numbers of objects in the images, their poses in 3D, the occlusion, and so on.
- Super-CLEVR can be used to test VQA algorithms on data on which it has been trained. But it can also be used to test how VQA tests algorithms to generalize to out-of-distribution domains.
- Super-CLEVR can be extended to test VQA with adversarial examiners (but this has not been done).

Generalization: Decompose and analyze

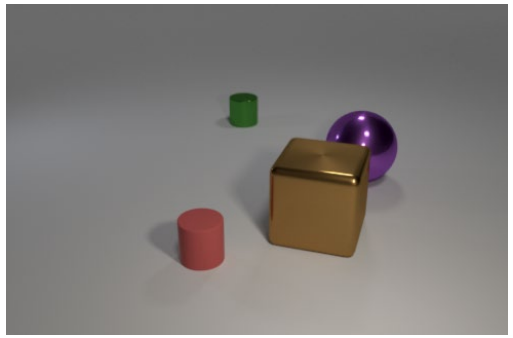
Four robustness factors in VQA domain shifts:

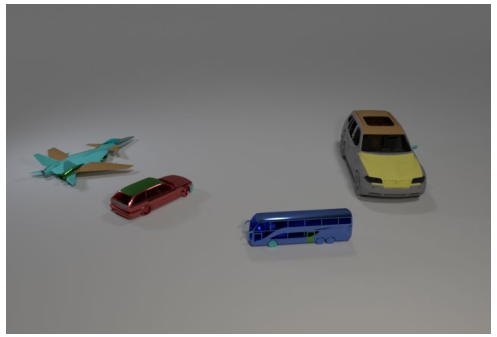
- visual complexity
- question redundancy
- concept distribution
- concept compositionality

Decompose VQA domain shifts into 4 factors

- **visual complexity**

  how hard is the image





Easy → Hard

- question redundancy
- concept distribution
- concept compositionality

Decompose VQA domain shifts into 4 factors

- visual complexity
- **question redundancy**

  the question may contain unnecessary information

What does the ~~little~~ boy ~~in front of the table~~ hold?

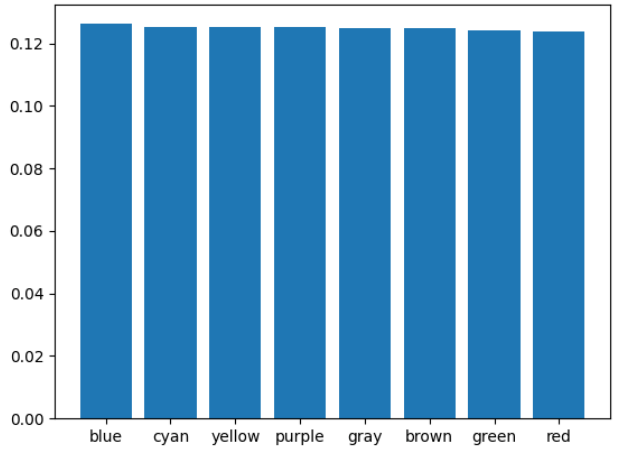
What is feeding the ~~large~~ animal ~~behind the fence~~?

- concept distribution
- concept compositionality

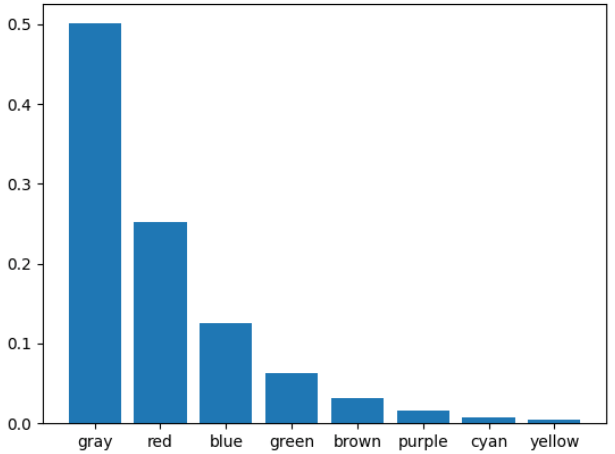
Decompose VQA domain shifts into 4 factors

- visual complexity
- question redundancy
- **concept distribution**

  The distribution the concepts (objects names and attributes)



Well-balanced



Long-tail distributed

- concept compositionality

Domain A

Domain B

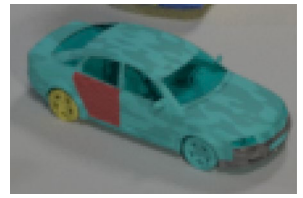
Super-CLEVR

“What color is the bus?”

Visual Complexity

easy

middle



hard

Question Redundancy

- redundancy

standard

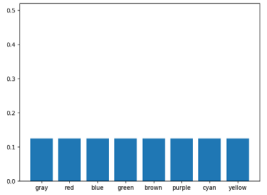
+ redundancy

“What color is the bus?”

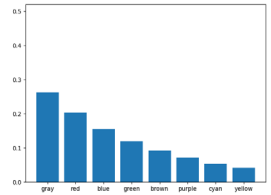
“What color is the large bus?”

“What color is the large bus behind the cyan car?”

