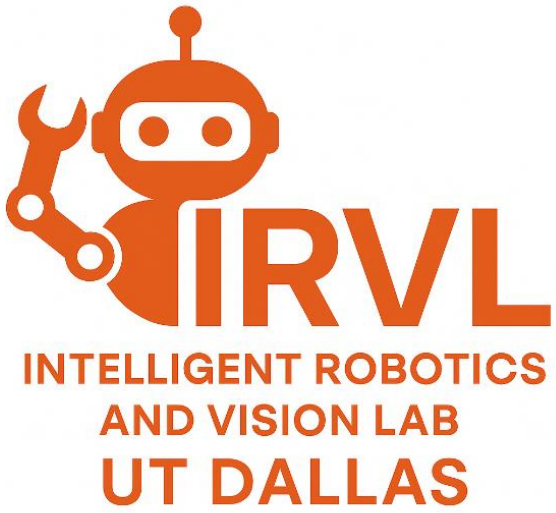
From Modular Robotics Pipelines to Vision-Language-Action Systems: Lessons from Real-World Manipulation



Yu Xiang

Assistant Professor

Intelligent Robotics and Vision Lab

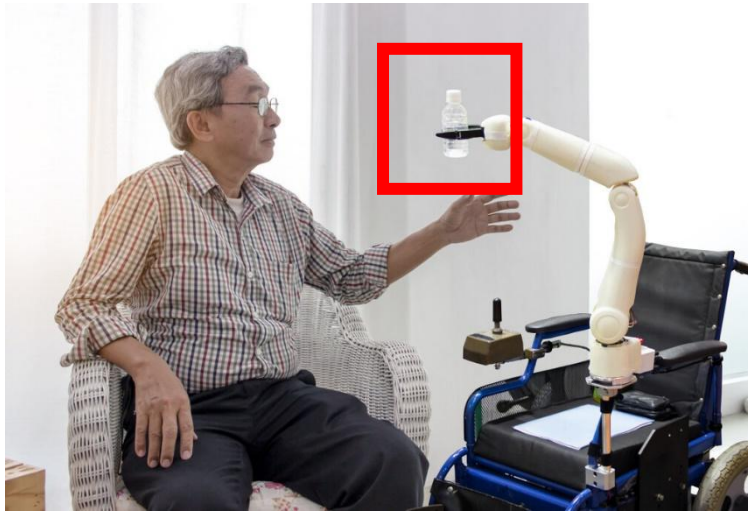
University of Texas at Dallas

6/1/2026

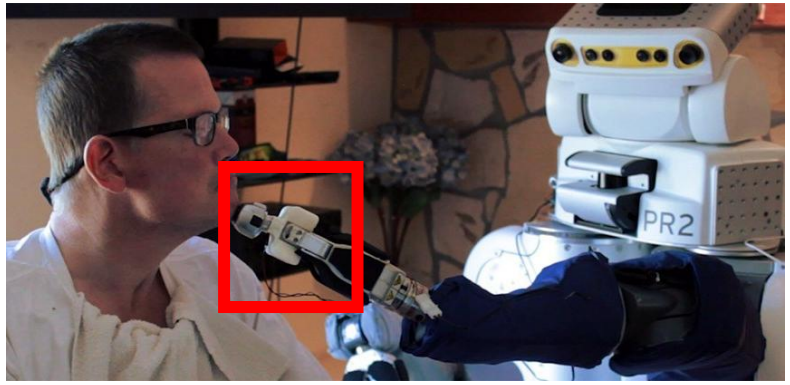
RIGOROUS Robot Perception @ ICRA 2026

Future Intelligent Robots in Human Environments

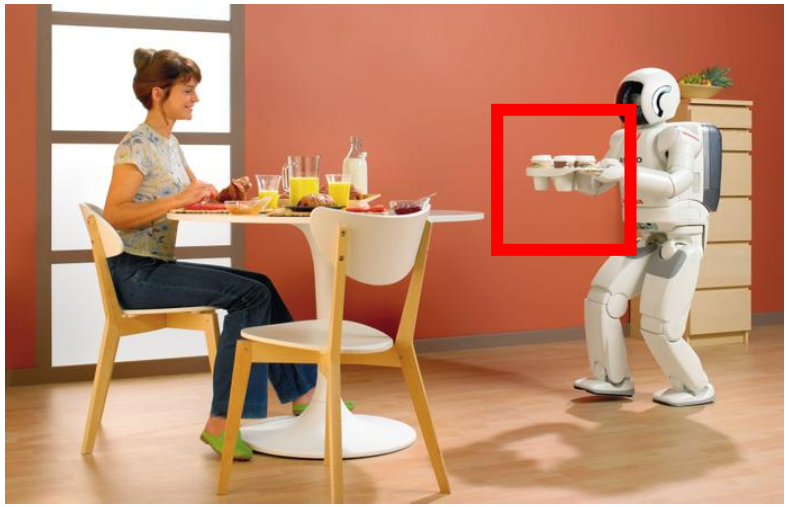
Manipulation



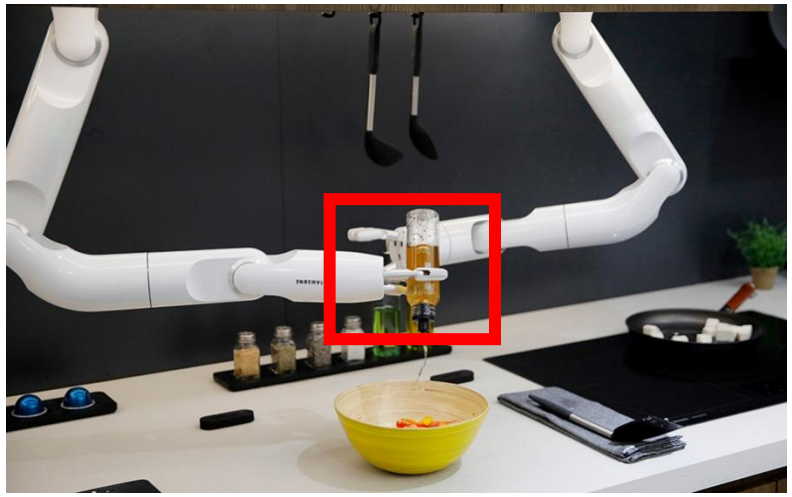
Senior Care



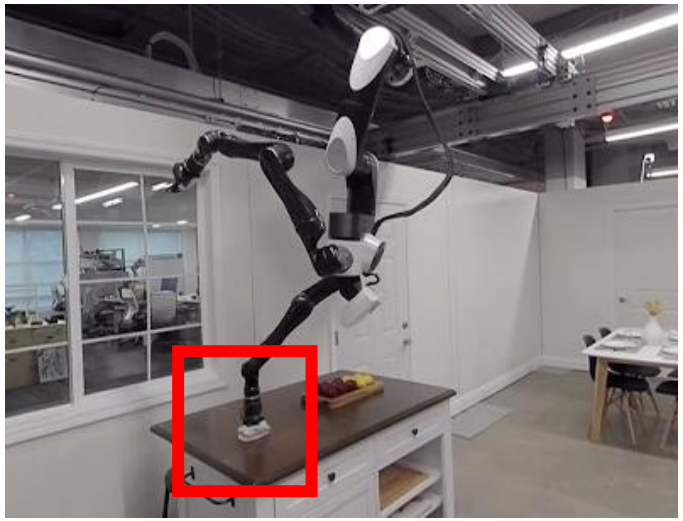
Assisting



Serving



Cooking



Cleaning

Dish washing

The Perception, Planning and Control Loop

Good Old Fashioned Engineering (GOFE)

How to Represent Objects?

How to Represent Objects?

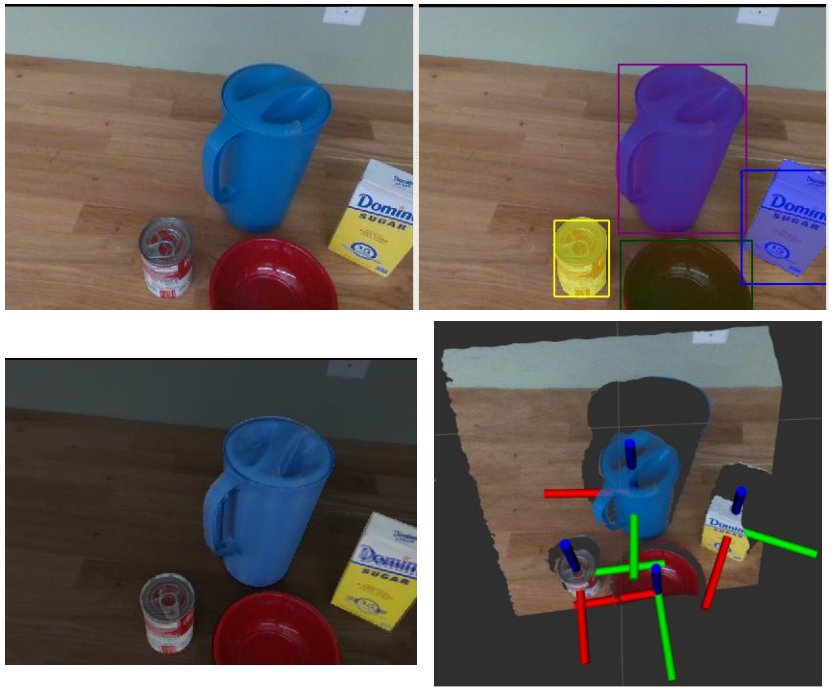
- 3D CAD models (Model-based)

- Point clouds (Model-free)

