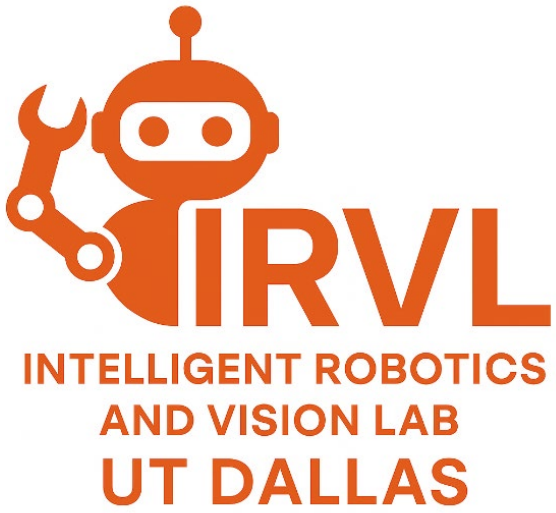
Toward Reproducible Real-World Evaluation of Robot Manipulation



Yu Xiang

Assistant Professor

Intelligent Robotics and Vision Lab

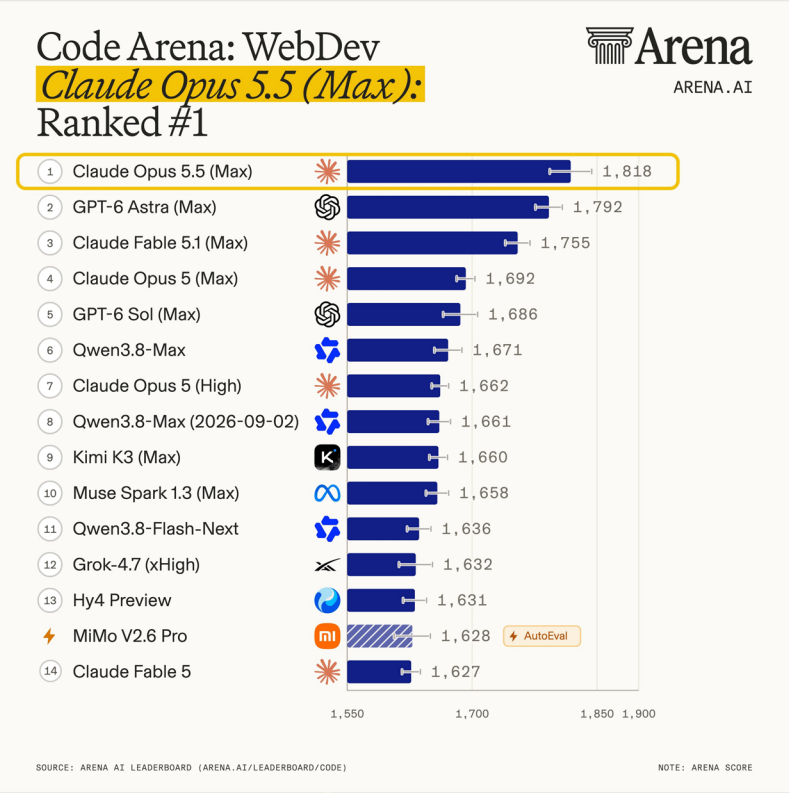
University of Texas at Dallas

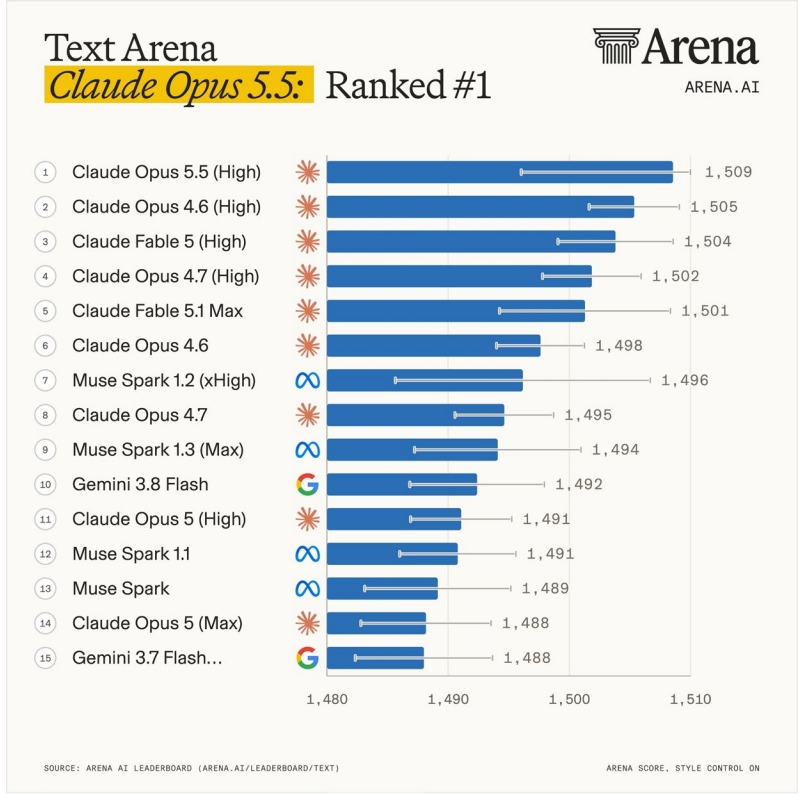
9/27/2026

Reproducing Robot Manipulation Workshop @ IROS 2026

Why do we need benchmarking?

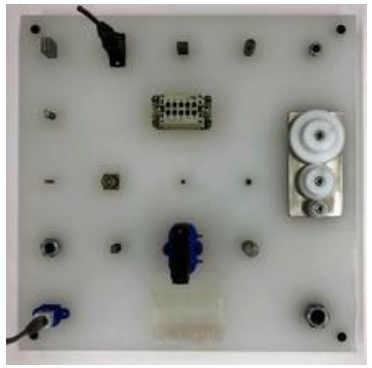
- We need to know whether we are actually making progress
  - Does it work?
  - Does it work better than existing methods?
  - Does it work beyond the conditions where it was developed?





Manipulation Benchmark

Benchmark	Type	Task	Objects	AR Tag-Free	Scene Reproducibility
Meta-World [11]	Simulation	50 tasks	Synthetic	✓	✓
RLBench [12]	Simulation	100 Tasks	Synthetic	✓	✓
robosuite [13]	Simulation	9 Tasks	Synthetic	✓	✓
Grasp Planning Protocol [10]	Real	Grasp Planning	YCB (single)	✓	✗
NIST Assembly [7]	Real	Assembly	Task Boards	✓	✓
FurnitureBench [14]	Real	Assembly	3D Printing	✗	✓
GRASPA [8]	Real	Grasping	YCB (clutter)	✗	✓
OCRTOC [15]	Real	Rearrangement	YCB + Others	✓	✗
RB2 [16]	Real	Pouring, Scooping, Zipping, Insertion	Others	✓	✗
Box and Blocks Test [17]	Real	Pick-and-Place	Blocks	✓	✗
SceneReplica (Ours)	Real	Pick-and-Place	YCB (clutter)	✓	✓



