NVIDIA

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

LEVERAGING SIMULATION TO TEACH OBJECT MANIPULATION TASKS

Dieter Fox, NVIDIA and University of Washington

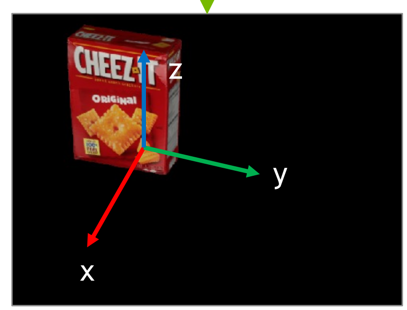
INGREDIENTS OF A MANIPULATION SYSTEM

- Task and motion planning
  - Determine sequence of high-level commands and collision-free trajectories to achieve goal configuration
- State estimation and perception
  - Infer relevant quantities from sensor data (objects, drawers, doors, manipulator, contacts, …)
- Object grasping and placement
  - Determine good grasps for objects given constraints (gripper, local geometry, placement)
- Trajectory generation and control
  - Real-time, reactive generation of control commands to safely move robot / gripper toward goals

PICK-AND-PLACE KITCHEN MANIPULATION SYSTEM

All objects are known, articulated kitchen model available, no clutter

6D OBJECT POSE ESTIMATION



6D Object Pose

3D Translation

3D Orientation

STATE ESTIMATION VIA OPTIMIZATION

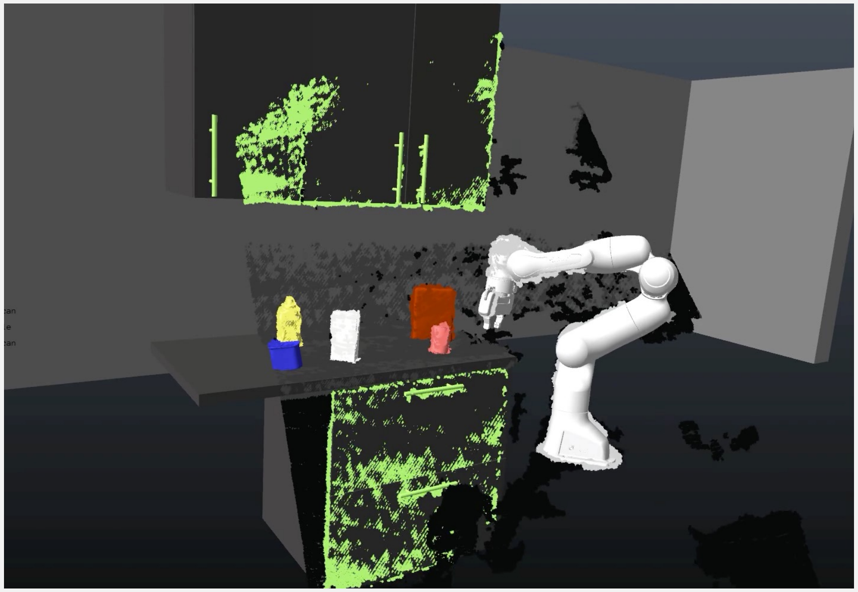
State θ includes camera pose, cabinet doors, drawers, object poses, robot base and manipulator

Depth camera: optimize articulation parameters to minimize point distance from model

Physical constraints: contacts and non-interpenetration added as loss terms

Object detections: decaying loss term

Robot and manipulator pose: decaying loss term



$$L(\theta) = L_{match}(\theta) + L_{physics}(\theta) + L_{detect}(\theta) + L_{base}(\theta)$$

TASK AND MOTION PLANNING WITH REACTIVE BEHAVIOR EXECUTION

- TAMP plans over high-level actions, pre-conditions / effects, and continuous trajectories
- Real-time kitchen, robot and object tracking
- Robust logical-dynamical systems perform real-time switching of behaviors based on pre-conditions computed from state
- Real-time reactive motion generation using Riemannian Motion Policies

[Cheng-Mukadam-Issac-Birchfield-Fox-Boots-Ratliff: WAFR-18]

[Garrett-Paxton-Lozano-Perez-Kaelbling-Fox: ICRA-20]

[Paxton-Ratliff-Eppner-Fox: IROS-19]

MODEL-FREE GRASPING AND PLACING OF UNKNOWN OBJECTS

MODEL-BASED VS MODEL-FREE GRASPING

Model-Based Grasping: Estimate Object Pose and use Inferred Pose to Transform Grasps

Observation

3D Model of Object

Pose Estimation Model

R,T

Predicted Object Pose

Annotated Grasps

Transform Grasps

Final Grasps

MODEL-BASED VS MODEL-FREE GRASPING

Model-Free Grasping: Directly Predict Final Grasp Pose

Observation

Grasp Generation Model

Grasps

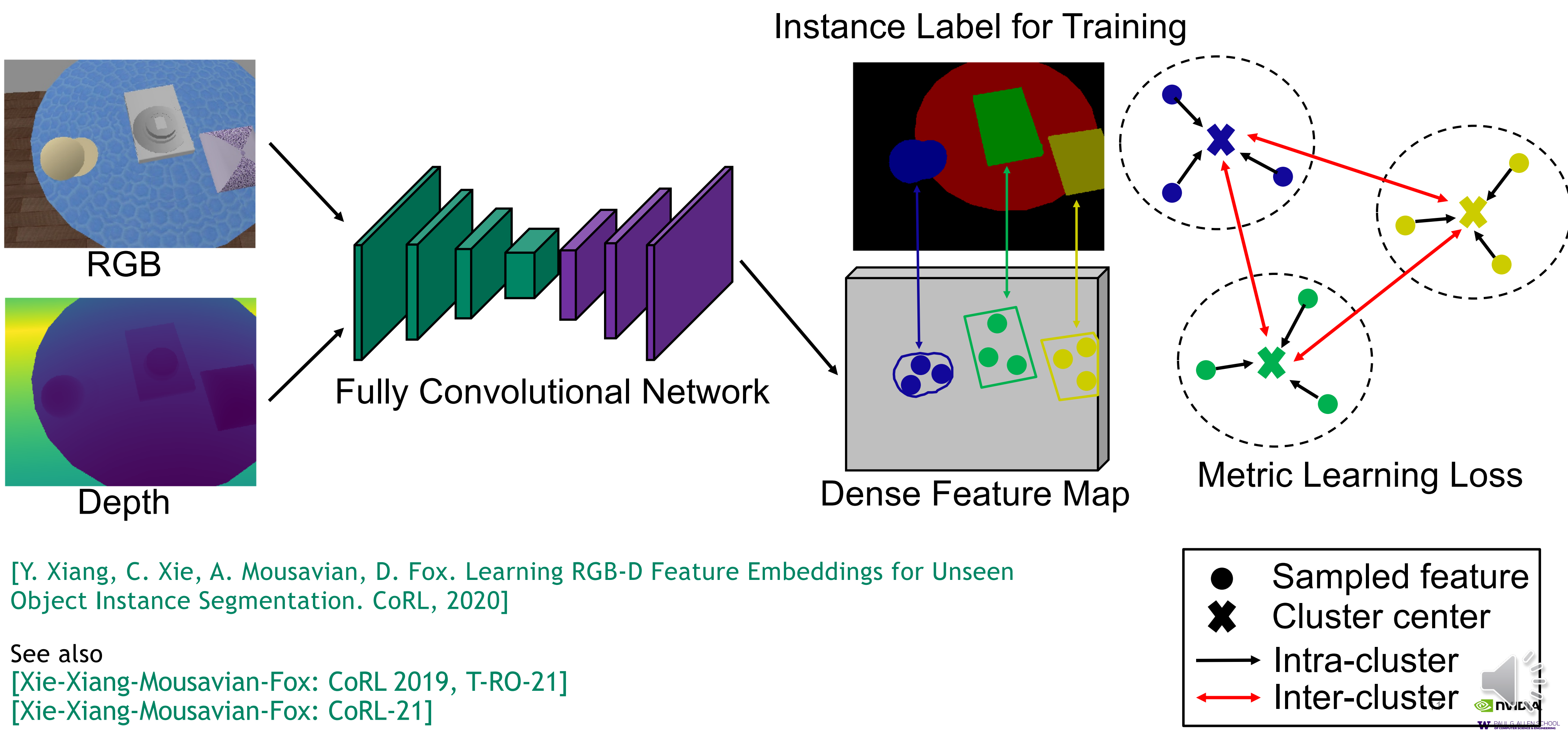
GETTING AN OBJECT OUT OF CLUTTER

Need to Segment Scene, Generate Grasps, and Check for Collisions

External view

Gripper camera view

UNKNOWN OBJECT INSTANCE SEGMENTATION



[Y. Xiang, C. Xie, A. Mousavian, D. Fox. Learning RGB-D Feature Embeddings for Unseen Object Instance Segmentation. CoRL, 2020]

See also
[Xie-Xiang-Mousavian-Fox: CoRL 2019, T-RO-21]
[Xie-Xiang-Mousavian-Fox: CoRL-21]

PHOTOREALISTIC SYNTHETIC TRAINING DATA

350K Rendered Images Along with Segmentation and Object Id

[Mousavian-Manuelli-Okorn-Xiang-Eppner-Murali-Fox, 2023]

OBJECTSEEKER INSTANCE SEGMENTATION

On Par with SOTA on Tabletop Datasets and SOTA on Non-Tabletop Scenes

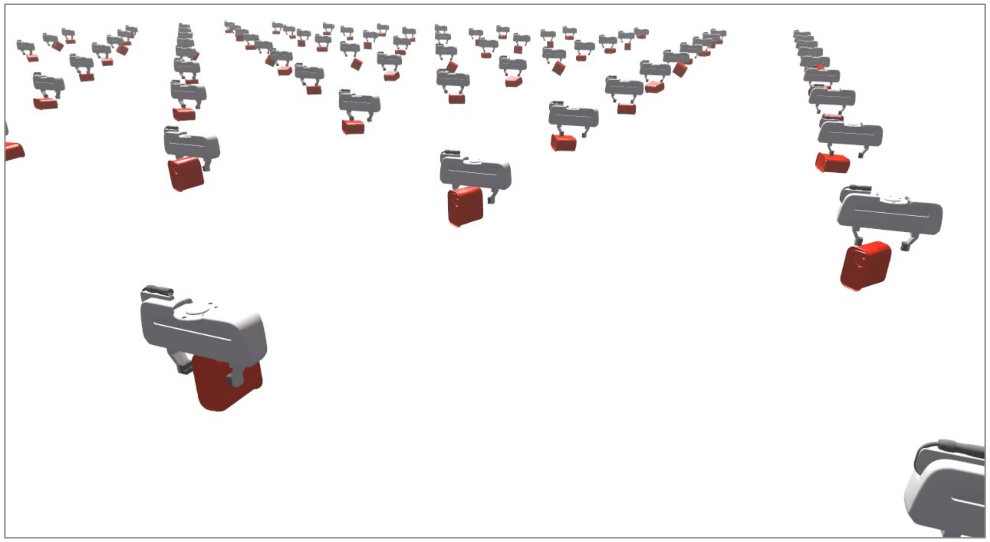
RGB input

ObjectSeeker
350K sim images
3.5M segments

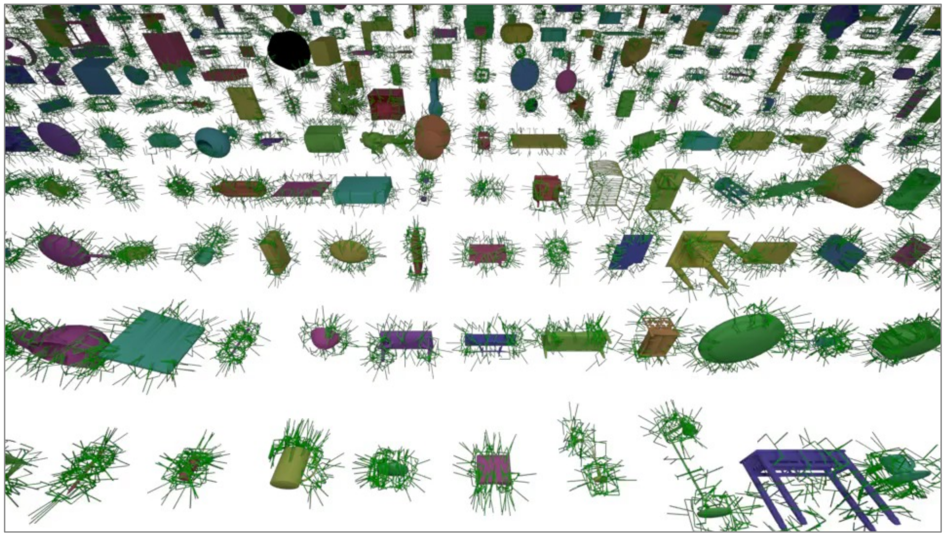
[Mousavian-Manuelli-Okorn-Xiang-Eppner-Murali-Fox, 2023]