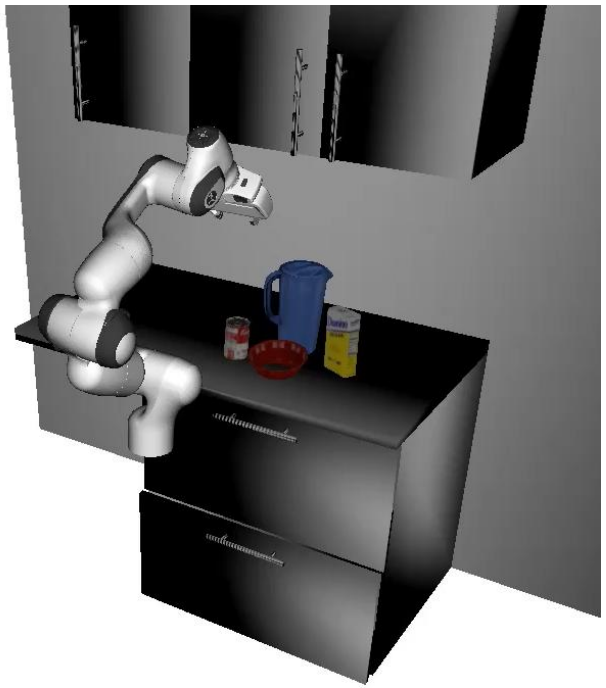
Using 3D Object Models

Perception → Planning → Control

6D object pose estimation

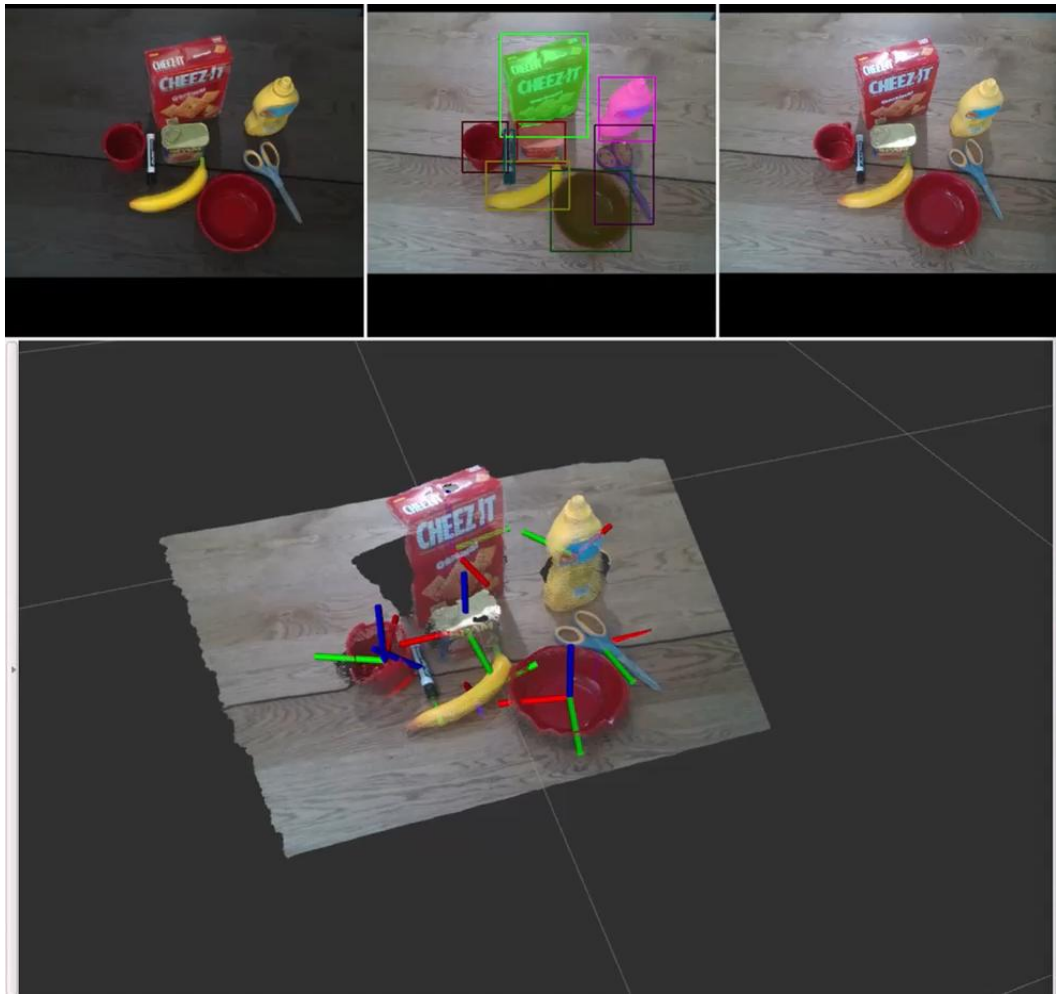


Grasp planning and motion planning



Manipulation trajectory following

6D Object Pose Estimation



FoundationPose: Unified 6D Pose Estimation and Tracking of Novel Objects

Bowen Wen, Wei Yang, Jan Kautz, Stan Birchfield

NVIDIA

CVPR 2024

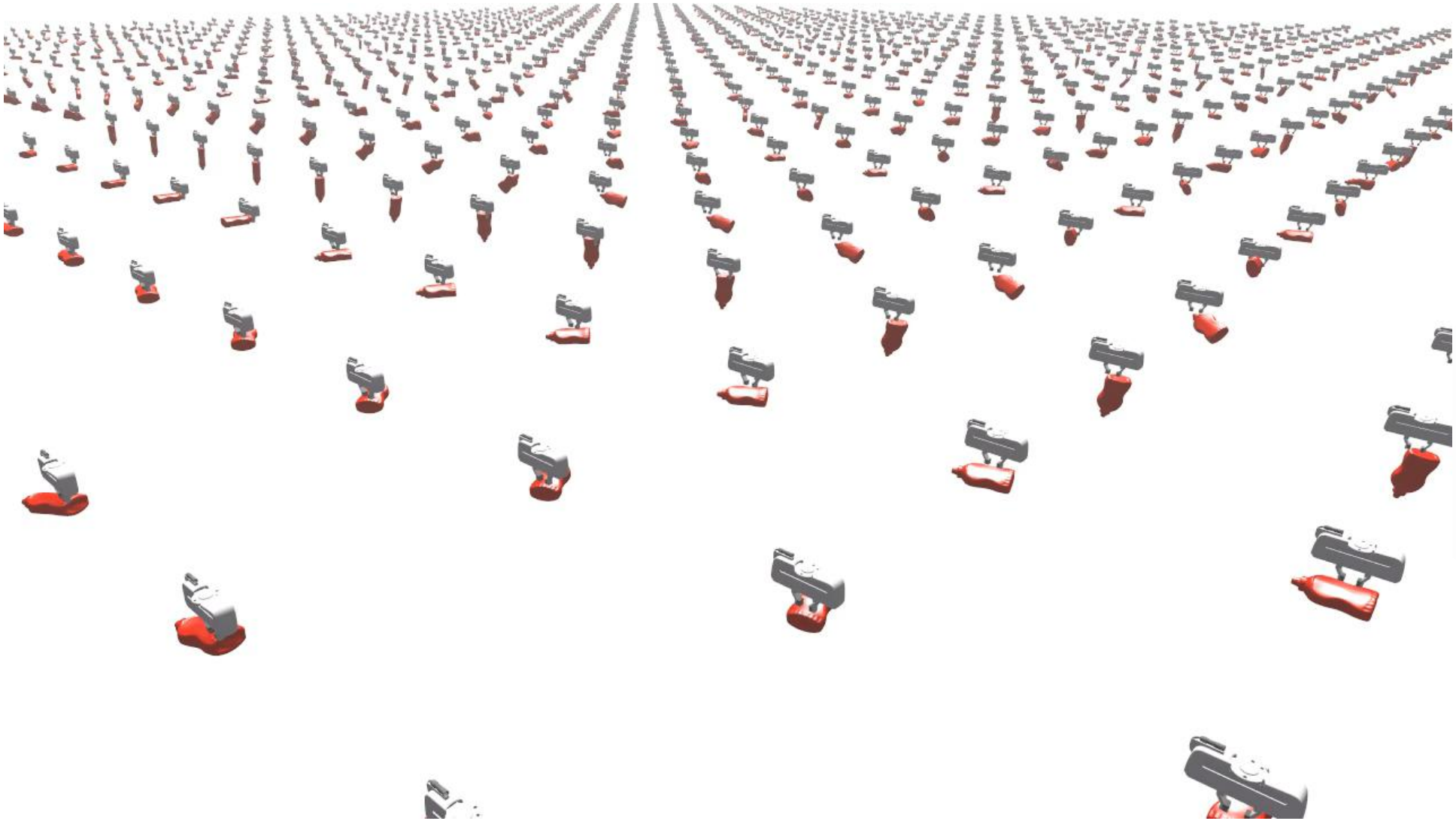
- PoseCNN, RSS'17
- DeepIM, ECCV'18
- DOPE, CoRL'18
- PoseRBPF, RSS'19, T-TO'21
- Self-supervised 6D Pose, ICRA'20
- LatentFusion, CVPR'20

Grasp Planning: GraspIt!

GraspIt! https://graspit-simulator.github.io/

Andrew Miller and Peter K. Allen. "Graspit!: A Versatile Simulator for Robotic Grasping". IEEE Robotics and Automation Magazine, V. 11, No.4, Dec. 2004, pp. 110-122.

Grasp Planning: A Physics-based Approach



Eppner-Mousavian-Fox, ISRR'19

MultiGripperGrasp

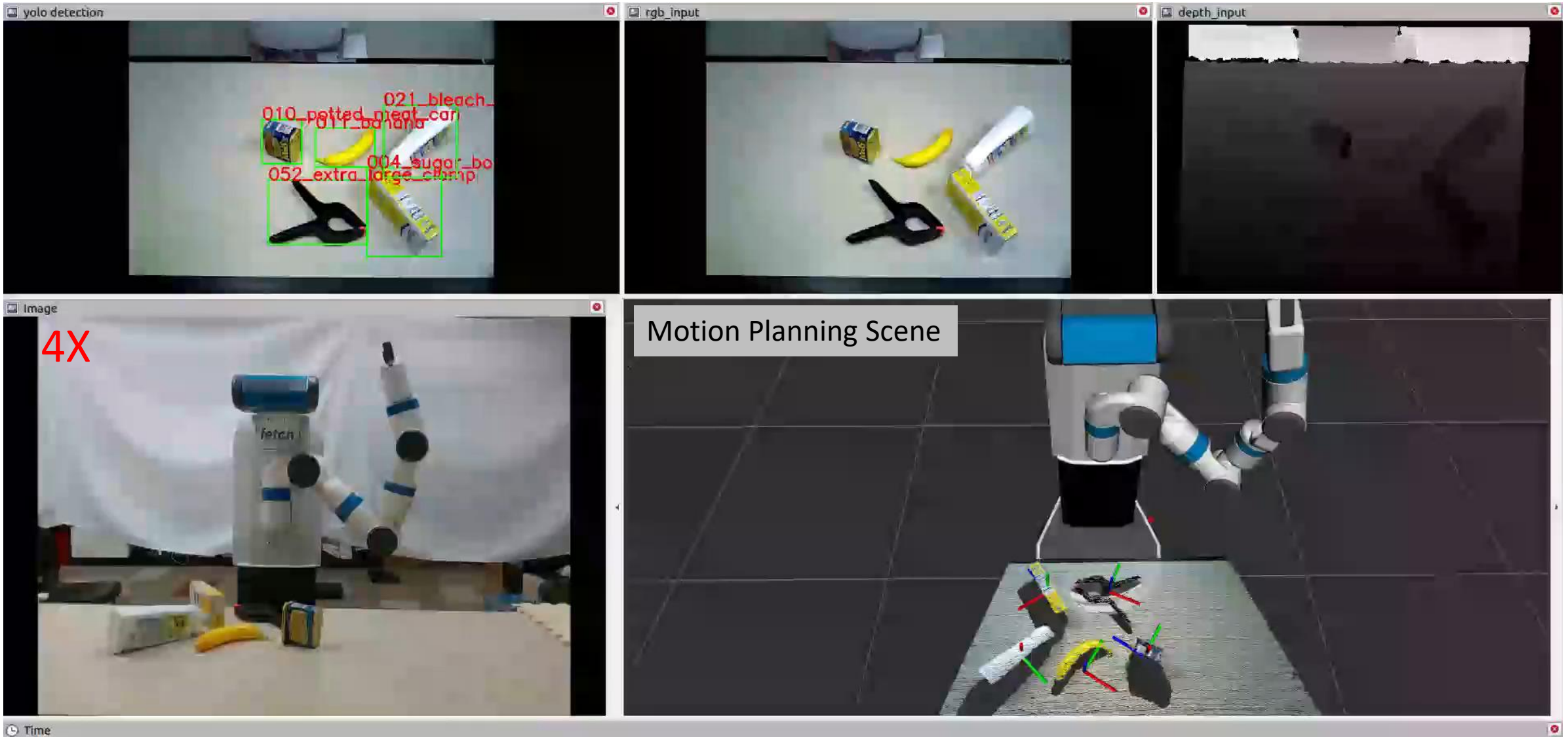
- A large-scale dataset for robotic grasping
  - 11 grippers, 345 objects, 30M grasps

Luis Felipe Casas

Ninad Khargonkar

MultiGripperGrasp: A Dataset for Robotic Grasping from Parallel Jaw Grippers to Dexterous Hands
Luis Felipe Casas Murrilo*, Ninad Khargonkar*, Balakrishnan Prabhakaran, Yu Xiang (*equal contribution)
In IROS, 2024.

