# Generalizable Autonomy

## Representations for Embodied FMs

Animesh Garg

Professor of AI Robotics
Georgia Tech

# Generalizable Autonomy

1950s
“AI”

The Dartmouth AI Project

# Generalizable Autonomy

1950s
"AI"

1968: Aspirational Robotic Assistant

# Generalizable Autonomy

1950s "AI" → 1960s Robots → 2013 Walking → 2019 Cooking → 2024 Humanoids

# The Computing Stack

Digital AI

App

General-Purpose Applications
Ease of Use

OS

Platform-Agnostic OS
Modular Utilities

Driver

Hardware-Specific Drivers
Optimize Performance

# The Computing Stack

## Digital AI

App
General-Purpose Applications
Ease of Use

OS
Platform-Agnostic OS
Modular Utilities

Driver
Hardware-Specific Drivers
Optimize Performance

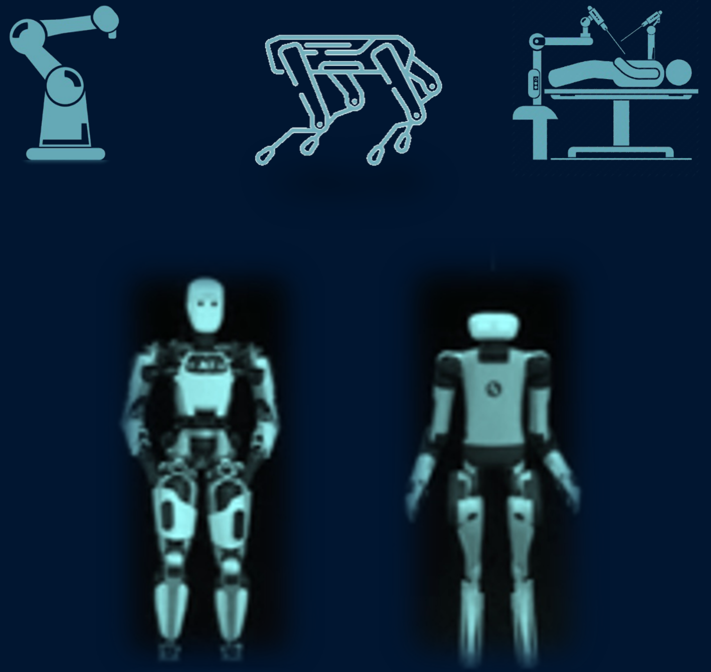
## Physical AI

App
Natural Interaction Interface
Ease of Use

OS
Platform-Agnostic Planning
Modular Tool-Use

Driver
Hardware-Specific Skills
Optimize for Morphology

# The Computing Stack
## Physical AI



Natural Interaction Interface
Ease of Use

App

Platform-Agnostic Planning
Modular Tool-Use

OS

Hardware-Specific Skills
Optimize for Morphology

Driver

# The Computing Stack

## Physical AI

Internet Data
Language, Image, Video
$, Very Diverse

Synthetic Data
Simulation
$$, Engineered Designs

Real World Data
Teleoperation
$$$$, Limited Diversity

Natural Interaction Interface
Ease of Use

App

Platform-Agnostic Planning
Modular Tool-Use

OS

Hardware-Specific Skills
Optimize for Morphology

Driver

# The Computing Stack

## Physical AI

### Planning with LLMs

Learning Planning Domains

Using Agentic Frameworks with Verifiers for correctness

Automated Iterative completion of Domain Specification

Natural Interaction Interface
Ease of Use

App

Platform-Agnostic Planning
Modular Tool-Use

OS

Hardware-Specific Skills
Optimize for Morphology

Driver

# CLIMB

## Language-Guided Continual Learning for Task Planning with Iterative Model Building

Walker Byrnes, Miroslav Bogdanovic, Avi Balakirsky, Stephen Balakirsky, Animesh Garg

# CLIMB Objectives

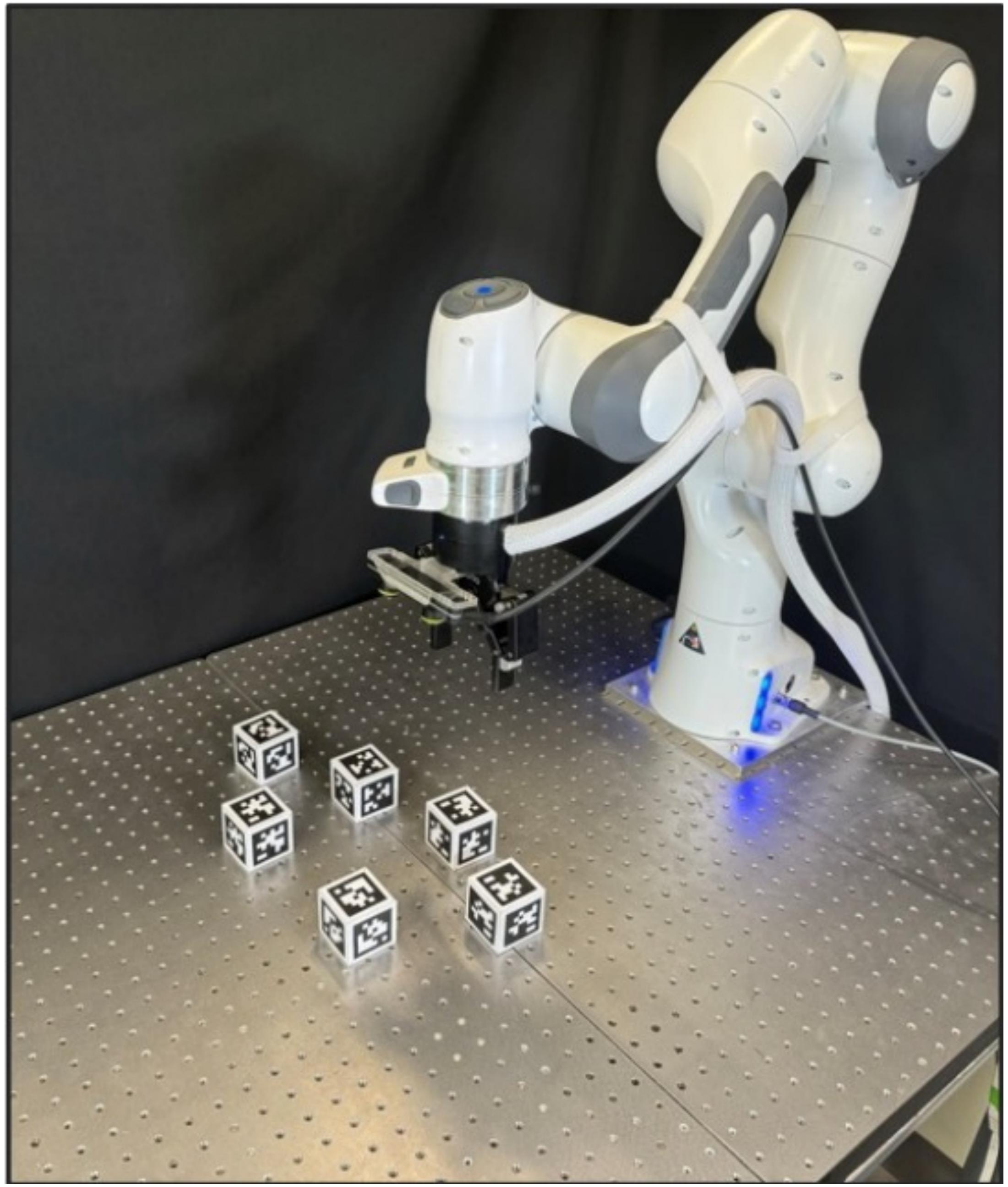
Given a domain description and task in **natural language**, learn a world model which represents domain state constraints through autonomous interaction with environment

Inputs:

- Domain Description
- Logical Action set
- Tasks to complete

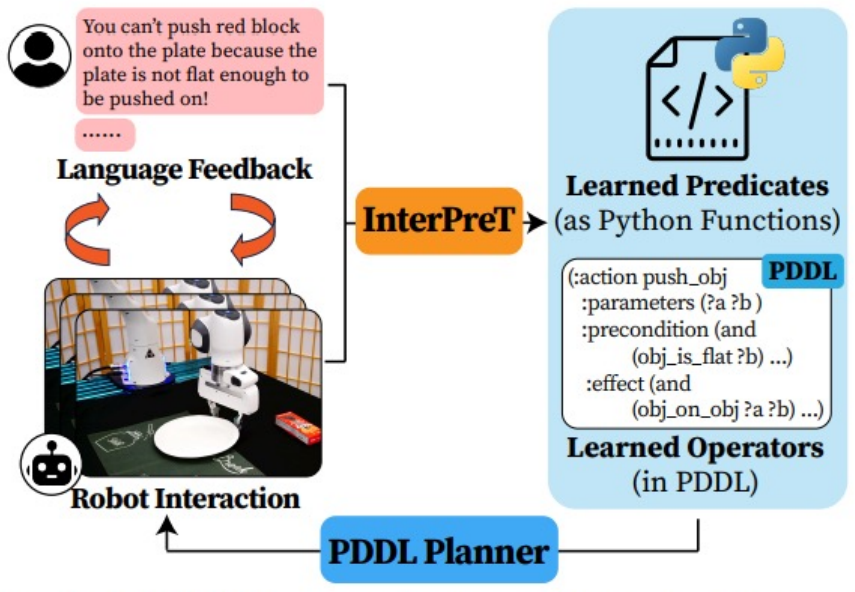
Outputs:

- World representation encoding task requirements
- Completed tasks in curriculum



# Prior Efforts

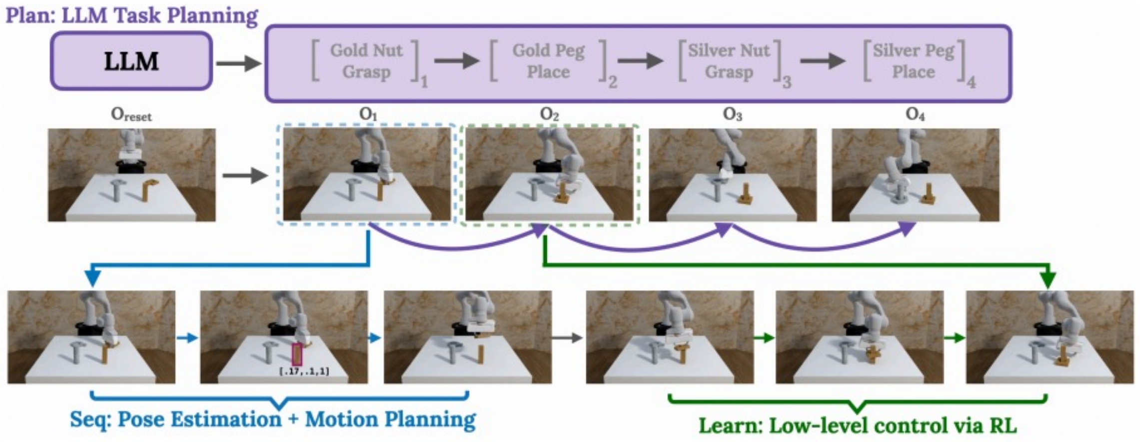
- **InterPreT** grounds language instructions in Problem Domain Definition Language (PDDL)
  - Relies on human input for corrective actions
- **Plan-Seq-Learn**
  - Learns atomic skills to accomplish dexterous actions, sequenced with classical task planner




INTERPRET: Interactive Predicate Learning from Language Feedback for Generalizable Task Planning


Muzhi Han[1], Yifeng Zhu[2], Song-Chun Zhu[1], Ying Nian Wu[1], Yuke Zhu[2]
[1]University of California, Los Angeles, [2]The University of Texas at Austin
https://interpret-robot.github.io


PLAN-SEQ-LEARN: LANGUAGE MODEL GUIDED RL FOR SOLVING LONG HORIZON ROBOTICS TASKS

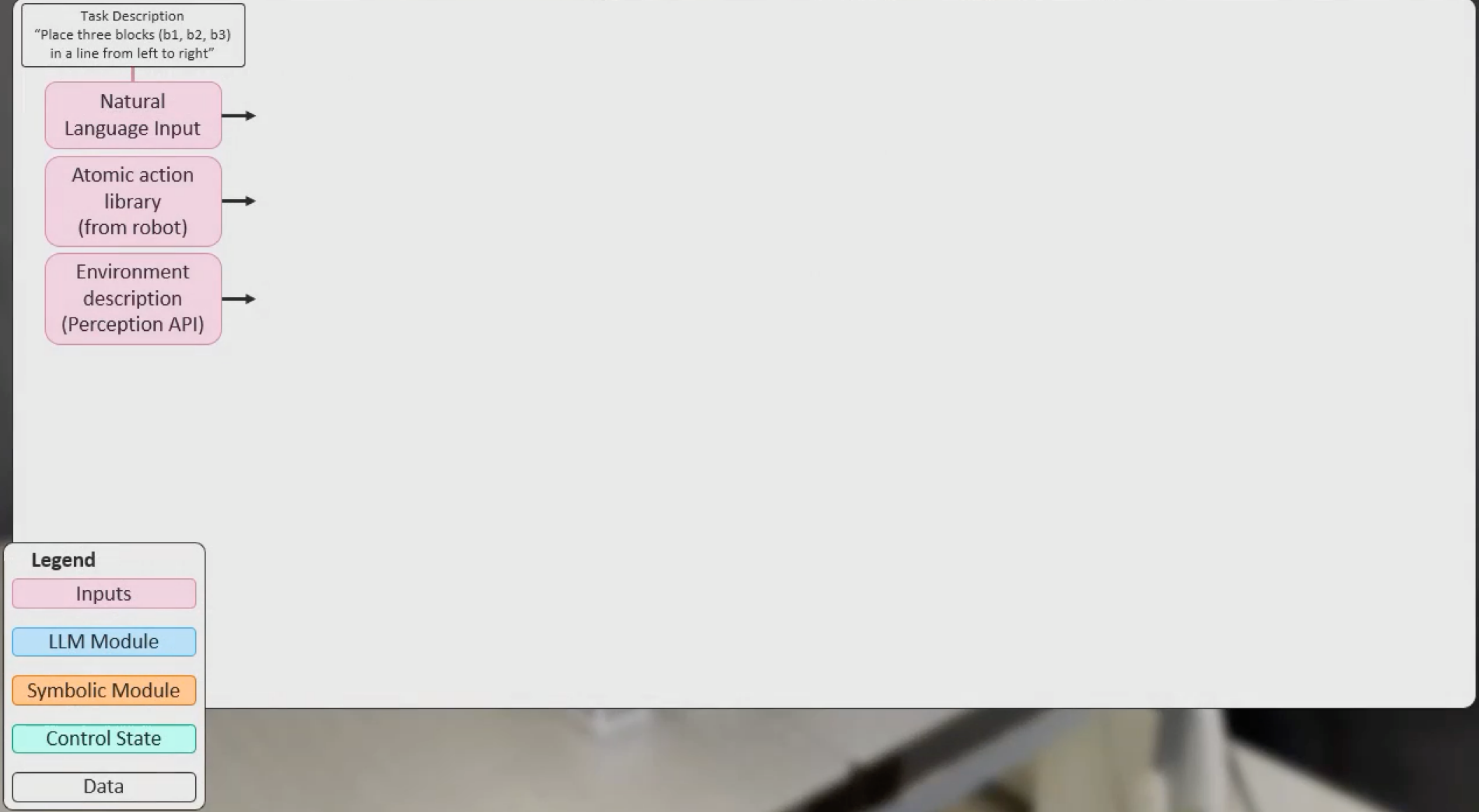


# CLIMB

## Language-Guided Continual Learning for Task Planning with Iterative Model Building

Walker Byrnes[1,2], Miroslav Bogdanovic[3], Avi Balakirsky[4], Stephen Balakirsky[2], Animesh Garg[1,3,5]

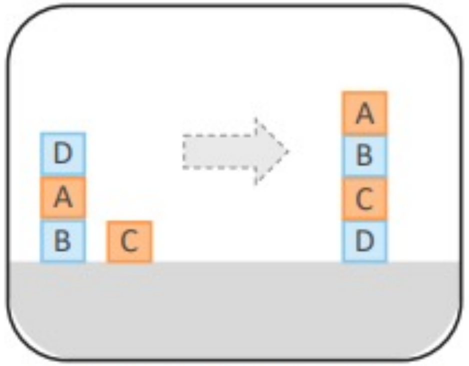
[1]Georgia Institute of Technology, [2]Georgia Tech Research Institute, [3]University of Toronto, [4]Ohio State University, [5]Nvidia


Task Description
“Place three blocks (b1, b2, b3) in a line from left to right”
Natural Language Input
Atomic action library (from robot)
Environment description (Perception API)
Legend
Inputs
LLM Module
Symbolic Module
Control State
Data

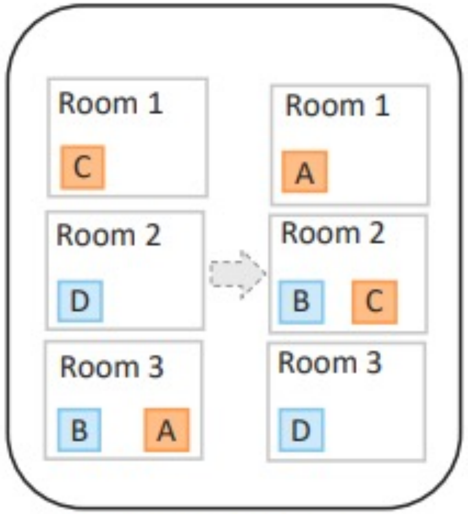
# Experiments

Three levels of fidelity: logical, simulated, and real

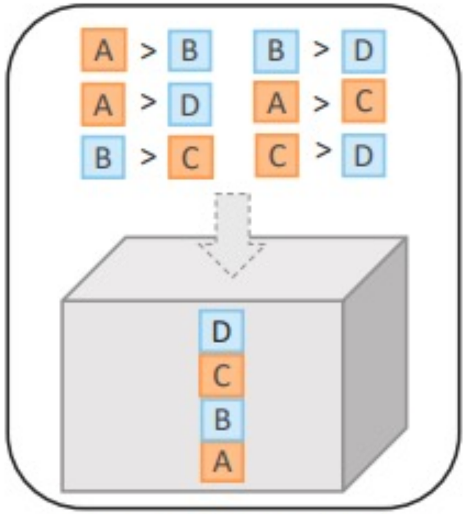
Logical Planning Environments




*Blocksworld* (stacking)

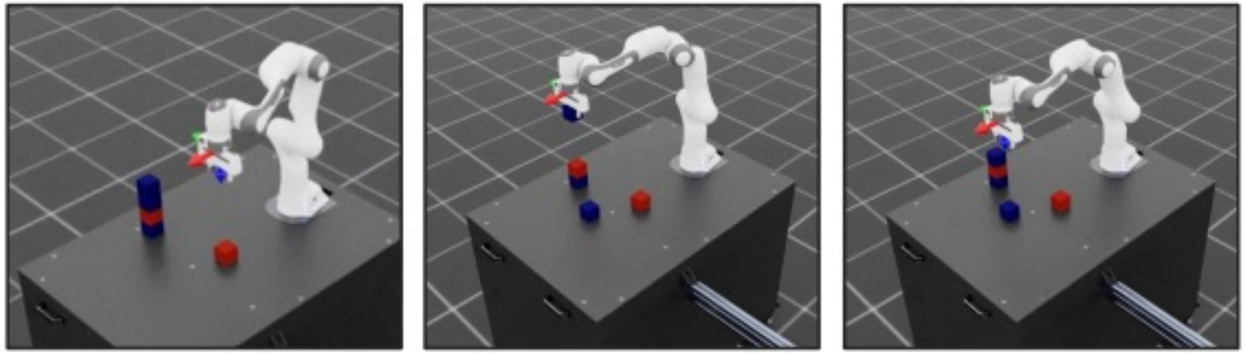



*Grippers* (transport)

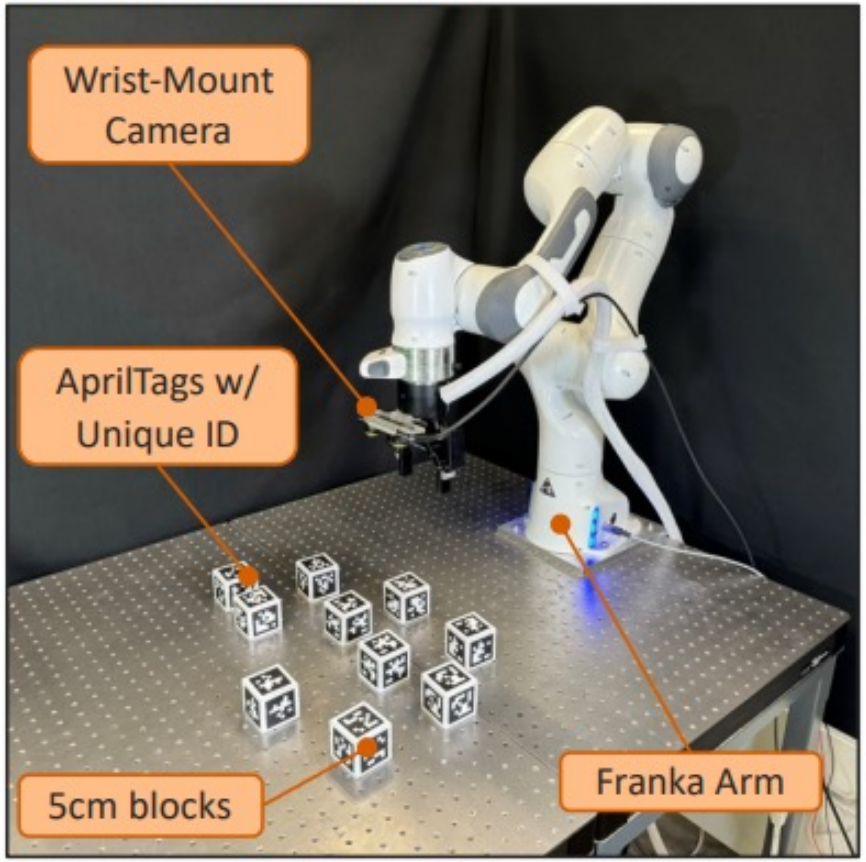



*Heavy* (bin packing)

IsaacLab Blocks Simulation



Real Robot Environment



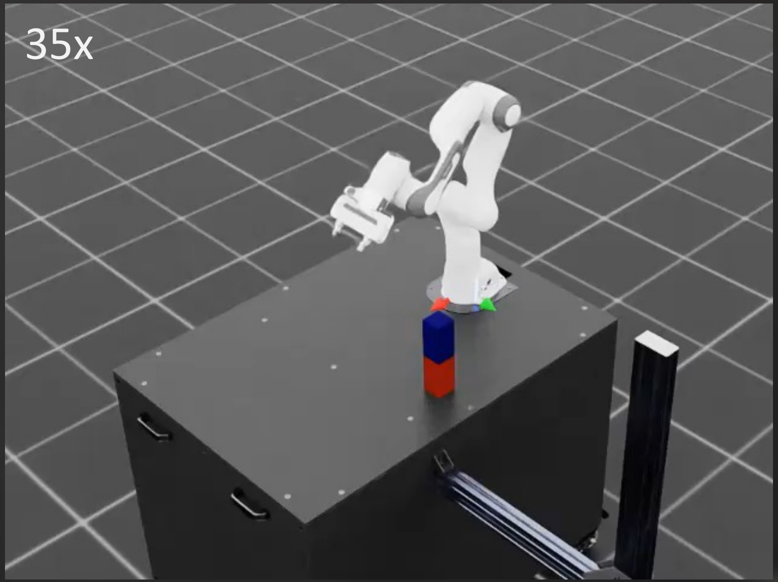
# Results – Logical Domain

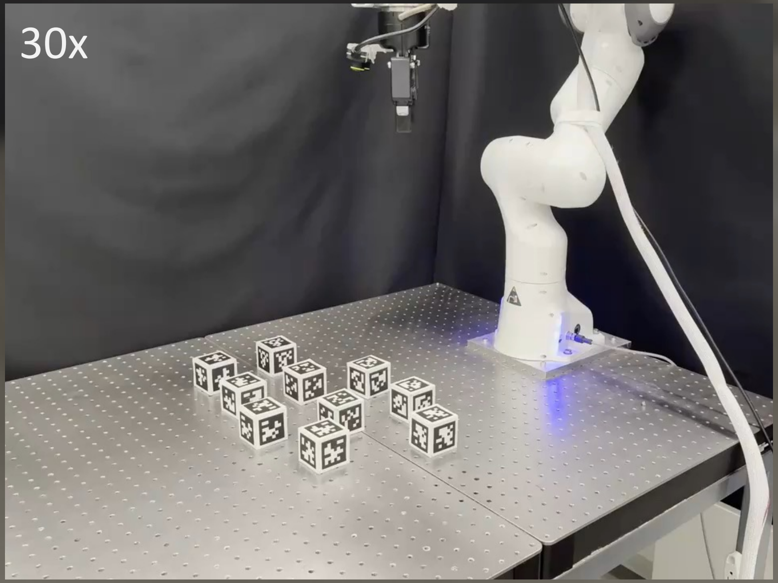
- Compared to naïve LLM prompt based planning, CLIMB's planning structure increases 0-shot performance for some classes of problems
- With few-shot (N=5) re-prompting we can match or improve performance in all evaluated cases
- Reliant on accurate predicate grounding, which can be mitigated with few-shot syntax and semantic corrections

| **Dataset** | LLM Plan | CLIMB 0-Shot | CLIMB Few-Shot |
|---|---|---|---|
| BLW | 0.12<br>(0.05, 0.20) | 0.40<br>(0.28, 0.53) | **0.80**<br>**(0.70, 0.90)** |
| GRP | 0.10<br>(0.03, 0.18) | 0.53<br>(0.42, 0.67) | **0.93**<br>**(0.87, 0.98)** |
| HVY | **0.68**<br>**(0.57, 0.80)** | 0.17<br>(0.08, 0.27) | **0.67**<br>**(0.55, 0.78)** |

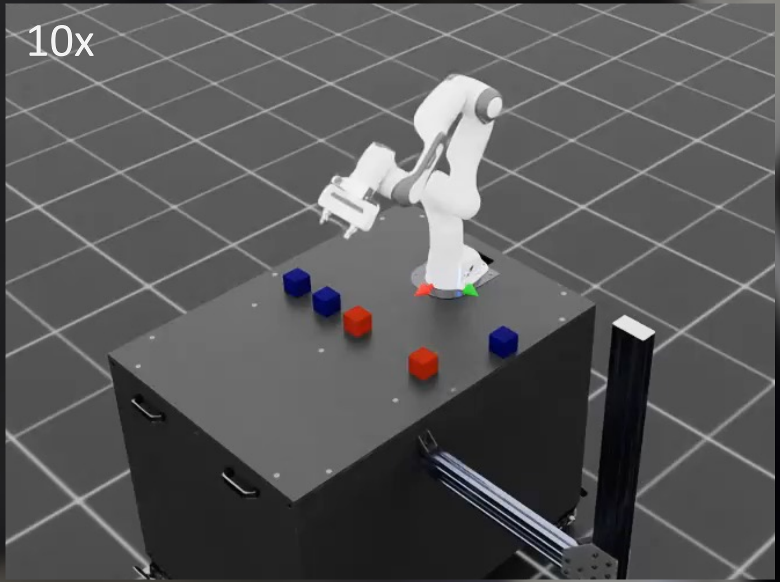
| Predicate | Zero-Shot | With Syntax Fixing |
|---|---|---|
| on-table | 0.95 | **1.00** |
| on | 0.25 | **0.45** |
| holding | 0.50 | **0.65** |
| arm-empty | 0.10 | **0.35** |
| clear | 0.60 | **0.80** |