PHYSICS-SIMULATION OF GRASPING

Isaac Sim can Assess Thousands of Grasps in Parallel



Sample Potential Grasps and Run Simulations to Assess Stability

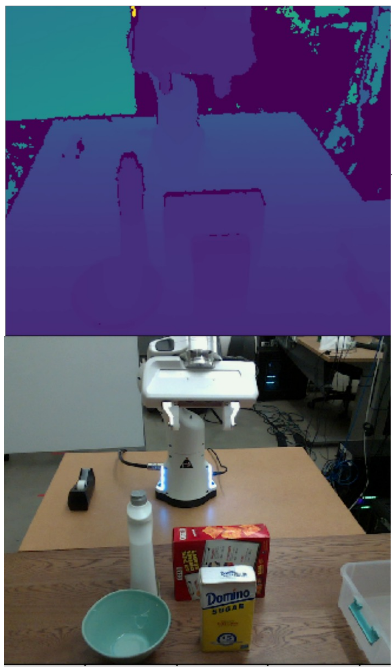


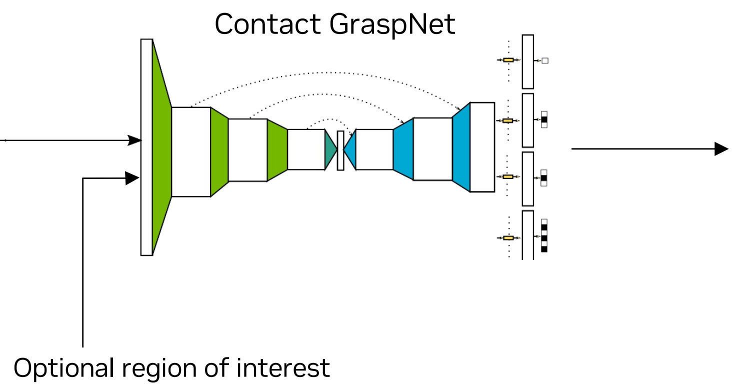
8,872 Objects Annotated with Successful Grasps

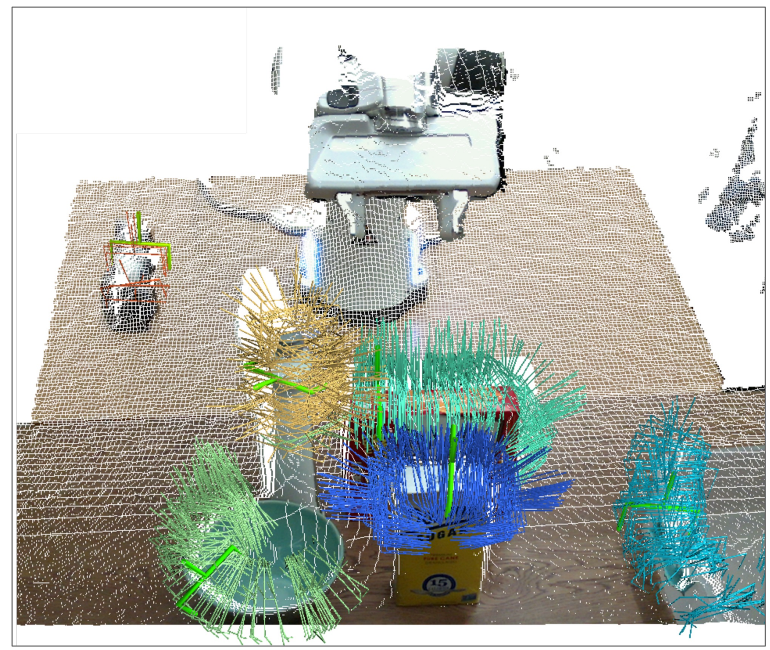
ACRONYM: [Eppner-Mousavian-F: ISRR-19, ICRA-21
ContactGraspNet: [Sundermeyer-Mousavian-Triebel-F: ICRA 2021]
GraspNet: [Mousavian-Eppner-F: ICCV-19]

Contact Graspnet

Generate 6D Grasp Poses from Input Point Clouds

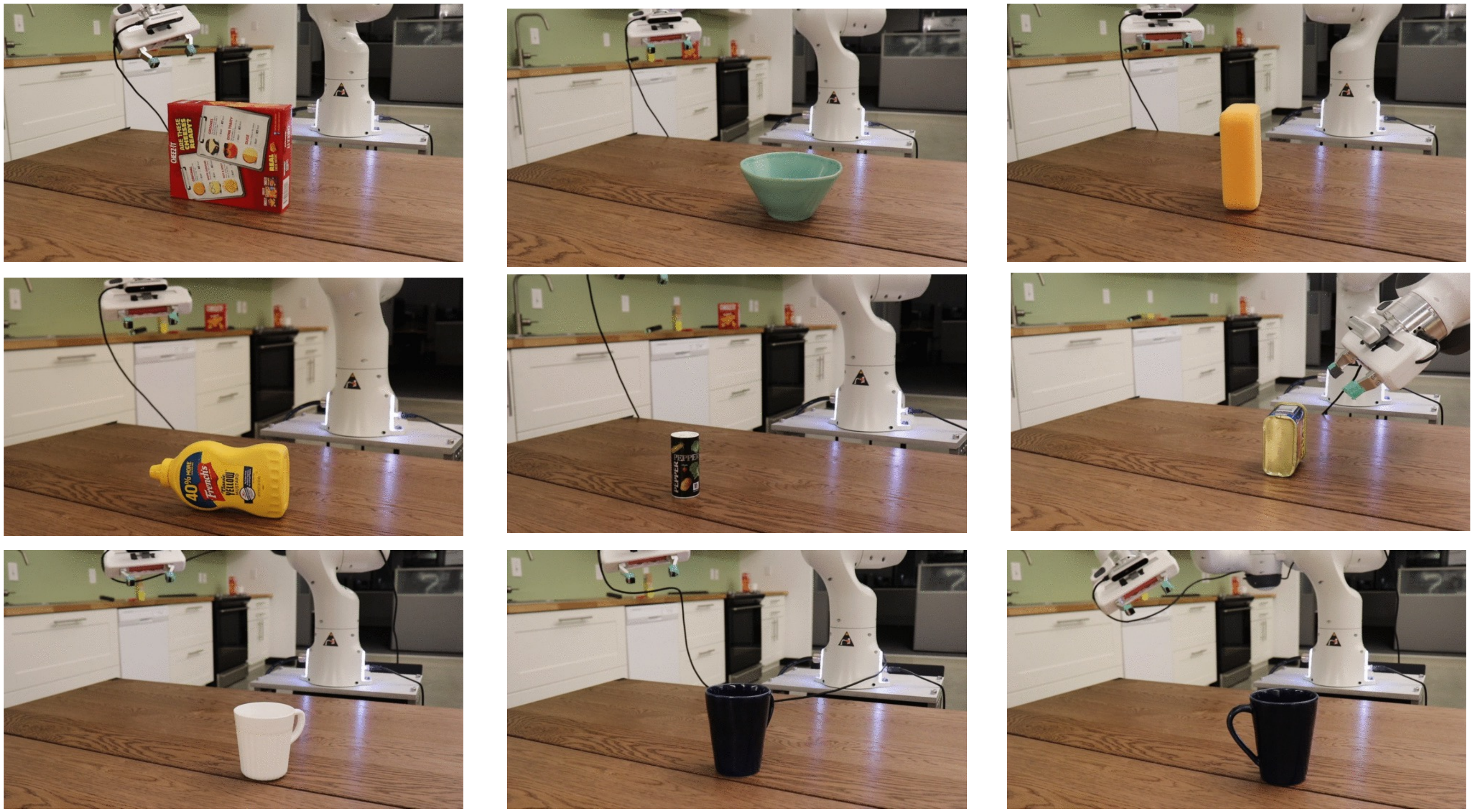






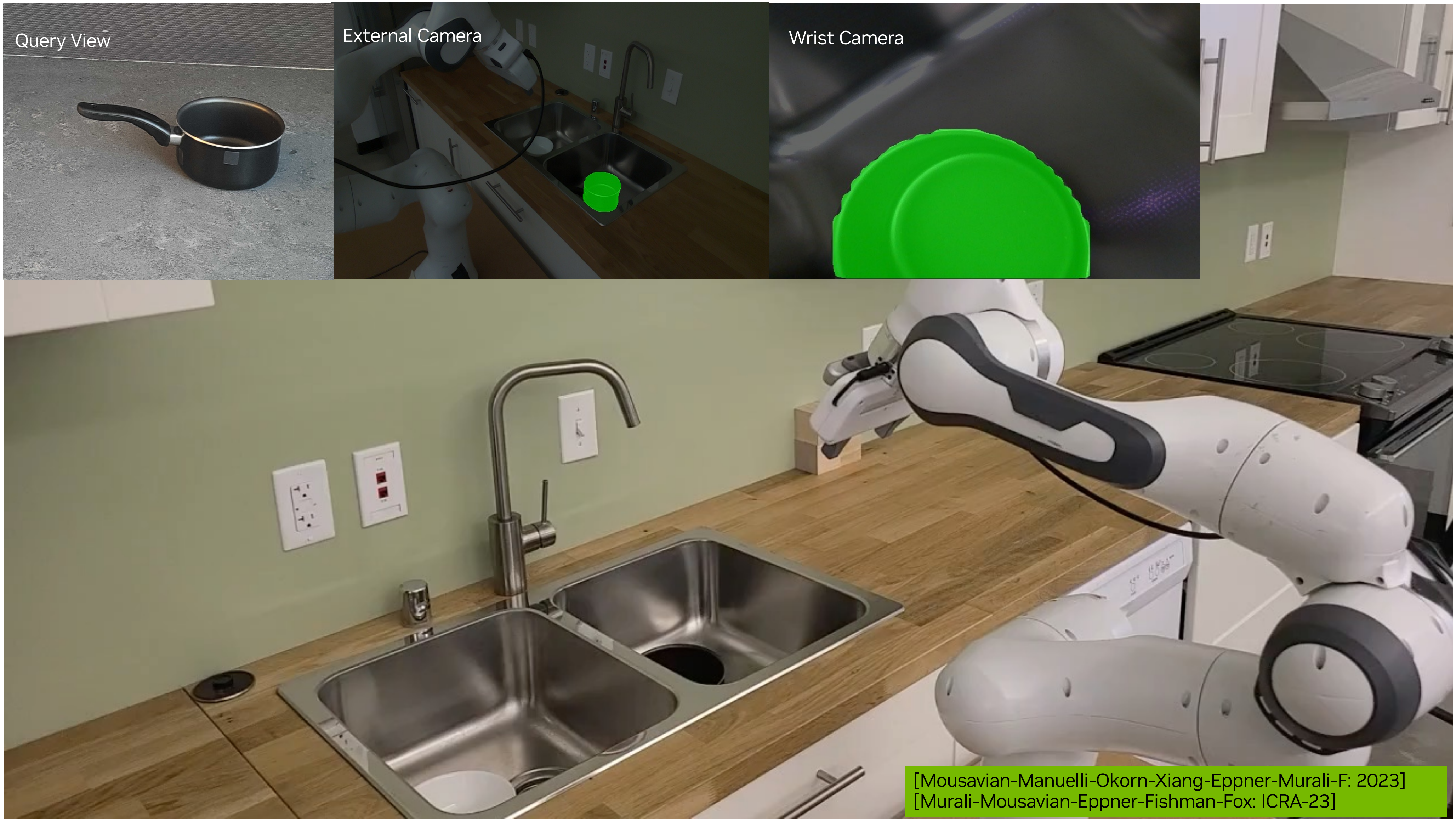
See also: [Mousavian-Eppner-F: ICCV-19]

