Learning Robotic Manipulation from Human Demonstration Videos

Yu Xiang

Assistant Professor

Intelligent Robotics and Vision Lab

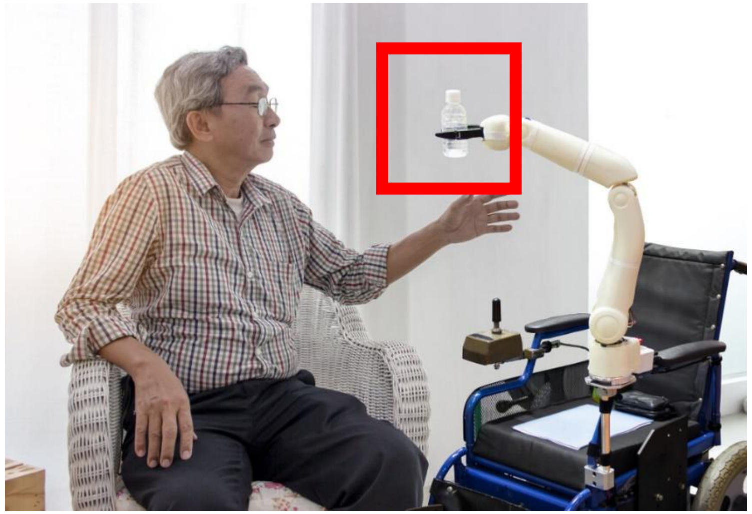
The University of Texas at Dallas

5/19/2025

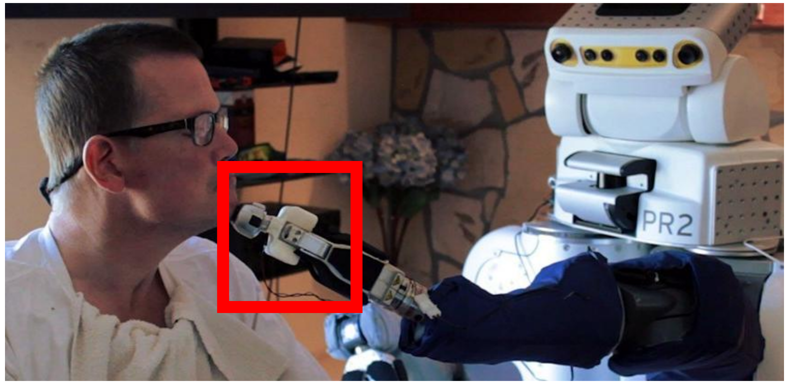
Stanford Vision and Learning Lab

Future Intelligent Robots in Human Environments

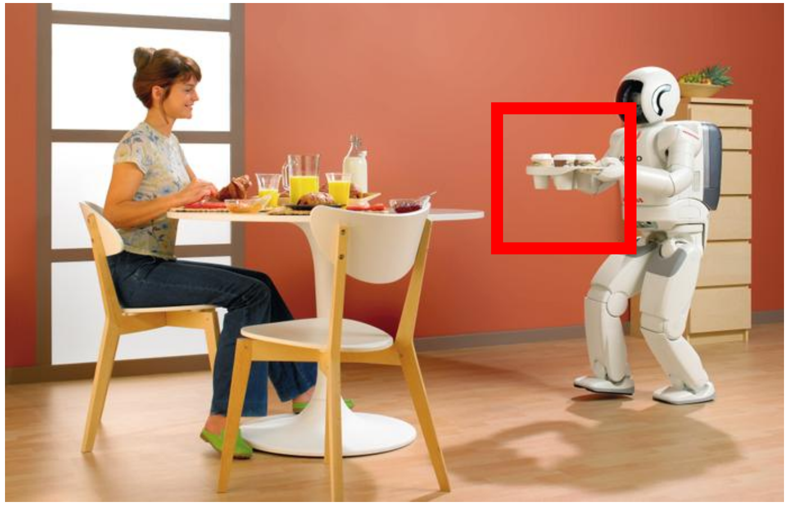
Manipulation



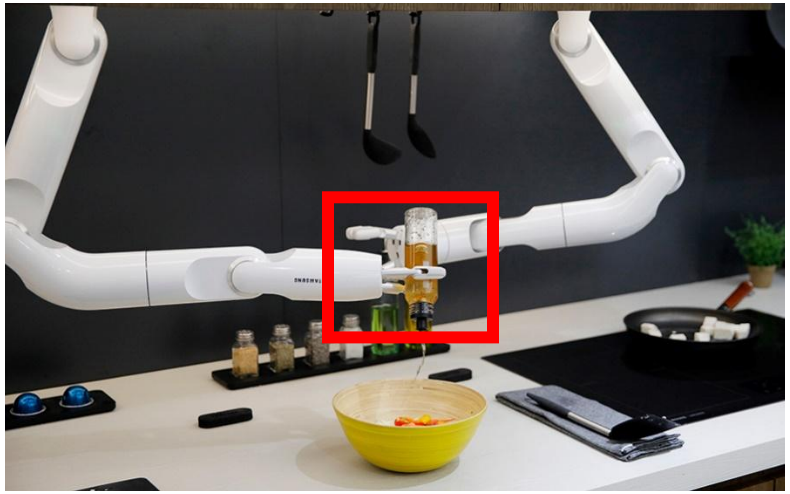
Senior Care



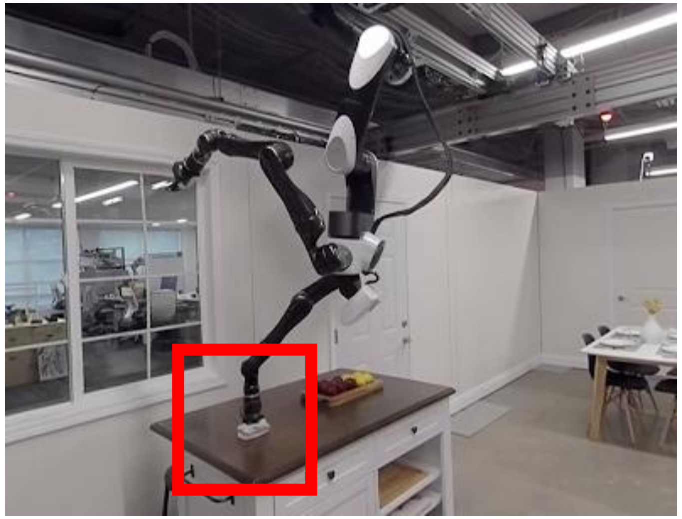
Assisting



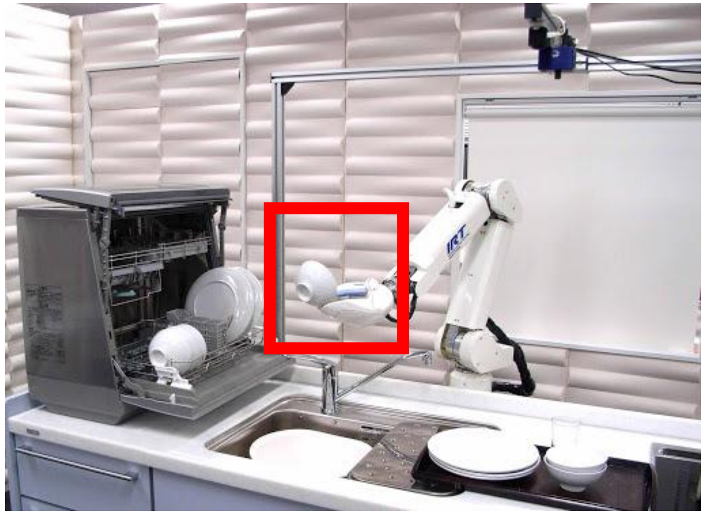
Serving



Cooking



Cleaning

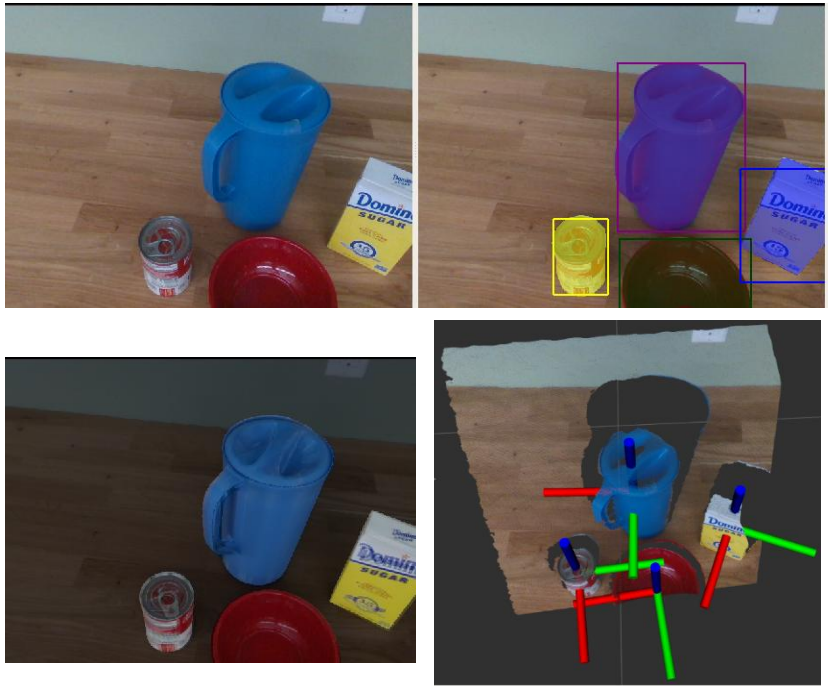


Dish washing

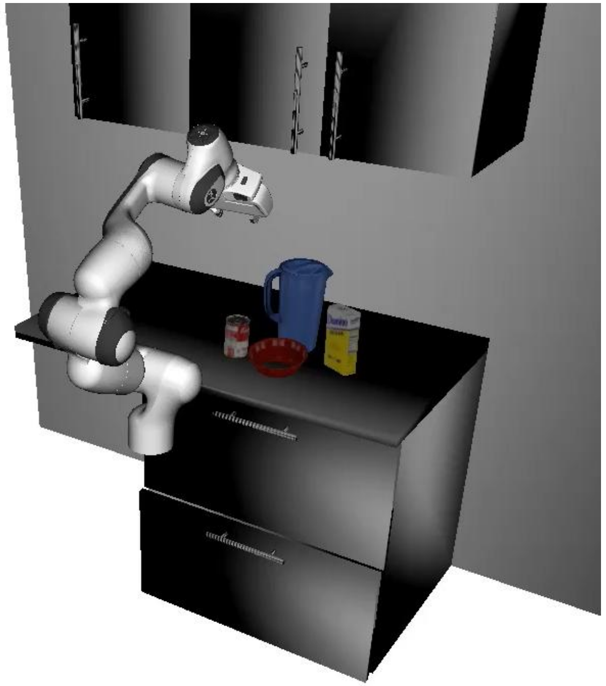
“Traditional” Approach for Robot Manipulation

Perception → Planning → Control

6D object pose estimation



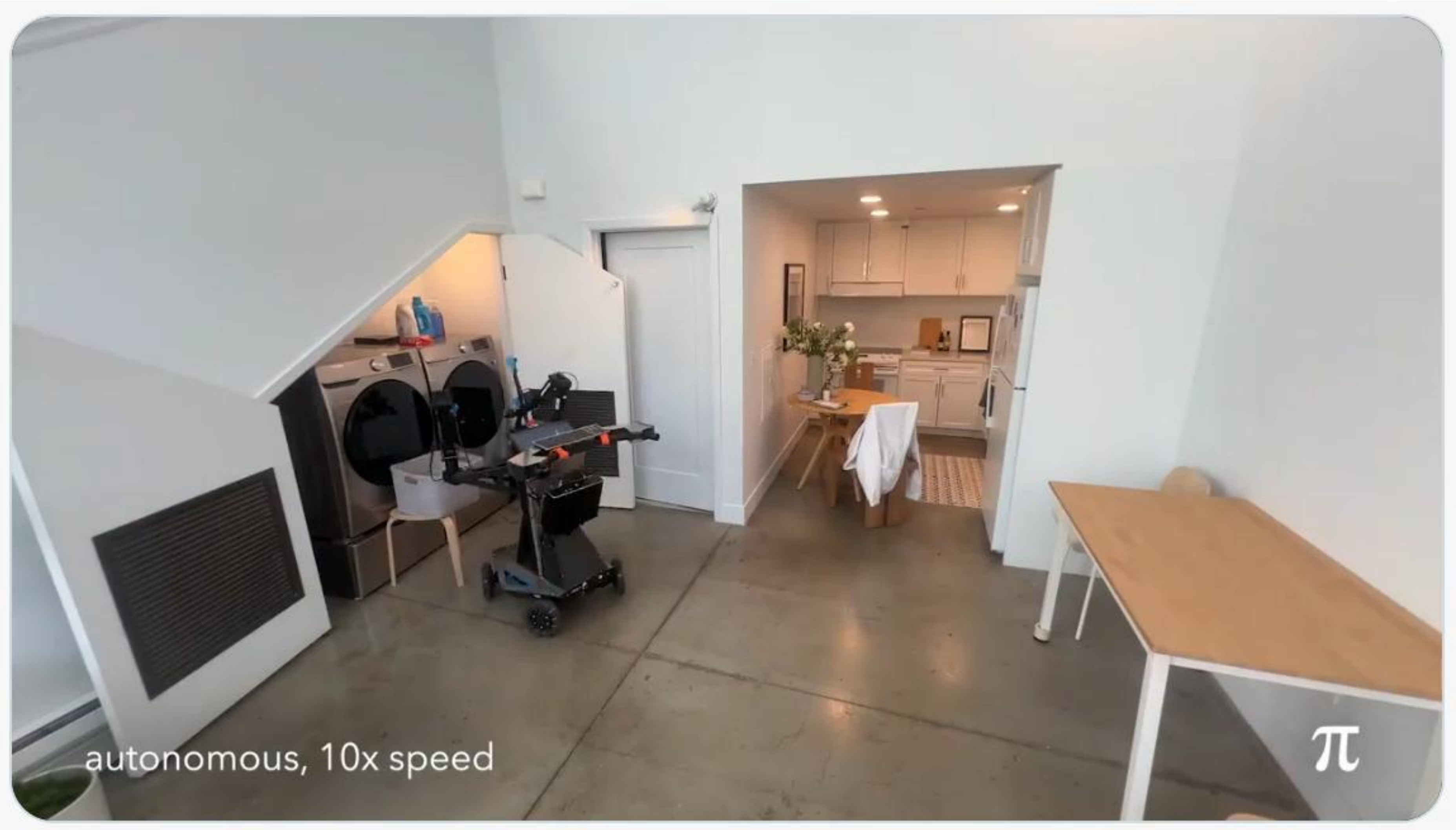
Grasp planning and motion planning



Manipulation trajectory following

Hard code the logics for manipulation based on perception and planning

Some Recent Breakthroughs



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

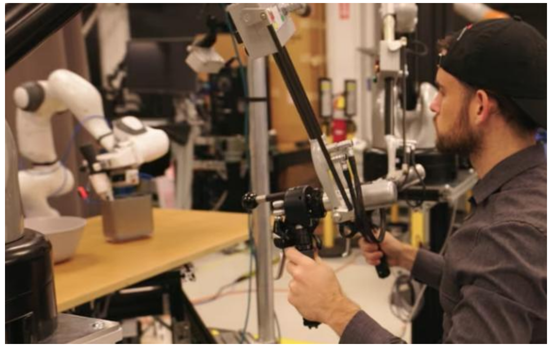
Some Recent Breakthroughs

Mobile ALOHA, Stanford, Zipeng Fu, Tony Zhao, Chelsea Finn https://mobile-aloha.github.io/

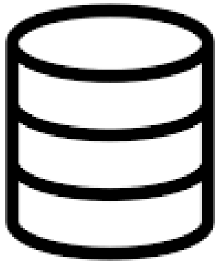
Key Ingredient: Imitation Learning

Kinesthetic Teaching

Teleoperation

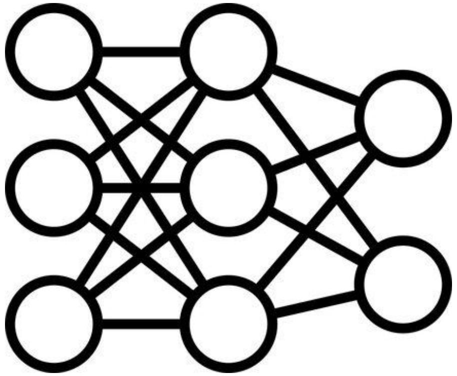


Collect Demonstrations



(state, action)

