Learning Robotic Manipulation from Human Demonstration Videos

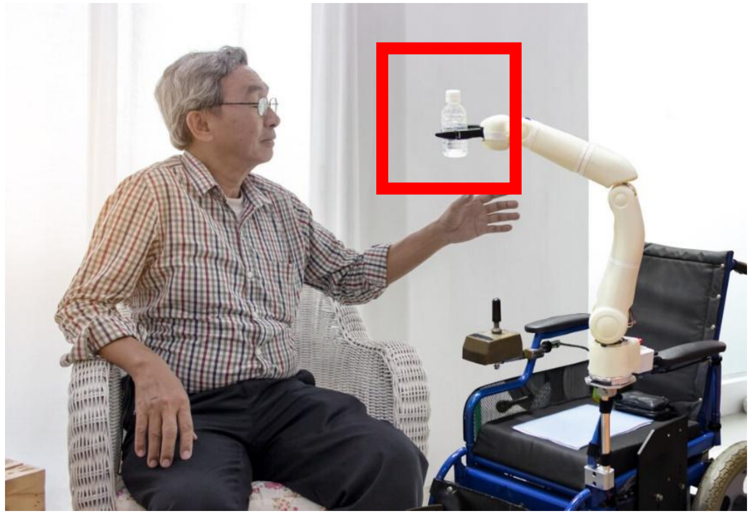
Yu Xiang
Assistant Professor
Intelligent Robotics and Vision Lab
The University of Texas at Dallas

10/15/2025

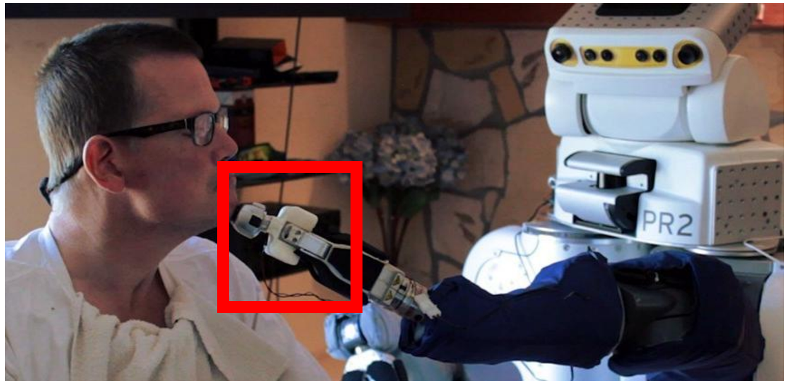
UNC Guest Lecture

Future Intelligent Robots in Human Environments

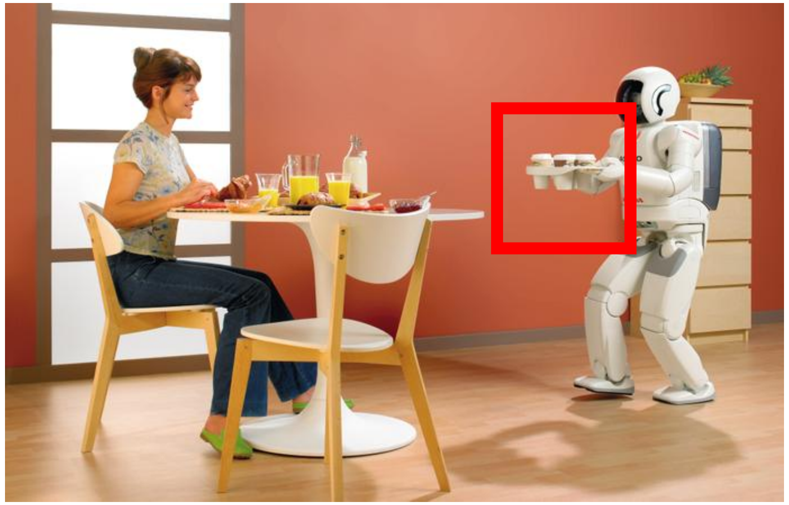
Manipulation



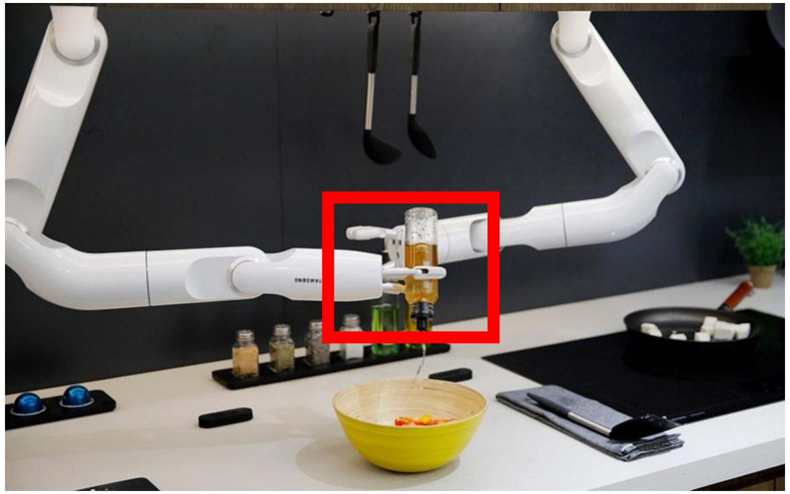
Senior Care



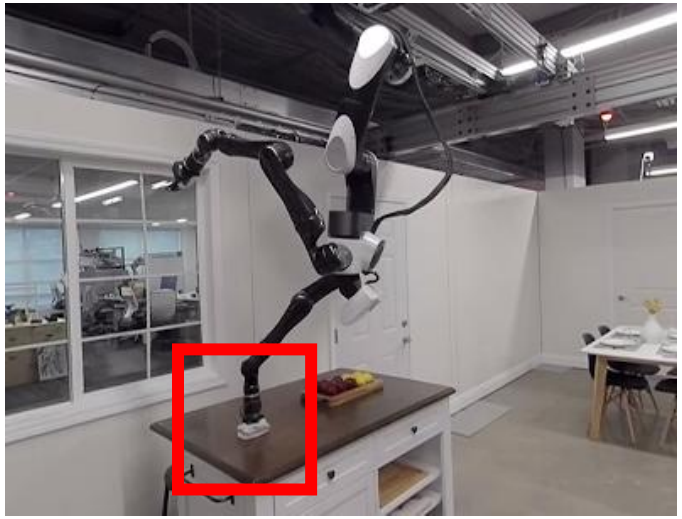
Assisting



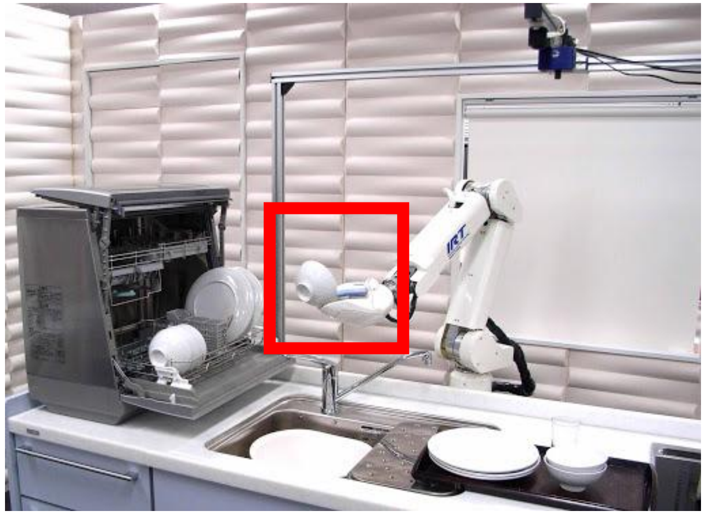
Serving



Cooking



Cleaning

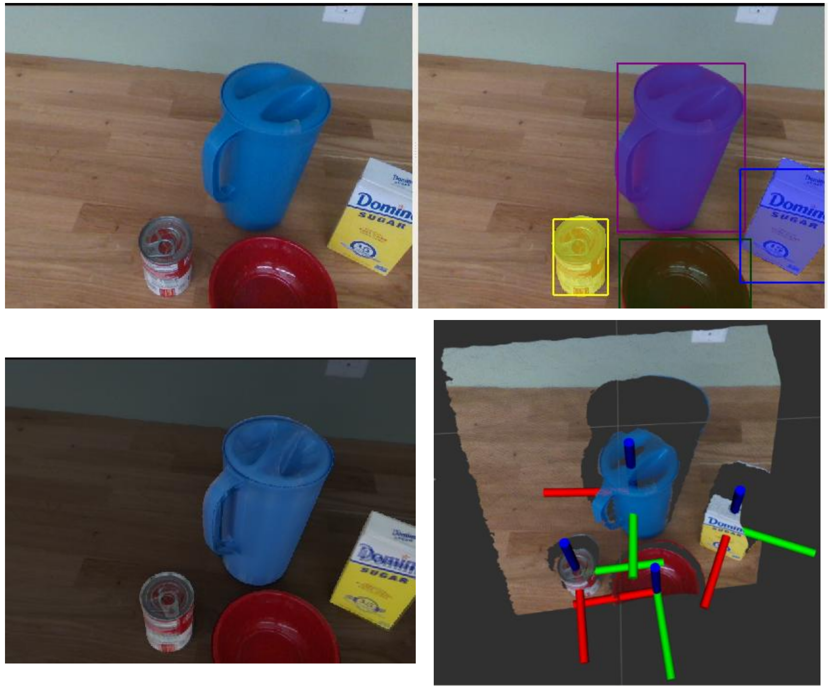


Dish washing

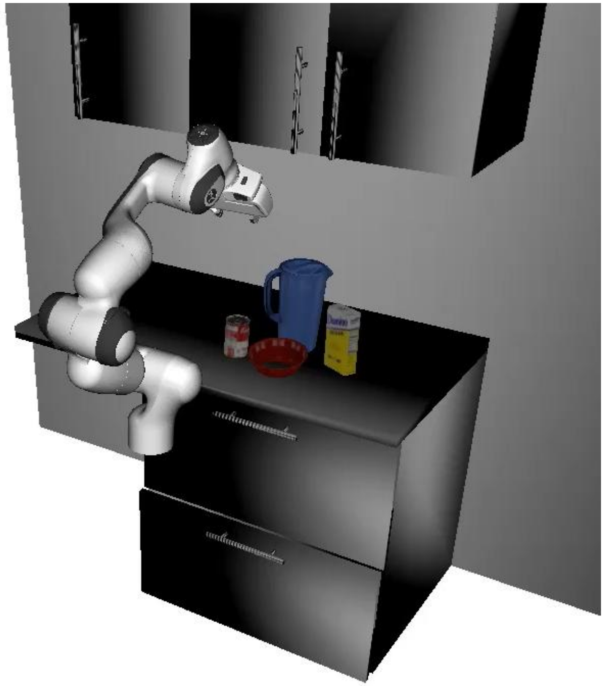
“Traditional” Approach for Robot Manipulation

Perception → Planning → Control

6D object pose estimation



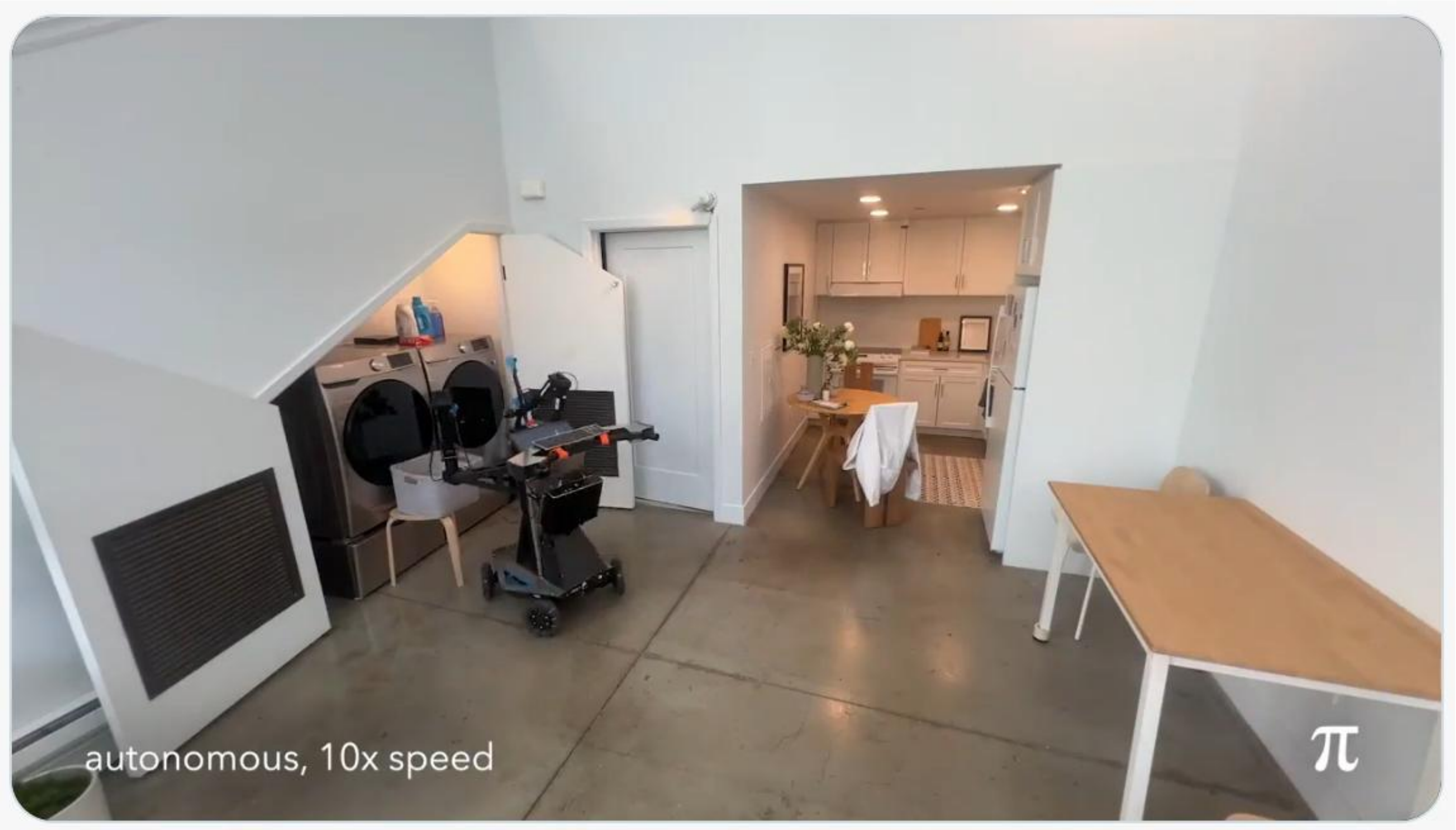
Grasp planning and motion planning



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

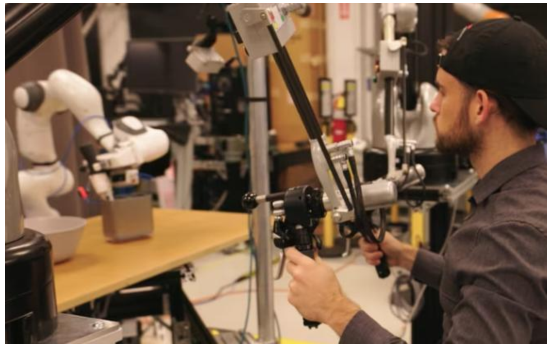
Some Recent Breakthroughs

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

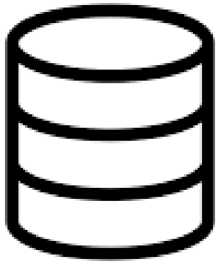
Key Ingredient: Imitation Learning

Kinesthetic Teaching

Teleoperation

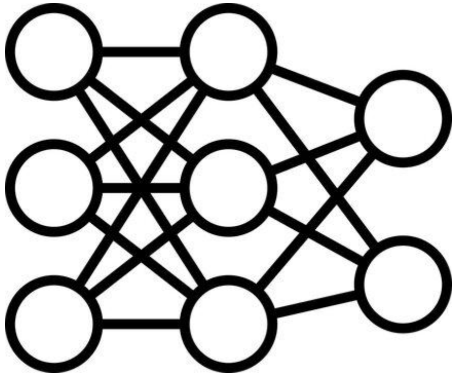


Collect Demonstrations



(state, action)

