1 JOHNS HOPKINS UNIVERSITY
2 UNIVERSITY OF SOUTHERN CALIFORNIA 1880
3 mpi max planck institut informatik
4 universität freiburg

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Super-CLEVR: A Virtual Benchmark to Diagnose Domain Robustness in Visual Reasoning

Highlight | WED-PM-249 | Jun 21, 2023

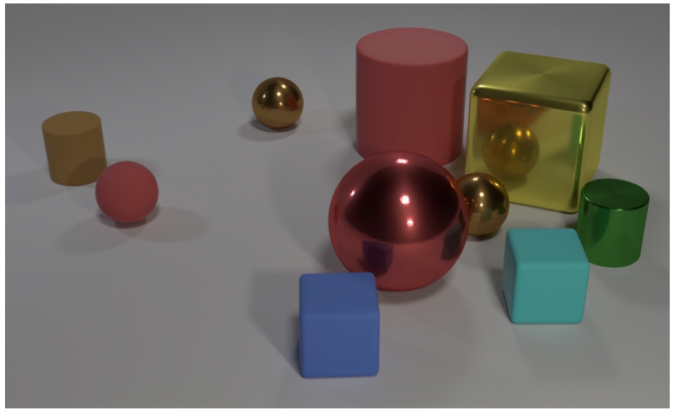





Zhuowan Li[1] Xingrui Wang[2] Elias Stengel-Eskin[1] Adam Kortylewski[3,4] Wufei Ma[1] Benjamin Van Durme[1] Alan Yuille[1]

Visual Reasoning

CLEVR



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

VQA2.0

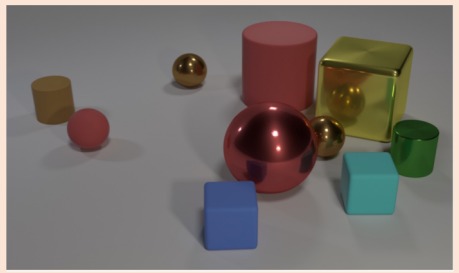
What color are her eyes?

GQA

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

Visual reasoning models suffer on out-of-domain testing

Domain-1



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

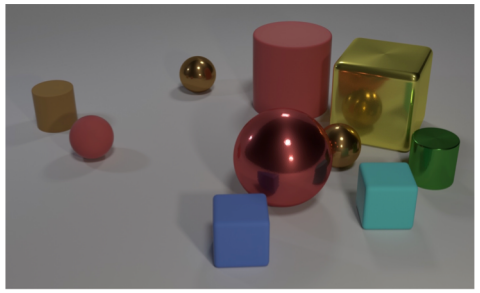
Domain-2

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

Training → **Testing**

poor performance

Domain gaps contain multiple factors



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

What color are her eyes?

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

How are they different from each other?

Multiple factors:

- Image styles?
- Question lengths?
- Concepts?
- …

Super-CLEVR:

study each domain shift factor separately

Domain A

Domain B

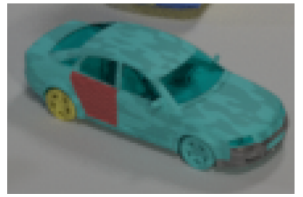
Super-CLEVR

"What color is the bus?"

Visual Complexity

easy

middle



hard

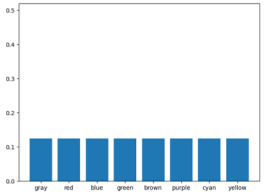
Question Redundancy

- redundancy: "What color is the bus?"

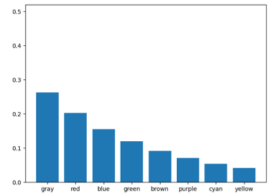
standard: "What color is the large bus?"

+ redundancy: "What color is the large bus behind the cyan car?"

