Hand-held Data Collection for Learning of Manipulation Tasks

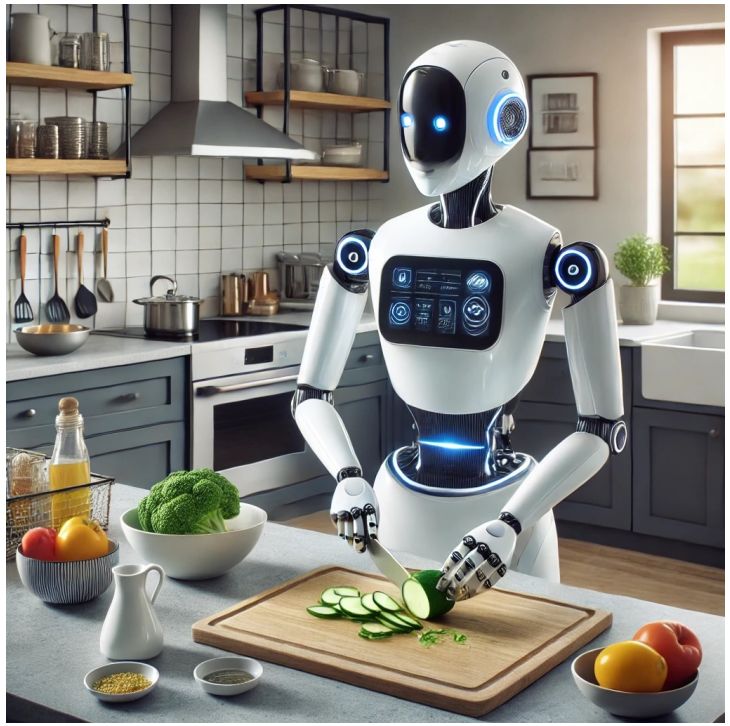
Jannik Jorge Grothusen

Betreuer: - Prof. Dr.-Ing. Robert Seifried (TUHH)
- Dr.-Ing. Daniel-André Dücker (TUM)

Munich Institute for Robotics and Machine Intelligence
Technical University of Munich

Institute of Mechanics and Ocean Engineering
Hamburg University of Technology

Motivation


The real world is much more complicated than the abstractions we can impose on it.

Can we solve this problem without manually understanding and programming behavior?

Why is it that we can build machines that **win at chess and go**, that **paint paintings**, and that can **explain jokes**, but we still don‘t have **robot housekeepers?**

Expert demonstrations are difficult to collect in the robotic domain.

Reason 1 - Embodiment Gap:

the fundamental difference between human and robot capabilities makes action transfer hard

Reason 2 - High Cost and Expertise Requirements:

SOTA teleoperation is resource-intensive and require specialized knowledge to operate

Problem Statement of this Thesis

Develop a **suitable concept** for an intuitive and effective **data collection interface** that
(1) reduces the embodiment gap, (2) captures transferable data,
and (3) supports the learning of diverse manipulation skills for robots.

Formalizing Manipulation Learning

Partially-Observable Markov Decision Processes (POMDP) models the relationship between an agent and its environment:

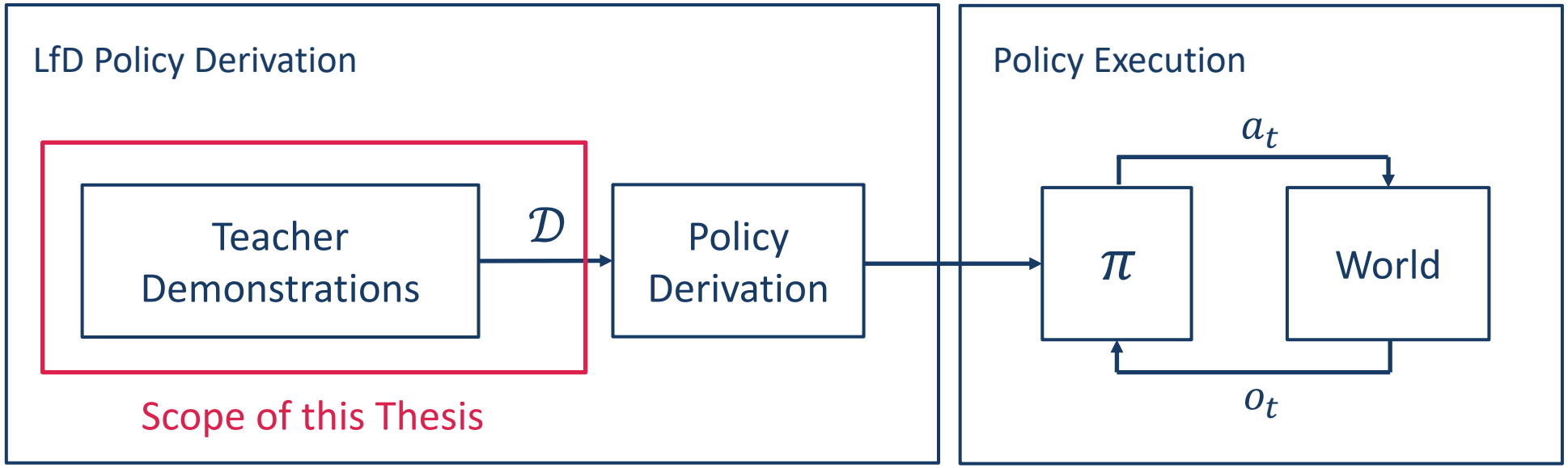
$$(S, A, T, R, \Omega, O, \gamma)$$

The goal of learning is to find a **deterministic control policy**:

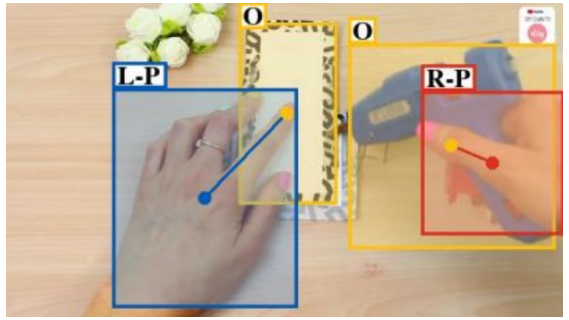
$$\pi : O \rightarrow A$$

(development by hand is challenging)

Learning from Demonstrations (LfD) provides the learner with examples $d_i = (o_i, a_i) \in \mathcal{D}$



Review on Data Collection Methods

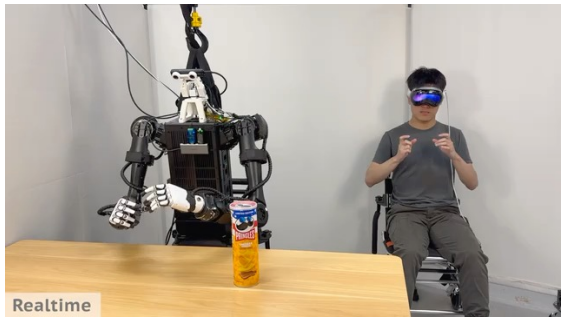


Shan et. al., 2020

Human Video Data

+ inexpensive
+ scalable
+ diverse by nature

- lack explicit action info
- large embodiment gap

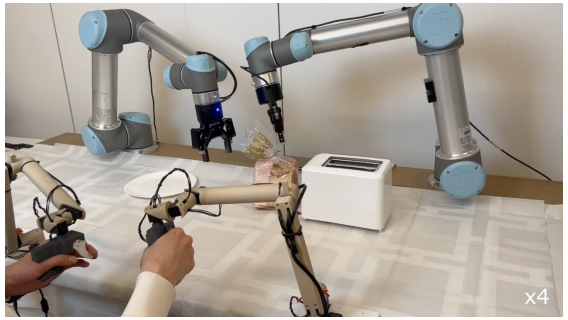


Cheng et. al., 2024 (a)

Shadowing

+ able to capture high complexity (fingers etc.)

