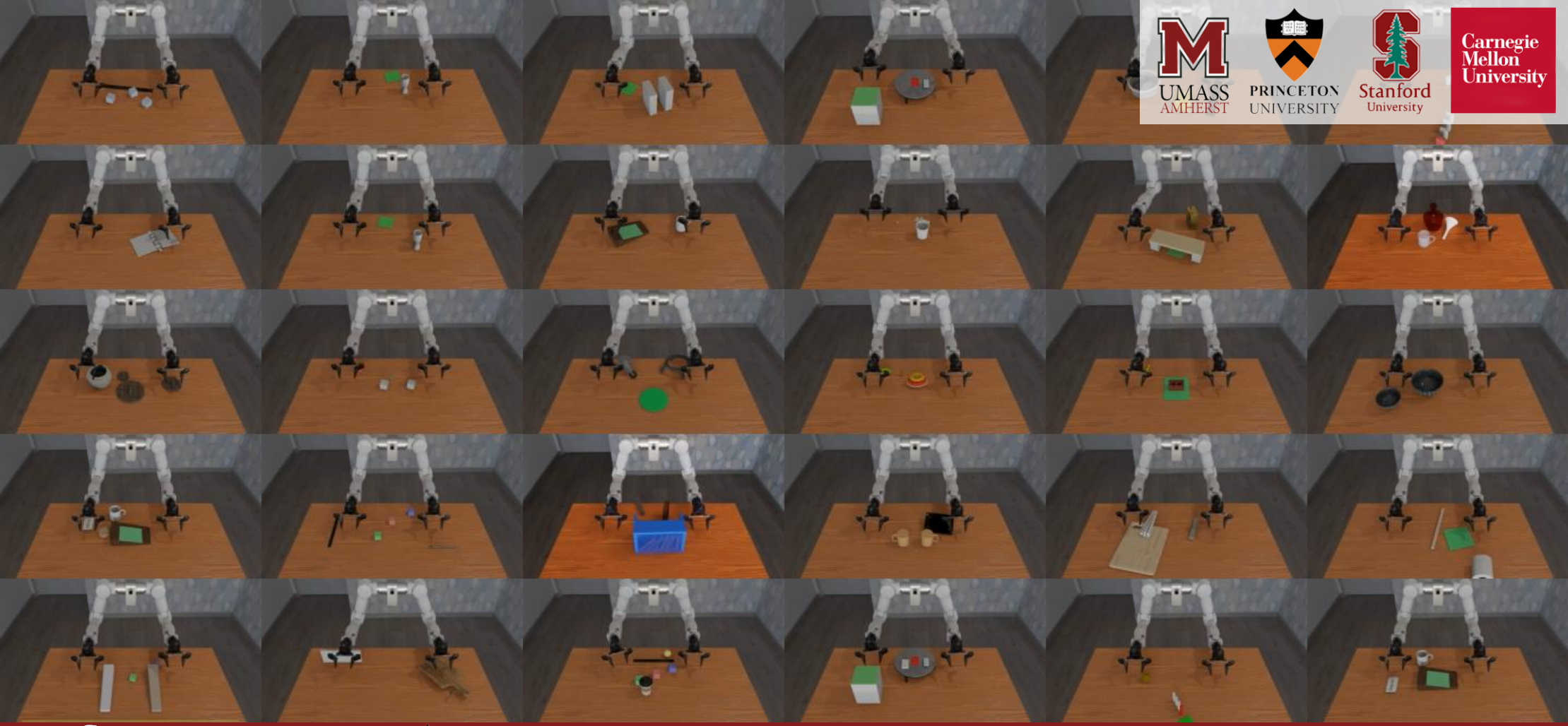


UMASS AMHERST · PRINCETON UNIVERSITY · Stanford University · Carnegie Mellon University

RobWits

Unexpected Challenges for Robotic Creative Problem Solving

Chunru Lin*, Hongxin Zhang*, Fenghao Yu, Zhehuan Chen, Thomas L. Griffiths, Yejin Choi, David Held, Chuang Gan

Content

Background

- Why build a new benchmark?
- What types of tasks should it include?

Automated Task Generation

- How to generate robotic tasks about creative problem solving at scale?

Benchmark Overview

- Task suites to evaluate creative problem solving & strategy adaptation

Evaluation Results

- How does recent models perform on RoboWits benchmark?

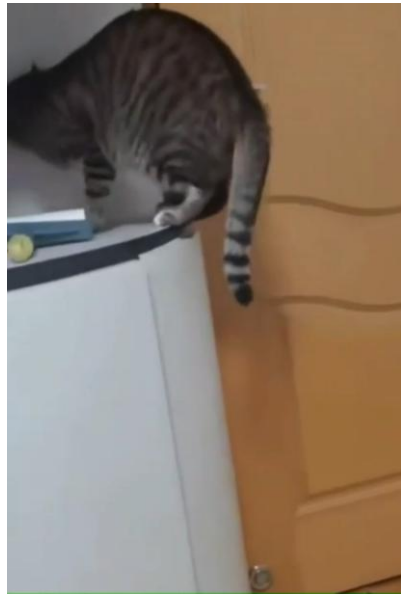
Background

Motivation

The ability to **creatively solve** problems and **adapt to** unexpected challenges is fundamental to **intelligent behavior.**



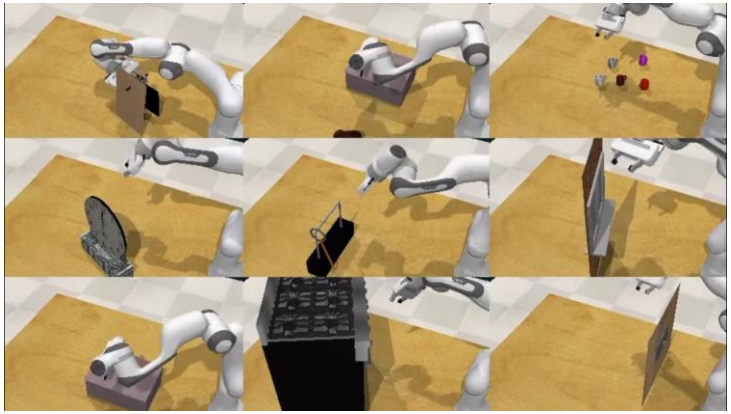






References: https://www.youtube.com/watch?v=gSnIjxVKlqo, https://www.youtube.com/watch?v=Cag05YVl-cM, http://xhslink.com/o/3J2MFjzGhjX, http://xhslink.com/o/5Sqiyni86xY