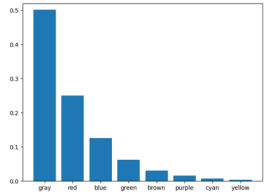
Concept Distribution



balanced

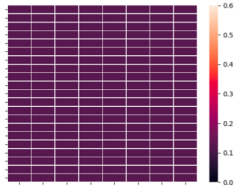


unbalanced

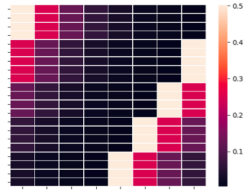


Long-tail

Concept Compositionality



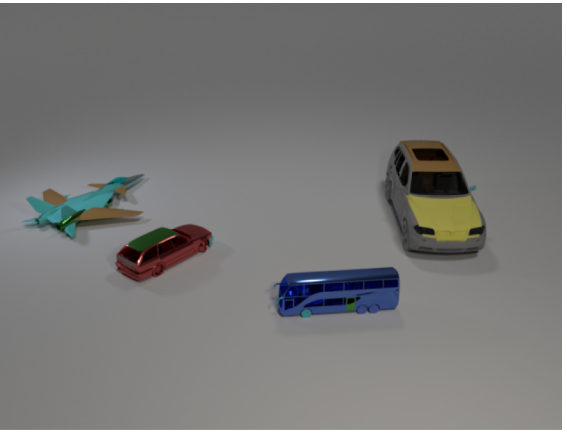
well-composed



correlated

Super-CLEVR: a controllable dataset

The dirtbike that is the same size as the brown motorbike is what color?



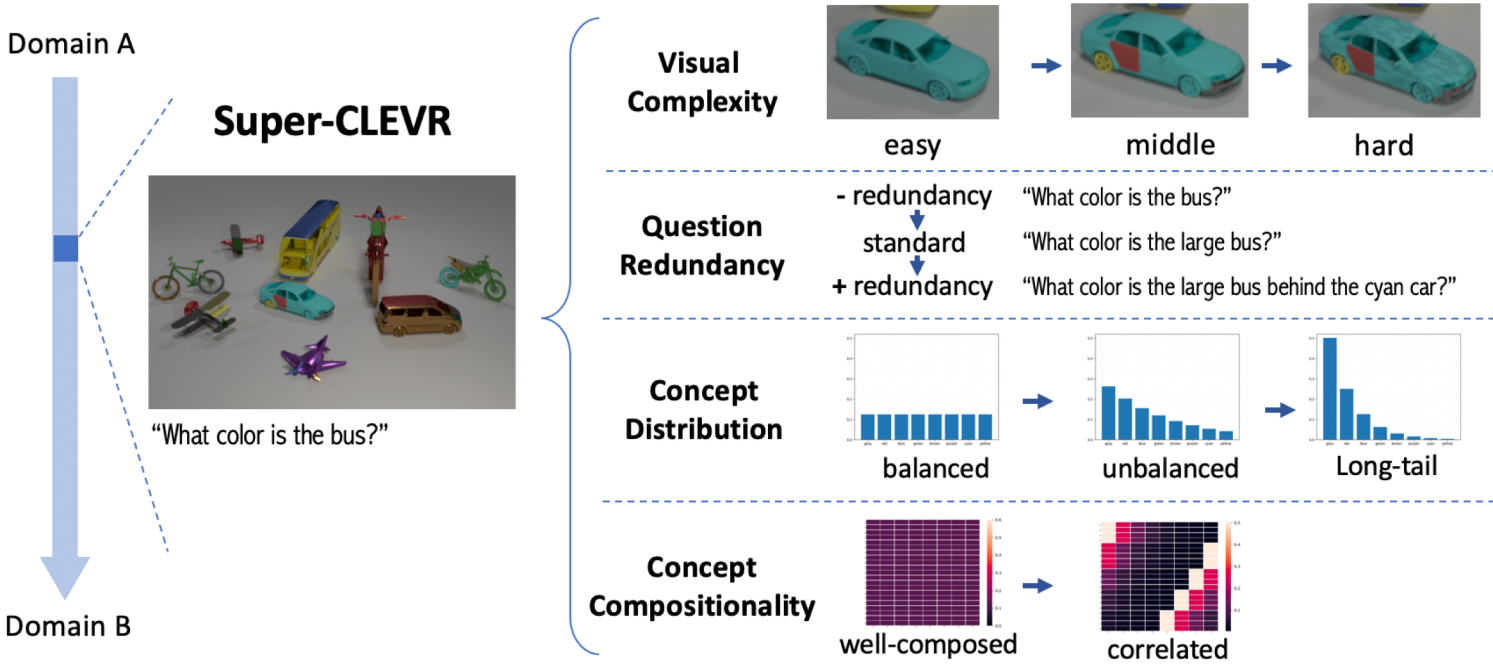
What is the color of the small car?