- explicit actions are hard to derive
- often un-intuitive use due to high latency



Wu et. al., 2023

Teleoperation

+ direct action transfer (identity mapping)
+ intuitive use

- often expensive
- relies on robot for demonstrations (limited portability)

Cheng et. al., 2024 (b)

Hand-held Tools

+ direct action-transfer in end-effector pose
+ portable
+ easy to use

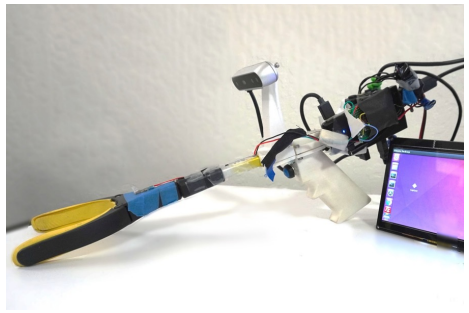
- able to record hardware-infeasible actions
- action recovery with SfM is challenging

Discussion of Data Collection Methods

	Scalability	Transferability	Complexity	Cost
Human Video	++	--	+	++
Shadowing	-	+	-	o
Teleoperation	-	++	o	-
Hand-held Tools	+	+	+	+

Chosen for further investigation in this thesis.

Discussion of Hand-held Devices



Song et. al., 2020

Shafiullah et. al., 2023

Cheng et. al., 2024

Universal Manipulation Interface (UMI)

Data Collection

Policy Rollout (75% Success)

UMI Usage:

1. GoPro Preparation
2. Timecode Sync
3. Mapping Video
4. Gripper Calibration
5. Data Collection

+ accessible (only 3D-prints, GoPro)
+ captures visual, action, and embodiment diversity

- limited sensing (not real-time, no depth)
- suboptimal workflow (bi-manual requires 2 operators)
- specific to one end-effector

Concept: A Hand-held Demonstration Interface

Overarching goals for data collection:

(1) acquire transferable demonstrations (o_i, a_i) from human motion

(2) be portable and intuitive to use

Goals that are addressing shortcomings of recent systems:

(3) enable real-time data processing and depth capture

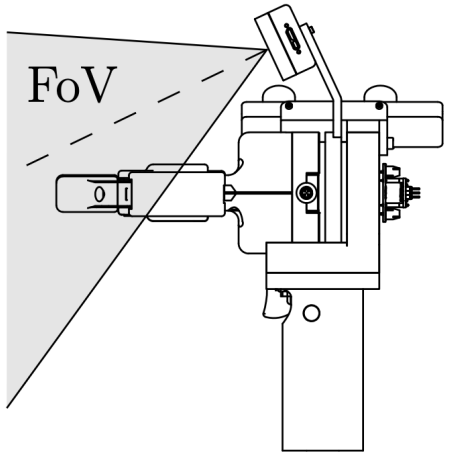
(4) improve workflow control for efficiency