Motion Planning



The Open Motion Planning Library in MoveIt https://ompl.kavrakilab.org/index.html

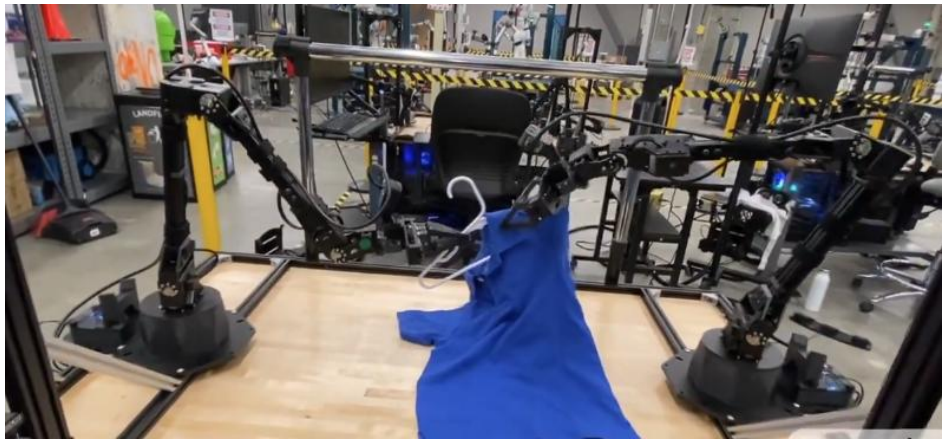
Using 3D Object Models

- Pros
  - Encodes appearance, 3D shape, affordance, physical properties for perception, planning and simulation

SAM 3D Objects

Maybe we can in the future

- Cons
  - We cannot build 3D models for all objects

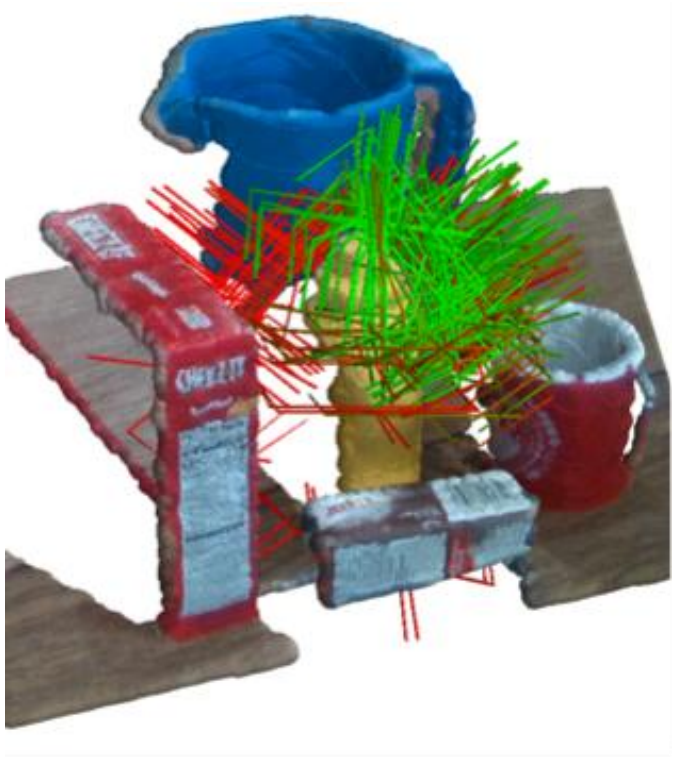


ALOHA Unleashed
Google DeepMind

Using 3D Point Clouds

Perception → Planning → Control

object instance segmentation



Grasp planning from point clouds

Control to reach grasp

Figure Credit: Murali-Mousavian-Eppner-Paxton-Fox, ICRA'20

Object Segmenting: Unseen Objects

Yangxiao Lu

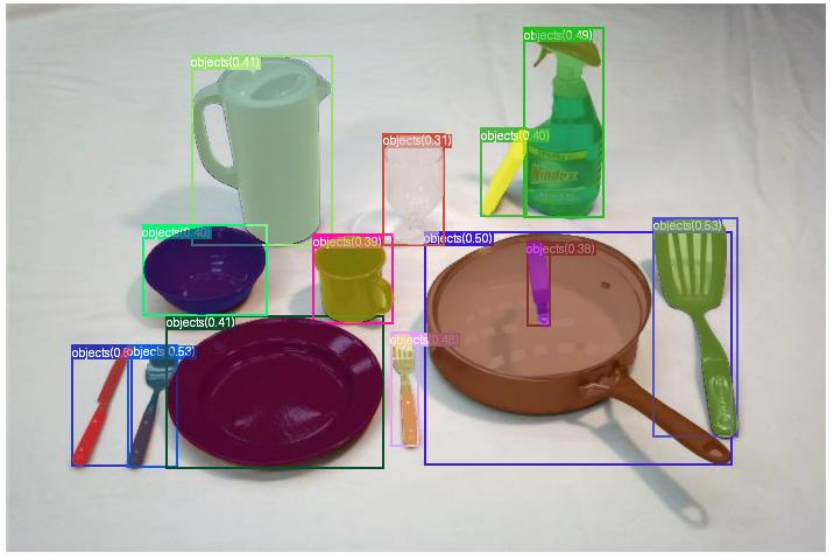
Input Image

Output Label

Xie-Xiang-Mousavian-Fox, CoRL'19, T-RO'21, CoRL'21
Xiang-Xie-Mousavian-Fox, CoRL'20
Lu-Khargonkar-Xu-Averill-Palanisamy-Hang-Guo-Ruozzi-Xiang, RSS'23
Lu-Chen-Ruozzi-Xiang, ICRA'24
Qian-Lu-Ren-Wang-Khargonkar-Xiang-Hang, ICRA'24

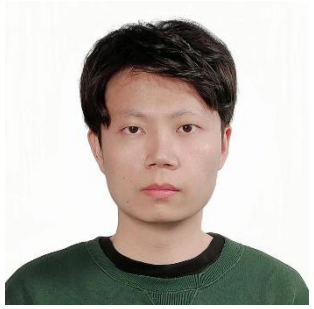
Object Segmentation: Language-guided

- Gounding Dino (object detection)
- SAM (object segmentation)

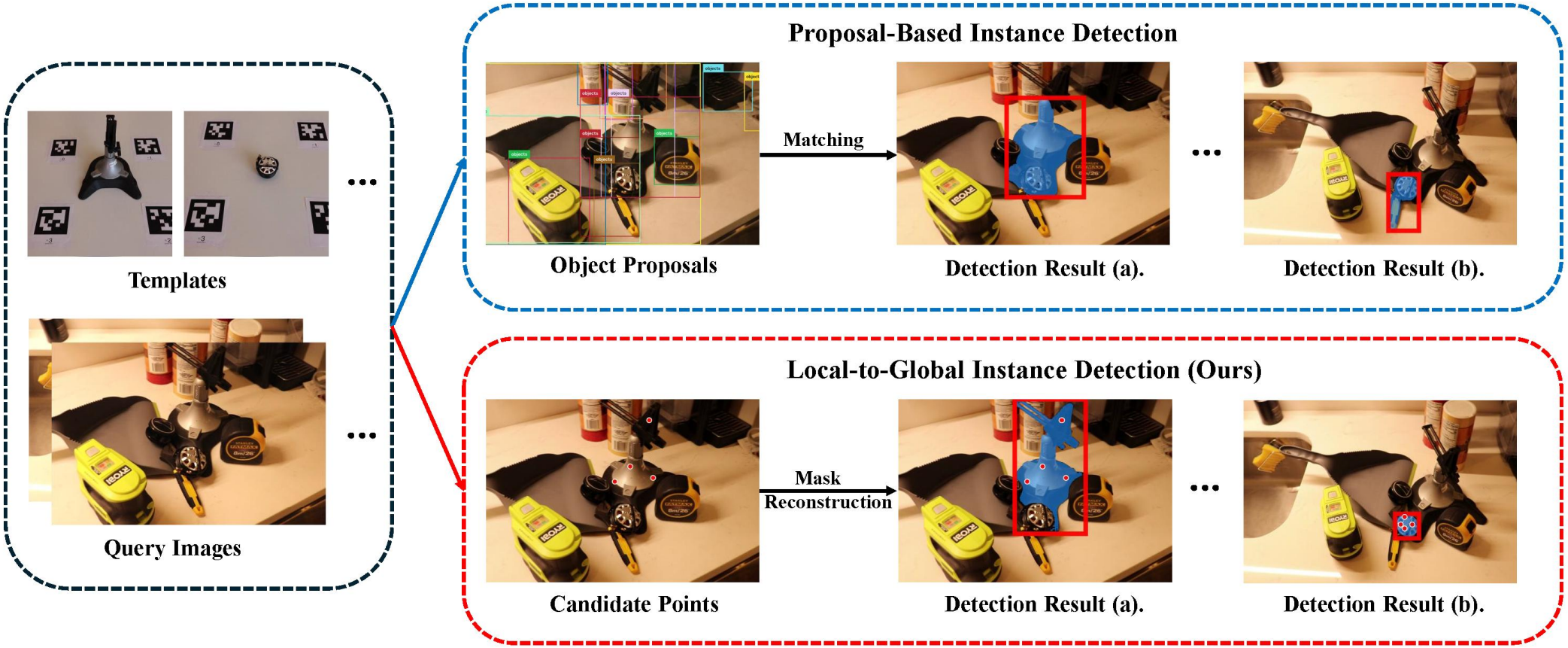


- Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. Liu et al., 2023
- Segment Anything. Kirillov et al., 2023

Object Segmenting: Template-Guided

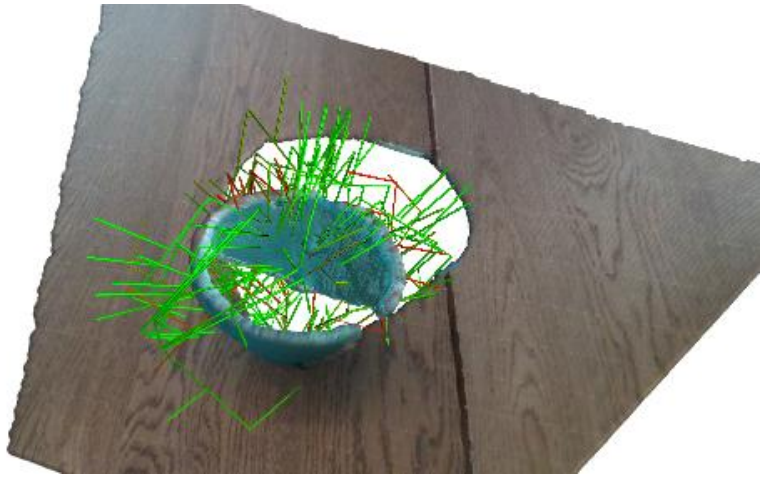


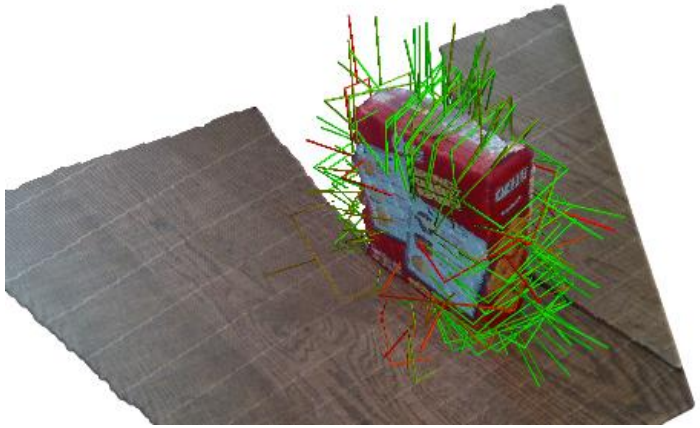
Qifan Zhang



From Local Matches to Global Masks: Template-Guided Instance Detection and Segmentation in Open-World Scenes
Qifan Zhang, Sai Haneesh Allu, Jikai Wang, Yangxiao Lu, Yu Xiang
In Robotics: Science and Systems (RSS), 2026.