- Separately study the 4 domain shift factors
- 5 models are studied
- Analysis finding:

  modular training + probabilistic execution ➡ best model

In more detail…

- 4 domain shift factors
- 5 models we studied
- Analysis Results



Study domain shift by decomposition

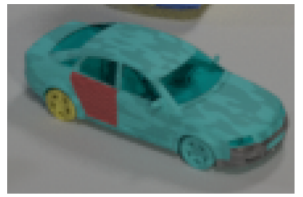
Factors contributes to domain shifts in visual reasoning:

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

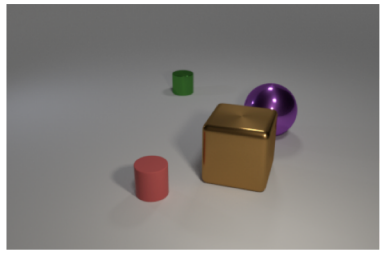
Decompose VQA domain shifts into 4 factors

- **visual complexity**
  how hard is the image





easy middle hard



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

- concept distribution
- concept compositionality

- redundancy — "What color is the bus?"

↓

standard — "What color is the large bus?"

↓

+ redundancy — "What color is the large bus behind the cyan car?"

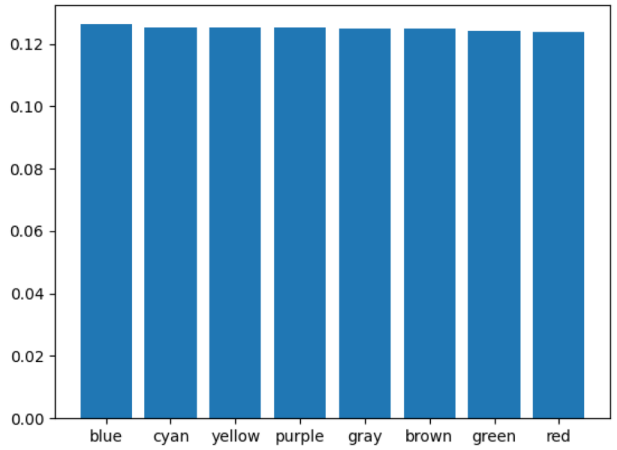
What does the ~~little~~ boy ~~in front of the table~~ hold?

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

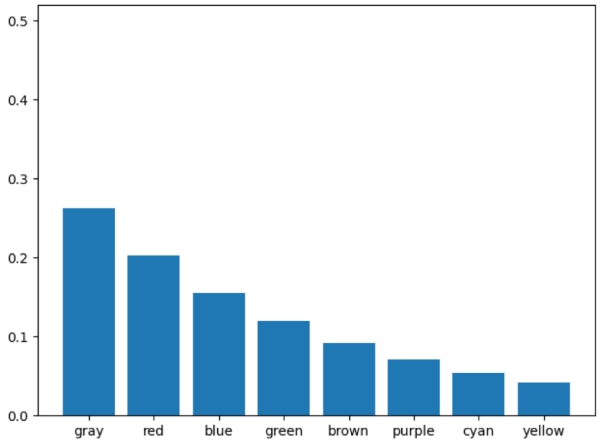
Decompose VQA domain shifts into 4 factors

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

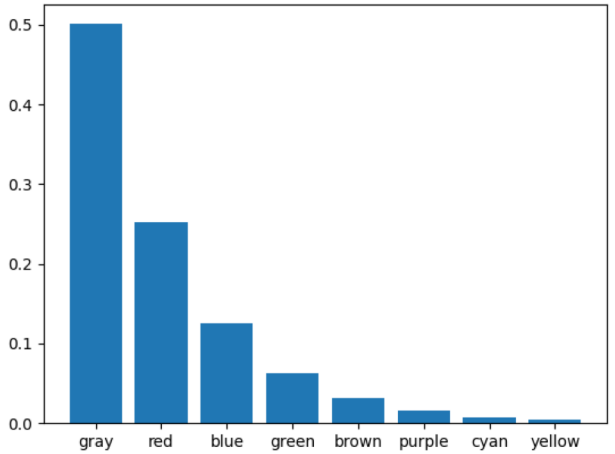
  The distribution the concepts (objects names and attributes)