NIST Assembly board

FurnitureBench

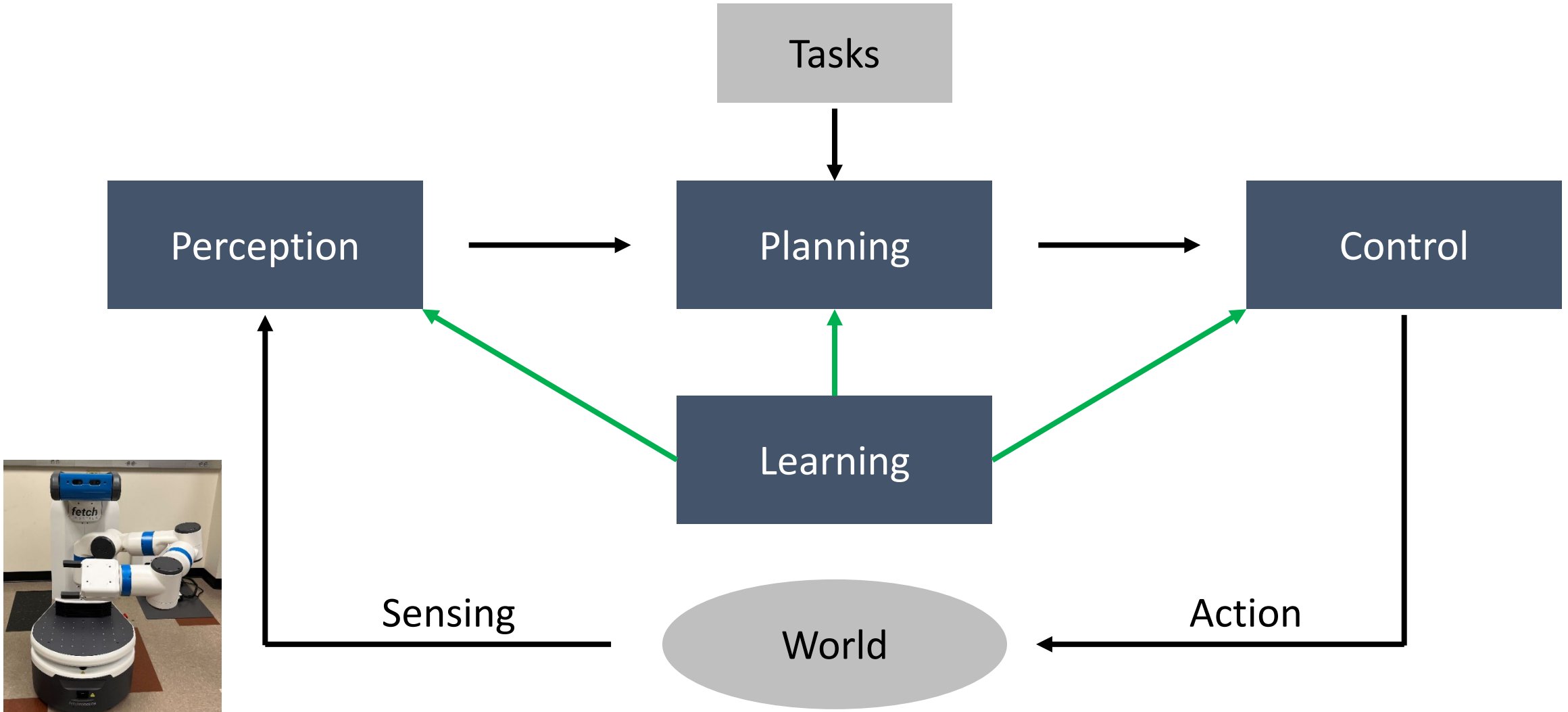
GRASPA

SceneReplica

SceneReplica: Benchmarking Real-World Robot Manipulation by Creating Reproducible Scenes
Ninad Khargonkar*, Sai Haneesh Allu*, Yangxiao Lu, Jishnu Jaykumar P, Balakrishnan Prabhakaran, Yu Xiang (*equal contribution)
In International Conference on Robotics and Automation (ICRA), 2024.

The Perception, Planning and Control Loop

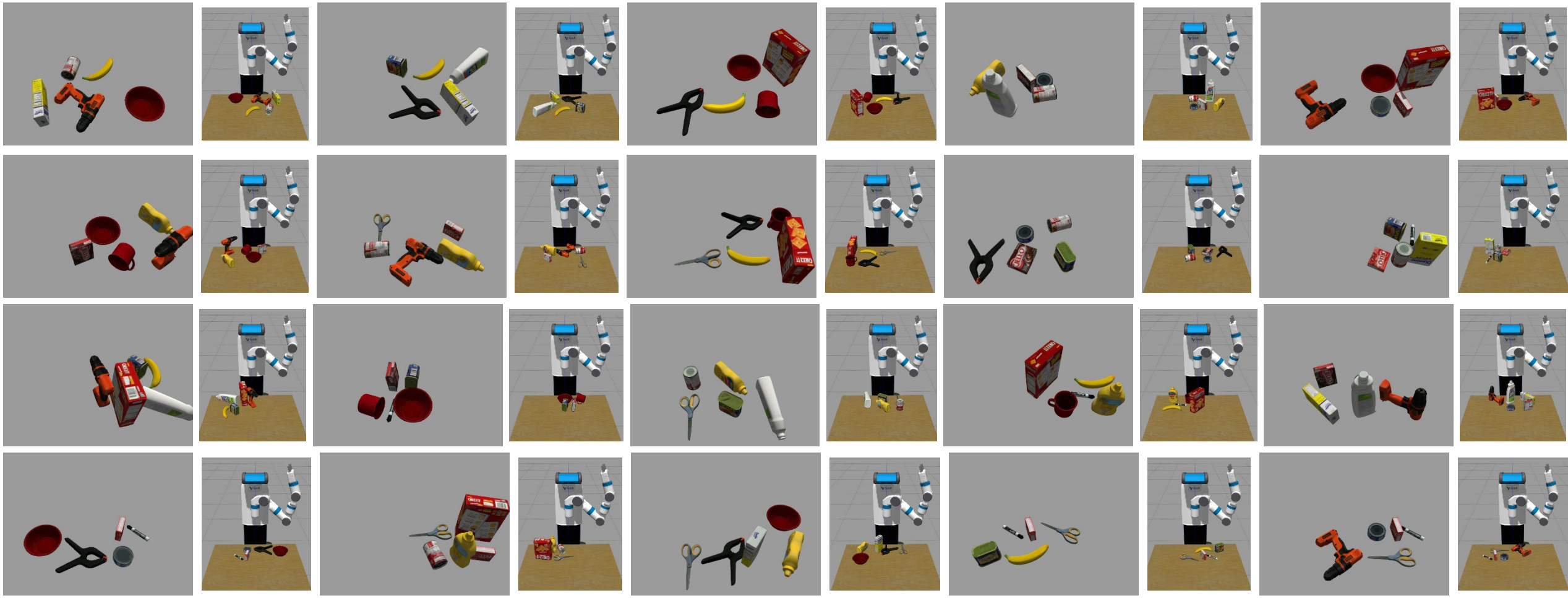
Good Old Fashioned Engineering (GOFE)