(5) enable use with different end-effectors

Observation Module

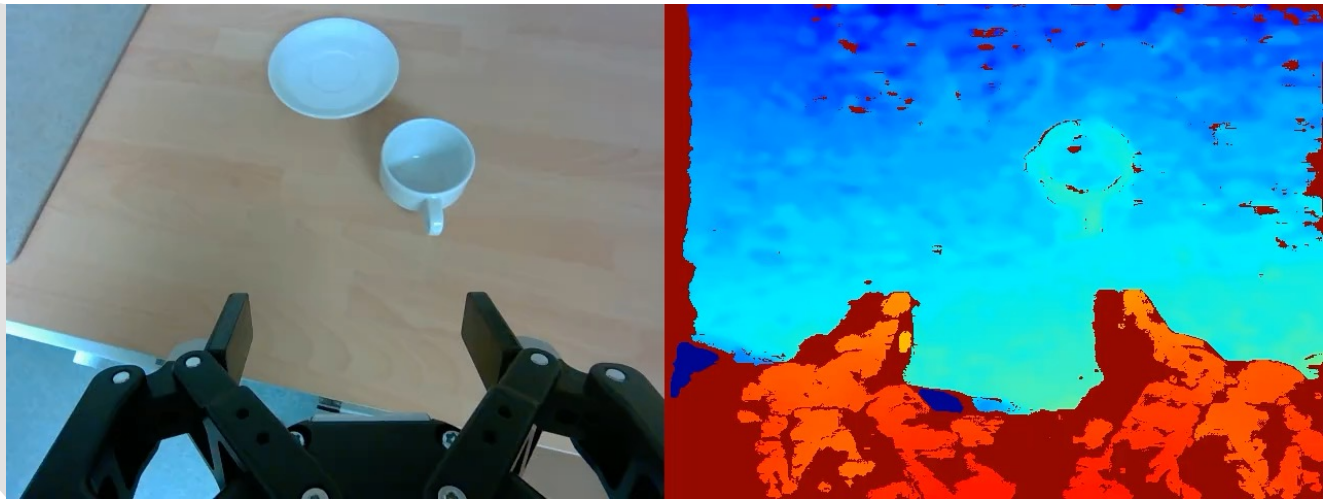
Action Handle

Recording of Transferable Observations



Wrist-mounted Position
minimizes observation embodiment gap

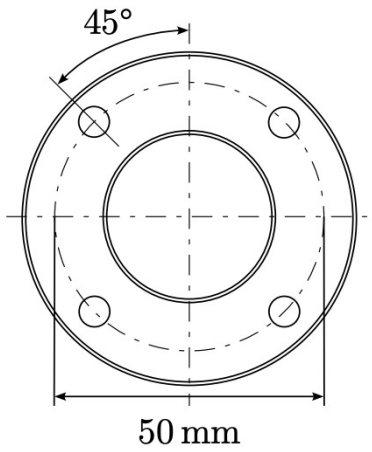
Capturing observations o_i incl. depth in real-time