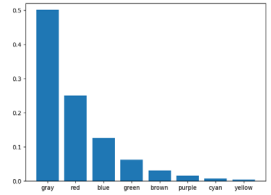
Concept Distribution



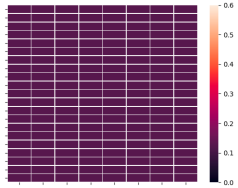
balanced



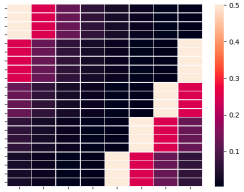
unbalanced



Concept Compositionality



well-composed



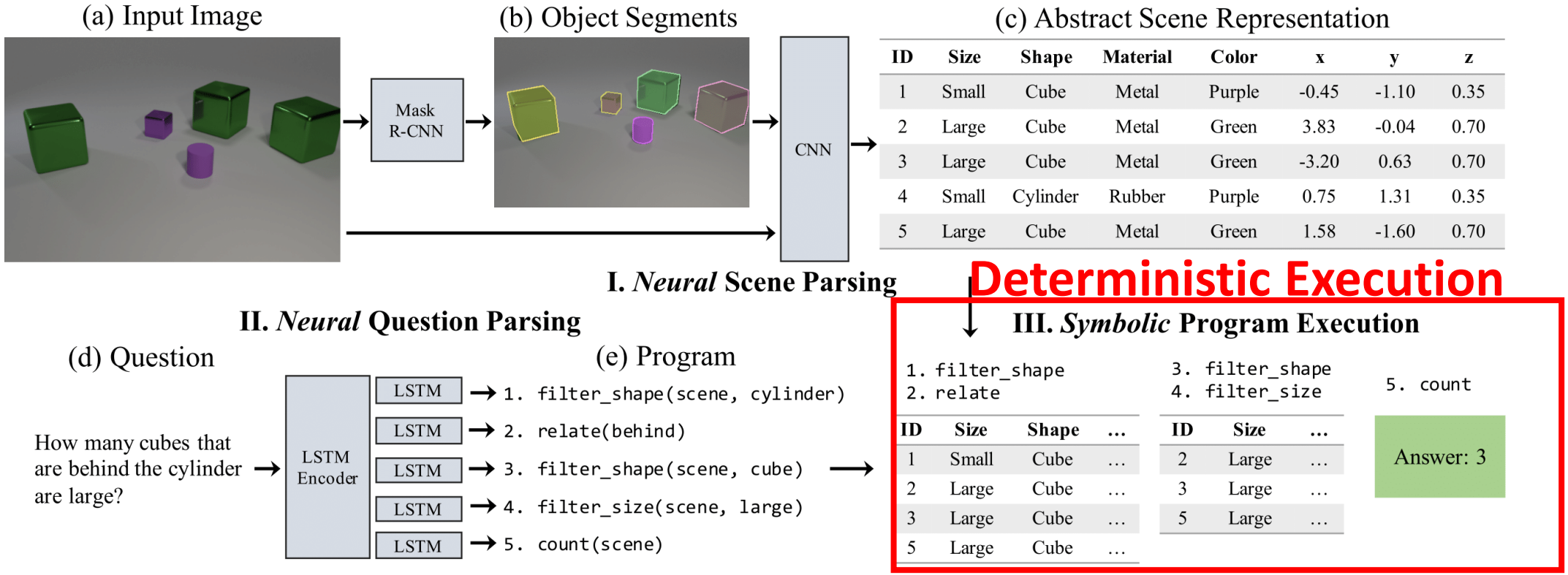
correlated

Five models are studied

non-modular
- FiLM
  two-stream feature merging method
- mDETR
  pretrained transformer model

modular
- NSCL
  neural modular method
- NSVQA
  neural modular method
- **Probabilistic NSVQA**
  our method

Recall Neural-Symbolic VQA (NSVQA)



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

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

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

Important – not emphasized in their paper – the training is modular. The language and the vision components are trained separately. No joint training.

Yi, Kexin, et al. "Neural-symbolic vqa: Disentangling reasoning from vision and language understanding." *NeurIPS* 2018.

Prob-NSVQA considers the confidence of the scene parser predictions.