SceneReplica Benchmark

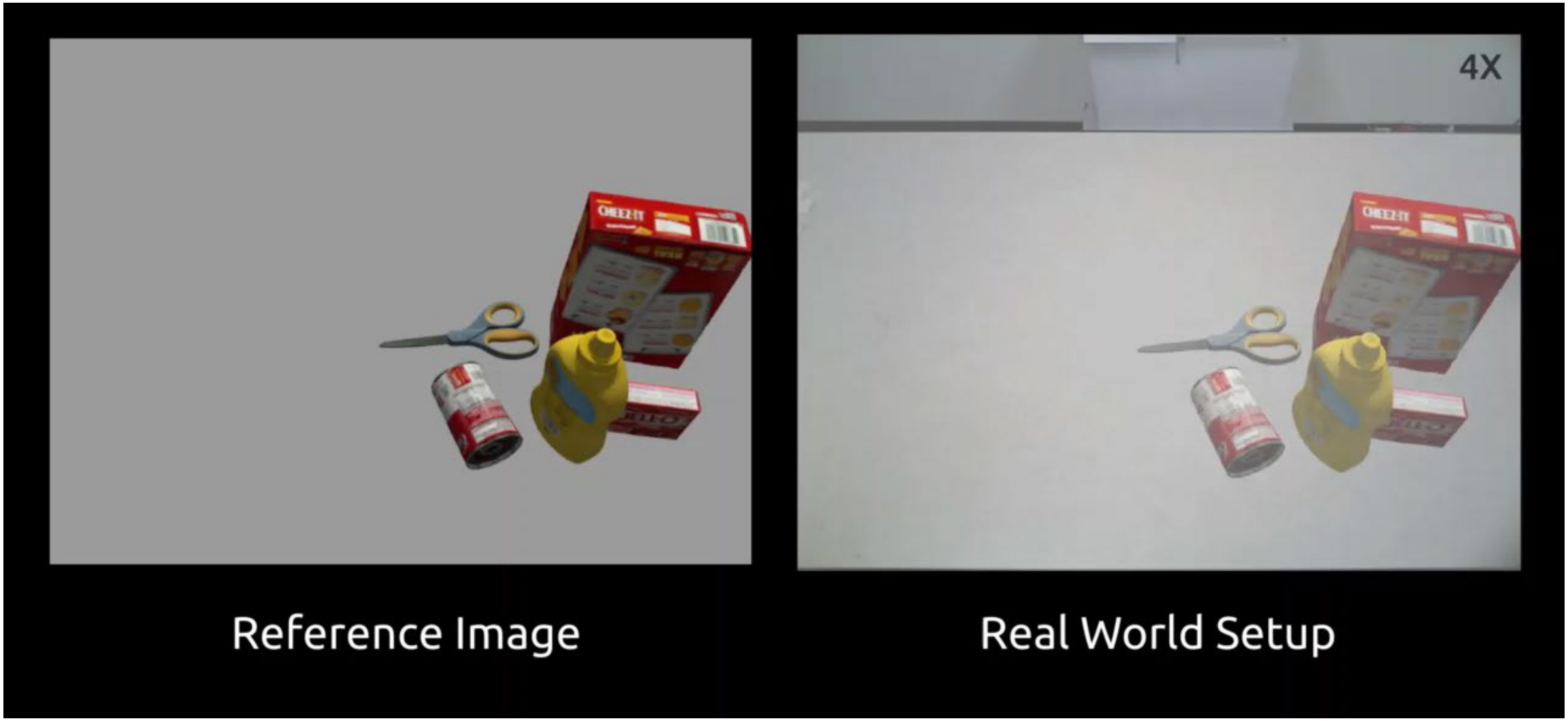
Ninad Khargonkar Sai Haneesh Allu

20 Scenes



SceneReplica: https://irvlutd.github.io/SceneReplica/

Real-World Scene Setup



Model-based Grasping vs Model-free Grasping

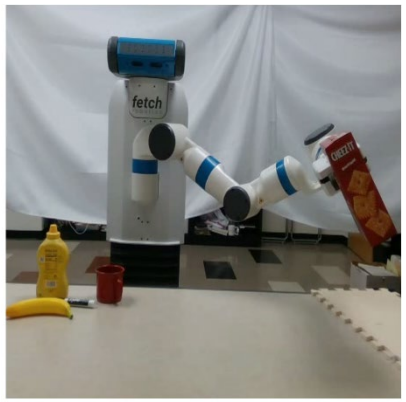
Input real world scene

6D Pose Estimation

Offline Grasp Database

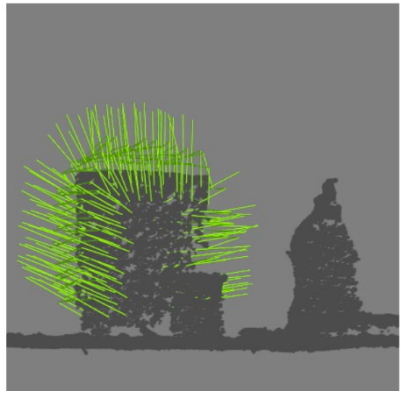
Motion Planning Setup

Grasping & Lifting

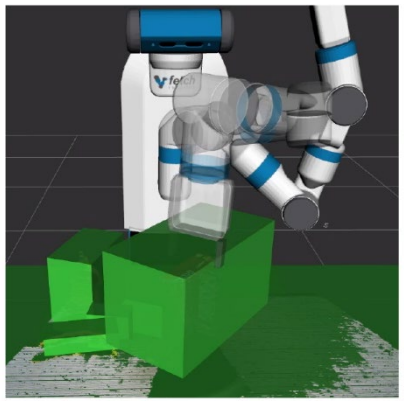


Moving arm for Dropoff

Unseen Object Segmentation

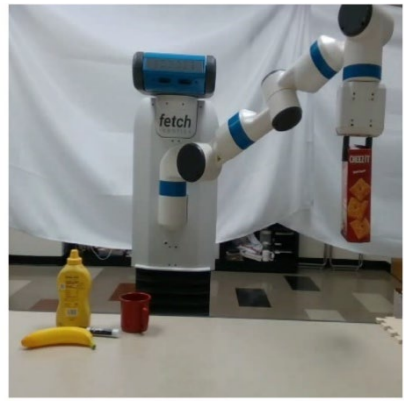


Model-free Grasp Planning



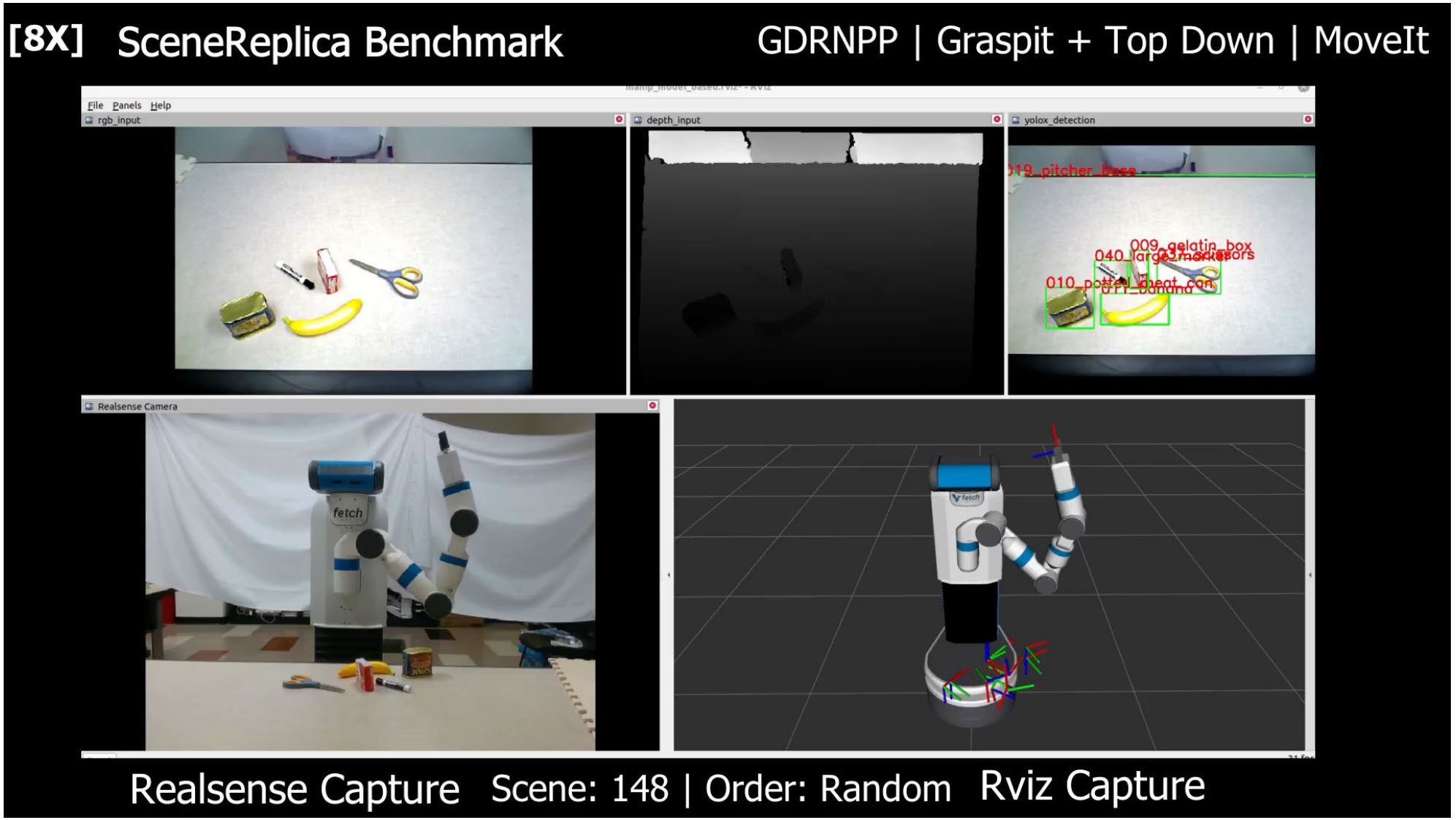
Motion Planning Setup

Grasping & Lifting

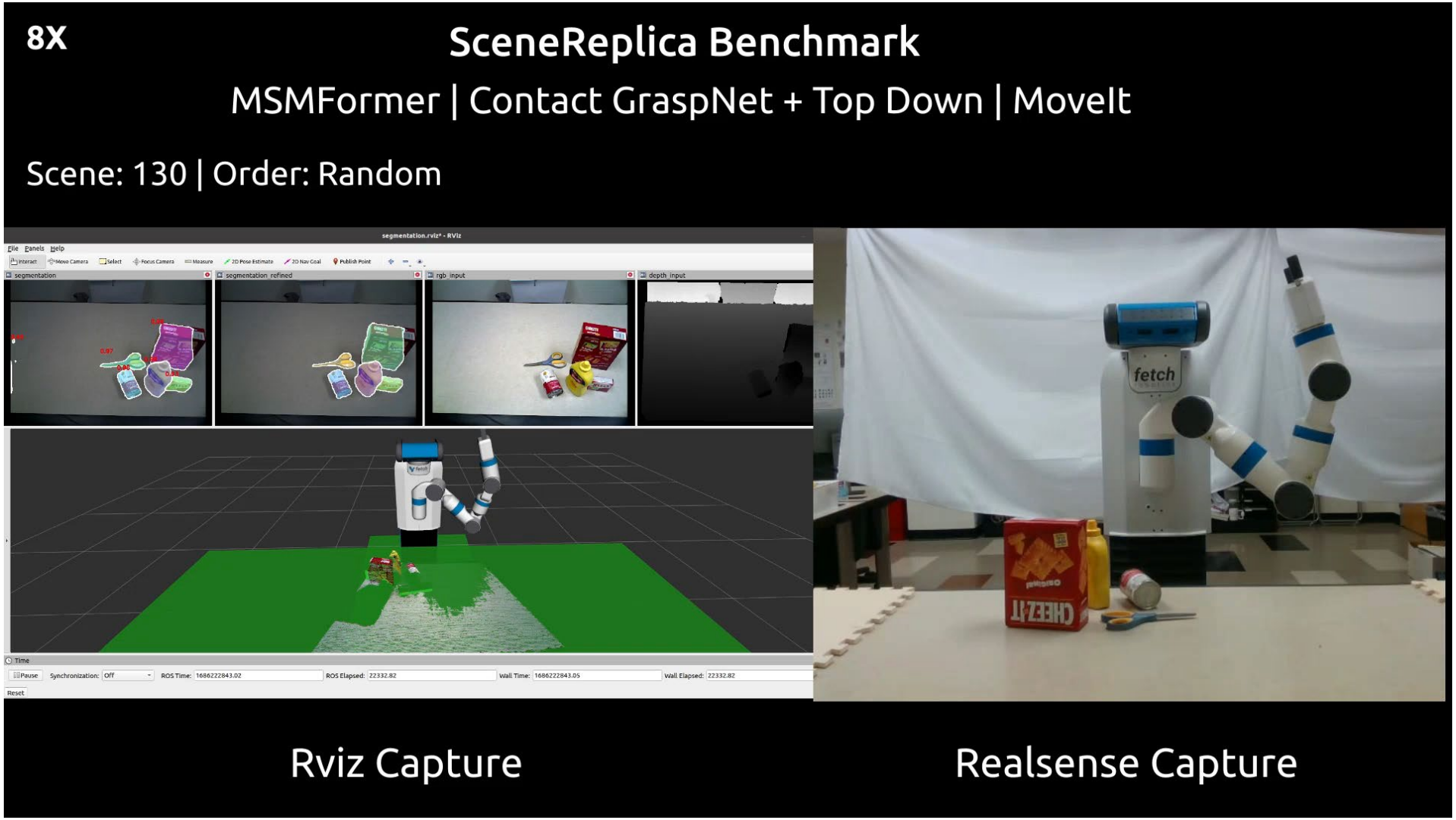


Moving arm for Dropoff

Model-based Grasping Example



Model-free Grasping Example

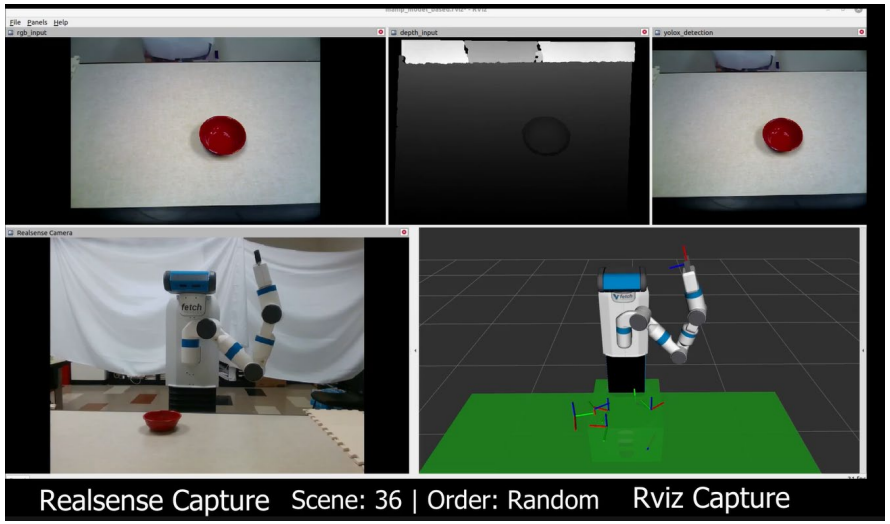


Current Leaderboard

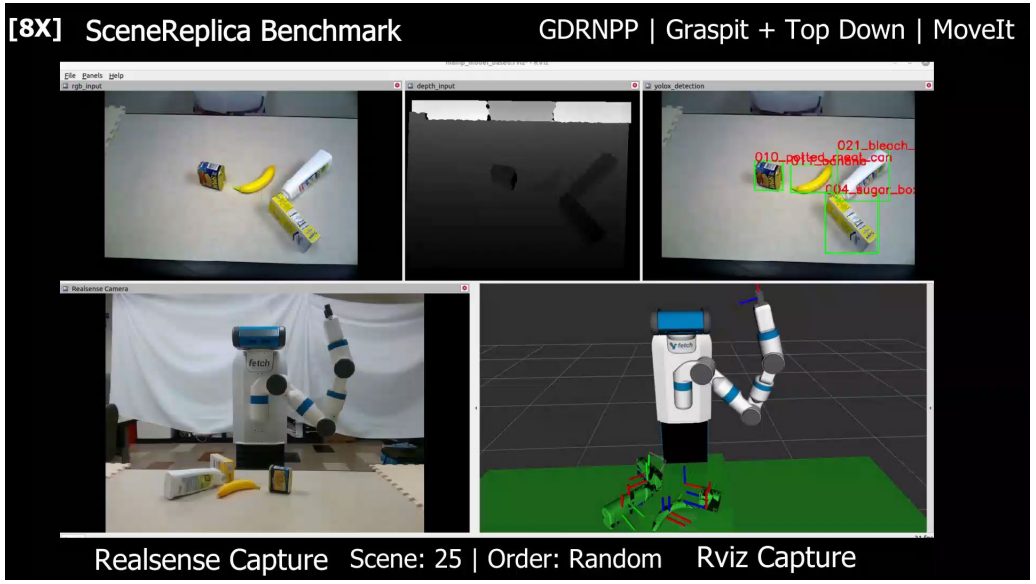
#	Perception	Grasp Planning	Motion Planning	Control	Ordering	Grasping Type	Pick & Place Success	Grasping Success	Videos
11	GDRNPP [9]	RFP [12] + Top-Down	OMPL [3]	MoveIt	Near-to-Far	Model-Based	70/100	73/100	
10	MSMFormer [8]	Contact-graspnet [7] + Top-Down	GTO [11]	MoveIt	Near-to-Far	Model-Free	65/100	71/100	