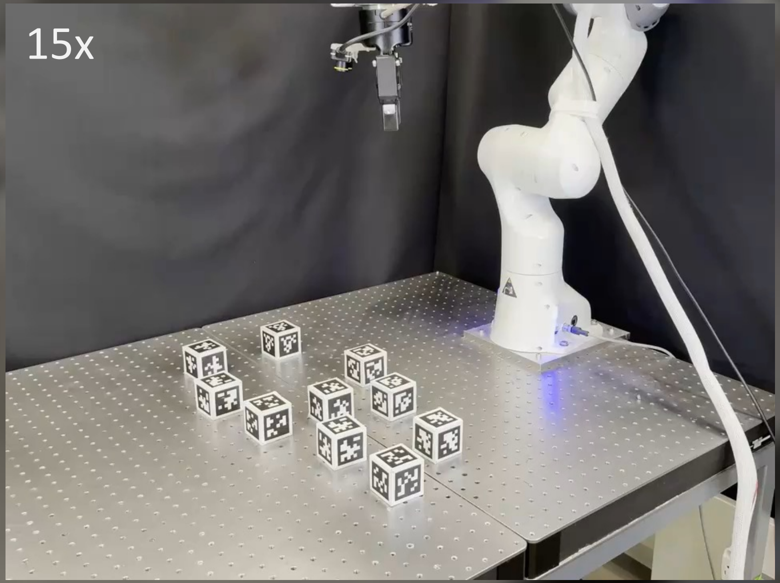
## Basic Stacking



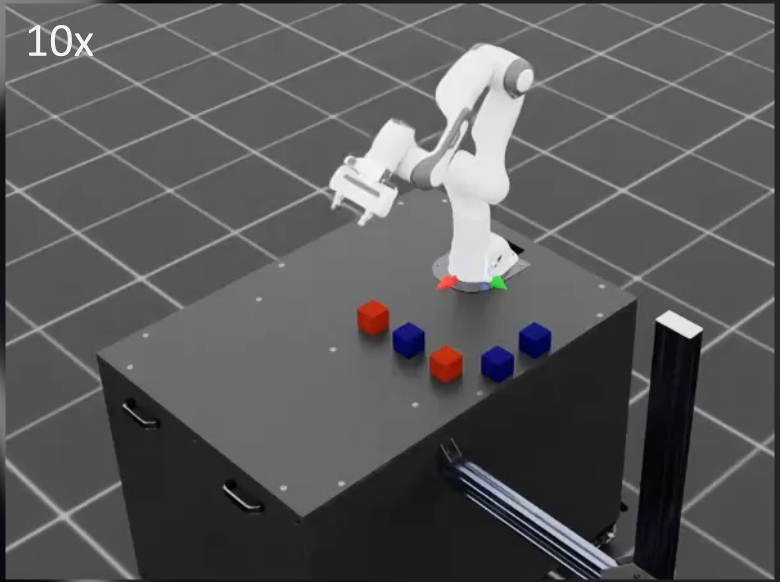



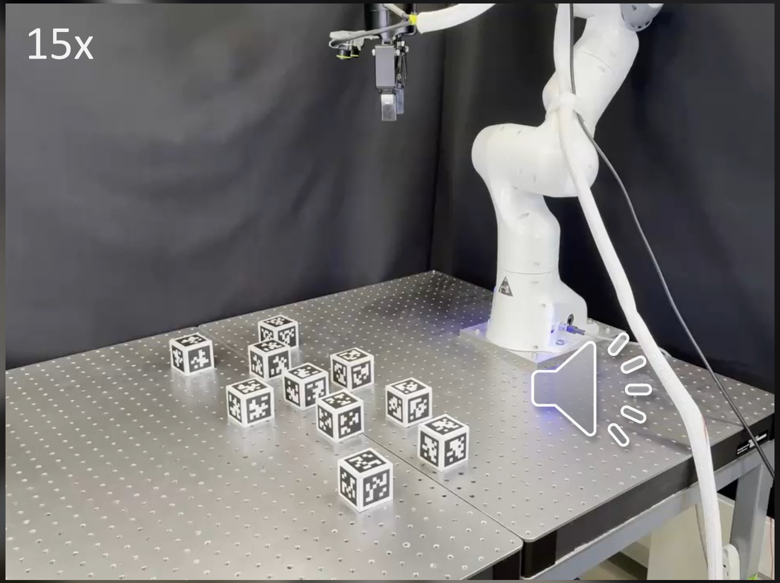

## 2D Arrangement



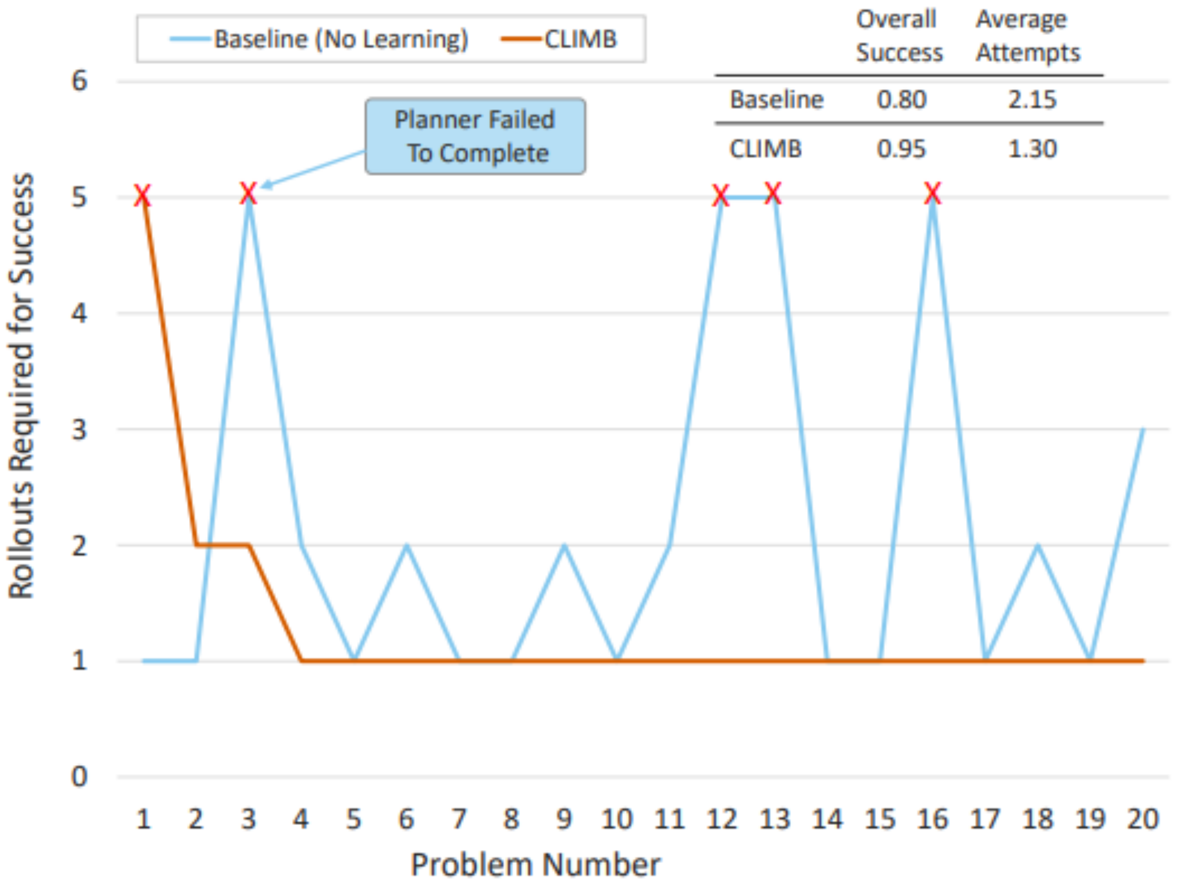

## Pyramid Stacking



# Results

**Logical BlocksWorld Dataset**

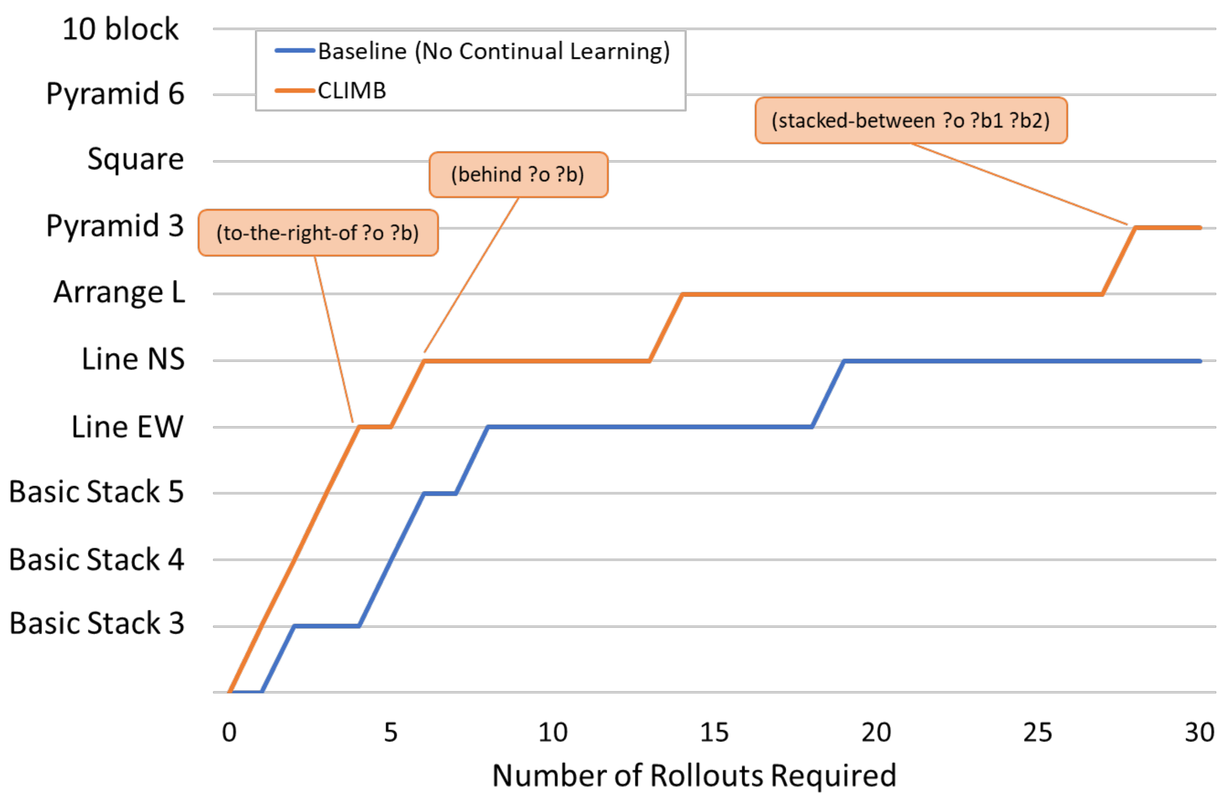



- Leveraging data from past instances improves overall success and reduces total rollouts required

- Once CLIMB obtains a complete domain, it can solve new problems zero-shot

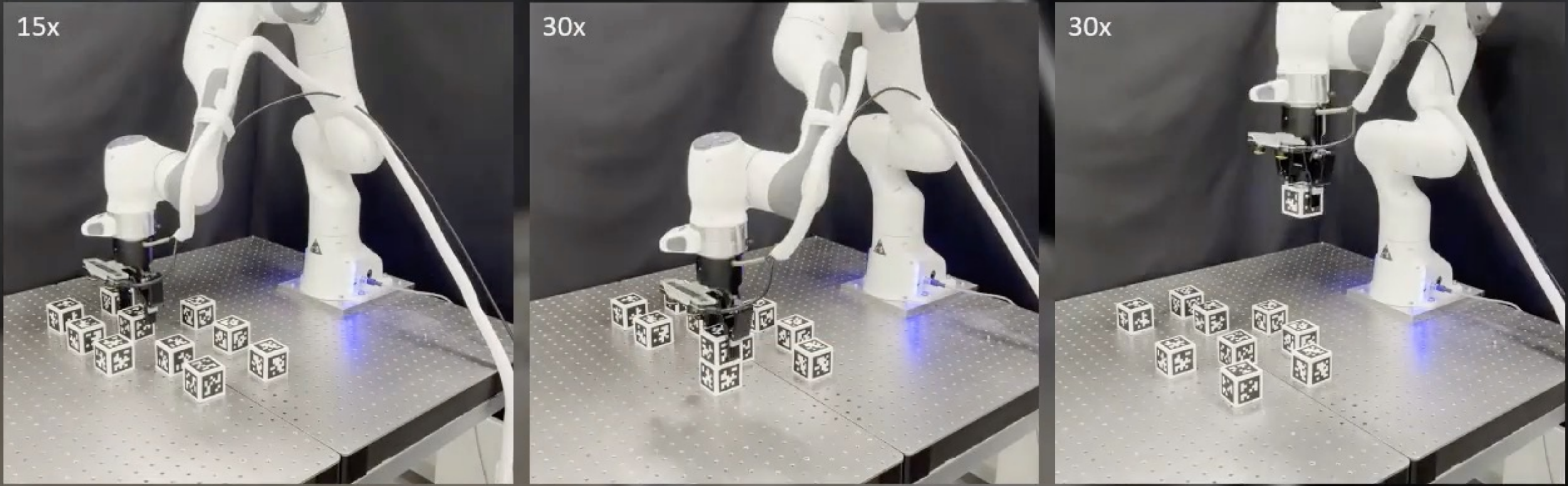
# Results

- Evaluation on curriculum of increasing complexity tasks
- CLIMB demonstrates understanding and incorporation of new world constraints and predicates with fewer rollouts than baseline



Basic Stack 3 | Basic Stack 4 | Basic Stack 6 | Line EW | Line NS | Arrange L | Pyramid 3

**Note:** All basic stack problems were successful on first rollout.



# The Computing Stack

## Physical AI

### Large Behavior Models

Large Scale Imitation Learning

Learned Task Planning and replanning Behavior

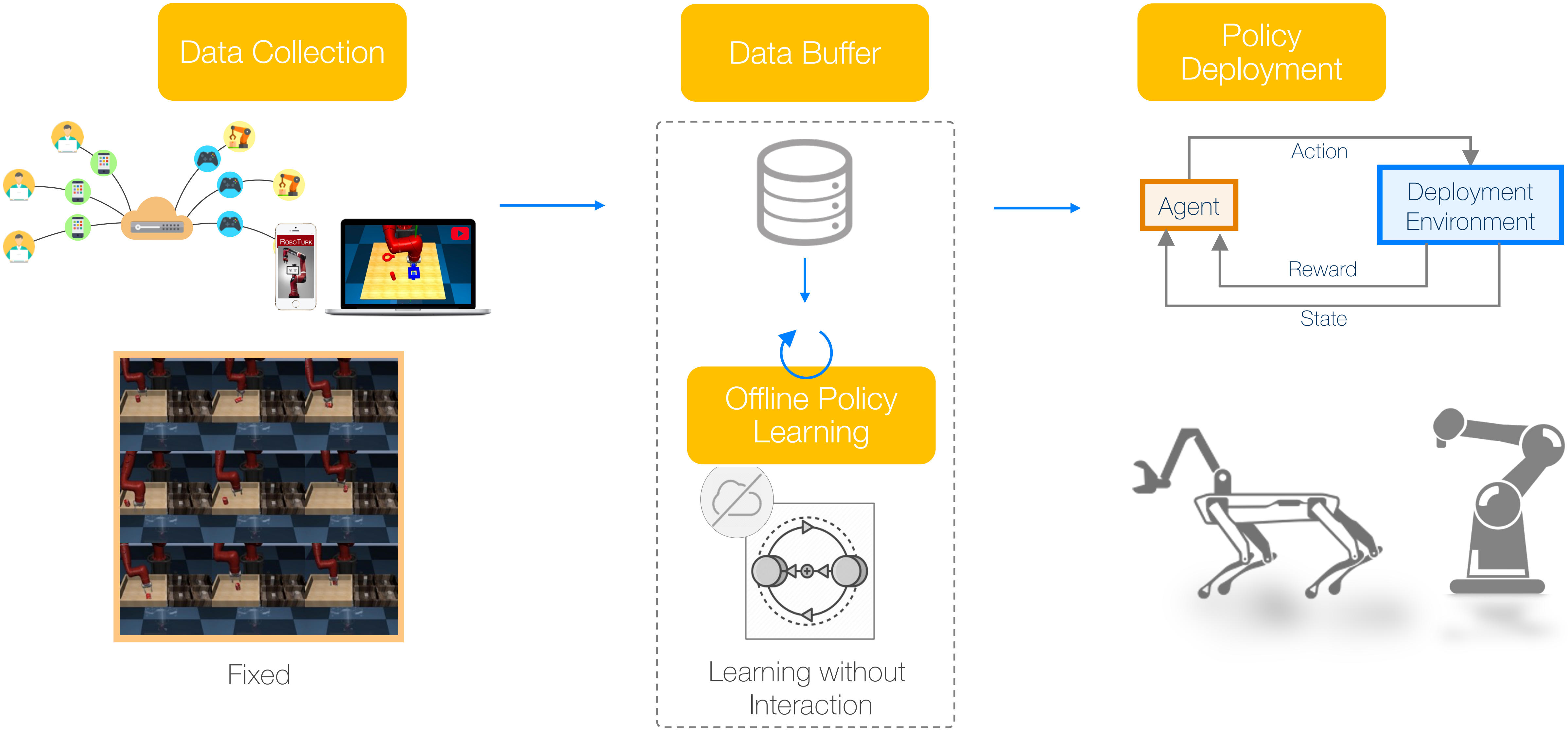
Fine-Tune Generalists for better Specialists using RL

**Natural Interaction Interface**
Ease of Use
App

**Platform-Agnostic Planning**
Modular Tool-Use
OS

**Hardware-Specific Skills**
Optimize for Morphology
Driver

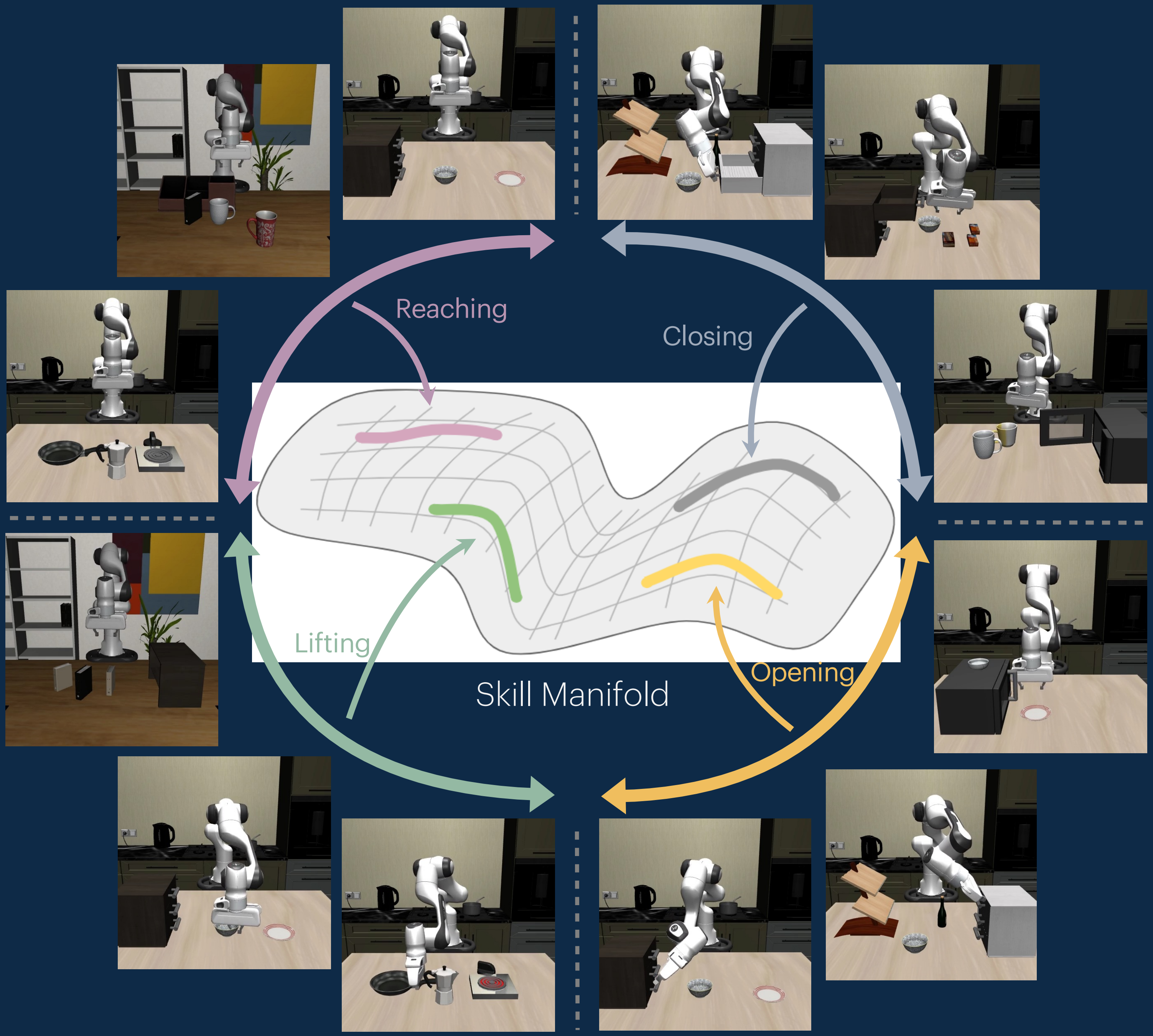
# Policy Learning from Offline Datasets



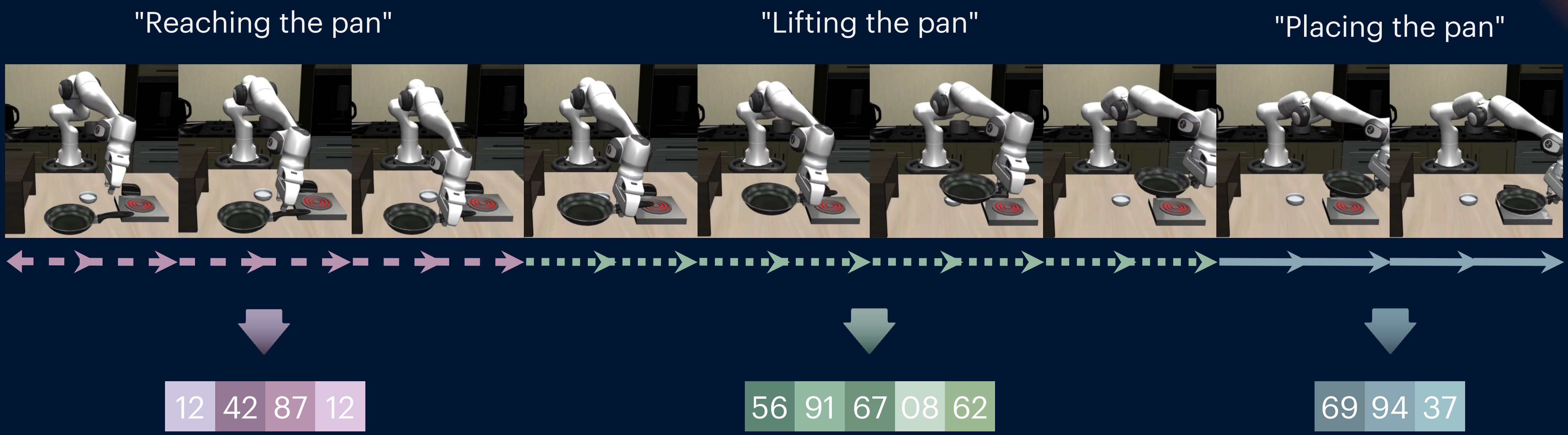
# QUEST

## Self-Supervised Skill Abstractions for Learning Continuous Control

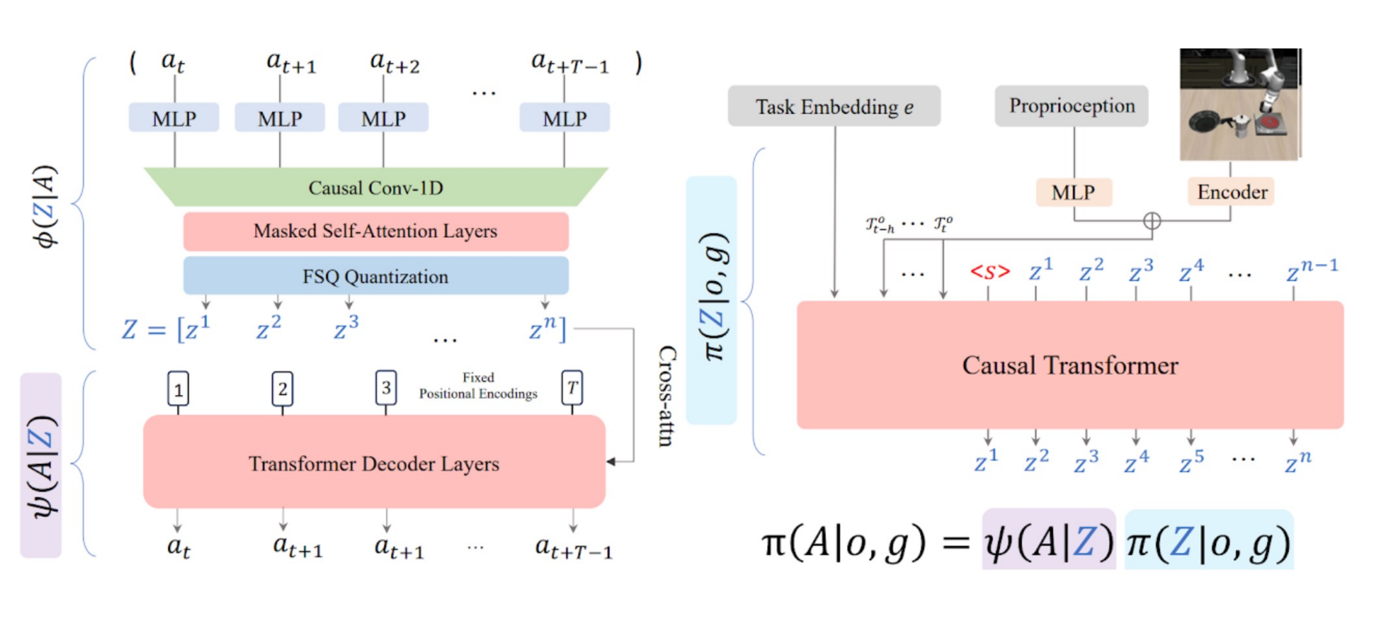
Atharva Mete, Haotian Xue, Albert Wilcox, Yongxin Chen, Animesh Garg