A Dataset of State-Action Pairs



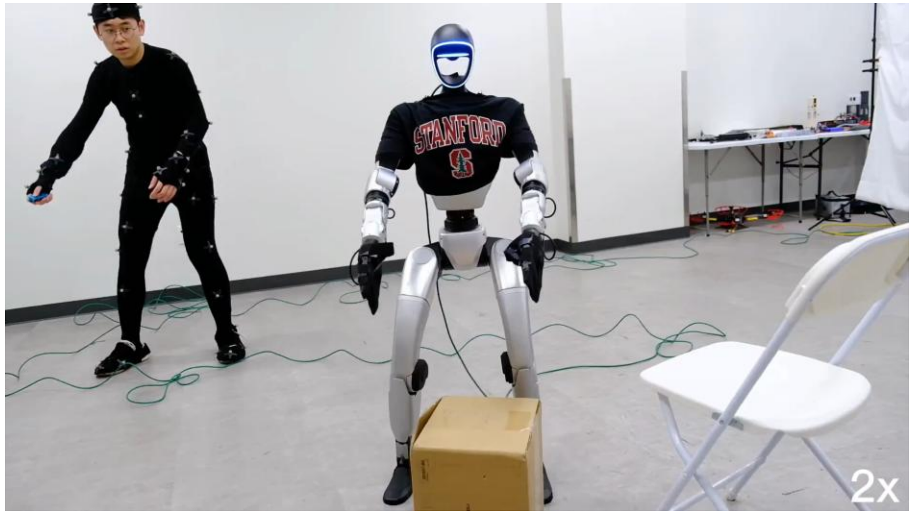
Train a Policy Network

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

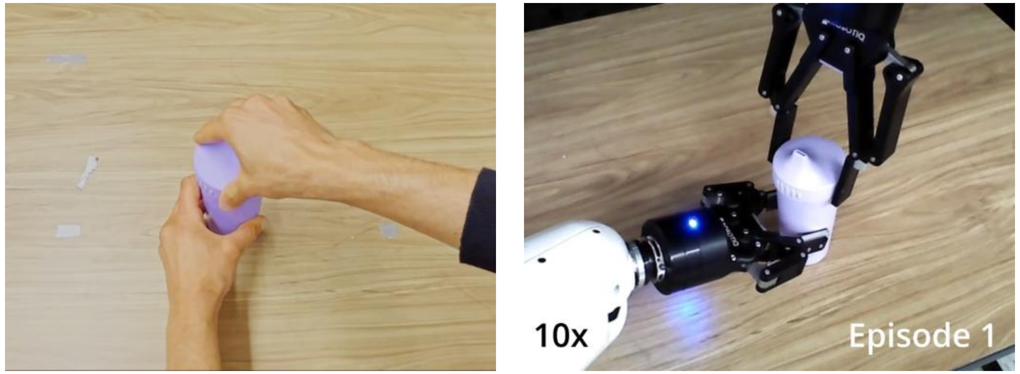
Learning Manipulation from Human Videos

Image generated by ChatGPT

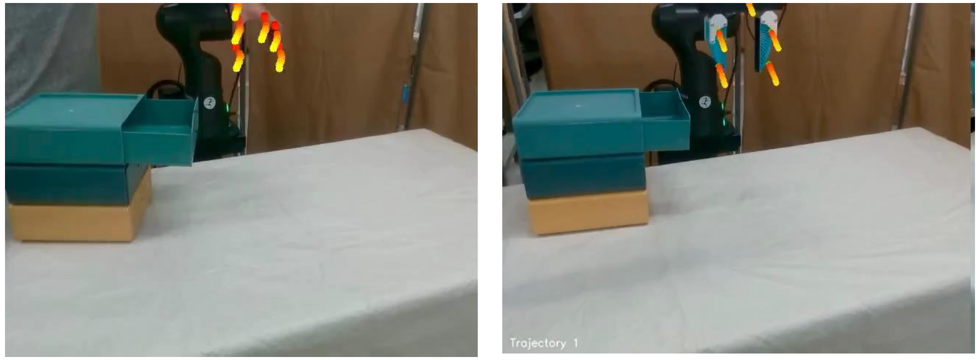
Learning Manipulation from Human Videos

- Imitation learning: convert human → robot actions, then imitate

DexMV, Qin et al. UCSD, ECCV 2022



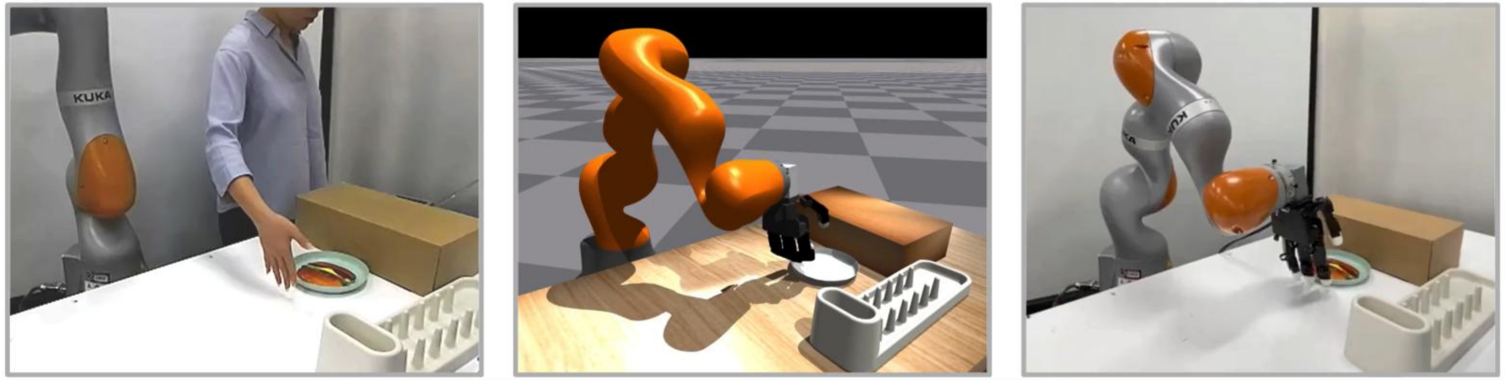
ScrewMimic, Bahety et al. UT Austin, RSS 2024



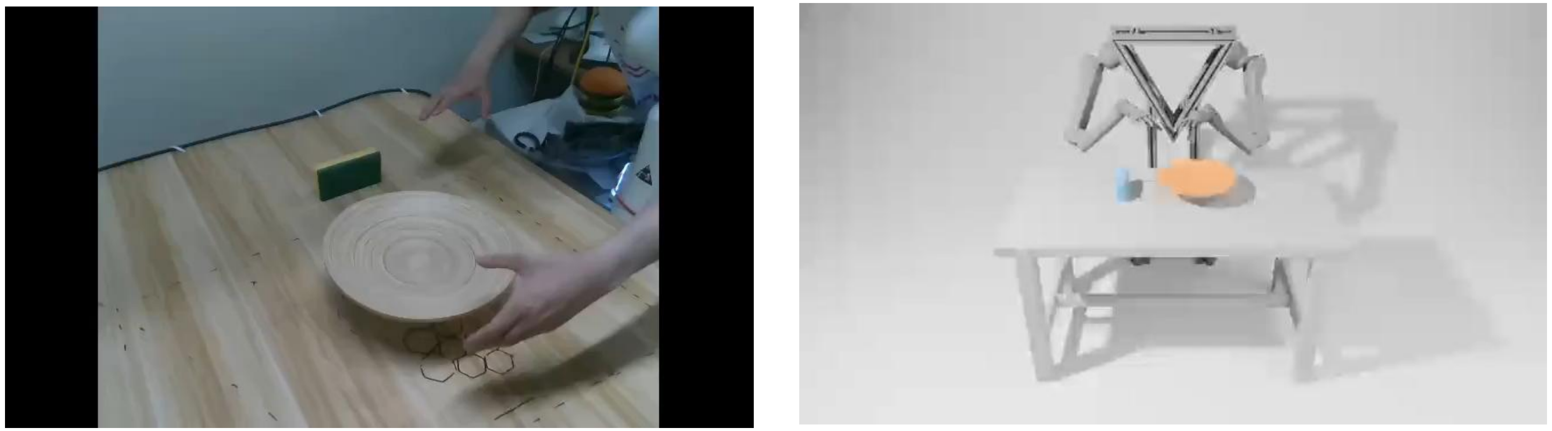
Motion Tracks, Ren et al. Cornell & Stanford, 2025

Learning Manipulation from Human Videos

- RL: replicate the environment in simulation, then train a policy



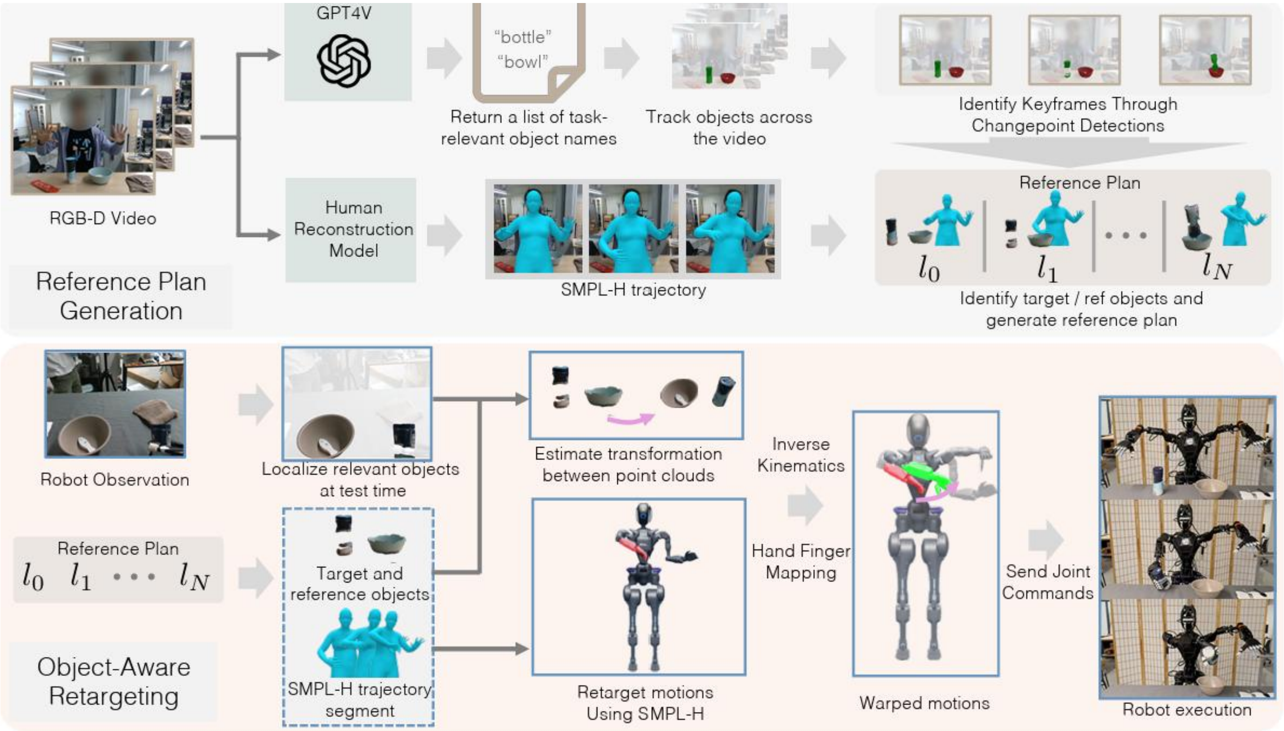
HUMAN2SIM2ROBOT, Lum et al. Stanford, 2025