Well-balanced



Slightly unbalanced



Long-tail distributed

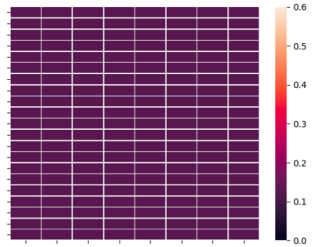
- concept compositionality

Decompose VQA domain shifts into 4 factors

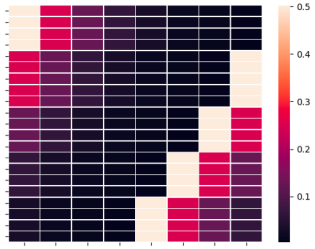
- visual complexity
- question redundancy
- concept distribution
- **concept compositionality**

  Some concepts always co-occur with another:

  - *Bananas* are usually *yellow bananas*
  - *Boys* are usually *little boys*
  - *Skies* are usually *blue skies*
  - …



well-composed



correlated

Domain A

Domain B

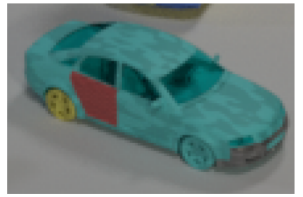
Super-CLEVR

"What color is the bus?"

Visual Complexity

easy

middle



hard

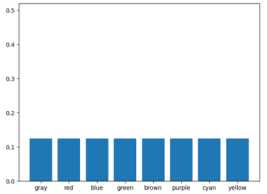
Question Redundancy

- redundancy: "What color is the bus?"

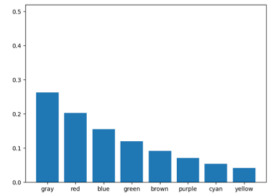
standard: "What color is the large bus?"

+ redundancy: "What color is the large bus behind the cyan car?"