Reaching
Closing
Lifting
Opening
Skill Manifold

# Multi-task Learning: QUEST

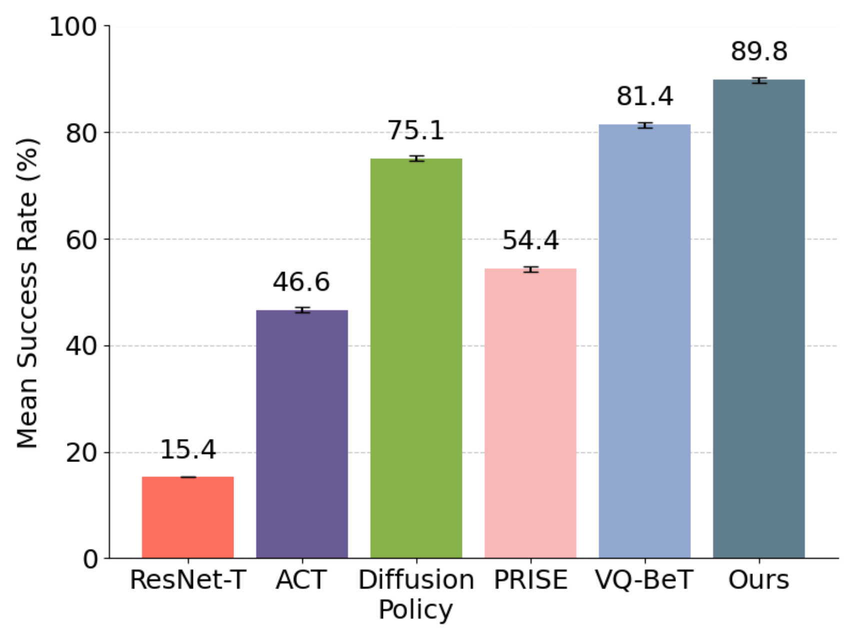


# Multi-Task Behavior Cloning
## Latent Variable Models



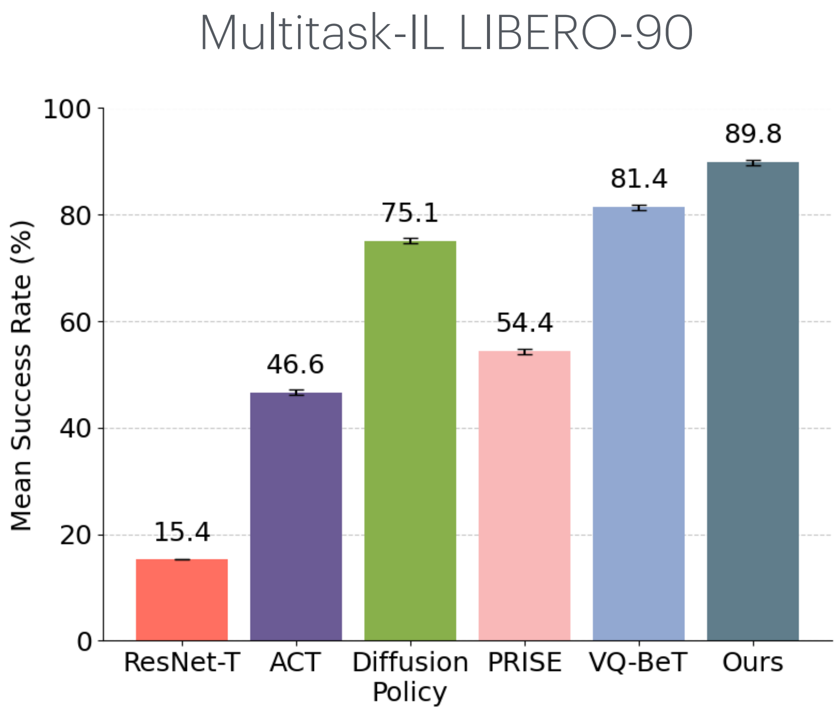
# Multi-task Learning: QUEST

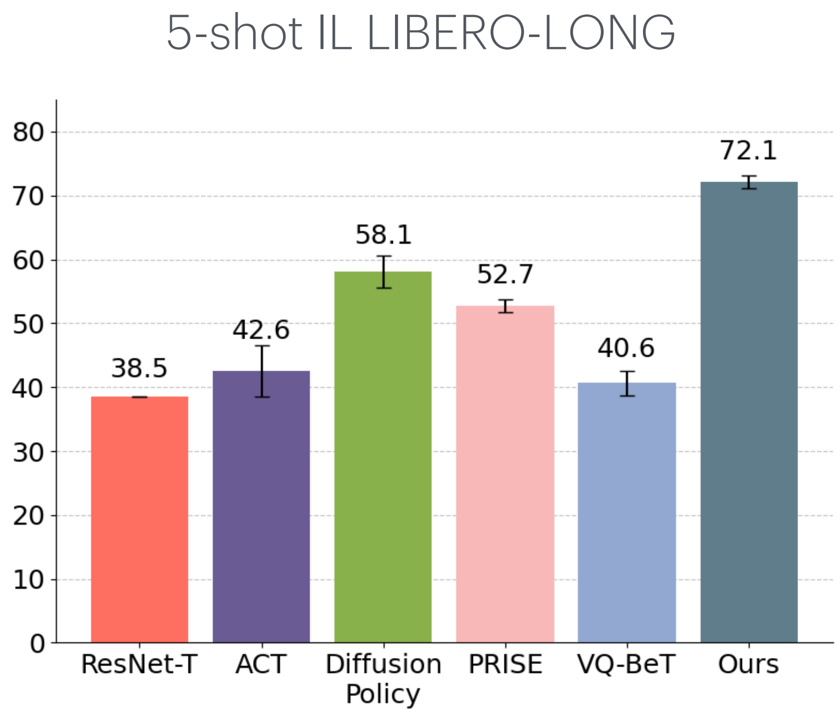
Multitask-IL LIBERO-90




**Multitask IL:** Relative improvement of **10.3%** over next best baseline

# Multi-Task Behavior Cloning
## Latent Variable Models




**Multitask IL:** Relative improvement of **10.3%** over next best baseline

**5-shot IL:** Relative improvement of **24%** over next best baseline

# Adapt3R

## Unified 3D Scene Representation for Domain Transfer in Imitation Learning

Albert Wilcox, Mohamed Ghanem, Masoud Moghani

# Motivation

- Robotics data is notoriously expensive,
- Collecting enough data to cover the full space of robotics deployment settings (all variations in scene, robot, etc) is proving to challenging

Generalization is often bounded
Change of robot embodiment
Change of camera pose

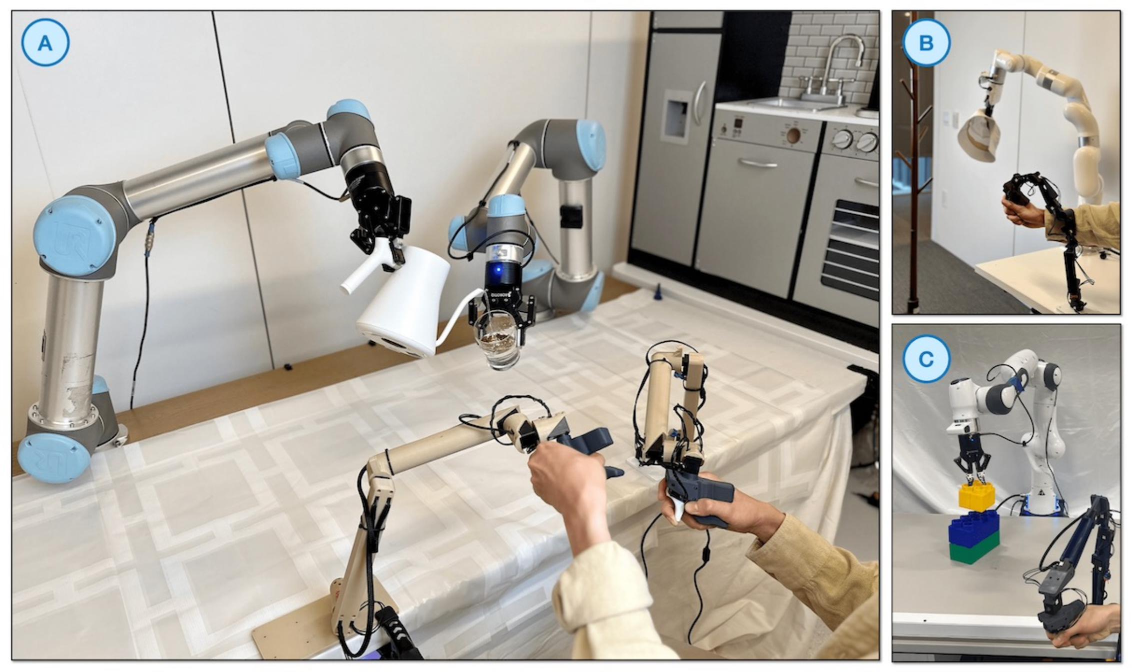
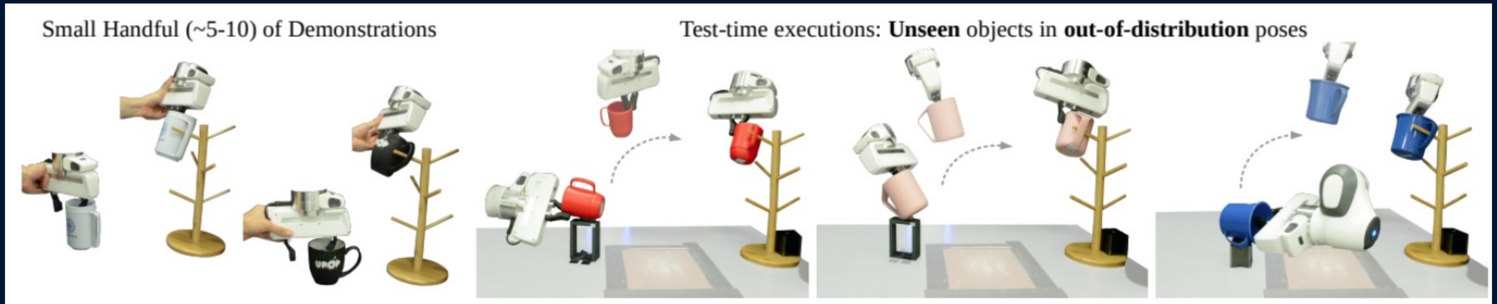
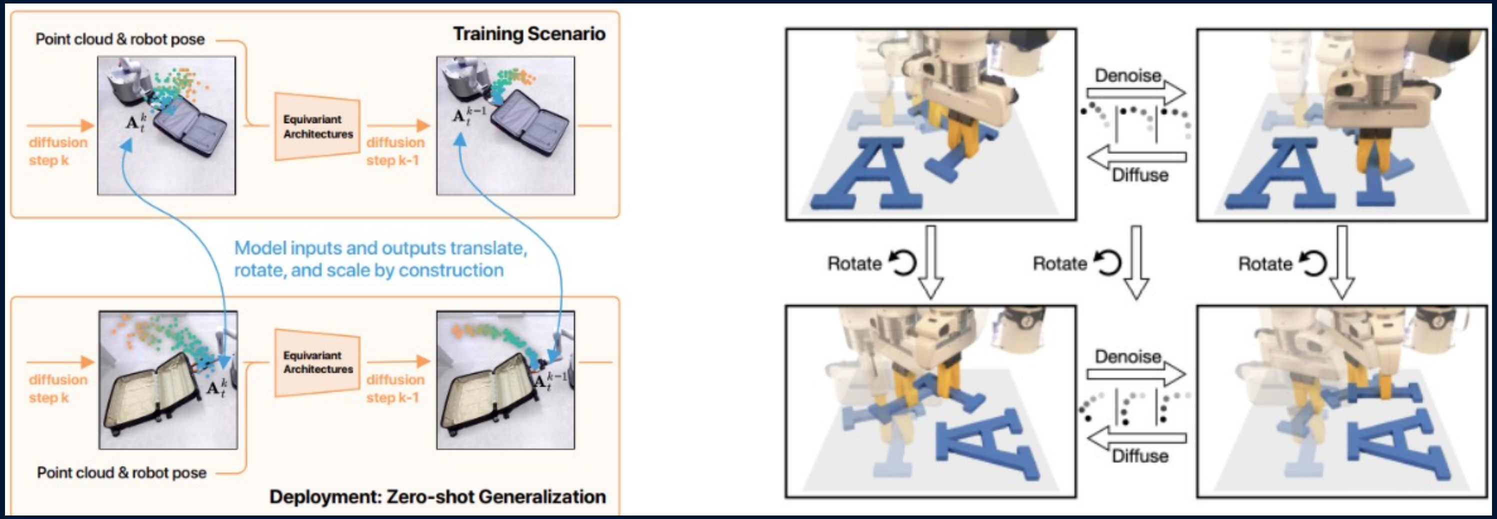



Image from GELLO Wu et al 2024

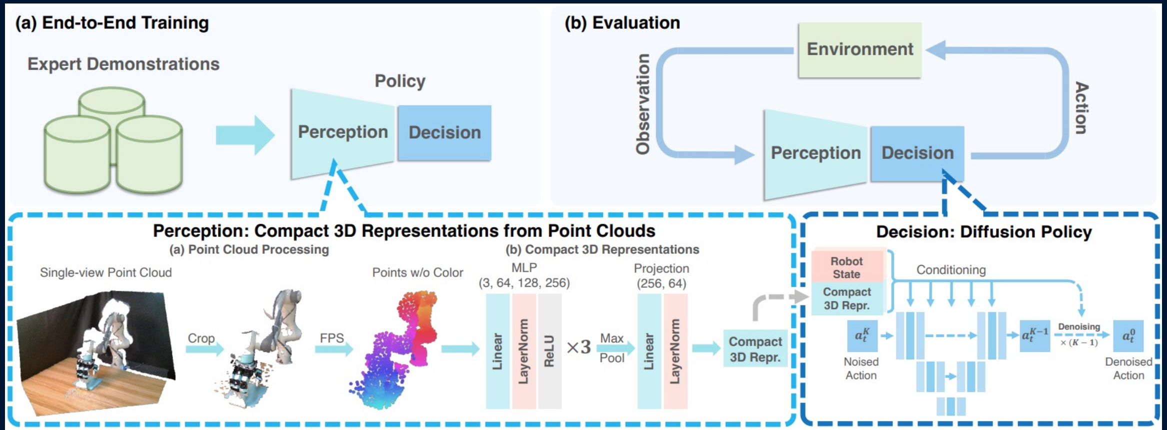
# Prior Work - SE(3) Equivariance

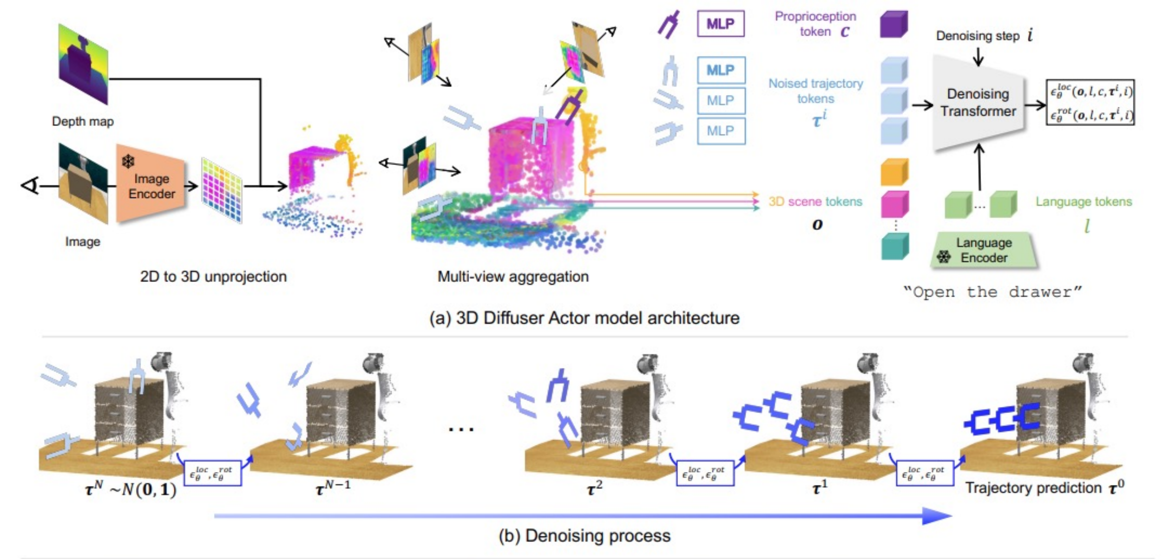



- SE(3) equivariance (pink and place): NDFs, TaxPose, KeyPoint-ViL, RiEMann, etc
- EquivAct, EquiBot / Equivariant diffusion policy
- Difficult to scale to settings with several objects
- Completely incompatible with modern BC methods

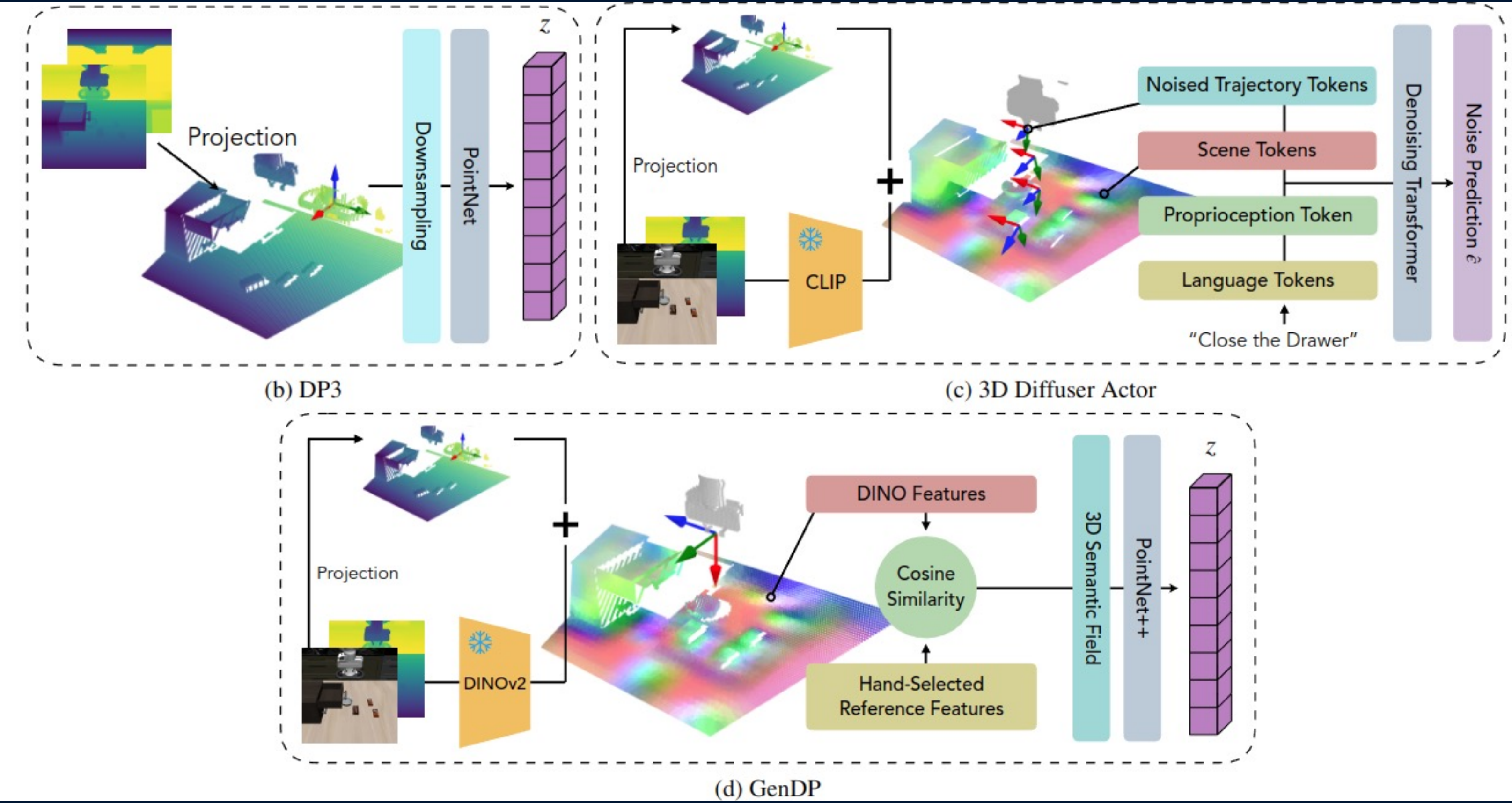
# Prior Work - 3D Diffusion Policy / iDP3




- 3D Diffusion Policy (DP3) and Improved 3D Diffusion Policy (iDP3) use colorless point clouds as a scene representation
- 3D Diffuser Actor lifts CLIP
- Omit semantic information or too slow to train and test
- Generalization to new camera poses is limited

# Challenges

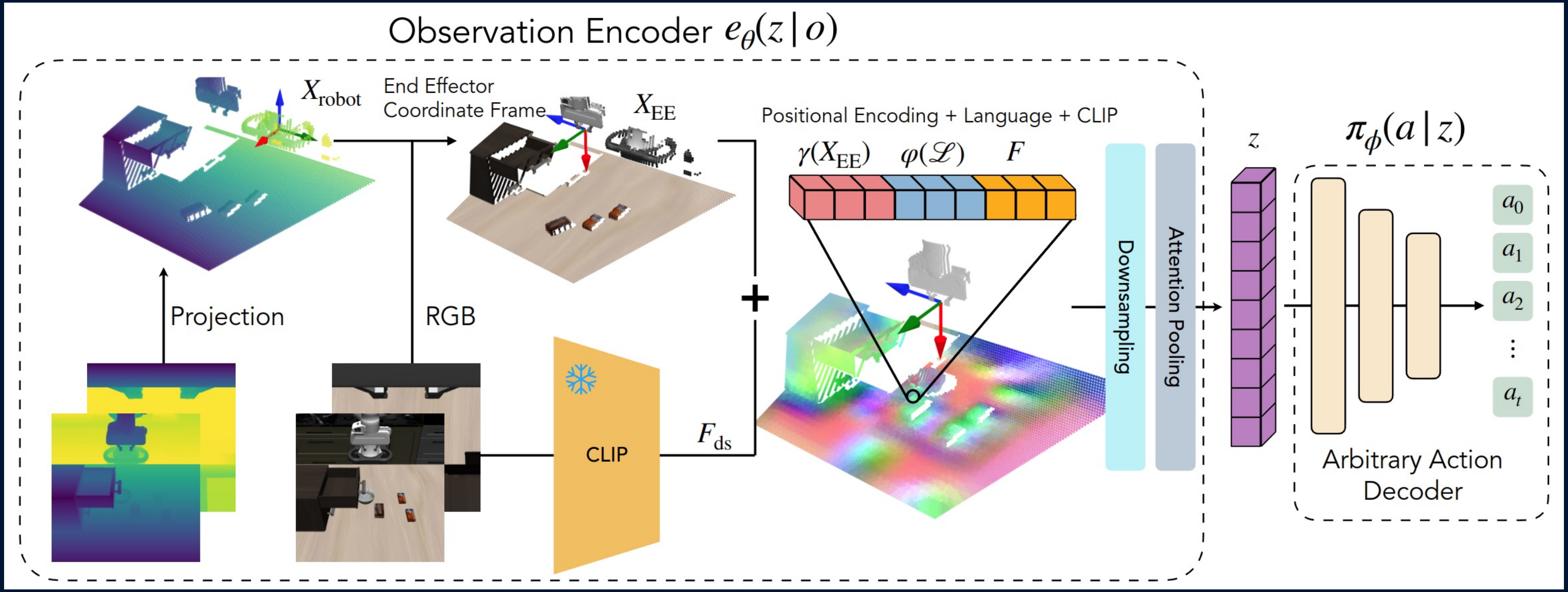



(b) DP3 (c) 3D Diffuser Actor

(d) GenDP

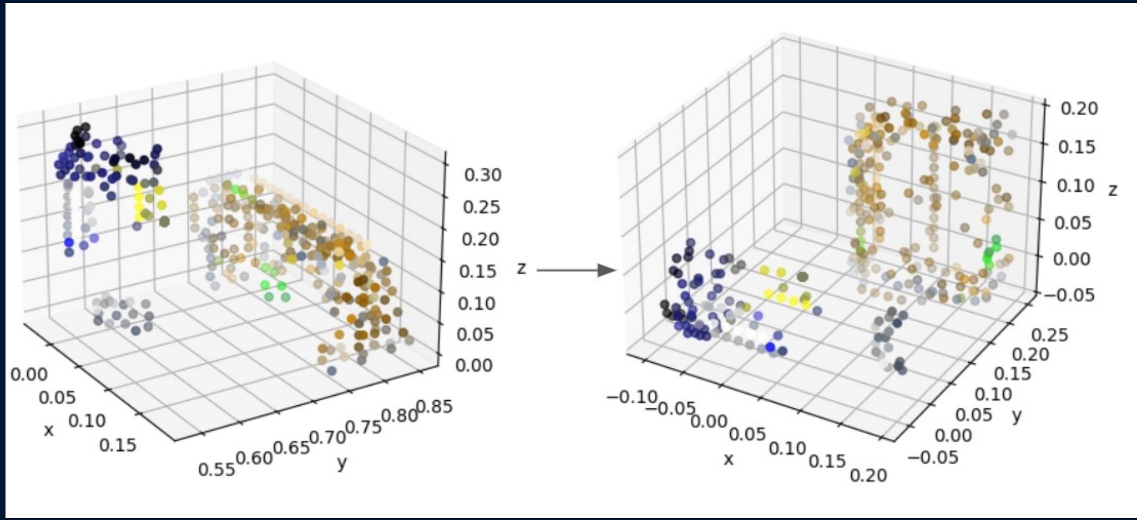
- DP3 uses a colorless point cloud and PointNet
- 3D diffuser actor cross attends between scene tokens and noised trajectory tokens
- GenDP requires hand-selected reference features

# Adapt3R Overview




- Lift CLIP encodings into a 3D semantic point cloud
- Use attention pooling over the cloud to extract a single conditioning vector
- Use that vector as input for an arbitrary policy and train end-to-end

# Adapt3R Overview




**1. Convert to End-Effector Frame**
Transform the point cloud from the robot base coordinate frame to the end effector's coordinate frame

**2. Object Centric Foreground cropping**
Remove points far behind the end effector

**3. CLIP-based down-sampling**
instead of Point Cloud based sub-sampling

**4. Positional embedding for PC**
Transform the point cloud from from XYZ to sinusoidal positional encoding

# Experiments - Multitask IL

| Algorithm | LIBERO-90 | MetaWorld |
|---|---|---|
| ACT + RGB | 0.912 | 0.697 |
| ACT + RGBD | 0.742 | 0.655 |
| ACT + DP3 | 0.759 | 0.402 |
| ACT + iDP3 | 0.764 | 0.419 |
| ACT + Adapt3R | **0.916** | **0.876** |
| DP + RGB | **0.907** | 0.605 |
| DP + RGBD | 0.867 | 0.540 |
| DP + DP3 | 0.703 | 0.446 |
| DP + iDP3 | 0.661 | 0.405 |
| 3D Diffuser-Actor | 0.837 | - |
| DP + Adapt3R | 0.899 | **0.866** |
| BAKU + RGB | 0.923 | 0.702 |
| BAKU + RGBD | 0.839 | 0.707 |
| BAKU + DP3 | 0.712 | 0.414 |
| BAKU + iDP3 | 0.742 | 0.460 |
| BAKU + Adapt3R | **0.931** | **0.869** |

**Drop-in replacement for BC methods!**

- Adapt3R achieves similar or better performance compared to baselines in all settings
- Notably, we achieve SOTA results on the LIBERO-90 benchmark

# Experiments
# Robot Change

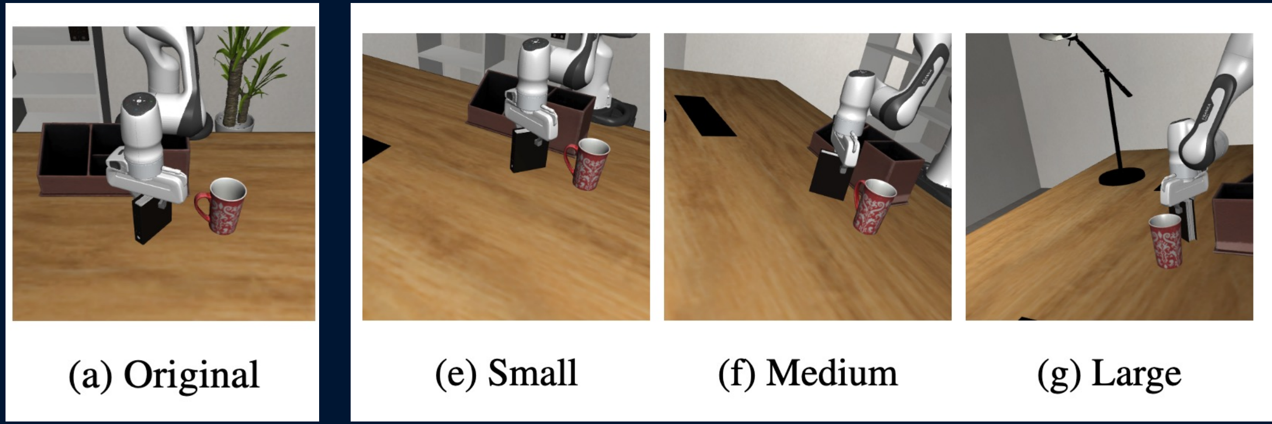
(a) Franka Panda (original) (b) UR5e (c) Kinova3 (d) IIWA