Existing Benchmarks



RLBench (ICRA 2020)



LIBERO-100 (NeurIPS 2023)

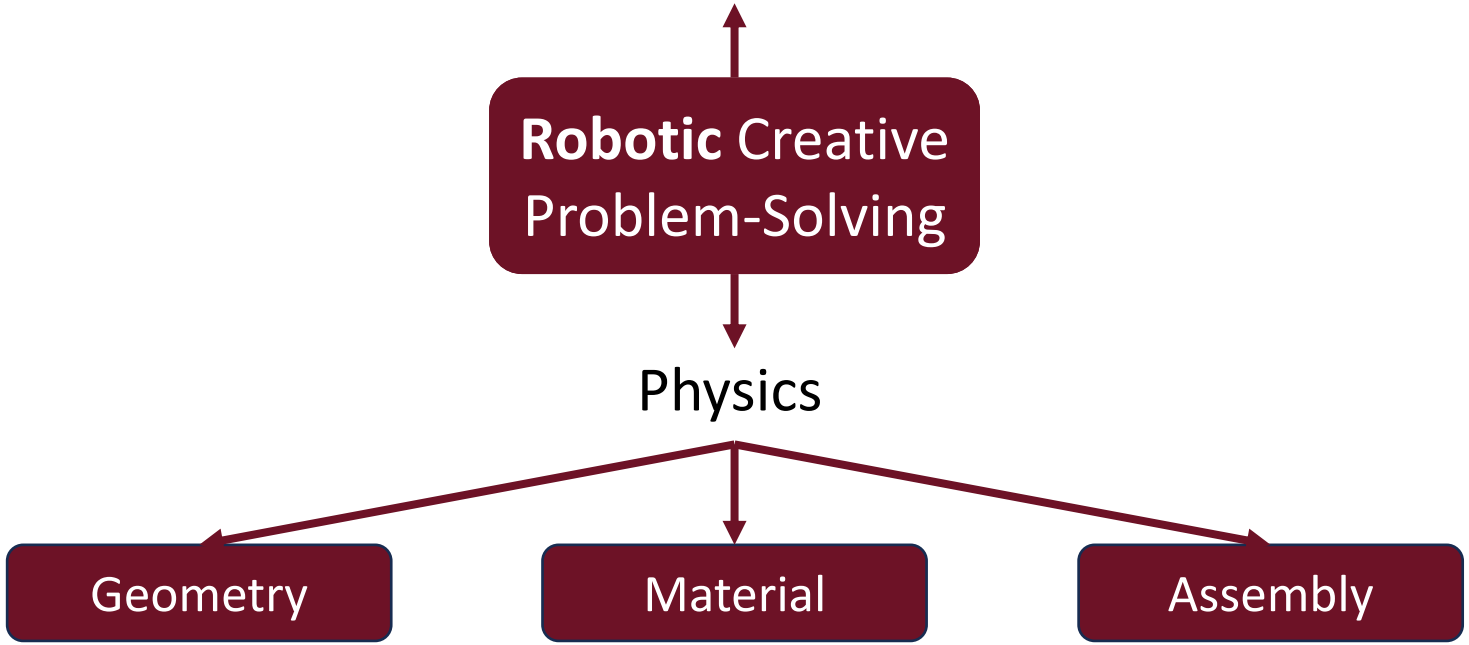


RoboTwin 2.0 (ICML 2026)

References: https://www.youtube.com/watch?v=og4c1urBoqA, https://libero-project.github.io/datasets, https://robotwin-platform.github.io/

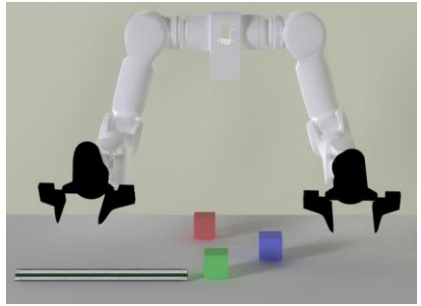
How to define “creative problem solving” for robots?

Trivial/Naïve solutions are invalidated or inefficient



Task Formulation

- **A simulation env** (objects with physical parameters and initial placement)
- **An instruction** (natural language)
- **An evaluation metric** (runable code)



"Align the three blocks perfectly in a straight line."

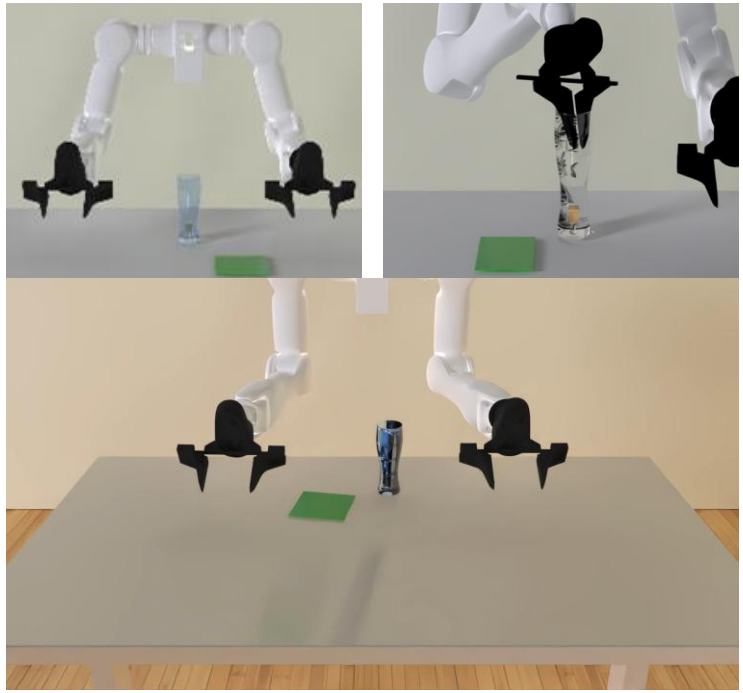
✅ All objects are on the table
✅ Three blocks are collinear
✅ Robot isn't holding any object

```
def is_completed(objs_info) -> bool:
    block = objs_info["block"]
    … …
    return on_same_line(blocks)
```