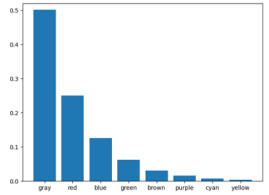
Concept Distribution



balanced

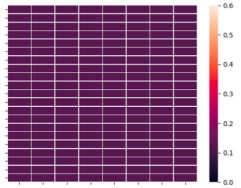


unbalanced

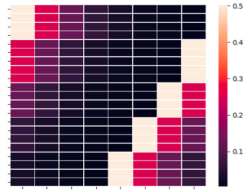


Long-tail

Concept Compositionality



well-composed

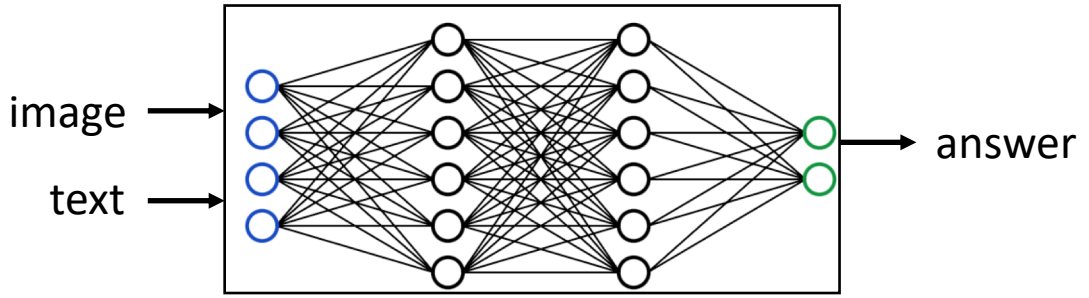


correlated

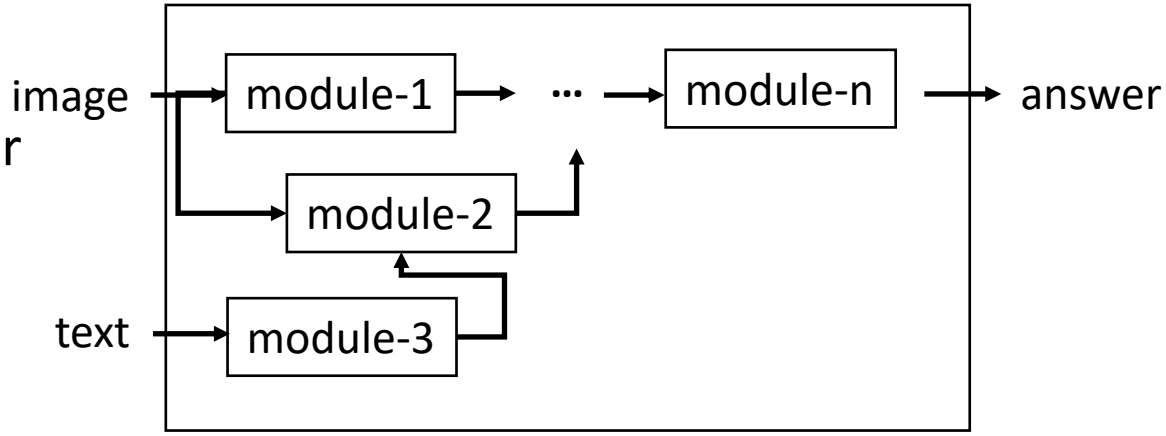
5 models are studied

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

modular
- NSCL
  neural symbolic concept learner
- NSVQA
  neural symbolic VQA
- **Probabilistic NSVQA**
  our method

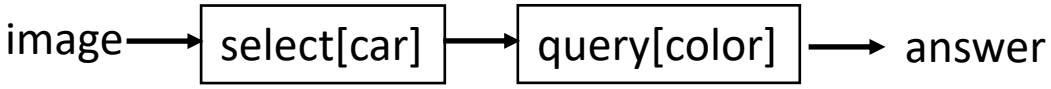


End-to-end



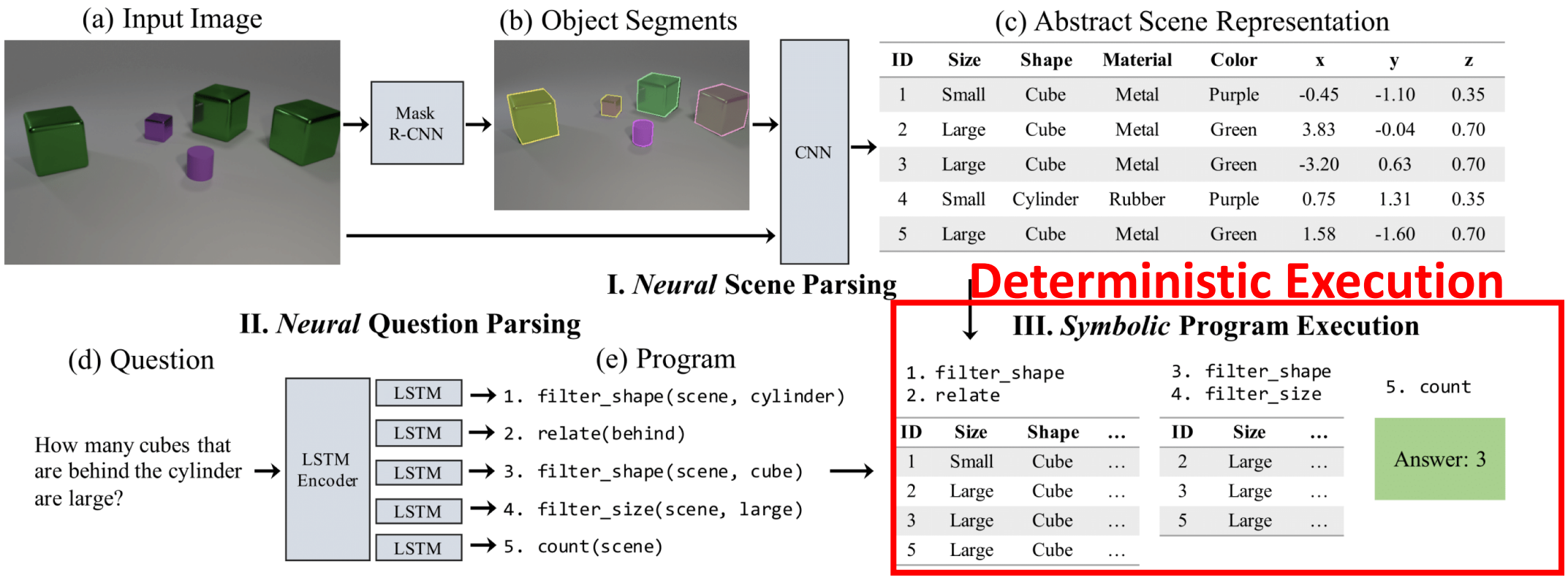
Neural modular method

What color is the car?