7	MSMFormer [8]	Contact-graspnet [7] + Top-Down	OMPL [3]	MoveIt	Fixed Random	Model-Free	61/100	70/100	
3	GDRNPP [9]	GraspIt! [2] + Top-Down	OMPL [3]	MoveIt	Near-to-Far	Model-Based	66/100	69/100	
7	MSMFormer [8]	Contact-graspnet [7] + Top-Down	OMPL [3]	MoveIt	Near-to-Far	Model-Free	57/100	65/100	
3	GDRNPP [9]	GraspIt! [2] + Top-Down	OMPL [3]	MoveIt	Fixed Random	Model-Based	62/100	64/100	
5	UCN [5]	Contact-graspnet [7] + Top-Down	OMPL [3]	MoveIt	Fixed Random	Model-Free	60/100	64/100	

https://irvlutd.github.io/SceneReplica/

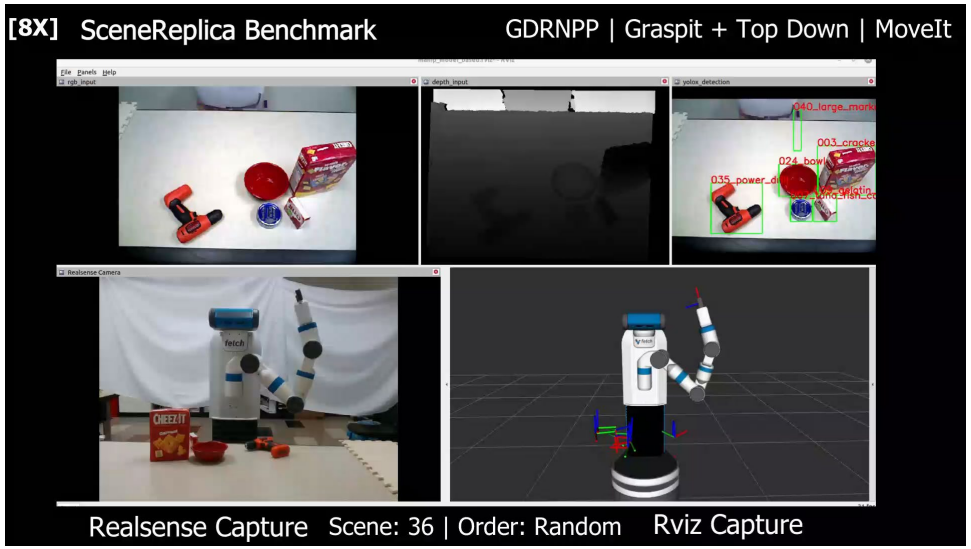
Failure Analysis (GDRNPP + GraspIt! + Top-Down)



Object Detection Error



Grasp Planning Error



Pose Estimation Error

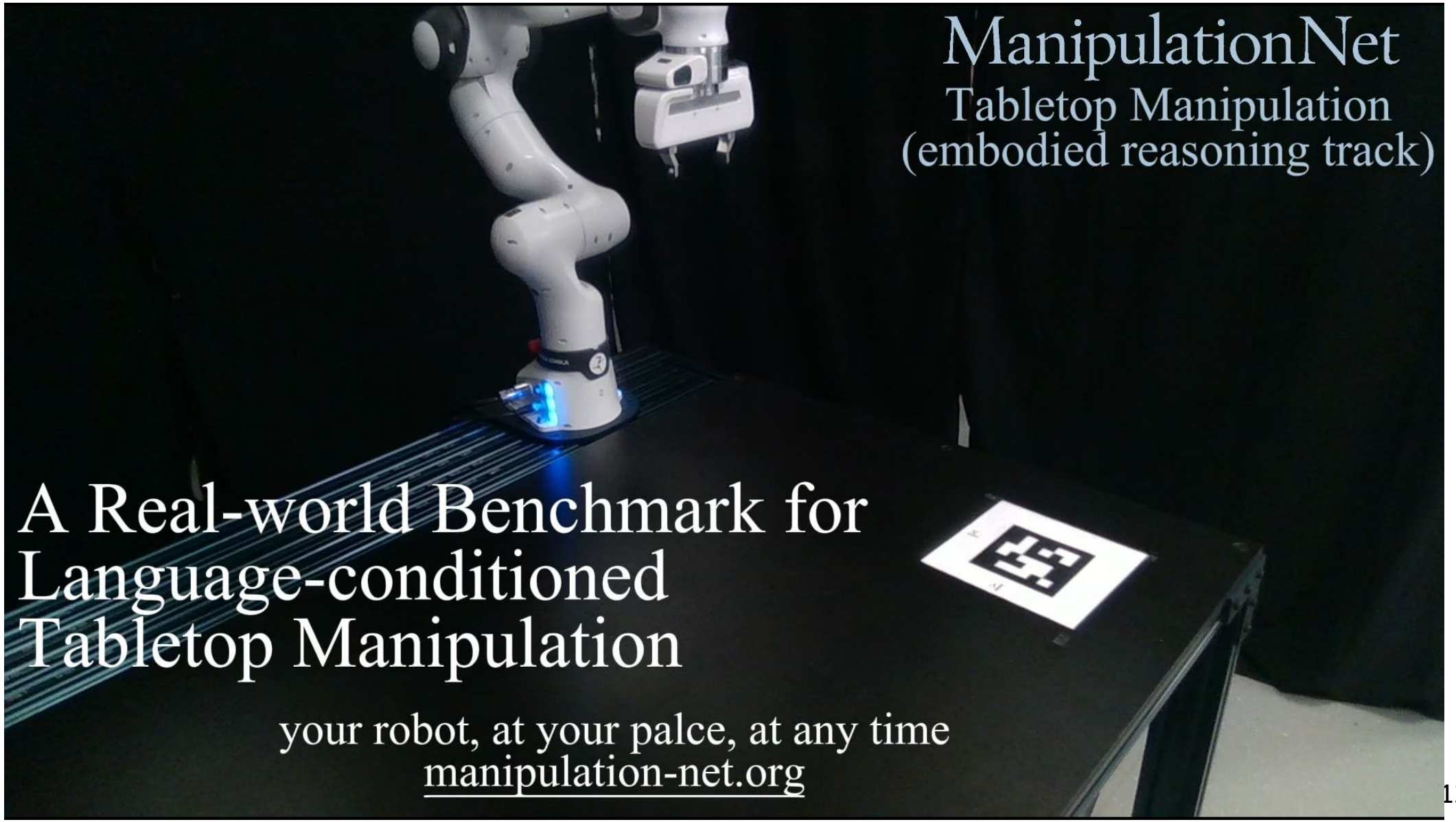
Object	Method 3			
	S	P_EF	P_LF	EF
003 cracker box	3	2	1	-
004 sugar box	5	-	-	-
005 tomato soup can	5	1	-	1
006 mustard bottle	7	-	-	-
007 tuna fish can	1	5	-	-
008 pudding box	5	-	-	-
009 gelatin box	6	-	1	-
010 potted meat can	7	-	-	-
011 banana	6	-	1	-
021 bleach cleanser	3	1	-	1
024 bowl	2	4	1	-
025 mug	4	-	1	-
037 scissors	4	3	-	-
035 power drill	1	3	2	1
040 large marker	2	4	-	-
052 extra large clamp	5	1	-	-
ALL	66	24	7	3