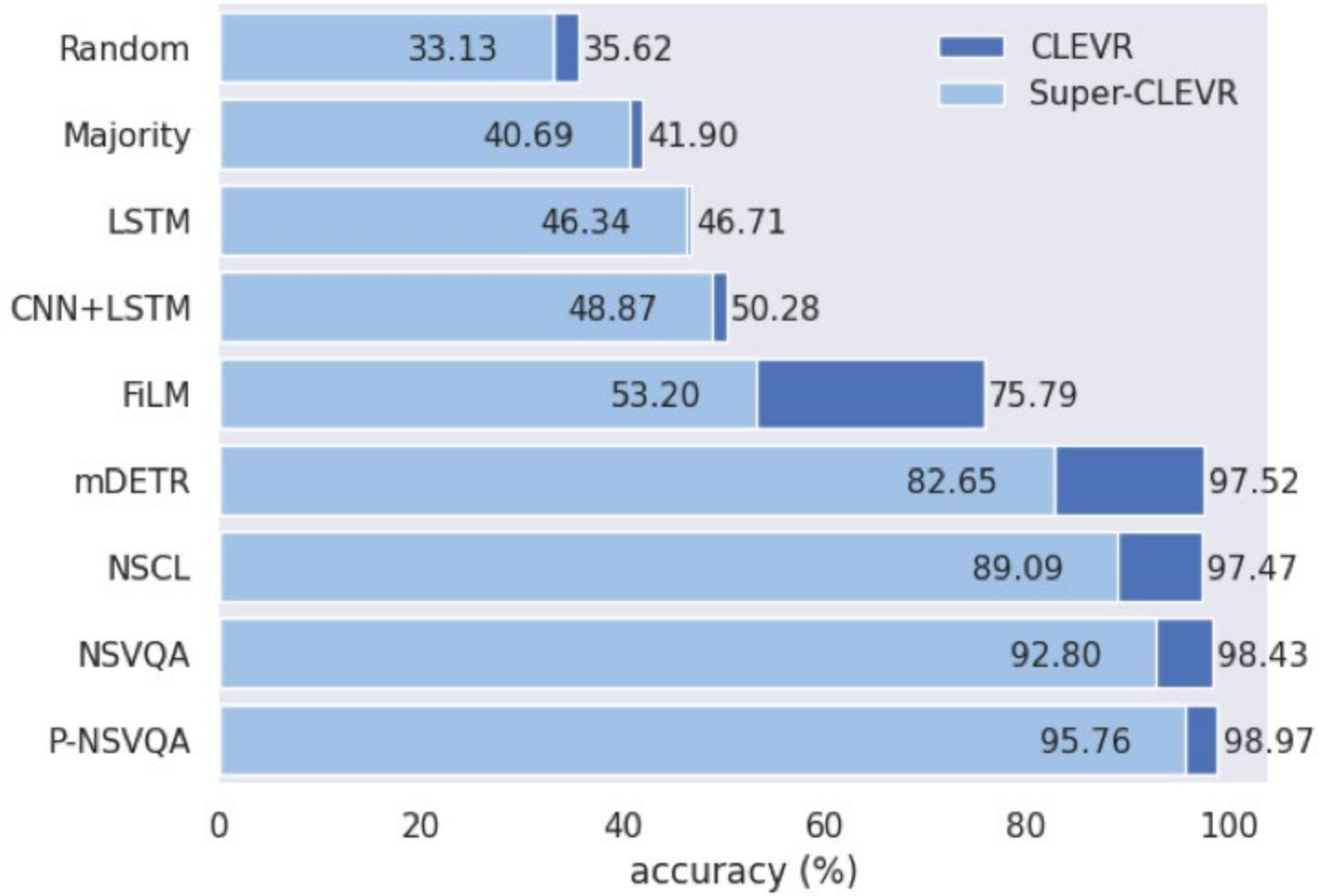
- We introduce probabilities to Prob-NSVQA. E.g., the vision modules output the probabilities that they have detected and classified an object.
- Given an image containing n objects, we maintain a vector of probs:

$$\mathbf{p} = [p^1, p^2, \ldots, p^n]$$

- $filter_{identifier}[attribute]$ (*e.g.* $filter_{color}[red]$)

e.g. for the filter operation we set $p^k = p^k * P^k_{attribute}$

Super-CLEVR is harder than CLEVR (I.I.D. testing)



Out-of-domain testing: Complete Results

	FiLM			mDETR			NSCL			NSVQA			Prob NSVQA		
Visual Complexity															
	easy	mid	hard	easy	mid	hard	easy	mid	hard	easy	mid	hard	easy	mid	hard
easy	**59.96**	53.95	50.66	**93.36**	84.30	82.97	**95.13**	92.31	90.81	**95.19**	94.19	94.09	**96.76**	95.98	96.37
mid	**57.41**	53.28	50.18	**83.34**	82.36	81.27	84.5	**89.10**	86.33	81.99	92.80	**93.78**	86.25	**95.76**	95.11
hard	**55.95**	53.11	50.47	79.71	79.94	**80.71**	76.85	78.66	**85.08**	73.11	79.71	**92.65**	79.81	86.47	**95.36**
Question Redundancy															
	rd-	rd	rd+	rd-	rd	rd+	rd-	rd	rd+	rd-	rd	rd+	rd-	rd	rd+
rd-	51.42	52.54	**53.51**	**83.94**	80.37	66.28	88.64	88.82	**90.33**	-	92.98	-	-	95.71	-
rd	50.39	53.28	**54.78**	**82.77**	82.36	70.36	88.45	89.10	**91.45**	-	92.80	-	-	95.76	-
rd+	46.14	52.30	**71.47**	78.48	84.05	**90.42**	87.94	88.34	**91.16**	-	92.69	-	-	95.73	-
Concept Distribution															
	bal	slt	long	bal	slt	long	bal	slt	long	bal	slt	long	bal	slt	long
bal	50.47	53.04	**54.35**	**80.71**	75.79	74.54	**85.08**	83.79	75.10	**92.65**	90.82	83.74	**95.36**	94.89	89.88
long	49.43	54.75	**62.96**	79.06	80.29	**90.66**	85.33	89.42	**91.10**	92.73	**93.38**	92.53	96.31	**96.32**	95.25
head	48.60	58.06	**61.60**	80.75	79.60	**87.46**	84.58	88.39	**90.19**	93.87	**94.82**	92.48	96.42	**96.80**	95.92
tail	**51.80**	48.70	50.08	**81.50**	70.88	60.94	**86.10**	80.27	60.55	**90.26**	89.20	75.32	**94.08**	93.20	82.68
oppo	**49.06**	48.93	46.68	**79.13**	68.37	56.98	**85.07**	77.86	55.14	**91.22**	88.65	71.32	**95.76**	94.09	79.74
Concept Compositionality															
	co-0	co-1	co-2	co-0	co-1	co-2	co-0	co-1	co-2	co-0	co-1	co-2	co-0	co-1	co-2
co-0	53.28	**57.00**	56.1	**83.36**	77.03	82.43	**89.1**	82.52	83.77	**92.80**	90.11	91.59	**95.76**	94.02	95.12
co-1	52.41	**60.57**	56.67	79.46	82.45	**83.93**	78.89	**87.18**	84.2	78.74	89.99	**90.67**	87.12	94.53	**94.78**
co-2	52.96	57.37	**60.53**	80.03	77.41	**87.24**	78.40	81.55	**88.84**	77.85	89.28	**92.23**	87.19	93.49	**95.61**