A Dataset of State-Action Pairs



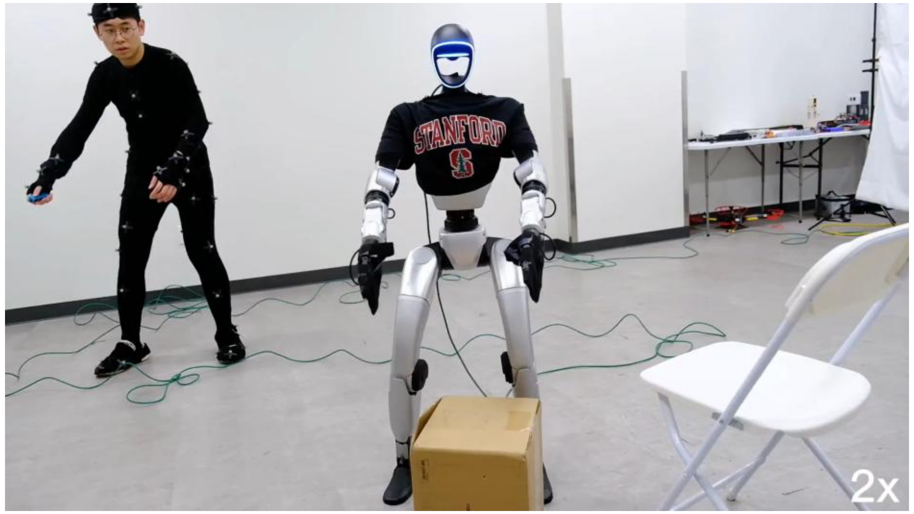
Train a Policy Network

Deploy the Policy Network

Key Ingredient: Teleoperation for Data Collection

https://mobile-aloha.github.io/

https://mobile-tv.github.io/



https://yanjieze.com/TWIST/

Tesla

Key Ingredient: Teleoperation for Data Collection

- Requires specific hardware

- Requires human expertise

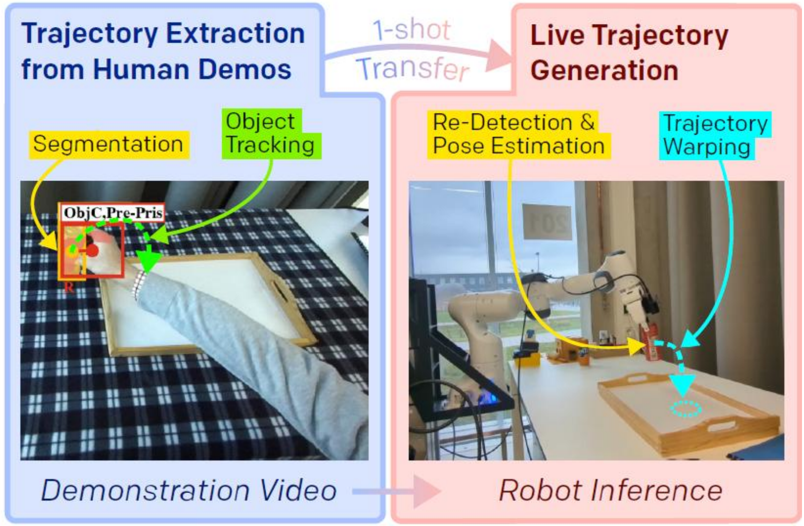
- Difficult to scale up

Learning Manipulation from Human Videos

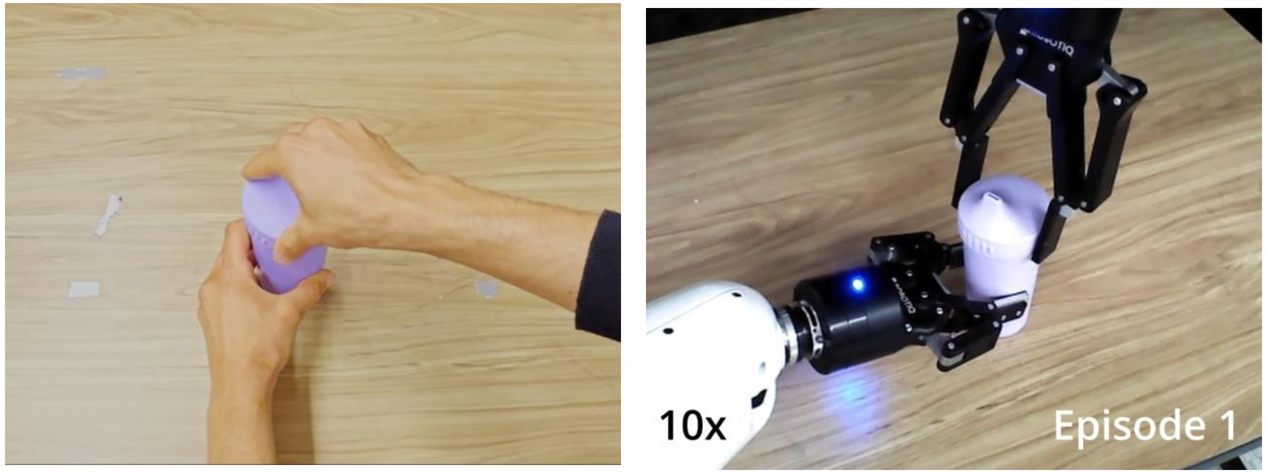
Image generated by ChatGPT

Learning Manipulation from Human Videos

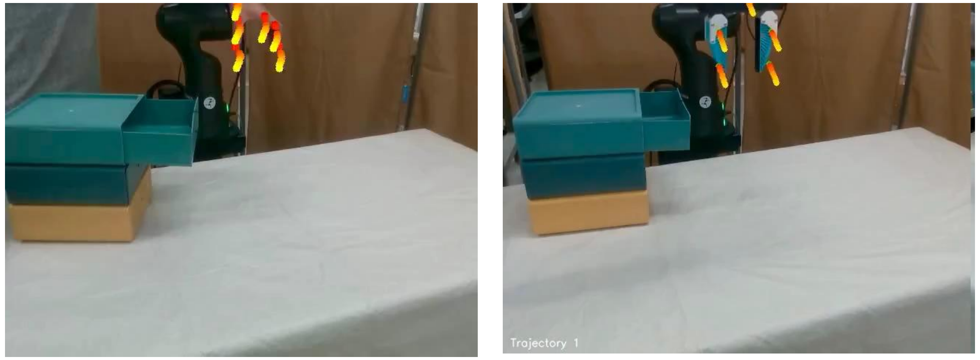
DexMV, Qin et al. UCSD, ECCV 2022



Trajectory Transfer, Heppert et al. University of Freiburg, IROS 2024

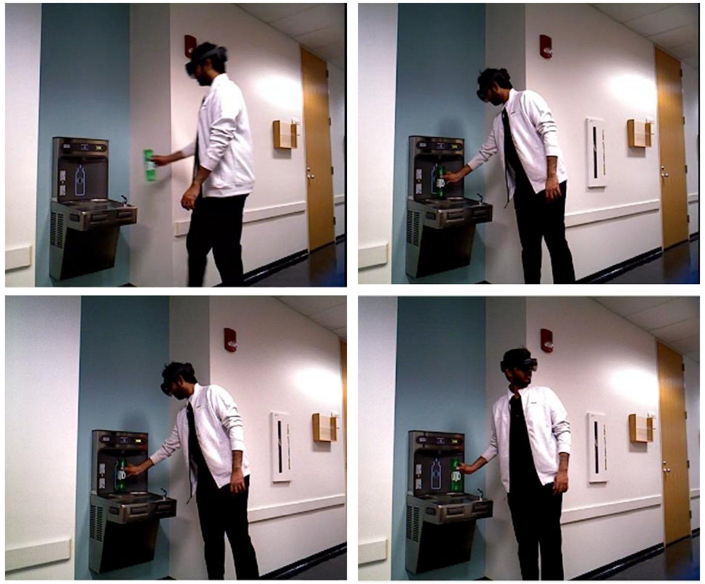


ScrewMimic, Bahety et al. UT Austin, RSS 2024



Motion Tracks, Ren et al. Cornell & Stanford, 2025

Learning Manipulation from Human Videos



Human demonstration for task
"getting water from a drinking fountain"

Perception

- Object segmentation and tracking
- Hand pose estimation and tracking

Understand human demonstration videos

Object and hand Trajectory

Control

- Trajectory Optimization
- Policy Learning

Skill learning

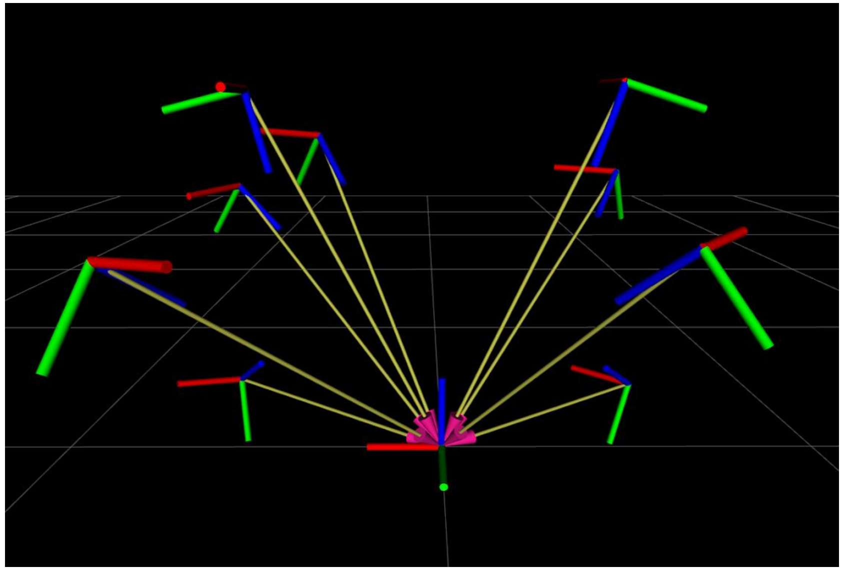
Goal: A robot learns to do the task from the demonstration video

Outline

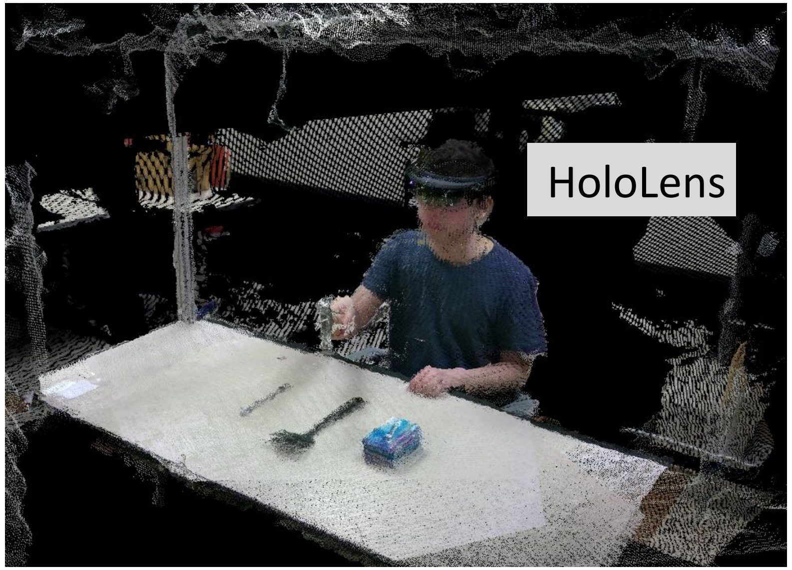
- HO-Cap: A low-cost capture system for hand-object interaction
- RobotFingerPrint: A unified gripper coordinate space for cross-embodiment grasp transfer
- An optimization framework for human-to-robot trajectory transfer