P_EF: #perception failure

P_LF: #planning failure

EF: #execution failure

SceneReplica Integrated into ManipulationNet



ManipulationNet Team

ASTM INTERNATIONAL · 清華大學 Tsinghua University · TEXAS The University of Texas at Austin · NVIDIA · UT DALLAS

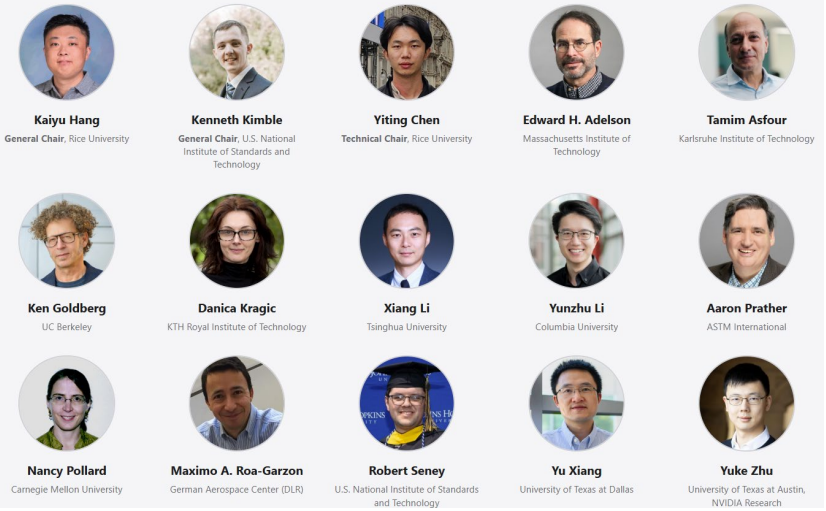
COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK · KIT Karlsruhe Institute of Technology · Carnegie Mellon University · DLR Deutsches Zentrum für Luft- und Raumfahrt German Aerospace Center · KTH Kungl Tekniska Högskolan · AUTODESK

Massachusetts Institute of Technology · RICE UNIVERSITY · Berkeley UNIVERSITY OF CALIFORNIA

A Global Infrastructure for Comparable Manipulation Research Supported By

U.S. NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY

Founding Committee

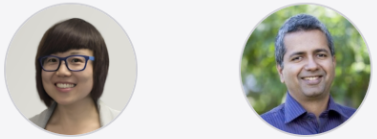


Kaiyu Hang — General Chair, Rice University
Kenneth Kimble — General Chair, U.S. National Institute of Standards and Technology
Yiting Chen — Technical Chair, Rice University
Edward H. Adelson — Massachusetts Institute of Technology
Tamim Asfour — Karlsruhe Institute of Technology
Ken Goldberg — UC Berkeley
Danica Kragic — KTH Royal Institute of Technology
Xiang Li — Tsinghua University
Yunzhu Li — Columbia University
Aaron Prather — ASTM International
Nancy Pollard — Carnegie Mellon University
Maximo A. Roa-Garzon — German Aerospace Center (DLR)
Robert Seney — U.S. National Institute of Standards and Technology
Yu Xiang — University of Texas at Dallas
Yuke Zhu — University of Texas at Austin, NVIDIA Research

Founding Developer Team

Yiting Chen — Rice University
Podshara Chanrungmaneekul — Rice University
Shihefeng Wang — Tsinghua University
Kaifeng Zhang — Columbia University
Ziyang Chen — UC Berkeley
Yash Chitambar — UC Berkeley
Shuo Sha — Columbia University