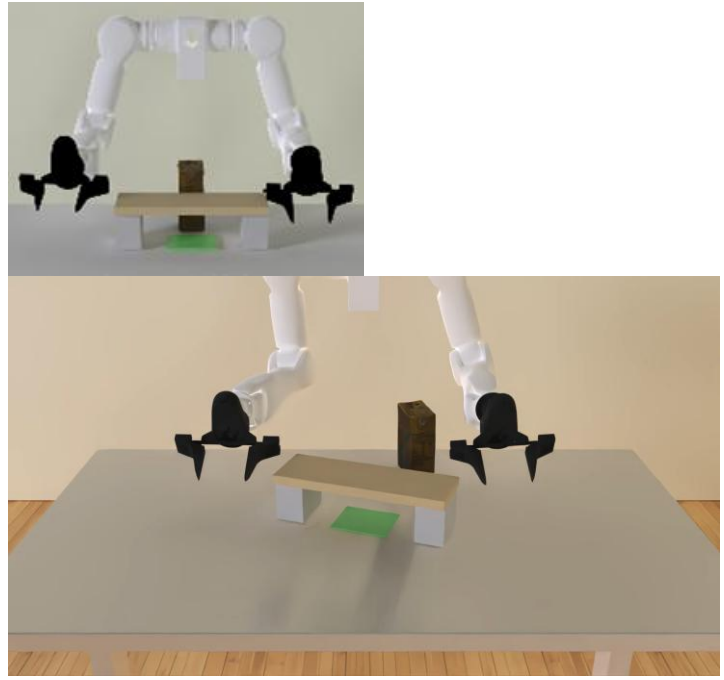
The reasoning comes from that **no solution hints** are in the instructions.

Example Tasks — Geometry

"Retrieve the cube from the container and place it on the green target area."

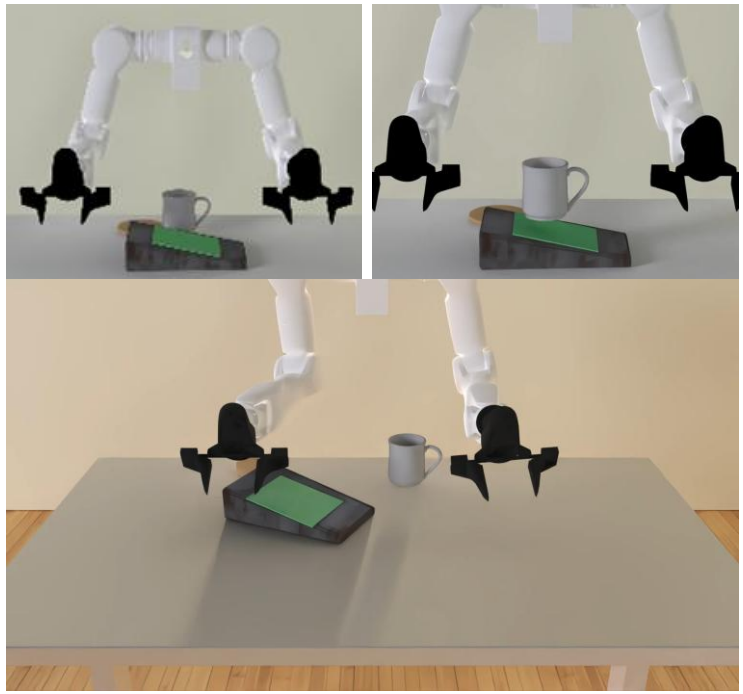


"Move the rectangular box onto the green target area."

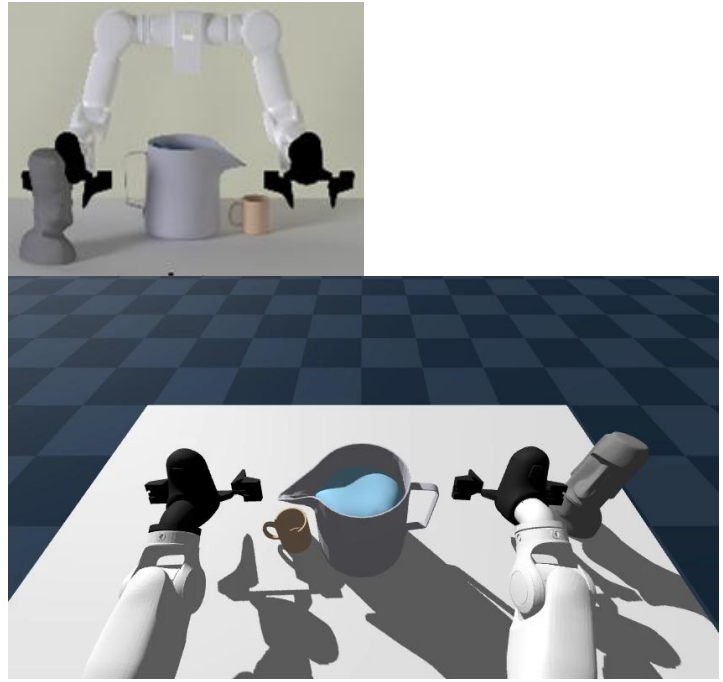


Example Tasks — Material

"Put the cup on the green target area on the slope without sliding down."

