Measuring CLEVRness: Blackbox Testing of Visual Reasoning Models

Spyridon Mouselinos[1] Henryk Michalewski[12]
Mateusz Malinowski[3]

[1]University of Warsaw, [2]Google, [3]DeepMind

ICLR 2022

The State of VQA / VR

Visual Question Answering (VQA):

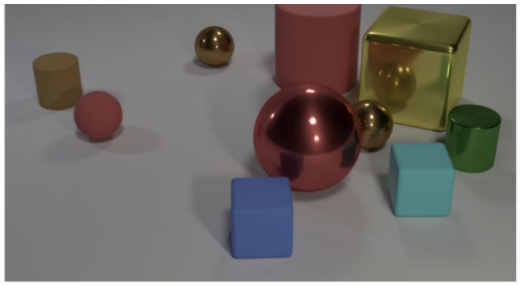
- Computer vision task.
- Text-based question about an image.
- Examples: DAQUAR[1], VQA[2], GQA[3]...

[1]A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input: Malinowski et al 2014.
[2]VQA: Antol et al 2015.
[3]GQA: Hudson et al. 2019

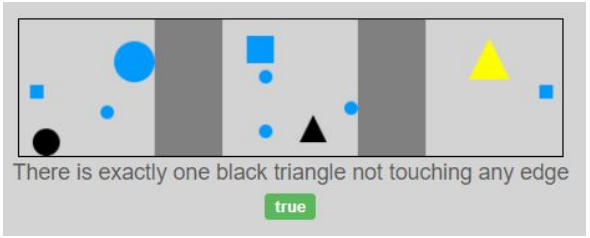
The State of VQA / VR

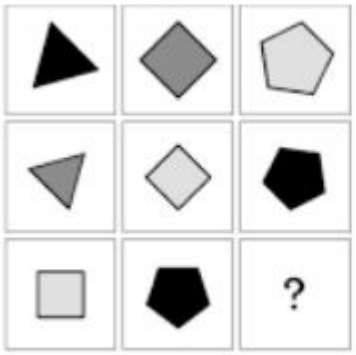
Visual Reasoning (VR):

- Same setup as VQA
- Questions require logical steps to be taken.
- Examples: CLEVR[1], NVLR[2], RAVEN[3]...



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





[1]CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning, Johnson et al. 2016
[2]NVLR: Natural Language for Visual Reasoning, Shur et al 2017.
[3]RAVEN: A Dataset for Relational and Analogical Visual rEasoNing, Zhang et al. 2019

The Problem

Current models:

- Handle effectively VQA / VR tasks.
- Achieve super-human performance.

Clever Hans in 1907

But...

- Exploitation of dataset biases.
- Entangled learning representations.

Is this a modern Clever Hans situation?

A concerning example: Robot assistant

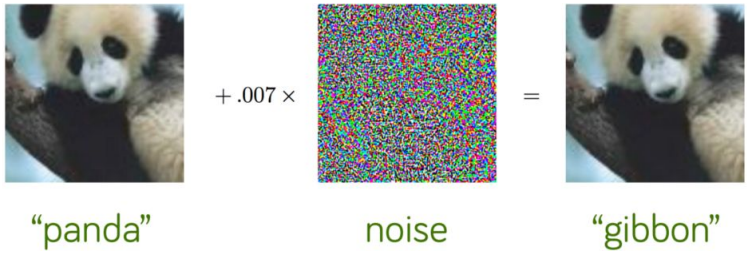
Robot: Pick Sphere!

Robot: ???

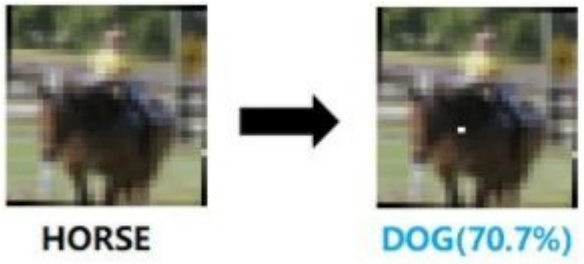
Concerns on other real-world examples (e.g autonomous driving, medical robots...)