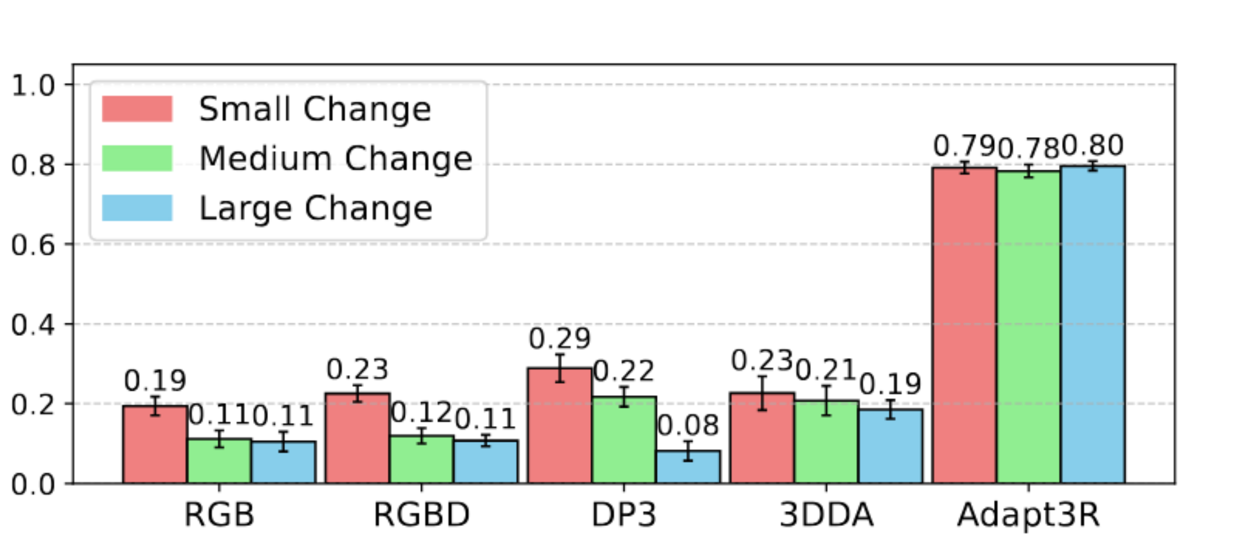
| Algorithm | UR5e | Kinova3 | IIWA |
|---|---|---|---|
| ACT + RGB | 0.594 | 0.667 | 0.557 |
| ACT + RGBD | 0.536 | 0.554 | 0.454 |
| ACT + DP3 | 0.596 | 0.586 | 0.617 |
| ACT + iDP3 | 0.571 | 0.517 | 0.548 |
| ACT + Adapt3R | **0.824** | **0.796** | **0.760** |
| DP + RGB | 0.564 | 0.545 | 0.433 |
| DP + RGBD | 0.478 | 0.457 | 0.361 |
| DP + DP3 | 0.574 | 0.409 | 0.452 |
| DP + iDP3 | 0.499 | 0.377 | 0.411 |
| 3D Diffuser-Actor | 0.737 | **0.766** | **0.636** |
| DP + Adapt3R | **0.761** | 0.568 | 0.522 |
| BAKU + RGB | 0.427 | 0.457 | 0.346 |
| BAKU + RGBD | 0.411 | 0.393 | 0.368 |
| BAKU + DP3 | 0.581 | 0.489 | 0.509 |
| BAKU + iDP3 | 0.541 | 0.422 | 0.490 |
| BAKU + Adapt3R | **0.813** | **0.757** | **0.696** |

- We train only on the Franka Panda
- evaluate zero-shot on
  - UR5e
  - Kinova3
  - Kuka IIWA
- Adapt3R shows a strong improvement compared to most baselines in this experiment

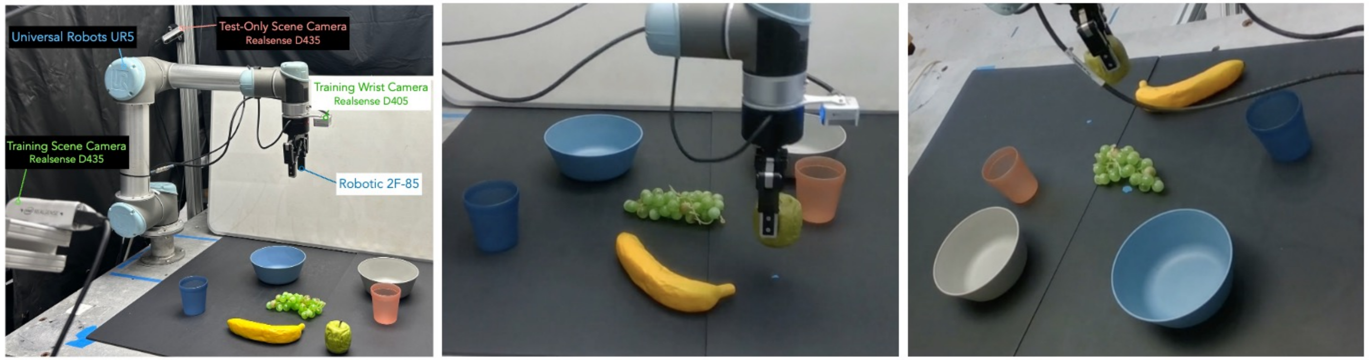
# Experiments
# Camera Pose Change

- We train only on the one camera pose
- evaluate zero-shot on new camera poses
- Adapt3R shows a strong performance preservation across camera poses



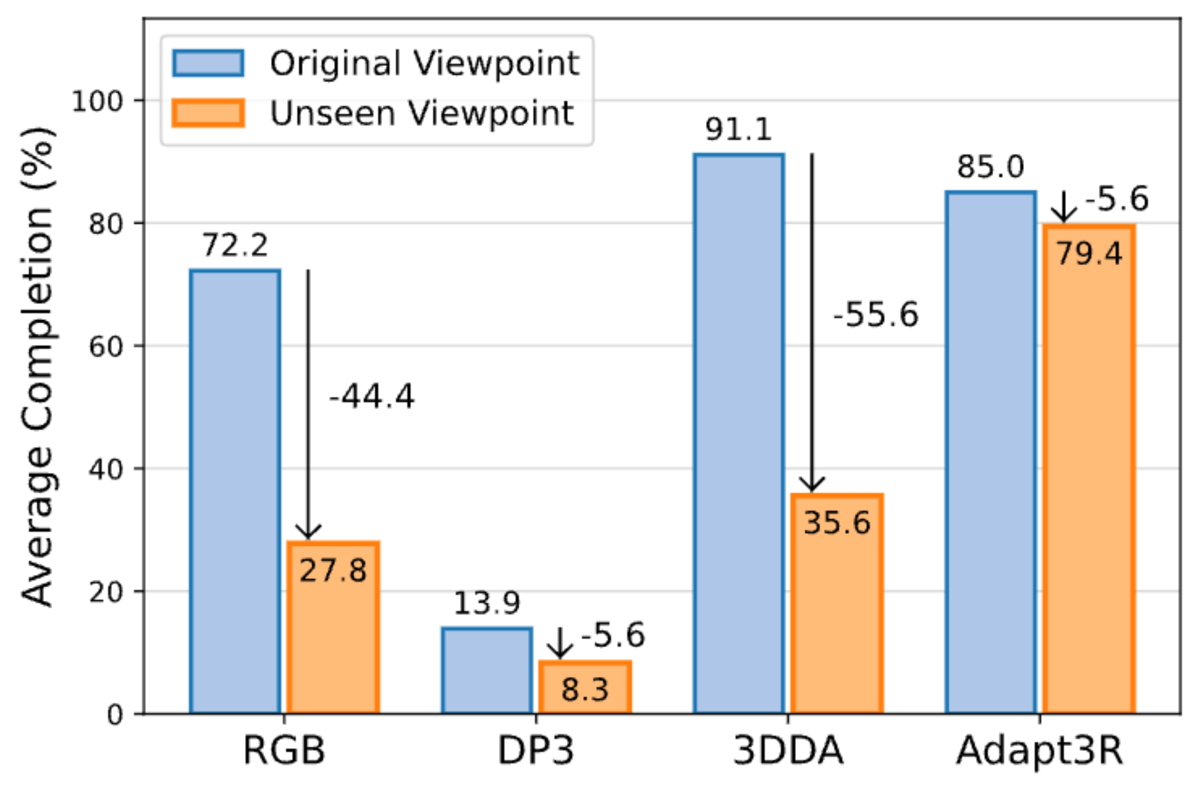



# Experiments
# Camera Pose Change



(a) Real experiment setup (b) Training Viewpoint (c) Zero-Shot Evaluation Viewpoint

- We train only on the UR5
- evaluate zero-shot on new poses
- Adapt3R shows a strong robustness

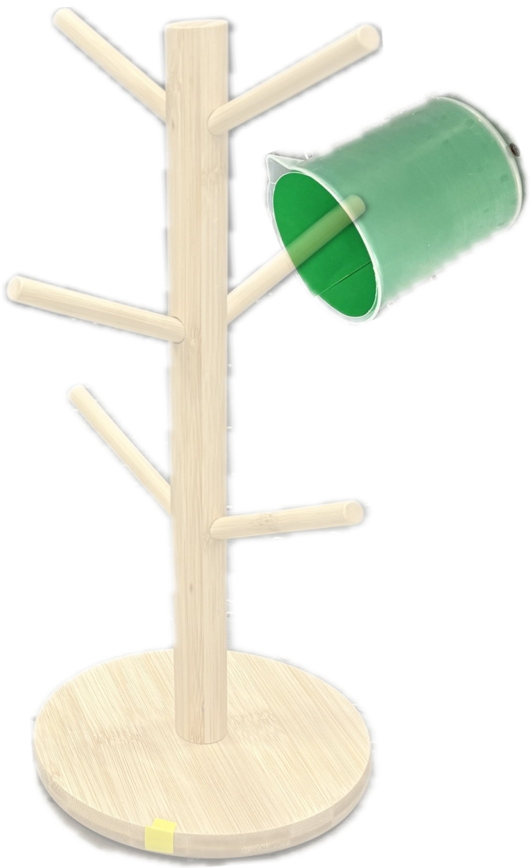


# AnyPlace

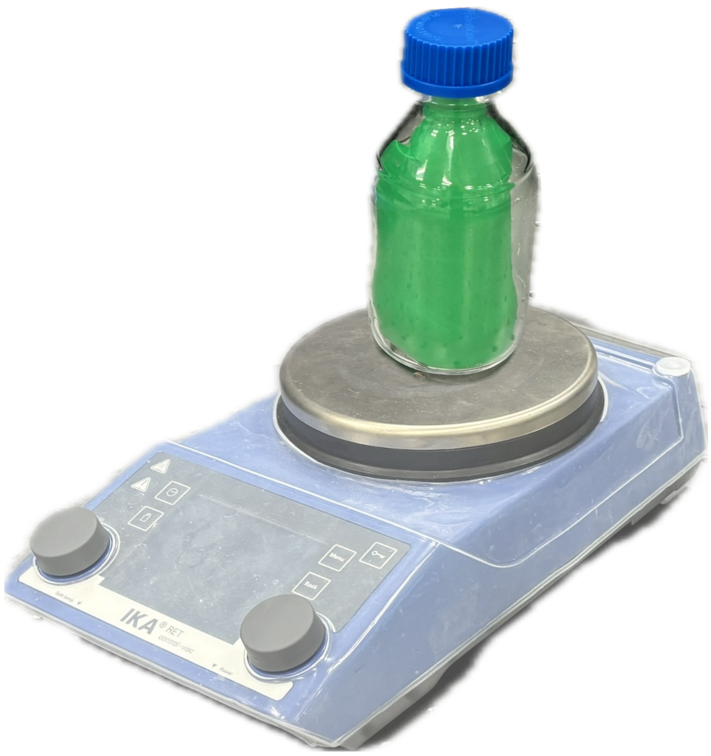
## Learning Generalized Object Placement for Robot Manipulation

Yuchi(Allan) Zhao, Miroslav Bogdanovic, Chengyuan Luo, Steven Tohme, Kourosh Darvish, Alán Aspuru-Guzik, Florian Shkurti, Animesh Garg

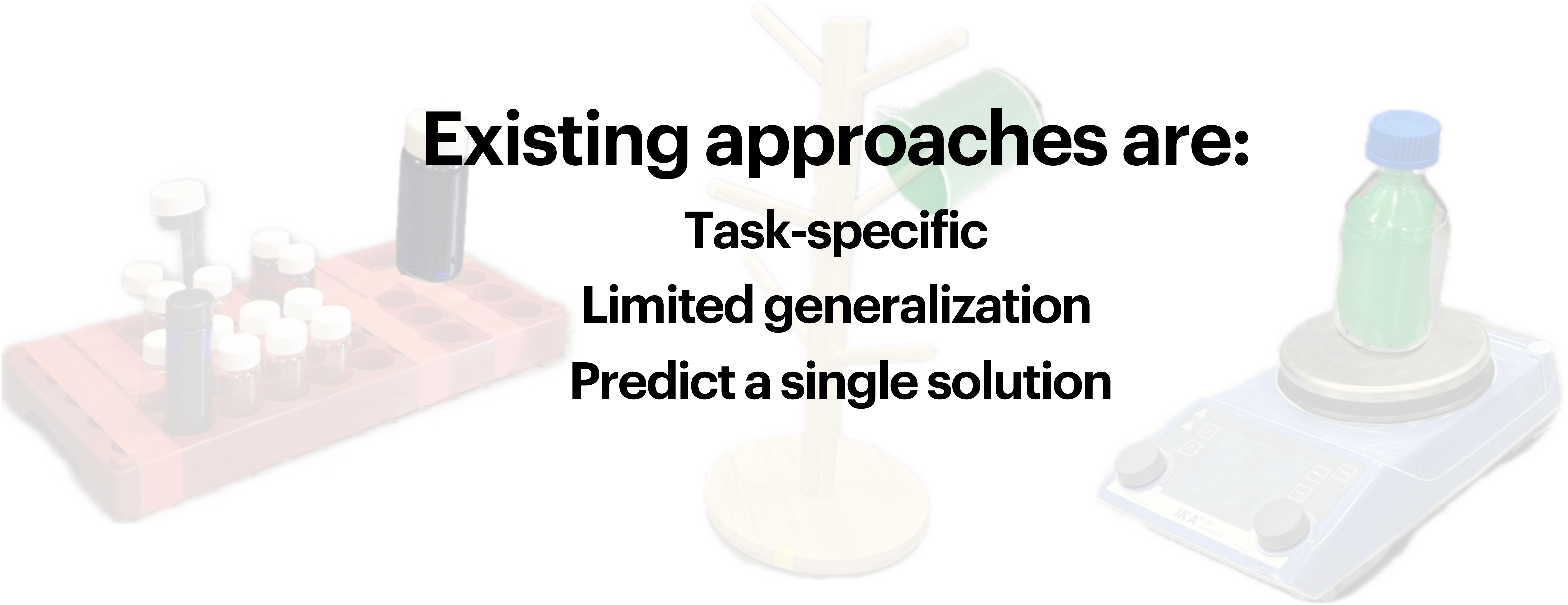
Inserting



Hanging




Stacking



# Existing approaches are:

Task-specific

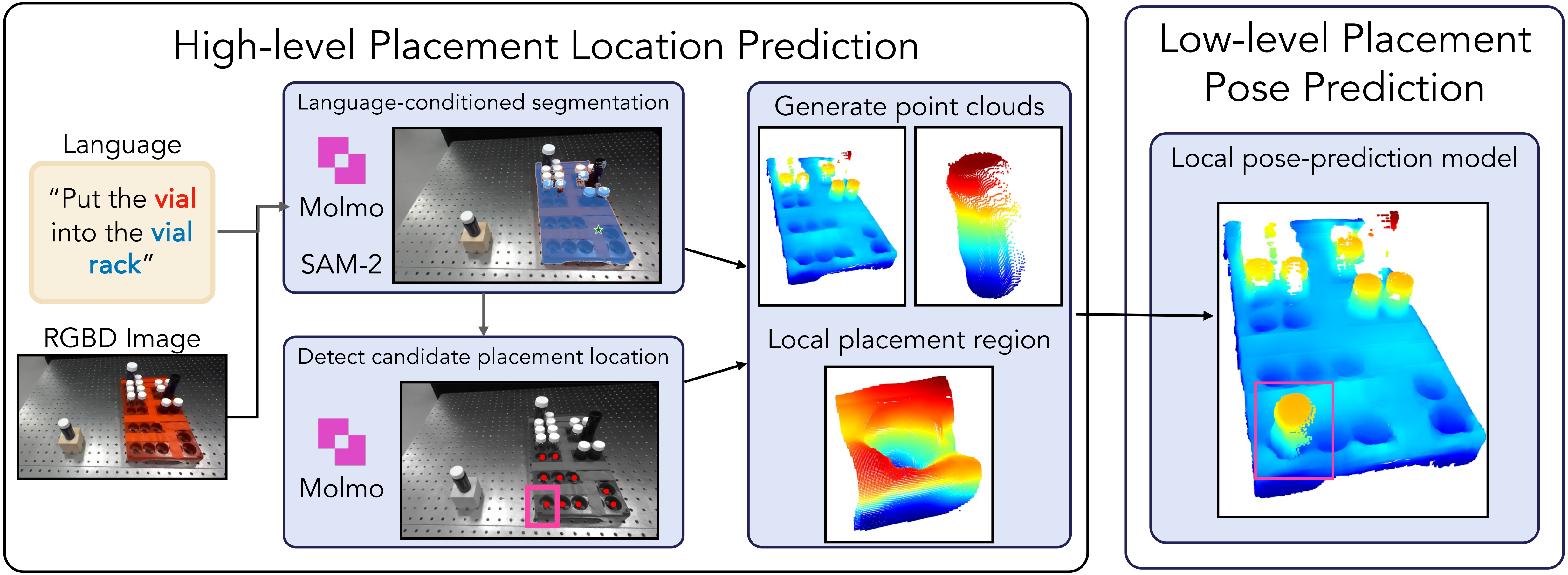
Limited generalization

Predict a single solution

How to enable robots to place objects in a generalizable and robust manner?

AnyPlace

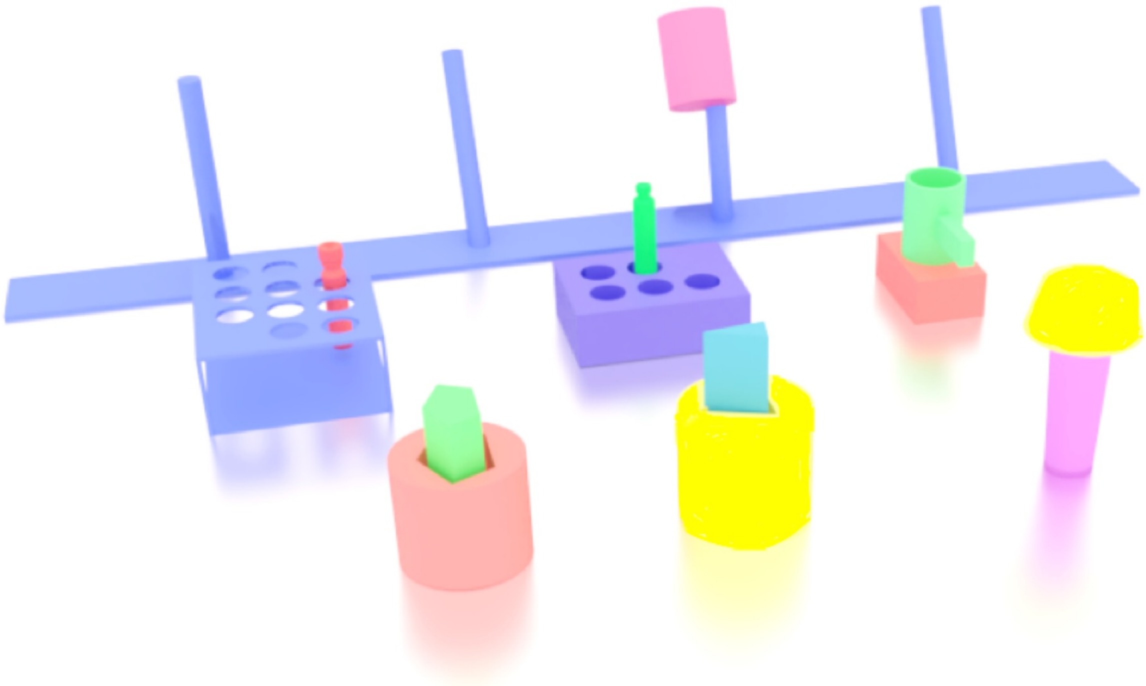
# AnyPlace Architecture

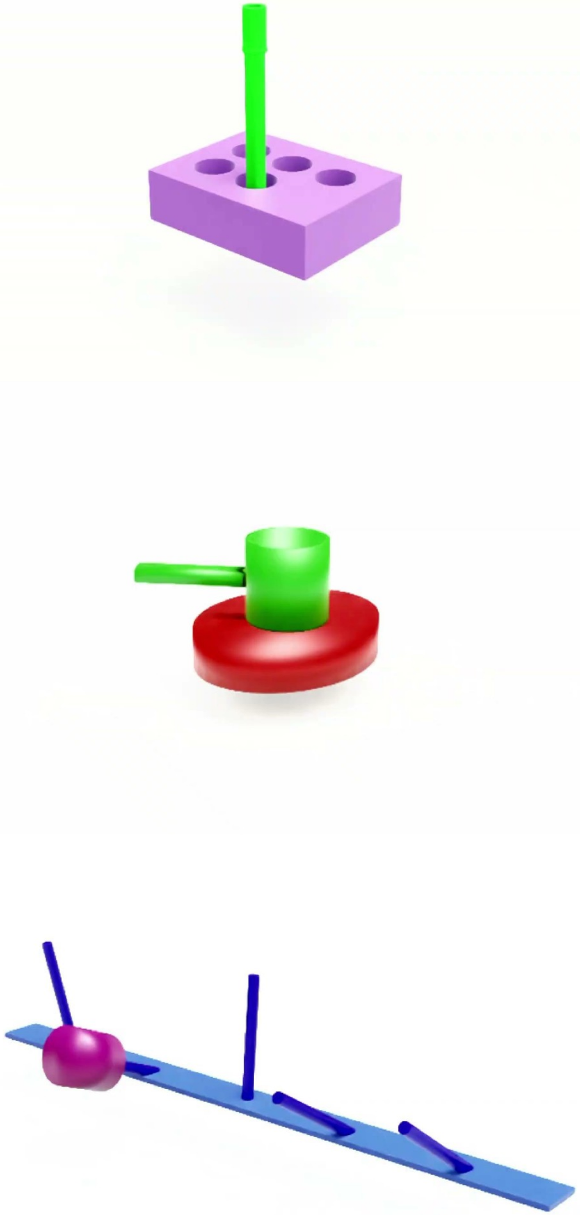



Zhao, Bogdanovic, Luo, Tohme, Darvish, Aspuru-Guzik, Shkurti, Garg. AnyPlace: Learning Generalized Object Placement for Robot Manipulation (2025)