HO-Cap: Hardware Setup

(a) Our hardware setup and objects



(b) Visualization of the camera poses



(c) Point clouds from the cameras

8x 

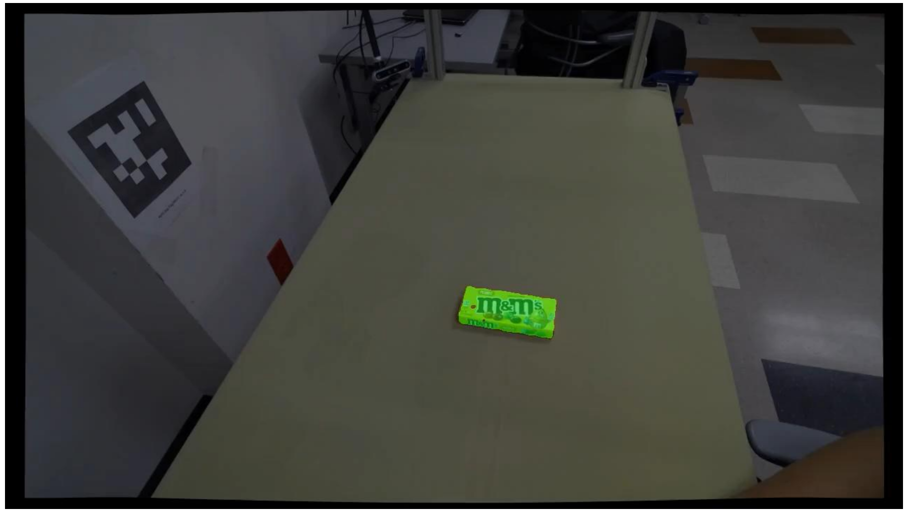
1x 

1x 

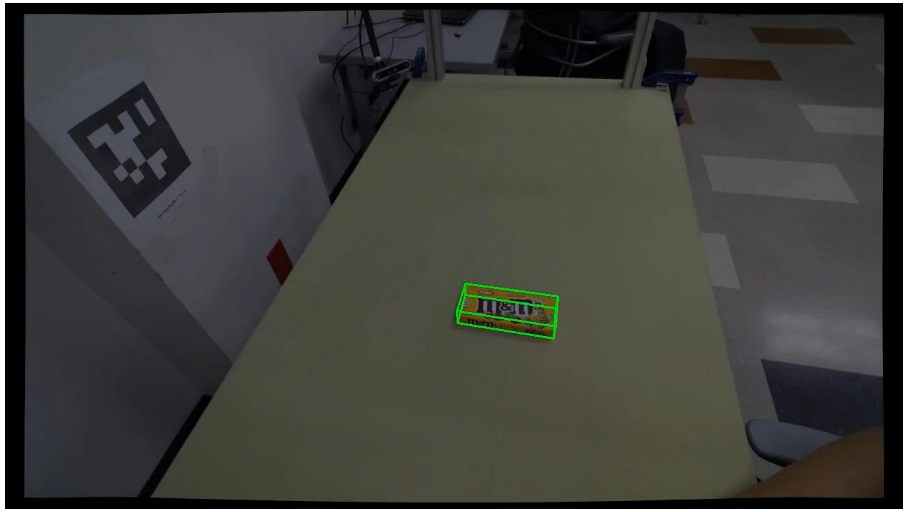
HO-Cap: Object Shape Reconstruction

RGB

Mask



6D Pose

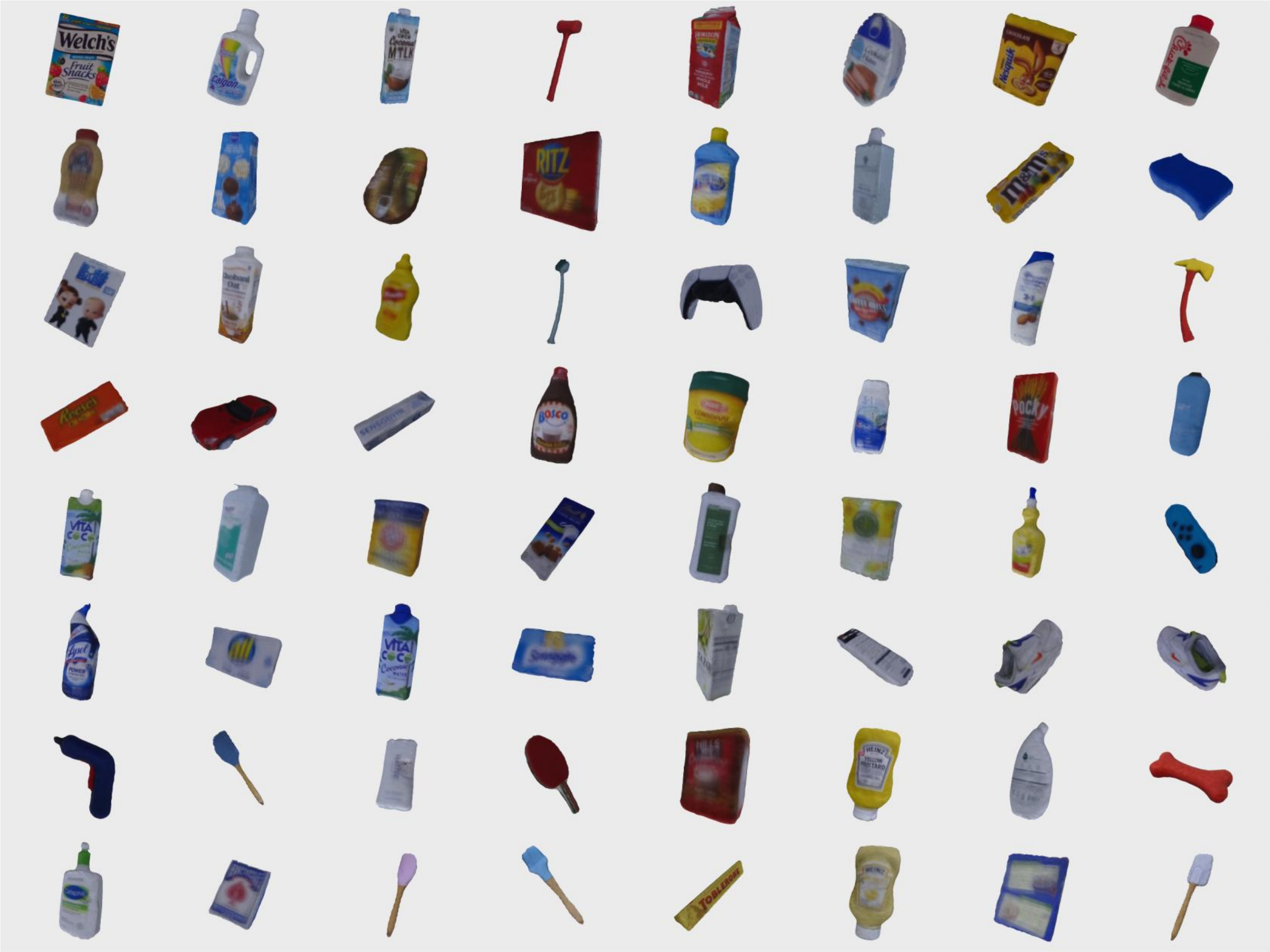


3D textured mesh

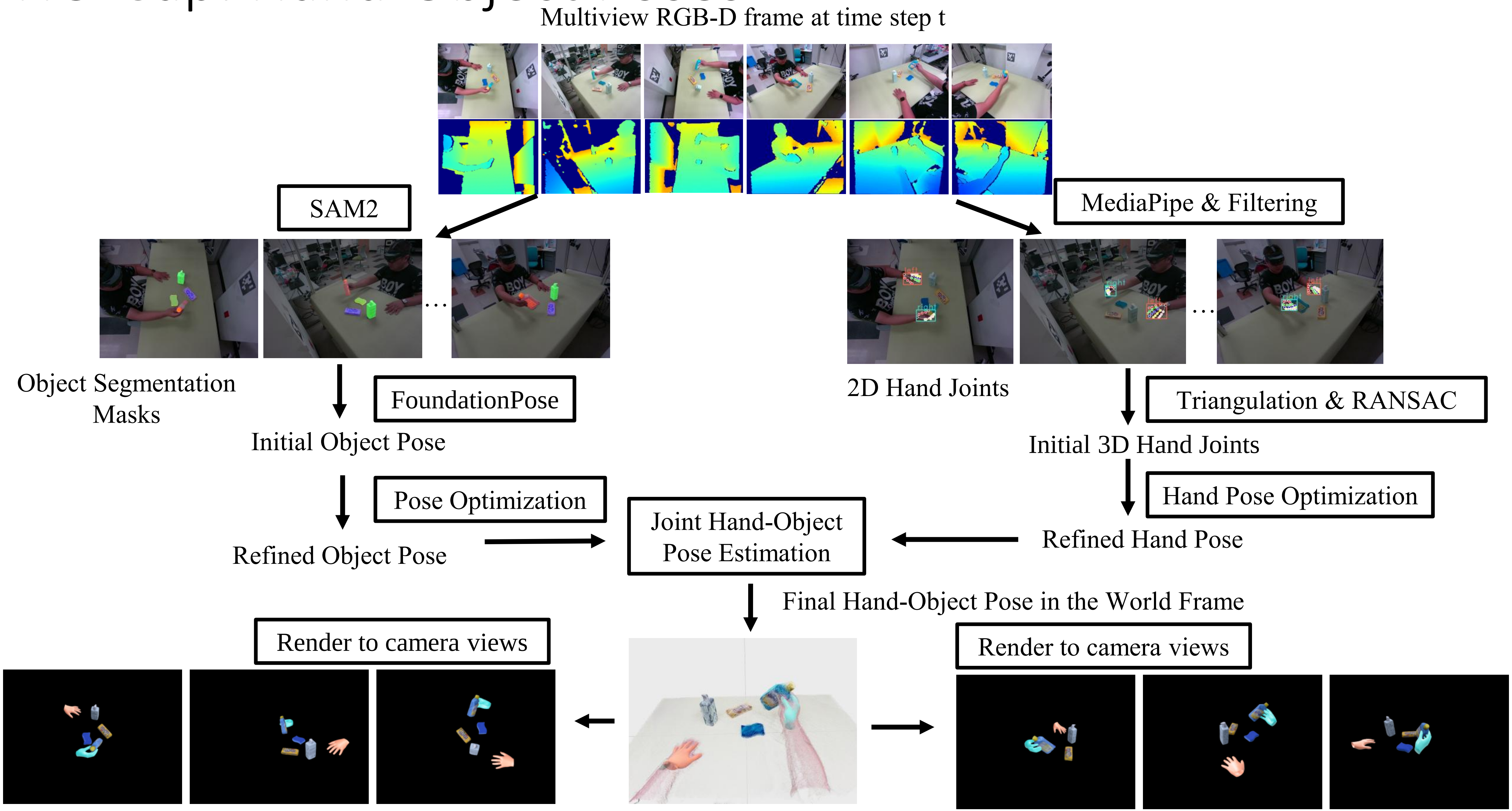
BundleSDF: Neural 6-DoF Tracking and 3D Reconstruction of Unknown Objects. Bowen Wen, Jonathan Tremblay, Valts Blukis, Stephen Tyree, Thomas Müller, Alex Evans, Dieter Fox, Jan Kautz, Stan Birchfield. In CVPR, 2023.

HO-Cap: Object Shape Reconstruction

64 Objects



HO-Cap: Hand-Object Poses

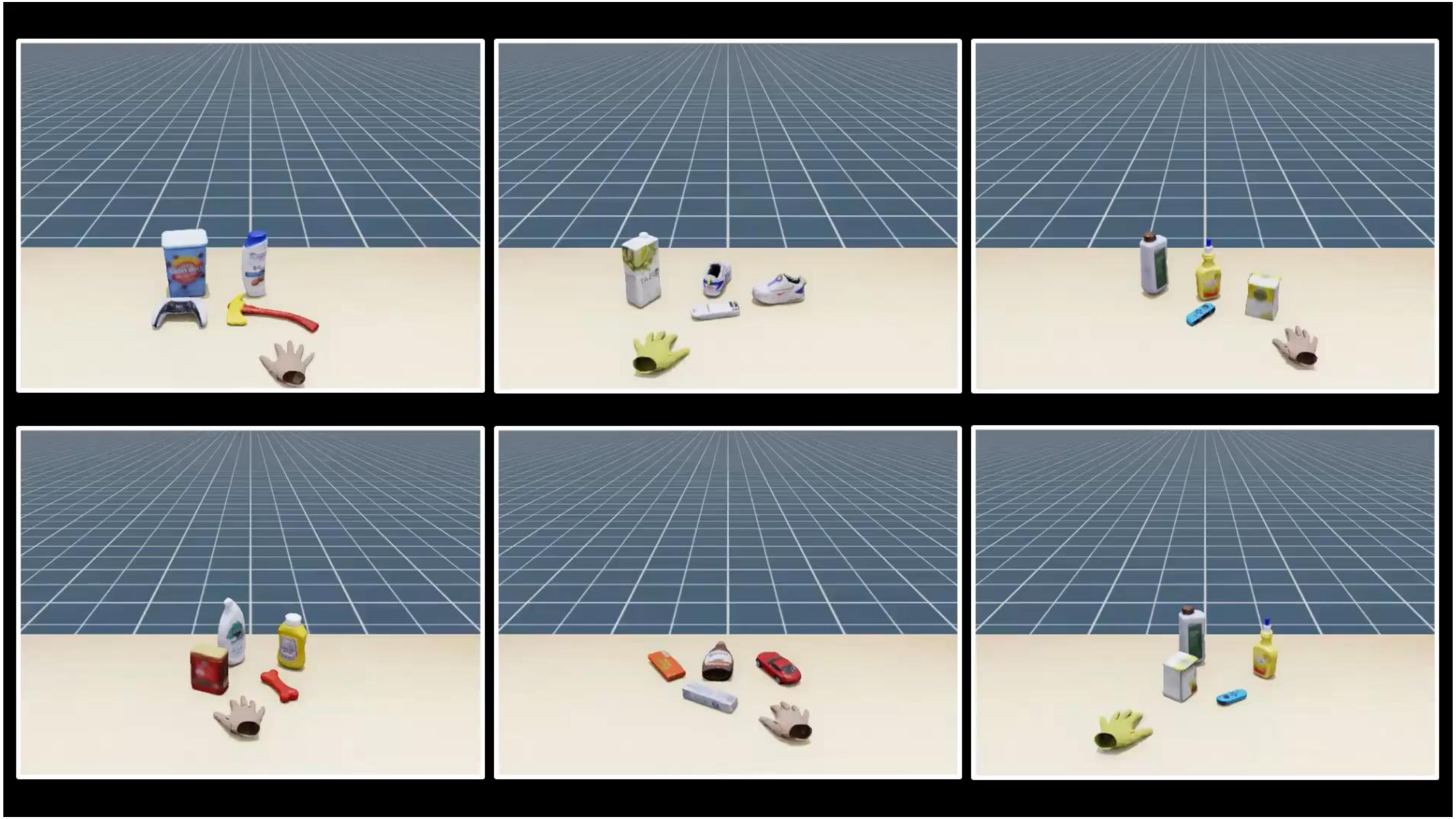


HO-Cap: Pick-and-Place

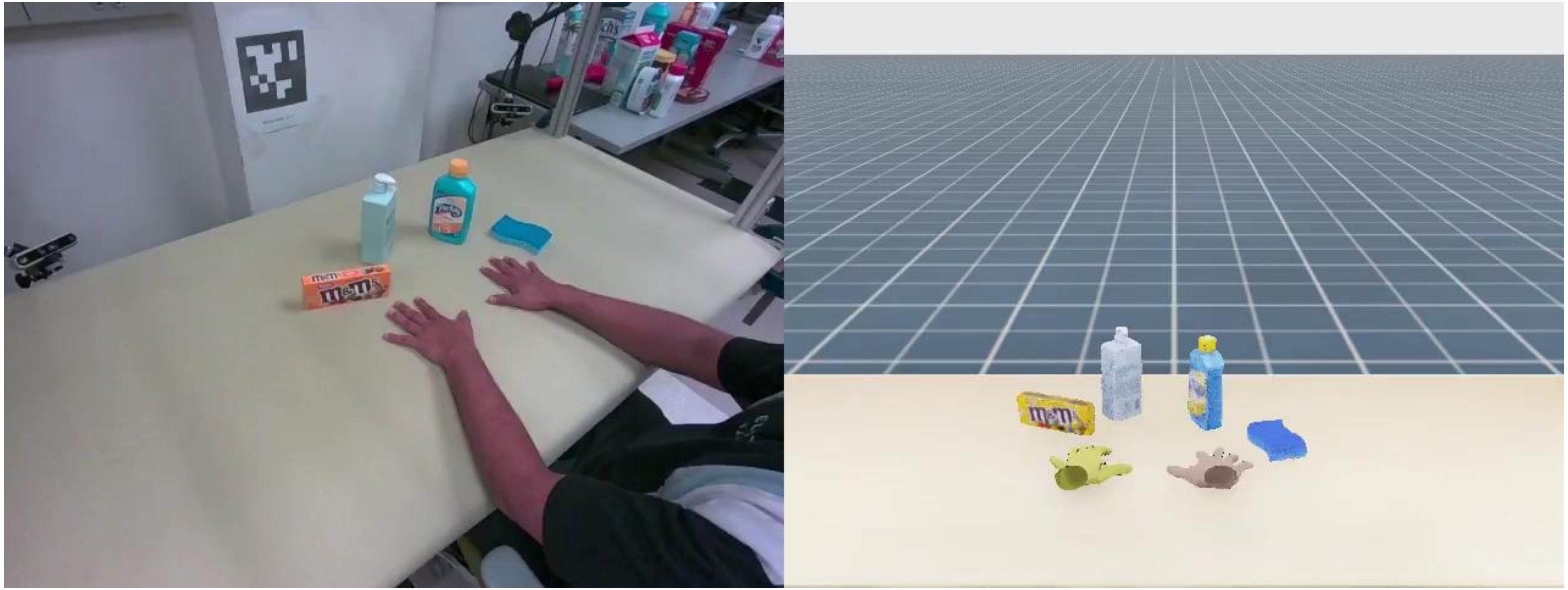
HO-Cap: Handover

HO-Cap: Affordance Usage

HO-Cap: Isaac Sim Replay



HO-Cap



We can use the HO-Cap data as human demonstrations for robots.

HO-Cap: A Capture System and Dataset for 3D Reconstruction and Pose Tracking of Hand-Object Interaction.
Jikai Wang, Qifan Zhang, Yu-Wei Chao, Bowen Wen, Xiaohu Guo, Yu Xiang. In arXiv, 2025 (under submission).