Fooling visual systems

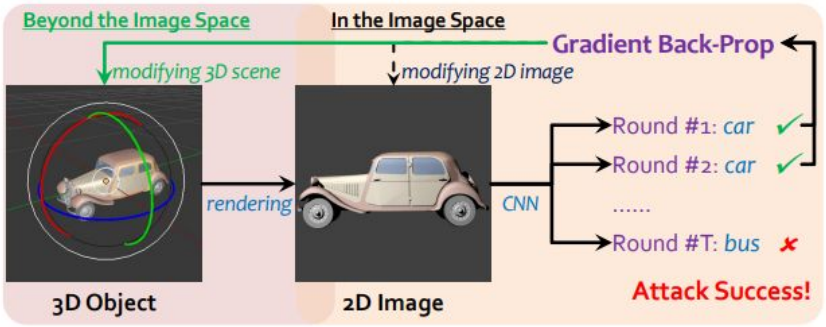
General Adversarial Attacks
(Goodfellow et al. 2015)



Pixel-Attacks
(Su et al. 2019)



Adversarial Attacks Beyond the Image Space (Zheng et al. 2019)



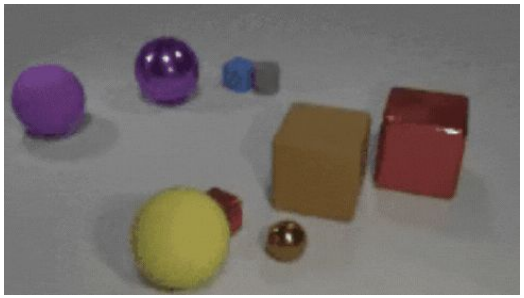
Limitations

Previous approaches need:

- Access to model internals / gradients.
- Access to model output probabilities.
- Access to differential renderer.
- A specific fooling target.

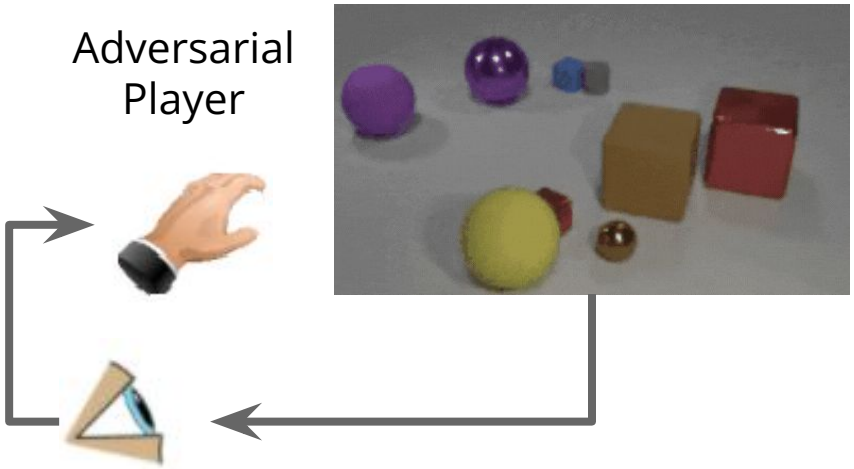
From static images to dynamic games

Static VQA



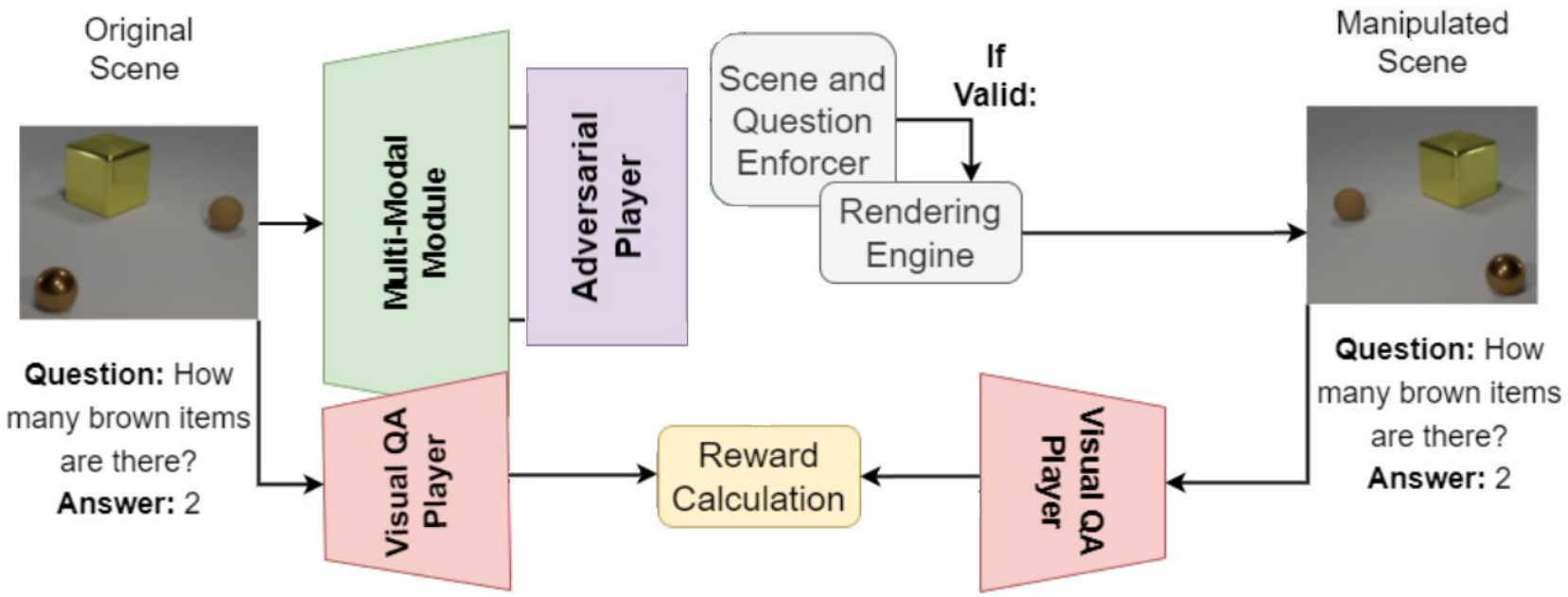
Question: What number of ....
Correct Answer: 0