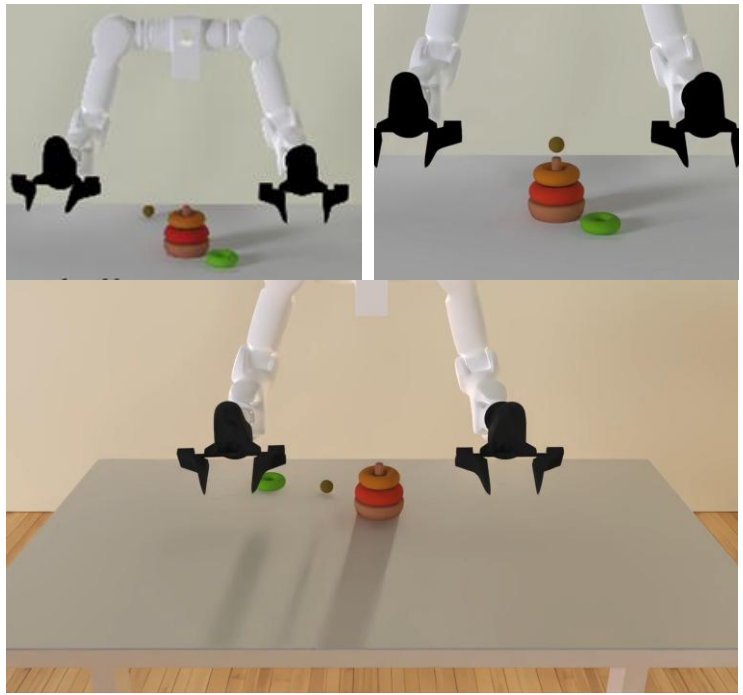


"Collect water in the mug without moving the pitcher."

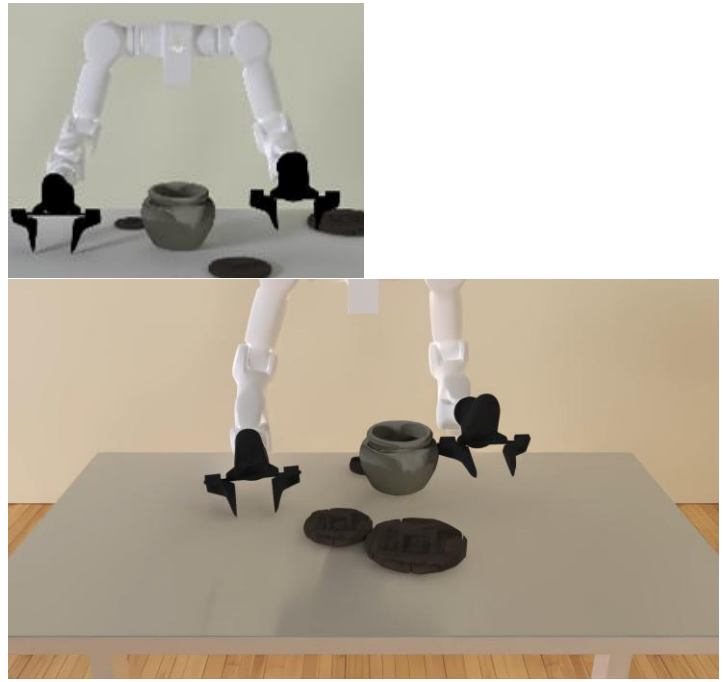


Example Tasks — Assembly

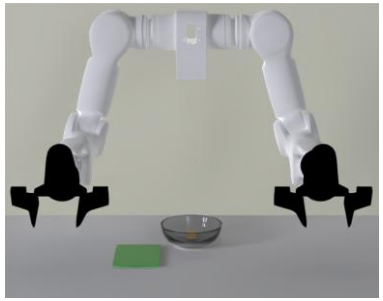
“Place the ball stably on top of the tower.”



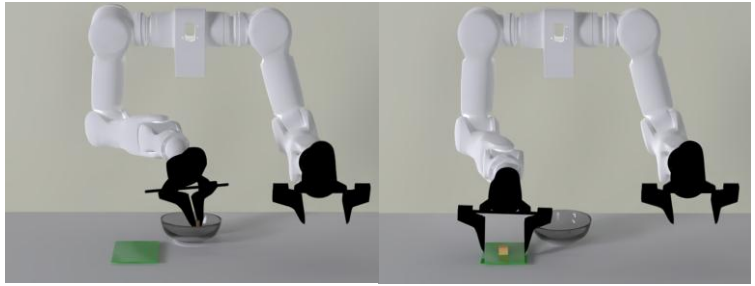
“Cover the pot precisely using one lid.”

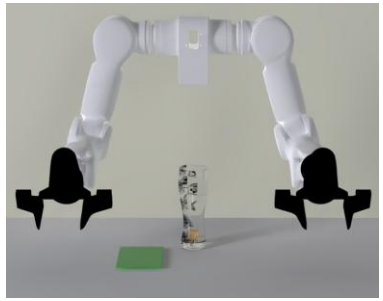


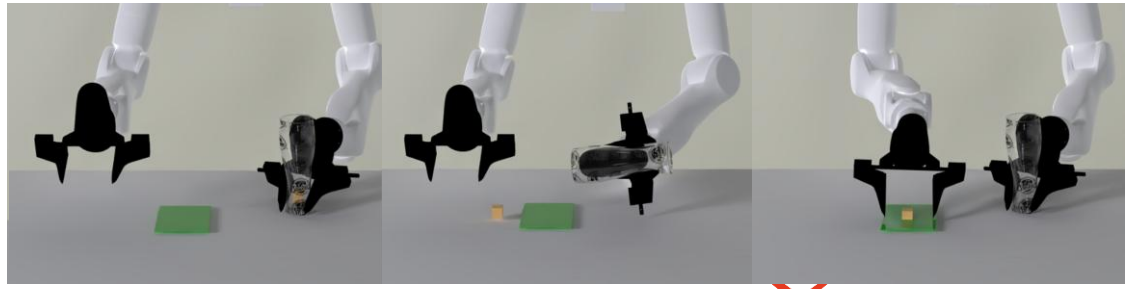
Task mutations in the reasoning level



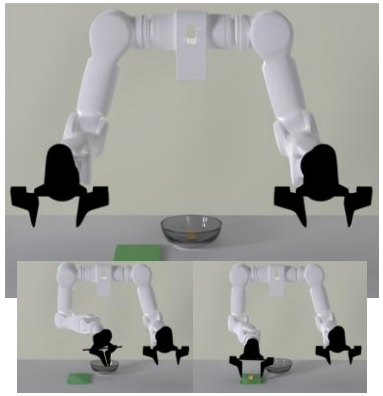
Instruction: Retrieve the cube from the container and place it on the green target area.