# AnyPlace

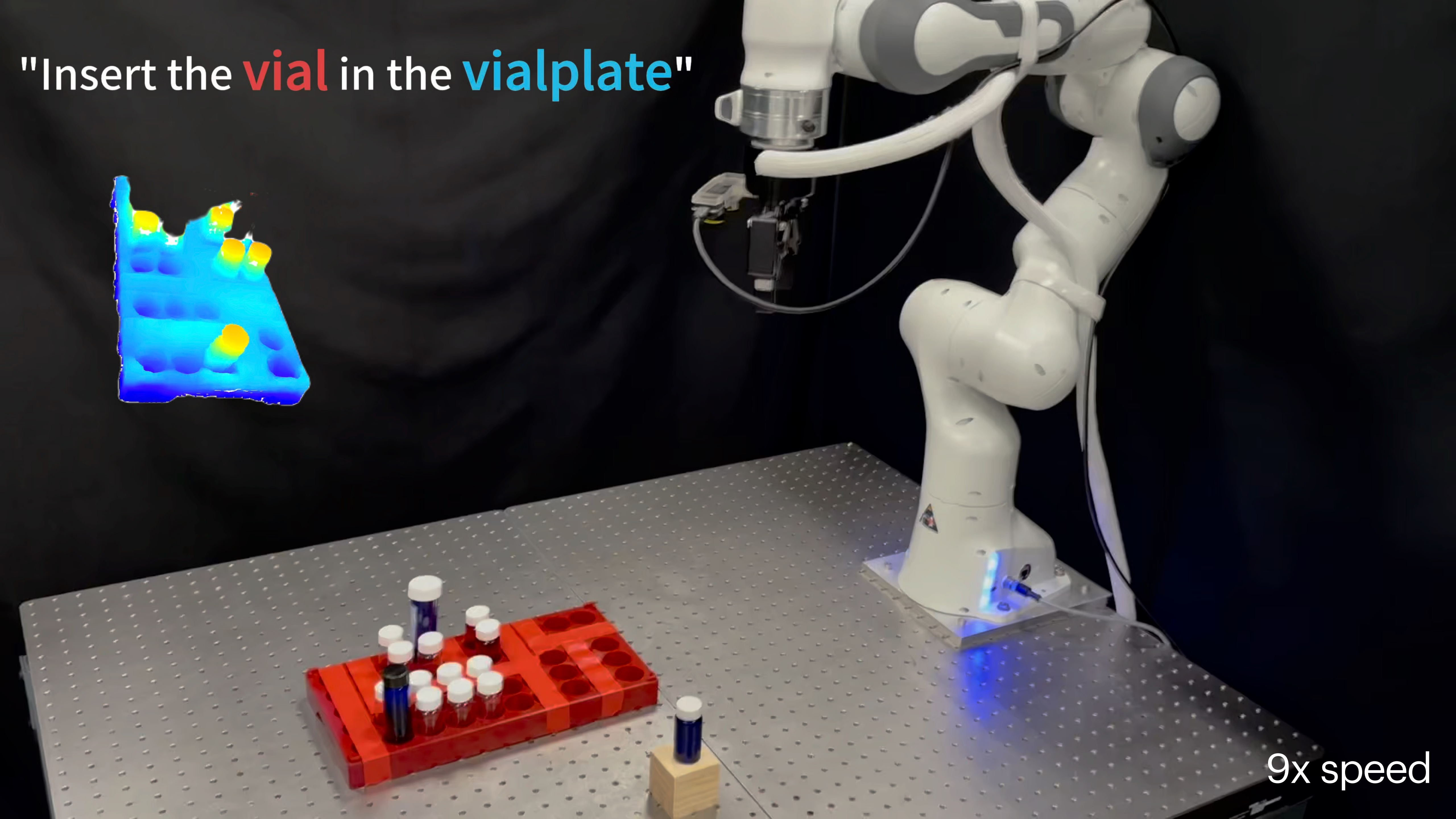
## Synthetic Dataset Generation



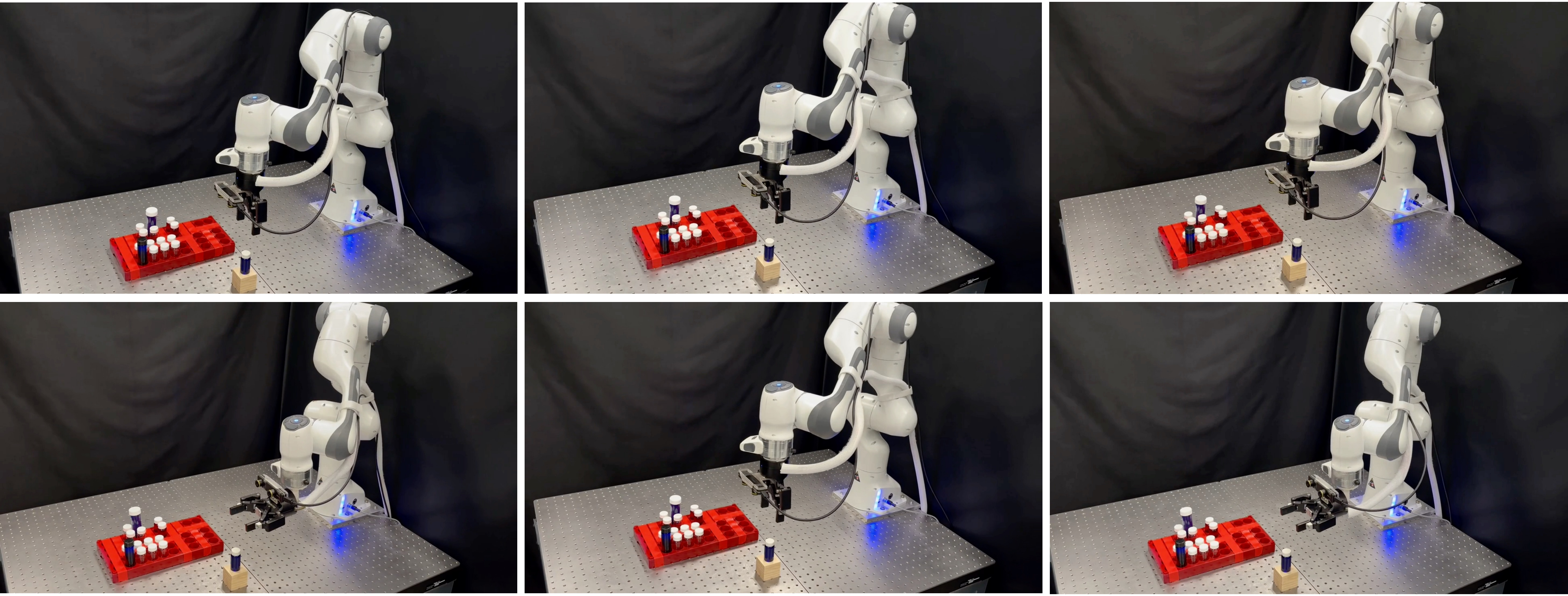


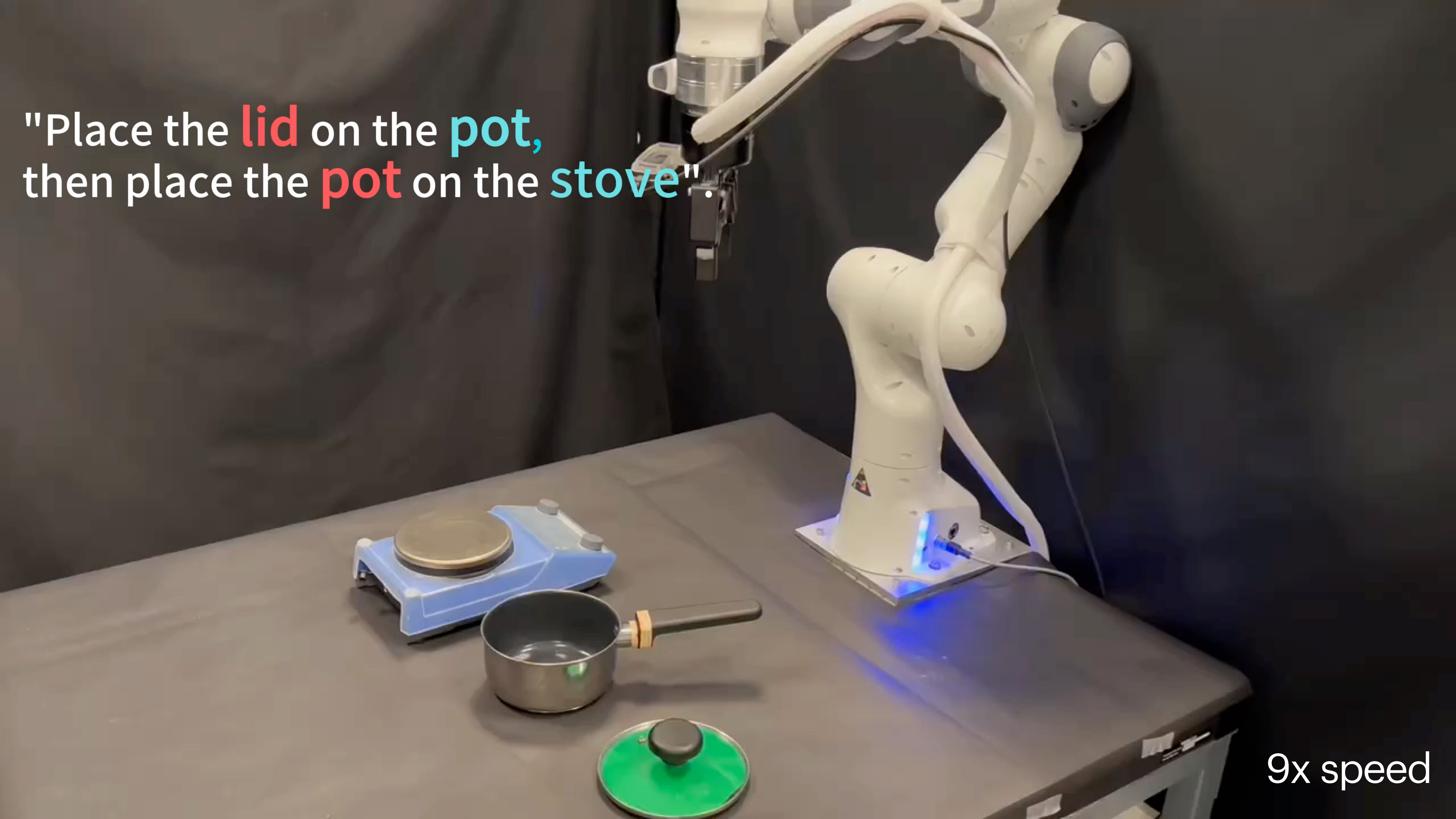
1,489 Objects
5,370 Placement

Zhao, Bogdanovic, Luo, Tohme, Darvish, Aspuru-Guzik, Shkurti, Garg. AnyPlace: Learning Generalized Object Placement for Robot Manipulation (2025)

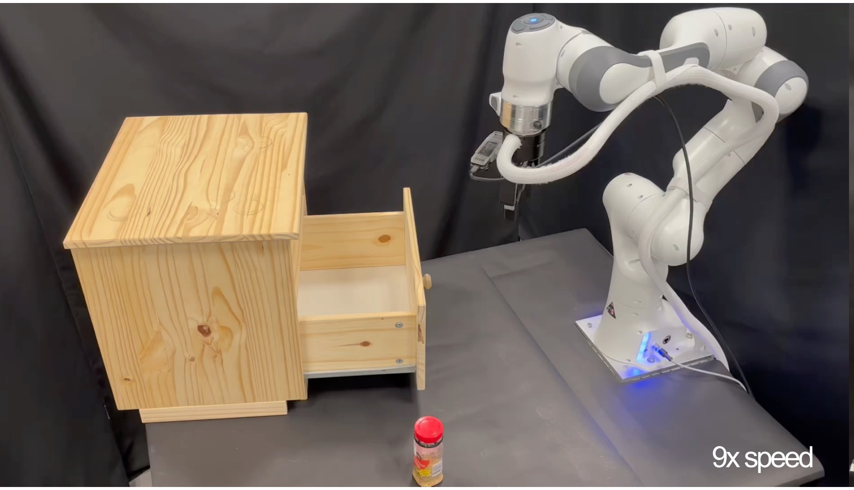

"Insert the vial in the vialplate"
9x speed

# Insert vials into different holes on the vial plate

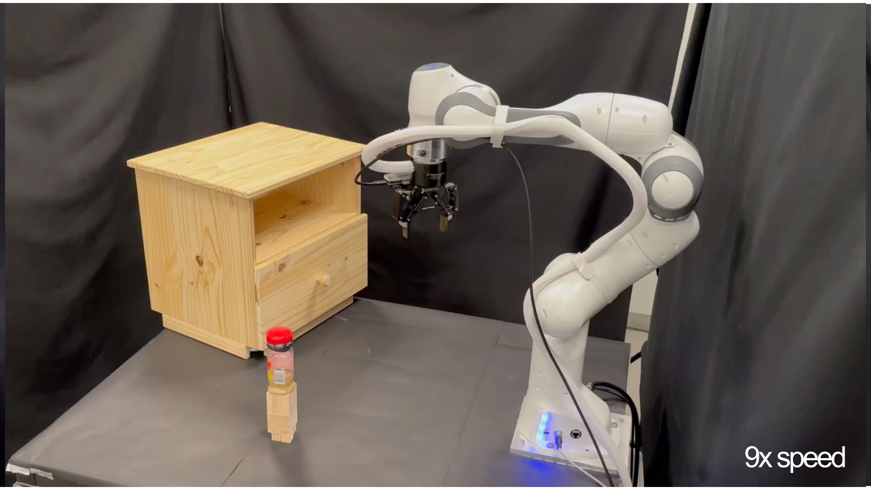



"Place the lid on the pot,
then place the pot on the stove".
9x speed

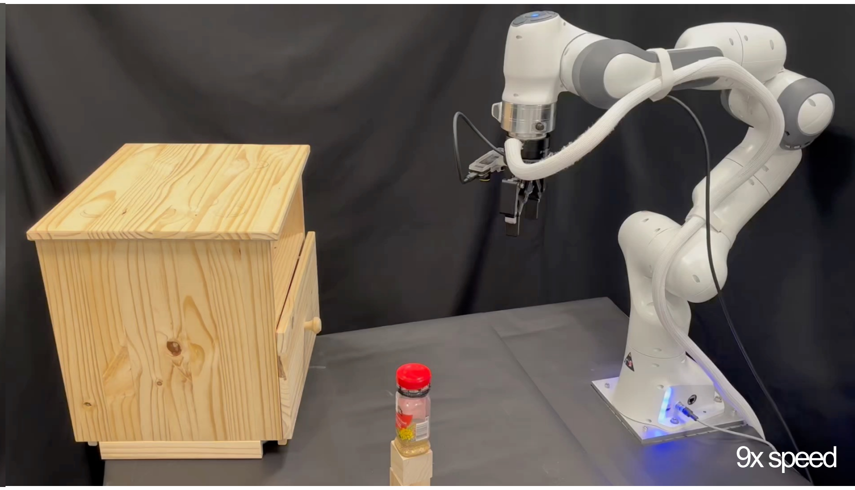
"Place the bottle in the drawer"




"Place the bottle on the middle shelf"




"Place the bottle on the top shelf"



# OG-VLA

## 3D-Aware Vision Language Action Model via Orthographic Image Generation

Ishika Singh, Ankit Goyal, Stan Birchfield, Dieter Fox, Animesh Garg, Valts Blukis

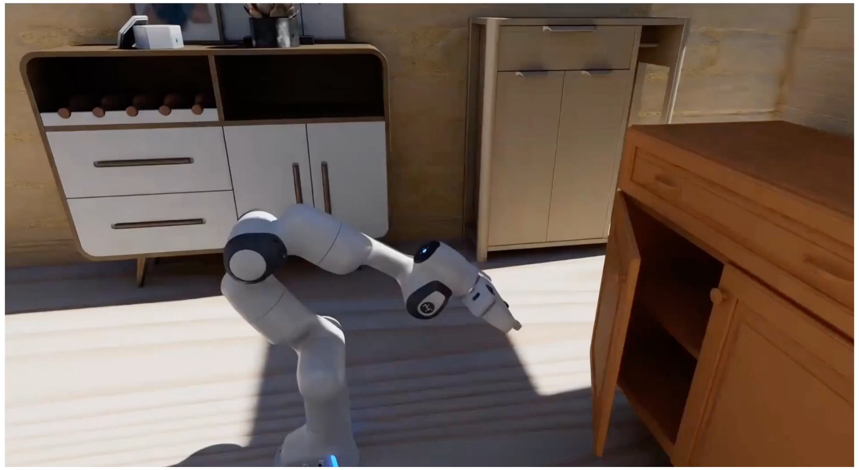
# Problem Statement

Pickup Object

*Lift the bottle* **30cm** *from the ground*

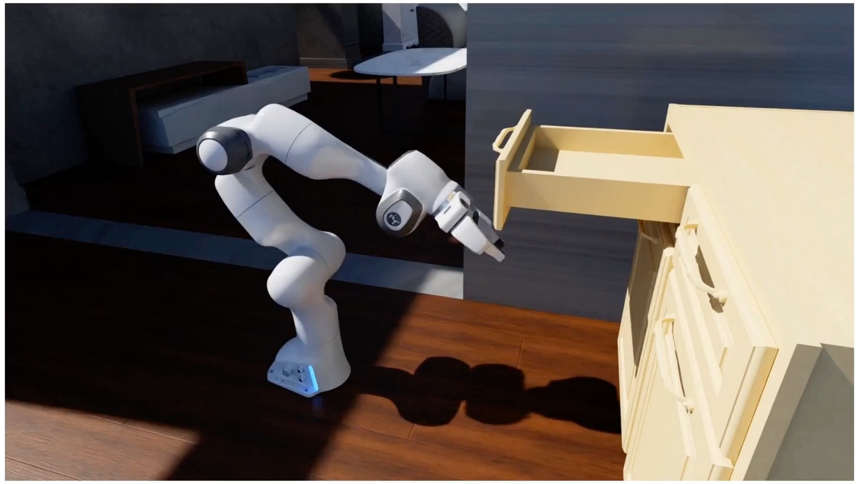
Open Cabinet

*Pull the* ***left*** *cabinet* **half** *open*



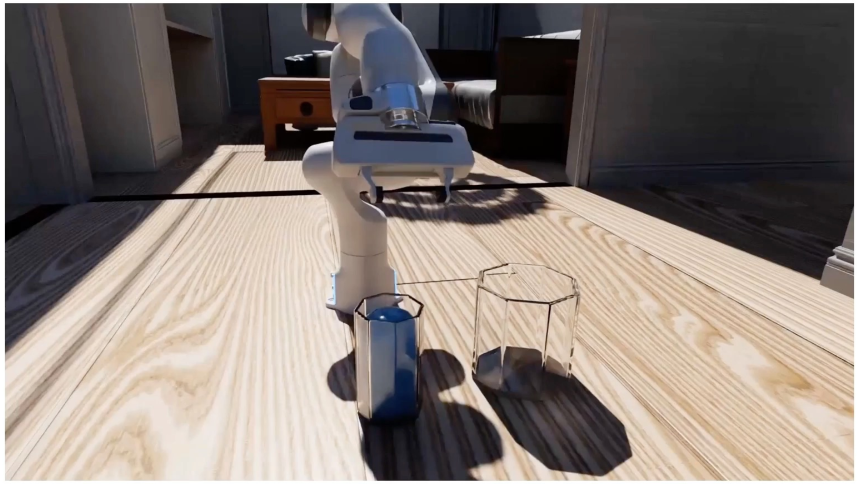
Close Drawer

*Push the* ***top left*** *dresser* **entirely** *closed*



Transfer Water

*Pour* **80%** *of the liquid to the cup*

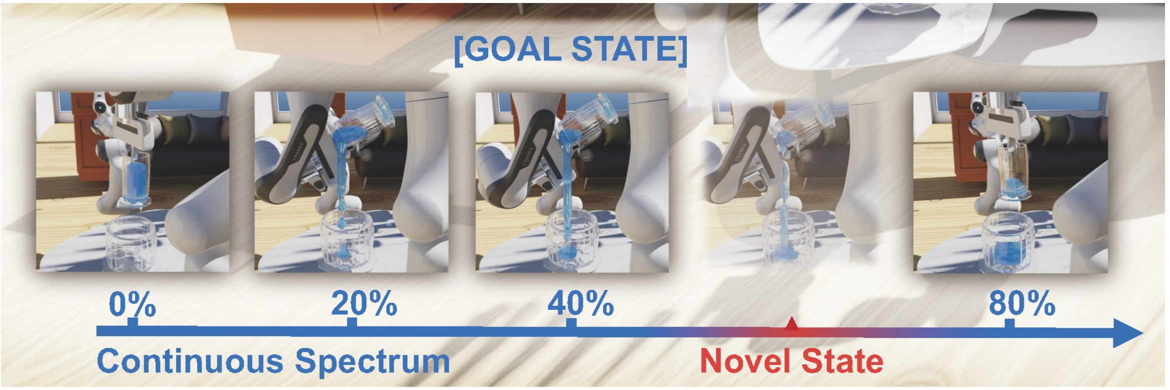


Media source: Arnold, ICCV 2023

# Generalization

Novel environments

Novel objects




Unseen language instructions

Media source: Arnold, ICCV 2023

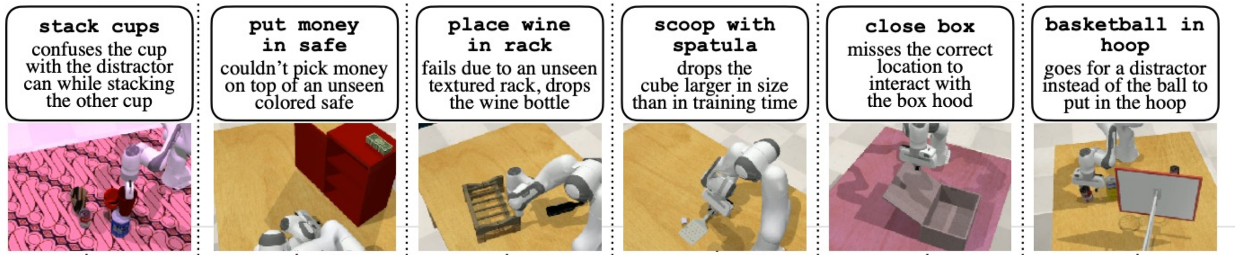
# Prior Works struggle with Generalization

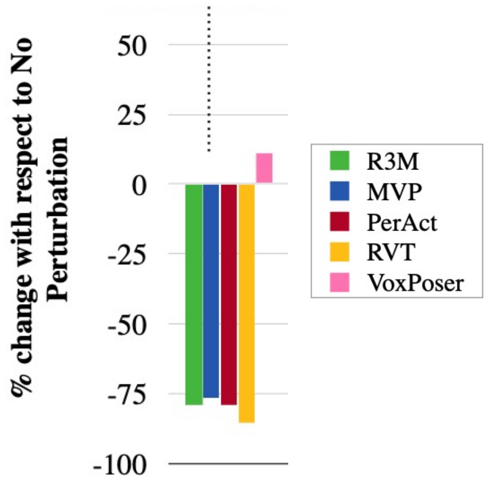
## ARNOLD

| | | |
|---|---|---|
| PerAct (MT) | 29.14 | 44.90 |
| Novel Object | 14.53 | 23.48 |
| Novel Scene | 19.00 | 34.37 |
| Novel State/Instr | 4.58 | 6.80 |

**Multi-Task Peract**
average across all Arnold benchmark tasks

## COLOSSEUM




Why? They are trained from scratch and overfit to their training data

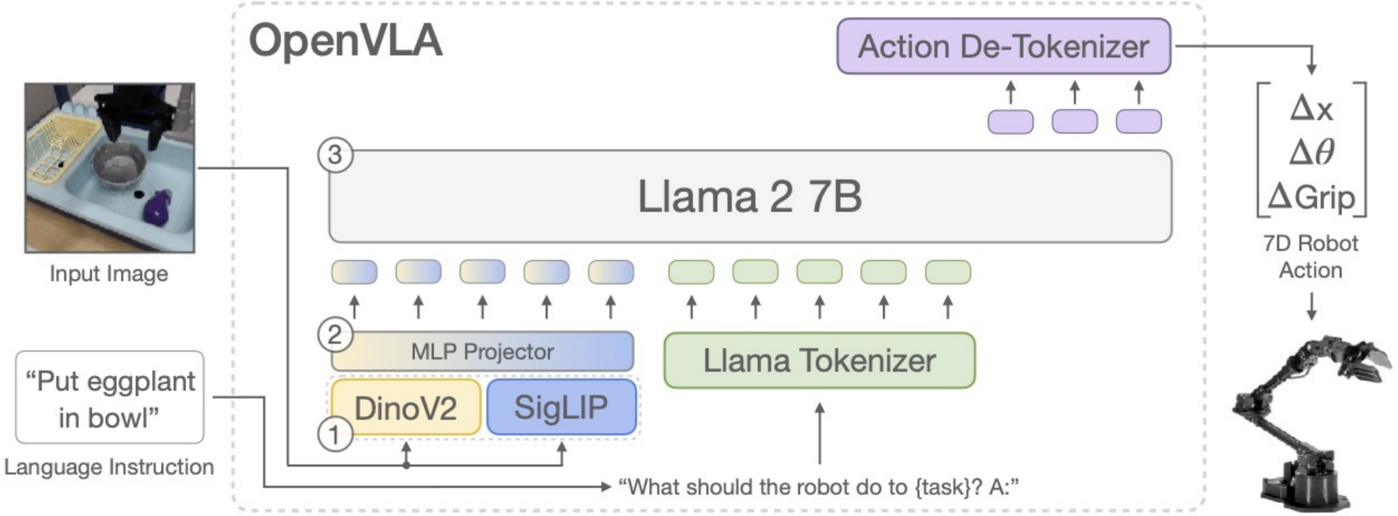
# What about VLAs?

Most use 2D single-view input.

Output space is actions vectors expressed as text tokens.

Not very efficient for learning tasks that are inherently 3D.

Perhaps we can leverage built-in priors used in 3D BC methods like RVT or Act3D for learning VLAs?



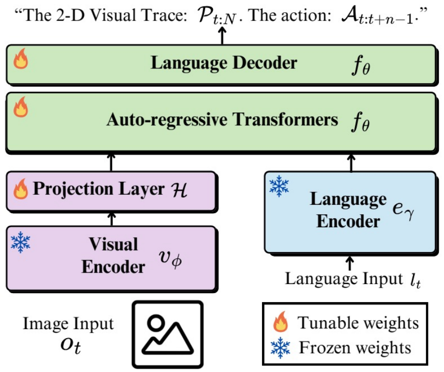
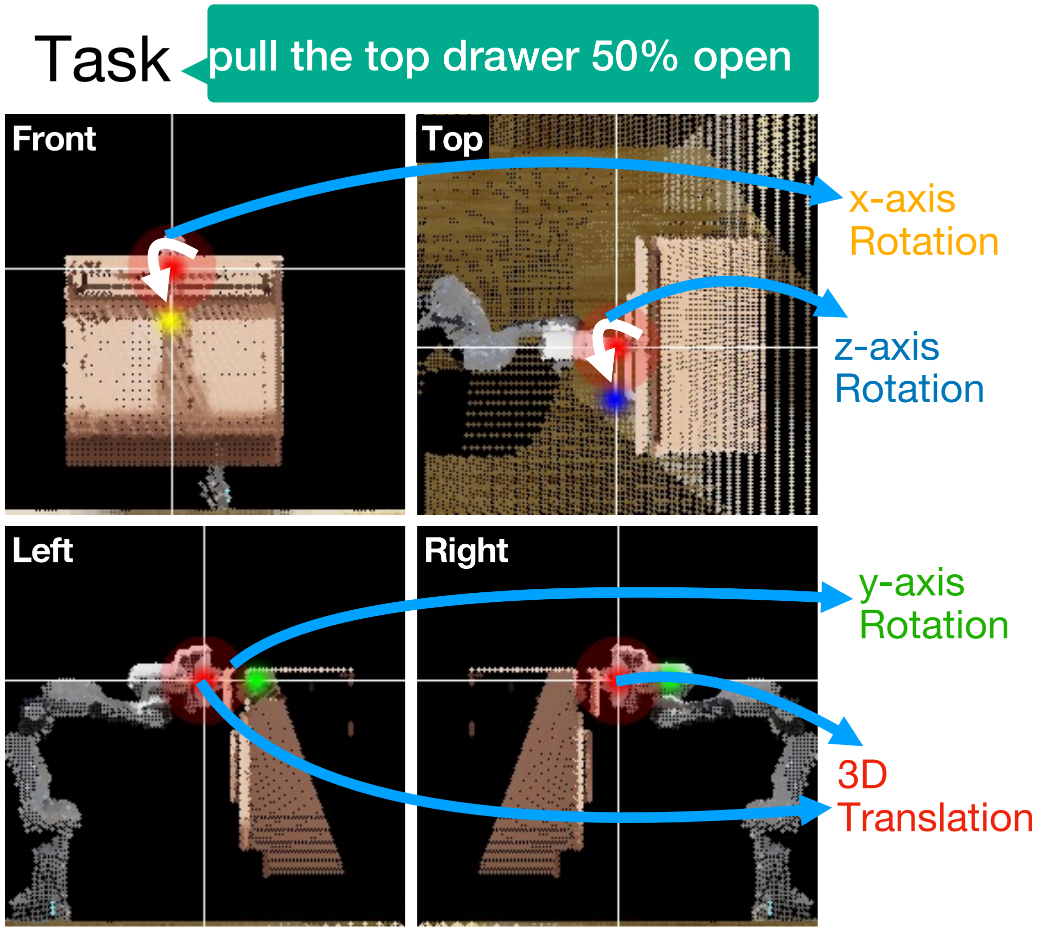
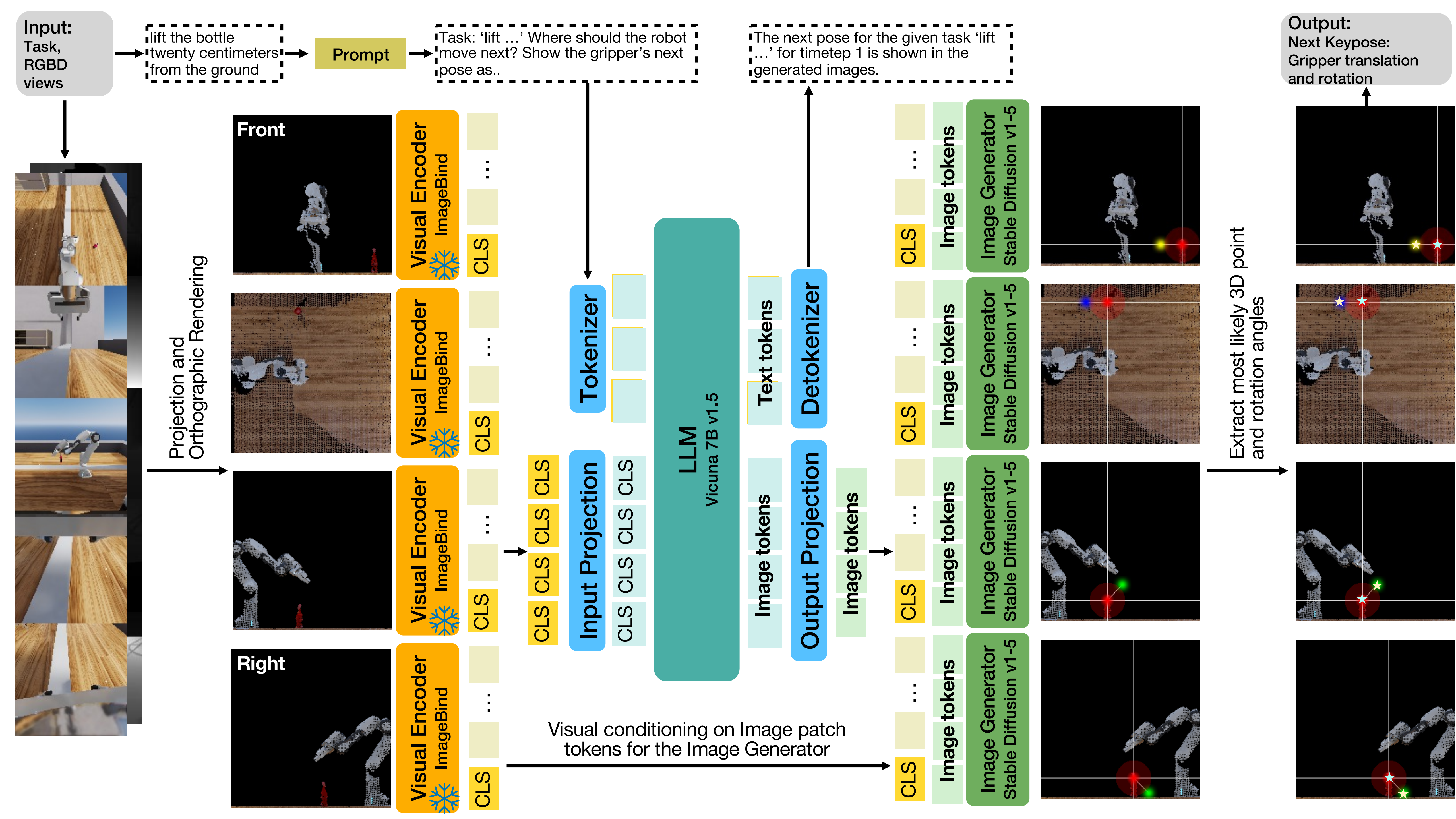