Dynamic Game



Question: What number of ....
VQA Player answer before: 0
VQA Player answer after: ???

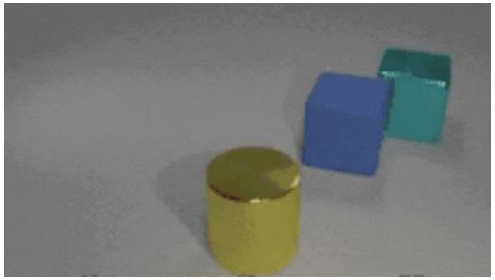
Method


Original Scene
Question: How many brown items are there?
Answer: 2
Multi-Modal Module
Adversarial Player
Visual QA Player
Scene and Question Enforcer
If Valid:
Rendering Engine
Manipulated Scene
Question: How many brown items are there?
Answer: 2
Reward Calculation
Visual QA Player

Dataset: CLEVR

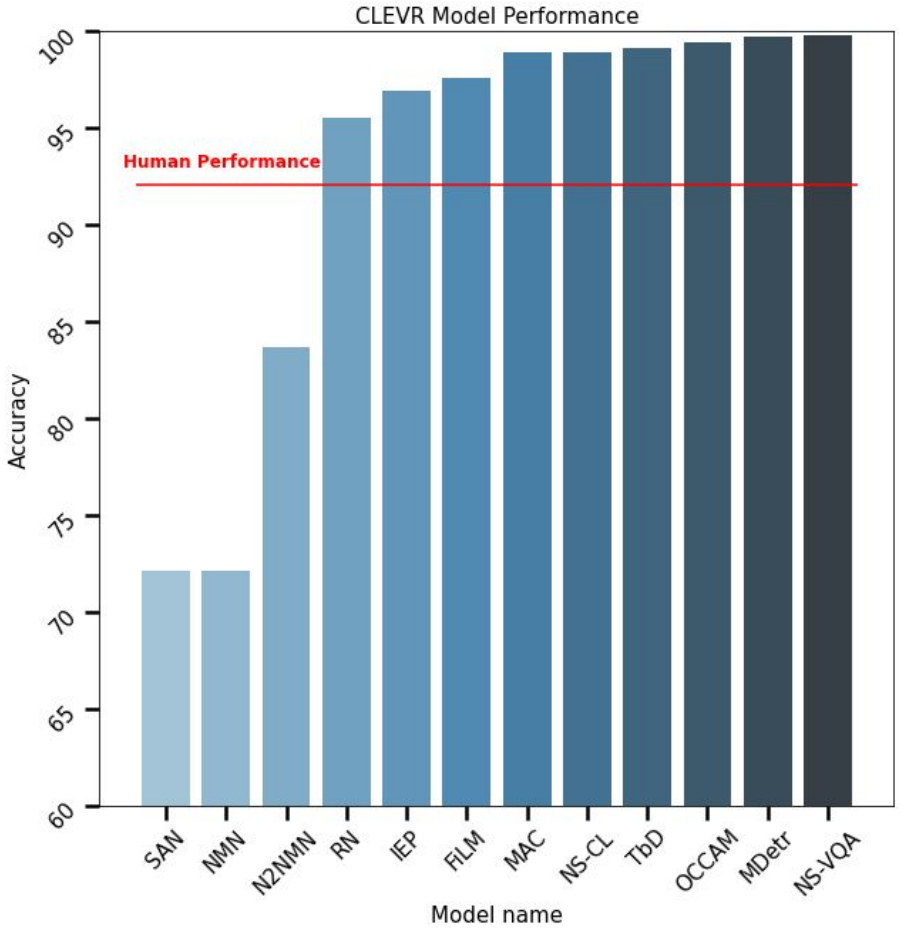
For our experiments we used the extensively studied CLEVR dataset.