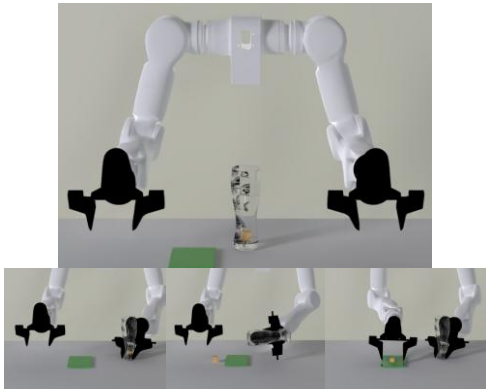




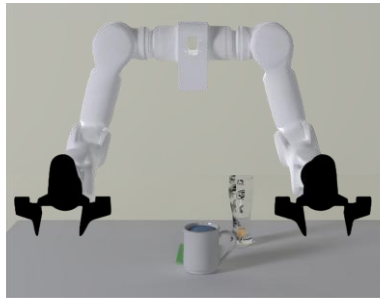


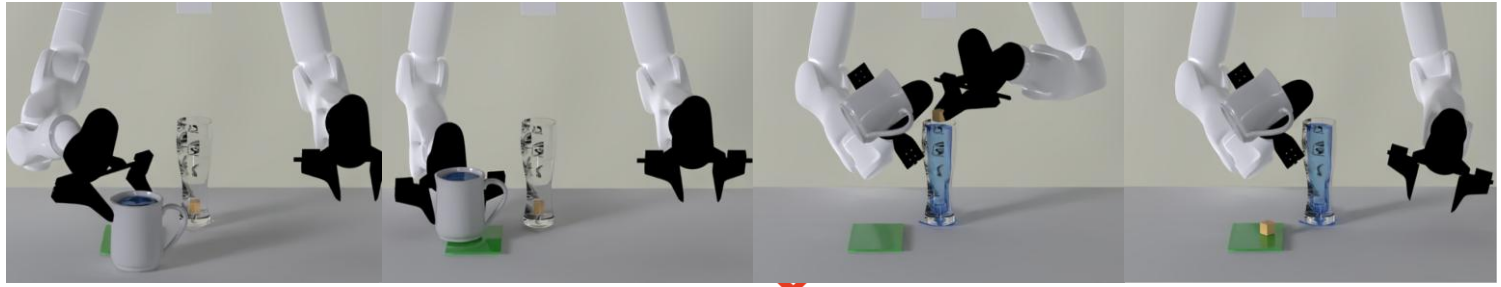
Task mutations in the reasoning level





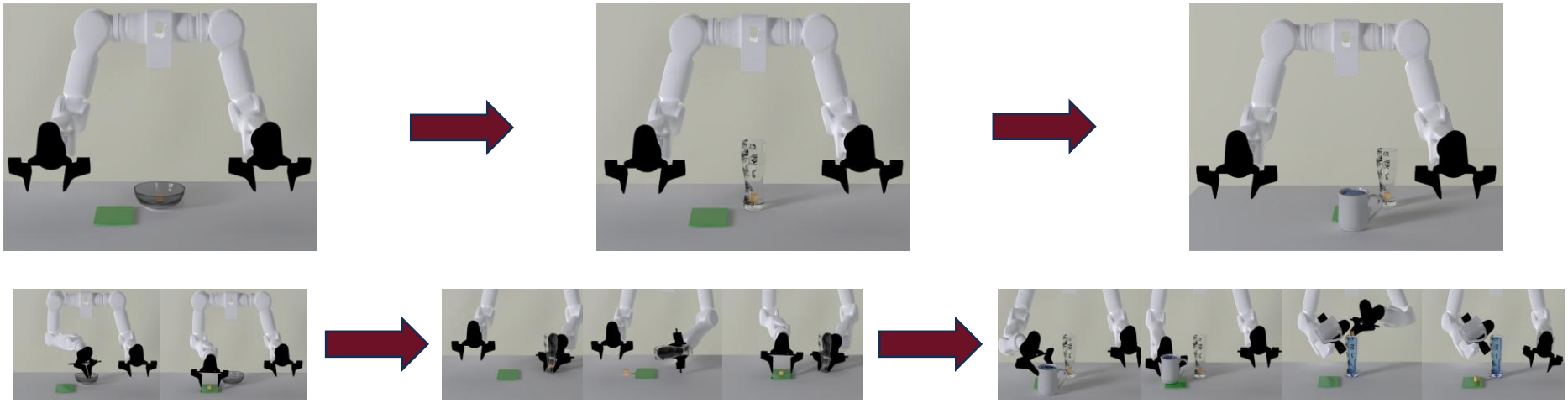
Instruction: Retrieve the cube from the container and place it on the green target area.



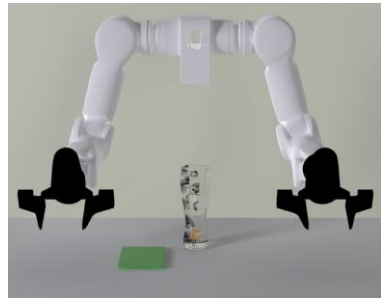
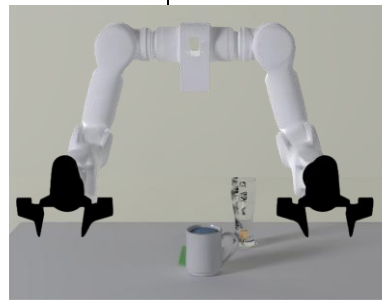


Task mutations in the reasoning level

Instruction: Retrieve the cube from the container and place it on the green target area.