Probabilistic NSVQA (our model)

Recall 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	…

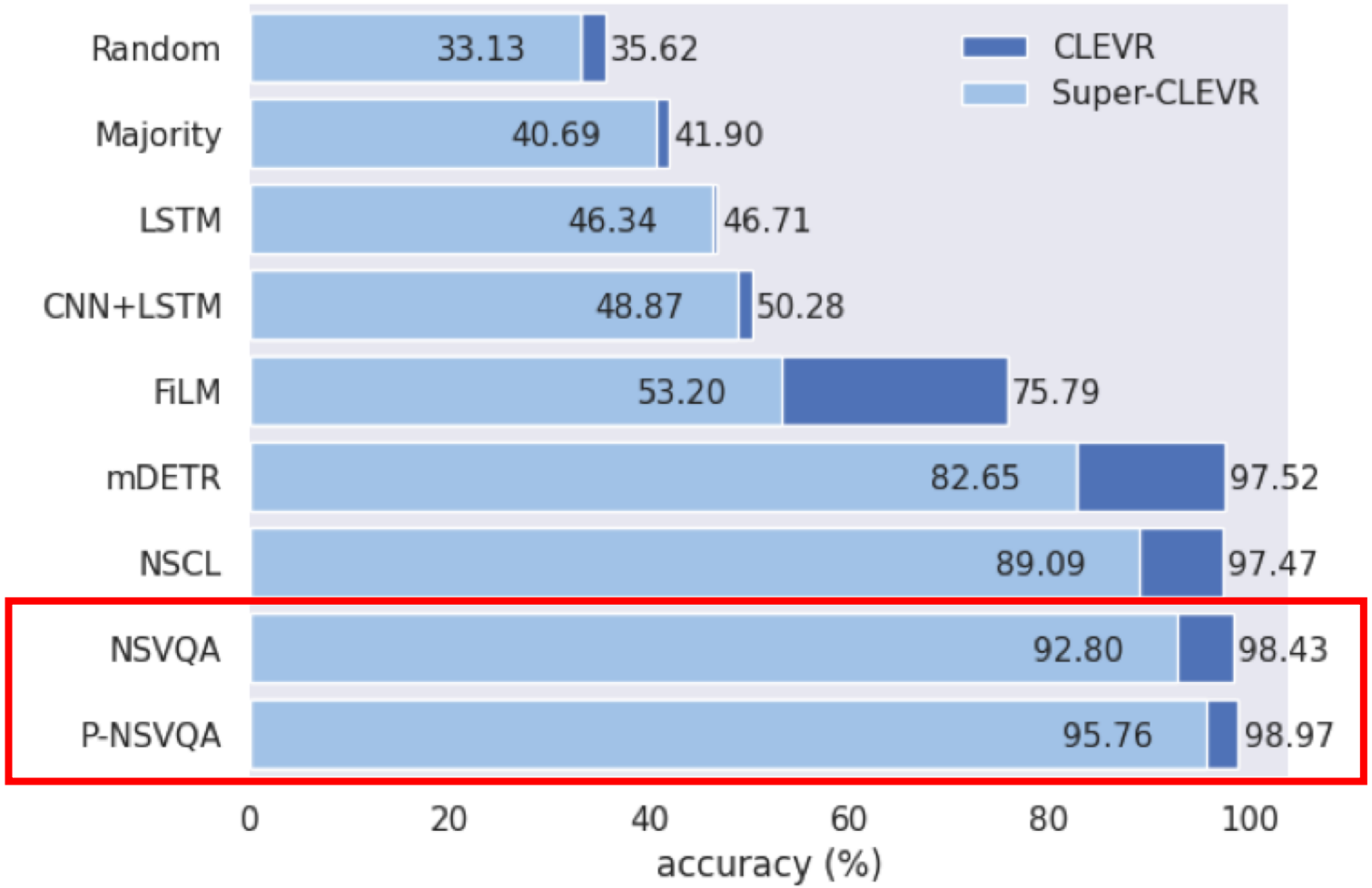
ID	Size	…
2	Large	…
3	Large	…
5	Large	…

Make it probabilistic!