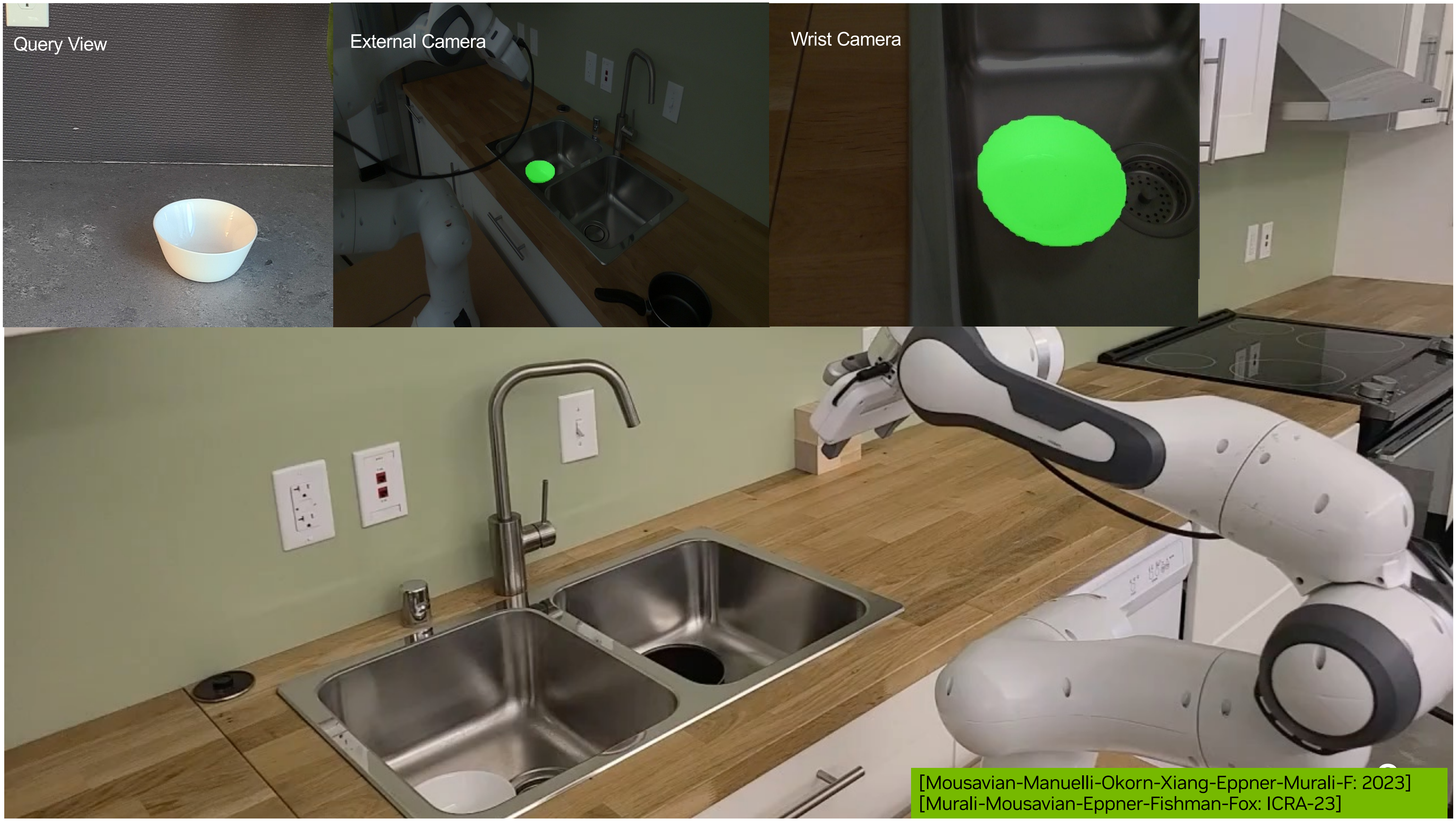
[Sundermeyer-Mousavian-Triebel-F: ICRA 2021]