Mutation Strategies

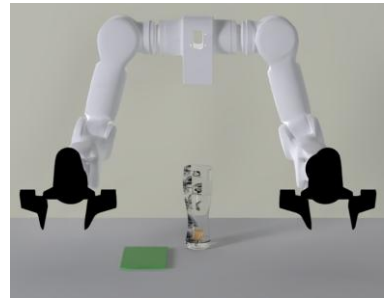
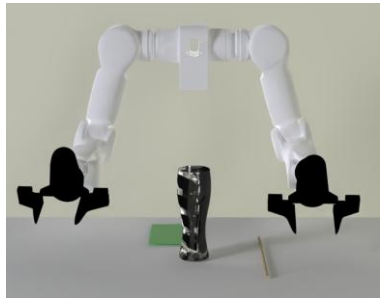



- **Pivot**

  Block the current solution and enable another

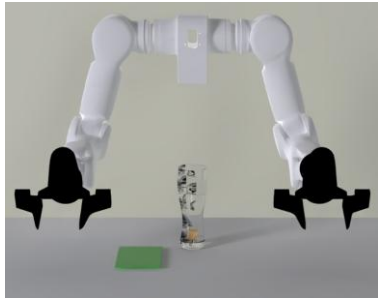
- **Trap**

  Add objects that appear useful but are functionally ineffective

- **Add**

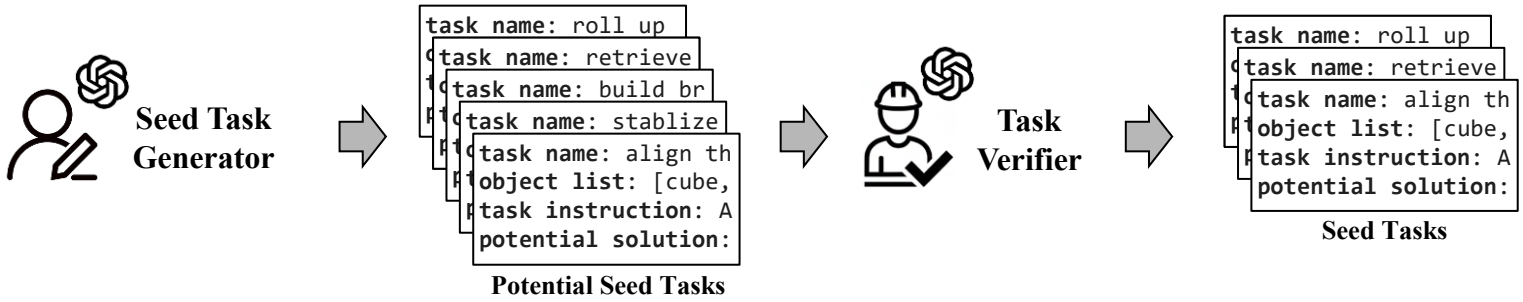
  Add irrelevant objects




Automated Task Generation

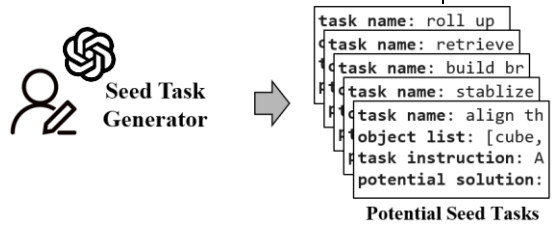
Automated Task Generation

1. Seed Task Generation



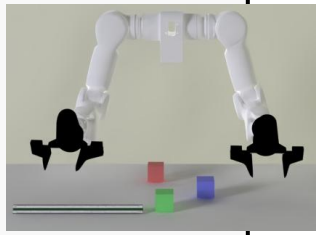
2. Task Mutation and Instantiation

Seed Task Generator

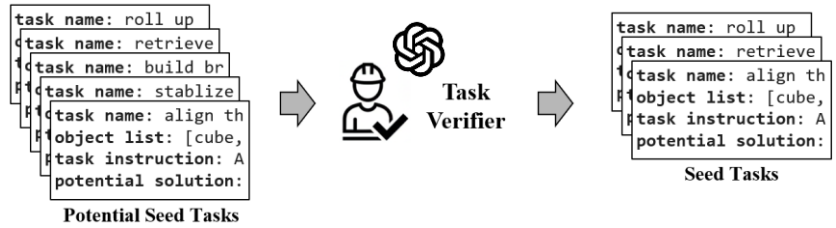


Task Schema

```
{
  "task_name": "a concise and meaningful name of the task",
  "object_list": [
    {
      "object_name": "abstract functional name of the object",
      "appearance_attribute": ["shape, size, color, high or short, ..."],
      "functional_attribute": ["material, density, friction, movability, ..."],
      "potential_instances":  ["at least 3 daily object names with required attributes"]
    },
    ... // continue for all necessary objects
  ],
  "initial_scene_setup": "describe the required position and/or relationship of all objects in the object_list",
  "task_instruction": "the goal of the task without hinting the solution",
  "success_criteria": "the signals to show the task is completed",
  "potential_solution": "how the task should be solved with objects provided",
  "task_description": "describe the task in detail"
}
```



Task Verifier



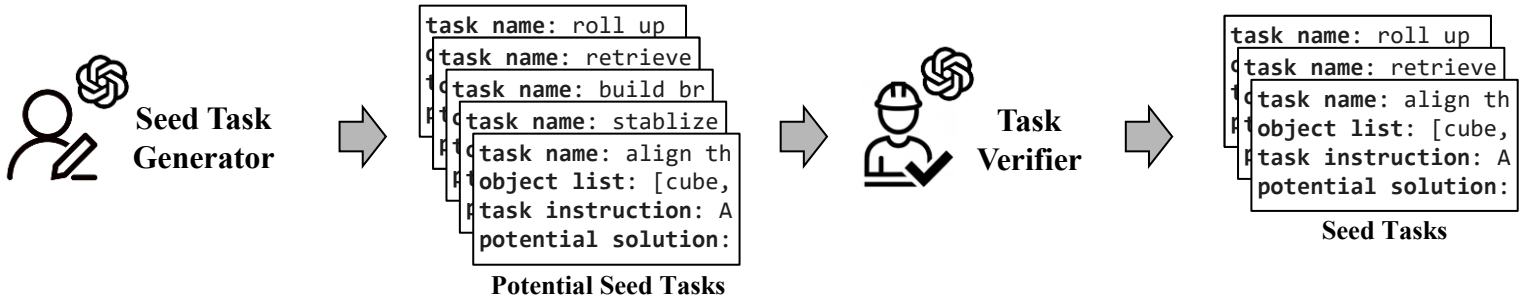
Verification Result Format

```
{
  "completeness": {
    "completeness": "yes"/"no",
    "reason": "..."
  },
  "simulatability": {
    "difficulty": "easy"/"hard"/
                  "impossible",
    "reason": "..."
  },
  "solution_feasibility": {
    "feasibility": "very feasible"/
                   "kind of feasible"/
                   "not feasible",
    "infeasible_step": "...",
    "reason": "..."
  },
  "solution_efficiency": {
    "efficiency": "yes"/"no",
    "bypass_solution": "..."
  },
  "difficulty": {
    "score": "1-5",
    "reason": "..."
  }
}
```