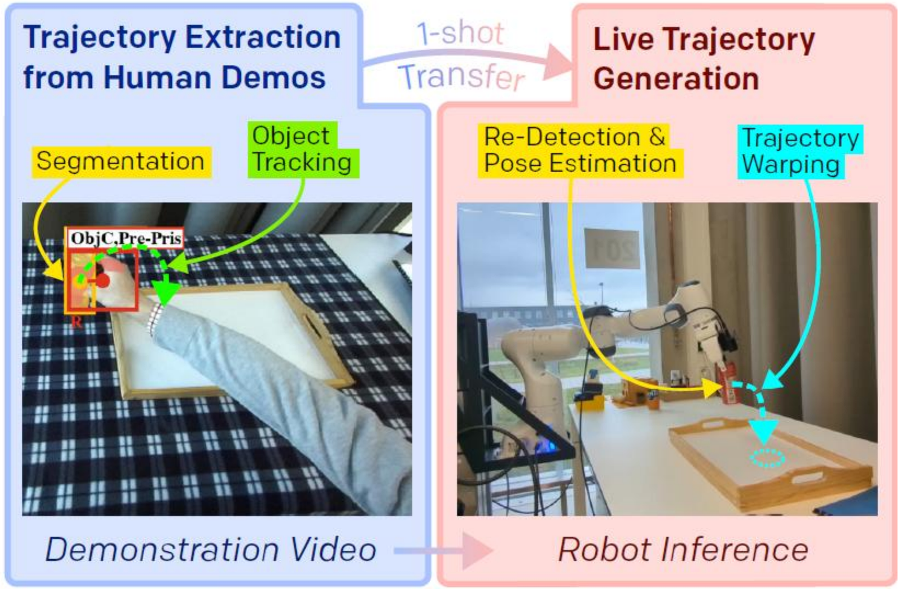
HERMES, Yuan et al., Tshinghua, 2025

Learning Manipulation from Human Videos

- Training-free: use perception + geometry to transfer trajectories



OKAMI, Li et al. UT Austin, CoRL 2024

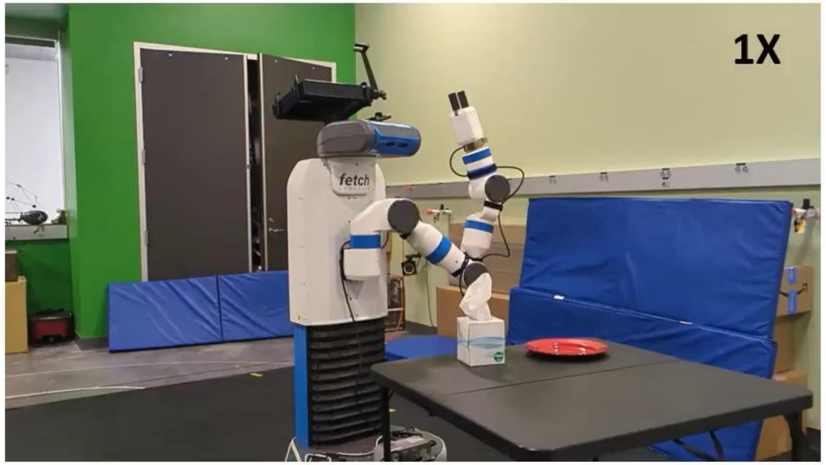


Trajectory Transfer, Heppert et al. University of Freiburg, IROS 2024

Our Work: One-Shot Human-to-Robot Trajectory Transfer

One-shot Human Demonstration

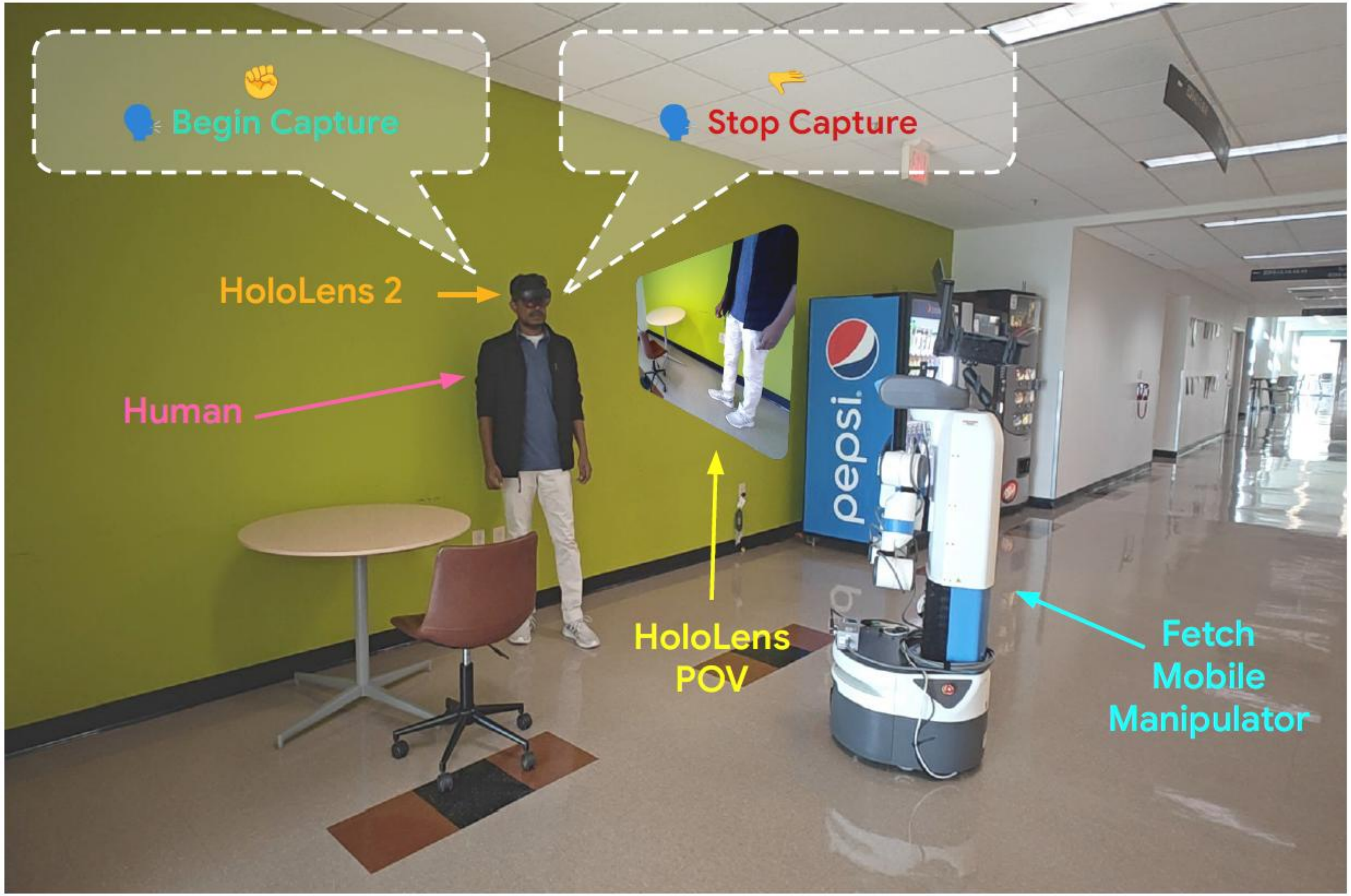
Robot Execution in Different Environments



Sai Haneesh Allu

Jishnu Jaykumar P

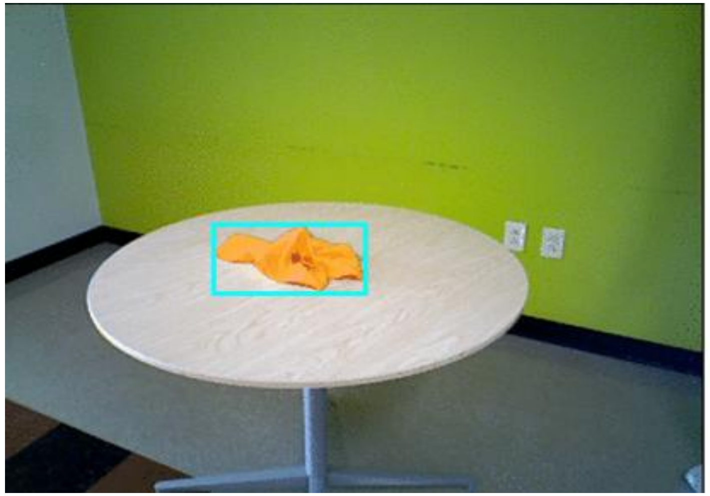
Human Demonstration Collection



Human Demonstration Collection

Human Demonstration Collection

Human Demonstration Collection



Clean table using Towel

Close jar with Red Lid

Pour Tumbler

Understanding of the Human Demonstrations

Text Prompt:
"Brown Chair"

Grounding
DINO

SAM2

Understanding of the Human Demonstrations

HaMeR

Optimization using Depth

Human-to-Robot Grasp Transfer

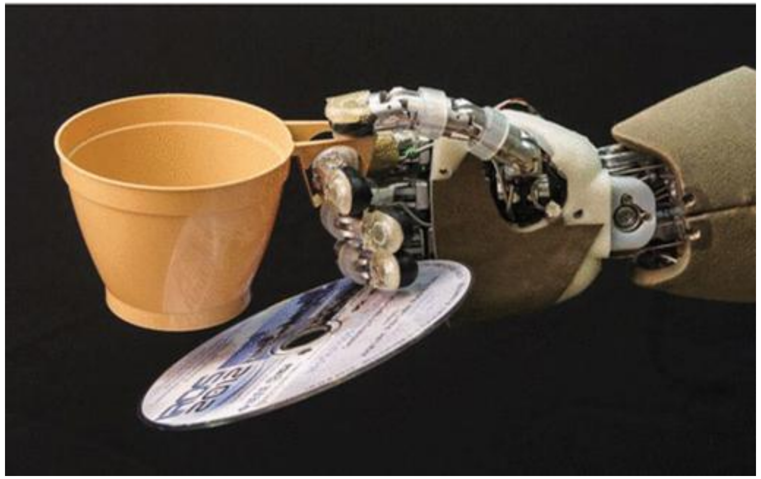
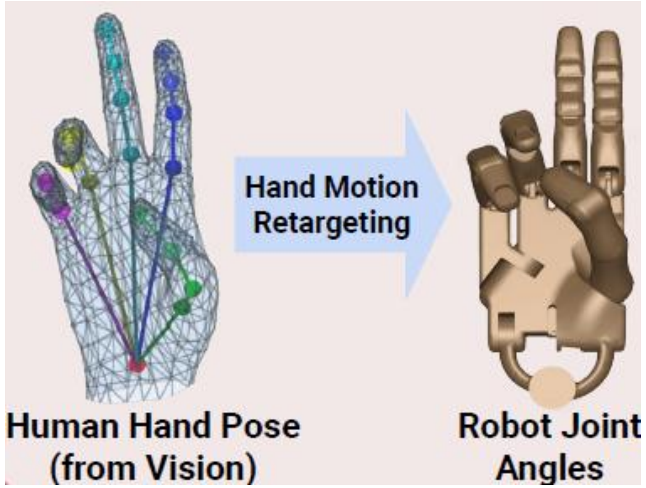