- Visual reasoning.
- 700k training and 150k validation images.
- 2-10 Objects of 7 colors, 2 sizes, 2 materials and 3 shapes.
- Synthetic examples rendered by the Blender Rendering Engine.

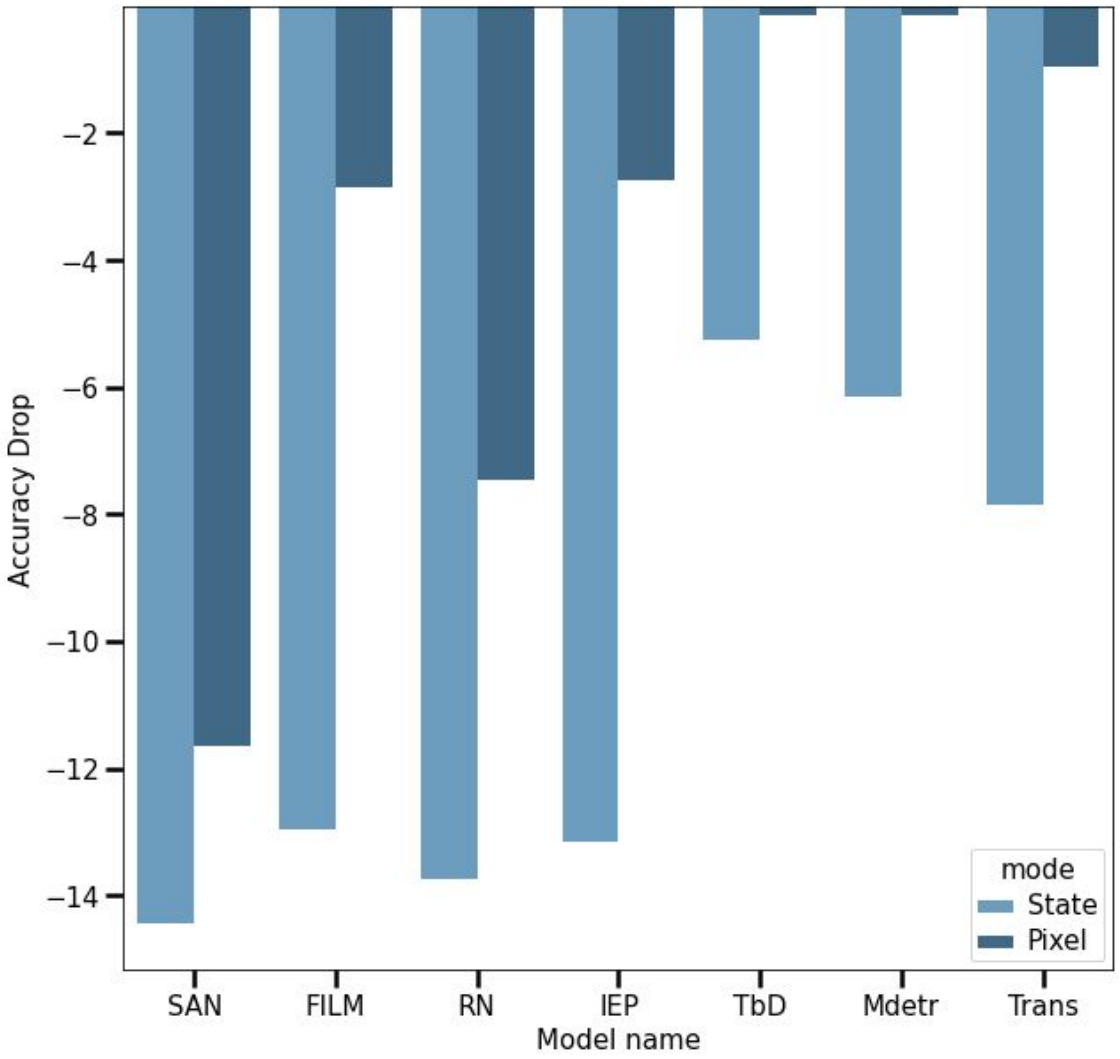


Example of CLEVR dataset:

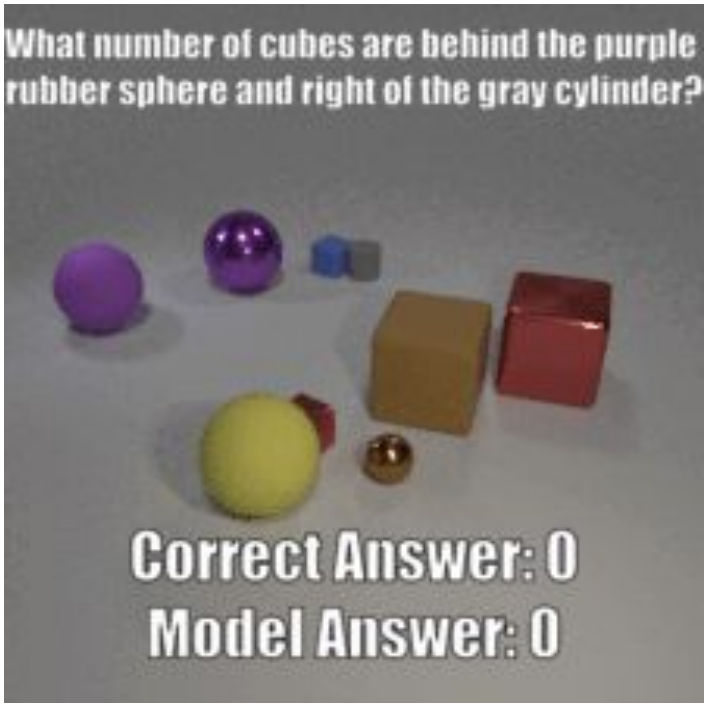
Question: How many cubes are there?
Answer: 2