Human-to-Robot Grasp Transfer

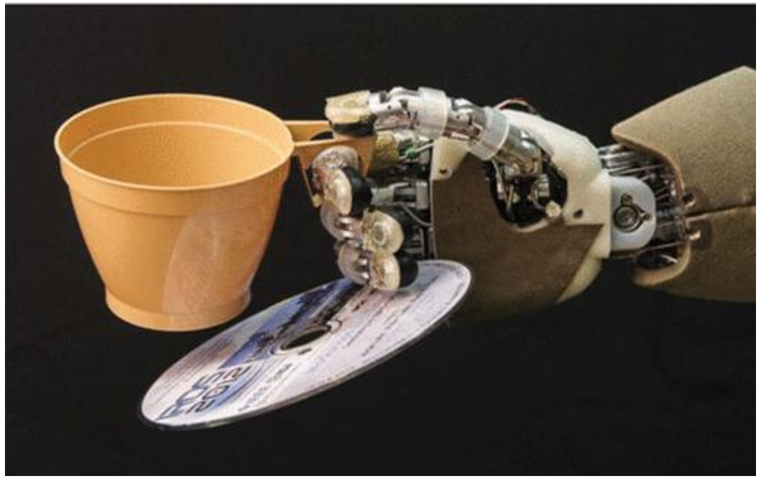
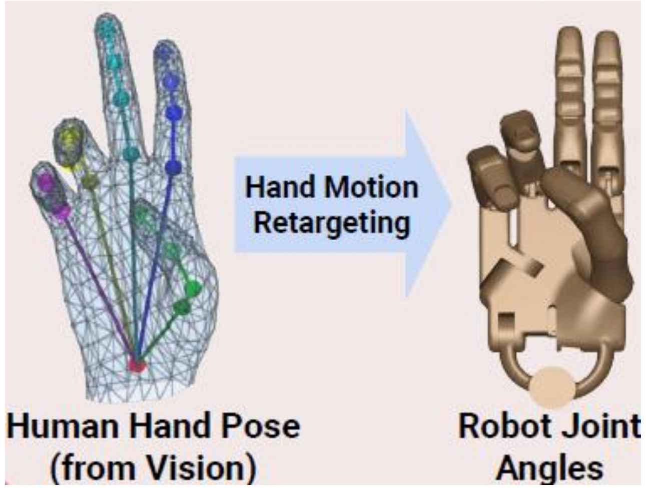


Image generated by ChatGPT

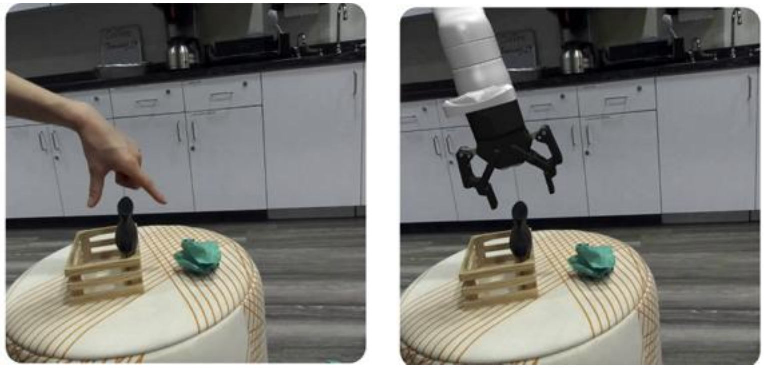
Human-to-Robot Grasp Transfer

- Retargeting



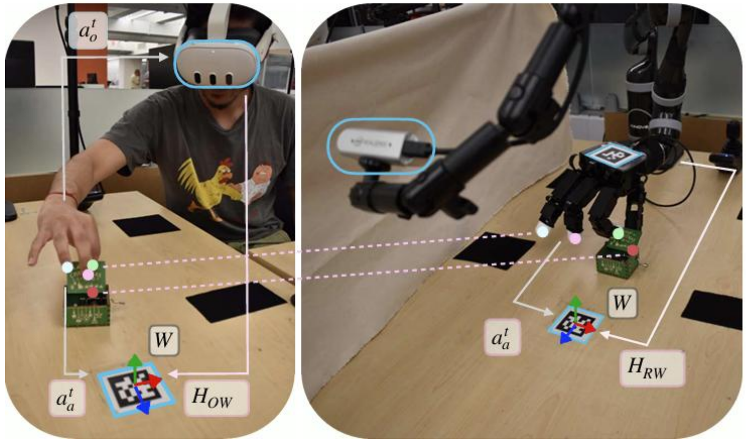
DexMV, Qin et al. UCSD, ECCV 2022

https://yzqin.github.io/dexmv/



Phantom, Lepert et al. Stanford 2025

https://phantom-human-videos.github.io/

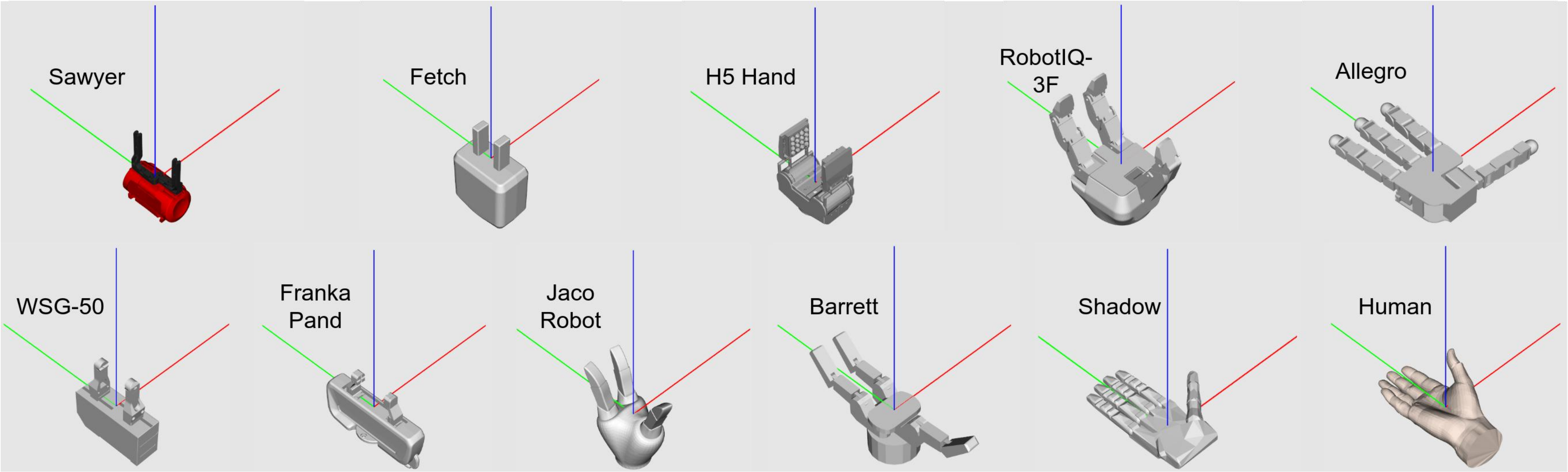


HuDOR, Guzey et al. NYU 2025

https://object-rewards.github.io/

A Common Grasping Space

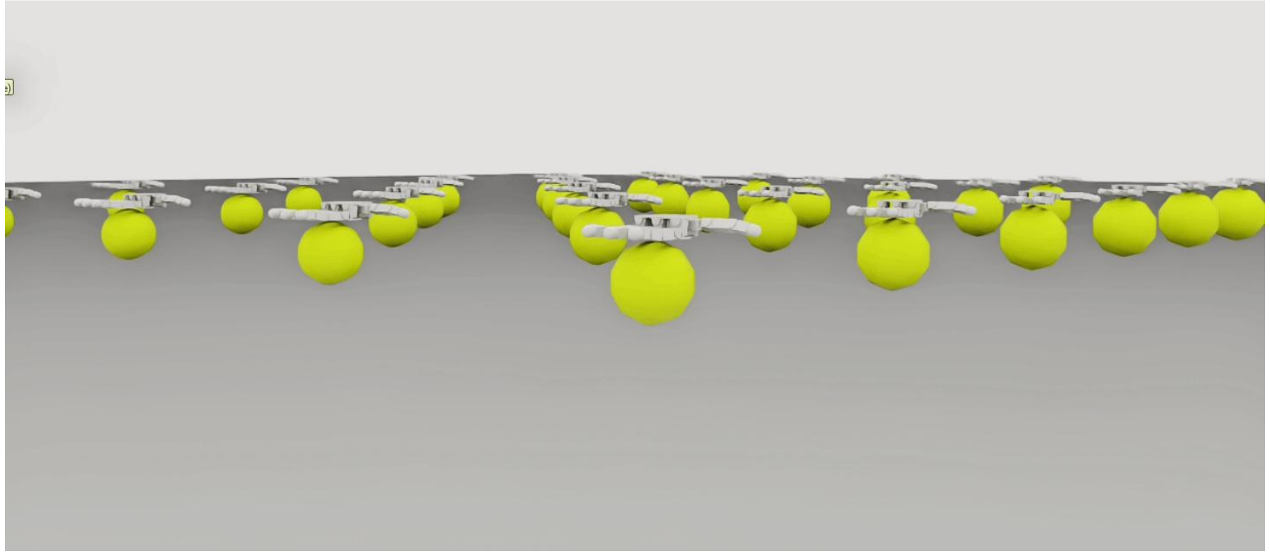
- Can we find a common grasping space for all the grippers?

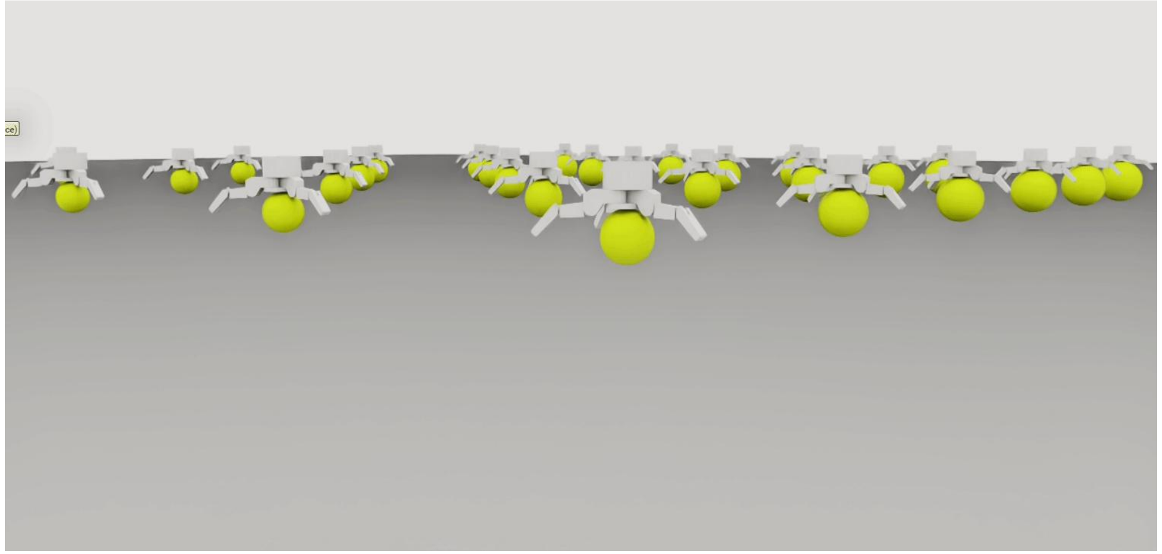


- We can align the palm orientations
- How to map fingers?

A Common Grasping Space

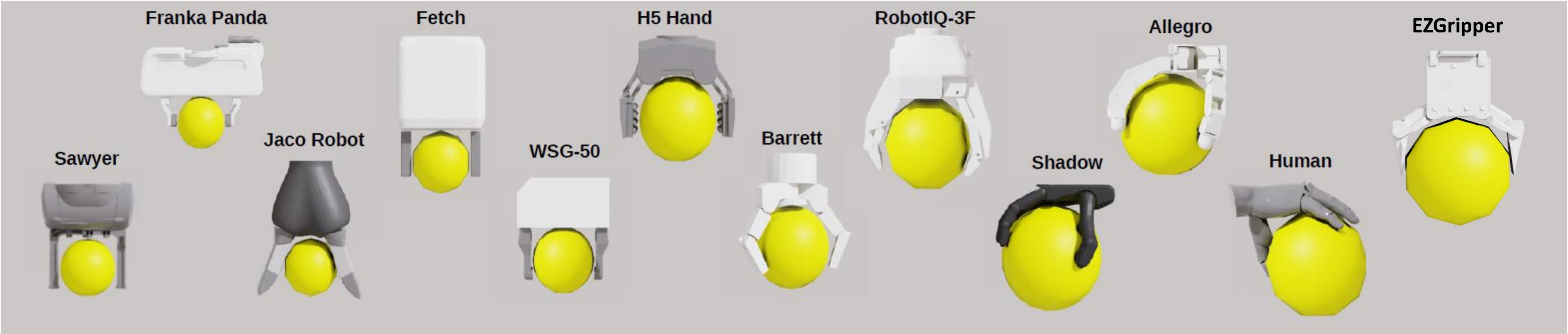
- Having the hands to grasp a common sphere
- Using contact maps on the sphere for retargeting
- Maximal sphere test in simulation





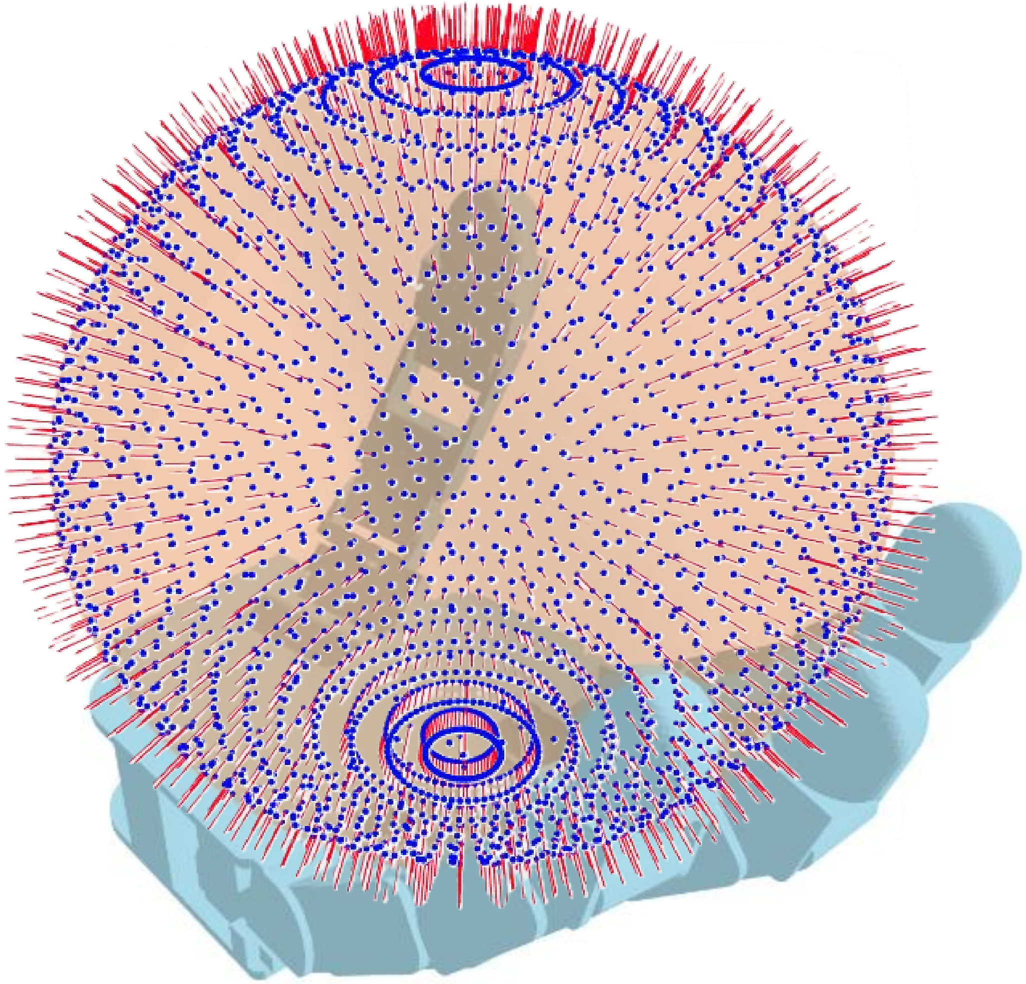
A Common Grasping Space

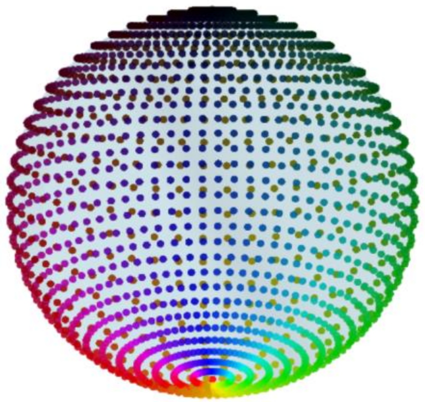
- Maximal spheres for each gripper



A Unified Gripper Coordinate Space

- Map spherical coordinates to the gripper

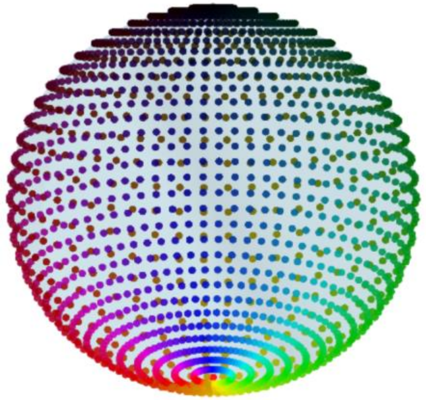




(λ, ϕ)