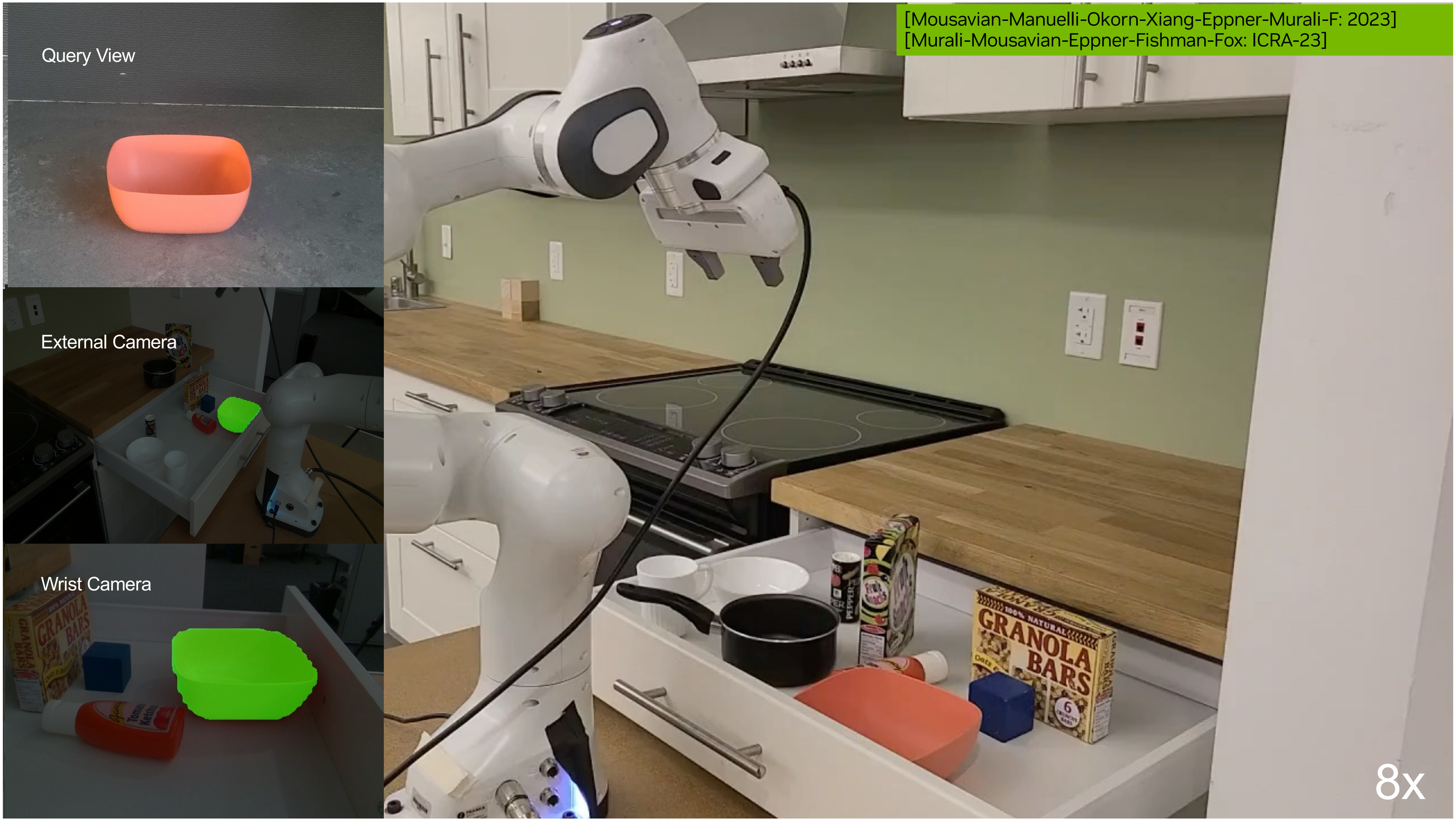


88% first attempt grasp success on unknown objects

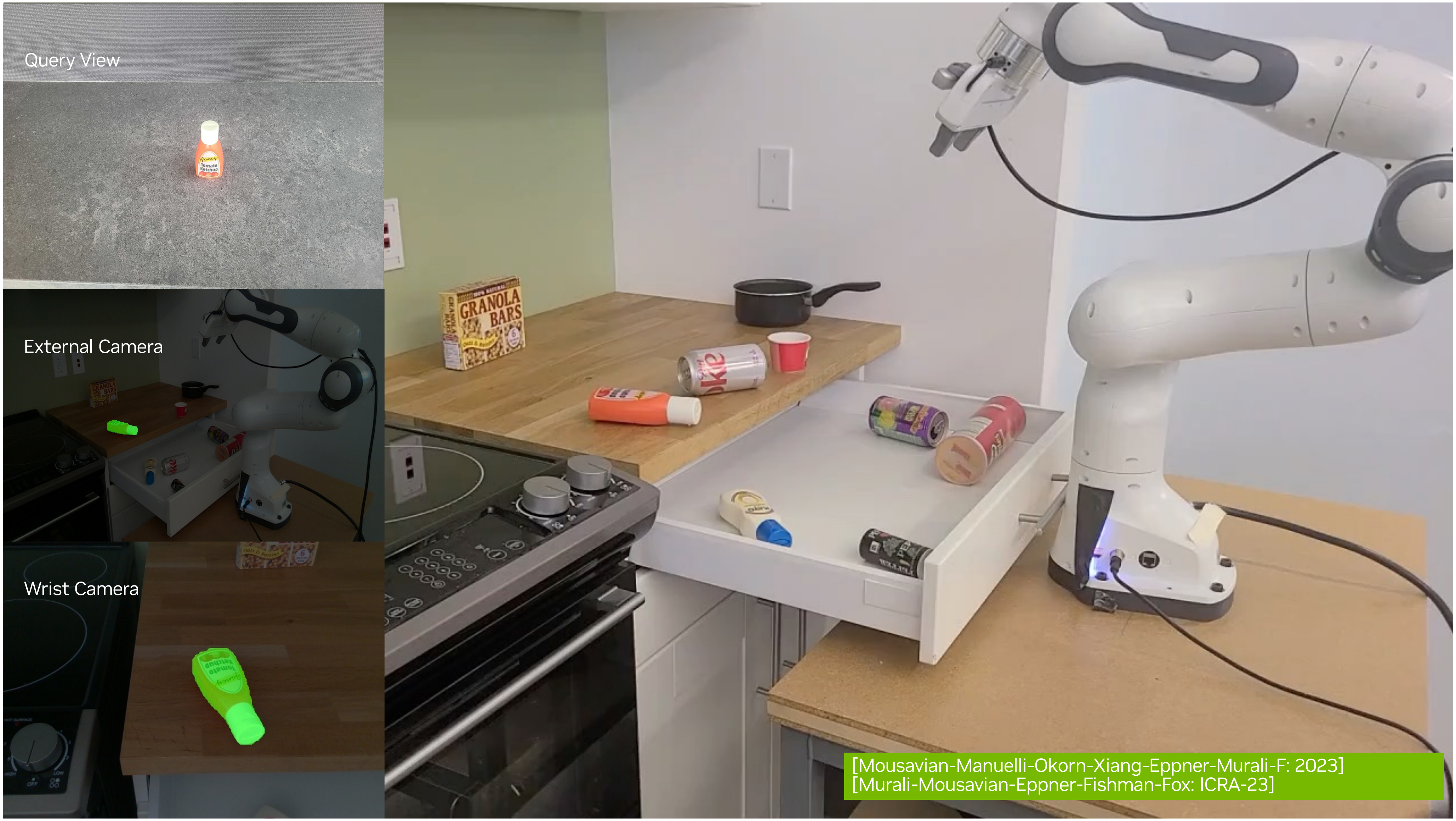
[6DOF-GraspNet: Mousavian-Eppner-F: ICCV-19]
Code and data available at: https://github.com/NVlabs/6dof-graspnet/


Query View
External Camera
Wrist Camera
[Mousavian-Manuelli-Okorn-Xiang-Eppner-Murali-F: 2023]
[Murali-Mousavian-Eppner-Fishman-Fox: ICRA-23]


Query View
External Camera
Wrist Camera
[Mousavian-Manuelli-Okorn-Xiang-Eppner-Murali-F: 2023]
[Murali-Mousavian-Eppner-Fishman-Fox: ICRA-23]



Query View
External Camera
Wrist Camera
[Mousavian-Manuelli-Okorn-Xiang-Eppner-Murali-F: 2023]
[Murali-Mousavian-Eppner-Fishman-Fox: ICRA-23]
8x
GRANOLA BARS



[Mousavian-Manuelli-Okorn-Xiang-Eppner-Murali-F: 2023]
[Murali-Mousavian-Eppner-Fishman-Fox: ICRA-23]

6-DOF GRASPING FOR CLUTTERED SCENES

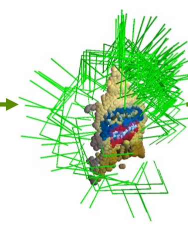
Extending Single Object Grasping to Cluttered Scenes

RGB-D Observation

Instance Segmentation

Cropped 3D Point Cloud

Grasps on Object



Grasps filtered by CollisionNet

CollisionNet efficiently reasons about gripper collisions with the scene, considering occluded areas as well

Instance segmentation: [Xie-Xiang-Mousavian-Fox: CoRL 2019, T-RO-21]; [Xiang-Xie-Mousavian-Fox: CoRL 2020]; [Xie-Xiang-Mousavian-Fox: CoRL-21]

[Murali-Mousavian-Eppner-Paxton-Fox, ICRA 2020]

GETTING AN OBJECT OUT OF CLUTTER

Deep Network Trained to Segment Scene, Generate Grasps, and Check for Collisions

External view

Gripper camera view

Target object is initially not reachable;
grasps will collide with surrounding clutter

[Murali-Mousavian-Eppner-Paxton-Fox: ICRA-20]

GETTING AN OBJECT OUT OF CLUTTER

Deep Network Trained to Segment Scene, Generate Grasps, and Check for Collisions

External view

Gripper camera view

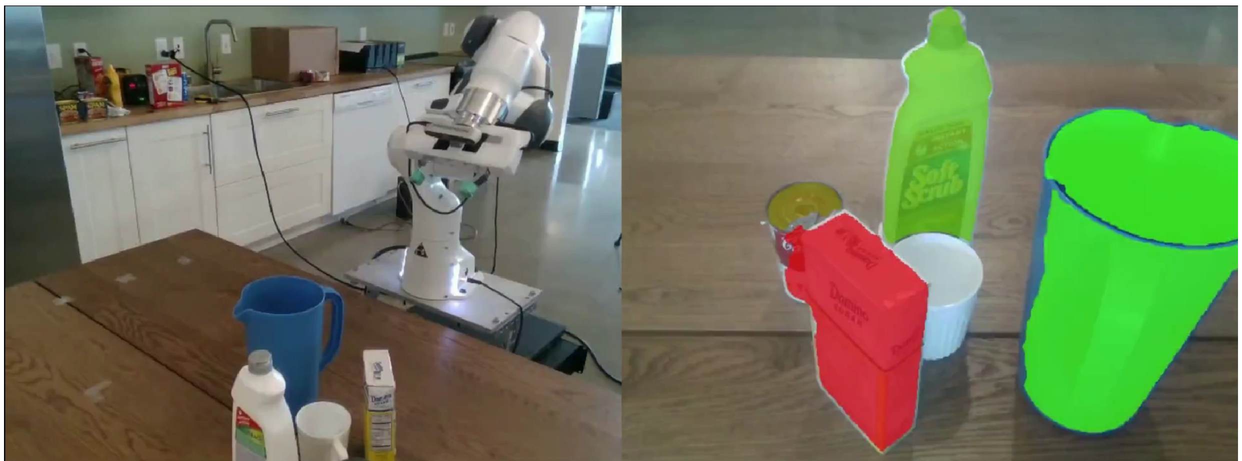
Target object is initially not reachable;
grasps will collide with surrounding clutter

[Murali-Mousavian-Eppner-Paxton-Fox: ICRA-20]

GETTING AN OBJECT OUT OF CLUTTER

Deep Network Trained to Segment Scene, Generate Grasps, and Check for Collisions

External view | Gripper camera view



Blocking objects are ranked
(**red** has the highest score and **green** is the lowest)

[Murali-Mousavian-Eppner-Paxton-Fox: ICRA-20]

GETTING AN OBJECT OUT OF CLUTTER

Deep Network Trained to Segment Scene, Generate Grasps, and Check for Collisions

External view

Gripper camera view

Blocking object is selected

[Murali-Mousavian-Eppner-Paxton-Fox: ICRA-20]

GETTING AN OBJECT OUT OF CLUTTER

Deep Network Trained to Segment Scene, Generate Grasps, and Check for Collisions

External view

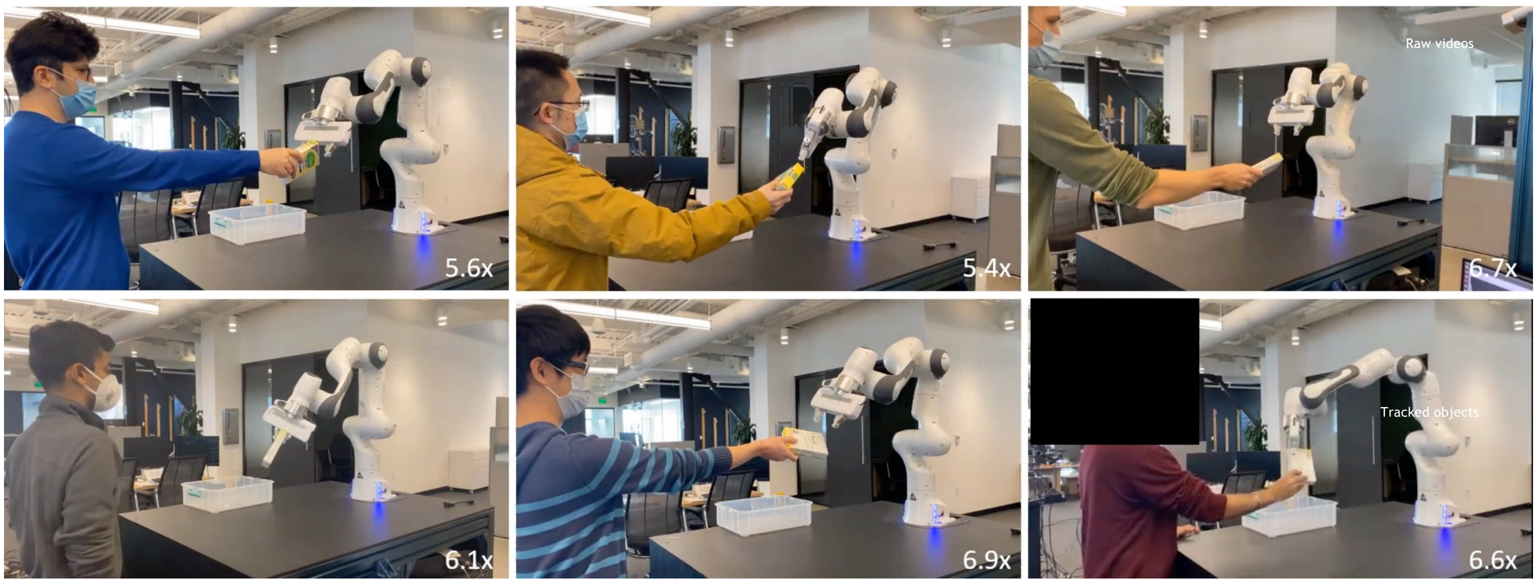
Gripper camera view

Blocking object is removed from the scene

[Murali-Mousavian-Eppner-Paxton-Fox: ICRA-20]

HANDOVER OF UNKNOWN OBJECTS

Continuously Detect Hand/Object, Determine Safe Grasp, and Control



- Tracking and segmentation of hand and objects enables robot to approach grasps that are safe and stable
- Large-scale data set for training and benchmarking hand tracking with object interactions

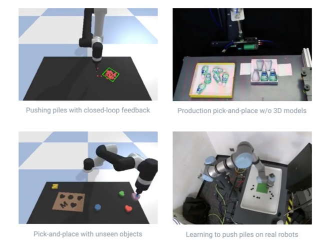
[Chao-Yang-Xiang-Molchanov-Handa-Tremblay-Narang-Van Wyk-Iqbal-Birchfield-Kautz-Fox: CVPR-2021]
[Yang-Paxton-Mousavian-Chao-Cakmak-Fox: ICRA-21]