Figure 2: **Architecture of LLARVA.**

# Our approach: OG-VLA

Generating actions on orthographic views



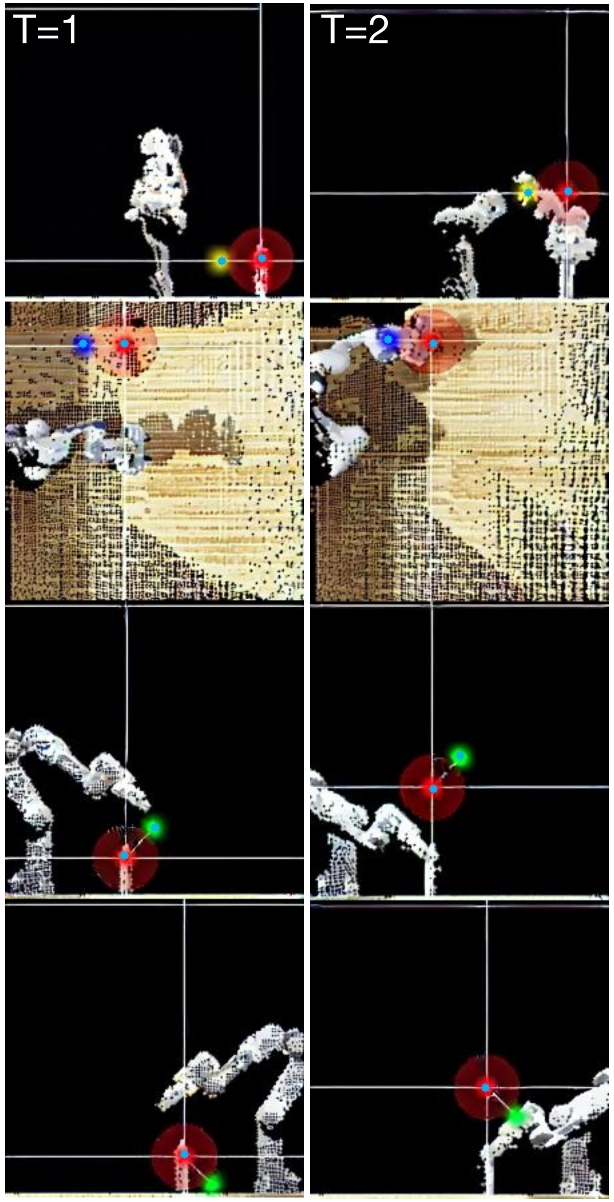

Input:
Task,
RGBD
views
lift the bottle
twenty centimeters
from the ground
Prompt
Task: 'lift …' Where should the robot
move next? Show the gripper's next
pose as..
The next pose for the given task 'lift
…' for timetep 1 is shown in the
generated images.
Output:
Next Keypose:
Gripper translation
and rotation
Projection and
Orthographic Rendering
Front
Right
Visual Encoder
ImageBind
CLS
Tokenizer
Input Projection
LLM
Vicuna 7B v1.5
Text tokens
Image tokens
Detokenizer
Output Projection
Image tokens
Image Generator
Stable Diffusion v1-5
Extract most likely 3D point
and rotation angles
Visual conditioning on Image patch
tokens for the Image Generator

# Results: Eval on Arnold (3D-manipulation)

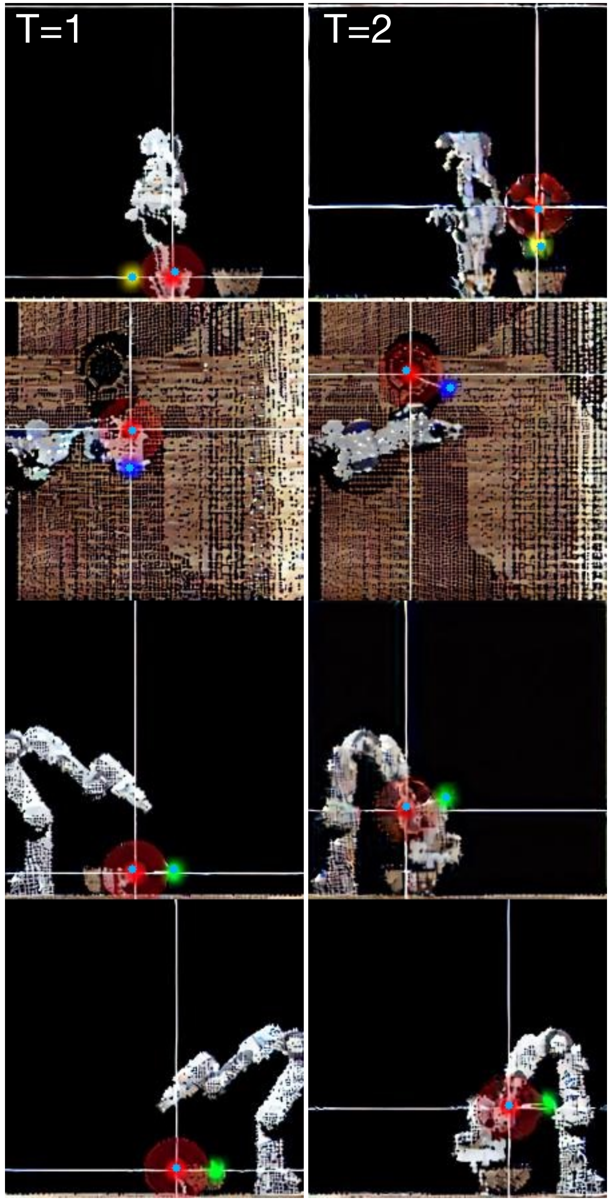
| Model | Pickup Object | Reorient Object | Open Drawer | Close Drawer | Open Cabinet | Close Cabinet | Pour Water | Transfer Water | Overall |
|---|---|---|---|---|---|---|---|---|---|
| **6D-CLIPort** | 6.7 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.8 |
| -Novel Object | 8.4 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| -Novel Scene | 10.4 | 0.0 | 0.0 | 0.0 | 0.0 | 1.3 | 0.0 | 0.0 | 1.5 |
| -Novel State | 0.00 | 0.0 | 0.0 | 0.7 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 |
| **PerAct** | 83.3 ± 2.4 | **16.7** ± 6.2 | 30.0 ± 10.8 | 31.7 ± 8.5 | **25.0** ± 0.0 | **30.0** ± 0.0 | **36.7** ± 6.2 | 18.3 ± 2.4 | 34.0 ± 3.1 |
| -Novel Object | 75.0 ± 0.0 | 3.3 ± 2.4 | 0.0 ± 0.0 | 23.3 ± 13.1 | 0.0 ± 0.0 | 0.0 ± 0.0 | 30.0 ± 4.1 | 1.7 ± 2.4 | 16.7 ± 2.6 |
| -Novel Scene | **75.0** ± 4.1 | **13.3** ± 2.4 | 13.3 ± 9.4 | 30.0 ± 14.1 | 0.0 ± 0.0 | **6.7** ± 2.4 | **26.7** ± 6.2 | 3.3 ± 2.4 | 21.0 ± 3.1 |
| -Novel State | **16.7** ± 2.4 | 1.7 ± 2.4 | 5.0 ± 0.0 | 11.7 ± 6.2 | 0.0 ± 0.0 | 0.0 ± 0.0 | 5.0 ± 0.0 | 11.7 ± 2.4 | 6.5 ± 1.2 |
| **OG-VLA@30k** | 86.7±2.9 | 15.0±8.7 | 38.3±2.9 | 51.7±2.9 | 0.0±0.0 | 16.7±2.9 | 25.0±5.0 | 16.7±7.6 | 31.2±2.9 |
| -Novel Object | **85.0**±5.0 | 0.0±0.0 | **1.7**±2.9 | 55.0±13.2 | **1.7**±2.9 | **5.0**±5.0 | 18.3±2.9 | 6.7±7.6 | 21.7±0.7 |
| -Novel Scene | 70.0±2.9 | 1.7±2.8 | 26.7±11.5 | 36.7±5.8 | **1.7**±2.9 | 1.7±2.9 | 16.7±11.5 | 8.3±2.9 | 20.8±1.3 |
| -Novel State | 0.0±0.0 | **13.3**±7.6 | 13.3±2.9 | **20.0**±0.0 | 0.0±0.0 | 0.0±0.0 | **8.3**±7.6 | 13.3±2.9 | 8.5±1.9 |
| **OG-VLA@100k** | **88.3** ± 2.4 | **16.7** ± 9.4 | **48.3** ± 2.4 | **56.7** ± 2.4 | 6.7 ± 4.7 | 23.3 ± 16.5 | 33.3 ± 6.2 | **28.3** ± 2.4 | **37.7** ± 0.6 |
| -Novel Object | 65.0 ± 8.2 | **15.0** ± 4.1 | **1.7** ± 2.4 | **58.3** ± 12.5 | 0.0 ± 0.0 | **5.0** ± 4.1 | **45.0** ± 8.2 | **8.3** ± 4.7 | **24.8** ± 1.2 |
| -Novel Scene | **75.0** ± 7.1 | **13.3** ± 8.5 | **31.7** ± 4.7 | **51.7** ± 2.4 | **1.7** ± 2.4 | 5.0 ± 4.1 | **26.7** ± 2.4 | **25.0** ± 7.1 | **28.8** ± 0.5 |
| -Novel State | 0.0 ± 0.0 | **13.3** ± 2.4 | **25.0** ± 7.1 | 15.0 ± 4.1 | 0.0 ± 0.0 | 0.0 ± 0.0 | 6.7 ± 4.7 | **20.0** ± 7.1 | **10.0** ± 0.9 |

# Ablation Study

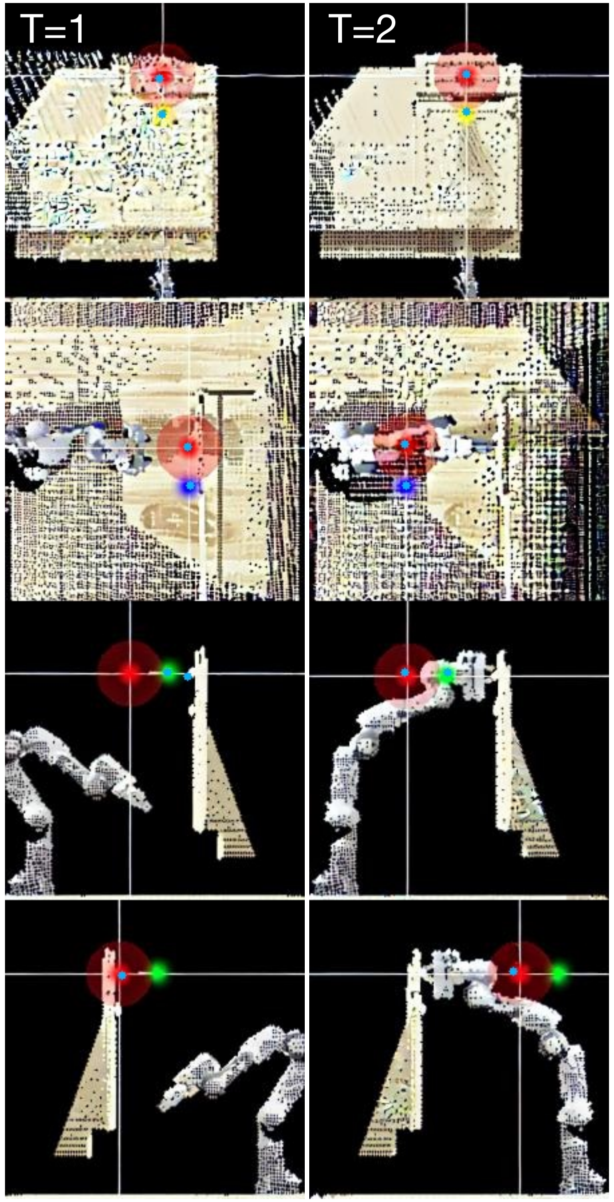
| Model | Pickup Object | Reorient Object | Open Drawer | Close Drawer | Open Cabinet | Close Cabinet | Pour Water | Transfer Water | Overall |
|---|---|---|---|---|---|---|---|---|---|
| **+Tiled Views** | 75.0 ± 7.1 | 6.7 ± 4.7 | 33.3 ± 2.4 | 28.3 ± 6.2 | 1.7 ± 2.4 | 20.0 ± 4.1 | 15.0 ± 10.8 | 18.3 ± 6.2 | 24.8 ± 0.3 |
| -Novel Object | 58.3 ± 8.5 | 15.0 ± 7.1 | 1.7 ± 2.4 | 30.0 ± 0.0 | 0.0 ± 0.0 | 10.0 ± 4.1 | 40.0 ± 4.1 | 6.7 ± 6.2 | 20.2 ± 1.3 |
| -Novel Scene | 60.0 ± 7.1 | 15.0 ± 10.8 | 45.0 ± 7.1 | 21.7 ± 6.2 | 5.0 ± 4.1 | 5.0 ± 4.1 | 26.7 ± 6.2 | 1.7 ± 2.4 | 22.5 ± 0.5 |
| -Novel State | 0.0 ± 0.0 | 10.0 ± 4.1 | 10.0 ± 7.1 | 18.3 ± 4.7 | 1.7 ± 2.4 | 1.7 ± 2.4 | 1.7 ± 2.4 | 16.7 ± 11.8 | 7.5 ± 2.6 |
| **-LLM** | 86.7 ± 2.4 | 5.0 ± 7.1 | 6.7 ± 4.7 | 33.3 ± 4.7 | 0.0 ± 0.0 | 6.7 ± 4.7 | 15.0 ± 0.0 | 6.7 ± 2.4 | 20.0 ± 1.5 |
| -Novel Object | 68.3 ± 6.2 | 1.7 ± 2.4 | 6.7 ± 9.4 | 40.0 ± 4.1 | 0.0 ± 0.0 | 3.3 ± 2.4 | 10.0 ± 4.1 | 8.3 ± 2.4 | 17.3 ± 1.6 |
| -Novel Scene | 71.7 ± 6.2 | 8.3 ± 6.2 | 18.3 ± 2.4 | 21.7 ± 8.5 | 3.3 ± 2.4 | 0.0 ± 0.0 | 15.0 ± 4.1 | 5.0 ± 4.1 | 17.9 ± 3.3 |
| -Novel State | 0.0 ± 0.0 | 0.0 ± 0.0 | 6.7 ± 2.4 | 16.7 ± 2.4 | 0.0 ± 0.0 | 1.7 ± 2.4 | 3.3 ± 2.4 | 10.0 ± 4.1 | 4.8 ± 0.8 |
| **+Tiled Views -LLM** | 71.7 ± 10.3 | 1.7 ± 2.4 | 13.3 ± 8.5 | 16.7 ± 4.7 | 0.0 ± 0.0 | 8.3 ± 2.4 | 15.0 ± 8.2 | 10.0 ± 0.0 | 17.1 ± 3.4 |
| -Novel Object | 56.7 ± 8.5 | 8.3 ± 2.4 | 1.7 ± 2.4 | 16.7 ± 8.5 | 0.0 ± 0.0 | 1.7 ± 2.4 | 15.0 ± 4.1 | 6.7 ± 2.4 | 13.3 ± 1.6 |
| -Novel Scene | 61.7 ± 6.2 | 5.0 ± 4.1 | 20.0 ± 4.1 | 11.7 ± 6.2 | 0.0 ± 0.0 | 3.3 ± 4.7 | 10.0 ± 0.0 | 10.0 ± 0.0 | 15.2 ± 1.2 |
| -Novel State | 0.0 ± 0.0 | 6.7 ± 2.4 | 10.0 ± 0.0 | 30.0 ± 4.1 | 1.7 ± 2.4 | 0.0 ± 0.0 | 6.7 ± 6.2 | 3.3 ± 4.7 | 7.3 ± 0.8 |
| **-Instruction to IG** | 71.7 ± 4.7 | 8.3 ± 2.4 | 20.0 ± 10.8 | 40.0 ± 12.2 | 1.7 ± 2.4 | 11.7 ± 9.4 | 5.0 ± 4.1 | 15.0 ± 4.1 | 21.7 ± 2.6 |
| -Novel Object | 66.7 ± 6.2 | 0.0 ± 0.0 | 1.7 ± 2.4 | 45.0 ± 4.1 | 0.0 ± 0.0 | 8.3 ± 2.4 | 20.0 ± 4.1 | 1.7 ± 2.4 | 17.9 ± 0.8 |
| -Novel Scene | 50.0 ± 8.2 | 11.7 ± 2.4 | 25.0 ± 0.0 | 26.7 ± 2.4 | 10.0 ± 0.0 | 11.7 ± 2.4 | 13.3 ± 4.7 | 5.0 ± 4.1 | 19.2 ± 2.1 |
| -Novel State | 0.0 ± 0.0 | 3.3 ± 4.7 | 8.3 ± 6.2 | 13.3 ± 8.5 | 0.0 ± 0.0 | 0.0 ± 0.0 | 1.7 ± 2.4 | 8.3 ± 2.4 | 4.4 ± 1.8 |



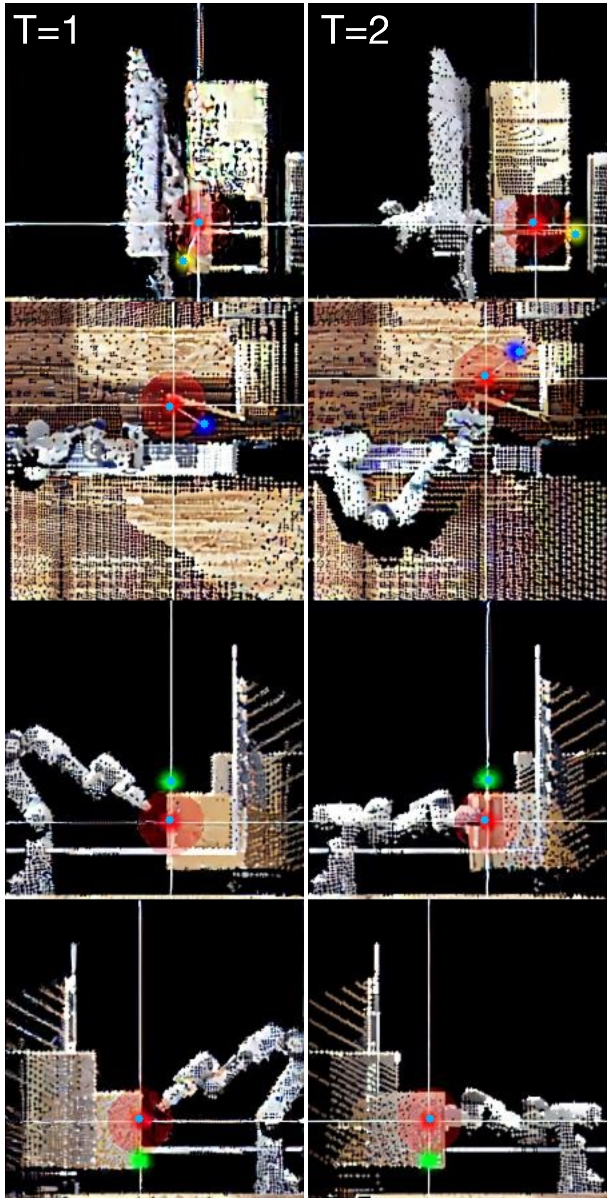

Task1: lift the bottle thirty centimeters from the ground



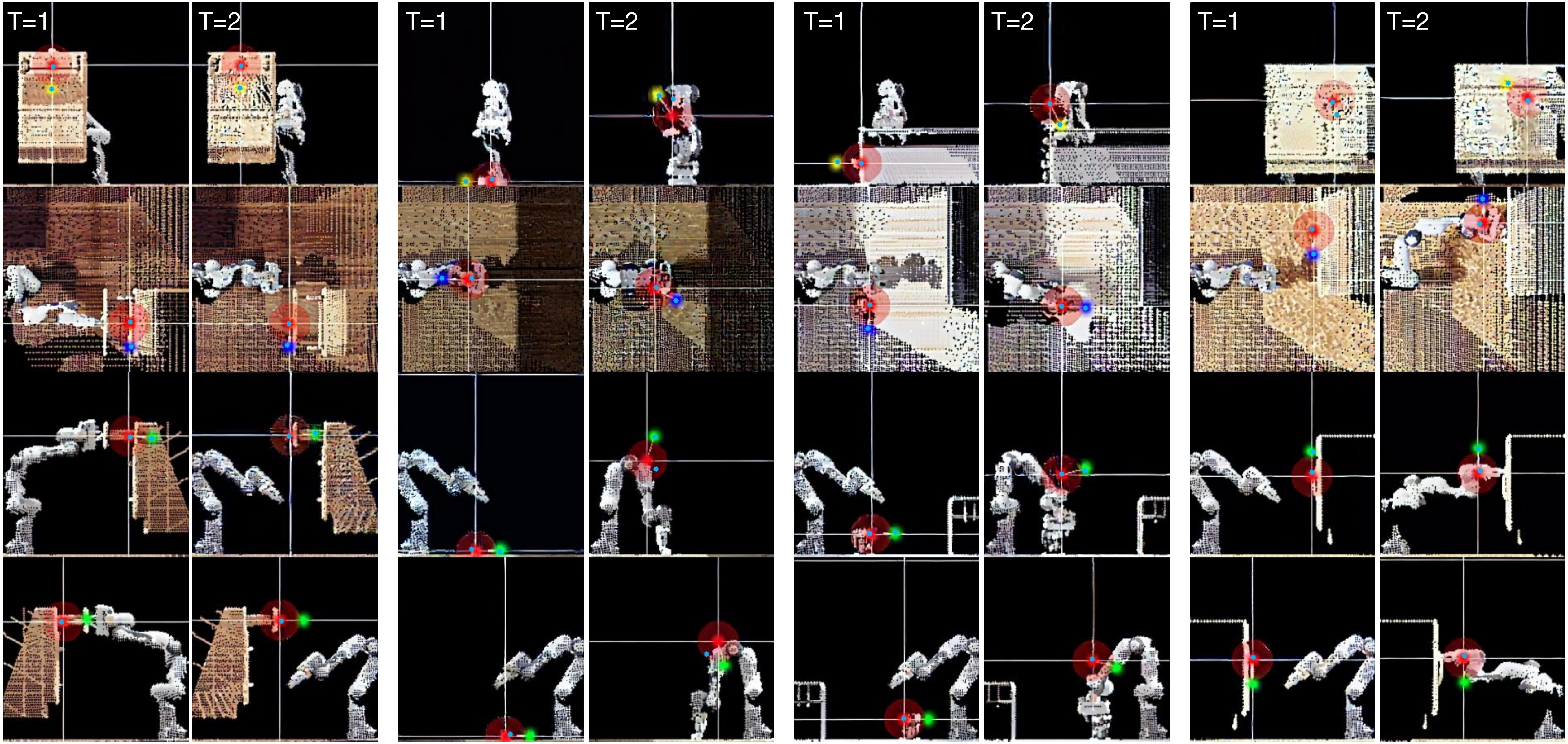

Task2: add forty percent of the liquid to the cup




Task3: pull the top dresser one hundred percent open




Task4: shut the cabinet half closed




Task: close the top dresser completely closed

Task: set the angle of the bottle forty-five degrees from the upward axis

Task: get seventy five percent water out of the glass

Task: pull the cabinet a quarter open

# The Computing Stack

## Physical AI

### Motion Generation Models

Fine-Tune Generalists for better Specialists.