- Do all mentioned objects appear in the object list?
- Does the physics engine support the required dynamics?
- Is the potential solution valid?
- Are there any trivial/easier solutions to solve the task?
- Rate the difficult score based on predefined criteria.

Automated Task Generation

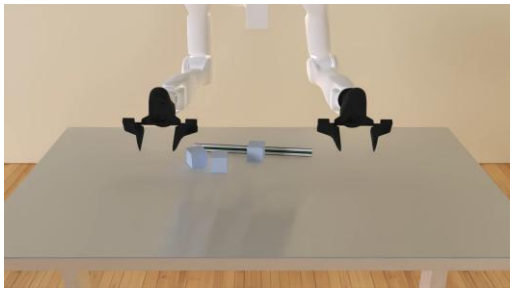
1. Seed Task Generation

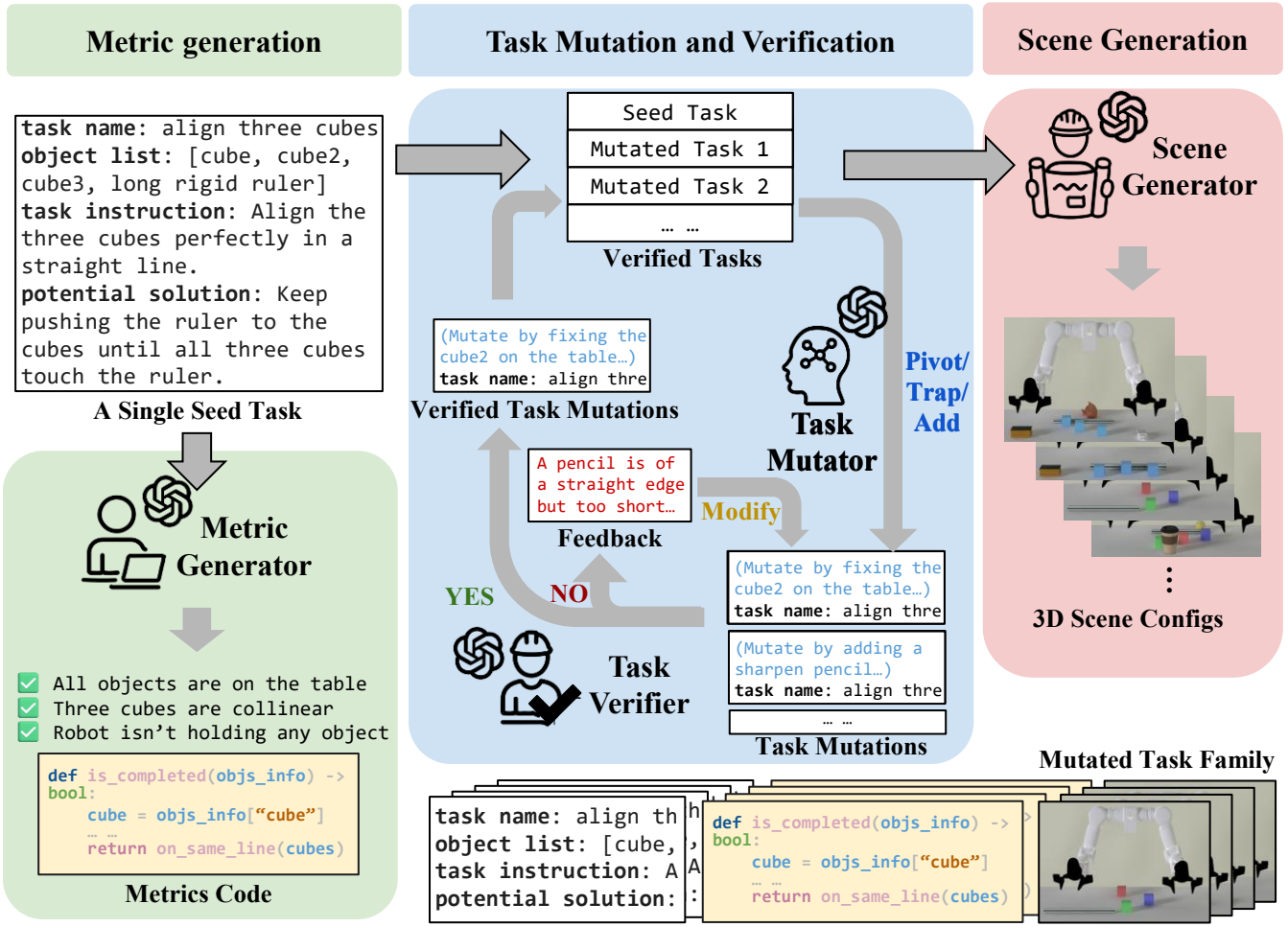


2. Task Mutation and Instantiation

Task Mutation and Instantiation

UMassAmherst | Manning College of Information & Computer Sciences





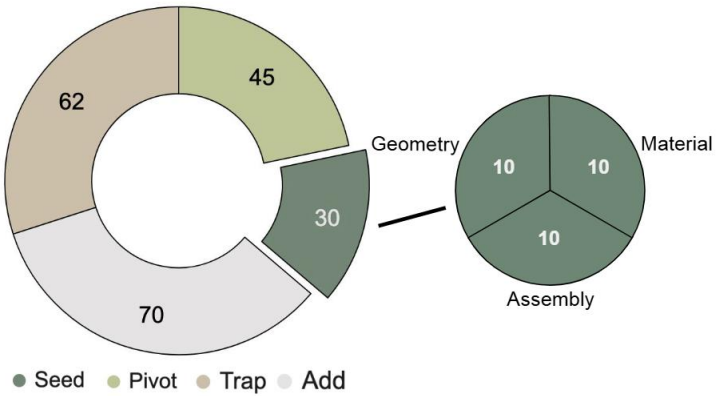
Benchmark Overview

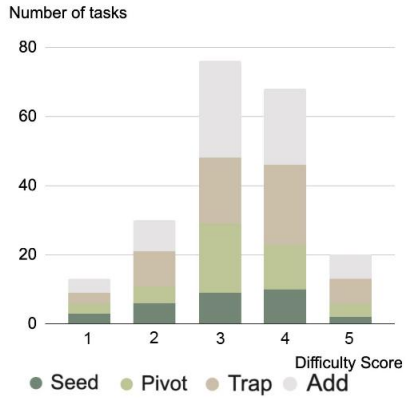
Task Statistics

208 bi-manual creative manipulation tasks

30 seed tasks across 3 categories

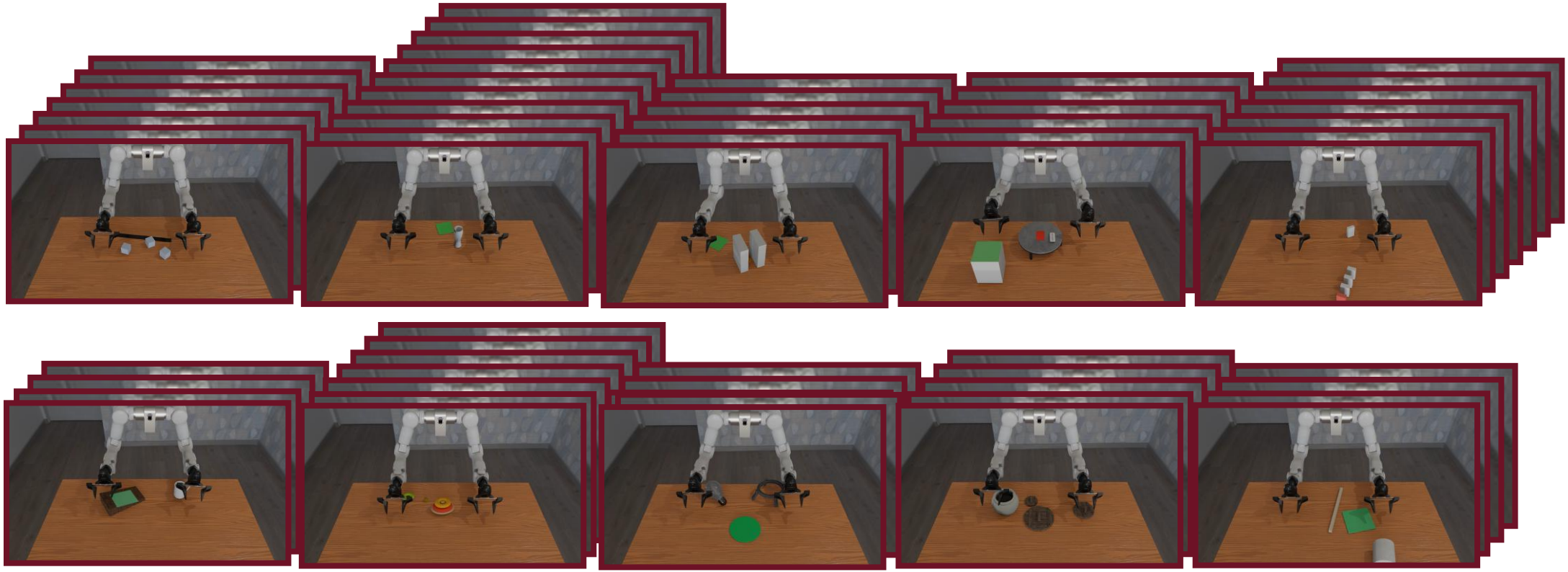
5 difficulty levels





	Pivot	Trap	Add
count	45	62	70
mean	0.26	0.42	0.09
max	3	3	1
min	-2	-1	-1
std	0.98	0.79	0.31

Task Suites



Evaluation Results

Can robotic foundation models creatively solve manipulation problems?

	Success Rate	Progress Score
ACT	11%	0.43
PI0	12.2%	0.46
PI0.5	12.6%	0.45