LEARNING
ACTION-CENTRIC ANIPULATION WITH LANGUAGE INSTRUCTIONS

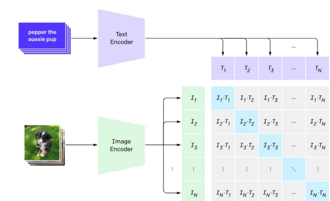
CLIPORT

Efficiently Teach Manipulation Tasks Leveraging Language Instructions

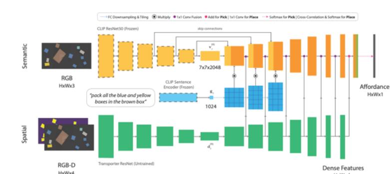
- **TransporterNets** learn precise pick-and-place skills
  - Actions specified in visual space
  - No object models, poses, or segmentations needed
  - No semantics, weak generalization, one network per task
- **CLIP** generates aligned image and text embeddings
  - Semantics via language-vision training, robust visual features
  - Not immediately suited for manipulation tasks
- **CLIPort** combines language reasoning with precise manipulation
  - Inherits manipulation capabilities from TransporterNets
  - Language enables training single, multi-task model
  - Some semantic transfer across tasks
  - Only 2D top-down manipulation (just like TransporterNets)



TransporterNets
[Zeng et. al, CoRL-2020]



CLIP
[Radford et. al, 2021]



CLIPort
[Shridhar et. al, CoRL-2021]

ıridhar-Manuelli-Fox: CoRL-2021]

DATA COLLECTION

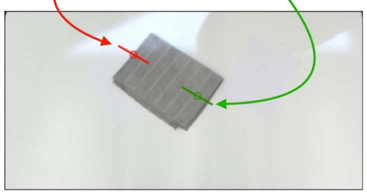
Folding Task

9 examples

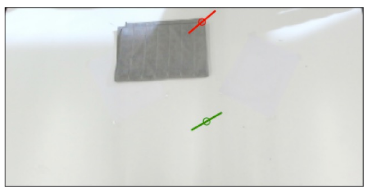
Data collection time: ~10 min

Pick Label

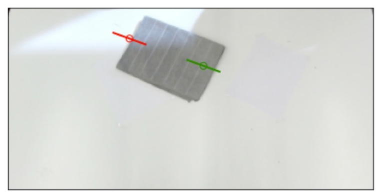
Place Label



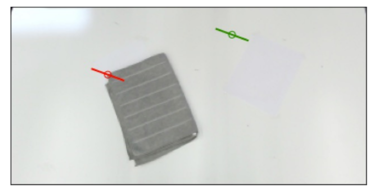
“fold the cloth in half”



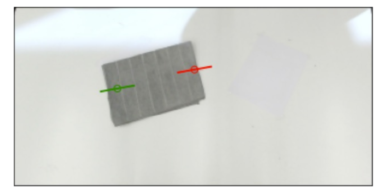
“unfold the cloth”



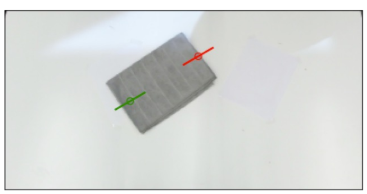
“fold the cloth in half”



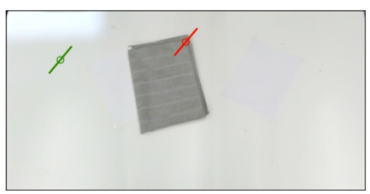
“unfold the cloth”



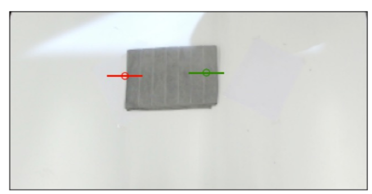
“fold the cloth in half”



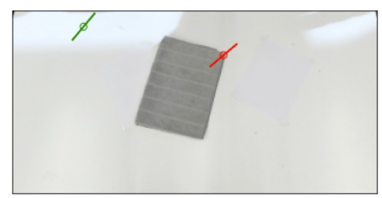
“fold the cloth in half”



“unfold the cloth”



“fold the cloth in half”

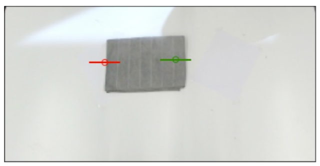


“unfold the cloth”

Data Collection

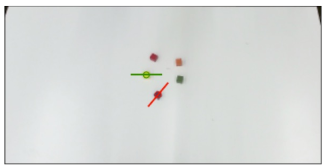
179 total examples

Folding Task
9 examples
~10 min



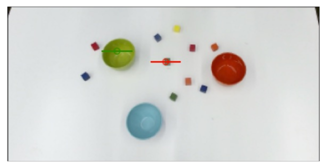
"fold the cloth in half"

Stacking Task
13 examples
~10 min



"put the blue block on the yellow block"

Put in Bowl Task
10 examples
~10 min

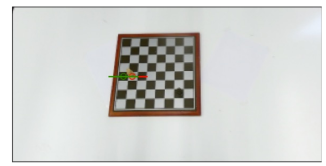


"put the green blocks in the green bowl"

Packing Task
31 examples
~30 min

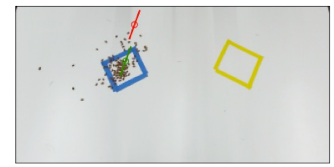
"pack the screwdriver in the brown box"

Move Rook Task
29 examples
~40 min



"move the rook one block right"

Sweeping Task
23 examples
~80 min

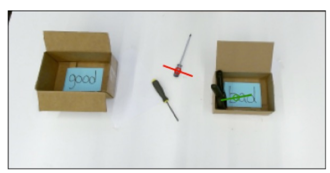


"sweep the beans into the blue zone"

Cherry Task
26 examples
~20 min

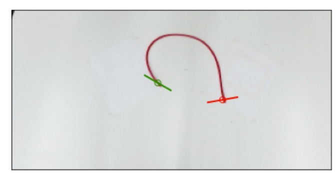
"pick all the cherries and put them in the box"

Reading Task
26 examples
~30 min

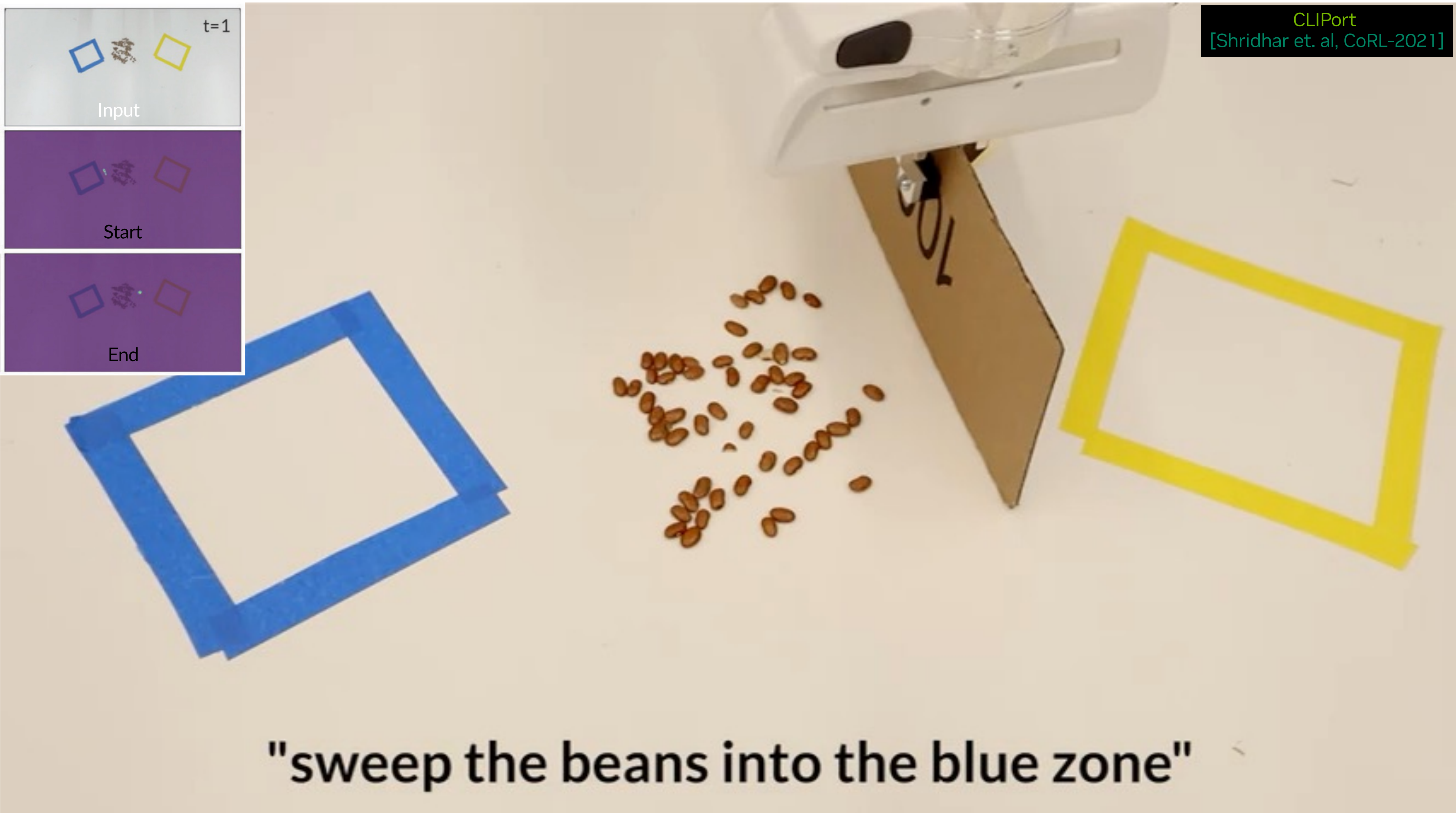


"put the blue screwdriver in the bad box"