Out-of-Domain Testing: Summary Results

non-modular

modular

	Visual	Redund.	Dist.	Comp.
FiLM	**4.03**	21.33	28.46	9.04
mDETR	9.81	19.05	36.34	9.45
NSCL	15.57	0.92	37.44	15.40
NSVQA	17.48	**0.08**	20.92	11.44
NSVQA + Prob	12.88	**0.08**	**13.72**	**7.00**

Table 2. *Relative Degrade* under domain shifts, *i.e.* the percentage of accuracy decrease when the model is tested with domain that differs with training. Lower *RD* means better robustness.

Findings: comparison between models

	Visual	Redund.	Dist.	Comp.
FiLM	**4.03**	21.33	28.46	9.04
mDETR	9.81	19.05	36.34	9.45
NSCL	15.57	0.92	37.44	15.40
NSVQA	17.48	**0.08**	20.92	11.44
NSVQA + Prob	12.88	**0.08**	**13.72**	**7.00**

- Neural symbolic methods are robust on question redundancy

  Question parsing is easy, and trained separately

Findings: comparison between models

	Visual	Redund.	Dist.	Comp.
FiLM	**4.03**	21.33	28.46	9.04
mDETR	9.81	19.05	36.34	9.45
NSCL	15.57	0.92	37.44	15.40
NSVQA	17.48	**0.08**	20.92	11.44
NSVQA + Prob	12.88	**0.08**	**13.72**	**7.00**

- Neural symbolic methods are robust on question redundancy
- Neural symbolic methods are **only** robust on question redundancy

  Why?]

  Maybe we need not only modular network, but also modular training

Findings: comparison between models

	Visual	Redund.	Dist.	Comp.
FiLM	**4.03**	21.33	28.46	9.04
mDETR	9.81	19.05	36.34	9.45
NSCL	15.57	0.92	37.44	15.40
NSVQA	17.48	**0.08**	20.92	11.44
NSVQA + Prob	12.88	**0.08**	**13.72**	**7.00**

- P-NSVQA is the most robust on 3 out of 4 factors
  - Probablistic + symbolic -> best model