Current VLAs struggle on RoboWits

Can robotic foundation models robustly handle unexpected challenges?

	Seed		Mutation	
	Success Rate	Progress Score	Success Rate	Progress Score
ACT	11%	0.43	4.8%	0.31
PI0	12.2%	0.46	6.5%	0.32
PI0.5	12.6%	0.45	9.8%	0.31

Current VLAs lack robustness under unexpected challenge

Where do the failures stem from?

	Pivot	Trap	Add(1)	Add(2)	Add(3)
ACT	1.7%	5.1%	13.3%	13.5%	4.3%
PI0	2.0%	6.4%	13.3%	17.6%	11.1%
PI0.5	3.7%	5.4%	15.0%	12.1%	7.7%

Pivot has lowest performance, showing model have merely no strategy transfer

Performance drops as more objects are introduced.

Can modular planners with Foundation Model think-out-of-the-box?

	Seed		Mutation	
	Success Rate	Progress Score	Success Rate	Progress Score
ACT	11%	0.43	4.8%	0.31
PI0	12.2%	0.46	6.5%	0.32
PI0.5	12.6%	0.45	9.8%	0.31
VLM Controller	7%	0.45	4.8%	0.36
VLM Planner	60.8%	/	44.4%	/

Foundation Model Planners are promising assuming error-free control and oracle feedback

UMassAmherst | Manning College of Information & Computer Sciences

Thanks for your attention!

Q & A