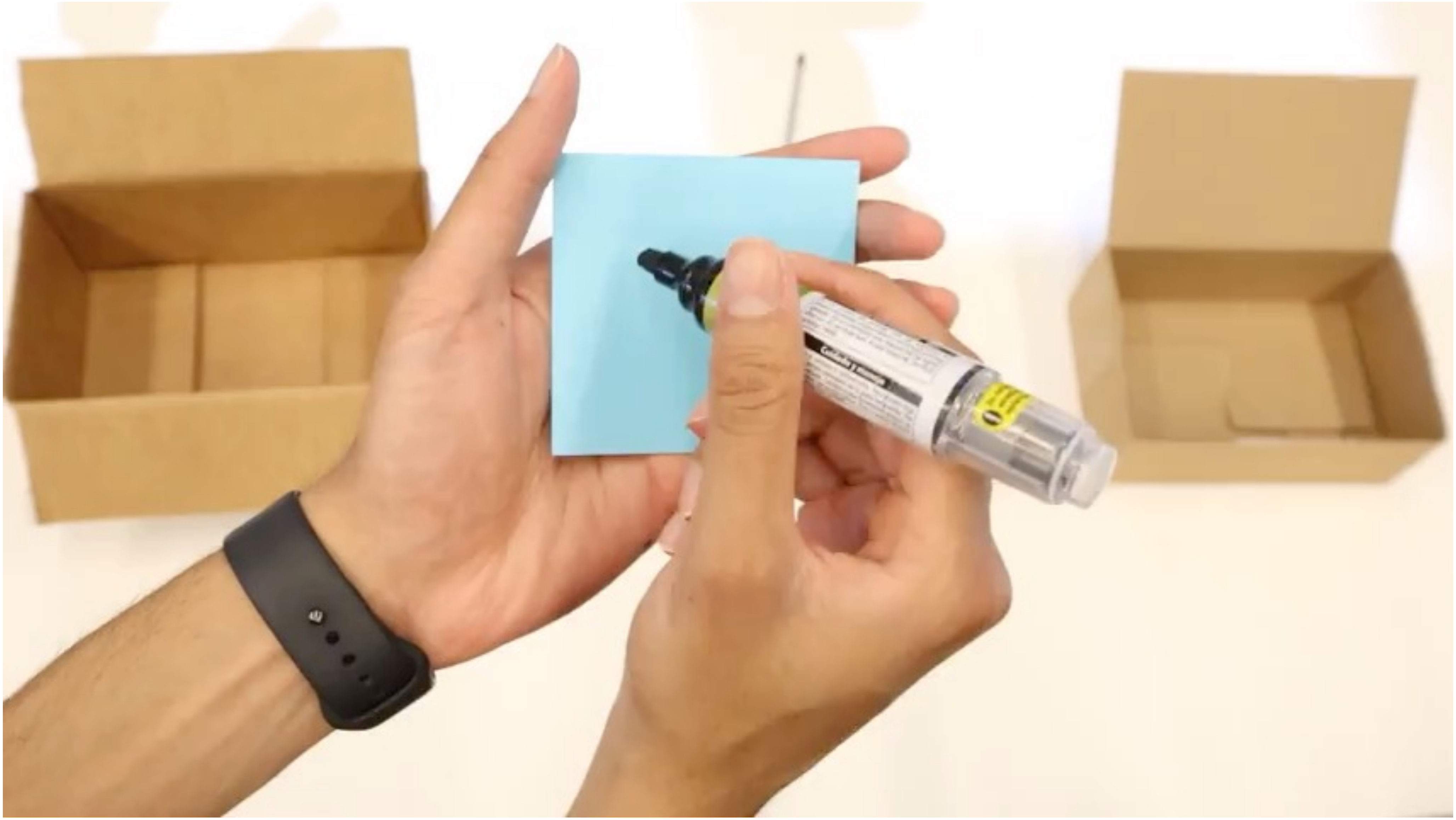
Rope Task
12 examples
~15 min



"close the loop"


t=1
Input
Start
End
CLIPort
[Shridhar et. al, CoRL-2021]
"sweep the beans into the blue zone"

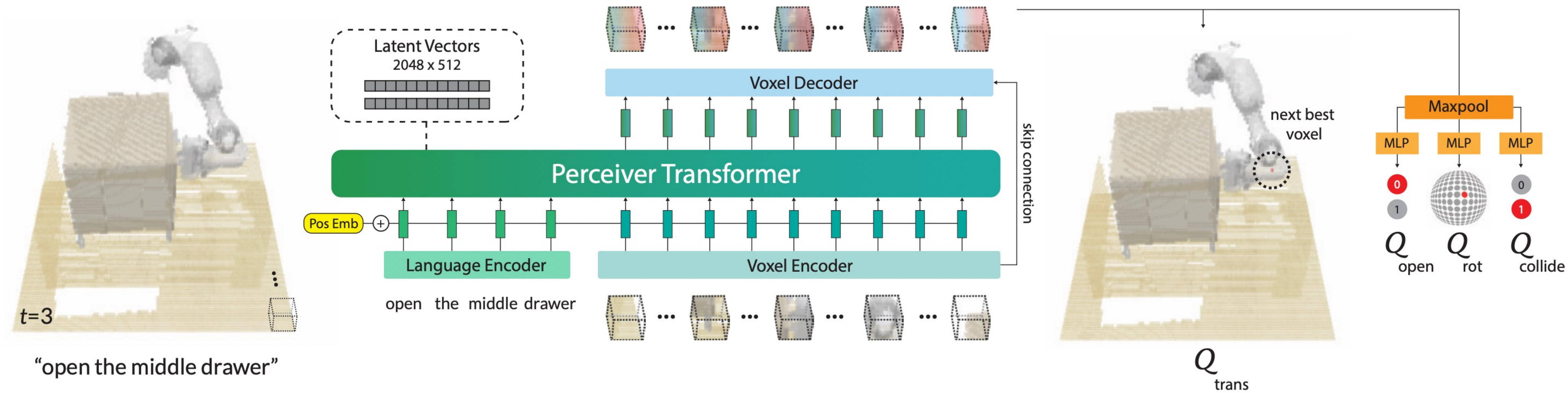

t=2
Input
Pick
Place
CLIPort
[Shridhar et. al, CoRL-2021]



[Shridhar-Manuelli-Fox: CoRL-2021]

PERCEIVER ACTOR

Predicting 3D Pose / 3D Orientation of Next Gripper Action



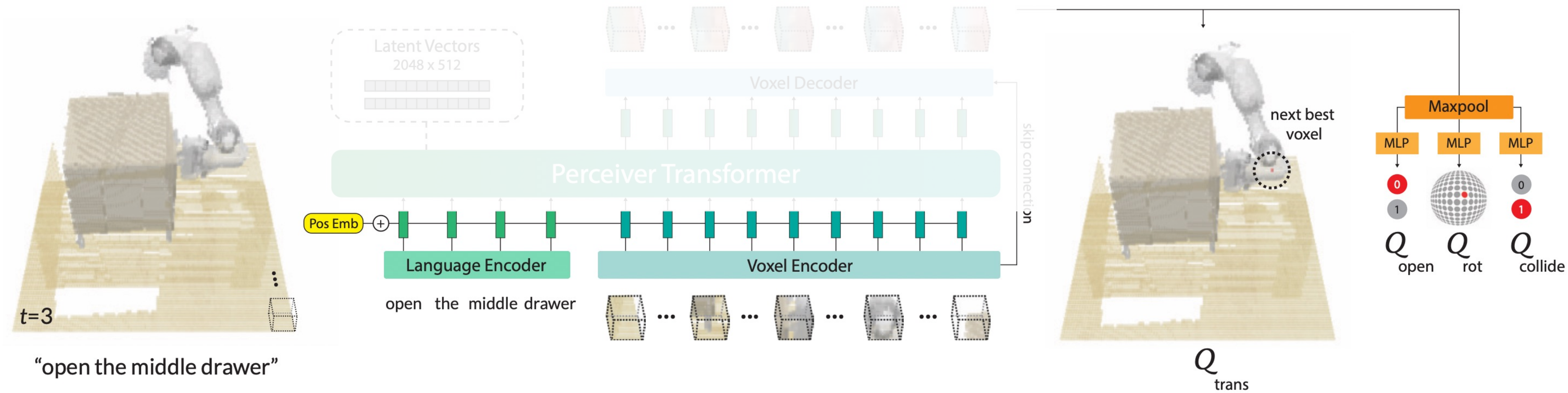
- Scene representation: 100^3 voxels at 1cm resolution (occupancy, color)
- Input: 20^3 = 8,000 tokens (each over 5^3 voxels) and text for task specification
- Output: Next gripper pose and status (softmax over voxels)
  (3D translation at 1cm resolution, 3D rotation at 5deg resolution)

- Significantly outperforms multi-level U-net structure of C2F-ARM [James etal: CVPR-22]

[Shridhar-Manuelli-Fox: CoRL-2021]

PERCEIVER ACTOR

Predicting 3D Pose / 3D Orientation of Next Gripper Action



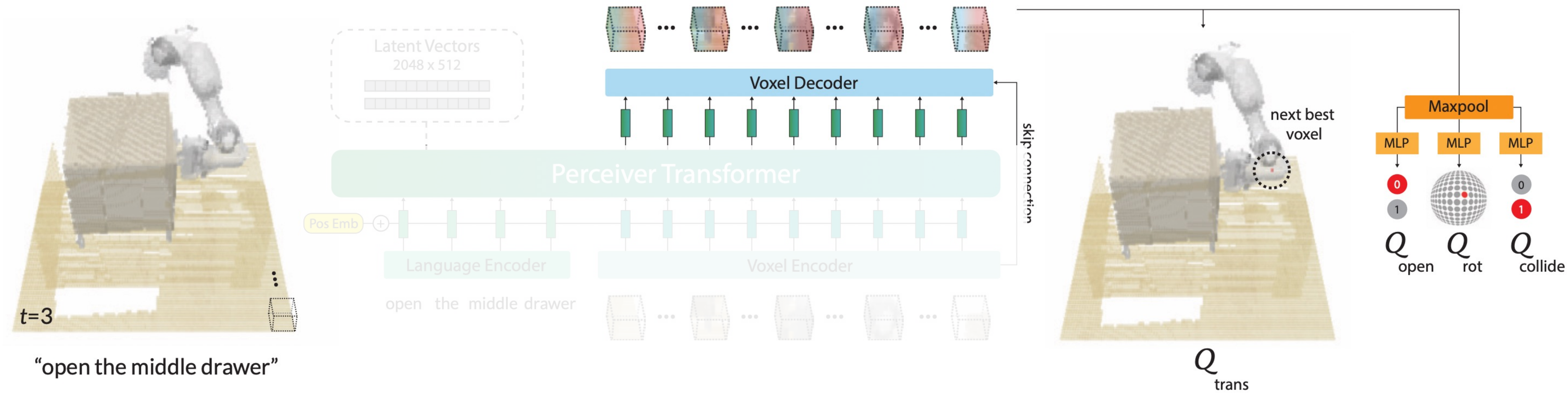
- Scene representation: 100^3 voxels at 1cm resolution (occupancy, color)
- Input: 20^3 = 8,000 tokens (each over 5^3 voxels) and text for task specification
- Output: Next gripper pose and status (softmax over voxels)
  (3D translation at 1cm resolution, 3D rotation at 5deg resolution)

- Significantly outperforms multi-level U-net structure of C2F-ARM [James etal: CVPR-22]

[Shridhar-Manuelli-Fox: CoRL-2021]