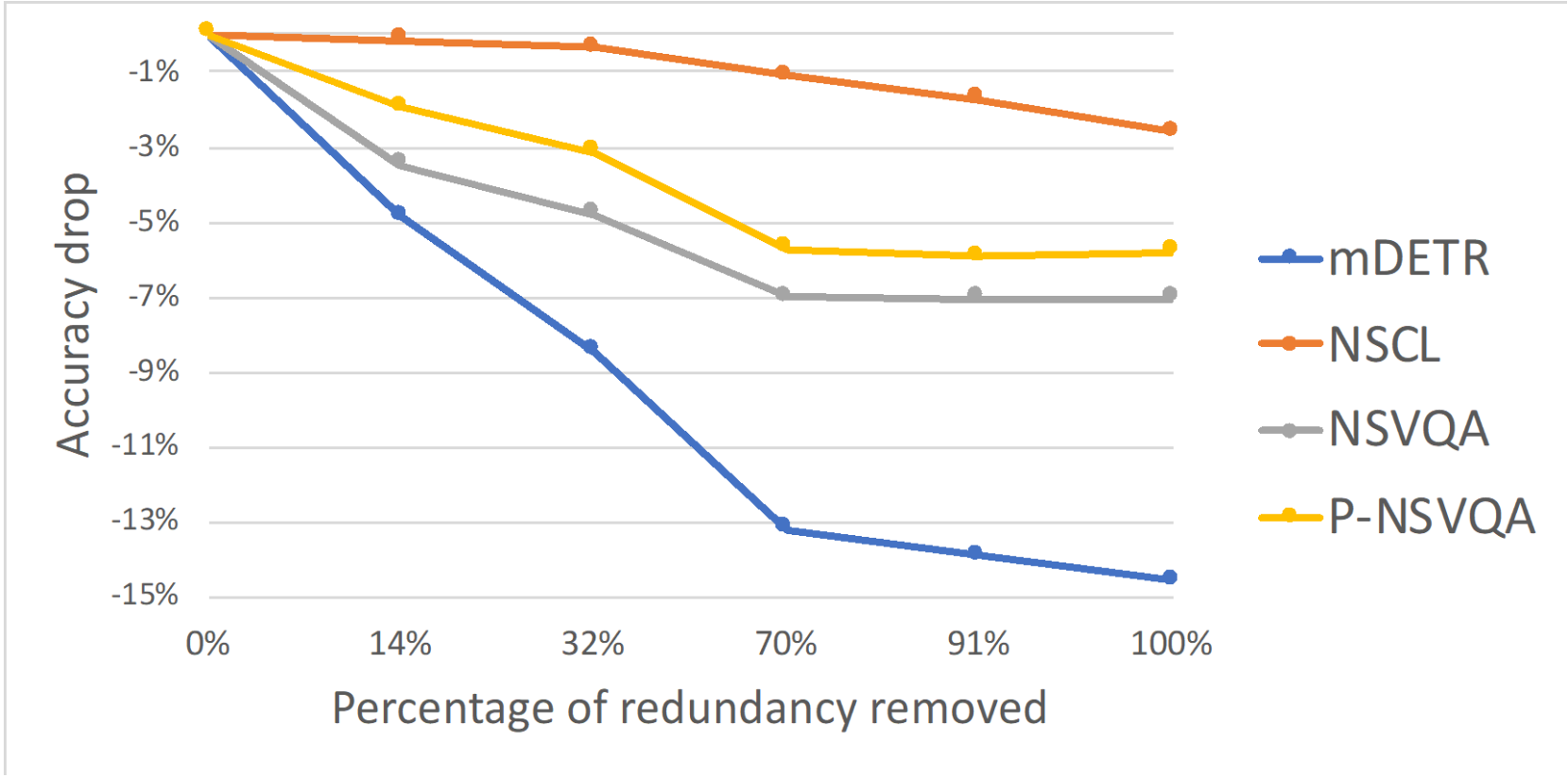
Findings: comparison between models

	Visual	Redund.	Dist.	Comp.
FiLM	**4.03**	21.33	28.46	9.04
mDETR	9.81	19.05	36.34	9.45
NSCL	15.57	0.92	37.44	15.40
NSVQA	17.48	**0.08**	20.92	11.44
NSVQA + Prob	12.88	**0.08**	**13.72**	**7.00**

- On visual complexity, end-to-end methods are more robust
  - mDETR has a more powerful visual component

Will the findings generalize to real data?

For question redundancy:



Progressively remove the redundant operations from questions in GQA dataset

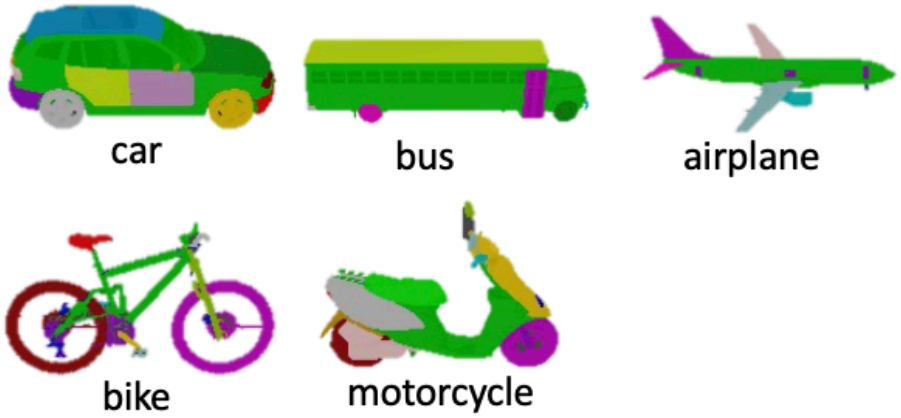
Extension: Super-CLEVR with Parts, 3D Pose, Occlusion

Ongoing work

Part questions

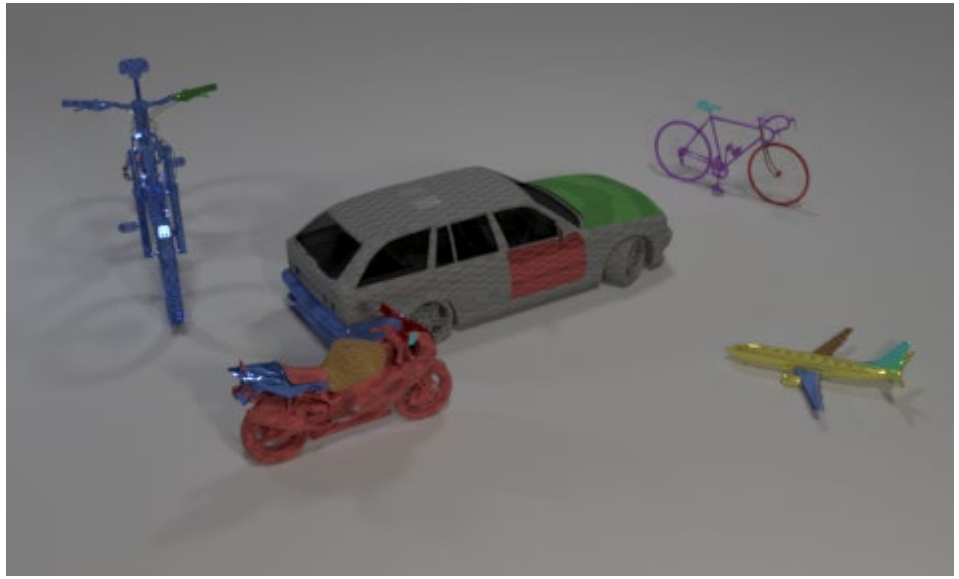
Objects from UDA-Part dataset

Object with parts:



Texture: dotted, checkered, stripped, none
Color: green, gray, brown, yellow, red, purple, cyan, blue
Size: large, small
Material: rubber, metal

Figure 2. Super-CLEVR contains 21 vehicle models belonging to 5 categories, with controllable attributes.