Image generated by ChatGPT

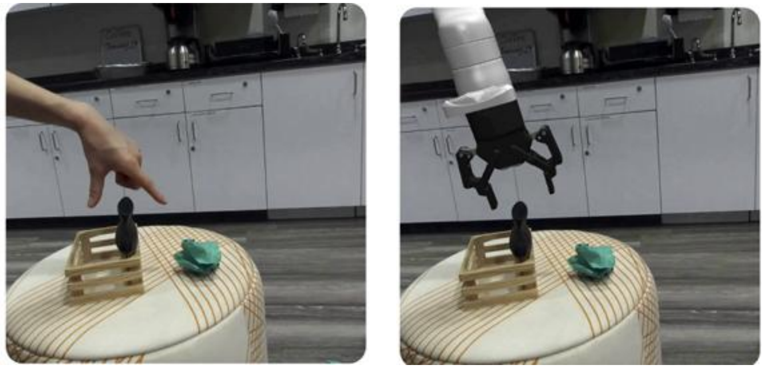
Human-to-Robot Grasp Transfer

- Retargeting



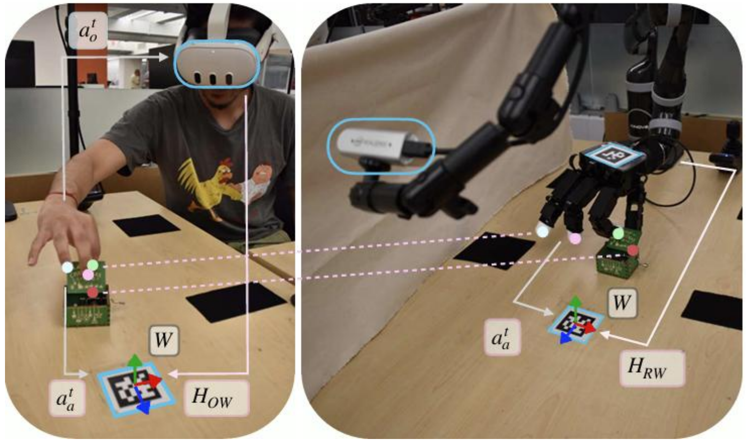
DexMV, Qin et al. UCSD, ECCV 2022

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



Phantom, Lepert et al. Stanford 2025

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

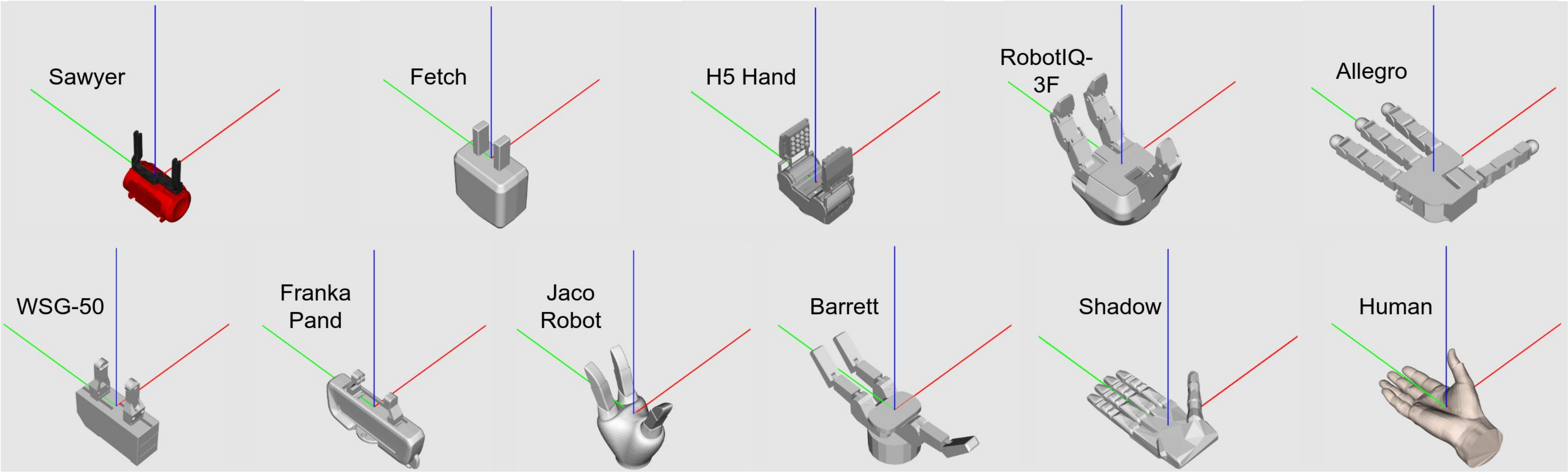


HuDOR, Guzey et al. NYU 2025

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

A Common Grasping Space

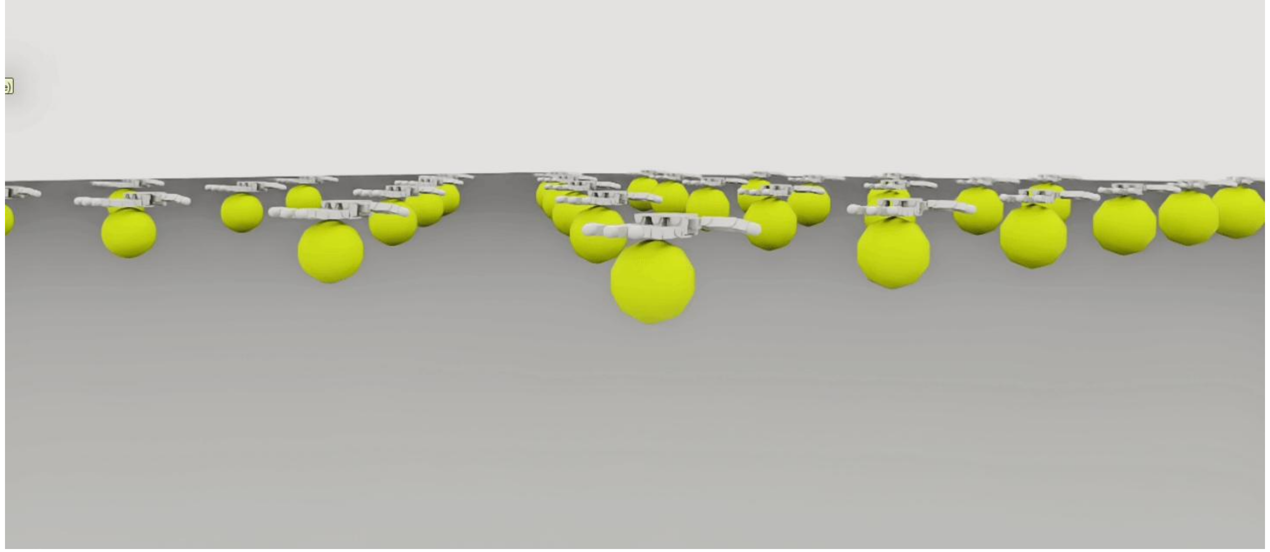
- Can we find a common grasping space for all the grippers?

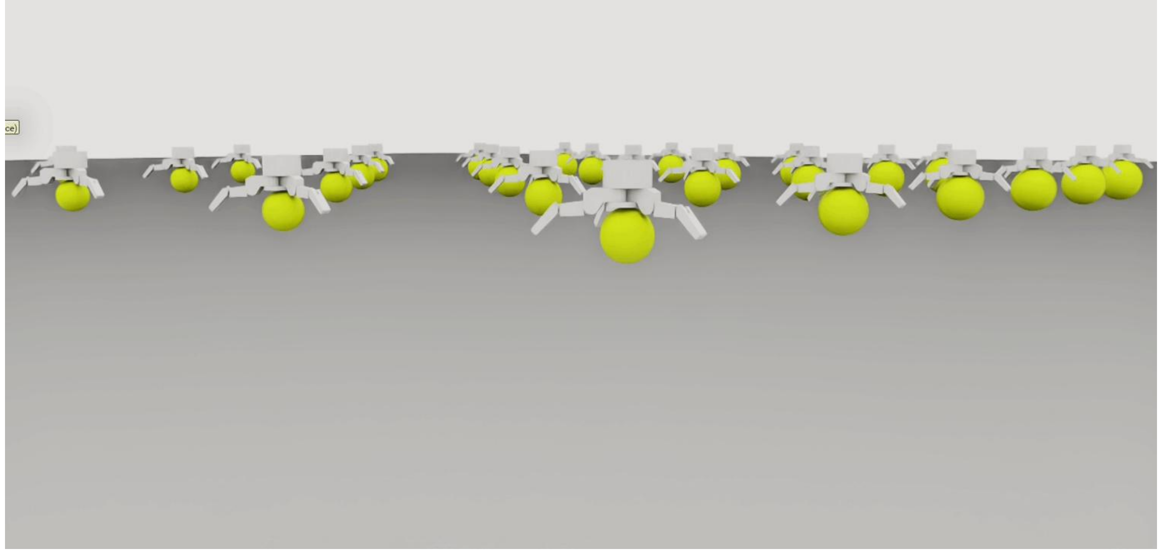


- We can align the palm orientations
- How to map fingers?

A Common Grasping Space

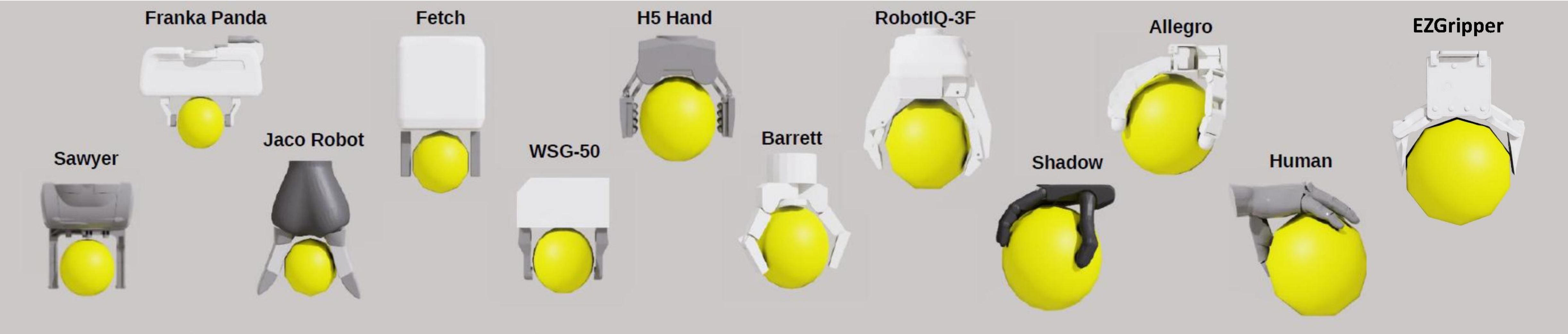
- Having the hands to grasp a common sphere
- Using contact maps on the sphere for retargeting
- Maximal sphere test in simulation





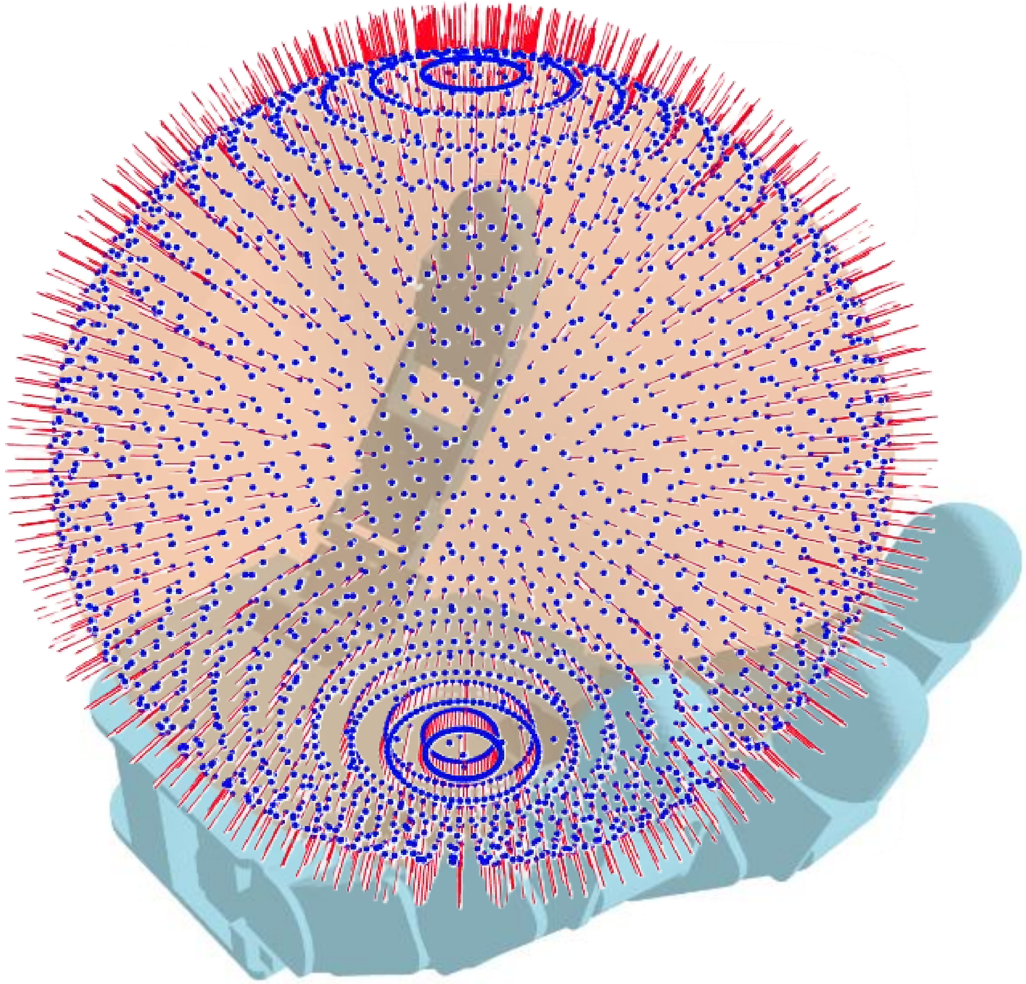
A Common Grasping Space

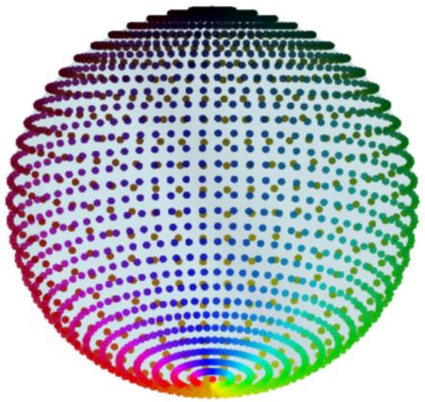
- Maximal spheres for each gripper



A Unified Gripper Coordinate Space

- Map spherical coordinates to the gripper

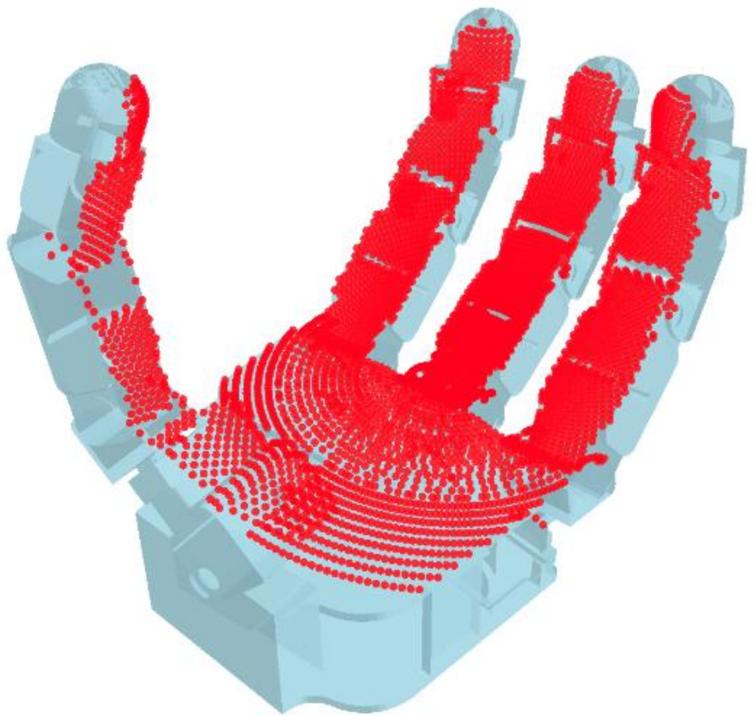


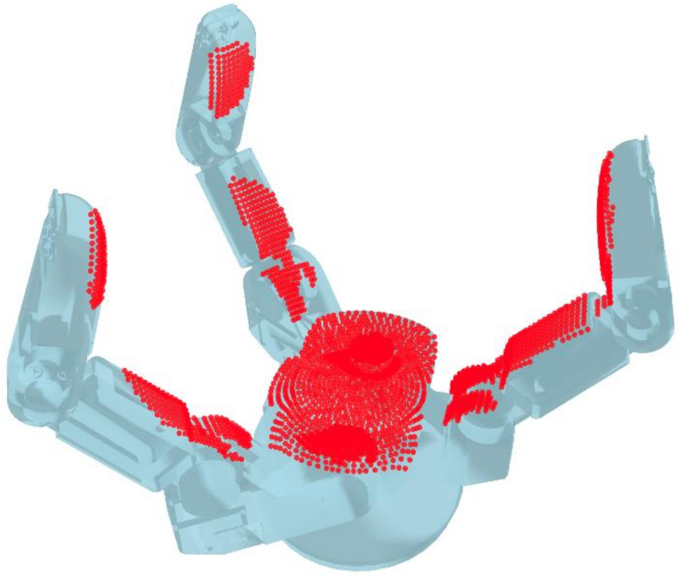


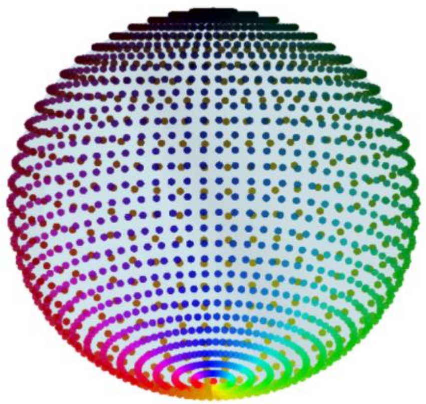
(λ, ϕ)