A Unified Gripper Coordinate Space

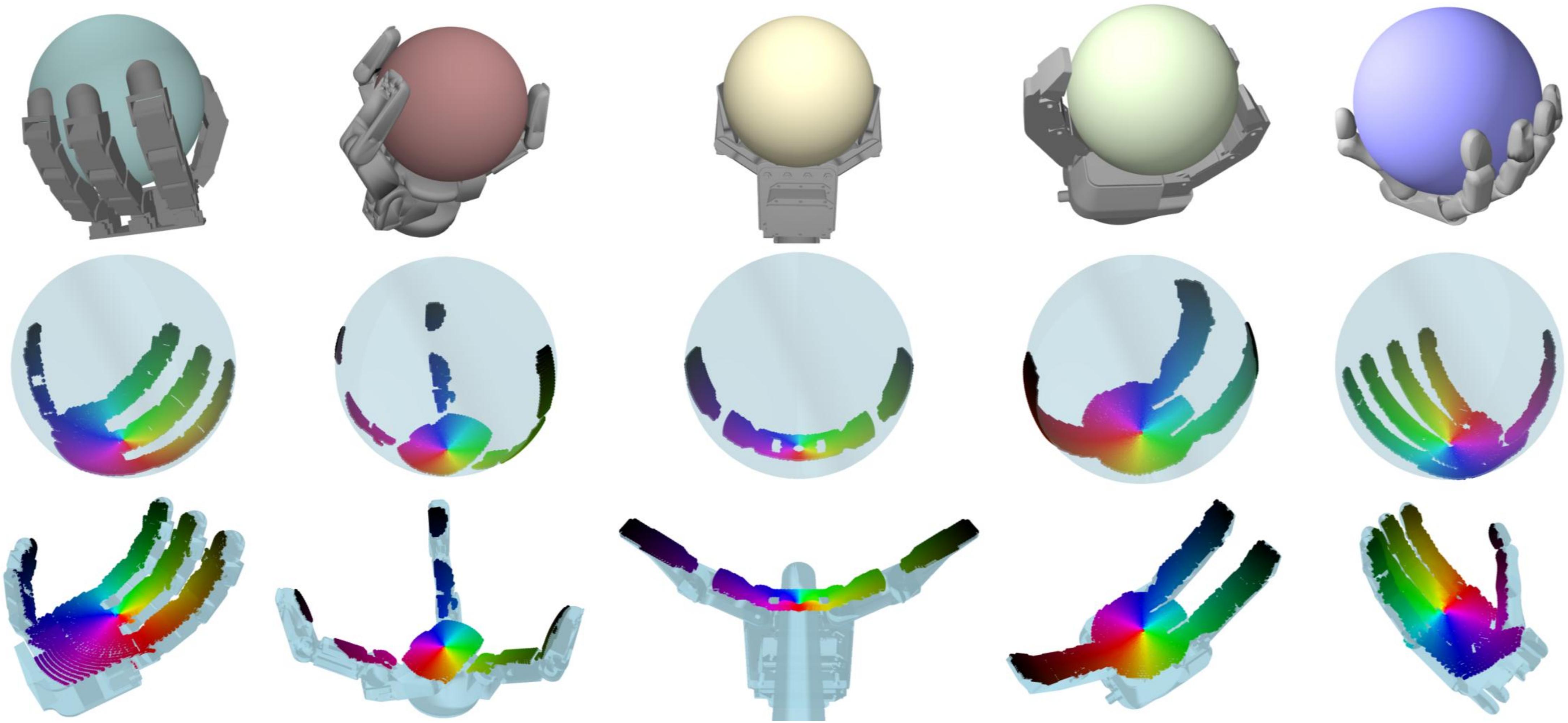
- Map spherical coordinates to the gripper



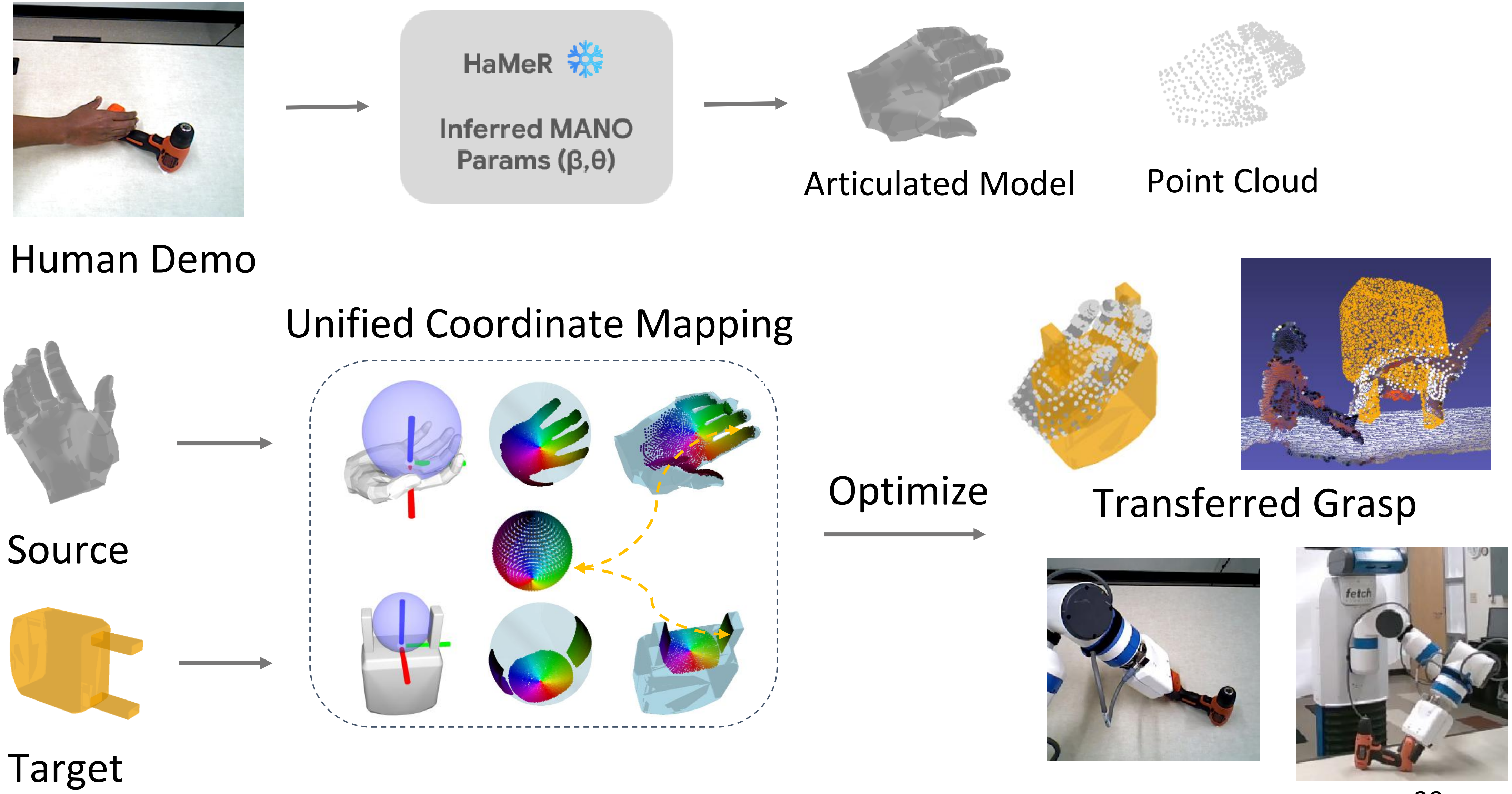
(λ, ϕ)

A Unified Gripper Coordinate Space

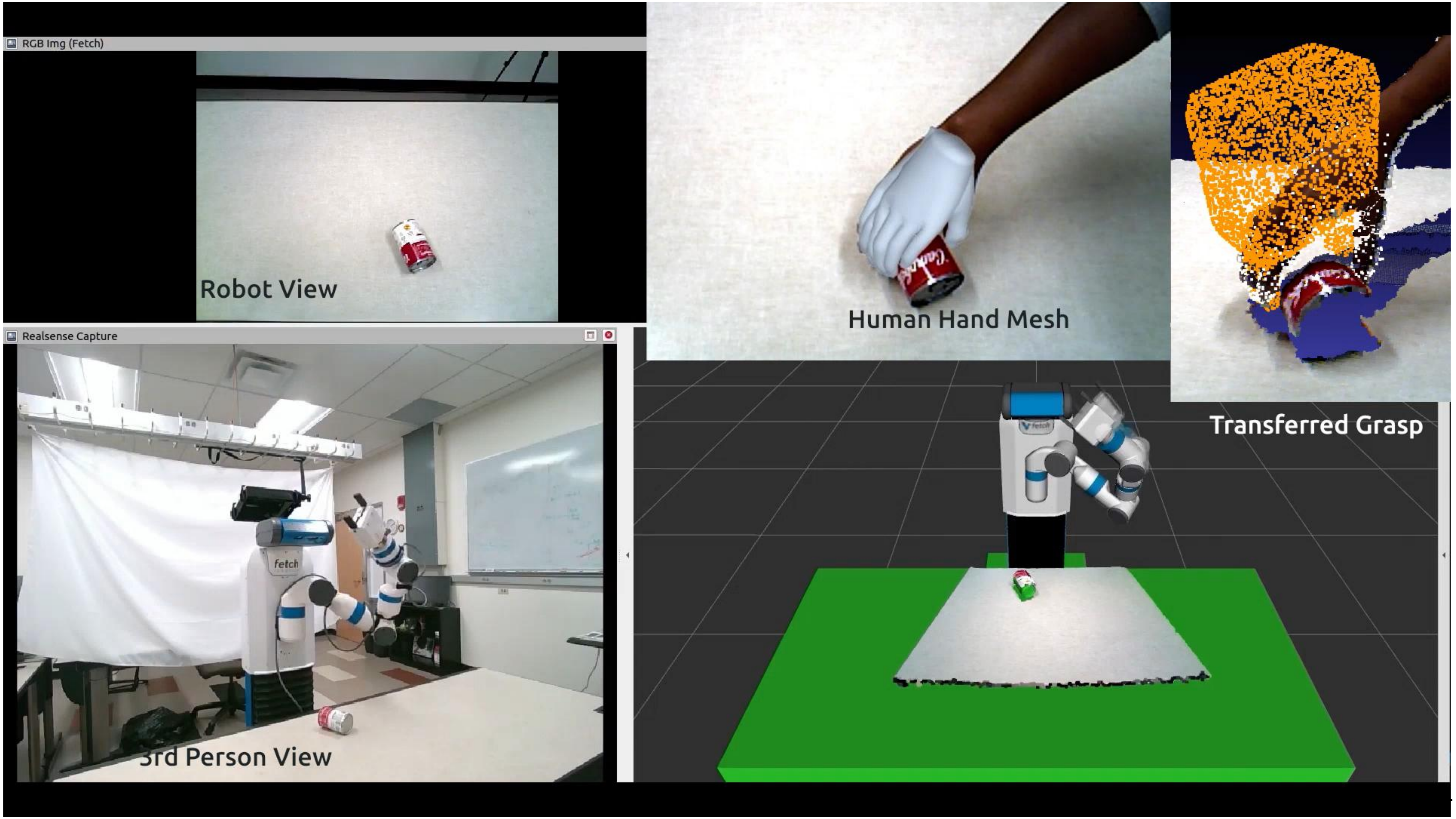
- Finger print: map spherical coordinates to the gripper



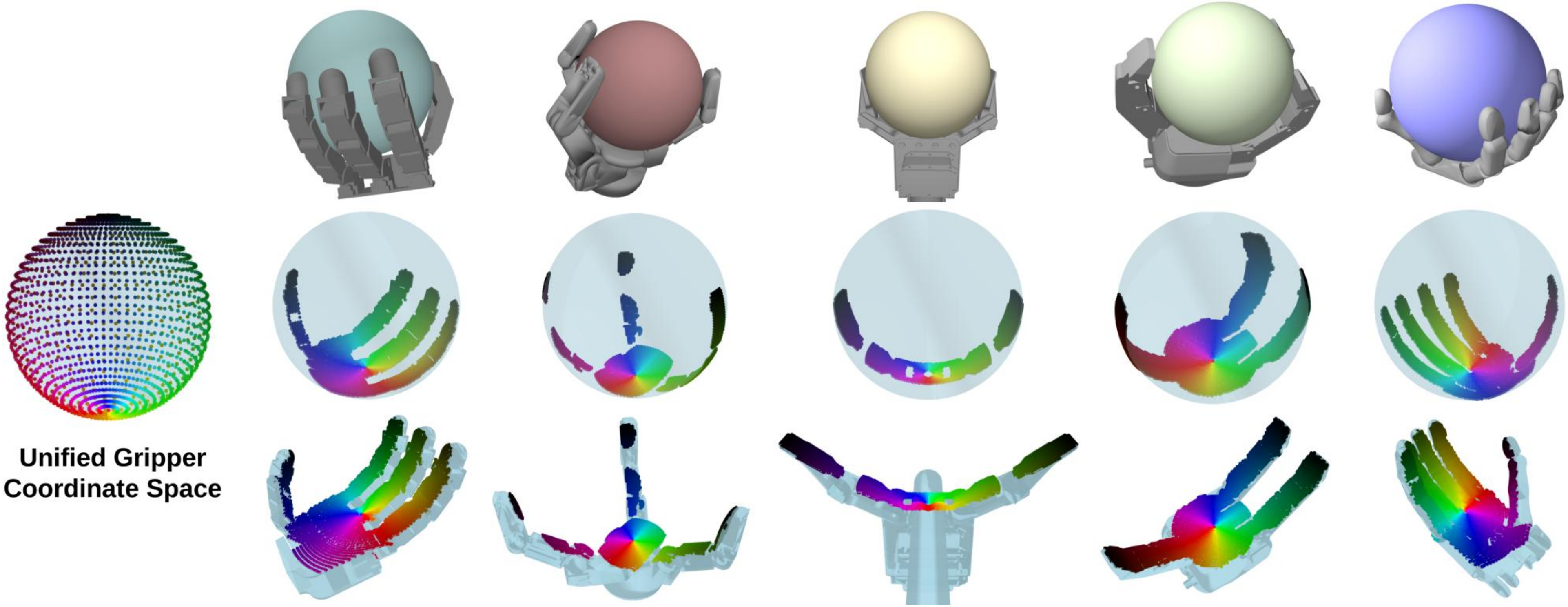
Grasp Transfer



Grasp Transfer



RobotFingerPrint



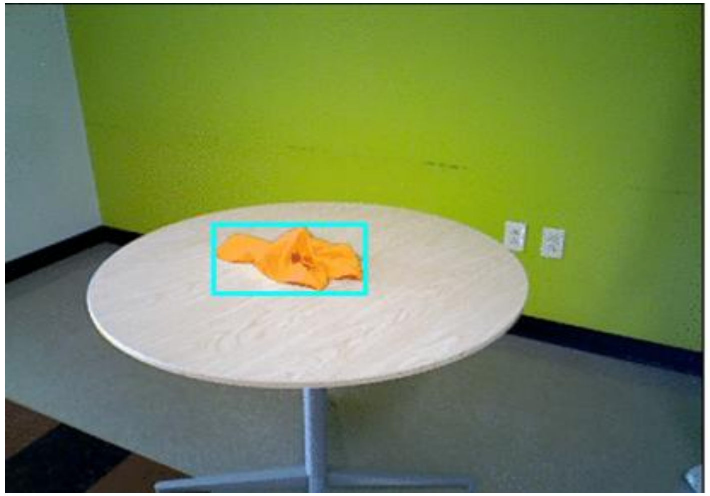
RobotFingerPrint: Unified Gripper Coordinate Space for Multi-Gripper Grasp Synthesis and Transfer.
Ninad Khargonkar, Luis Felipe Casas, Balakrishnan Prabhakaran, Yu Xiang. In arXiv, 2025 (under submission).