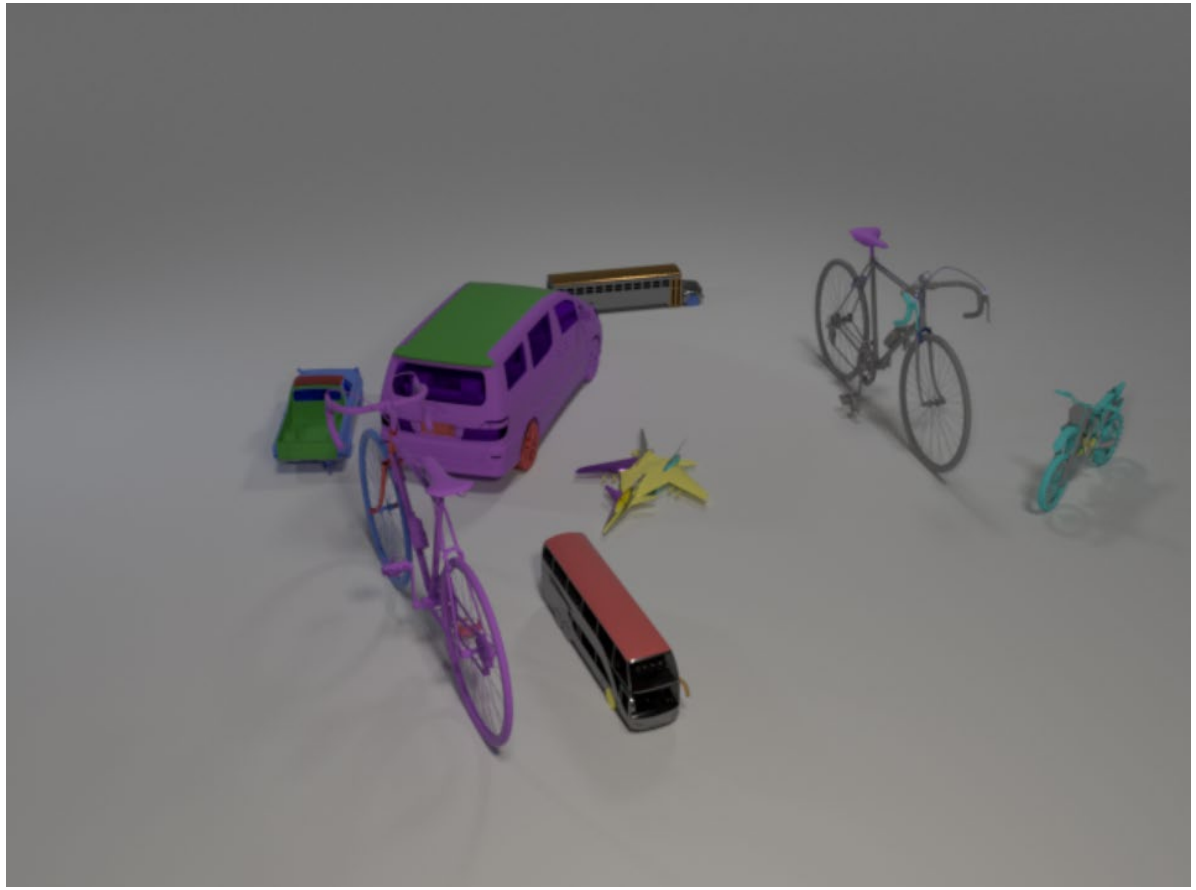
Q: What is the color of the front wheel of the small purple bicycle?
Q: What is the material of the yellow object that has a blue part?
Q: What is the color of the front wheel that belongs to the same object as the cyan seat?

UDA-Part dataset: Learning Part Segmentation through Unsupervised Domain Adaptation from Synthetic Vehicles. Qing Liu et al. CVPR 2022

Pose questions

Q: Which direction is the tiny blue school bus facing?
Q: What is the brown thing facing in the same direction as the tiny blue school bus?
Q: Is the plane and the tiny blue school bus facing in the opposite direction?