A Unified Gripper Coordinate Space

- Map spherical coordinates to the gripper







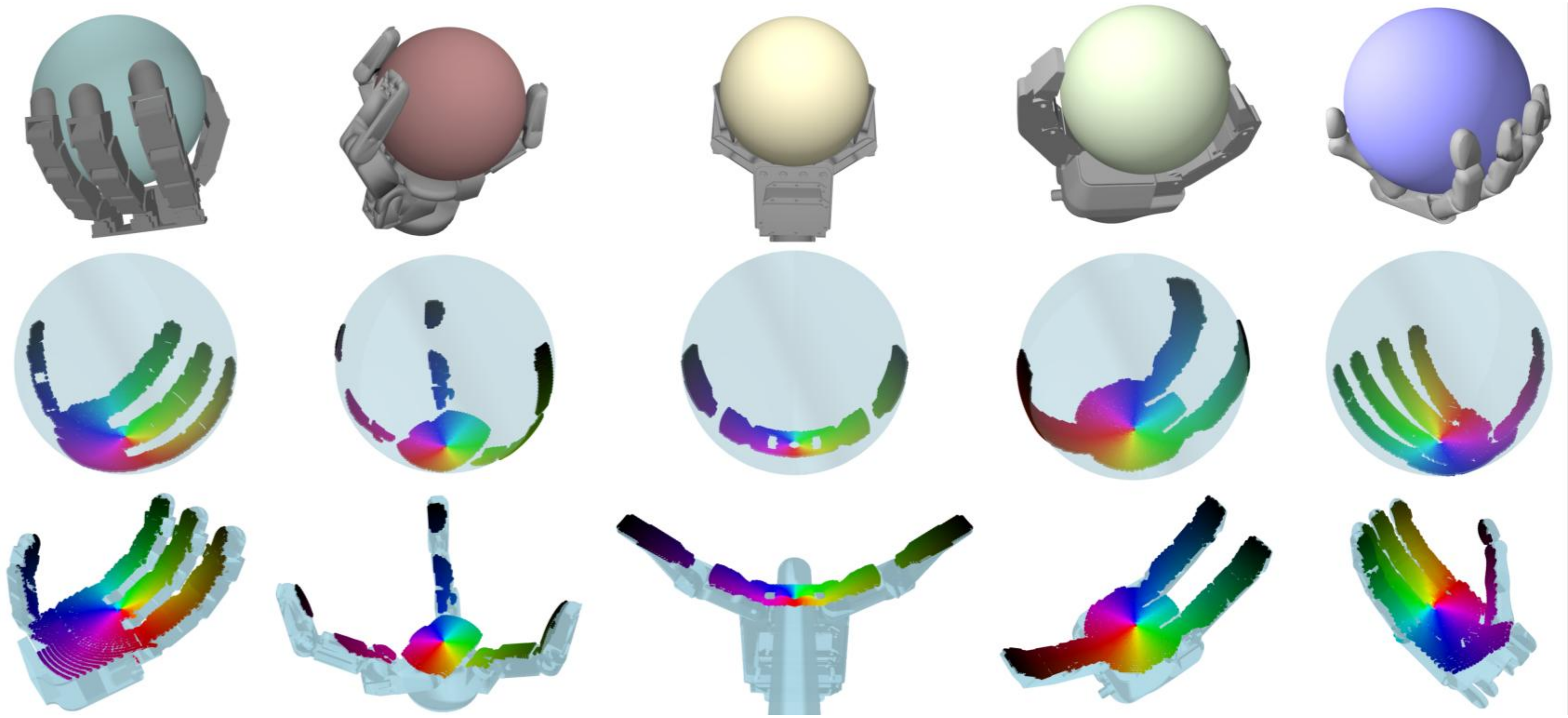
(λ, ϕ)

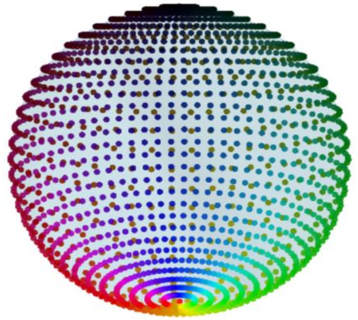
A Gripper is represented by a set of interior points $P_G = \{\mathbf{v}_g \mid \mathbf{v}_g \in \mathbb{R}^3\}$

Grasp configuration $\mathbf{q}$ changes the location of $\mathbf{v}_g$ $P_G(\mathbf{q})$

A Unified Gripper Coordinate Space

- Finger print: map spherical coordinates to the gripper



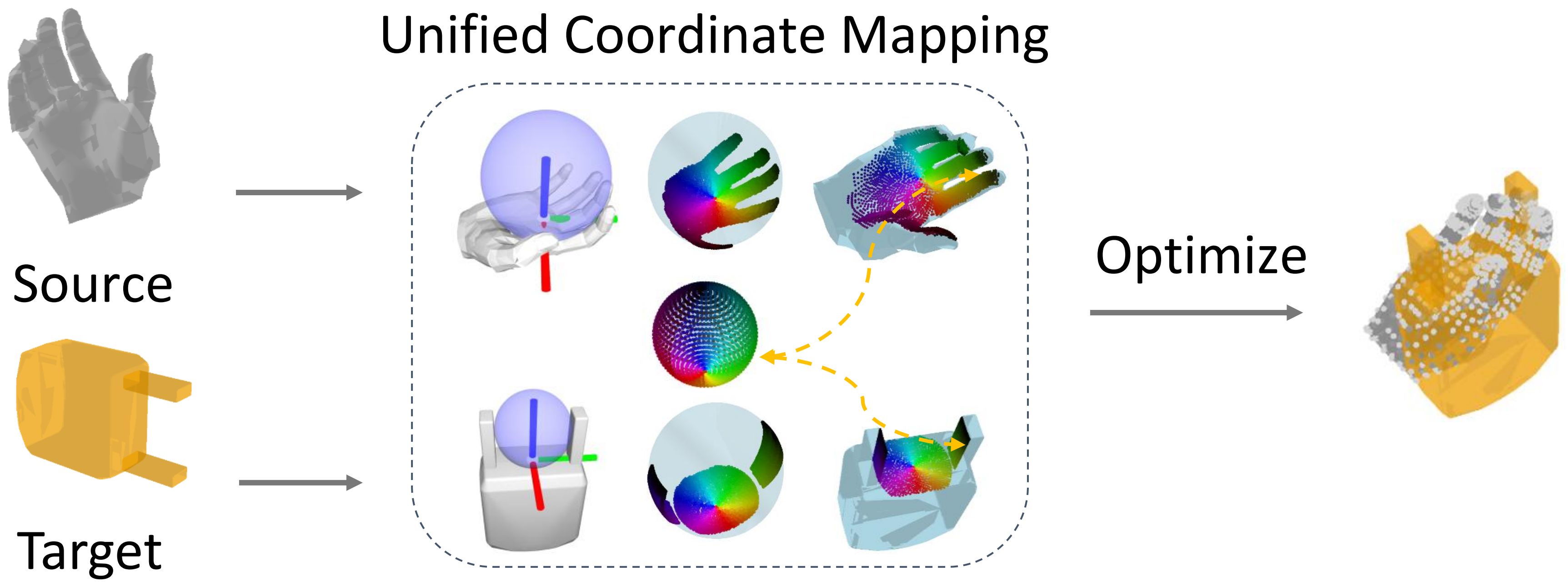


(λ, ϕ)

A UGCS coordinate is assigned to each point (fixed after assignment, independent of grasp)

$$\Phi_G = \{(\lambda_{\mathbf{v}_g}, \varphi_{\mathbf{v}_g}) \mid \mathbf{v}_g \in P_G \; ; \; \lambda_{\mathbf{v}_g}, \varphi_{\mathbf{v}_g} \in [0, 1]\}$$

Grasp Transfer



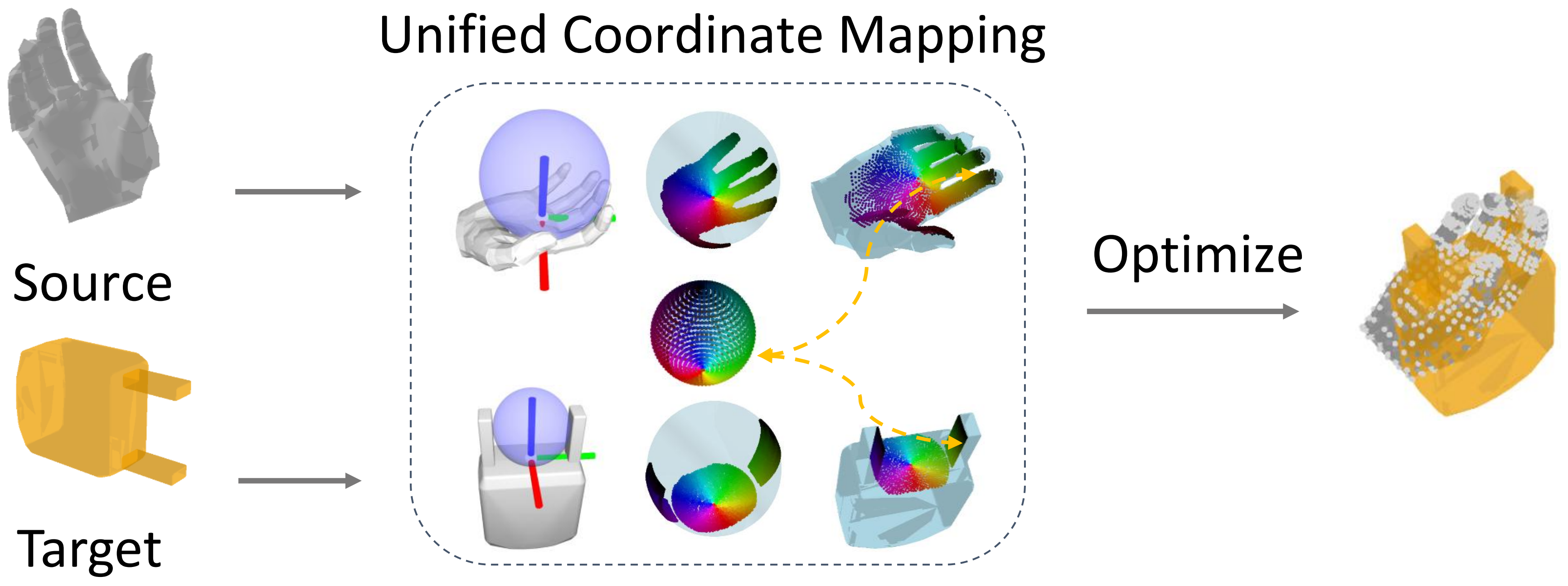
Two UGCS coordinate maps for two grippers

$$\Phi_{G_1} = \{(\lambda_{\mathbf{v}_g}, \varphi_{\mathbf{v}_g}) \mid \mathbf{v}_g \in P_{G_1} \; ; \; \lambda_{\mathbf{v}_g}, \varphi_{\mathbf{v}_g} \in [0,1]\}$$
$$\Phi_{G_2} = \{(\lambda_{\mathbf{v}_g}, \varphi_{\mathbf{v}_g}) \mid \mathbf{v}_g \in P_{G_2} \; ; \; \lambda_{\mathbf{v}_g}, \varphi_{\mathbf{v}_g} \in [0,1]\}$$

Matching their UGCS coordinates to establish correspondences (find mutually closest pairs)

$$P^c_{G_1} \subset P_{G_1}, P^c_{G_2} \subset P_{G_2}, |P^c_{G_1}| = |P^c_{G_2}|$$

Grasp Transfer



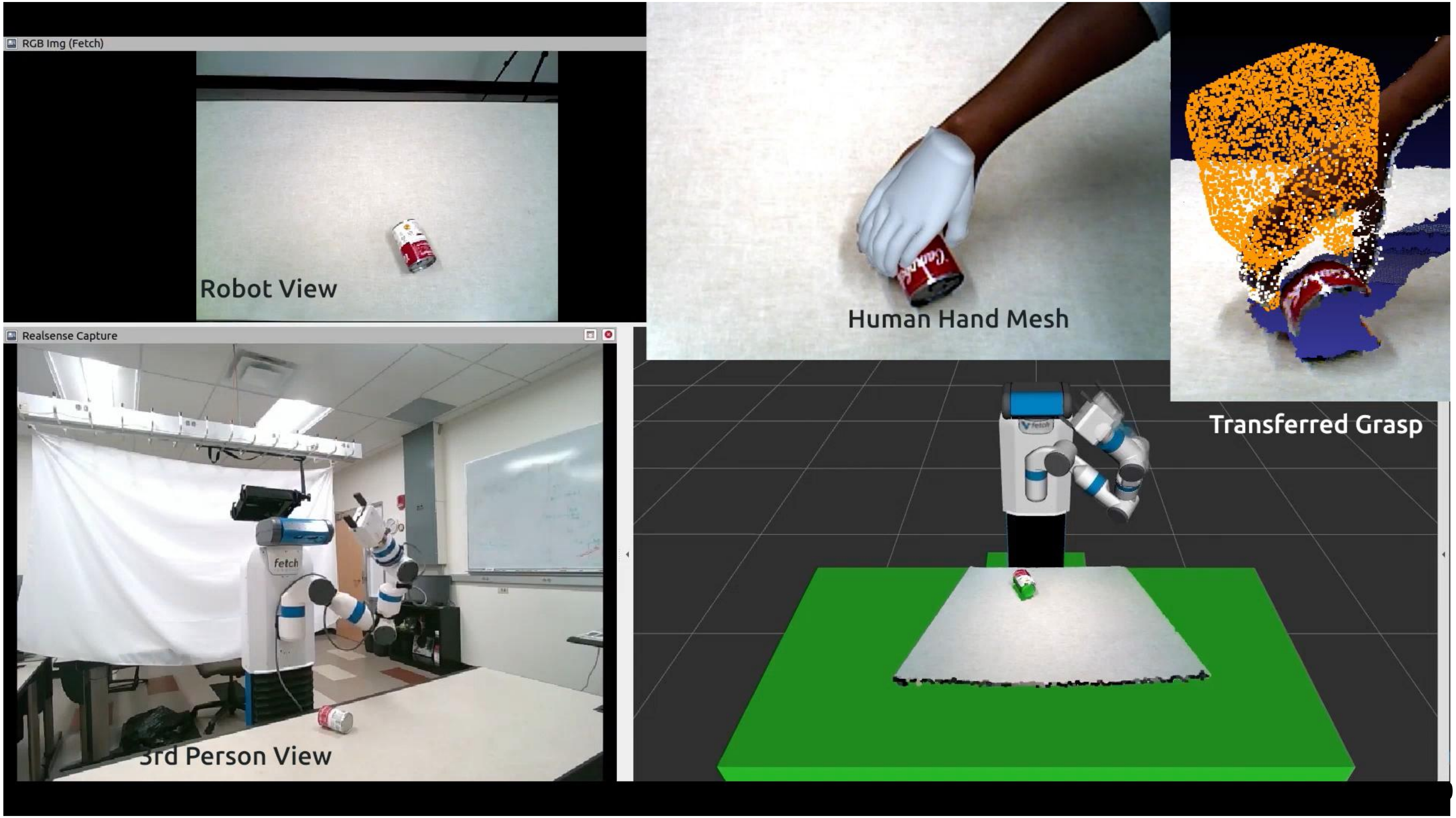
Optimize the target grasp

$$\mathbf{q}_F^* = \arg\min_{\mathbf{q}_F} \; E_{\text{dist}}(P_H^c(\mathbf{q}_H), P_F^c(\mathbf{q}_F)) + E_{\text{n}}(\mathbf{q}_F)$$