RGB

Depth

Human-to-Robot Action Mapping

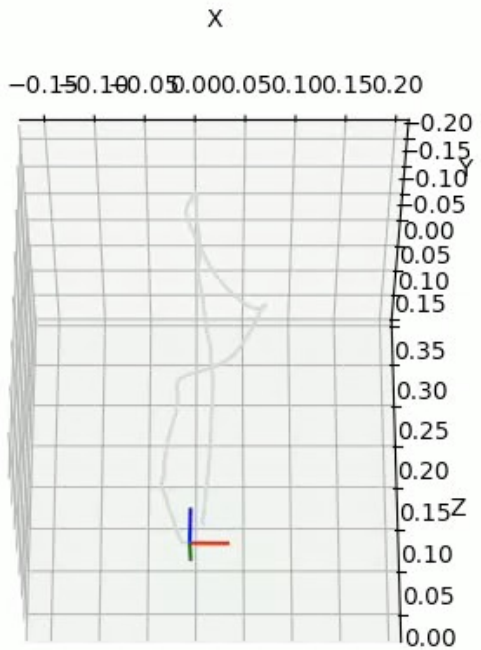


Modular Mounting Interface
ISO-9401-1-50-4-M6

RGB-D Image
+
IMU Data

Visual-Inertial-SLAM