Occlusion questions

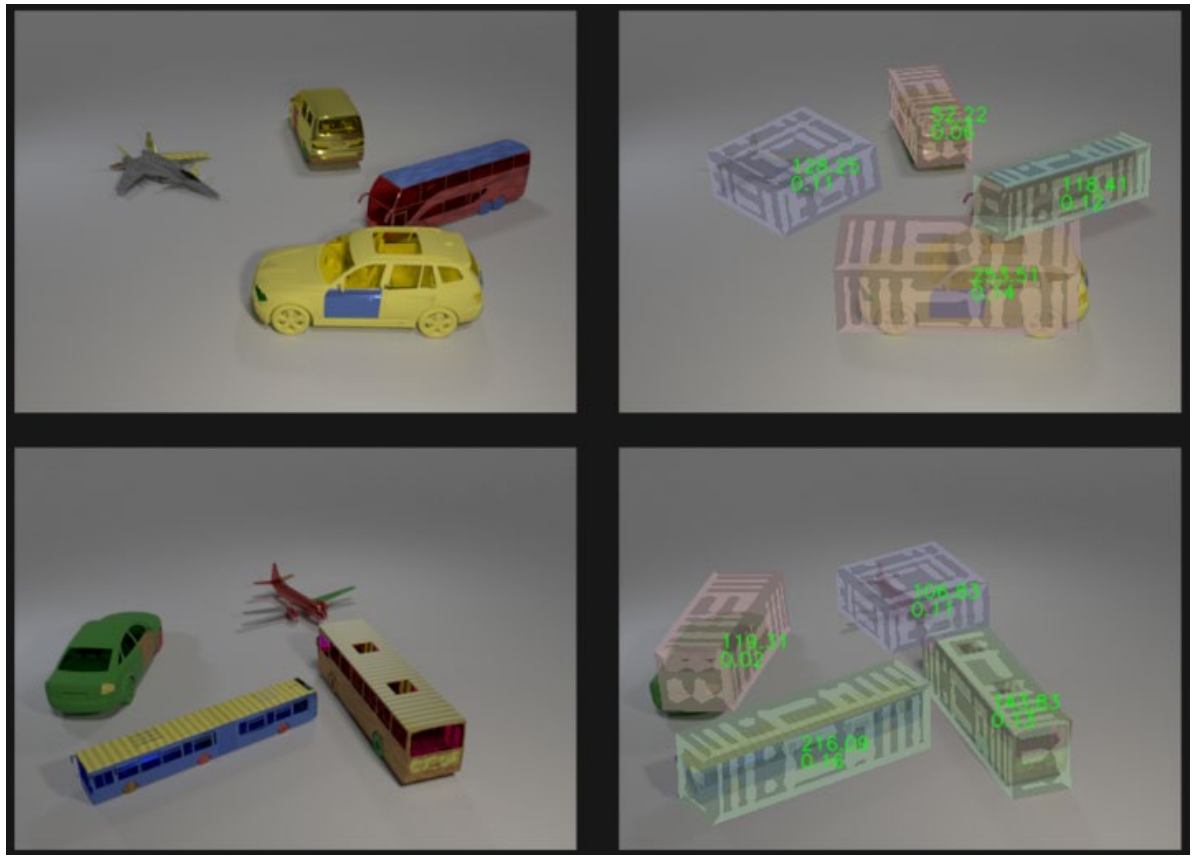


Q: What is the size of the purple object that is occluded?

Q: What part of the small rubber object is occluded?

Why part, pose, occlusion?

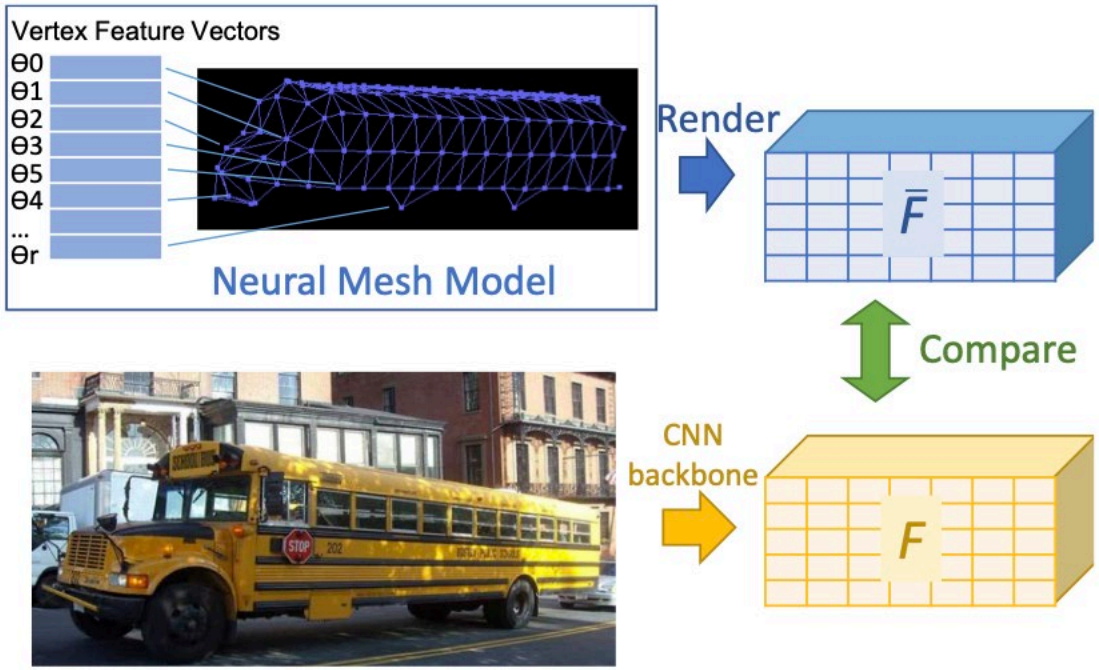
3D knowledge is necessary to answer those questions.