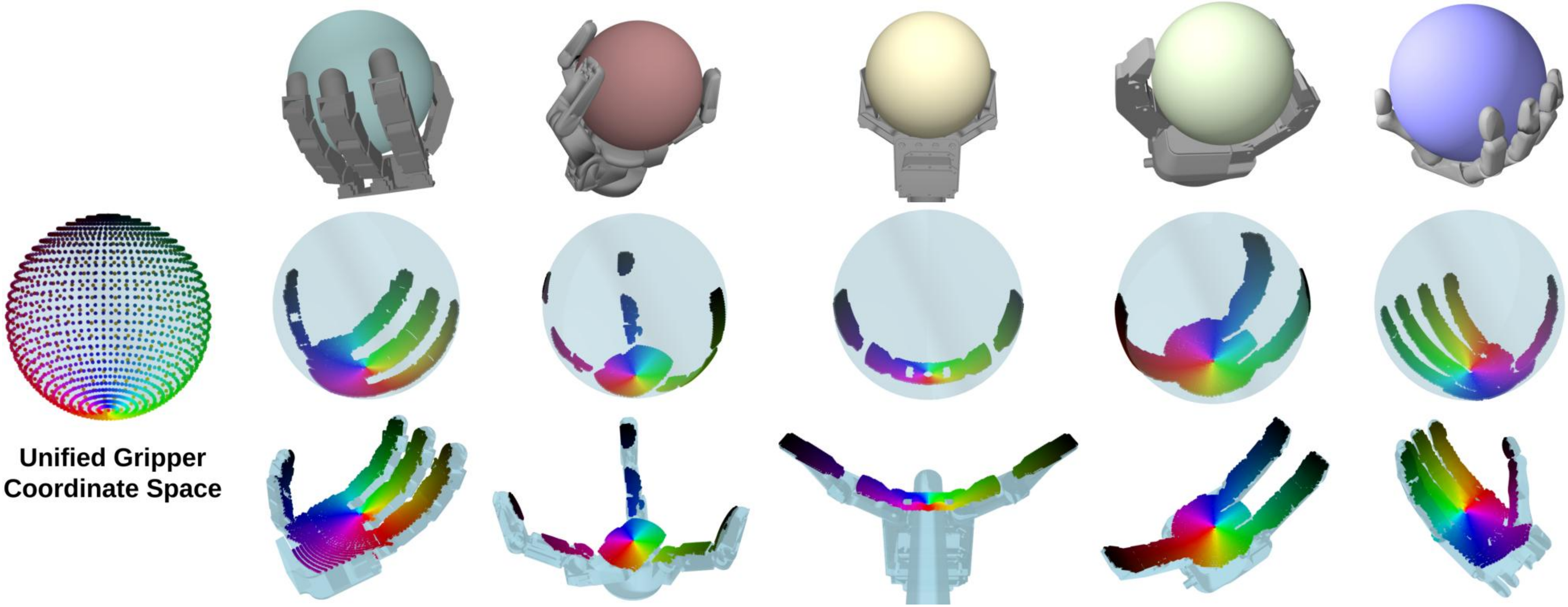
Reference grasp

Joint limits

Grasp Transfer

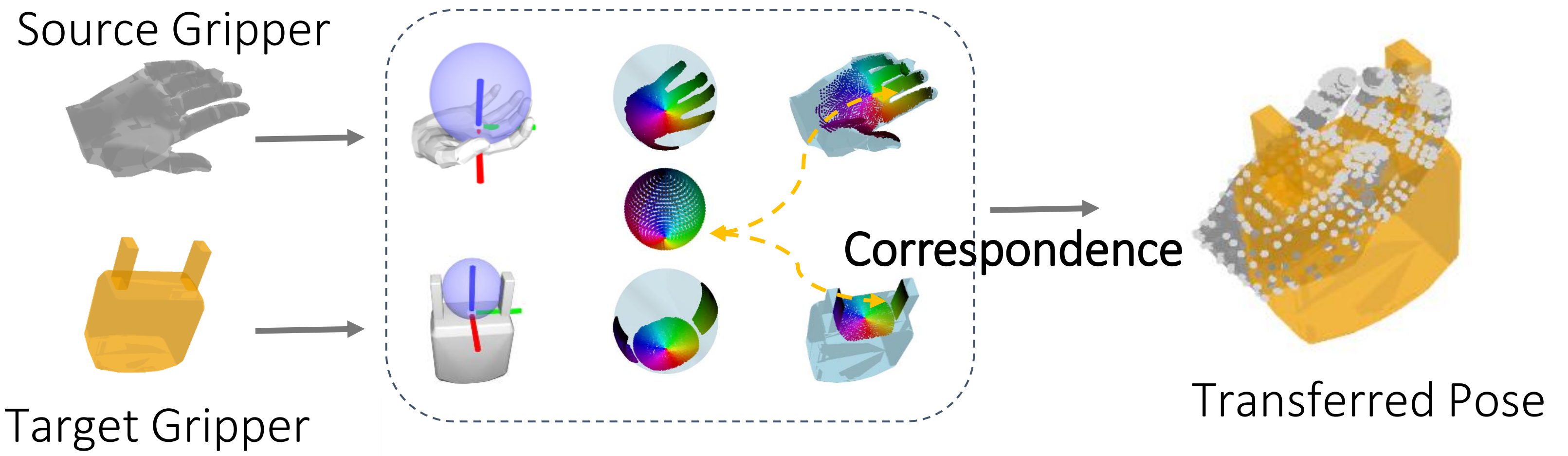


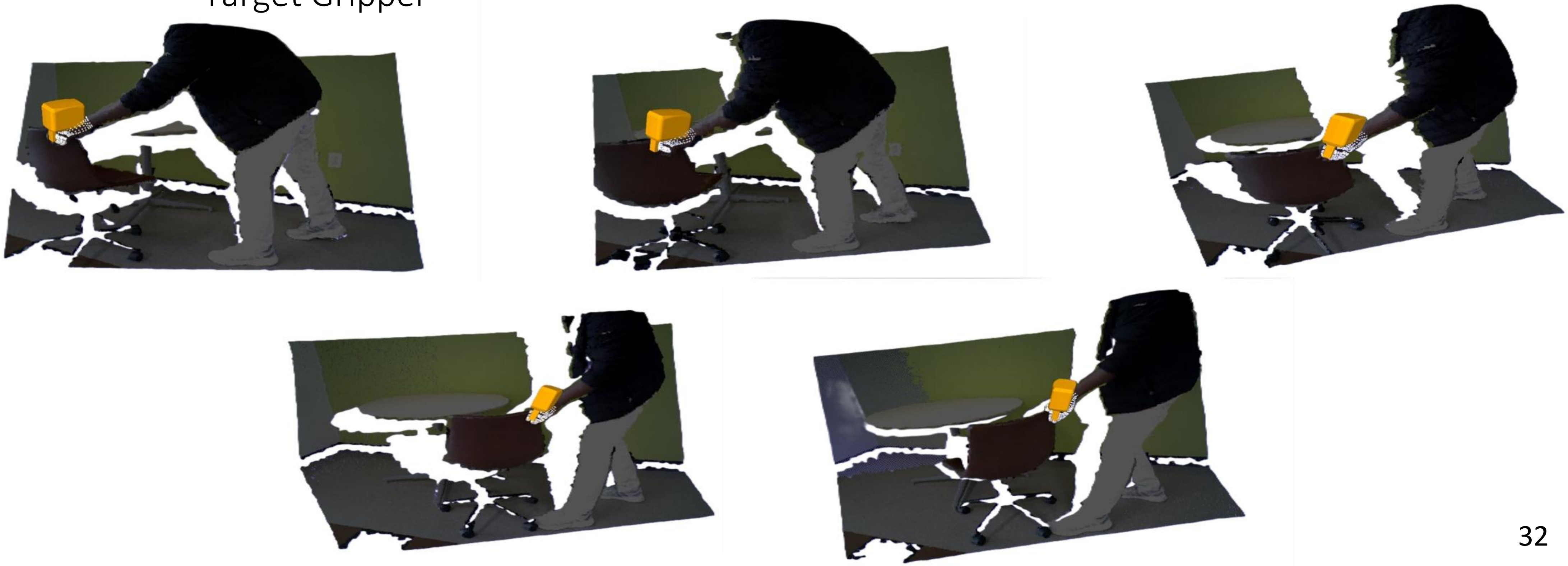
RobotFingerPrint



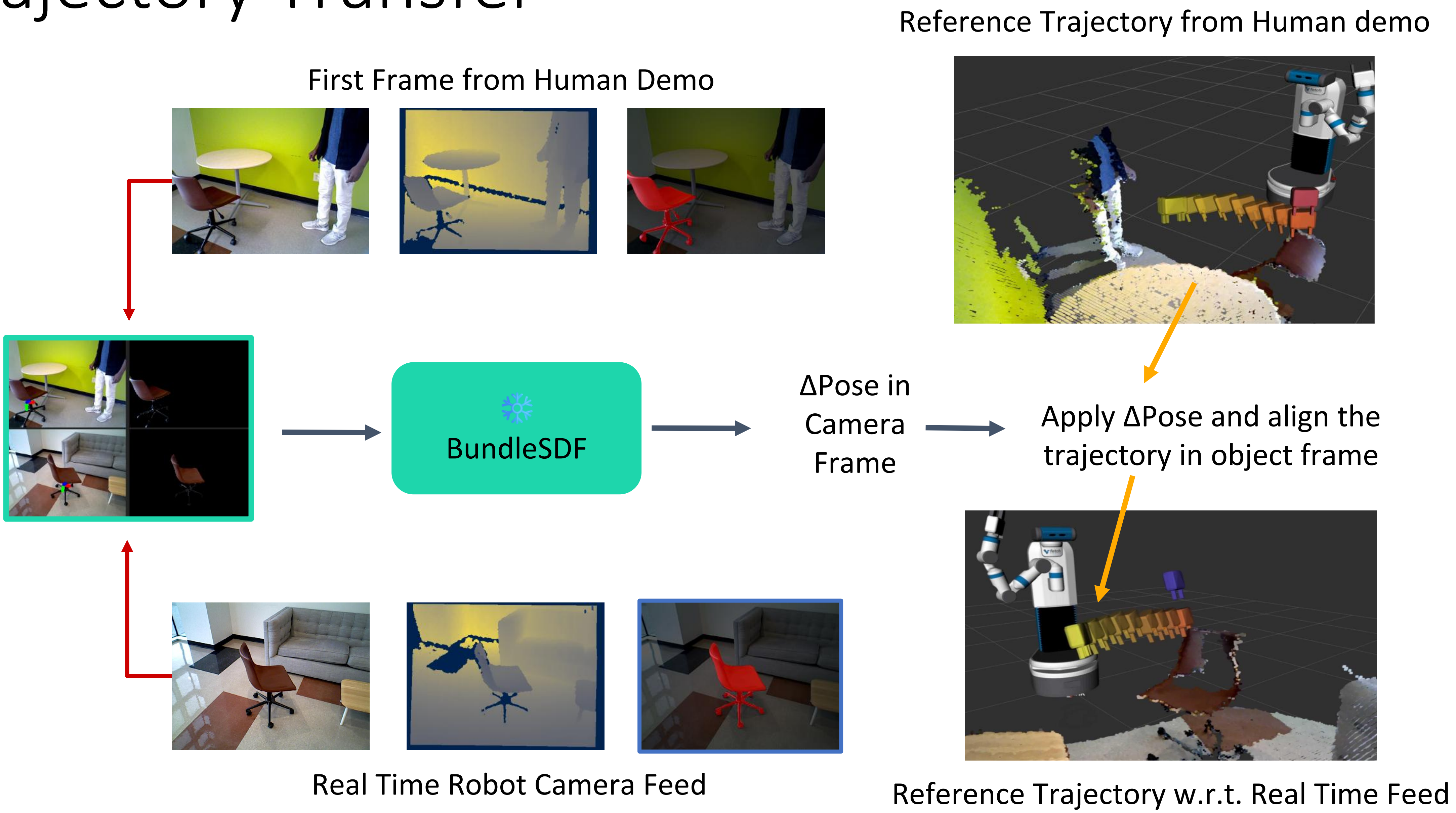
RobotFingerPrint: Unified Gripper Coordinate Space for Multi-Gripper Grasp Synthesis and Transfer.
Ninad Khargonkar, Luis Felipe Casas, Balakrishnan Prabhakaran, Yu Xiang. In IROS, 2025.

