JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

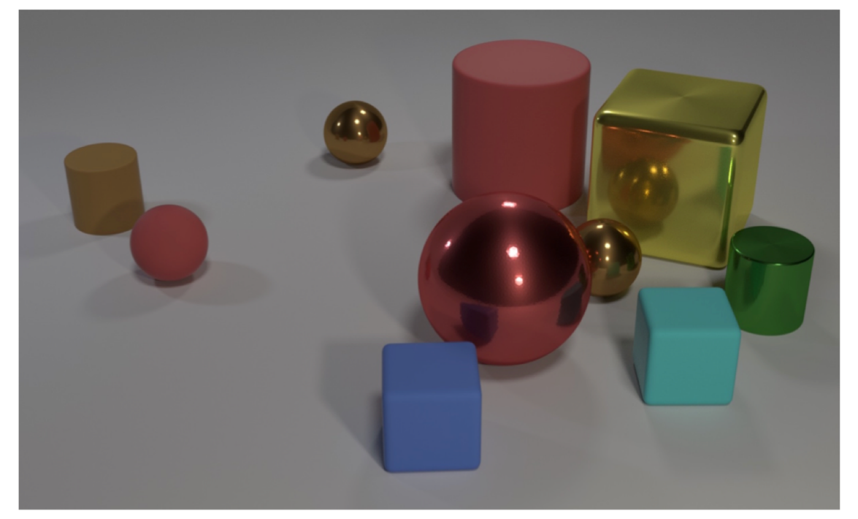
Super-CLEVR: A Virtual Benchmark to Diagnose Domain Robustness in Visual Reasoning

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

[1]Johns Hopkins University, [2]University of Southern California, [3]Max Planck Institute for Informatics, [4]University of Freiburg

Motivation: domain gaps in visual reasoning

CLEVR



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

VQA2.0

What color are her eyes?