Senior Advisory Board

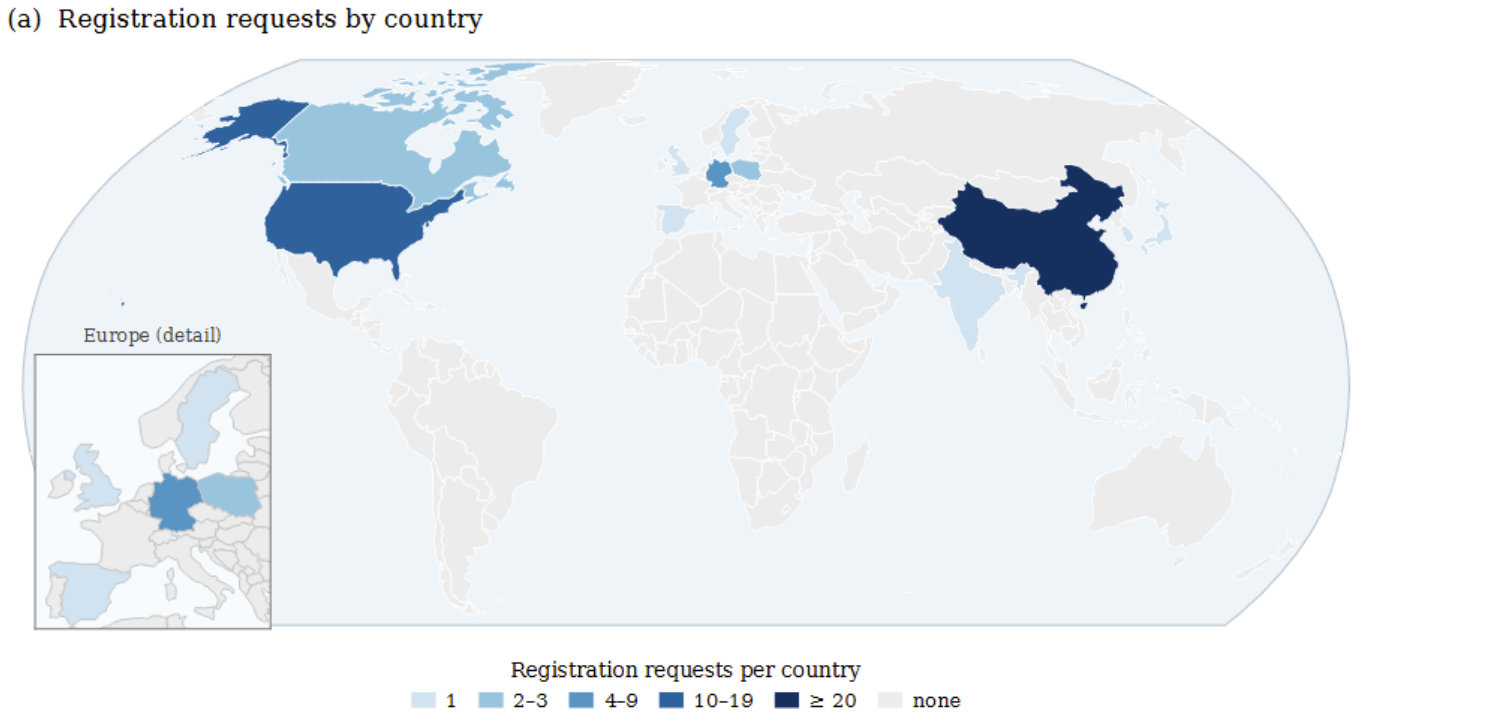


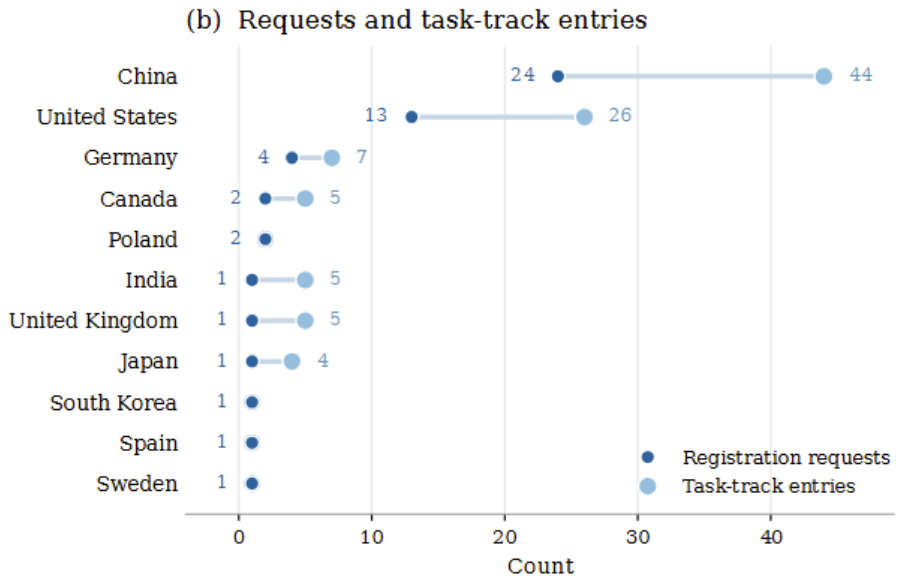
Hui Li — Autodesk
Sachin Chitta — Autodesk

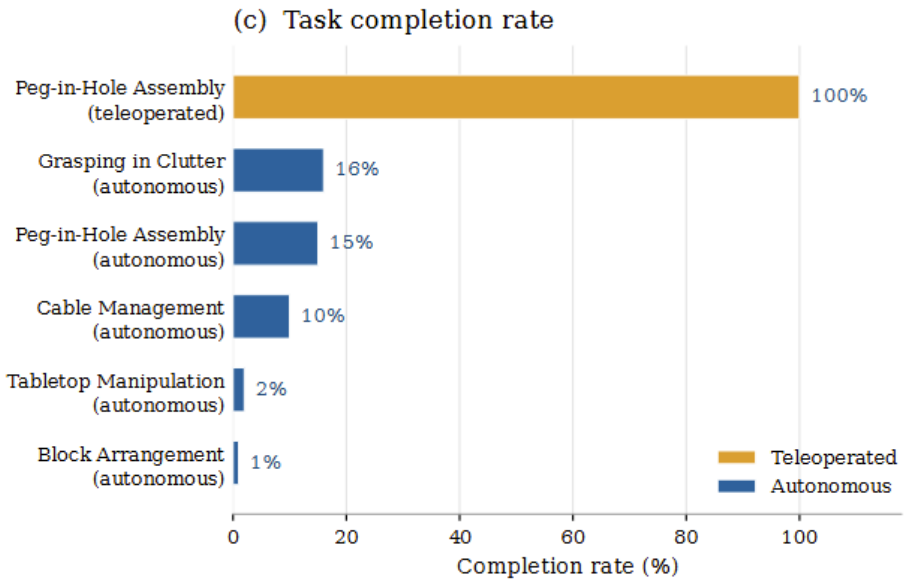
https://manipulation-net.org/

Progress

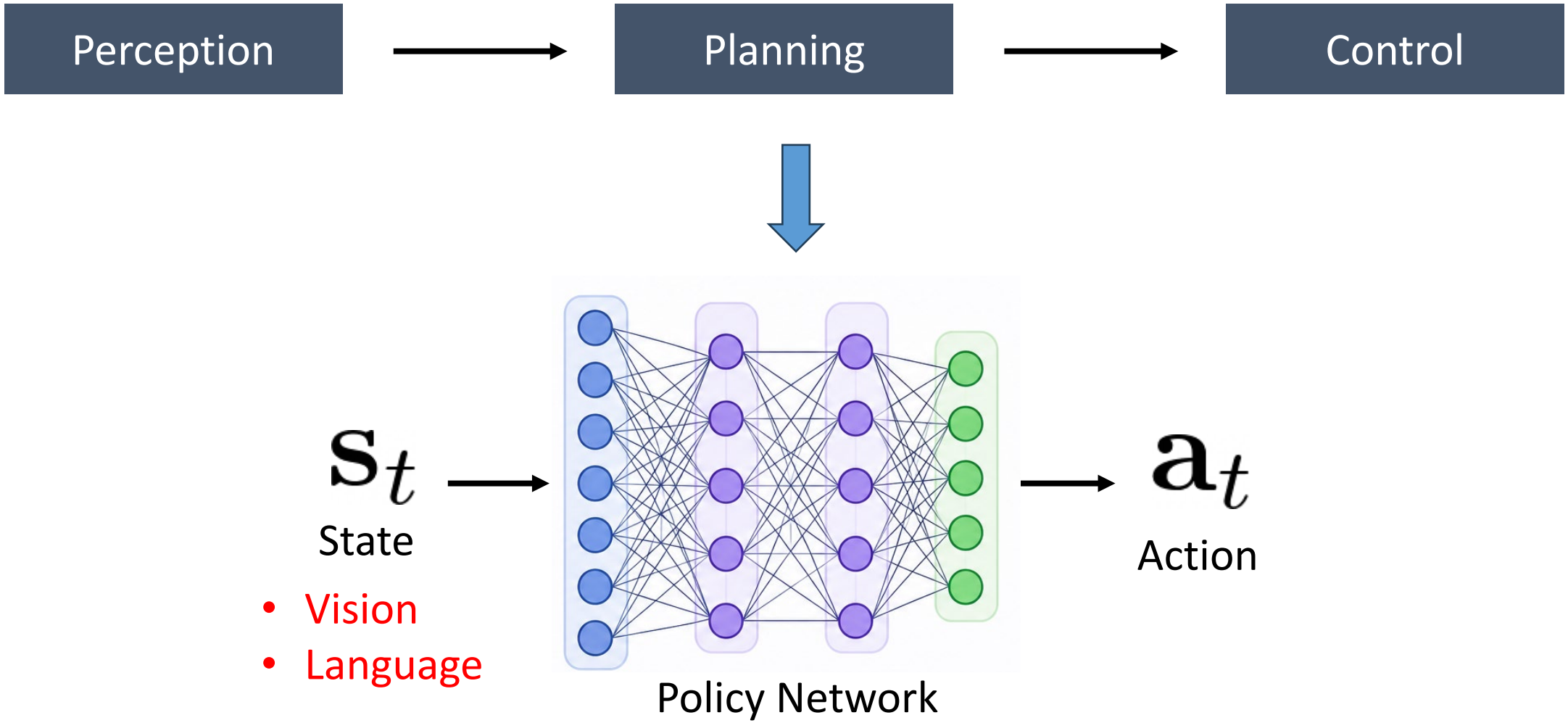
- 3 continents
- 10+ countries
- 100+ task registrations