PERCEIVER ACTOR

Predicting 3D Pose / 3D Orientation of Next Gripper Action

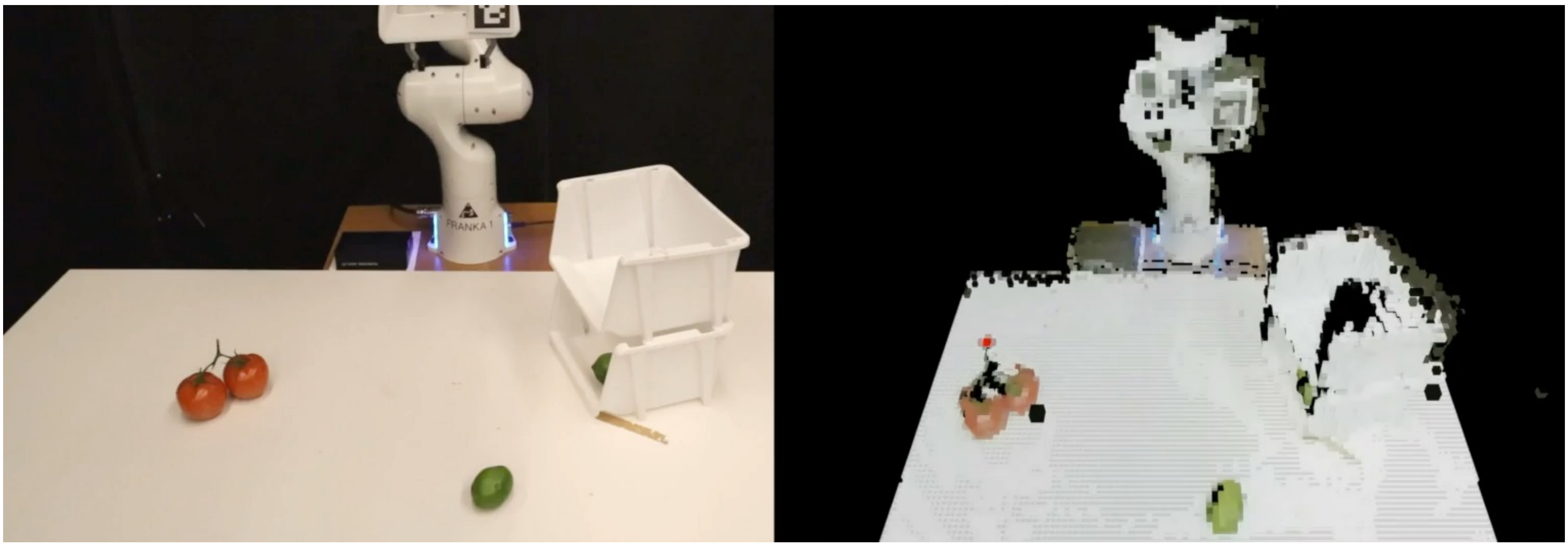


- Scene representation: 100^3 voxels at 1cm resolution (occupancy, color)
- Input: 20^3 = 8,000 tokens (each over 5^3 voxels) and text for task specification
- Output: Next gripper pose and status (softmax over voxels)
  (3D translation at 1cm resolution, 3D rotation at 5deg resolution)

- Significantly outperforms multi-level U-net structure of C2F-ARM [James etal: CVPR-22]

EXAMPLE EXECUTION: *PUT THE TOMATOES IN THE TOP BIN*

Single Command Input, at Each Step PerAct Predicts Next Gripper Pose



These clips are from **one multi-task agent**
trained with just **53 demos**

SIMULATION FOR ROBOT TRAINING AND DEVELOPMENT

- Models of kitchen cabinets, objects, and robot have to be physically accurate (masses, frictions, articulations, ...) and photorealistic
- Isaac Sim with Physics engine (Flex, PhysX)
- Johnny Costello: that was harder than building model of the death star

[Narang, Storey, Akinola, Macklin, Reist, Wawryzniak, Guo, State, Moravanszky, Lu, Handa, Fox: RSS 2022]

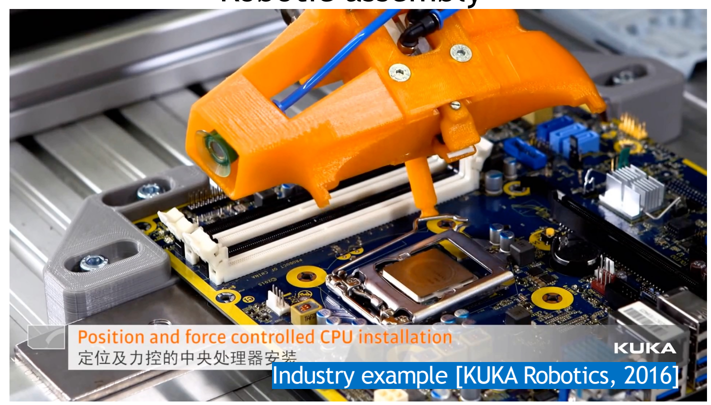
CONTACT-RICH ROBOTIC ASSEMBLY

Manual assembly (status quo)

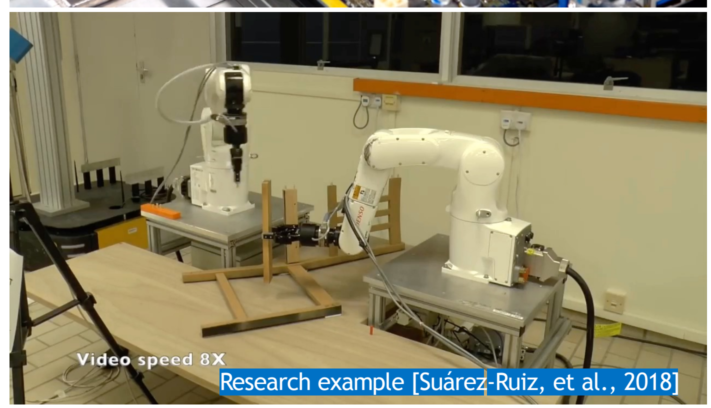

Automotive assembly [Assembly Magazine, 2021]


Aerospace assembly [Assembly Magazine, 2015]

Robotic assembly

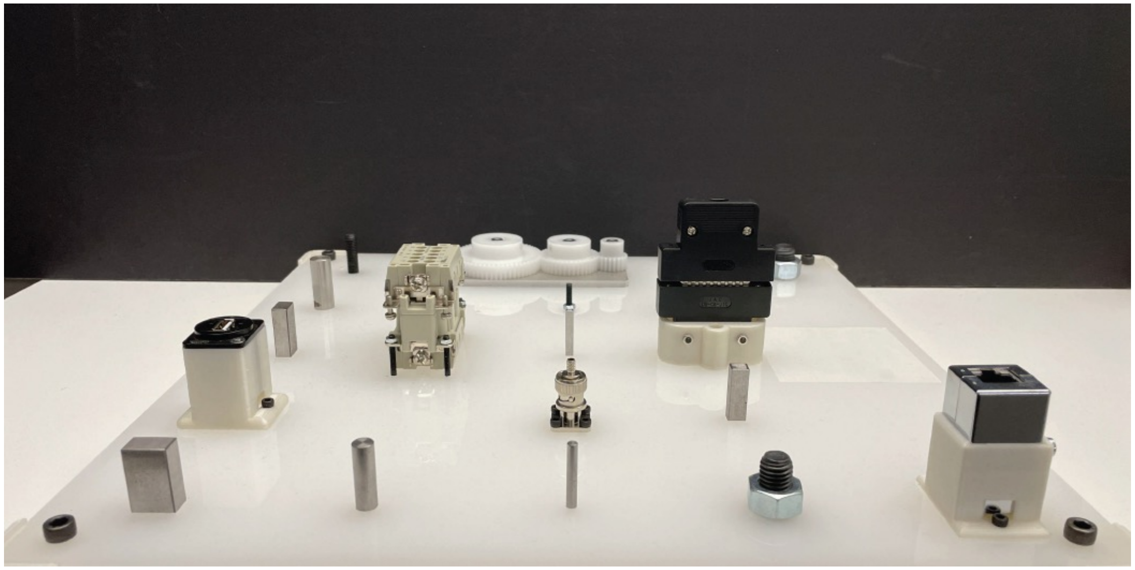


Industry example [KUKA Robotics, 2016]



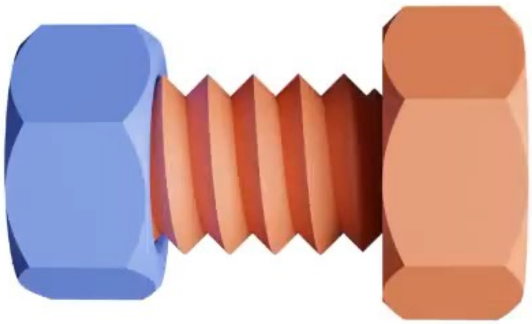
Research example [Suárez-Ruiz, et al., 2018]

INDUSTRIAL ASSEMBLY



NIST Benchmark for Assembly

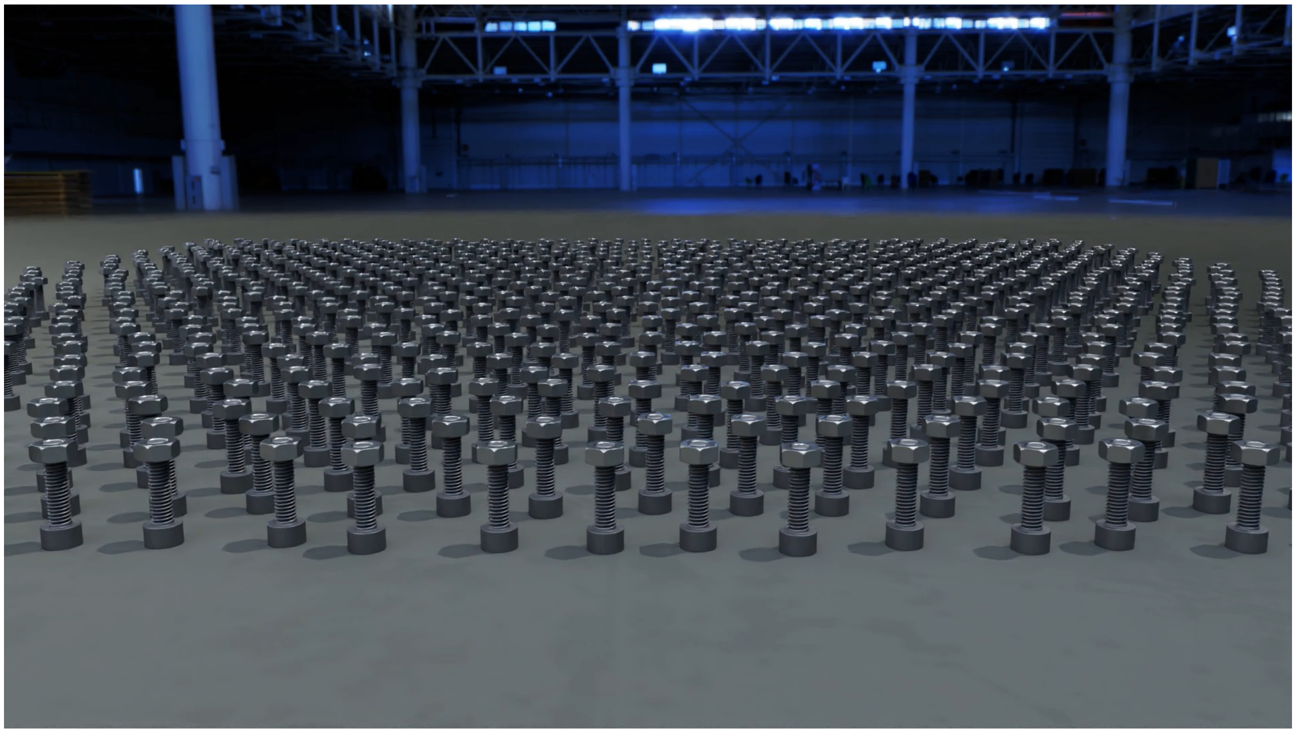
Round and rect. pegs/holes
Nuts/bolts
Gear assembly
Electrical connectors