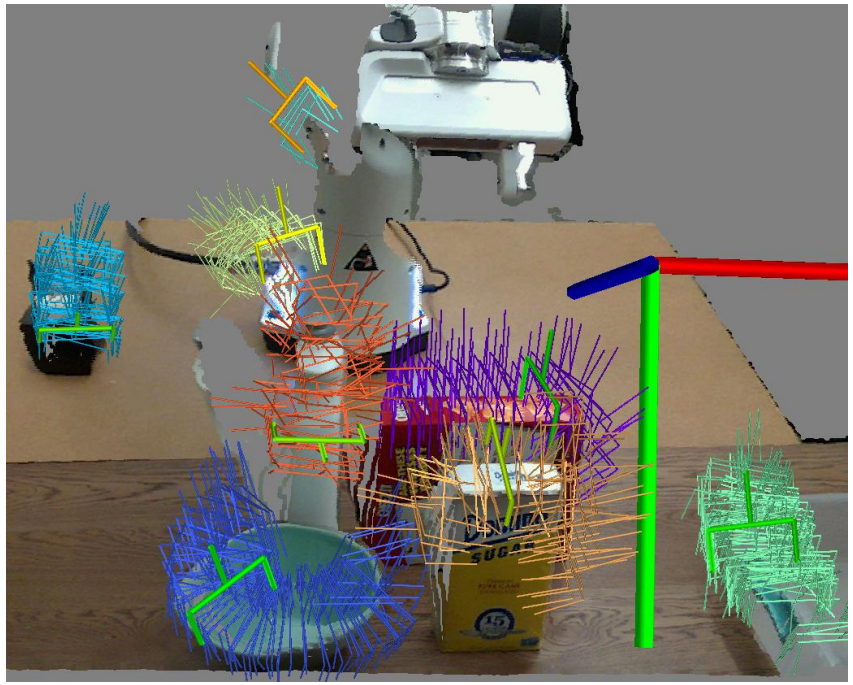
Grasp Planning with Point Clouds





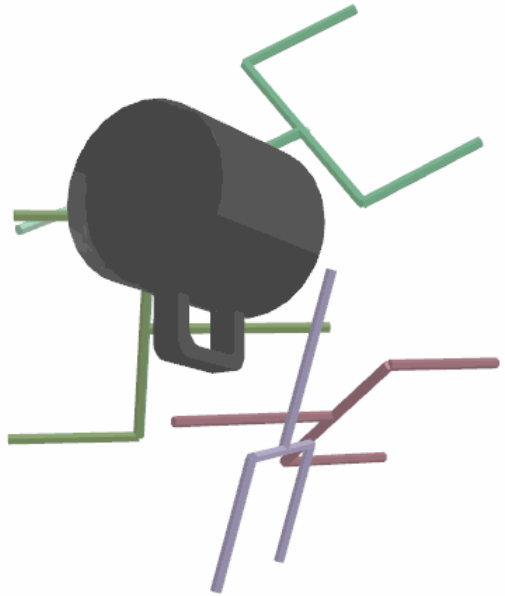
6D GraspNet

6-DOF GraspNet: Variational Grasp Generation for Object Manipulation. Mousavian et al., ICCV'19



Contact-GraspNet

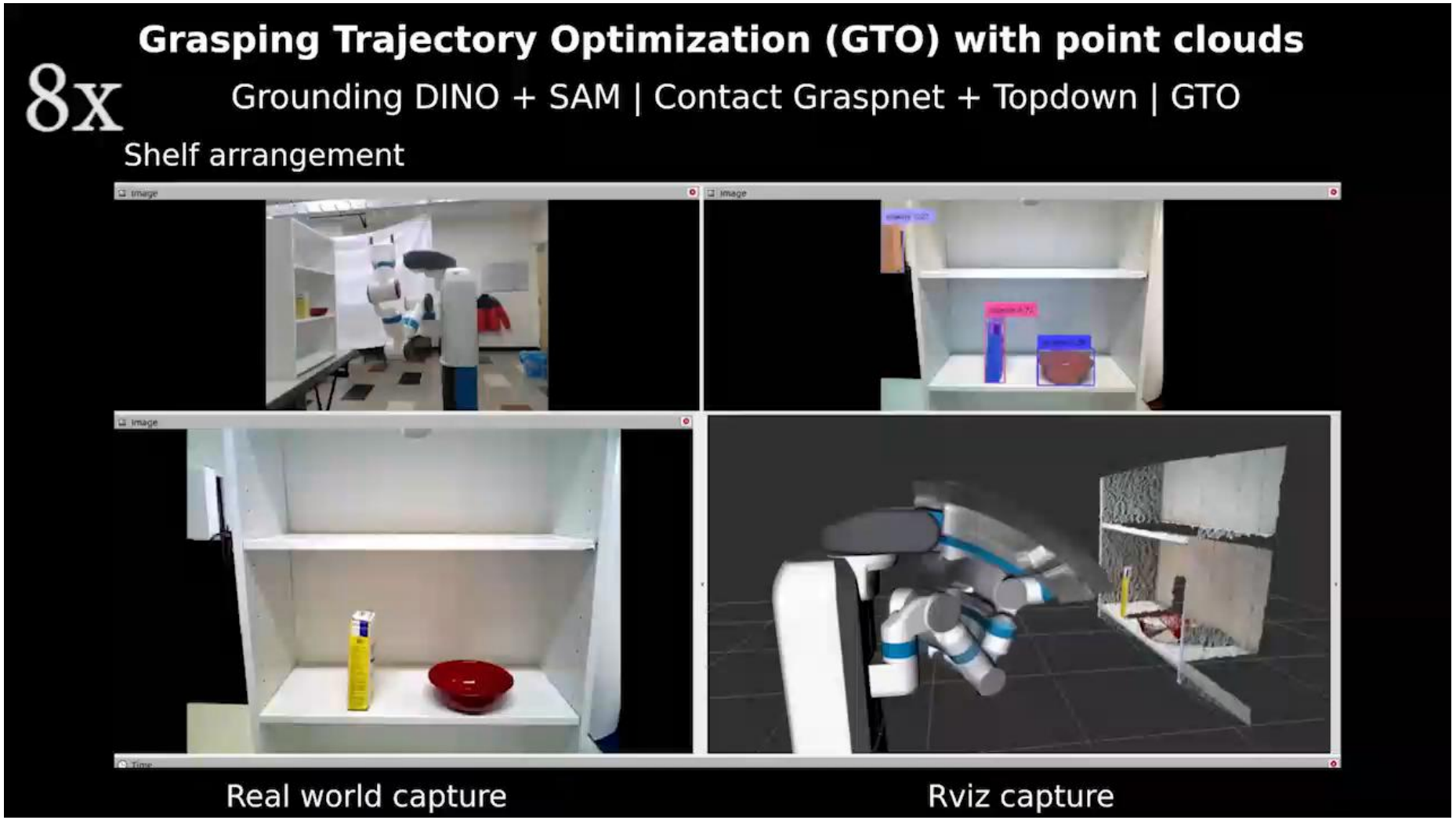
Contact-GraspNet: Efficient 6-DoF Grasp Generation in Cluttered Scenes. Sundermeyer, et al., ICRA'21



SE(3)-DiffusionFields

SE(3)-DiffusionFields: Learning smooth cost functions for joint grasp and motion optimization through diffusion. Urain et al., 2023

Grasping Trajectory Optimization with Point Clouds

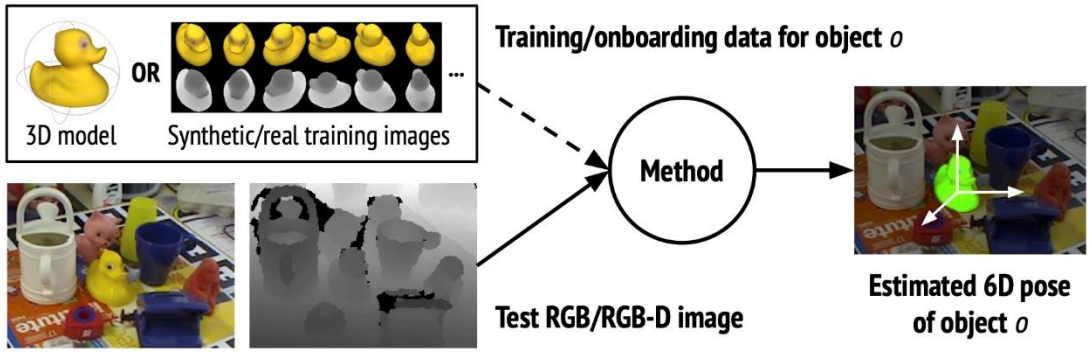


Grasping Trajectory Optimization with Point Clouds

Yu Xiang, Sai Haneesh Allu, Rohith Peddi, Tyler Summers, Vibhav Gogate. In IROS, 2024.

Benchmarking

- BOP: Benchmark for 6D object pose estimation https://bop.felk.cvut.cz/home/



Datasets: Core datasets LM LM-O T-LESS ITODD HB HOPE YCB-V RU-APC IC-BIN IC-MI TUD-L TYO-L

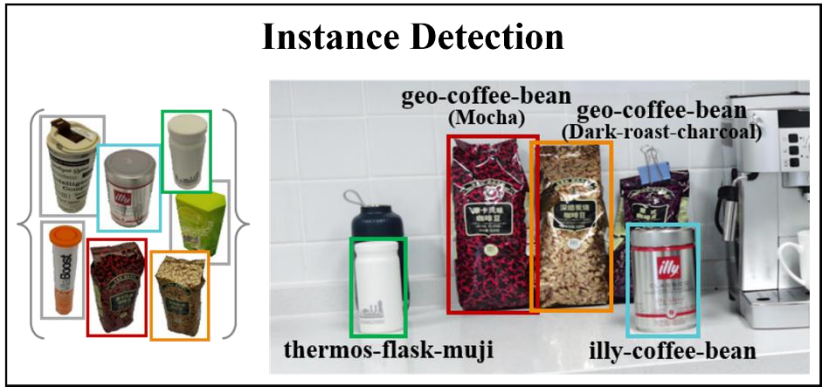
6D localization of seen objects – Core datasets

This leaderbord shows the overall ranking for Task 1 on the core datasets (LM-O, T-LESS, TUD-L, IC-BIN, ITODD, HB, YCB-V). For each method, the date of the latest considered submission is reported. If more submissions of a method are available for a dataset, the submission with the highest AR_{Core} score is considered. The reported time is the average image processing time averaged over the core datasets.

Show 50 entries Search:

	Date (UTC)	Method	Test image	AR_{Core}	AR_{LM-O}	AR_{T-LESS}	AR_{TUD-L}	AR_{IC-BIN}	AR_{ITODD}	AR_{HB}	AR_{YCB-V}	Time (s)
1	2023-09-24	GPose2023-OfficialDet	RGB-D	0.851	0.805	0.895	0.966	0.734	0.687	0.944	0.929	4.575
2	2022-10-15	GDRNPP-PBRReal-RGBD-MModel	RGB-D	0.837	0.775	0.874	0.966	0.722	0.679	0.926	0.921	6.263

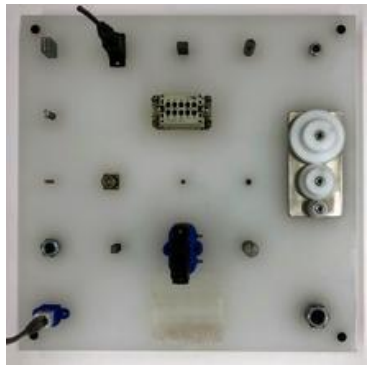
- Object instance detection https://eval.ai/web/challenges/challenge-page/2478/overview



Rank	Participant team	AP (↑)	AP50 (↑)	AP75 (↑)	AP_easy (↑)	AP_hard (↑)	AP_small (↑)	AP_medium (↑)	AP_large (↑)	AR_1 (↑)	AR_10 (↑)	AR_100 (↑)	AR_small (↑)	AR_medium (↑)	AR_large (↑)	Last submission at
1	Grounding-X	73.4	79.8	77.1	77.4	62.0	51.8	78.5	86.9	79.4	81.8	81.8	62.4	85.5	91.3	2 days ago
2	leinad (ZERO)	70.3	83.3	75.9	74.1	59.4	49.0	76.6	84.9	75.9	76.8	76.8	60.6	82.0	87.5	1 day ago
3	hust_zwb	69.5	80.6	74.8	74.0	56.8	45.6	74.3	87.5	73.7	73.7	73.7	50.8	78.9	90.3	8 days ago
4	IRVL (NIDS-Net2)	67.2	79.9	74.1	72.3	52.7	42.2	71.1	89.6	71.4	71.4	71.4	45.9	75.5	91.6	14 days ago