Human-to-Robot Trajectory Transfer

Sai Haneesh Allu

Jishnu Jaykumar P

One-shot imitation learning



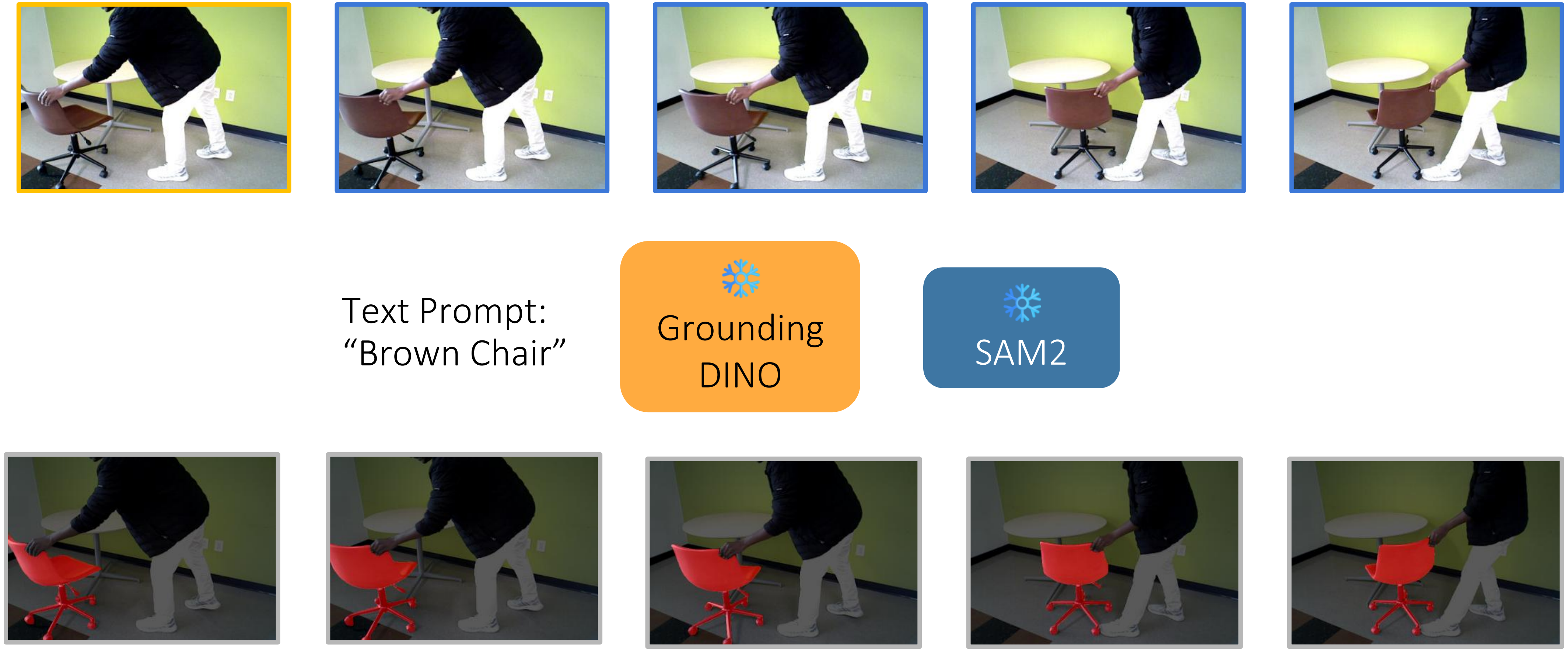
Clean table using Towel

Close jar with Red Lid

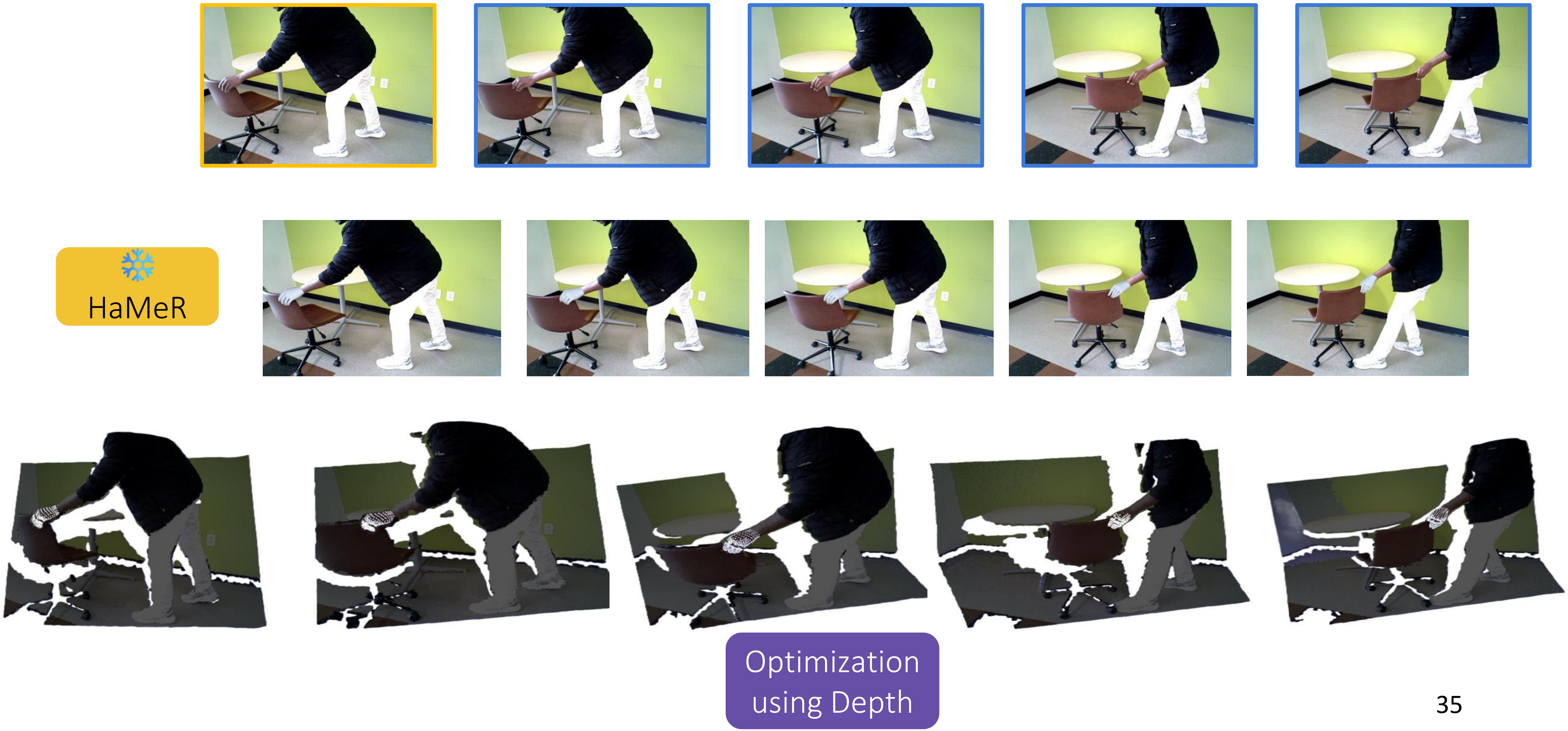
Pour Tumbler

On-going work

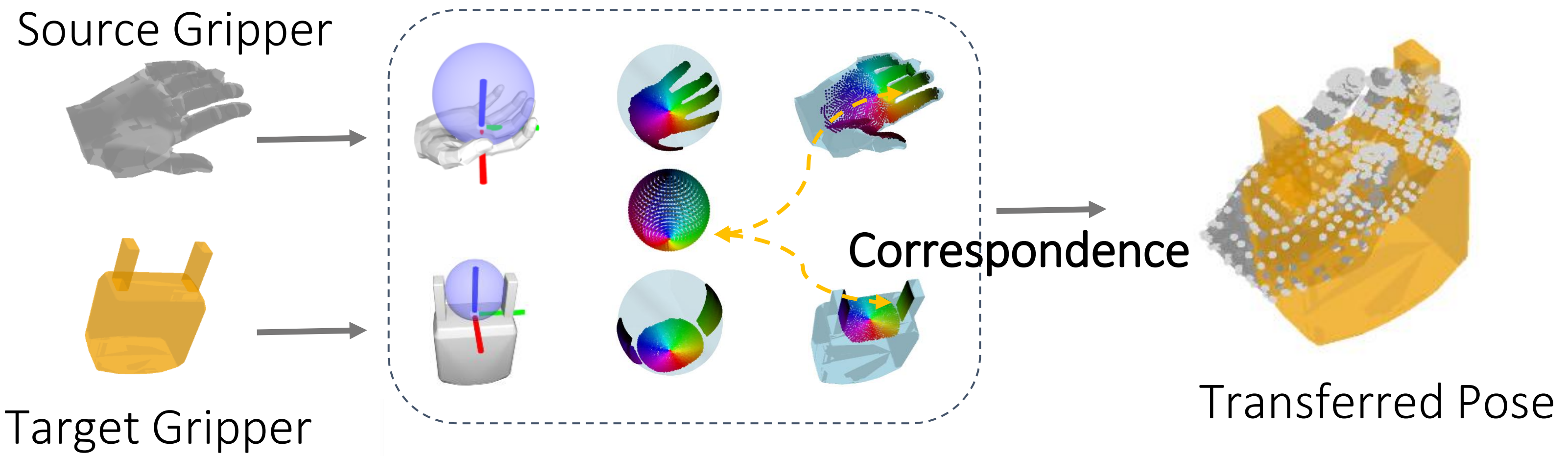
Understanding of the Human Demonstrations

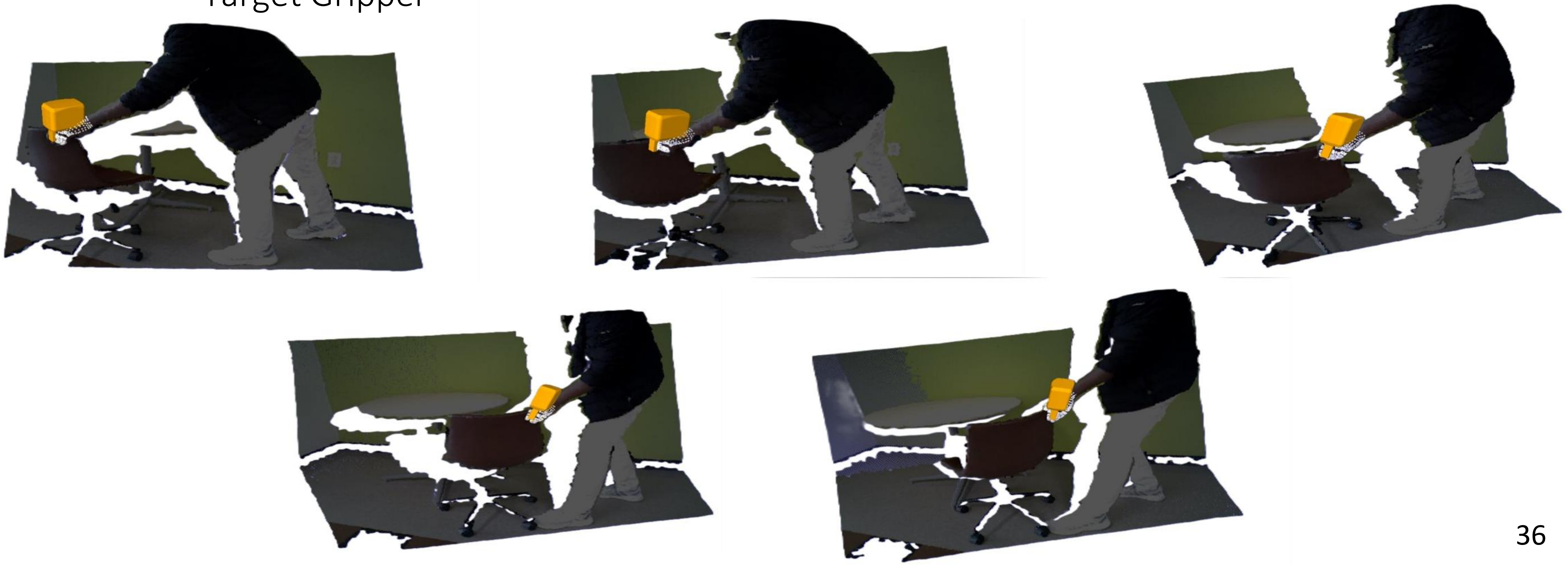


Understanding of the Human Demonstrations

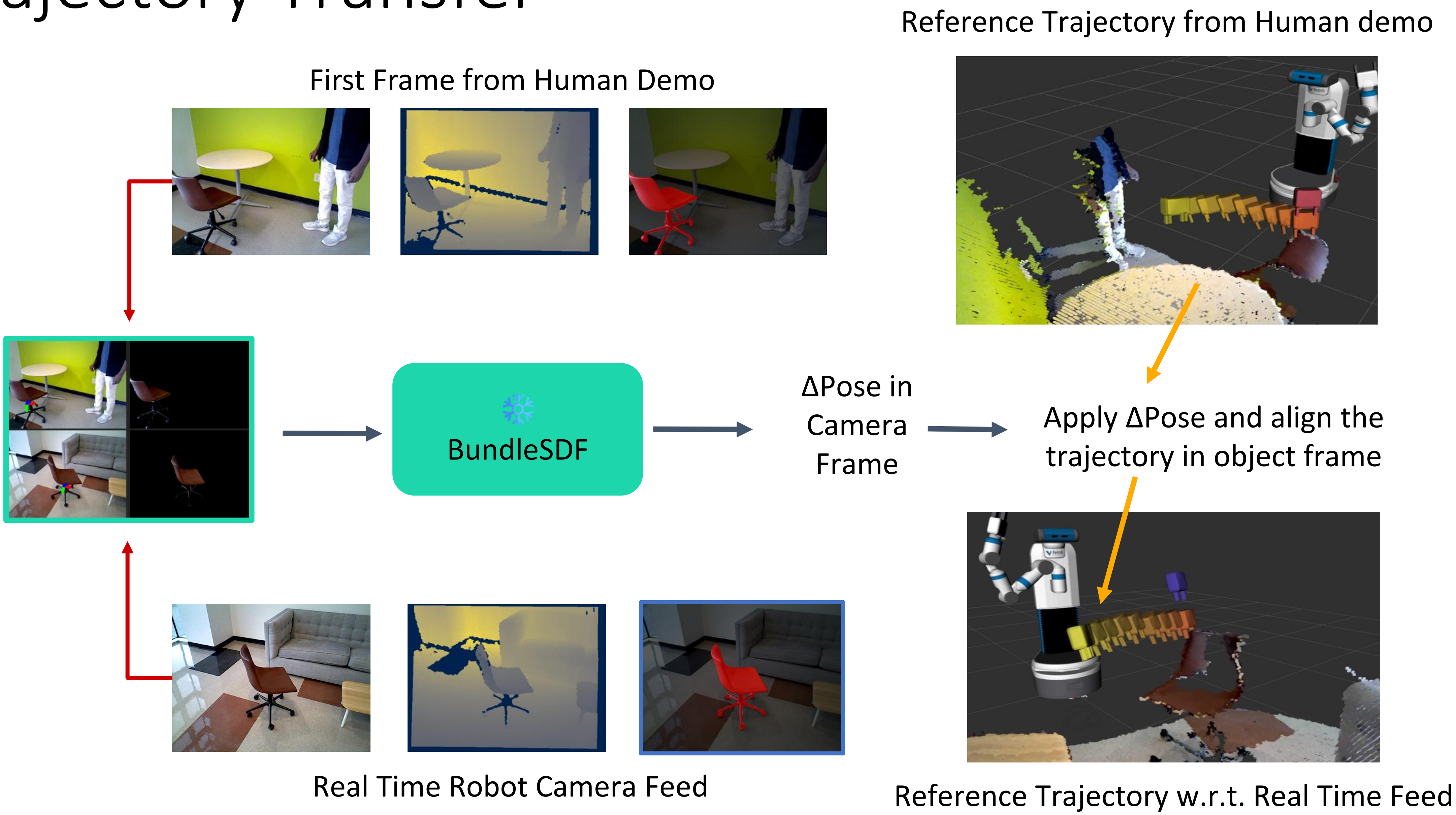


Understanding of the Human Demonstrations





Trajectory Transfer

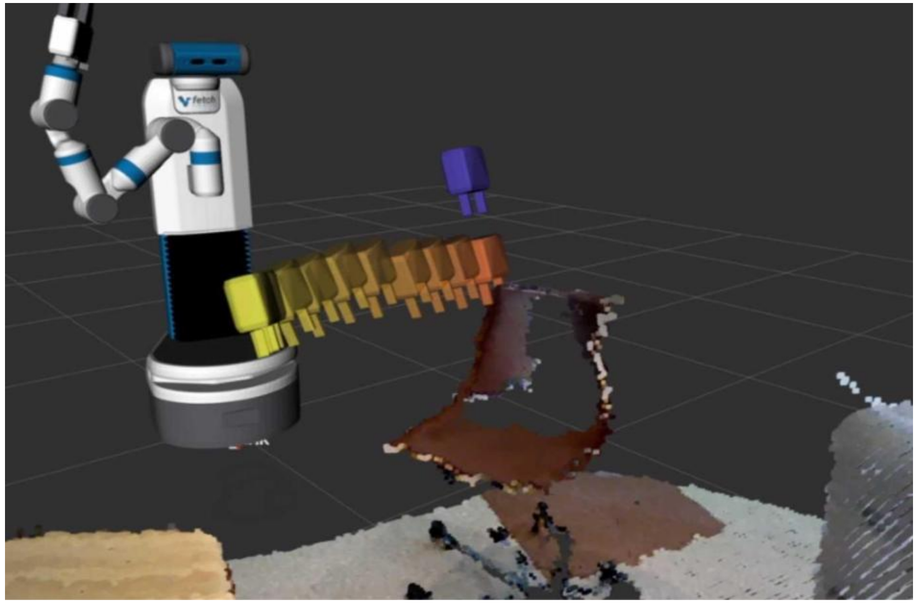


Trajectory Transfer

- How to follow the transferred gripper trajectory?