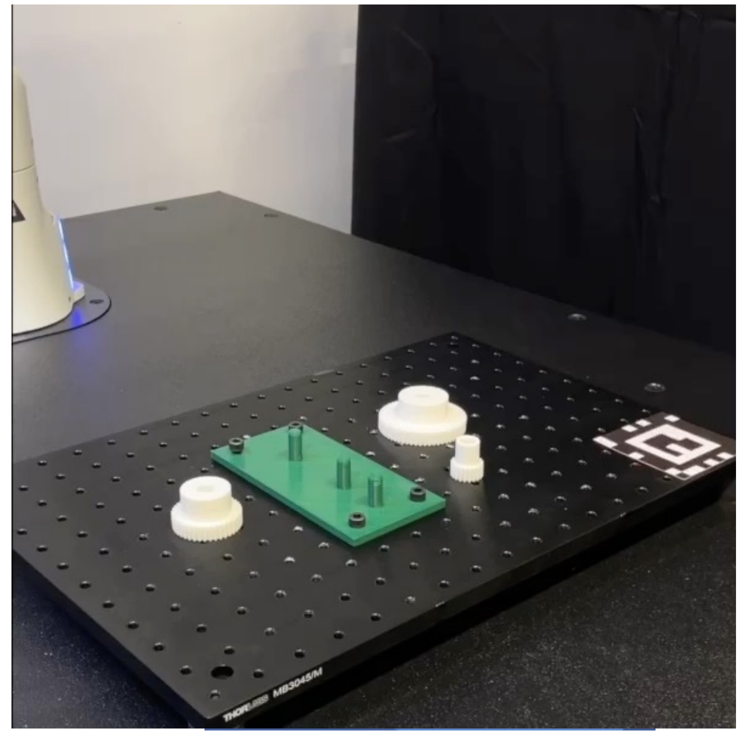


1/350 real-time [Ferguson, et al., 2020]

FACTORY / INDUSTREAL

GPU-optimized Simulation of Contact-Rich Tasks: 20,000 x Speedup + Higher Precision



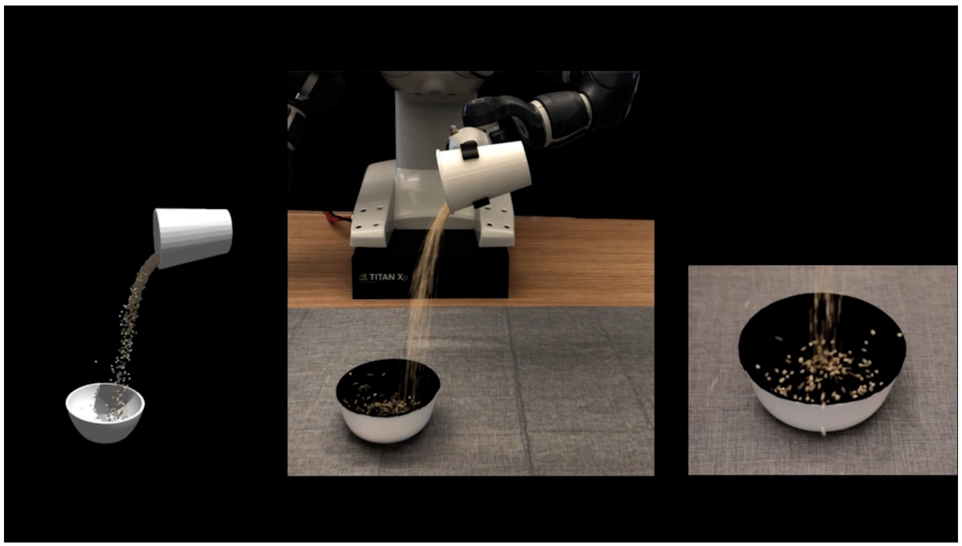


3 simulation environments spanning rigid NIST board tasks; includes 7 real-world robot controllers

[Narang-Akinola-Guo-Handa-Lu-Macklin-Moravanszky-Reiss-State-Storey-Wawrzyniak-F: RSS-22]
[Tang-Lin-Narang-Akinola-Handa-Sukhatme-Ramos-F: RSS-23]

SIMULATING GRANULAR MEDIA

Material Properties Estimated from Real Data

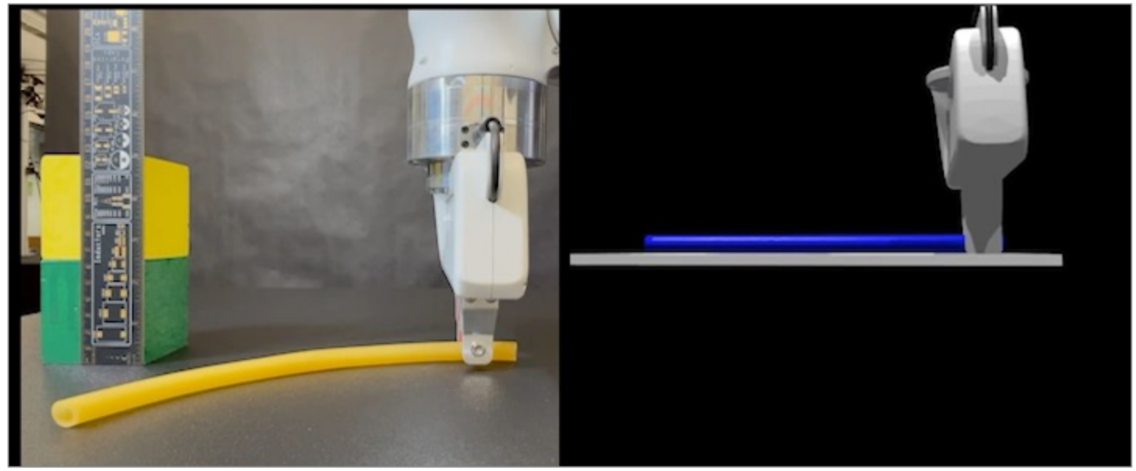


[Matl-Narang-Baijcsy-Ramos-F: ICRA-20]

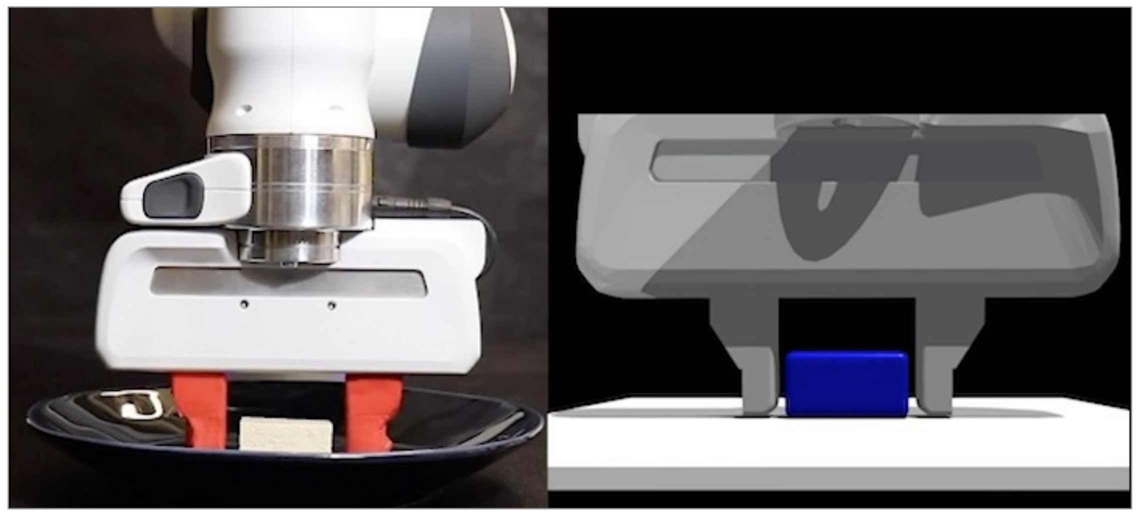
Deformable objects and granular media

- Simulation matches real world behavior very well (w/ off the shelf material parameters)
- Sim parameters can be adjusted to real world data

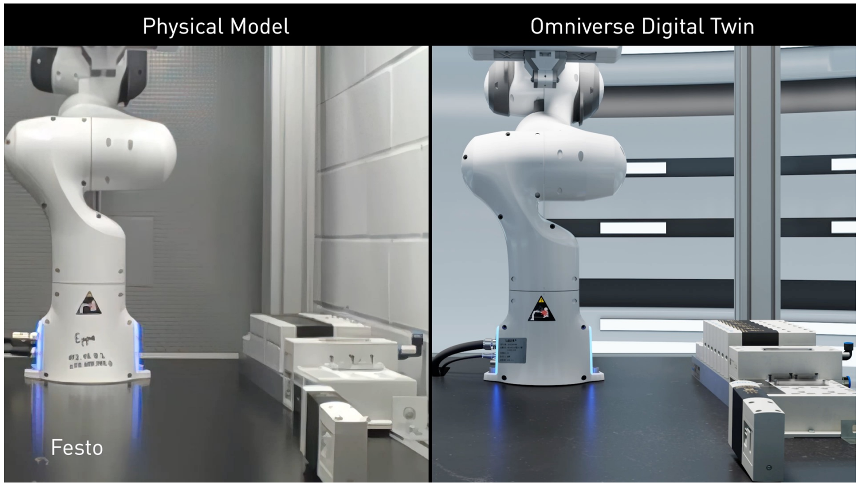
[Huang-Narang-Eppner-Sundaralingam-Macklin-Hermans-F: RA-L-22]
[Matl-Narang-Ramos-F: ICRA-20]
[Ramos-Posas-F: RSS-19]



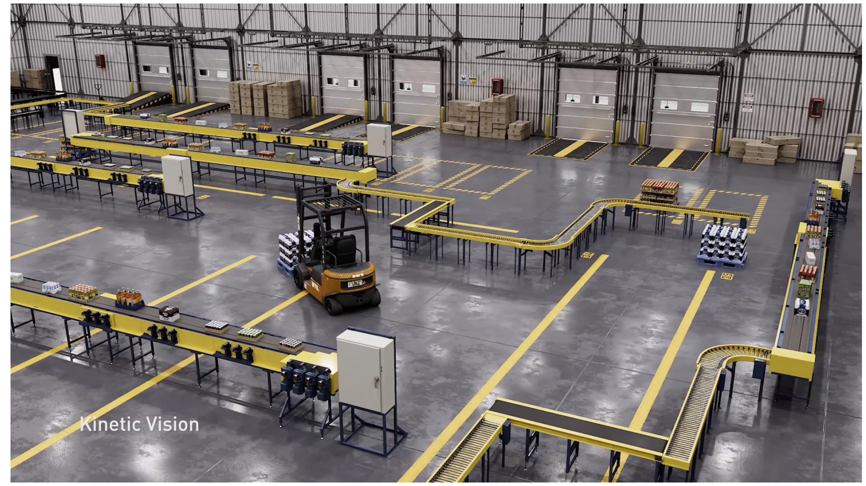
Tube deformation



Grasping and squeezing tofu



Festo



PepsiCo

Scaling via Omniverse and Isaac Sim

- Digprocesses
- Complete workflows to safely develop, train, and validate
- Introspection into what the robot observes and is planning
- ital Twins for designing and programming industrial

Amazon

TOWARD OBJECT MANIPULATION WITHOUT EXPLICIT MODELS

- Explicit object models enable reasoning for complex manipulation tasks, but models are often not available and modeling and object pose estimation errors result in brittle execution
- Learning to map raw observations (s.a. point clouds, images) directly to manipulation relevant properties (*e.g.* segmentation, grasps, collisions, spatial relations) enables robust manipulation of unknown objects
- CLIPort / PerAct: Combining pre-trained language-vision models with manipulation-specific representations enables highly data efficient teaching of manipulation tasks using **action-centric representations**

- Physics-based, photo-realistic simulation of manipulation tasks is within reach
- Allows safe and scalable training and development leveraging ground truth states for labeling and demonstration generation
- Controlled environments for development and benchmarking