Understanding of the Human Demonstrations





Trajectory Transfer

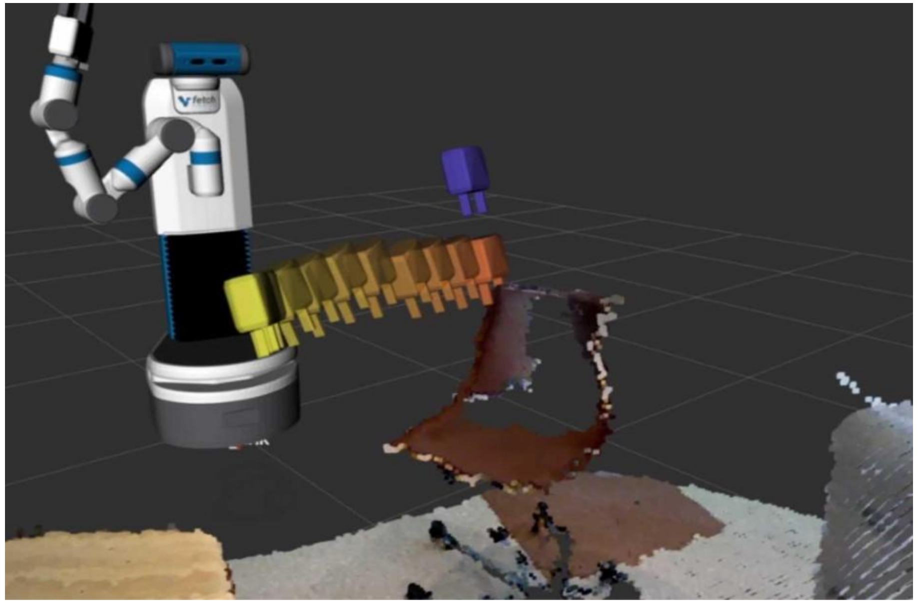


Trajectory Transfer

- How to follow the transferred gripper trajectory?

Task Space

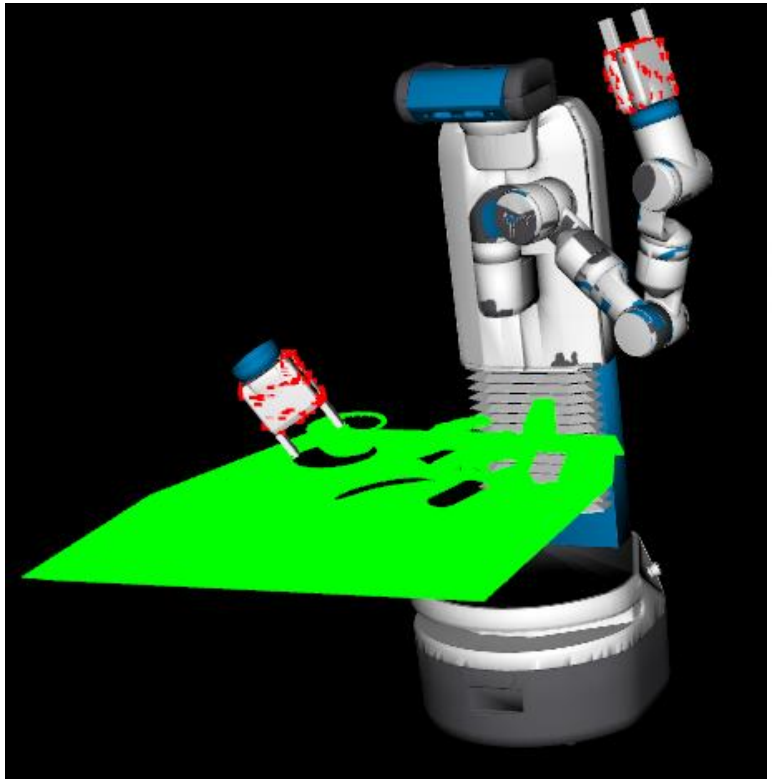
Robot View



Reference Trajectory w.r.t. Real Time Feed

Trajectory Optimization

- Point Cloud-based Cost Function for Goal Reaching



Gripper pose

Goal pose

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

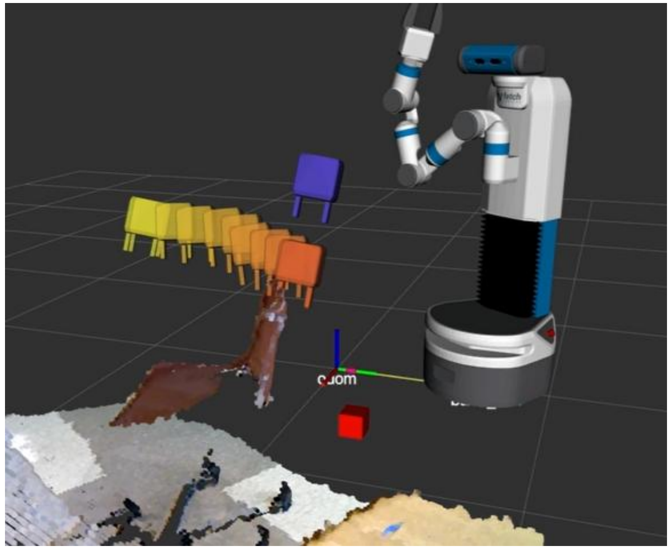
Points on the gripper

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



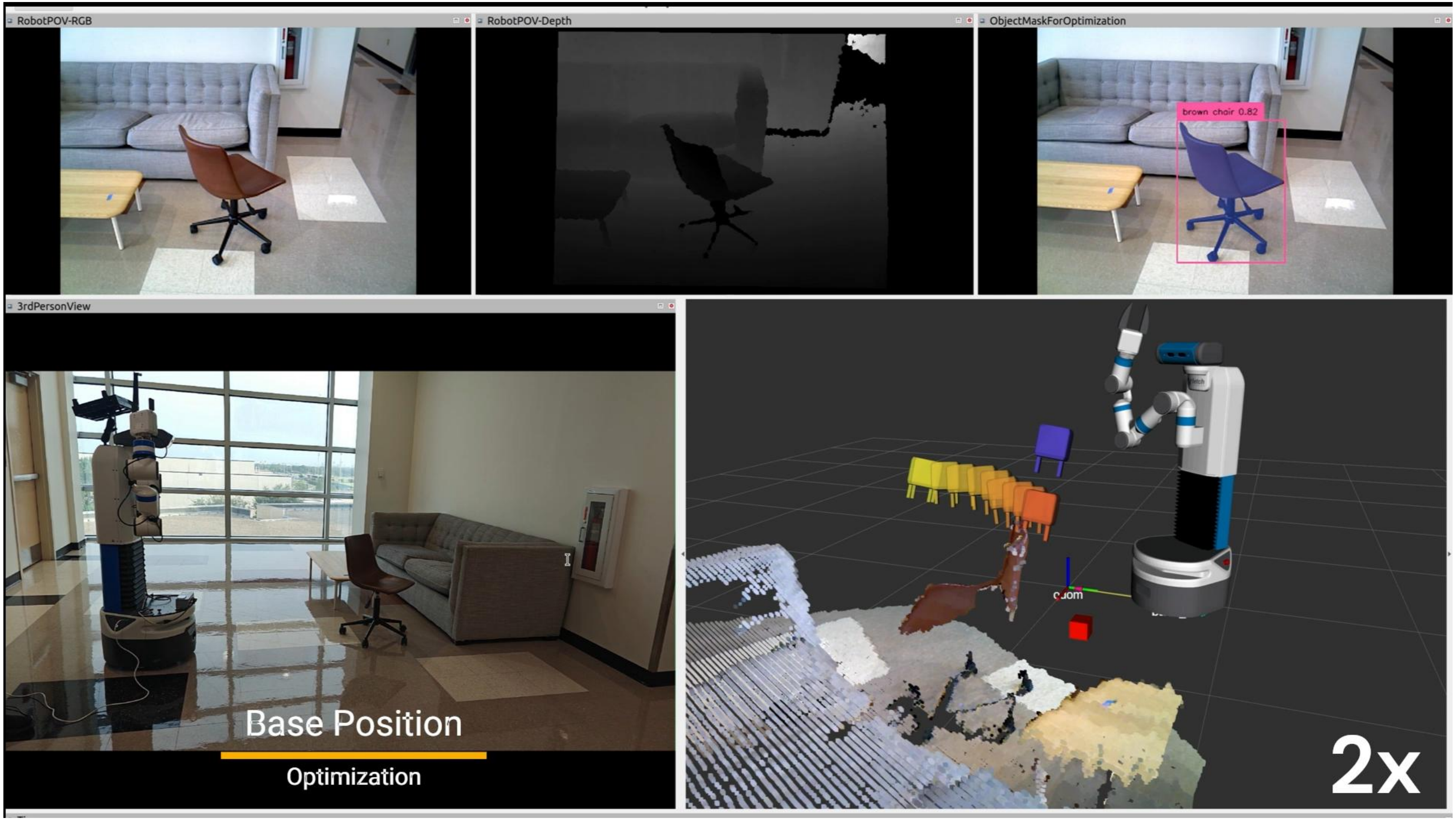
Gripper pose $\mathcal{T} = \{\mathbf{T}_1, \mathbf{T}_2 \dots, \mathbf{T}_N\}$ Known

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

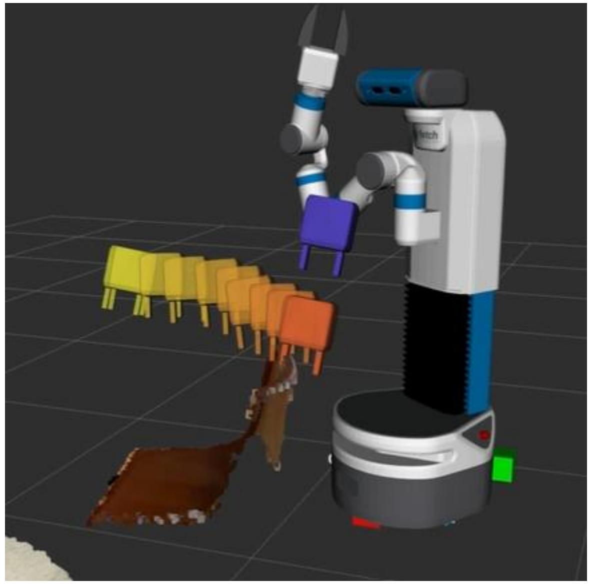
$$\begin{aligned} \arg\min_{\mathbf{x}, \mathcal{Q}} \quad & \lambda_{\text{effort}} \|\mathbf{x}\|^2 + \lambda_{\text{goal}} \textstyle\sum_{i=1}^{N} c_{\text{goal}}(\mathbf{T}(\mathbf{q}_i), \underline{\mathbf{T}(\mathbf{x}) \cdot \mathbf{T}_i}) \\ \text{s.t.,} \quad & \mathbf{x}_l \leq \mathbf{x}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

Gripper goal in new base

Optimizing the Robot Base Location



Optimizing the Robot Trajectory



- Find the trajectory to follow the gripper poses well

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

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

Gripper trajectory in new robot base

$$\arg\min_{\mathcal{Q}, \dot{\mathcal{Q}}} \sum_{t=1}^{T} c_{\text{goal}}(\mathbf{T}(\mathbf{q}_t), \mathbf{T}_t) + \lambda_1 c_{\text{collision}}(\mathbf{q}_t) + \lambda_2 \sum_{t=1}^{T} \|\dot{\mathbf{q}}_t\|^2$$

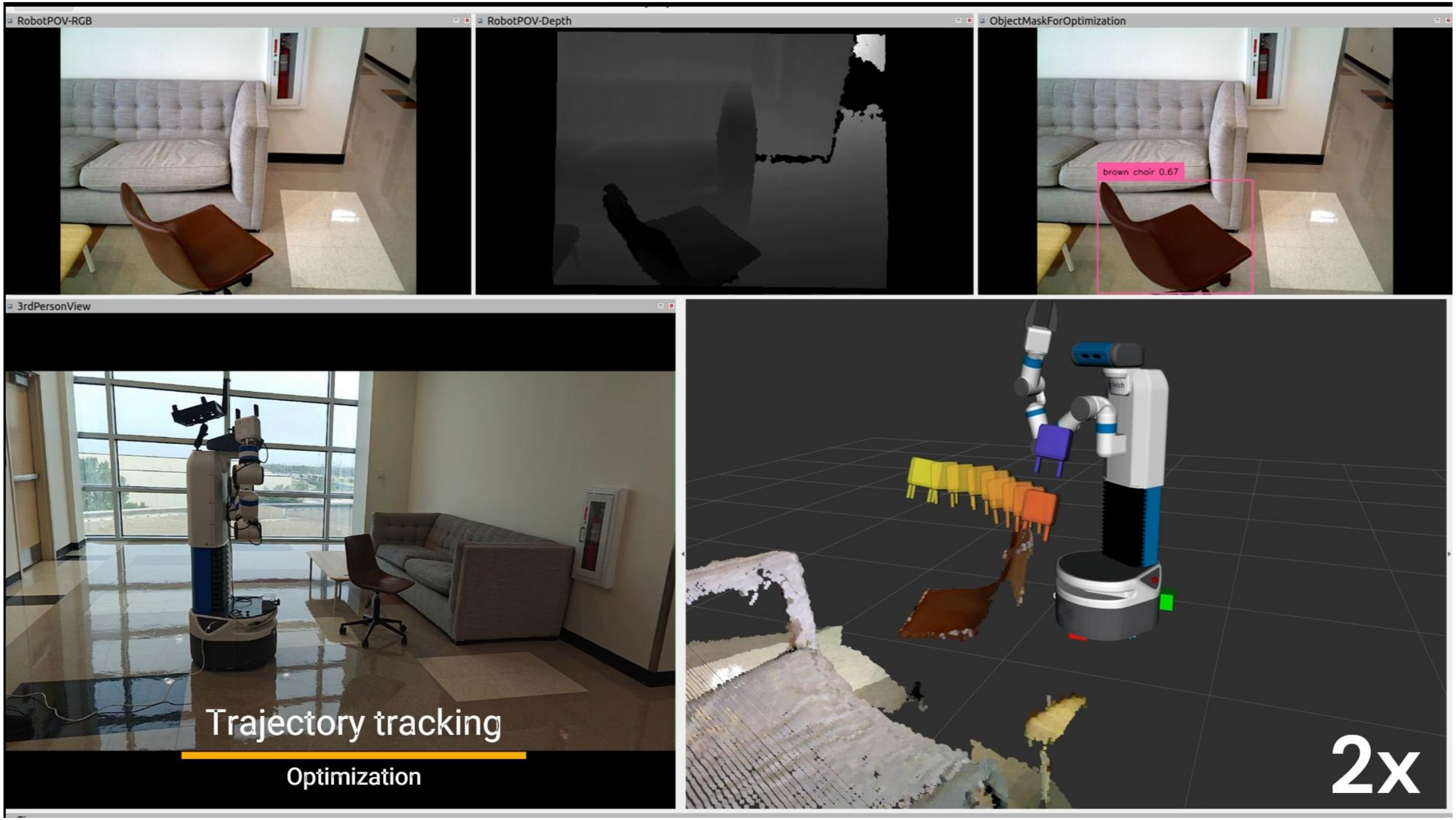
$$\dot{\mathbf{q}}_1 = \mathbf{0}, \dot{\mathbf{q}}_T = \mathbf{0}$$

$$\mathbf{q}_{t+1} = \mathbf{q}_t + \dot{\mathbf{q}}_t dt, t = 1, \ldots, T-1$$

$$\mathbf{q}_l \leq \mathbf{q}_t \leq \mathbf{q}_u, t = 1, \ldots, T$$

$$\dot{\mathbf{q}}_l \leq \dot{\mathbf{q}}_t \leq \dot{\mathbf{q}}_u, t = 1, \ldots, T$$