End-to-End Manipulation

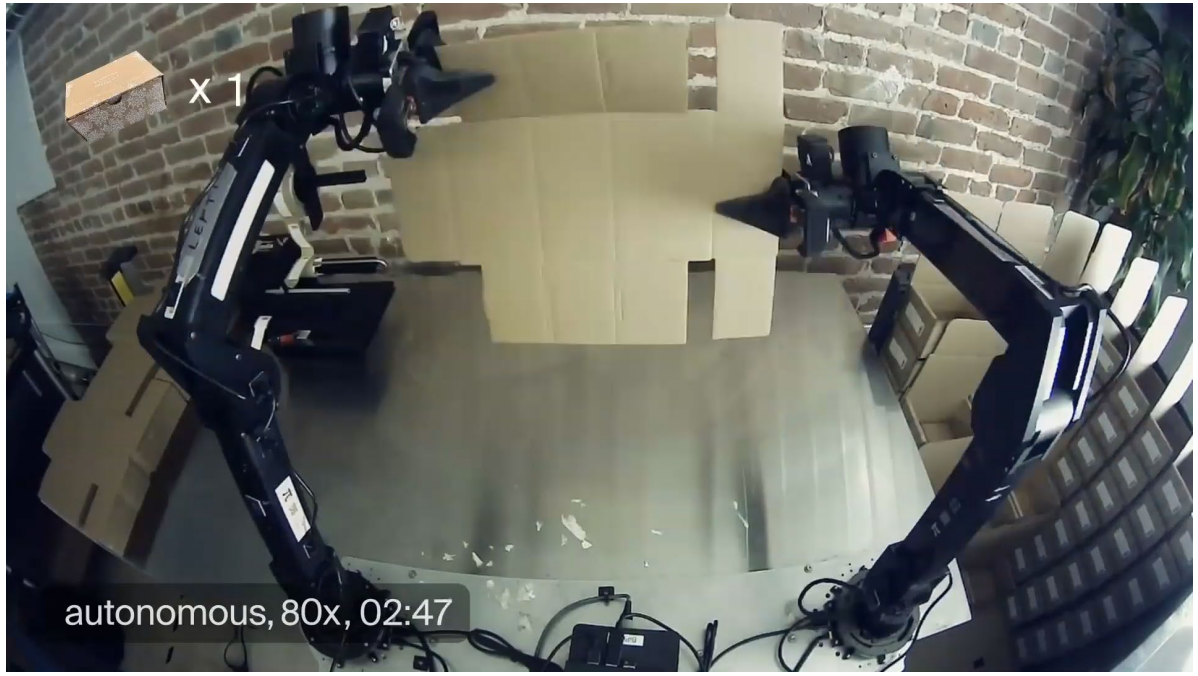


- Vision
- Language

Vision-Language-Action (VLA) Models

Some Recent Breakthroughs

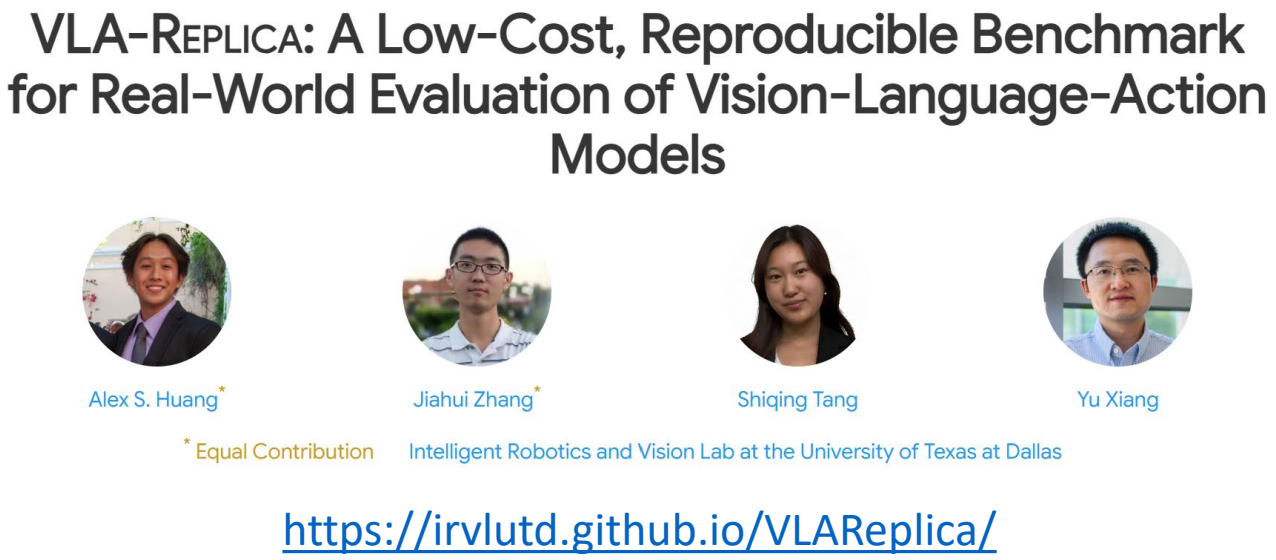
Sunday Robotics
https://www.sunday.ai/blog/act-2-preview



Physical Intelligence
https://www.pi.website/blog/pistar06

How Good are these VLA Models?

- We need benchmarking



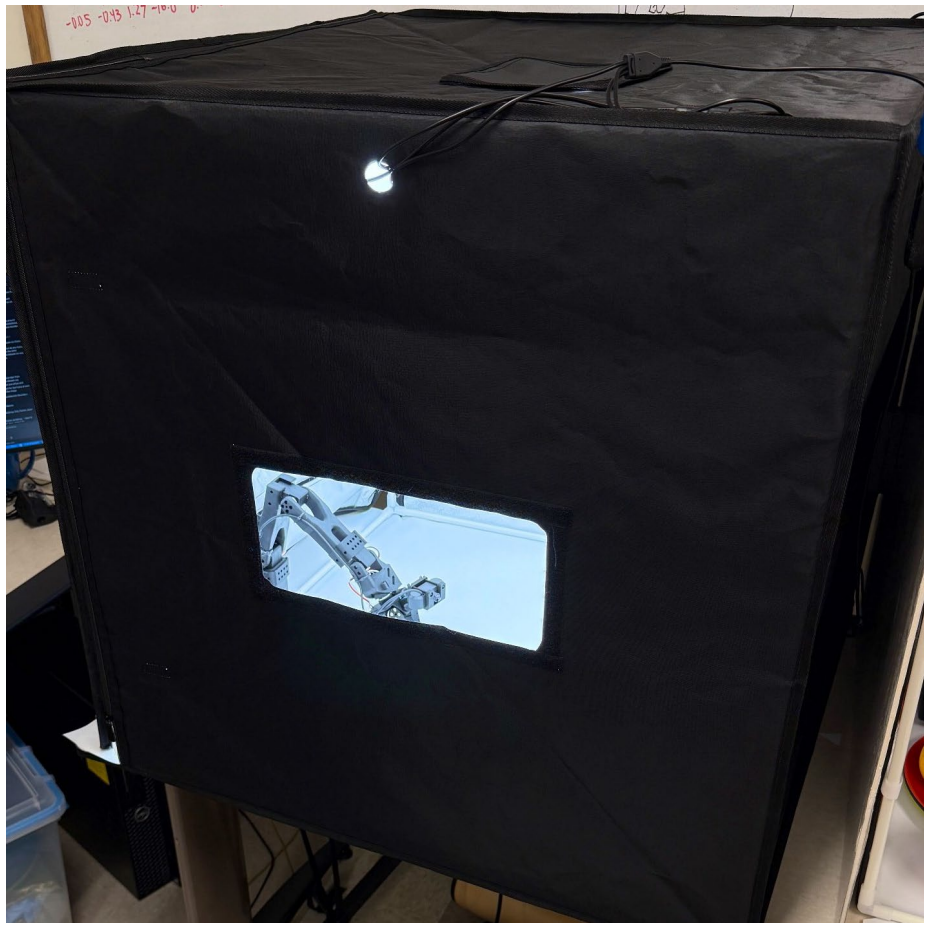
https://irvlutd.github.io/VLAReplica/

In NeurIPS Track on Datasets and Benchmarks, 2026.

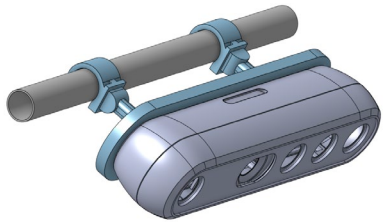
VLA-Replica Benchmark

Low-cost Components

SO-101 Arm ~$200

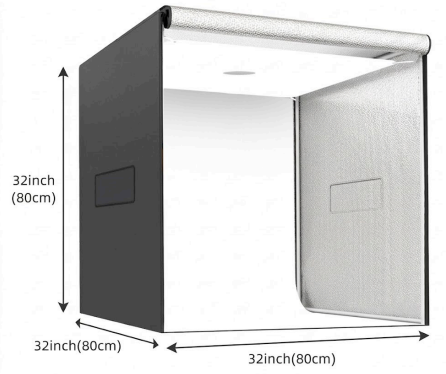


Real Set-up



D455 Camera ~$425

Webcam $13.98



Light Box $152.99

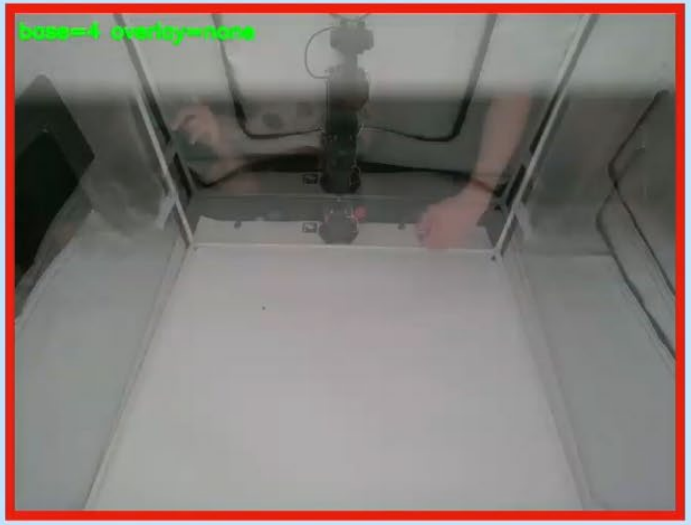
VLA-Replica Benchmark

VLA-Replica Benchmark

2X

Camera Calibration

Using reference images for camera calibration



Top Camera Calibration

VLA-Replica Benchmark

2X

Task Evaluation

Task 1: Fold the pink towel in half.