Considering the uncertainty of scene parsing

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

In-domain Results



- Super-CLEVR is harder than CLEVR
- P-NSVQA is the best

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**	**92.95**	92.94	92.67	95.66	**95.72**	95.43
rd	50.39	53.28	**54.78**	**82.77**	82.36	70.36	88.45	89.10	**91.45**	91.19	**92.78**	92.14	94.87	**95.72**	95.43
rd+	46.14	52.30	**71.47**	78.48	84.05	**90.42**	87.94	88.34	**91.16**	91.38	91.96	**92.80**	94.88	95.47	**95.72**
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**

Summary OOD testing results

Relative Degrade

- the percentage of accuracy decrease when the model is tested with domain that differs with training

$$RD = \frac{Acc_{iid} - Acc_{ood}}{Acc_{iid}}$$

	Visual	Redund.	Dist.	Comp.
non-modular	**FiLM**	**4.03**	21.33	28.46	9.04
non-modular	**mDETR**	9.81	19.05	36.34	9.45
modular	**NSCL**	15.57	0.92	37.44	15.40
modular	**NSVQA**	17.48	1.72	20.92	11.44
modular	**Prob NSVQA**	12.88	**0.84**	**13.72**	**7.00**

Table. *Relative Degrade* of models. Smaller is better.

Finding-1: modular models are very robust on redundancy

	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	1.72	20.92	11.44
Prob NSVQA	12.88	**0.84**	**13.72**	**7.00**

- Modular training is important
- Question component shouldn’t compensate for visual understanding

Finding-2: P-NSVQA is the best on 3 out of 4 factors

	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	1.72	20.92	11.44
Prob NSVQA	12.88	**0.84**	**13.72**	**7.00**

- Modularity + Probabilistic execution -> best model

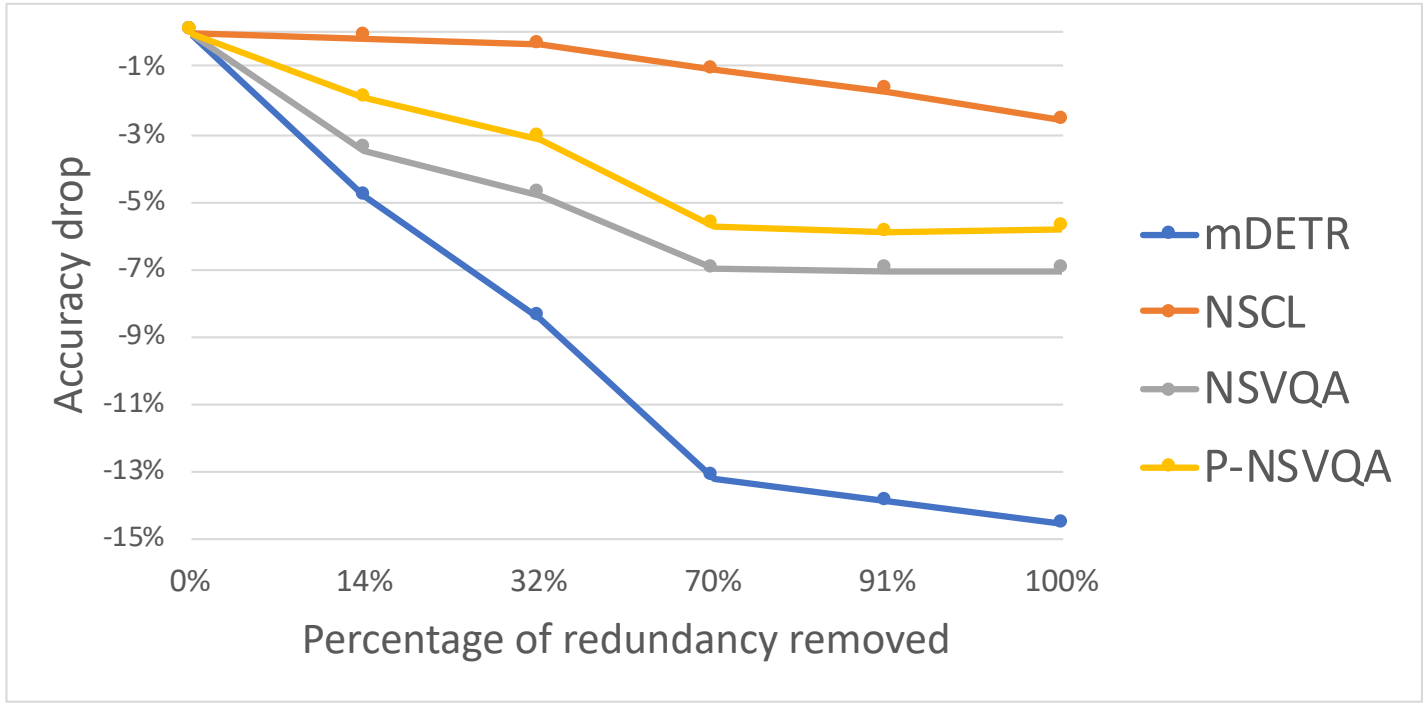
Finding-3: non-modular methods win on visual complexity

	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	1.72	20.92	11.44
Prob NSVQA	12.88	**0.84**	**13.72**	**7.00**

- mDETR has a more powerful pretrained visual component
- Visual scene parser (MaskRCNN) in modular methods can be improved.