Task Space

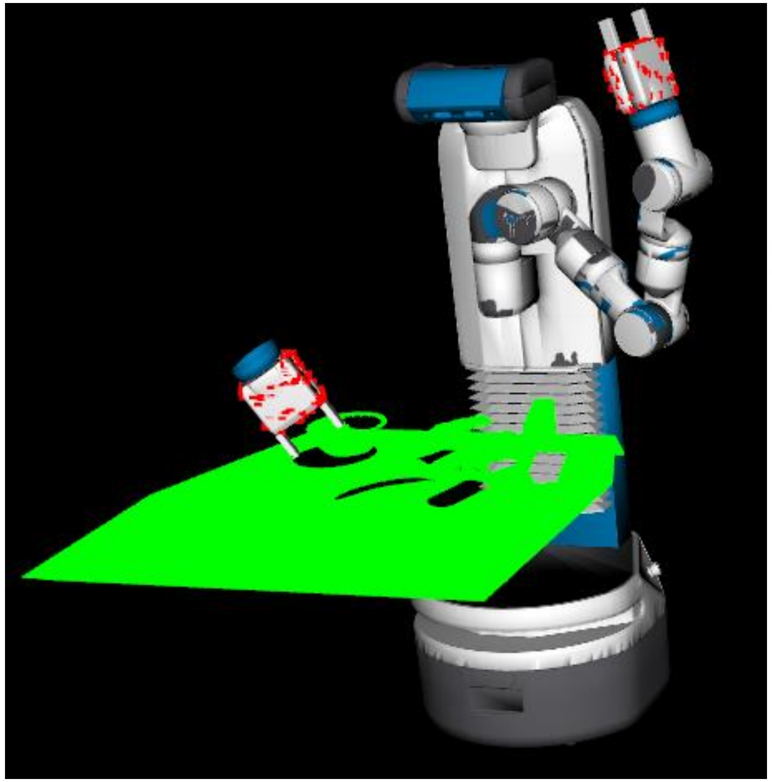
Robot View



Reference Trajectory w.r.t. Real Time Feed

Trajectory Optimization

- Point Cloud-based Cost Function for Goal Reaching



Gripper pose

Goal pose

$$c_{\text{goal}}(\mathbf{T}_T, \mathbf{T}_g)$$
$$= \sum_{i=1}^{m} \|(\mathbf{R}_T \mathbf{x}_i + \mathbf{t}_T) - (\mathbf{R}_g \mathbf{x}_i + \mathbf{t}_g)\|^2,$$

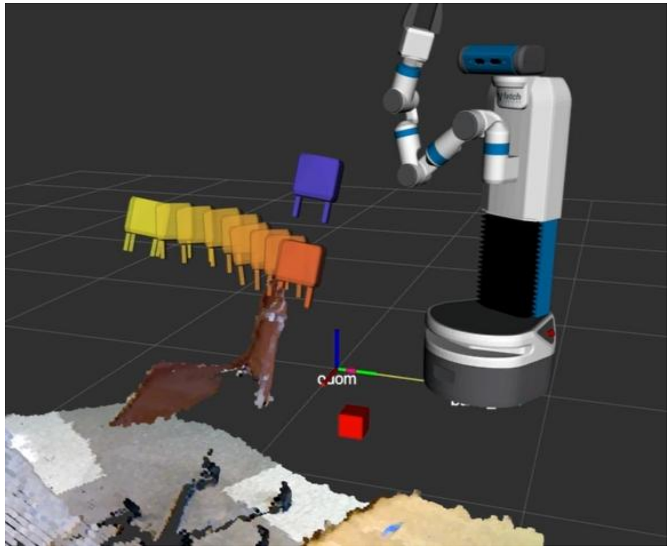
Points on the gripper

Grasping Trajectory Optimization with Point Clouds. Yu Xiang, Sai Haneesh Allu, Rohith Peddi, Tyler Summers, Vibhav Gogate. In IROS, 2024.

Optimizing the Robot Base Location

- Find the base position that can reach N gripper poses from the trajectory

Base $\mathbf{x} = \begin{bmatrix} x \\ y \\ \theta \end{bmatrix}$ $\mathbf{T}(\mathbf{x}) = \begin{bmatrix} \cos\theta & -\sin\theta & 0 & x \\ \sin\theta & \cos\theta & 0 & y \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$ Unknown



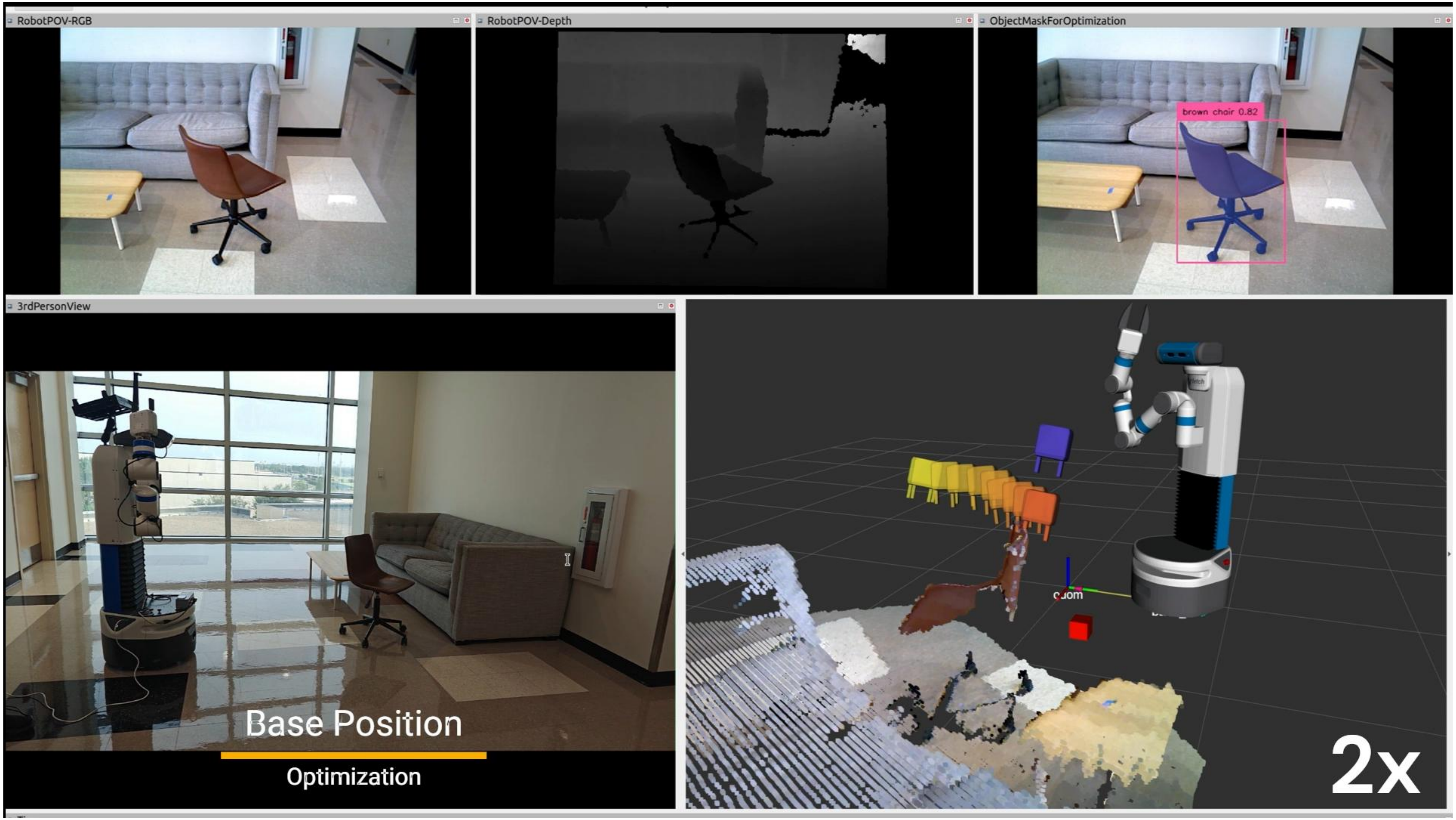
Gripper pose $\mathcal{T} = \{\mathbf{T}_1, \mathbf{T}_2 \ldots, \mathbf{T}_N\}$ Known

Arm configuration $\mathcal{Q} = \{\mathbf{q}_1, \mathbf{q}_2 \ldots, \mathbf{q}_N\}$ Unknown

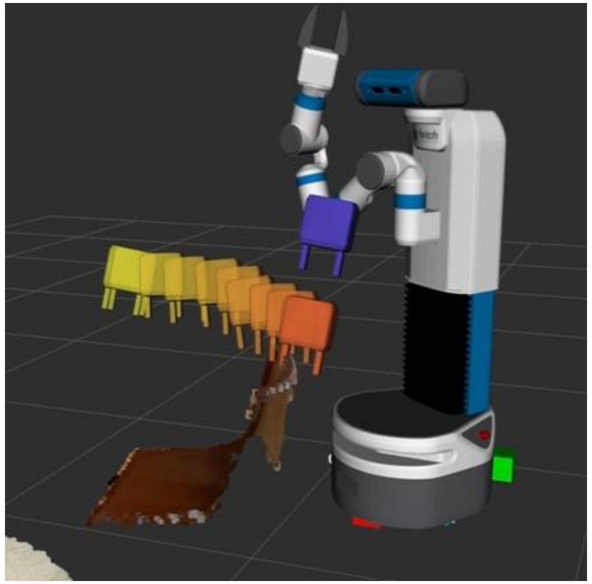
$$\begin{aligned} \arg\min_{\mathbf{x}, \mathcal{Q}} \quad & \lambda_{\text{effort}} \|\mathbf{x}\|^2 + \lambda_{\text{goal}} \textstyle\sum_{i=1}^{N} c_{\text{goal}}(\mathbf{T}(\mathbf{q}_i), \underline{\mathbf{T}(\mathbf{x}) \cdot \mathbf{T}_i}) \\ \text{s.t.,} \quad & \mathbf{x}_l \le \mathbf{x} \le \mathbf{x}_u \\ & \mathbf{q}_l \le \mathbf{q}_i \le \mathbf{q}_u, i = 1, \ldots, N \end{aligned}$$

Gripper goal in new base

Optimizing the Robot Base Location



Optimizing the Robot Trajectory



- Find the trajectory to follow the gripper poses well

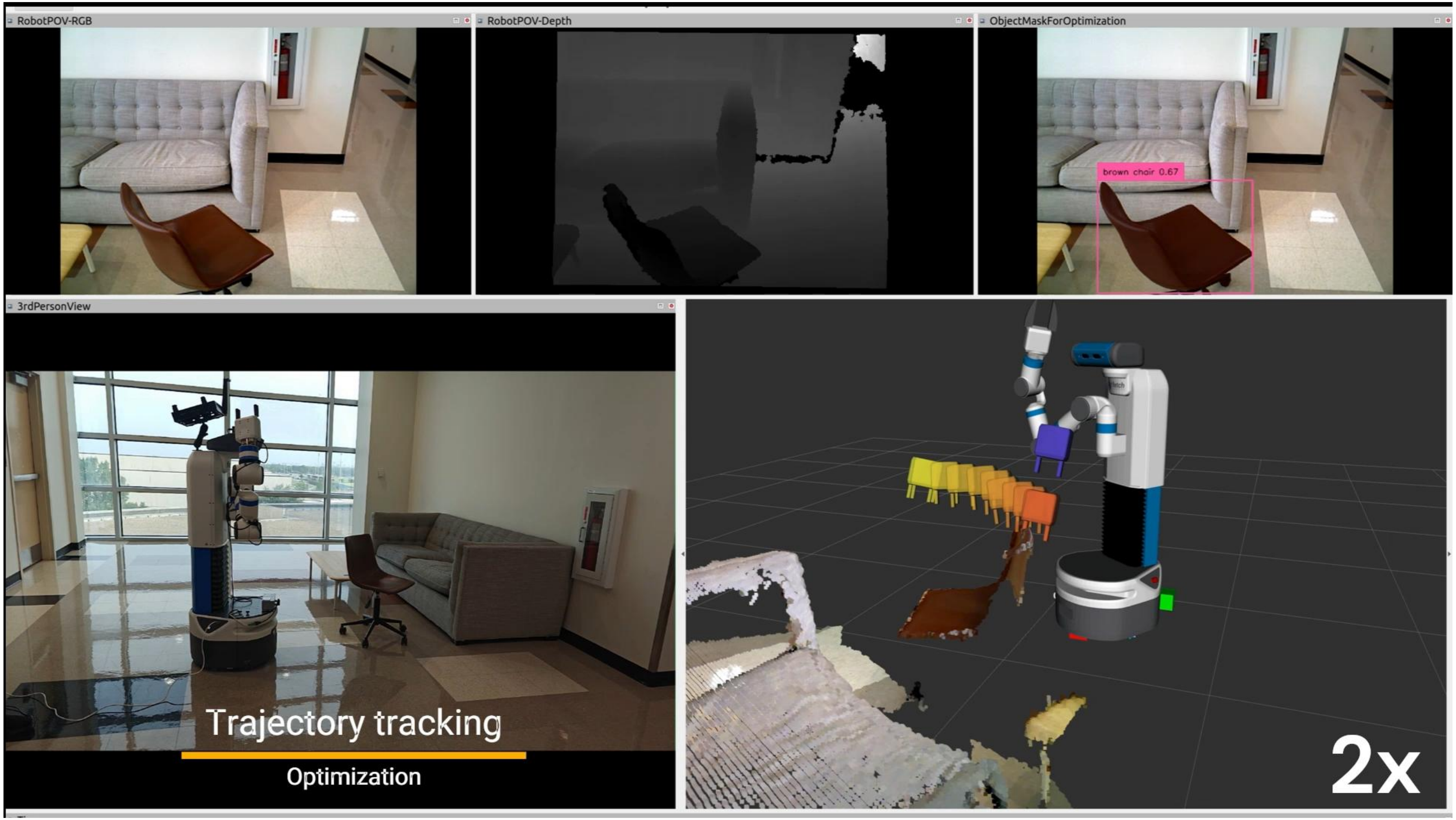
Unknown $\mathcal{Q} = (\mathbf{q}_1, \dots, \mathbf{q}_T) \quad \dot{\mathcal{Q}} = (\dot{\mathbf{q}}_1, \dots, \dot{\mathbf{q}}_T)$