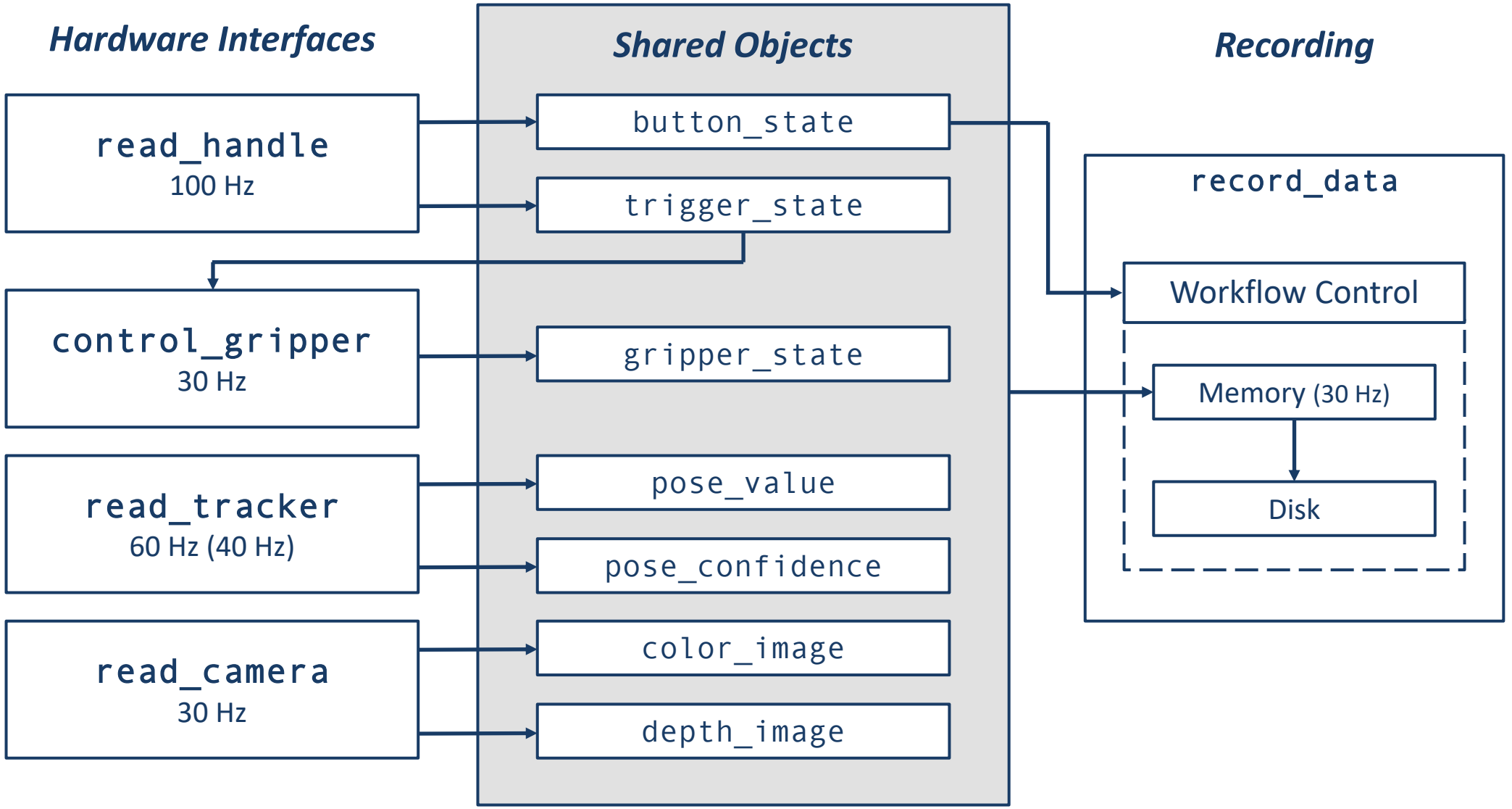
6 DoF End-effector Pose as Action



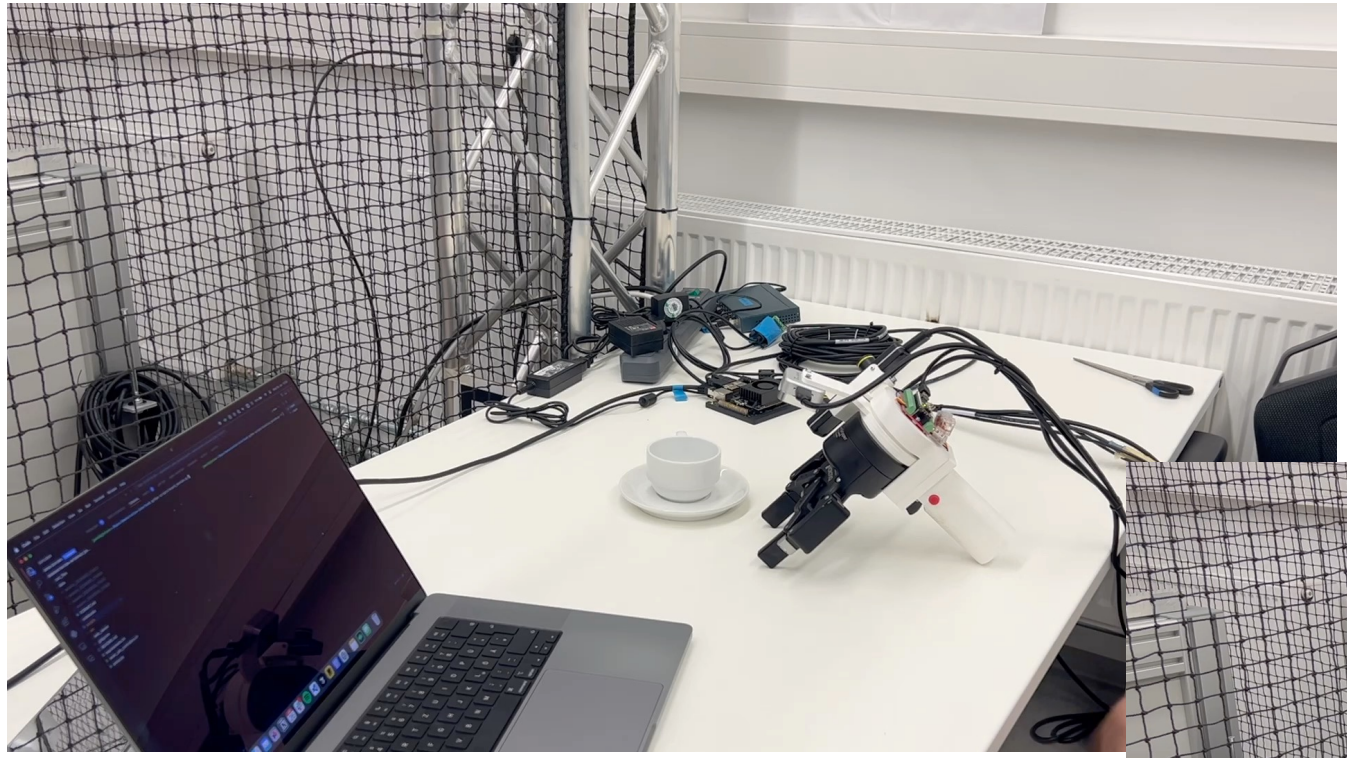
Controlling Gripper State
in normed range [0,100]