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

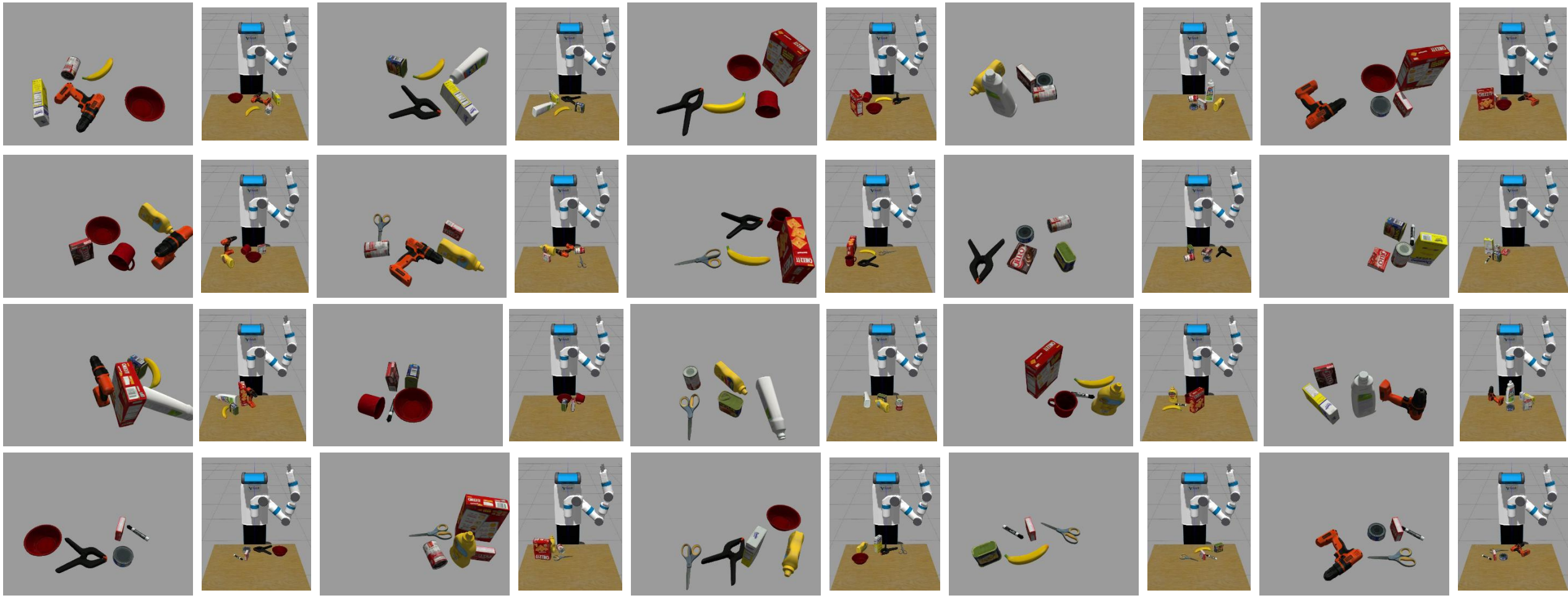
SceneReplica

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.