Known $\mathcal{T} = \{\mathbf{T}_1, \mathbf{T}_2 \dots, \mathbf{T}_T\}$

Gripper trajectory in new robot base

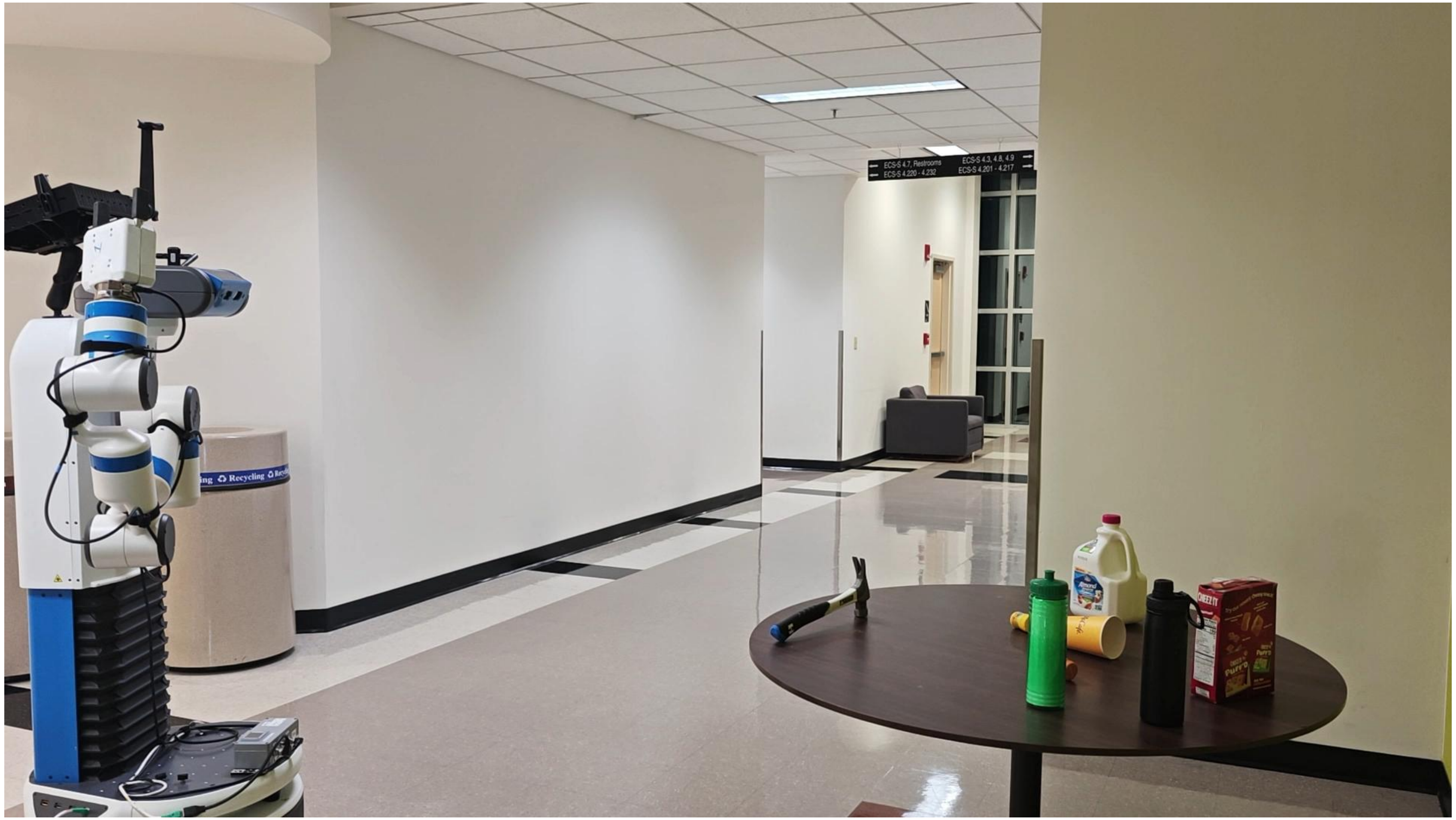
$$
\begin{aligned}
\arg\min_{\mathcal{Q}, \dot{\mathcal{Q}}} \quad & \textstyle\sum_{t=1}^{T} c_{\text{goal}}(\mathbf{T}(\mathbf{q}_t), \mathbf{T}_t) + \lambda_1 c_{\text{collision}}(\mathbf{q}_t) + \lambda_2 \sum_{t=1}^{T} \|\dot{\mathbf{q}}_t\|^2 \\
\text{s.t.,} \quad & \mathbf{q}_1 = \mathbf{q}_0 \\
& \dot{\mathbf{q}}_1 = \mathbf{0}, \dot{\mathbf{q}}_T = \mathbf{0} \\
& \mathbf{q}_{t+1} = \mathbf{q}_t + \dot{\mathbf{q}}_t dt, t = 1, \dots, T-1 \\
& \mathbf{q}_l \le \mathbf{q}_t \le \mathbf{q}_u, t = 1, \dots, T \\
& \dot{\mathbf{q}}_l \le \dot{\mathbf{q}}_t \le \dot{\mathbf{q}}_u, t = 1, \dots, T
\end{aligned}
$$

Optimizing the Robot Trajectory

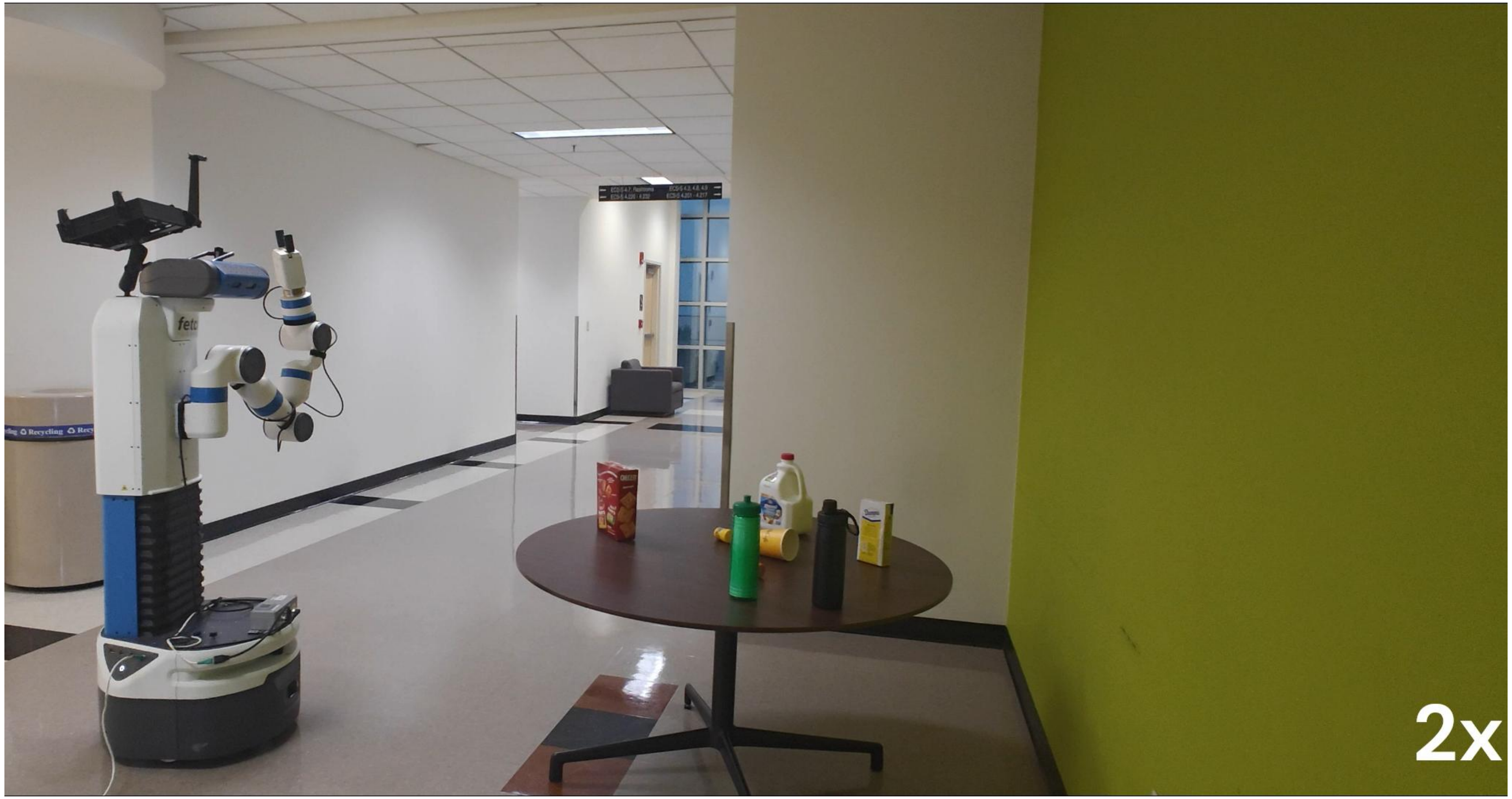


Trajectory Optimization to Follow the Reference

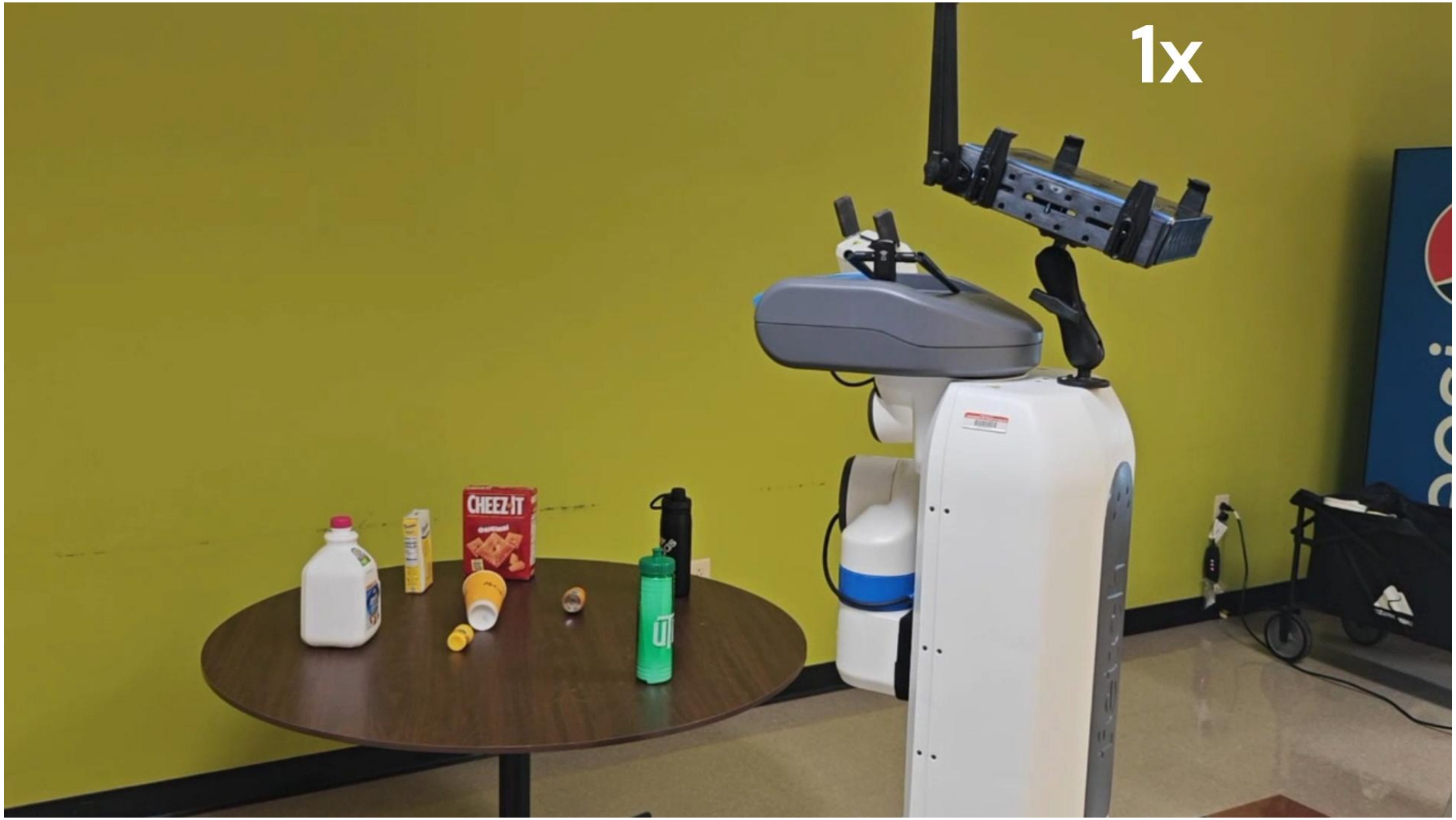
Trajectory Optimization to Follow the Reference



Trajectory Optimization to Follow the Reference



Failure Example

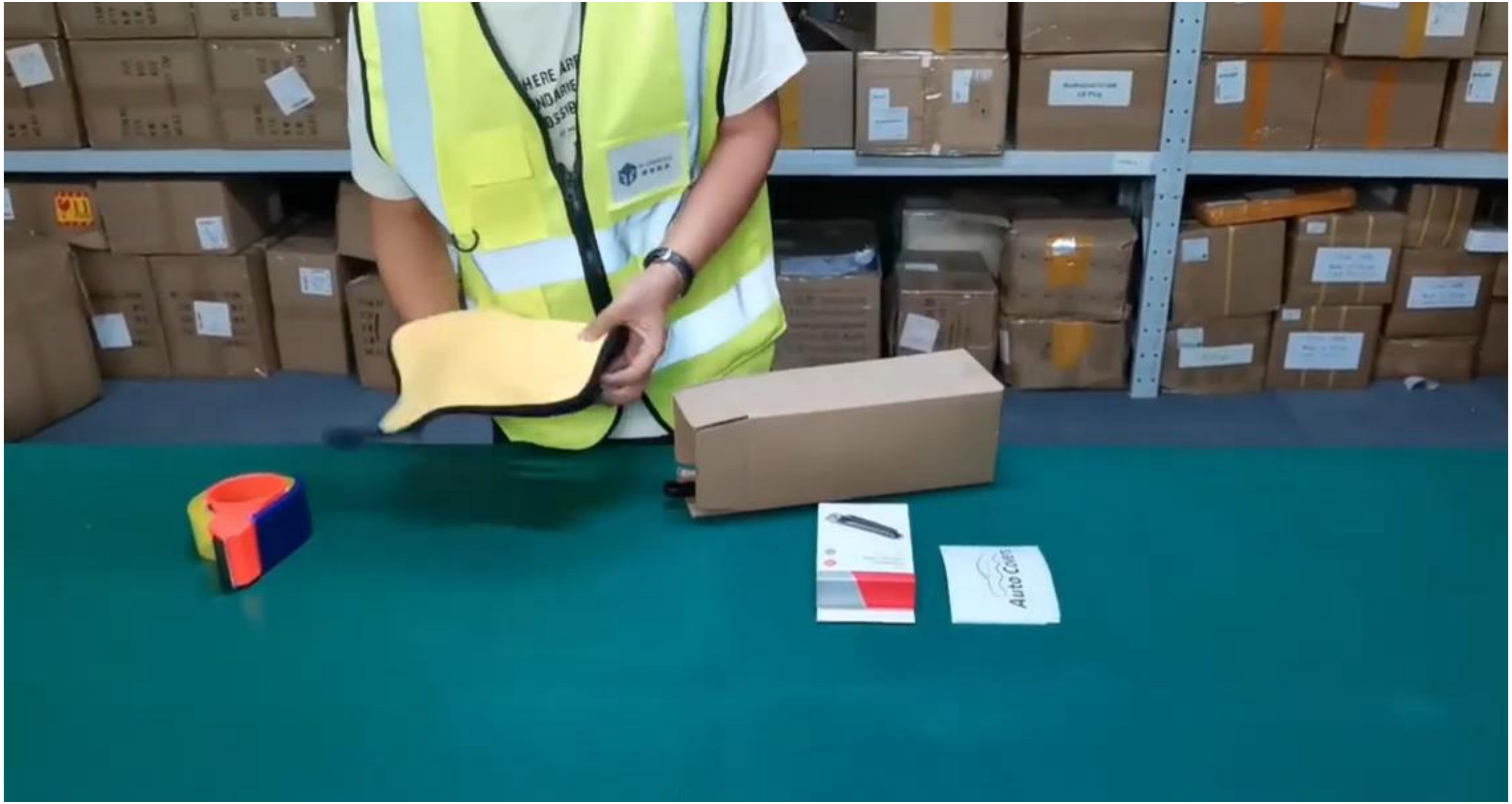


Frame 0

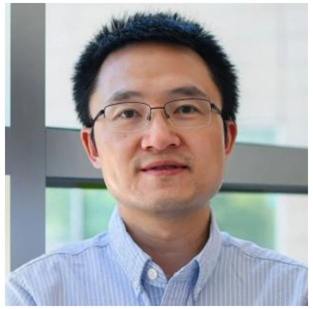
Challenges and Opportunities on Learning from Human Videos

- Understanding of human manipulation from videos is still challenging
  - 3D understanding
  - Deformable, articulated objects
  - Long-horizon tasks
- Trajectory transfer & optimization is slow
  - Better & faster optimization tools
  - Policy learning, e.g., using data from trajectory optimization
- Dexterous manipulation with multi-finger hands
  - Force feedback & tactile sensing
  - Bimanual manipulation

Robot Manipulation is still an Open Challenge



Intelligent Robotics and Vison Lab (IRVL)

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

Assisted by
Ms. Rhonda Walls

Thank you!