Intuitive Workflow Control
can be done with one hand

Dataflow

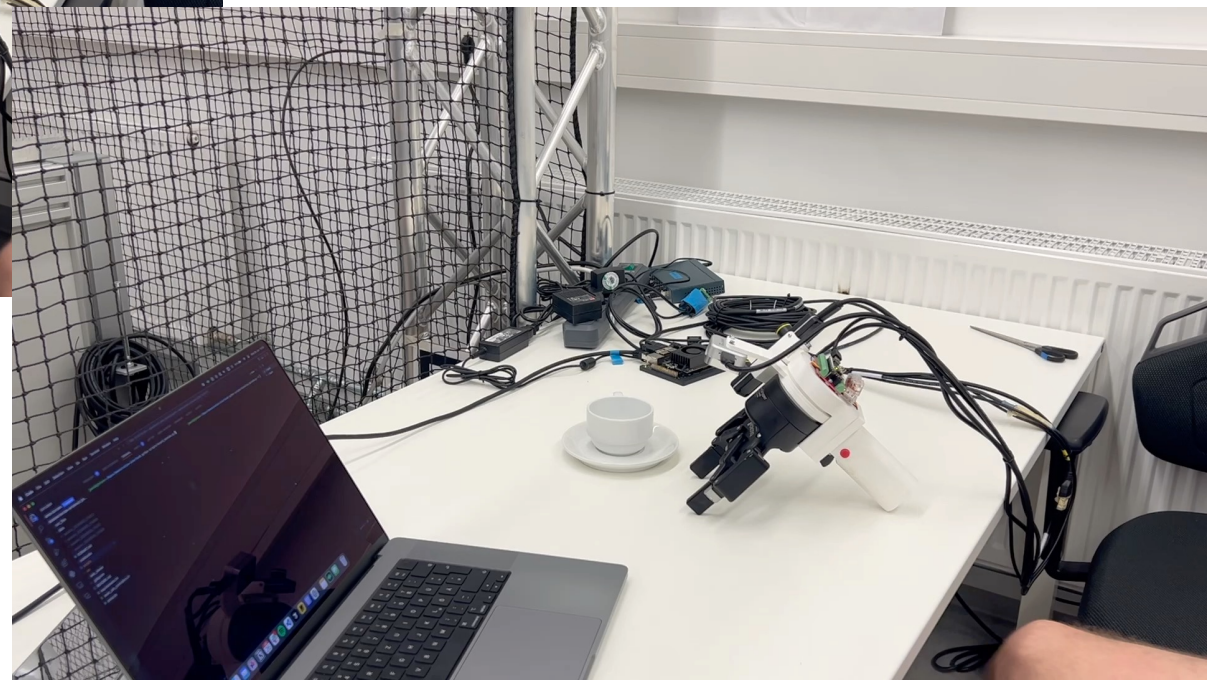

Hardware Interfaces
Shared Objects
Recording
read_handle
100 Hz
control_gripper
30 Hz
read_tracker
60 Hz (40 Hz)
read_camera
30 Hz
button_state
trigger_state
gripper_state
pose_value
pose_confidence
color_image
depth_image
record_data
Workflow Control
Memory (30 Hz)
Disk

Example Workflow (4x Speed)