Reinforcement Learning for Locomotion, WBC + Dexterity

Self-supervised learning without rewards

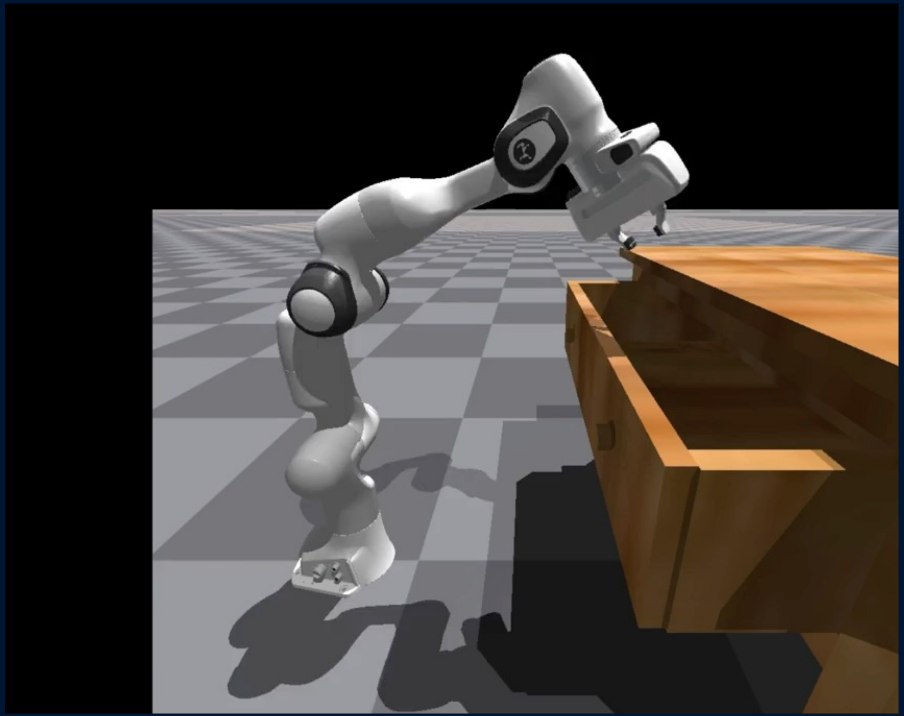
| | | |
|---|---|---|
| App | Natural Interaction Interface | Ease of Use |
| OS | Platform-Agnostic Planning | Modular Tool-Use |
| Driver | Hardware-Specific Skills | Optimize for Morphology |

# Act-AIM

## Discovering Robotic Interaction Modes with Discrete Representation Learning

Liquan Wang, Ankit Goyal, Haoping Xu, Animesh Garg

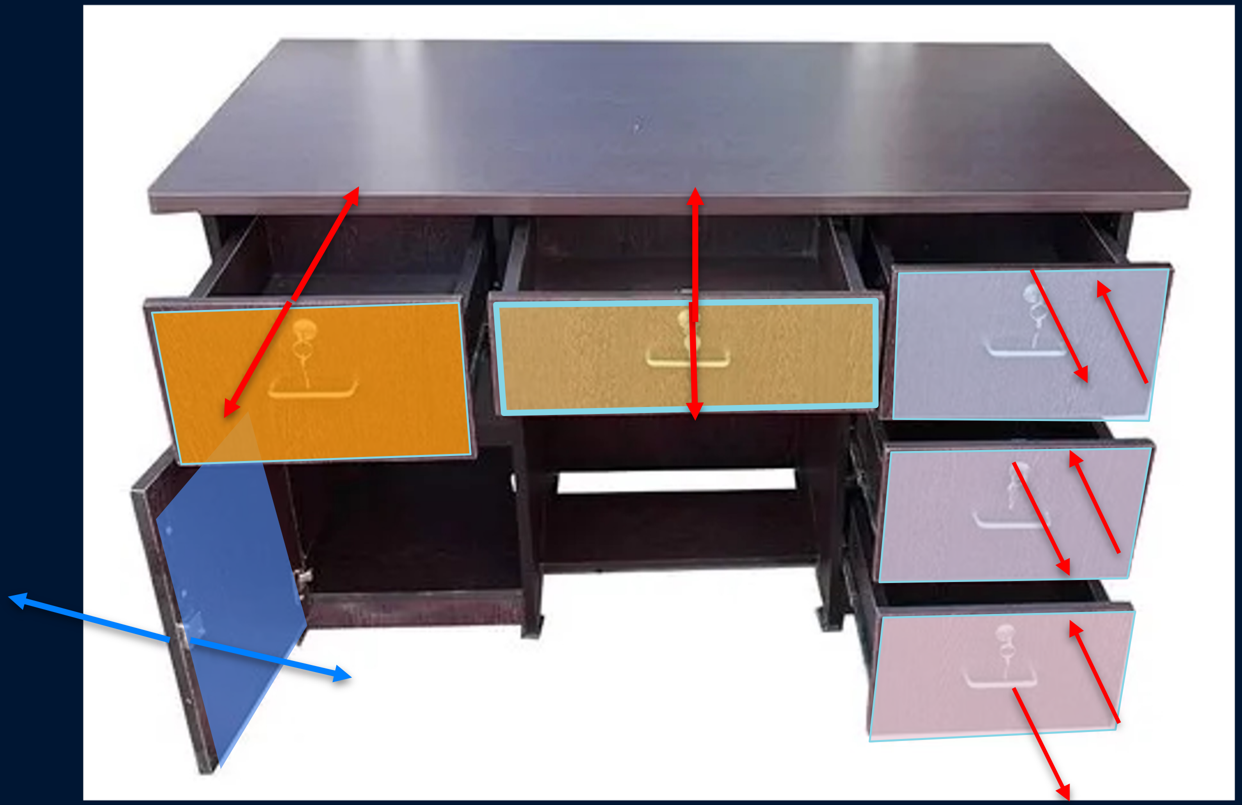
# Self-Supervised Learning



Learning to do “what can be done”
Learning from Self-Supervised Play

Transferring to Real world
Learning without data or rewards

# Learning without Supervision

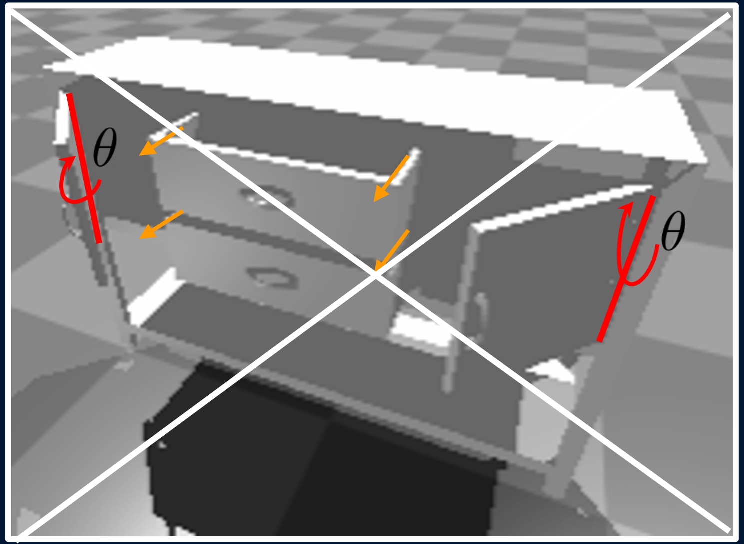


***Unsupervised Discovery** of **Interaction Modes***

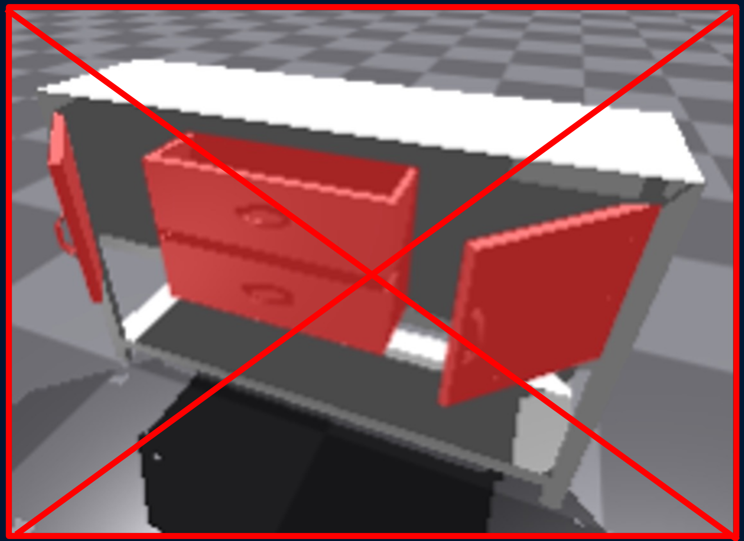
*Different Types of Motions (Revolute, Prismatic, …)*

*Variable Number of Links*

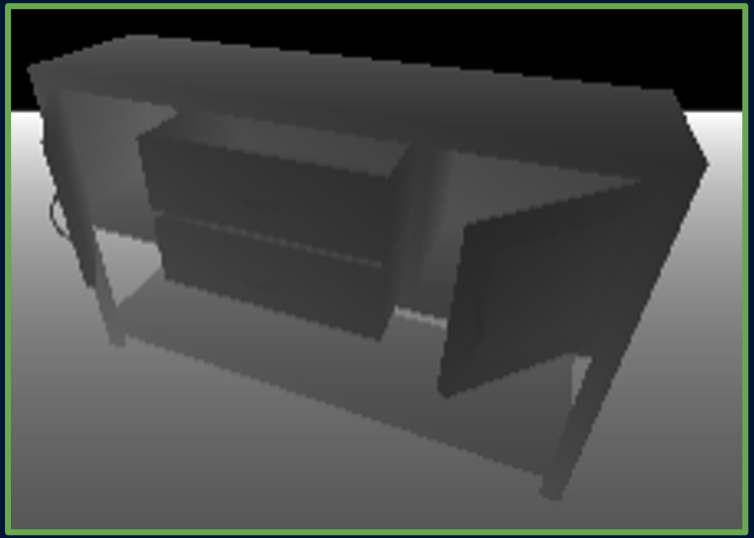
# Learning without Supervision




No Ground Truth
Articulation DoF

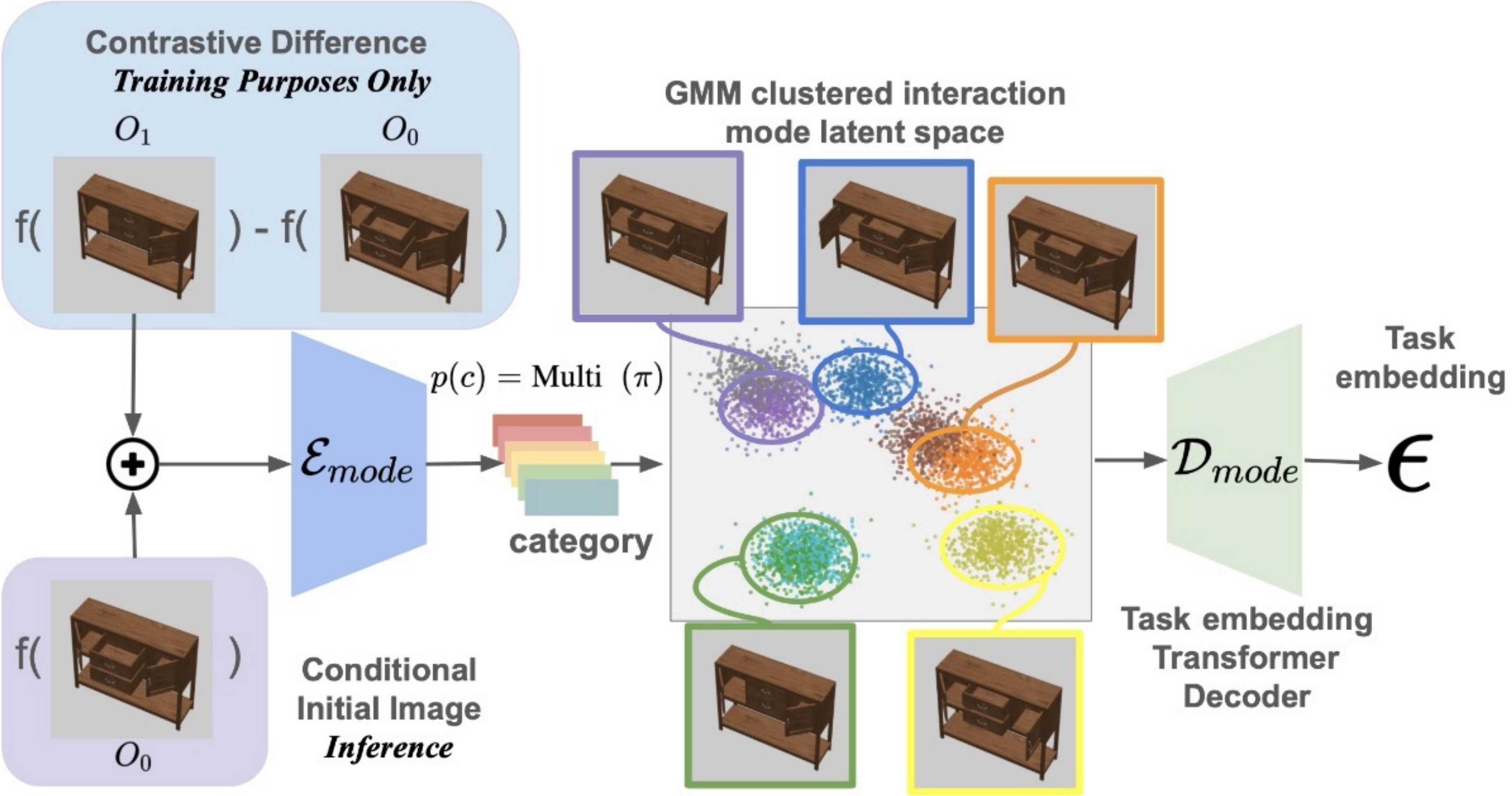


No Ground Truth
Part Segmentation

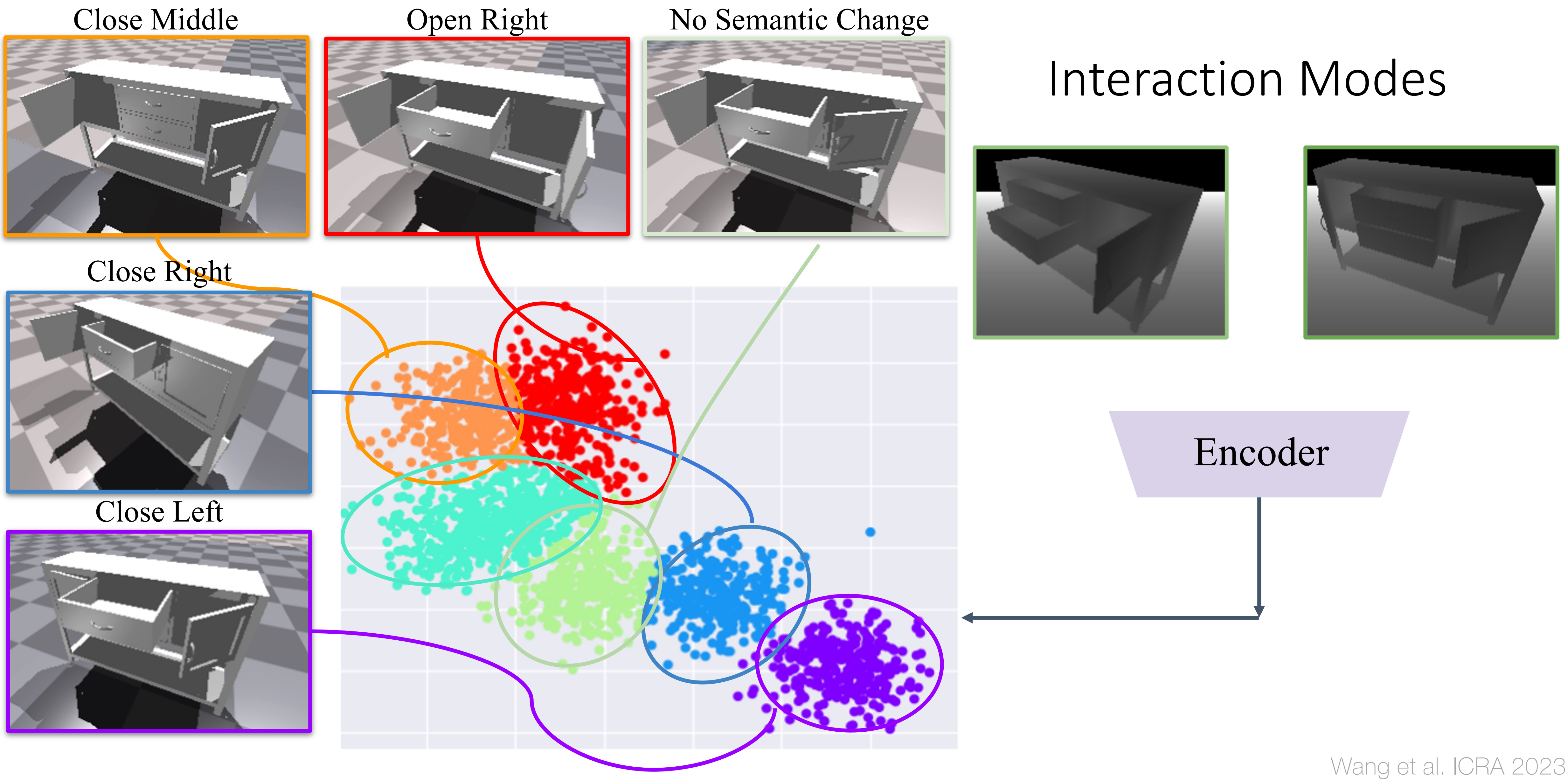


Visual Input
Object Depth

# Self-Supervised Data Collection

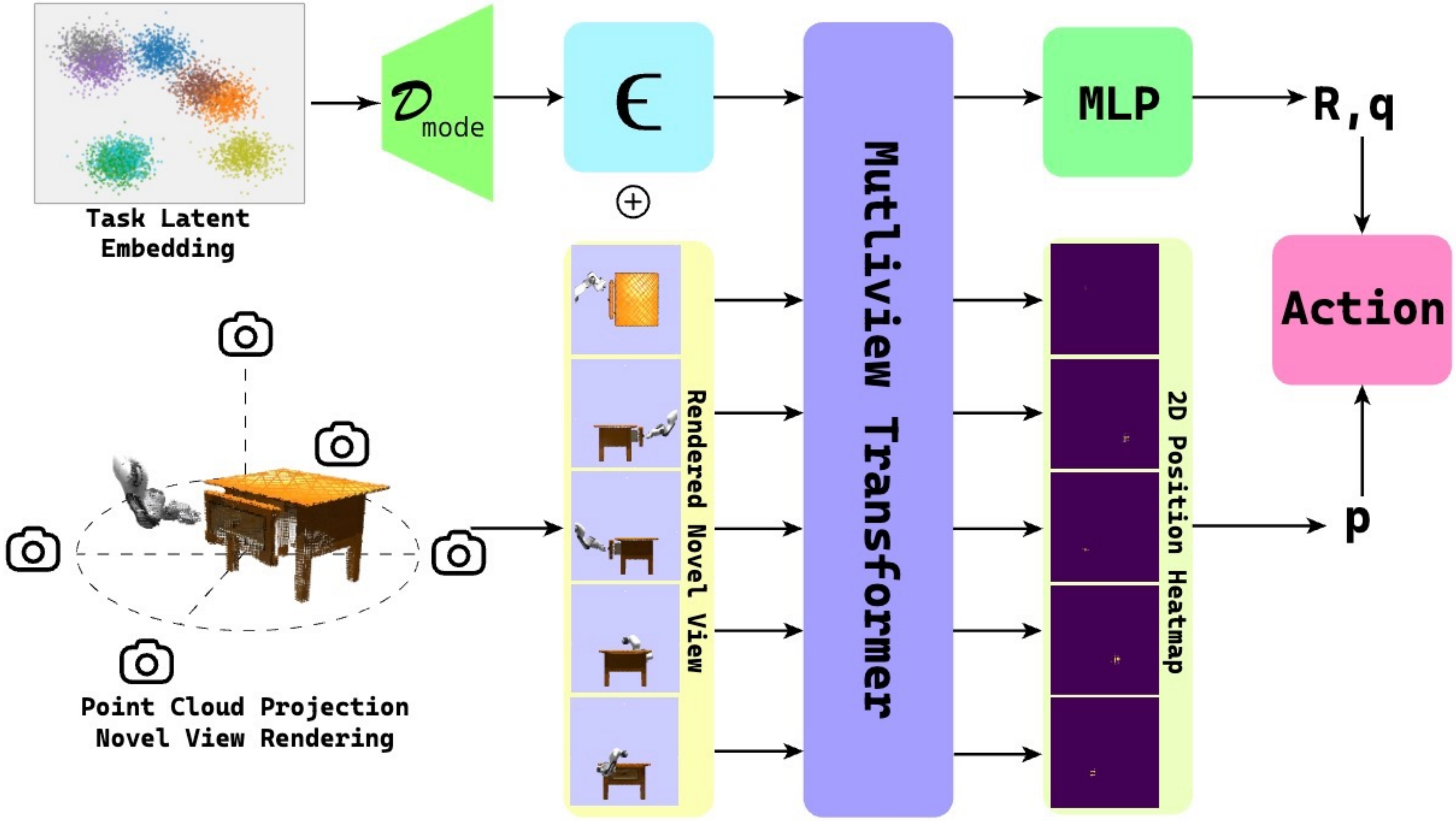


# Self-Supervised Data Collection

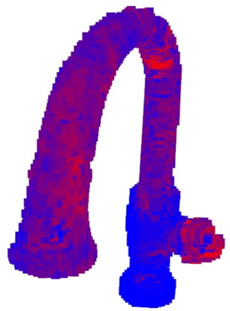



Wang et al. ICRA 2023

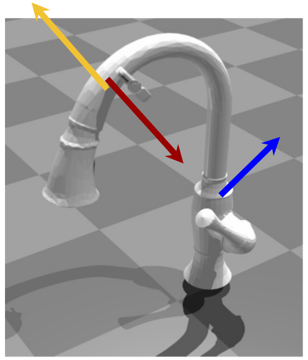
# Model Training: ActAIM-2

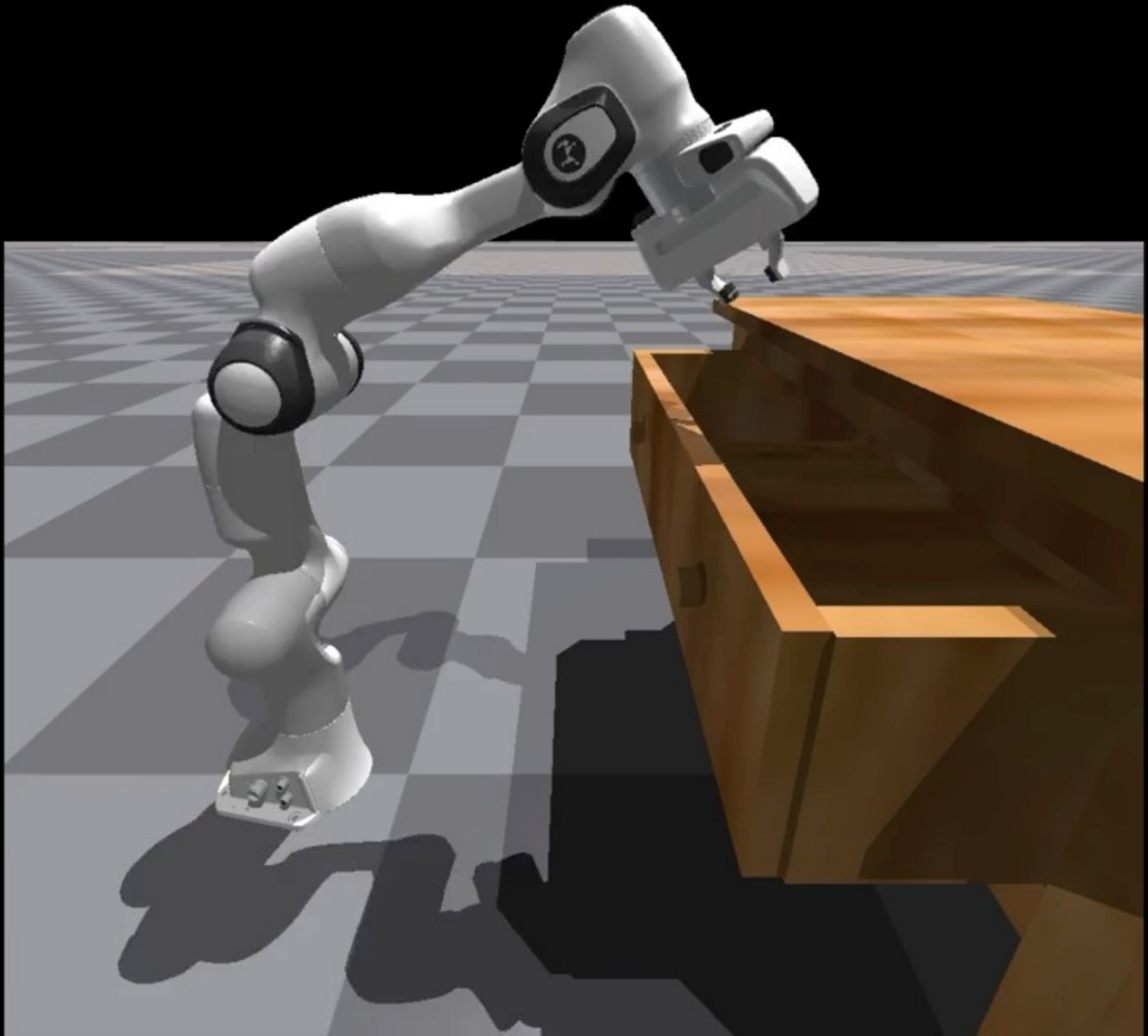



*Point Score* Prediction:
Interaction Distribution



*Action Predictor* Output:
Proposed Actions



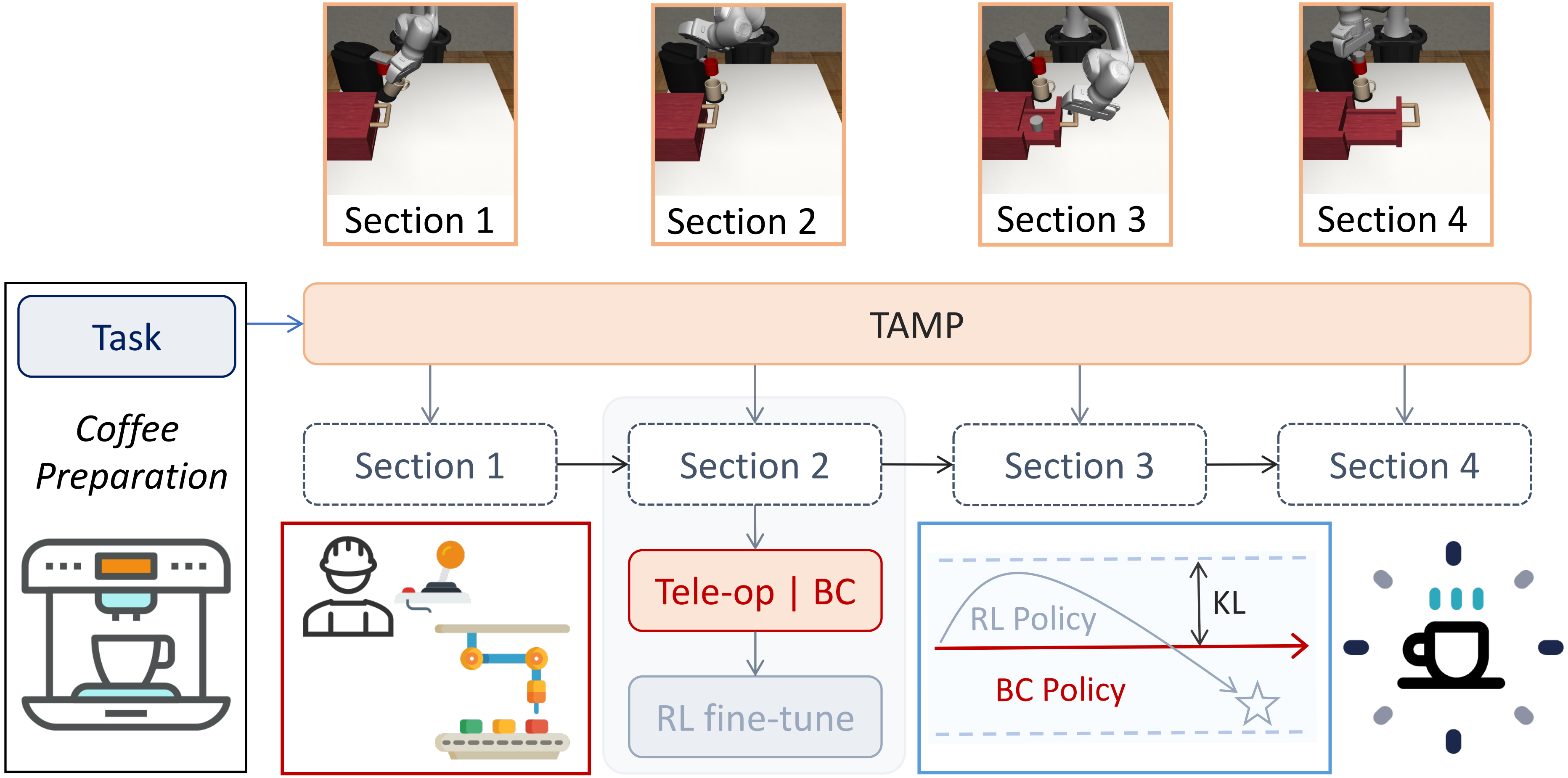


# SPIRE

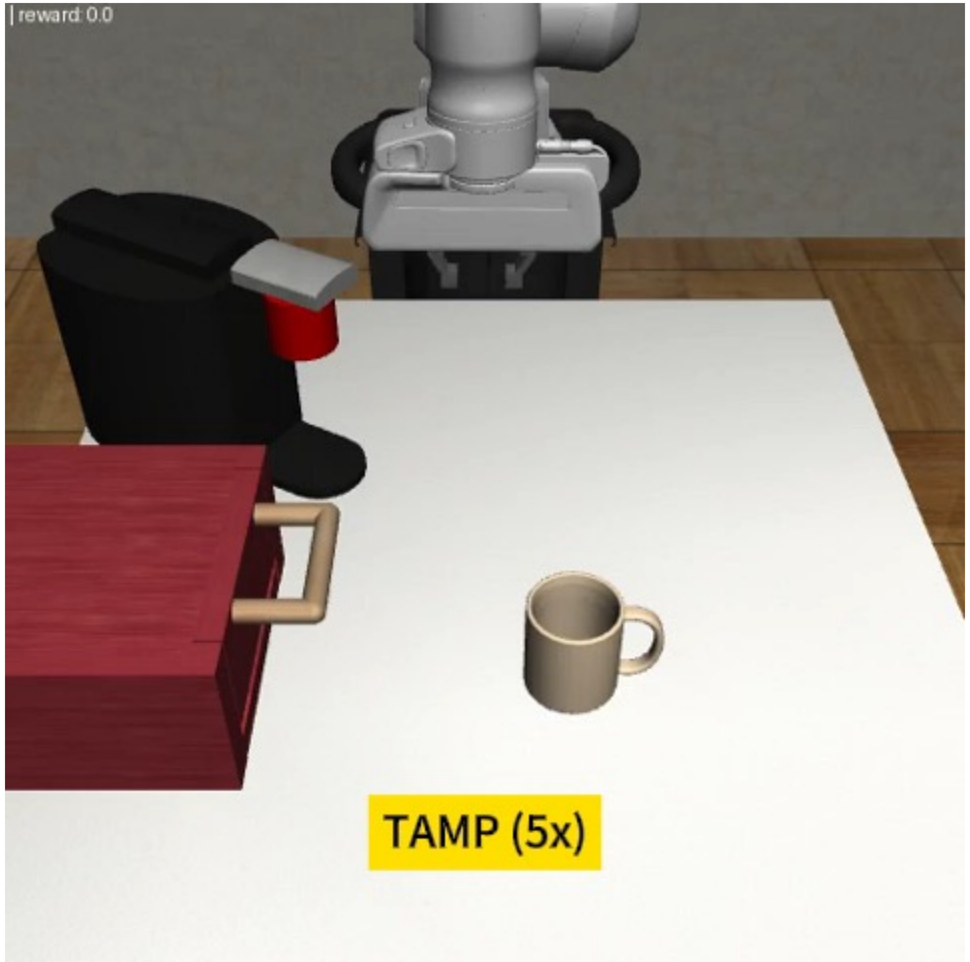
## Synergistic Planning, Imitation, and Reinforcement for Long-Horizon Manipulation