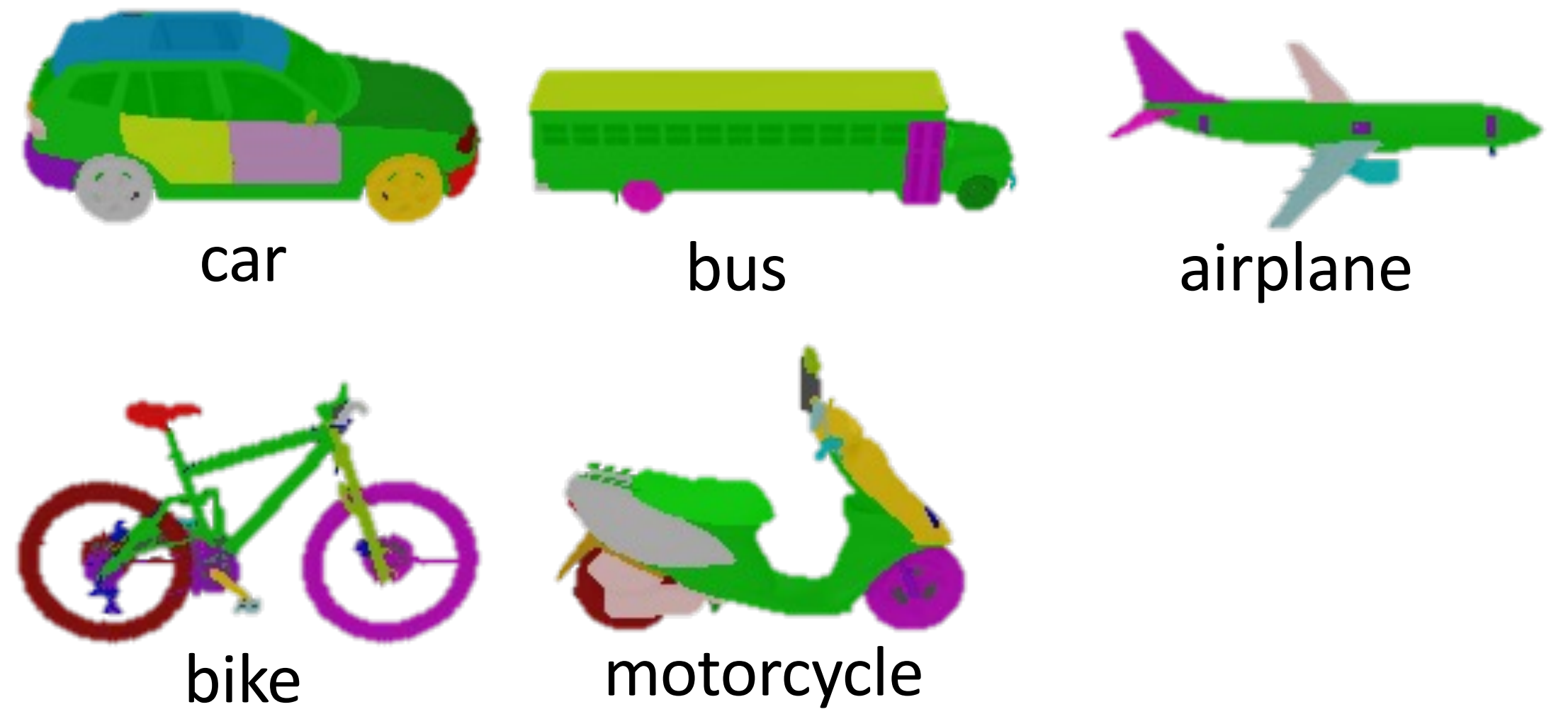
GQA

What color is the food on the red object left of the small girl holding a hamburger?

- Models suffer on out-of-domain testing.
- Due to multiple factors entangled with each other.

Dataset generation

Object with parts:



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

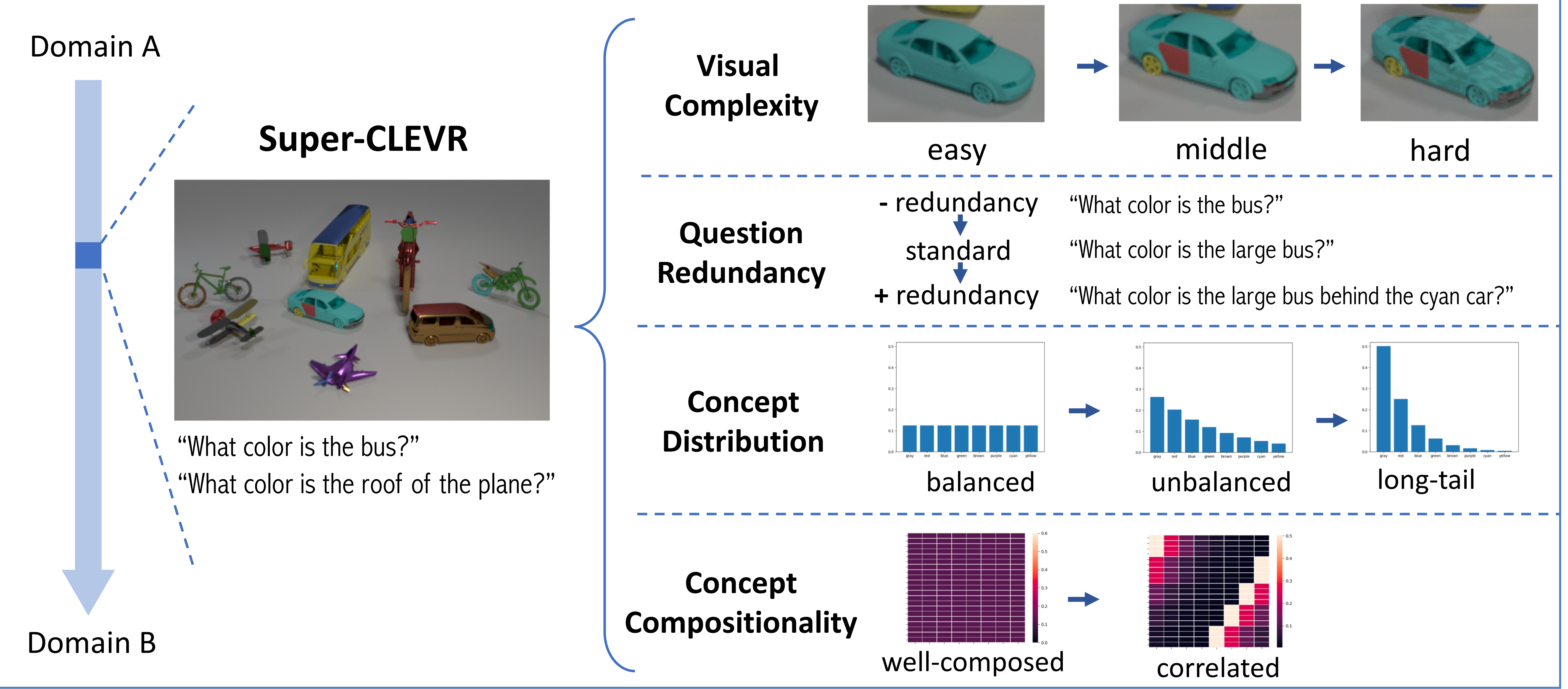
The dirtbike of the same size as the brown motorbike is what color?