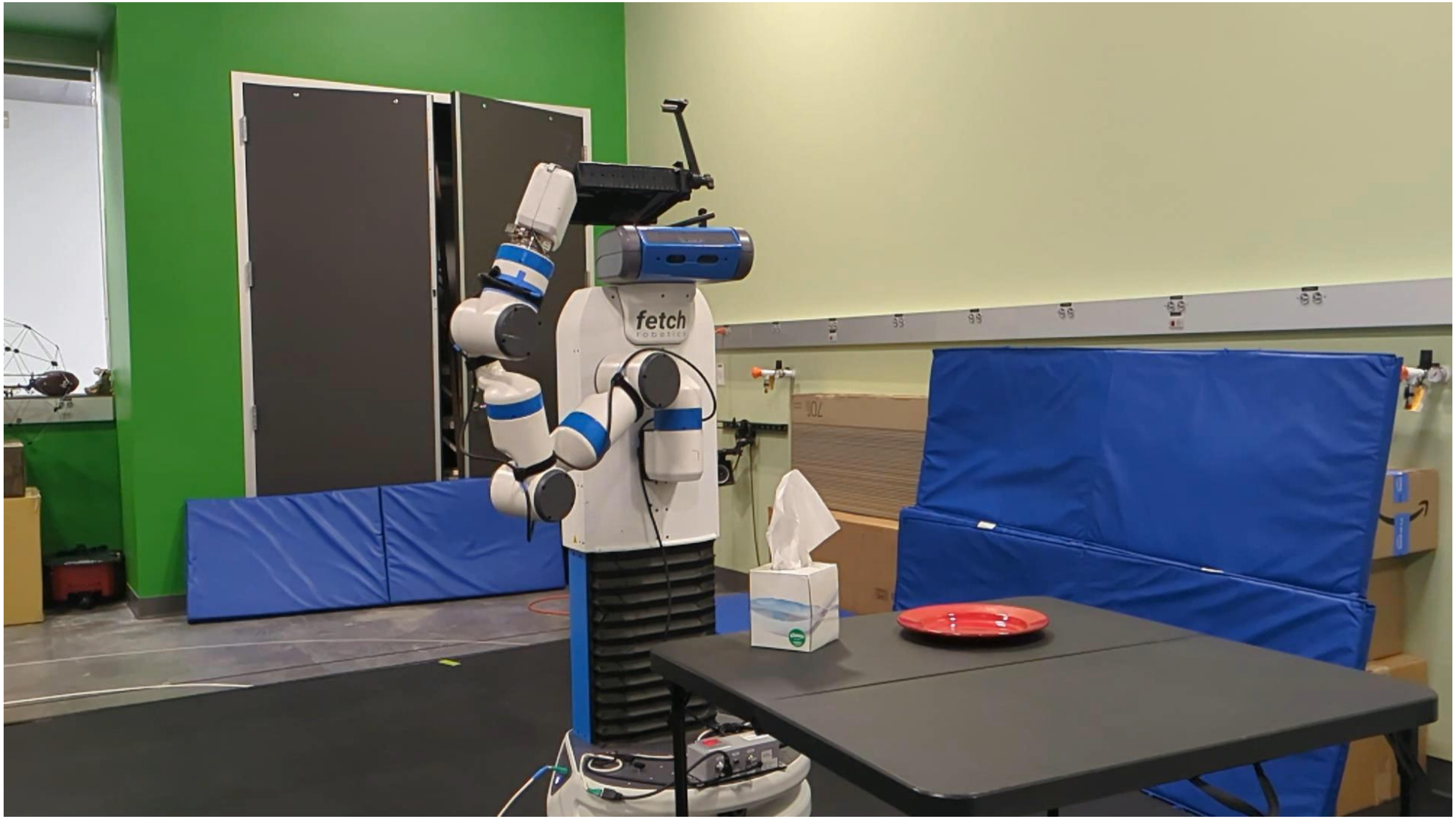
Optimizing the Robot Trajectory



Trajectory Optimization to Follow the Reference

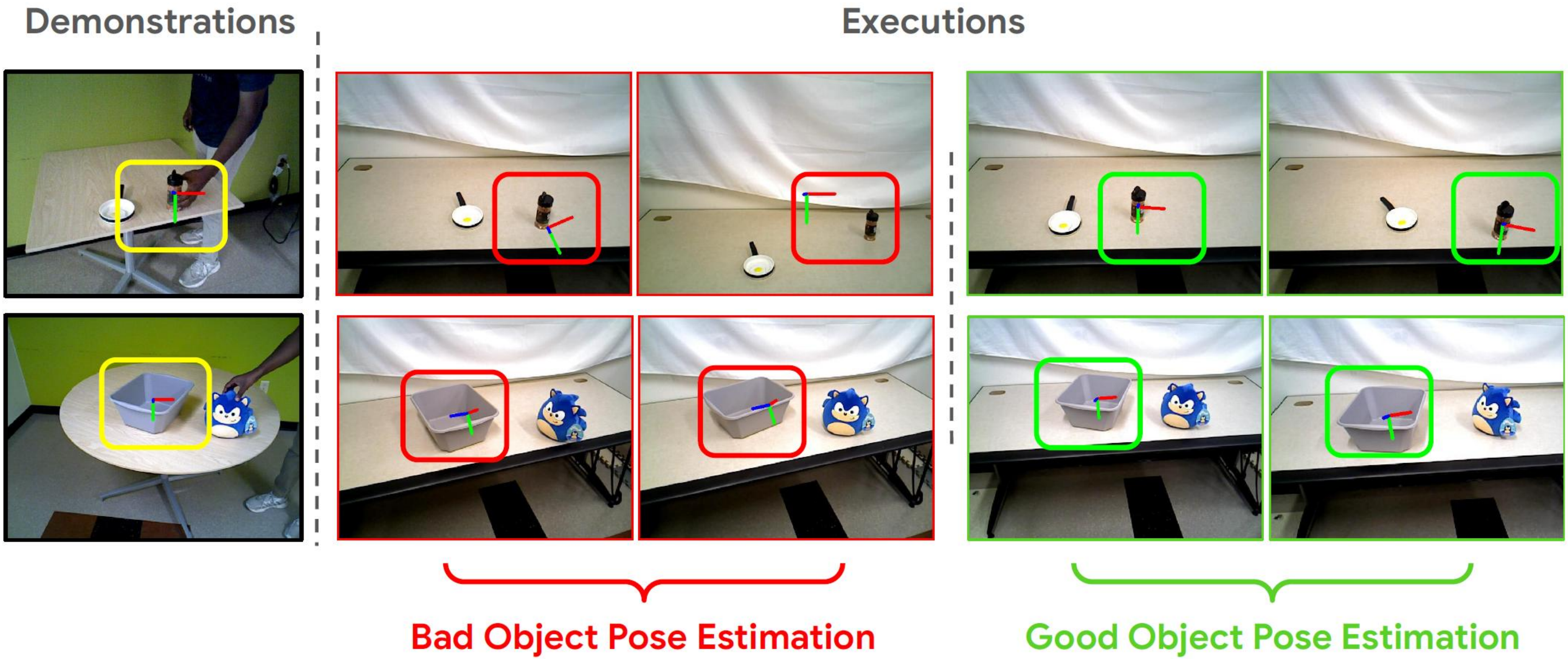
Trajectory Optimization to Follow the Reference

Trajectory Optimization to Follow the Reference



Failure Example

Object Pose Verification



Quantitative Evaluation

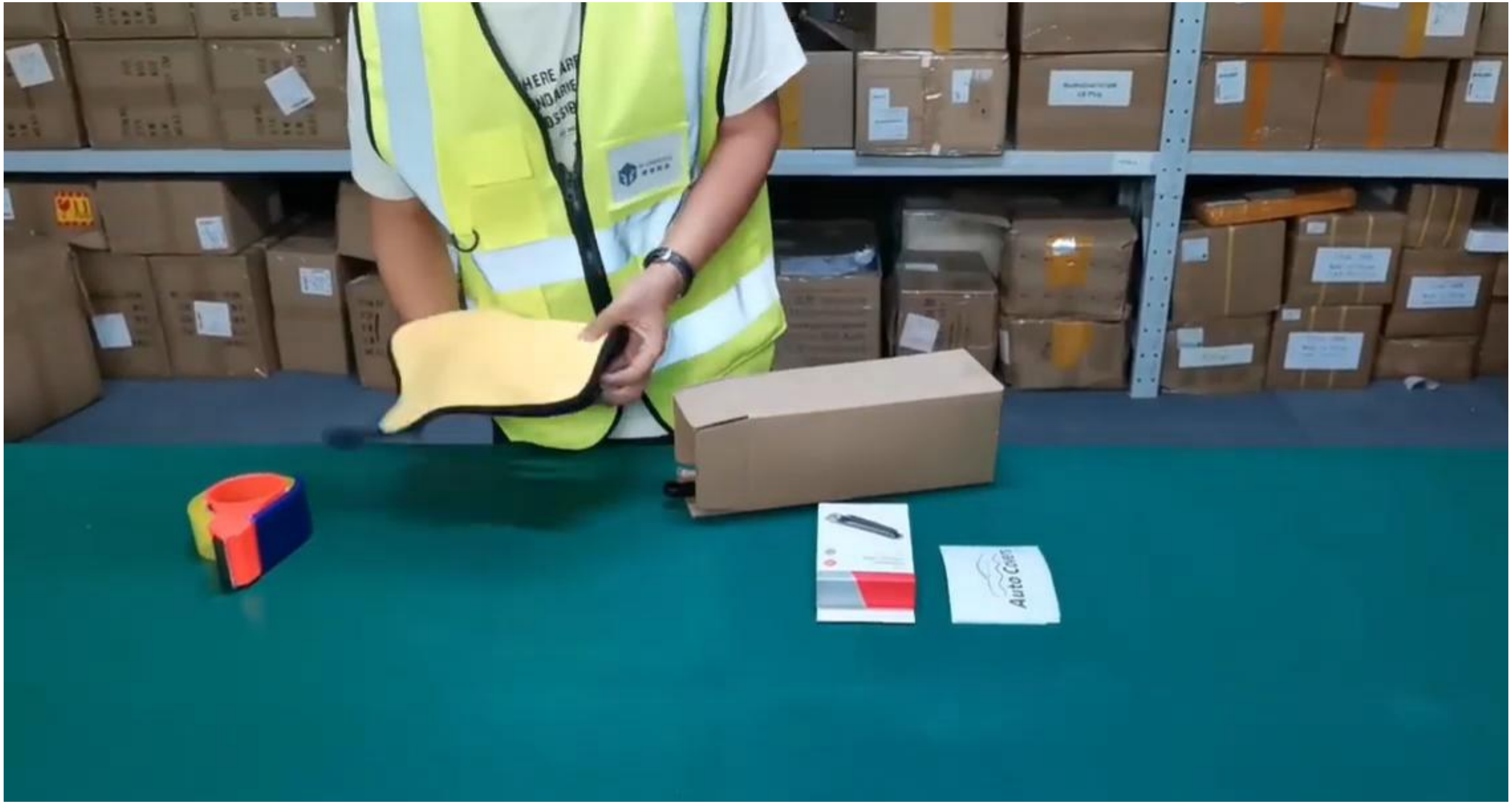
Skill	Grasp success		Task completion	
	Ours	DITTO [16]	Ours	DITTO [16]
Single object				
Move the chair	3	3	3	0
Close fire extinguisher door	3	0	3	0
Dual object				
Put toy in the bin	3	2	3	1
Put bread in the toaster	3	1	3	1
Put seasoning on the omelette	3	3	3	2
Put Lays on the red plate	2	2	2	1
Clean plate with brush	3	1	3	0
Clean plate with tissue	3	0	3	0
Clean plate with kitchen towel	3	2	3	1
Remove cap from wall hook	3	3	3	1
Hang cap onto wall hook	3	0	2	0
Take out sugar box from shelf	3	1	3	0
Rearrange sugar box in the shelf	3	2	2	0
Place bottle in the shelf	3	3	3	0
Close jar with a lid	3	2	2	0
Displace cracker box	3	3	3	3
Total	**47/48**	**28/48**	**44/48**	**10/48**

DITTO [16]:Trajectory Transfer, Heppert et al. University of Freiburg, IROS 2024

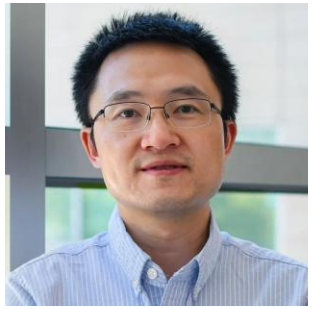
Challenges and Opportunities on Learning from Human Videos

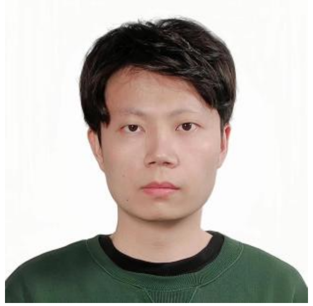
- Understanding of human manipulation from videos is still challenging
  - 3D understanding
  - Deformable, articulated objects
  - Long-horizon tasks
- Trajectory transfer & optimization is slow
  - Better & faster optimization tools
  - Policy learning, e.g., using data from trajectory optimization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

Robot Manipulation is still an Open Challenge



Intelligent Robotics and Vision Lab (IRVL)
















NSF DARPA XPENG SONY NVIDIA FRONTIER MINING

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

Assisted by
Ms. Rhonda Walls

Thank you!