Reference Placement

Align placement with reference frame

VLA-Replica Benchmark: Tasks

We provide:

- 500 demos across 10 tasks
- 10 In-distribution tasks
- 8 Out-of-distribution tasks
- 5 reference frames per task

VLA-Replica Benchmark: Tasks

Task #	Task Type	Training & ID Eval Task Example	OOD Eval Task Example
1	Pick-and-Place	Put bread in **red**/**blue** plate	Put bread on **yellow** plate
2		Put **red** bowl on **purple** coaster	Put **green** bowl on **yellow** coaster
3		Stack **blue** cube on **red** cube	Stack **green** cube on **green** cube
4		Put all **2 or 3** blocks in **blue** box	Put all **3 or 4** blocks in **pink** box
5	Object Interaction	Fold **pink**/**yellow** towel in half	Fold **blue** towel in half
6		Open oven	N/A
7		Clean whiteboard with eraser	N/A
8	Memory	Pour pepper **1,2 or 3** times to **red** plate	Pour pepper **4 or 5** times to **blue** plate
9		Lift **green**/**red**/**blue** bowl **1 or 3** times	Lift **yellow** bowl **2, 4 or 5** times
10		Press button **1 or 3** times	Press button **2, 4 or 5** times

VLA-Replica Benchmark: Leaderboard

VLA-Replica-ID | VLA-Replica-OOD

3 runs eval with std

Rank	Method	Average	Put bread on plate	Put bowl on coaster	Stack block on block	Put all blocks into box	Fold towel	Open oven	Erase whiteboard	Shake pepper times
1	$\pi_{0.5}$ [6]	**0.53 ± 0.05**	0.80 ± 0.20	0.87 ± 0.12	0.33 ± 0.12	0.47 ± 0.12	1.00 ± 0.00	0.53 ± 0.12	0.33 ± 0.12	0.33 ± 0.
2	GR00T N1.7 [8]	**0.42 ± 0.06**	0.80 ± 0.00	0.87 ± 0.12	0.00 ± 0.00	0.29 ± 0.08	1.00 ± 0.00	0.47 ± 0.31	0.40 ± 0.20	0.20 ± 0.
3	MolmoAct2-SO100_101 [7]	**0.39 ± 0.06**	0.93 ± 0.12	0.73 ± 0.12	0.27 ± 0.12	0.27 ± 0.12	0.73 ± 0.12	0.13 ± 0.23	0.53 ± 0.12	0.13 ± 0.
4	π_0 [5]	**0.32 ± 0.03**	0.67 ± 0.12	0.67 ± 0.31	0.00 ± 0.00	0.00 ± 0.00	0.67 ± 0.23	0.20 ± 0.00	0.33 ± 0.12	0.27 ± 0.
5	SmolVLA [3]	**0.29 ± 0.04**	0.80 ± 0.20	0.20 ± 0.20	0.27 ± 0.12	0.00 ± 0.00	0.73 ± 0.12	0.33 ± 0.12	0.20 ± 0.00	0.00 ± 0.
6	ACT [1]	**0.21 ± 0.08**	0.33 ± 0.12	0.07 ± 0.12	0.00 ± 0.00	0.00 ± 0.00	0.60 ± 0.20	0.33 ± 0.12	0.20 ± 0.00	0.27 ± 0.
7	X-VLA [4]	**0.13 ± 0.02**	0.27 ± 0.12	0.13 ± 0.12	0.00 ± 0.00	0.00 ± 0.00	0.53 ± 0.12	0.07 ± 0.12	0.00 ± 0.00	0.20 ± 0.
8	DiT-D [2]	**0.11 ± 0.05**	0.20 ± 0.20	0.13 ± 0.12	0.00 ± 0.00	0.07 ± 0.12	0.20 ± 0.00	0.20 ± 0.35	0.20 ± 0.00	0.07 ± 0.
9	DiT-F [2]	**0.07 ± 0.05**	0.13 ± 0.23	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.27 ± 0.12	0.13 ± 0.23	0.13 ± 0.12	0.00 ± 0.

https://irvlutd.github.io/VLAReplica/

VLA-Replica Benchmark: Leaderboard

VLA-Replica-ID | **VLA-Replica-OOD**

3 runs eval with std

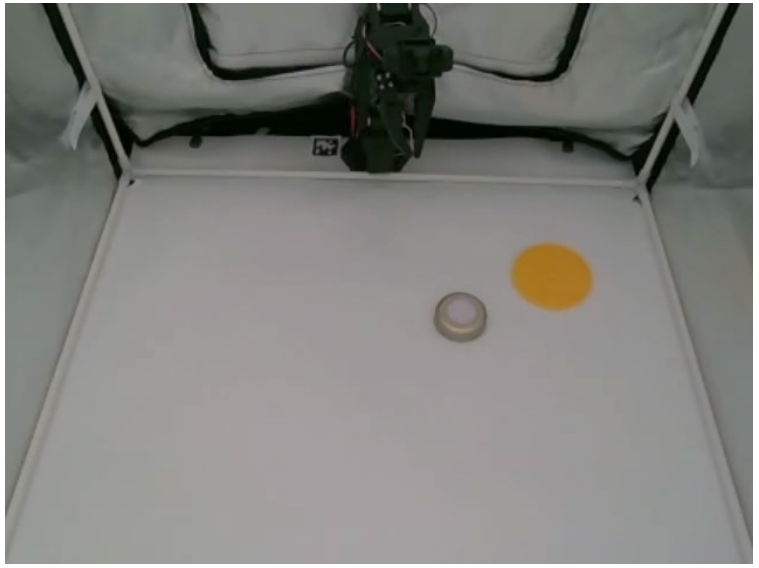
Rank	Method	Average	Put bread on plate	Put bowl on coaster	Stack block on block	Put all blocks into box	Fold towel	Shake pepper n times	Lift bowl n times	Press button n times
1	MolmoAct2-SO100_101 [7]	**0.47 ± 0.05**	0.87 ± 0.12	0.87 ± 0.12	0.60 ± 0.00	0.20 ± 0.20	0.87 ± 0.12	0.13 ± 0.12	0.13 ± 0.12	0.07 ± 0.
2	$\pi_{0.5}$ [6]	**0.39 ± 0.10**	0.93 ± 0.12	0.60 ± 0.20	0.20 ± 0.20	0.27 ± 0.12	0.80 ± 0.00	0.27 ± 0.23	0.00 ± 0.00	0.07 ± 0.
3	GR00T N1.7 [8]	**0.35 ± 0.00**	0.93 ± 0.12	0.67 ± 0.12	0.00 ± 0.00	0.13 ± 0.12	1.00 ± 0.00	0.00 ± 0.00	0.07 ± 0.12	0.00 ± 0.
4	π_0 [5]	**0.30 ± 0.02**	0.73 ± 0.12	0.67 ± 0.12	0.20 ± 0.20	0.00 ± 0.00	0.53 ± 0.06	0.22 ± 0.03	0.00 ± 0.00	0.07 ± 0.
5	SmolVLA [3]	**0.27 ± 0.04**	0.73 ± 0.12	0.33 ± 0.12	0.20 ± 0.20	0.13 ± 0.12	0.47 ± 0.12	0.00 ± 0.00	0.20 ± 0.00	0.07 ± 0.
6	ACT [1]	**0.08 ± 0.01**	0.40 ± 0.20	0.27 ± 0.12	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.
7	DiT-D [2]	**0.08 ± 0.04**	0.07 ± 0.12	0.33 ± 0.23	0.00 ± 0.00	0.00 ± 0.00	0.07 ± 0.12	0.07 ± 0.12	0.00 ± 0.00	0.07 ± 0.
8	X-VLA [4]	**0.05 ± 0.04**	0.33 ± 0.31	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.07 ± 0.12	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.
9	DiT-F [2]	**0.04 ± 0.01**	0.13 ± 0.12	0.07 ± 0.12	0.00 ± 0.00	0.07 ± 0.12	0.07 ± 0.12	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.

https://irvlutd.github.io/VLAReplica/

VLA-Replica Benchmark: Pi0.5 Behaviors

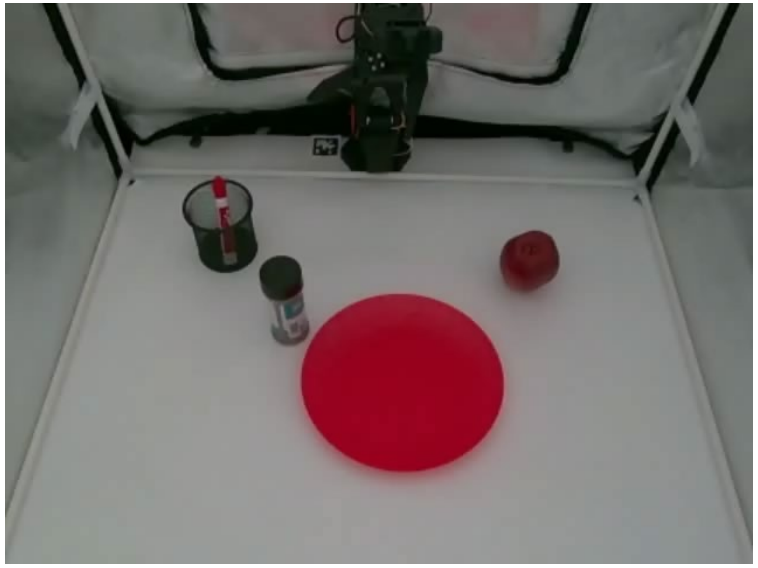
Stack the **yellow** block on the **blue** block

Precise manipulation is challenging



Press the button **three** times

Memory task is challenging



Pour **1** shake of pepper into the **red** plate

Failure recovery

Discussion: Integrating Components Together

- Modular pipelines
  - Perception-planning-control (ROS, ROS2)
  - Interpretability and debugging
  - Easier safety and constraint enforcement
  - Error accumulation
  - High latency usually, hard for closed-loop execution
- End-to-end manipulation
  - Training-testing (LeRobot)
  - Closed-loop manipulation
  - Failure recovery
  - Better for contact-rich dexterity
  - Poor interpretability and generalization
  - Hard to debug (More data? Better model?)

IRVL
INTELLIGENT ROBOTICS AND VISION LAB
UT DALLAS

NSF DARPA SONY NVIDIA XPENG FRONTIER MINING

https://labs.utdallas.edu/irvl/

Assisted by
Mr. Garvin R. Solomon

Thank you!