SceneReplica Benchmark

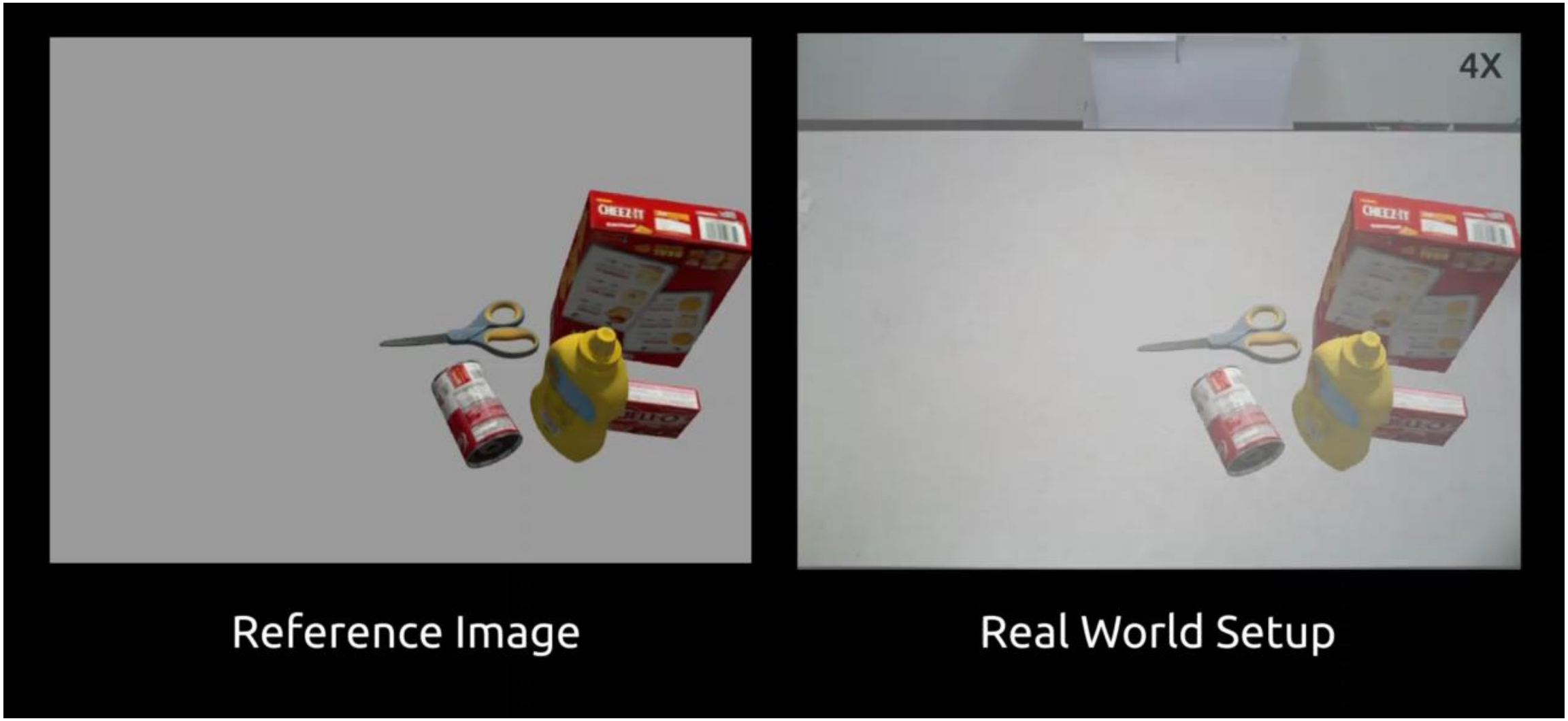
Ninad Khargonkar Sai Haneesh Allu

20 Scenes



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

Real-World Scene Setup



Model-based Grasping vs Model-free Grasping

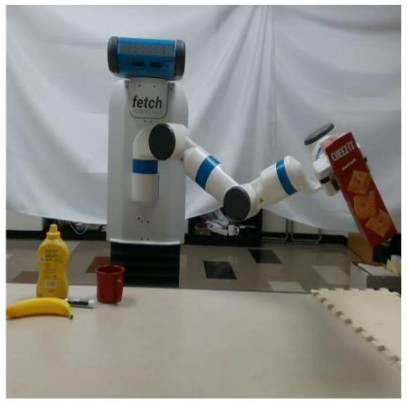
Input real world scene

6D Pose Estimation

Offline Grasp Database

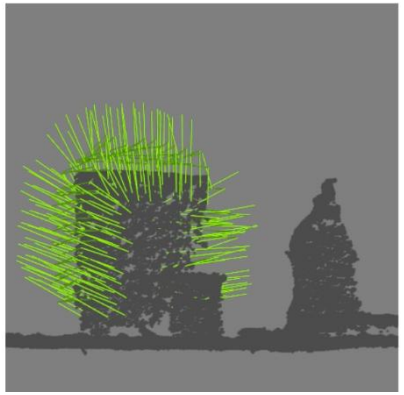
Motion Planning Setup

Grasping & Lifting

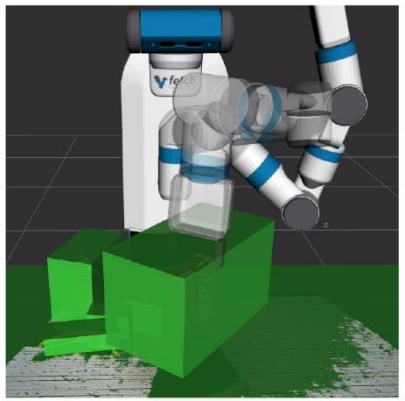


Moving arm for Dropoff

Unseen Object Segmentation

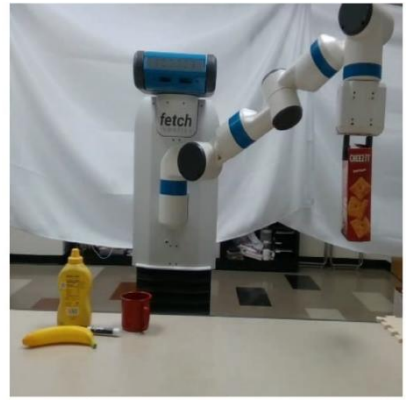


Model-free Grasp Planning



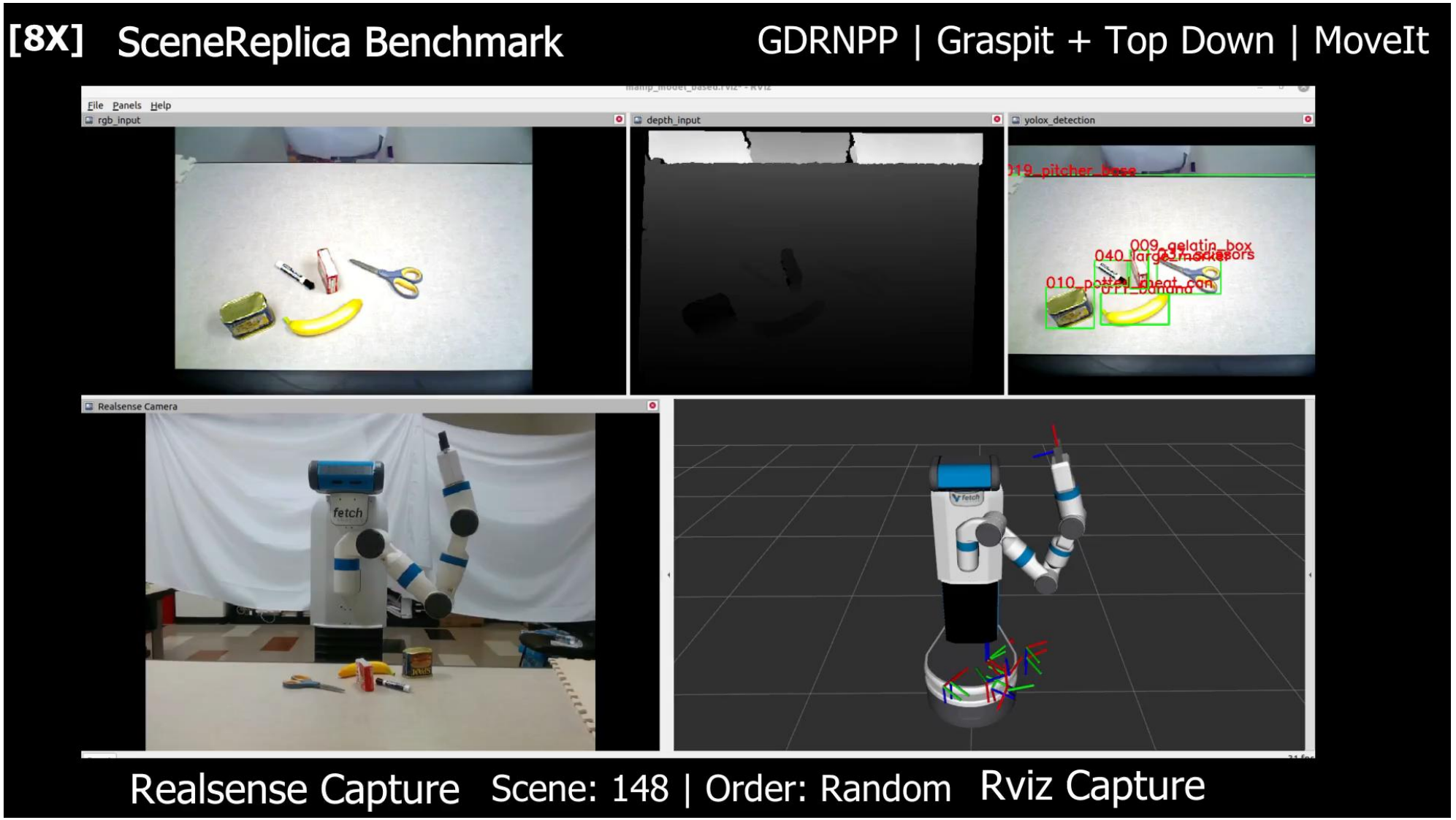
Motion Planning Setup

Grasping & Lifting

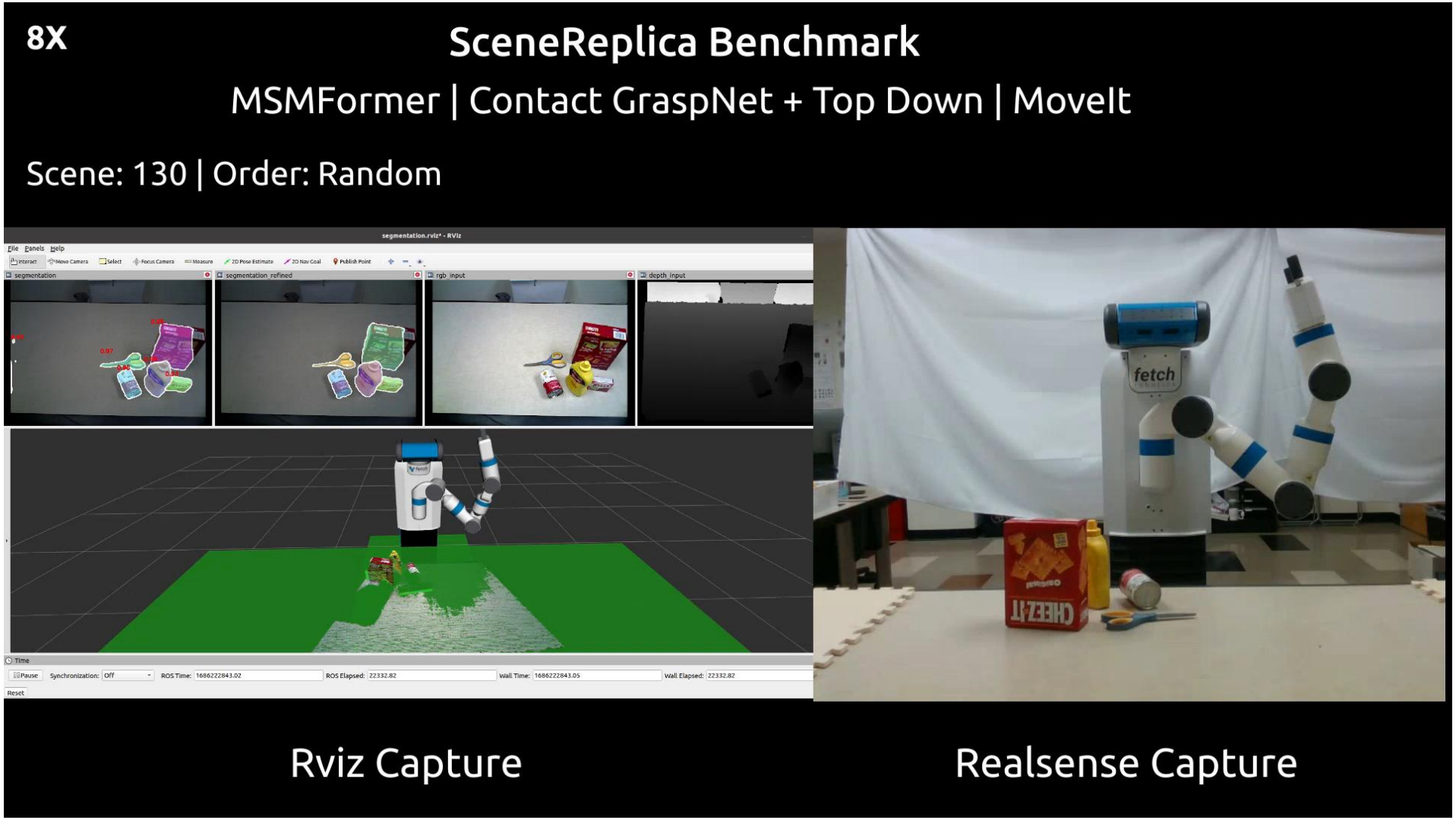


Moving arm for Dropoff

Model-based Grasping Example



Model-free Grasping Example

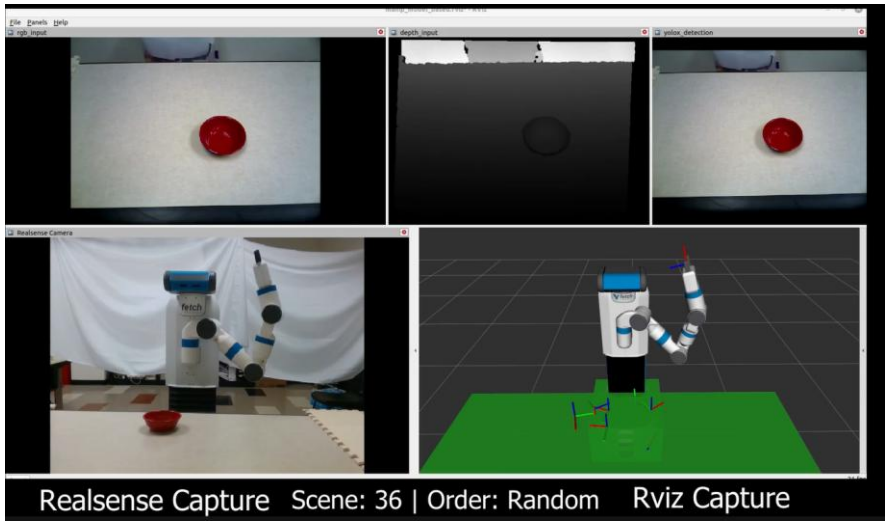


Current Leaderboard

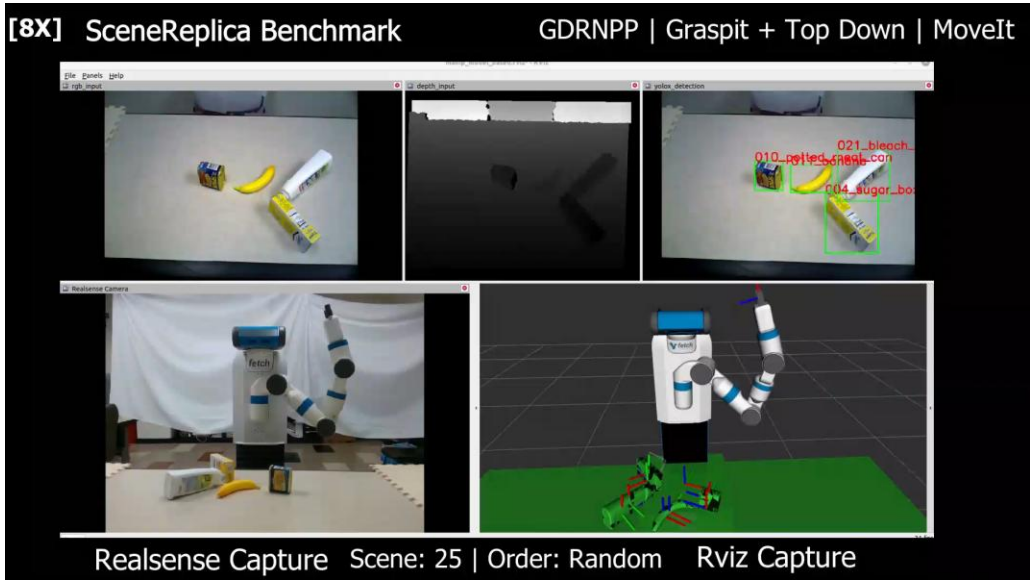
#	Perception	Grasp Planning	Motion Planning	Control	Ordering	Grasping Type	Pick & Place Success	Grasping Success	Videos
11	GDRNPP [9]	RFP [12] + Top-Down	OMPL [3]	MoveIt	Near-to-Far	Model-Based	70/100	73/100	
10	MSMFormer [8]	Contact-graspnet [7] + Top-Down	GTO [11]	MoveIt	Near-to-Far	Model-Free	65/100	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/

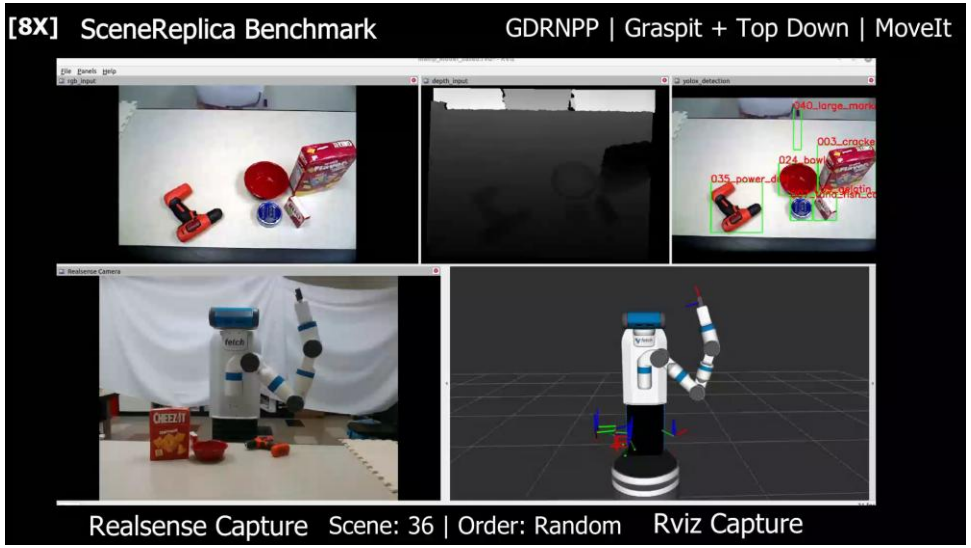
Failure Analysis (GDRNPP + GraspIt! + Top-Down)