Zihan Wang, Ajay Mandlekar, Caelan Garrett, Animesh Garg

# Overview

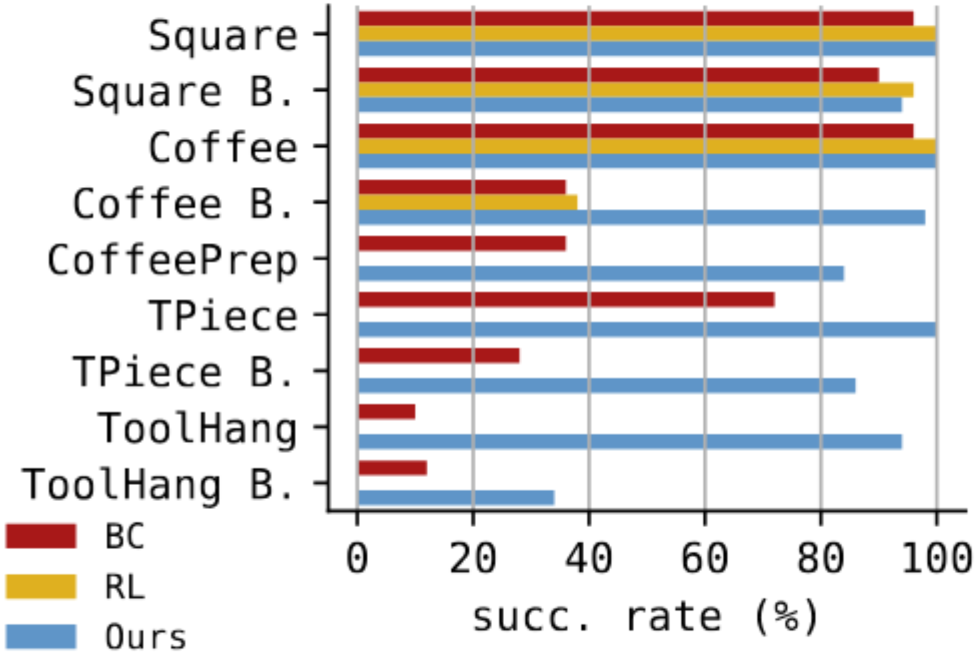

Section 1
Section 2
Section 3
Section 4
Task
TAMP
*Coffee Preparation*
Section 1
Section 2
Section 3
Section 4
Tele-op | BC
RL fine-tune
RL Policy
KL
BC Policy

# SPIRE Solves Long-Horizon Tasks (5 Sequential Subtasks)




* Red border indicates SPIRE agent-controlled sections

# SPIRE: Train proficient agents using a handful of data




SPIRE reaches 80% in 8/9 tasks, BC and RL have 3/9.

| **Episode Duration:** | **BC** [14] | **RL** [15] | **Ours** |
|---|---|---|---|
| Square | 18.1 | **8.3** | 11.6 |
| Square Broad | 24.5 | **8.4** | 13.6 |
| Coffee | 63.1 | **15.0** | 38.4 |
| Coffee Broad | 80.6 | **25.7** | 61.3 |
| Coffee Preparation | 193.3 | - | **168.5** |
| Three Piece | 58.7 | - | **34.0** |
| Three Piece Broad | 62.2 | - | **38.1** |
| Tool Hang | 81.8 | - | **61.7** |
| Tool Hang Broad | 130.5 | - | **109.8** |

SPIRE completes tasks in 60% of the time of BC agents.

# PWM

## Policy Learning with Multi-Task World Models

Ignat Georgiev, Varun Giridhar, Nicklas Hansen, Animesh Garg

# How to learn many things (better than data)

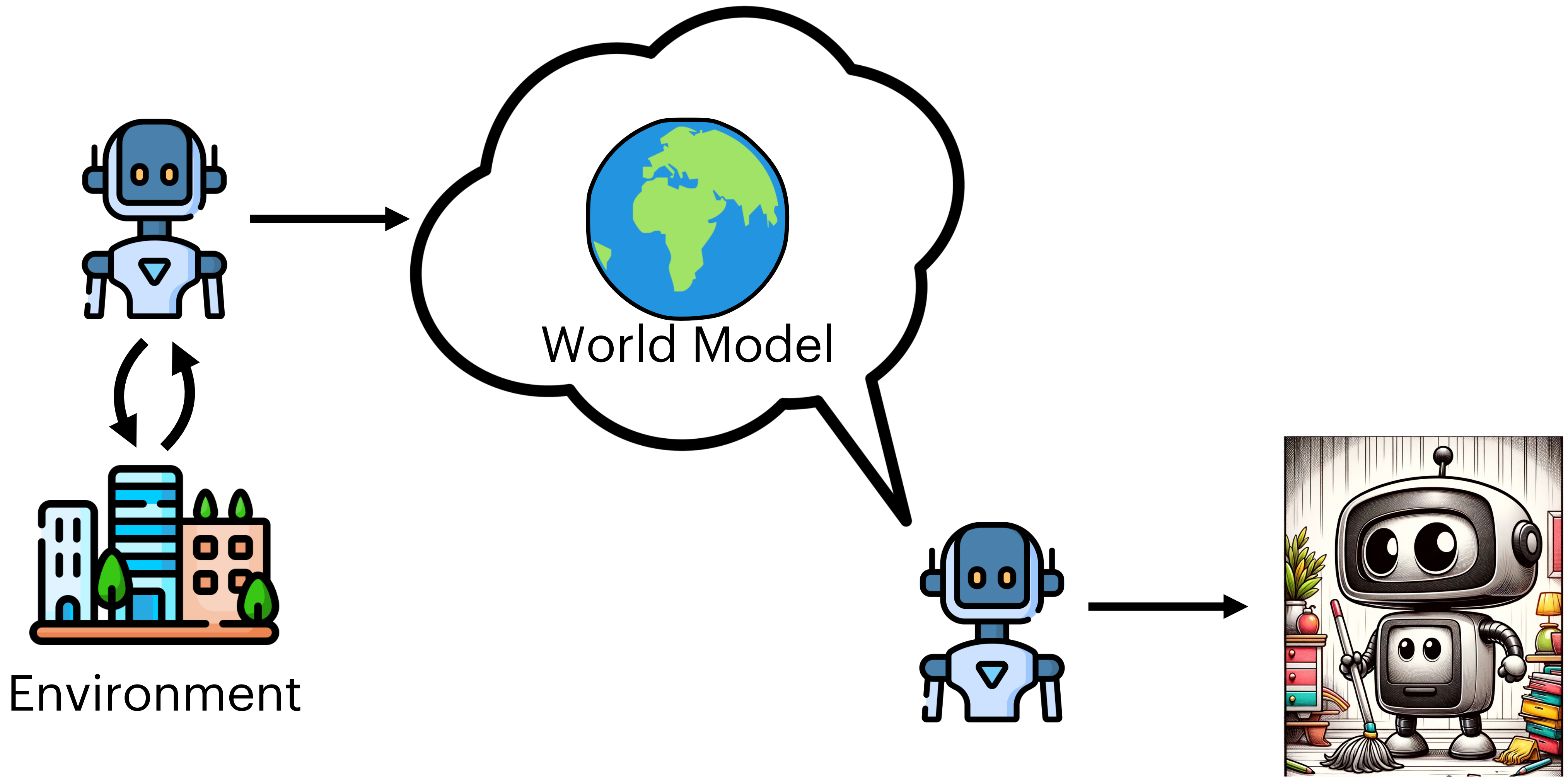
## PWM: Policy Learning with Large World Models

Large World Models

Multi-task

Efficient Policy Optimization

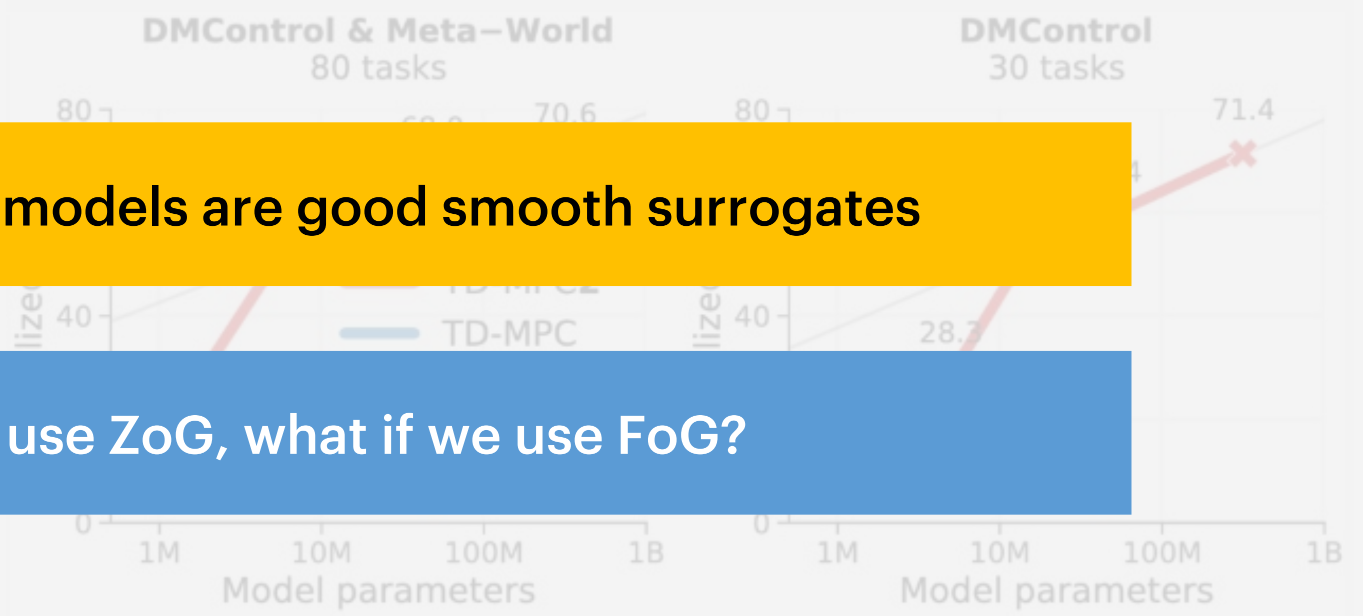
# World Model Framework



# TDMPC2

## A scalable multi-task world model approach

- First multi-task RL policy that scales to 80 different tasks
- Across MetaWorld and DMC
- F
planning
- L
e
params




**Turns out that TDMPC2 world models are good smooth surrogates**

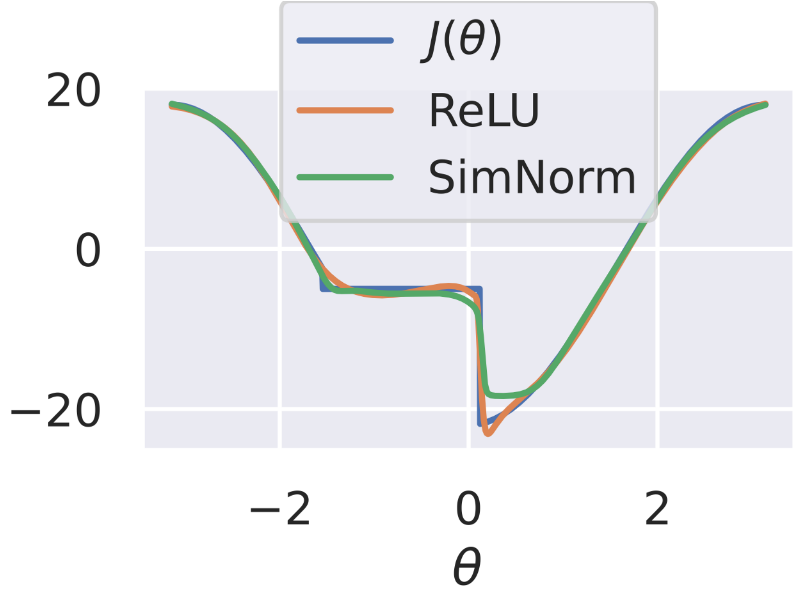
**But TDMPC2 choses to use ZoG, what if we use FoG?**

# World models are smooth surrogates

- When regularized correctly, world models can act as smooth surrogates
  - No sampling required!
- Maps $\boldsymbol{z}$ into $V$ $L$-dimensional simplices

$$\mathrm{SimNorm}(\boldsymbol{z}) := [\boldsymbol{g}_1, ..., \boldsymbol{g}_L], \quad \boldsymbol{g}_i = \mathrm{Softmax}(\boldsymbol{z}_{i:i+V})$$

- The key is not to make models accurate
  - But to make them smooth
  - And have a low optimality gap

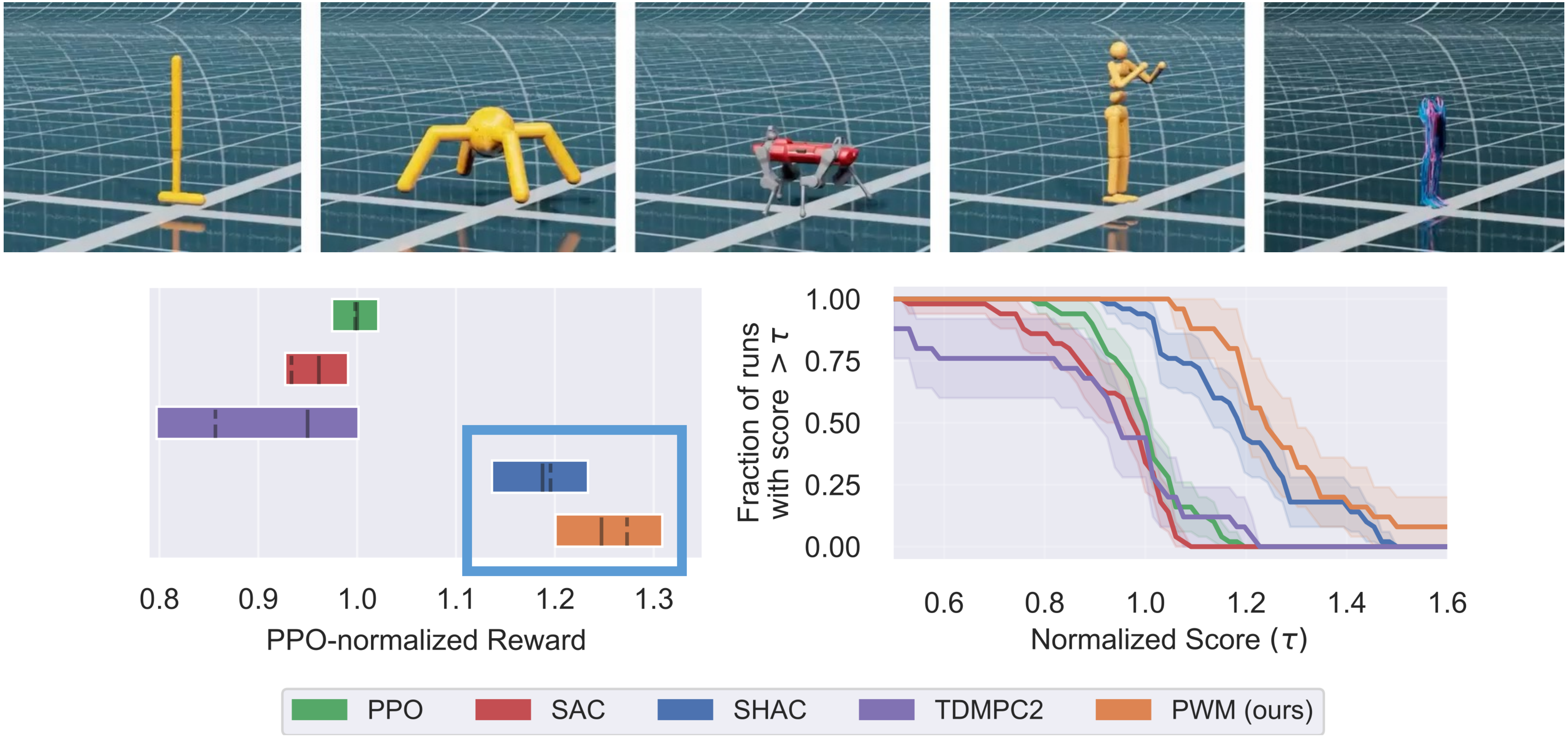



| Model | Model error | Opt. gap |
|---|---|---|
| True | 0.0 | 16.850 |
| ReLU | 0.707 | 16.046 |
| SimNorm | 1.131 | 3.473 |

**(c)** Model error and optimality gap.

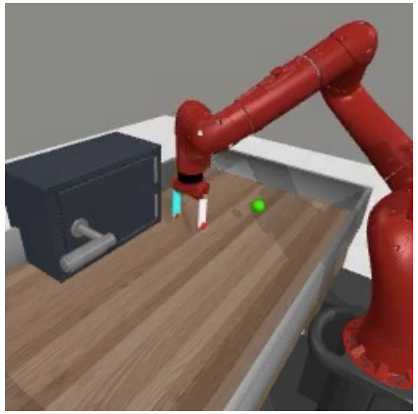

1. Regularized large models enable efficient policy learning
2. Use First-order optimization to train policies in <10m per task

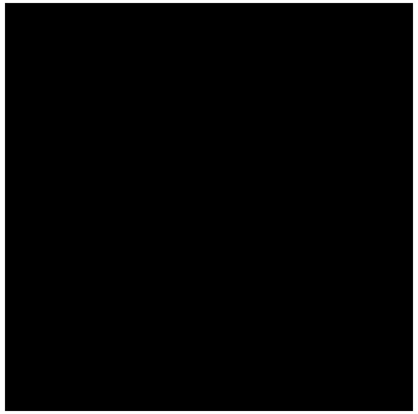
# High-dimensional single-task

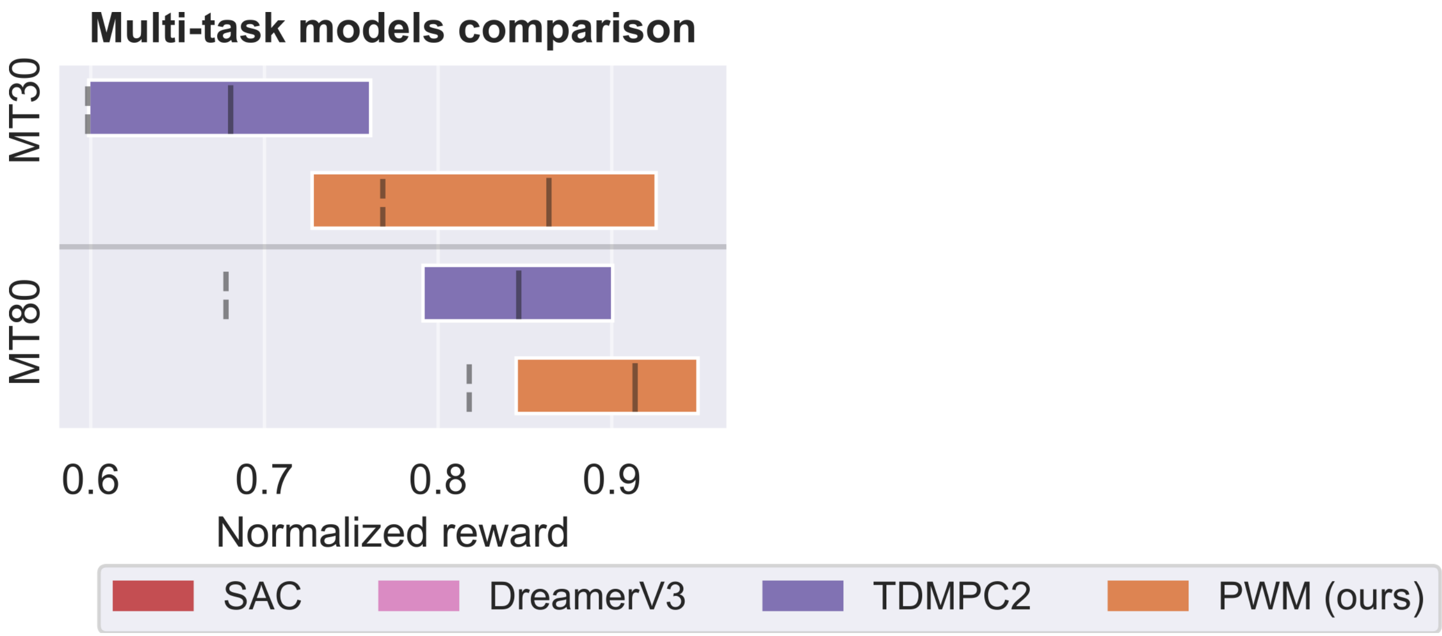



**Takeaway: optimizing over surrogate models obtains better policies than ground truth!**

# Multi-task experiments



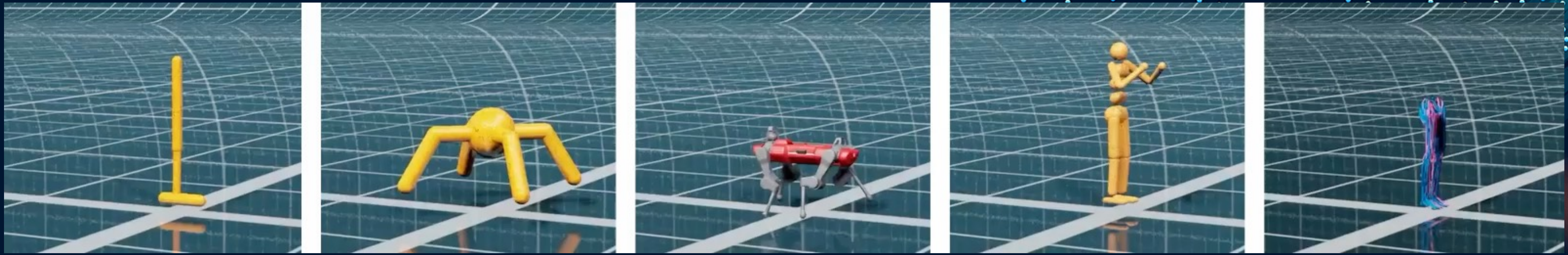


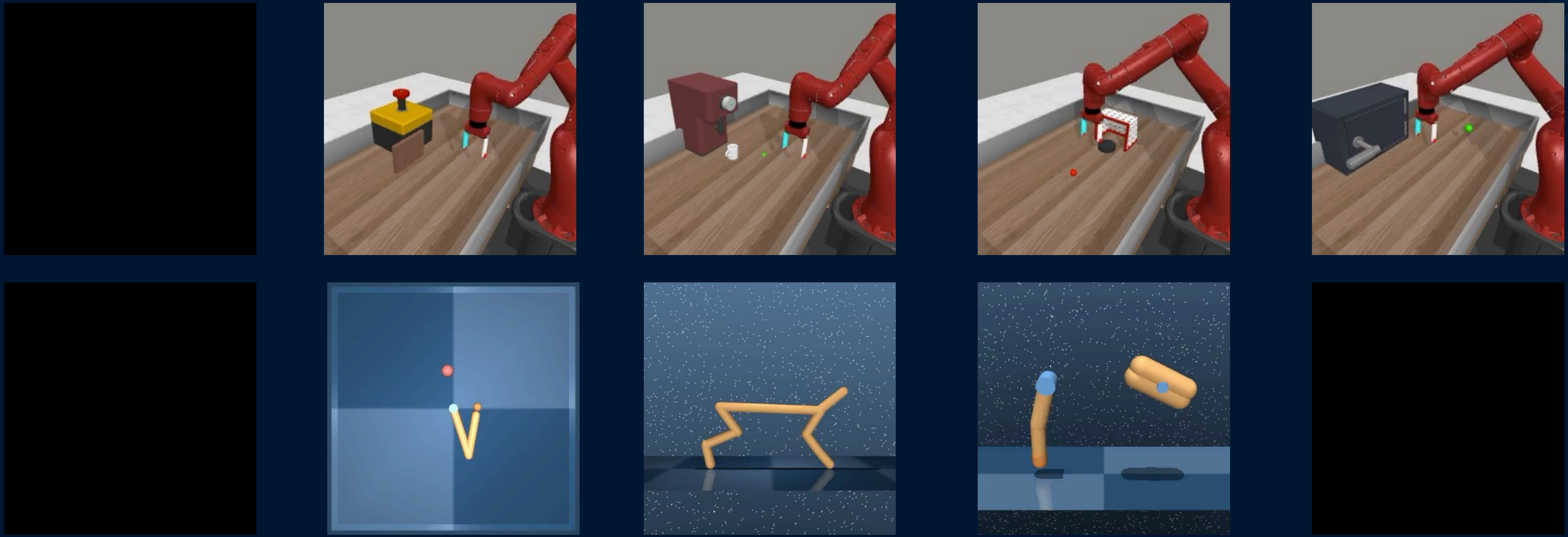



**Beats TDMPC2 without the need for online planning -> more scalable**

**Matches single-task experts without any online interaction**

# PWM learns over 80 tasks





# Generalizable Autonomy

**Structure**

**Data**

Generative AI to Enable Robotics

Innovations in better Models and larger datasets

# Generalizable Autonomy

Representations for Embodied FMs

Animesh Garg

Professor of AI Robotics
Georgia Tech

animesh.garg@gatech.edu
@animesh_garg