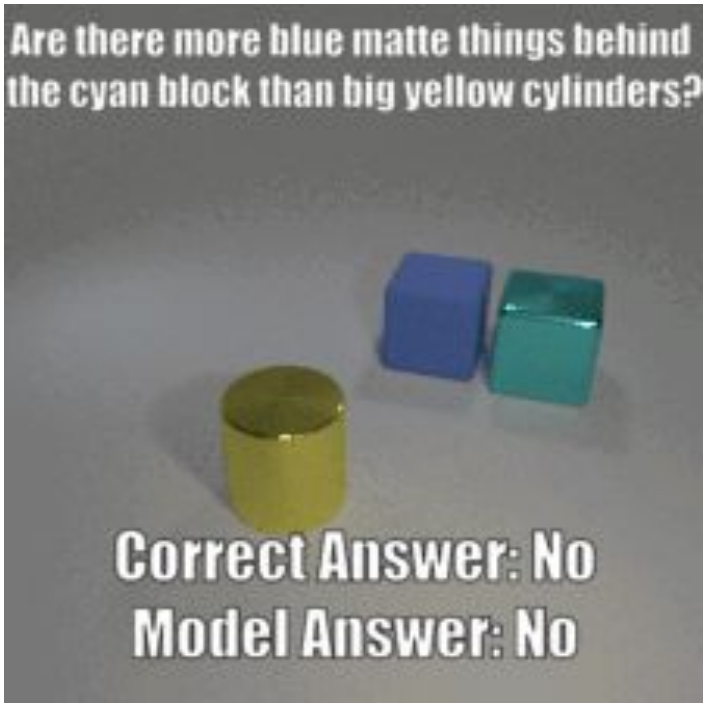


Performance Drop



Visual Fooling Demonstration





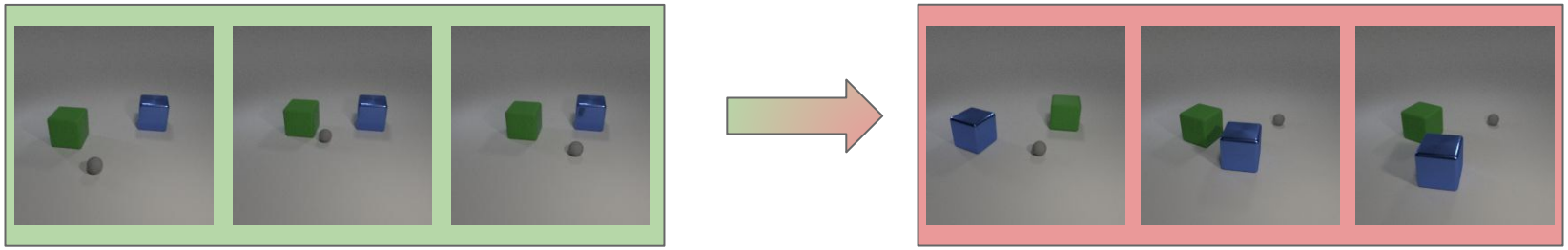
More data = Robustness?

If the VQA player was trained on all available perturbations → No fooling.

Do we really need all of those examples?

Let's simulate this on a smaller scale:

- [2-4] Objects / [1-10] Questions / [1-2] Reasoning steps per question.

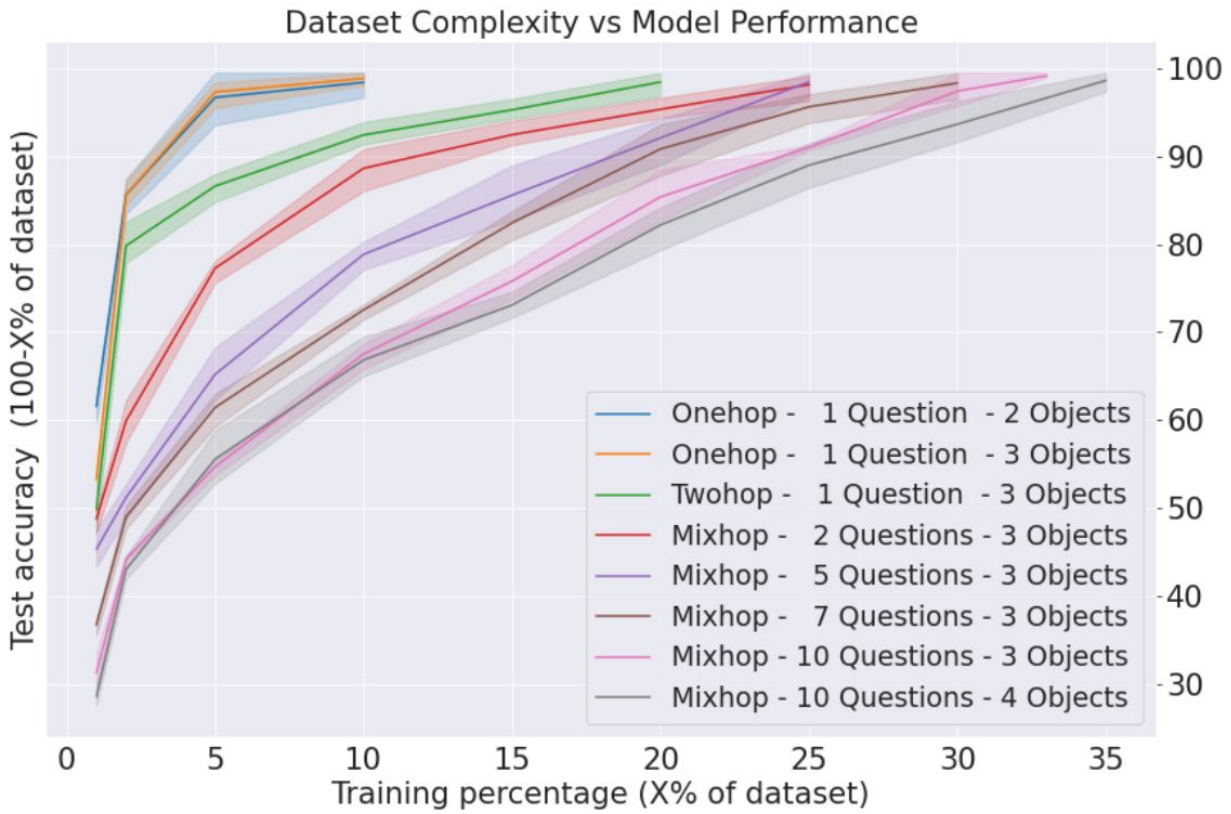


Training set (X% of all perturbations)

Test set (100-X% of all perturbations)

Data requirements

-



Models are susceptible to reasoning gaps and require extra data, ***proportional to their task complexity*** in order to generalize to all possible configurations.

Thank you!