Object Detection Error



Grasp Planning Error



Pose Estimation Error

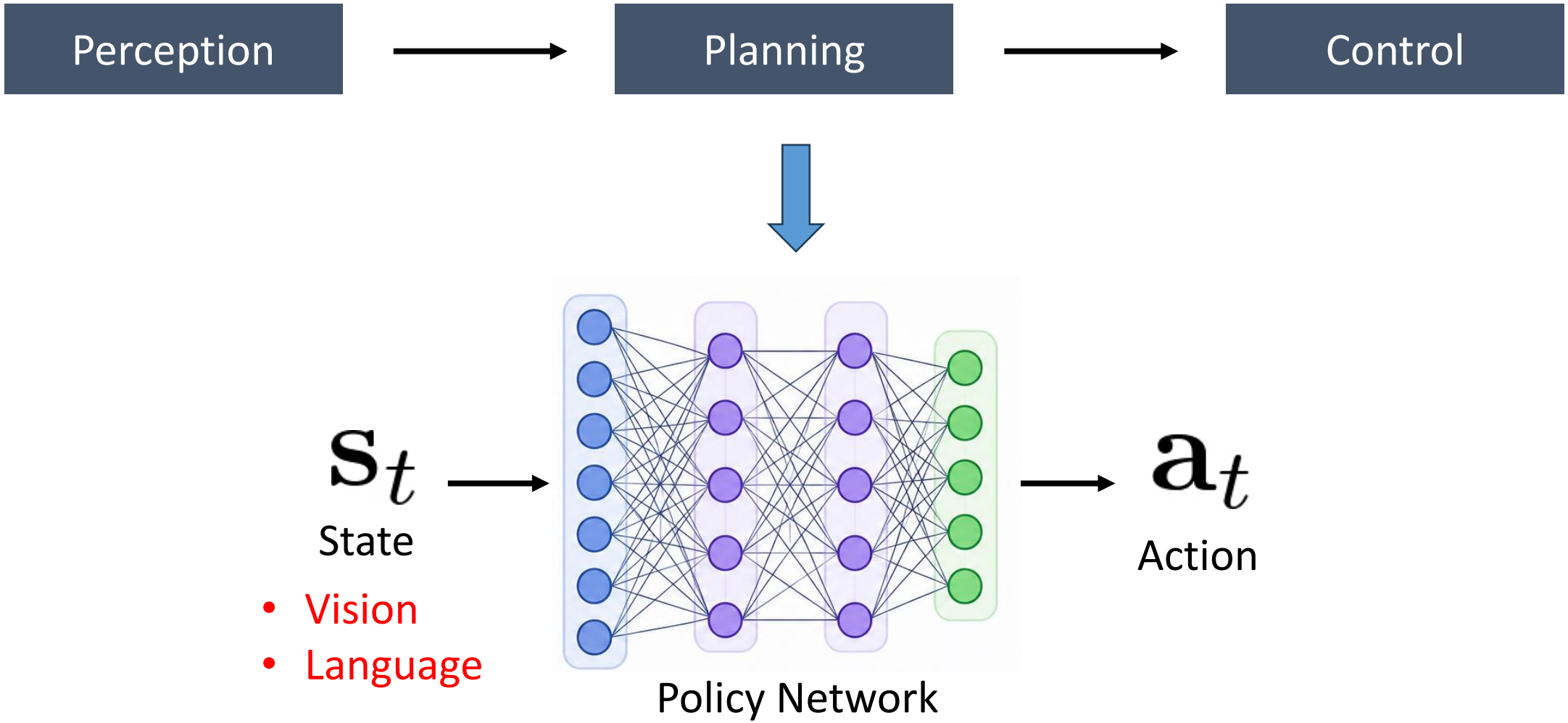
Object	Method 3			
	S	$P_E F$	$P_L F$	EF
003 cracker box	3	2	1	-
004 sugar box	5	-	-	-
005 tomato soup can	5	1	-	1
006 mustard bottle	7	-	-	-
007 tuna fish can	1	5	-	-
008 pudding box	5	-	-	-
009 gelatin box	6	-	1	-
010 potted meat can	7	-	-	-
011 banana	6	-	1	-
021 bleach cleanser	3	1	-	1
024 bowl	2	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

$P_E F$: #perception failure

$P_L F$: #planning failure

EF: #execution failure

End-to-End Manipulation



Vision-Language-Action (VLA) Models

Some Recent Breakthroughs

https://www.physicalintelligence.company/blog/pi0

Physical Intelligence: a startup with people from Berkeley, Stanford, etc.

How Good are these VLA Models?

- We need benchmarking

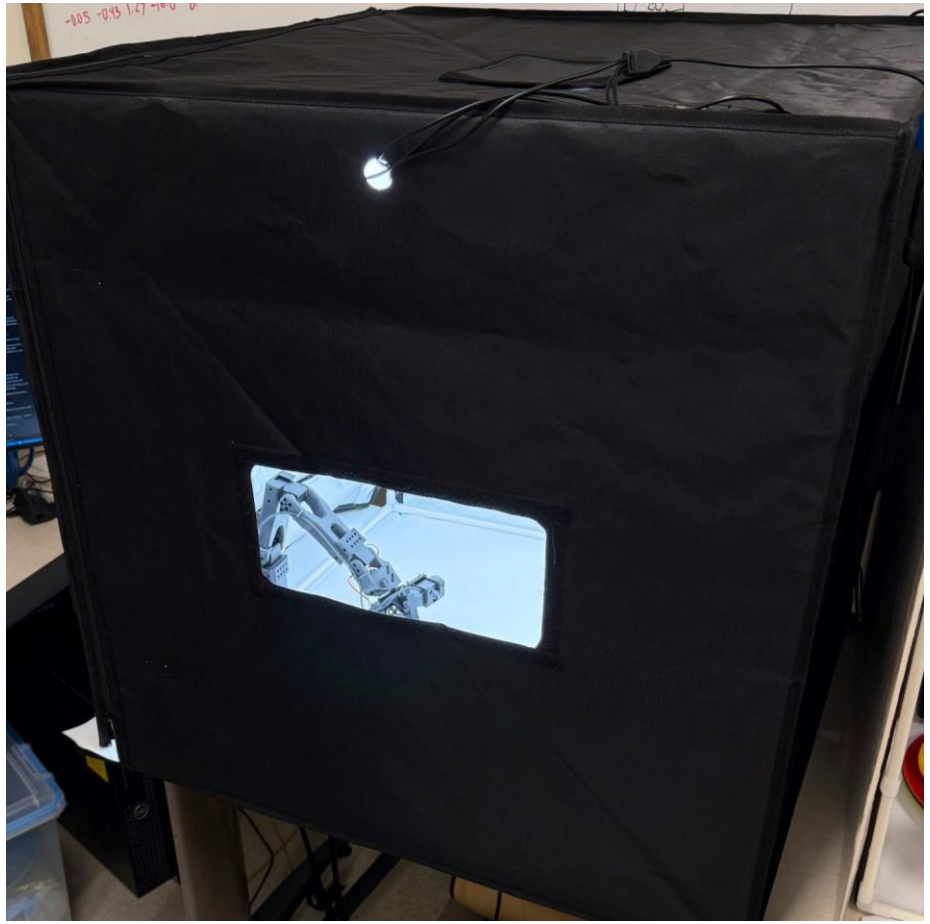
https://irvlutd.github.io/VLAReplica/

In arXiv, 2026.