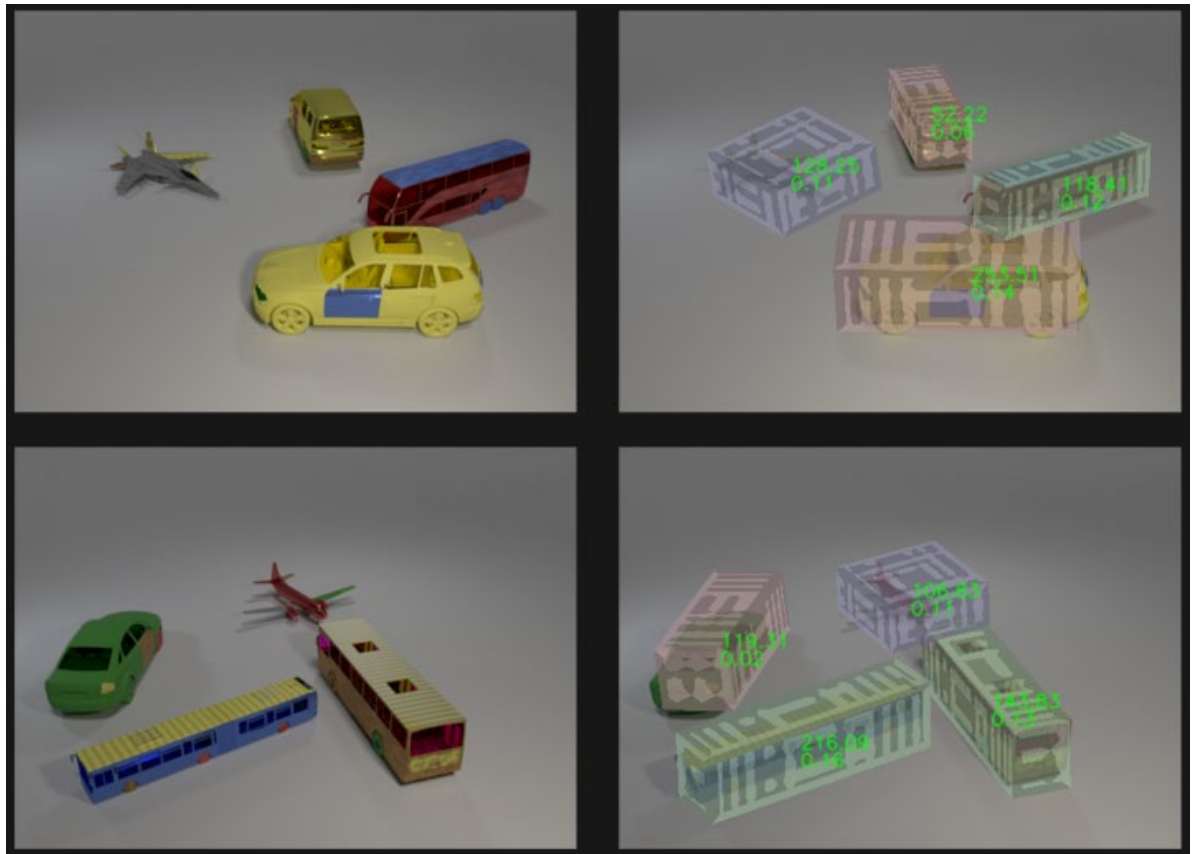
Introduce 3D Pose estimation in VQA

NeMo (Wang el at. ICLR 2021)

- A generative pose estimation model
- Robust to occlusion



NeMo Render-and-Compare: Feature Map



Ongoing results

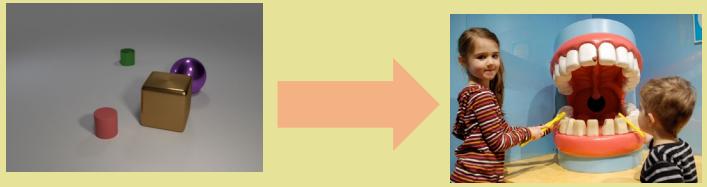
	Part	Pose	Occlusion	Part + Occlusion
mDETR	41.52	71.76	64.99	44.60
P-NSVQA with FRCNN	-	87.78	-	-
P-NSVQA with NeMO	-	86.40	-	-
P-NSVQA with NeMO & GT		91.34		

Table being completed.

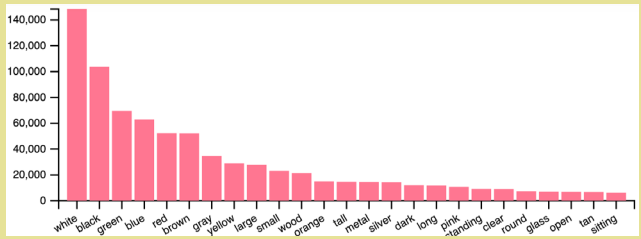
Towards Generalizable Visual Reasoning

Calibrating concepts and operations

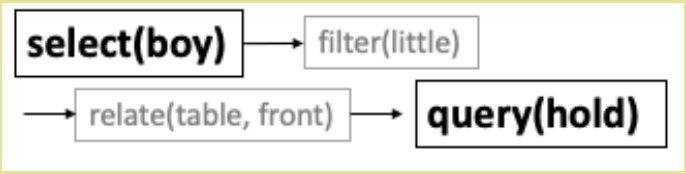
Synthetic-real gaps, and solutions



Long tail distribution



Unequal importance of operations



Super-CLEVR

systematically study domain robustness

Domain shift factors

- Visual complexity
- Question redundancy
- Concept distribution
- Concept compositionality
- …

3D-aware VQA

- Object parts
- 3D poses
- Occlusions
- …

Probabilistic + Symbolic → robust model

Future direction?

SS-CLEVR – more Realistic images.

Adversarial Testing

Stronger Vision Modules.

Thanks! Questions?

zli110@jhu.edu