Will the findings generalize to real data?

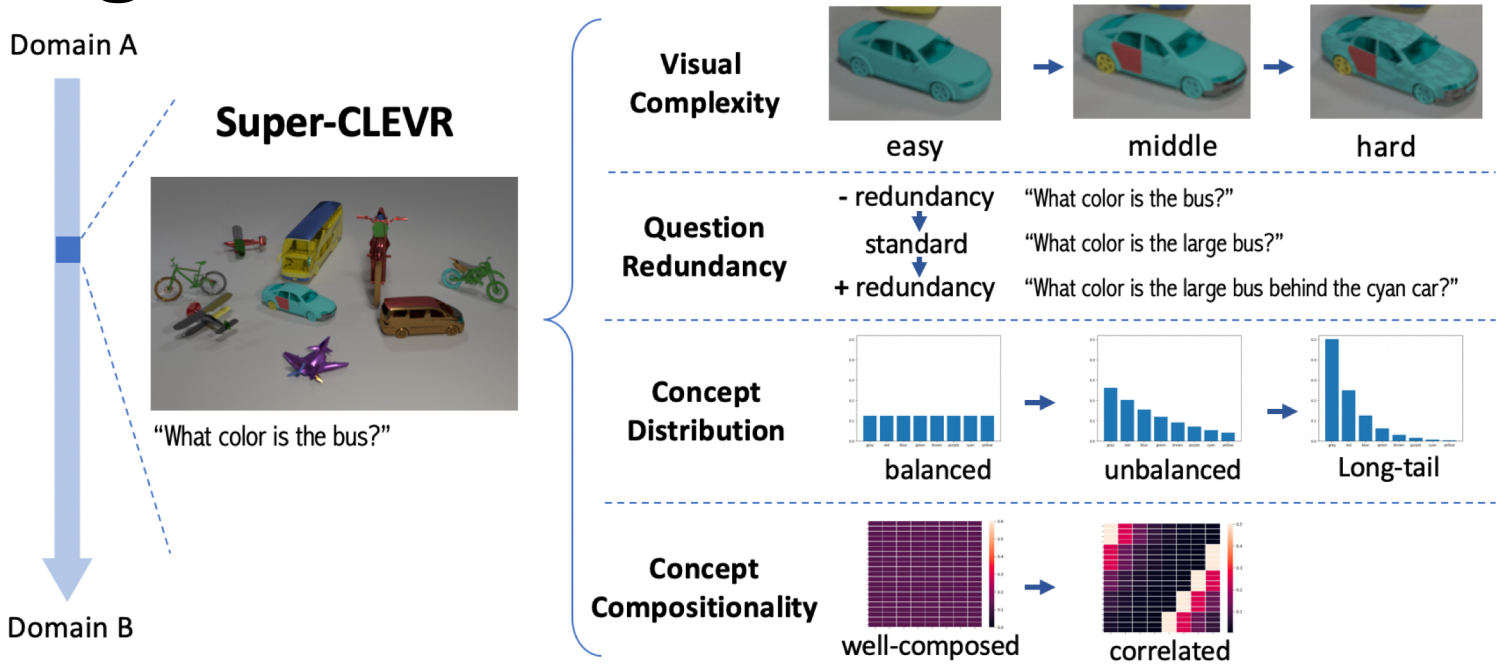
We verify findings on question redundancy on the real GQA dataset



When the redundant operations are progressively removed, the performance drops of modular methods are smaller than non-modular methods.

Take-home messages

- Super-CLEVR dataset



- Analysis findings:
  1. Modular models are very robust on question redundancy.
  2. P-NSVQA is the best on 3 out of 4 factors.
  3. Non-modular methods win on visual complexity.

➡ Modularity and probabilistic execution are important; we need better visual modules.

Welcome to our session!

WED-PM-249 (Highlight)

Jun 21, 2023