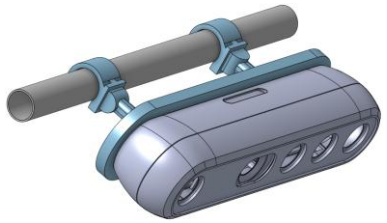
VLA-Replica Benchmark

Low-cost Components

SO-101 Arm ~$200

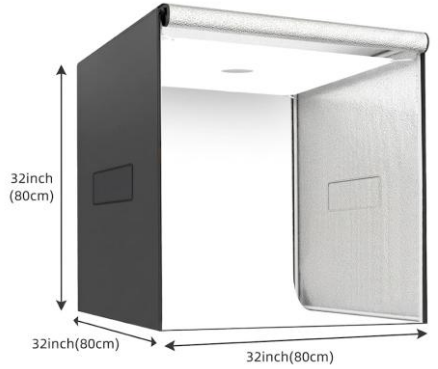


Real Set-up



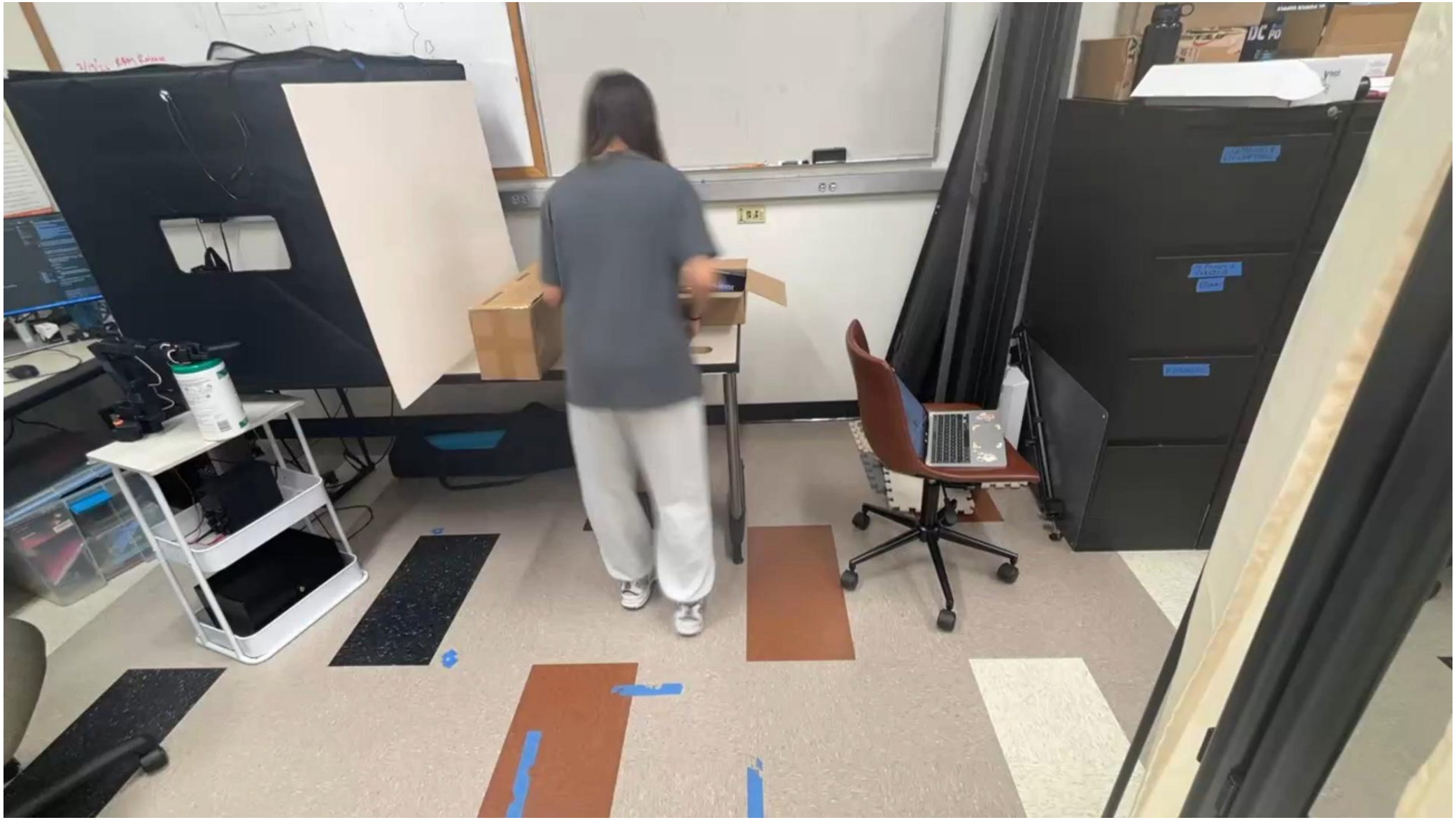
D455 Camera ~$425

Webcam $13.98



Light Box $152.99

VLA-Replica Benchmark



VLA-Replica Benchmark

2X

Camera Calibration

Using reference images for camera calibration

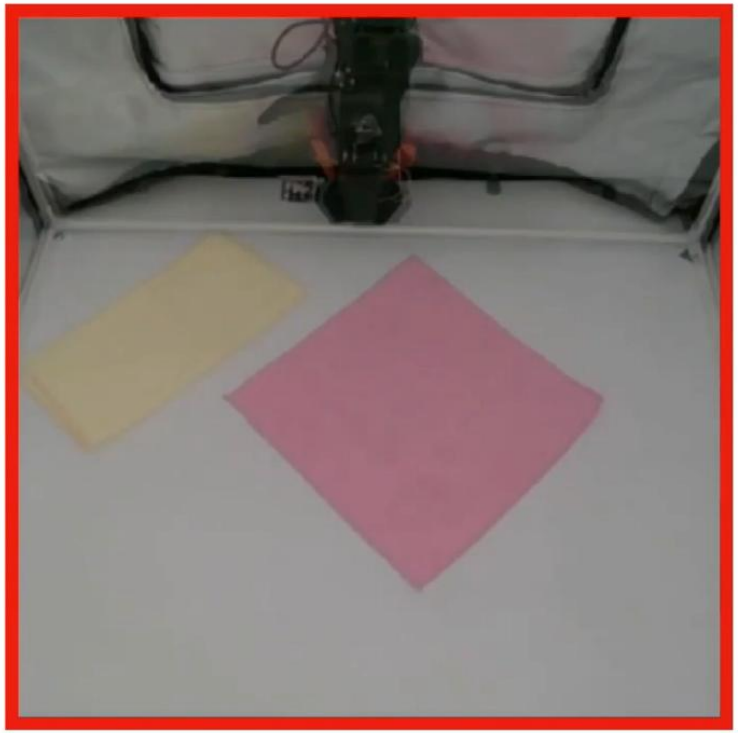
Top Camera Calibration

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

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 n times	Lift bowl n times	Press button n times	Video
1	$\pi_{0.5}$ [6]	**0.54**	0.8	0.8	0.4	0.4	1.0	0.6	0.4	0.4	0.4	0.2	🔗
2	π_0 [5]	**0.34**	0.8	0.6	0	0	0.8	0.2	0.4	0.2	0.2	0.2	🔗
3	SmolVLA [3]	**0.26**	0.6	0.2	0.2	0	0.6	0.4	0.2	0	0.2	0.2	🔗
4	ACT [1]	**0.18**	0.4	0	0	0	0.4	0.4	0.2	0.2	0.2	0	🔗
5	DiT-D [2]	**0.16**	0.4	0	0	0.2	0.2	0.6	0.2	0	0	0	🔗
6	X-VLA [4]	**0.14**	0.4	0.2	0	0	0.6	0	0	0.2	0	0	🔗
7	DiT-F [2]	**0.12**	0.4	0	0	0	0.2	0.4	0.2	0	0	0	🔗

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

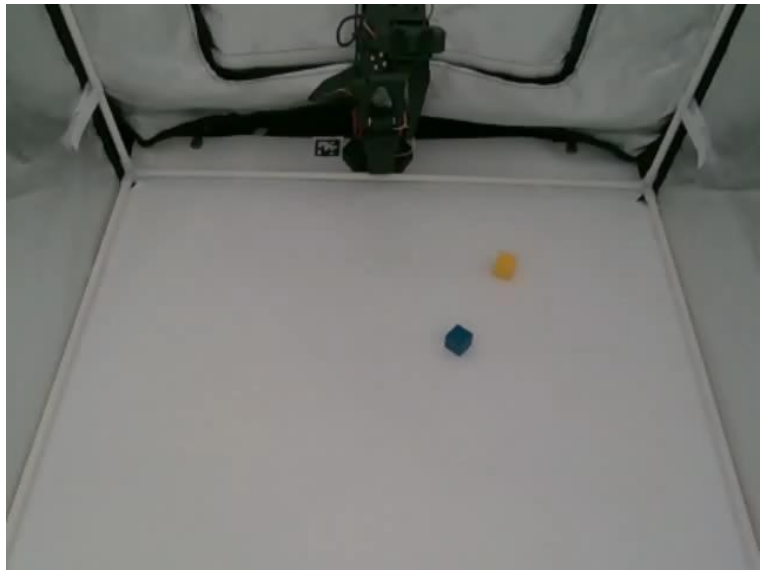
VLA-Replica Benchmark: Leaderboard

VLA-Replica-ID | VLA-Replica-OOD

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	Video
1	$\pi_{0.5}$ [6]	0.35	1.0	0.4	0	0.2	0.8	0.4	0	0	🔗
2	SmolVLA [3]	0.3	0.8	0.4	0.2	0.2	0.6	0	0.2	0	🔗
3	π_0 [5]	0.3	0.8	0.6	0.2	0	0.6	0.2	0	0	🔗
4	ACT [1]	0.075	0.4	0.2	0	0	0	0	0	0	🔗
5	X-VLA [4]	0.075	0.6	0	0	0	0	0	0	0	🔗
6	DiT-D [2]	0.05	0	0.2	0	0	0.2	0	0	0	🔗
7	DiT-F [2]	0.025	0.2	0	0	0	0	0	0	0	🔗

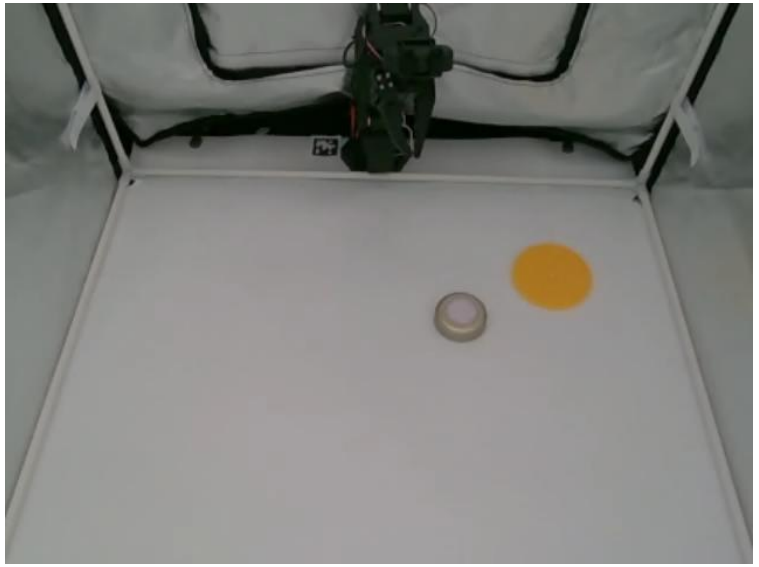
https://irvlutd.github.io/VLAReplica/

VLA-Replica Benchmark: Pi0.5 Behaviors



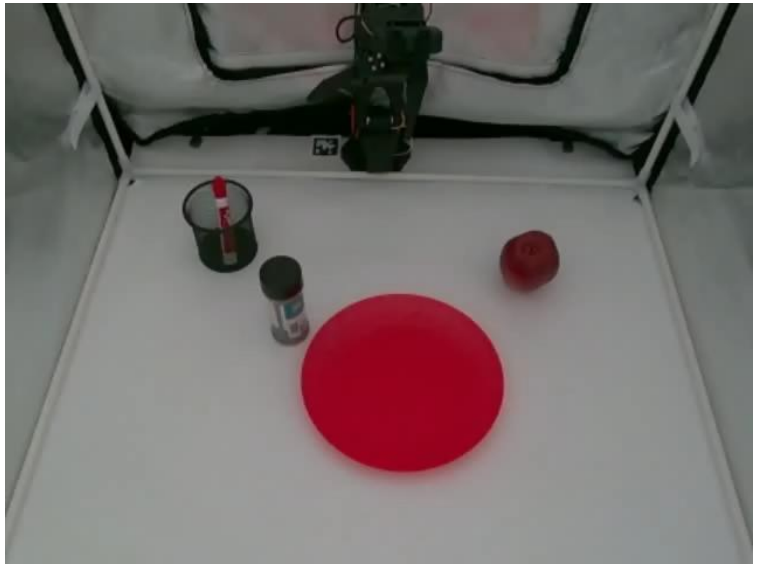
Stack the **yellow** block on the **blue** block

Precise manipulation is challenging



Press the button **three** times

Memory task is challenging



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

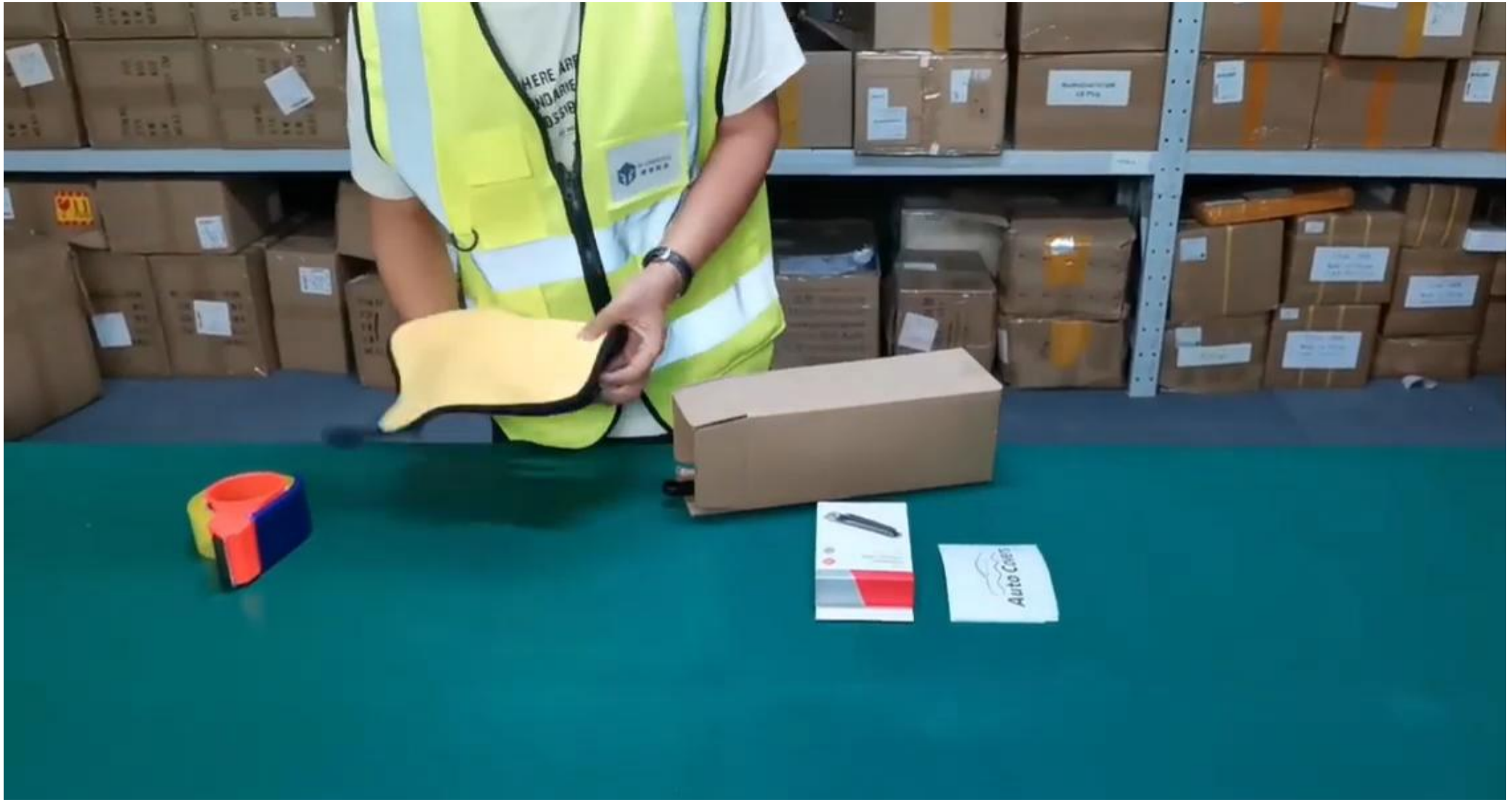
Failure recovery

Discussion

- Modular pipelines
  - Interpretability and debugging
  - Easier safety and constraint enforcement
  - Error accumulation
  - High latency usually, hard for closed-loop execution

- End-to-end manipulation
  - Closed-loop manipulation
  - Failure recovery
  - Better for contact-rich dexterity
  - Poor interpretability and generalization
  - Hard to debug (More data? Better model?)

Robot Manipulation is still an Open Challenge



IRVL
INTELLIGENT ROBOTICS AND VISION LAB
UT DALLAS

NSF DARPA SONY NVIDIA XPENG FRONTIER MINING

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

Assisted by
Ms. Rhonda Walls

Thank you!