5 models are studied

non-modular
- **FiLM**: two-stream feature merging
- **mDETR**: pretrained transformers

modular (symbolic)
- **NSCL**: neural symbolic concept learner
- **NSVQA**: neural symbolic VQA
- **P-NSVQA (ours)**: NSVQA + probability

Super-CLEVR: study each factor separately



"What color is the bus?"
"What color is the roof of the plane?"

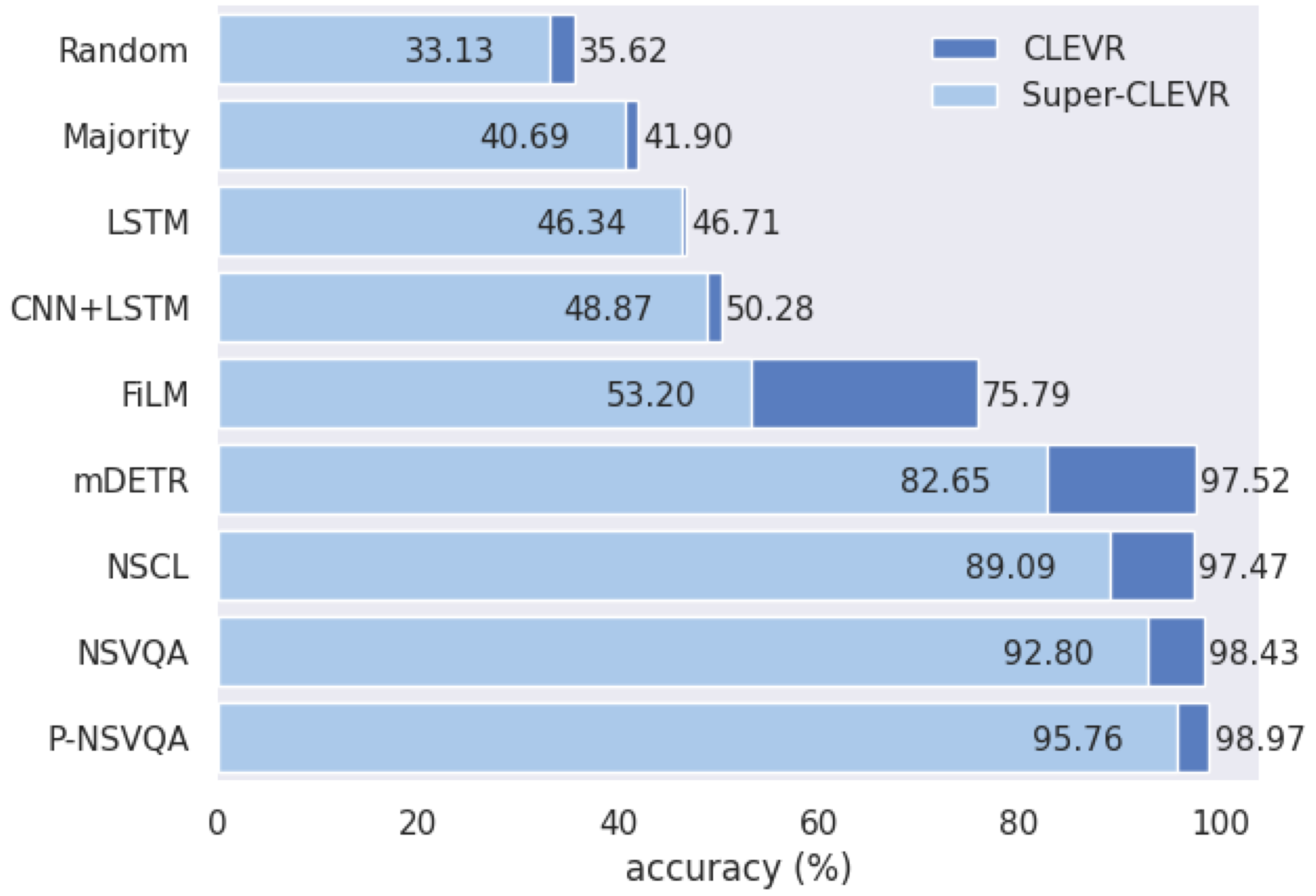
Visual Complexity: easy → middle → hard

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

In-domain results



- Super-CLEVR is challenging.
- P-NSVQA is the best.

Out-of-domain results

	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**

Relative Degrade (RD) of models' accuracy in OOD testing

1. Modular models are (only) 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

Conclusion

- The Super-CLEVR dataset.
- Modular + Probability -> best model