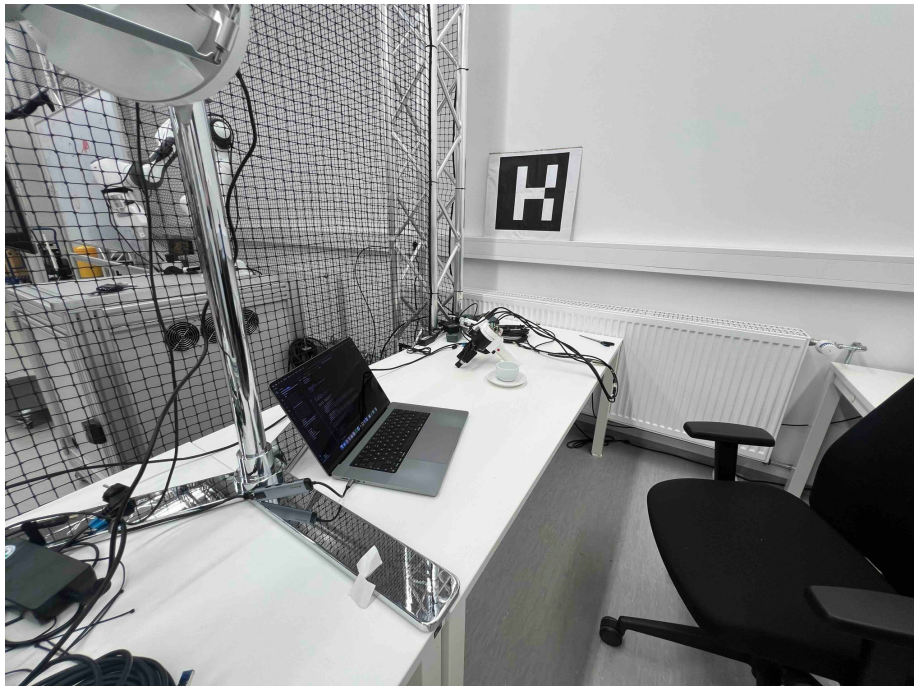


Typical data recording steps:

1. Execute a single Python script
2. Wait for initialization (~10s)
3. Press button to start recording

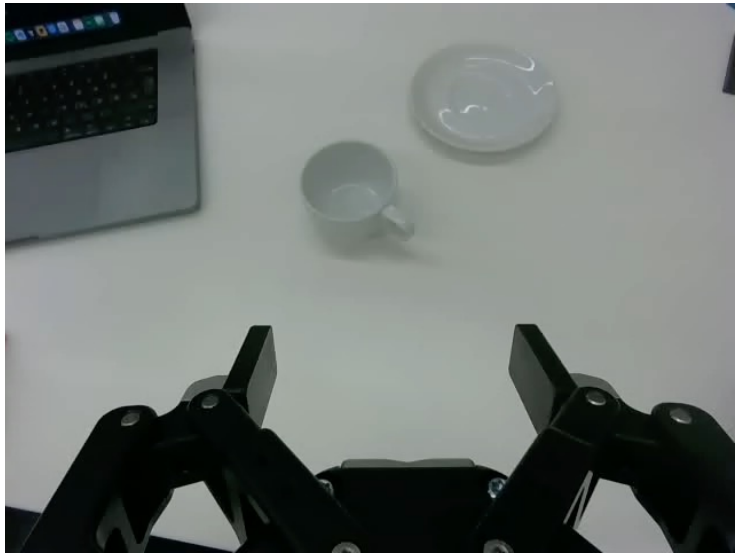


Starting Point: Development

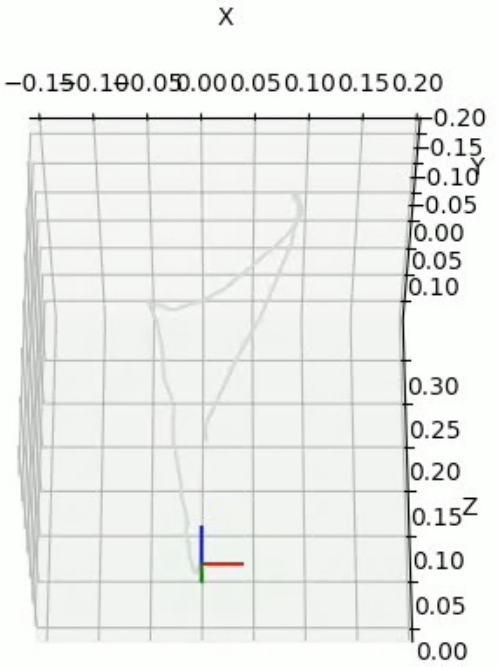


Lab (Small): the development environment

Initial Result:



Observation



Action

Evaluation Strategy

Accuracy of Action Recovery

How well can robot-native actions be recovered from human demonstration?

Metric is **Average Tracking Error (mm)**

Experiment: Compare to Ground Truth Motion using a Motion Capture System

Efficiency of Data Collection

How intuitive and fast can manipulation data be recorded?

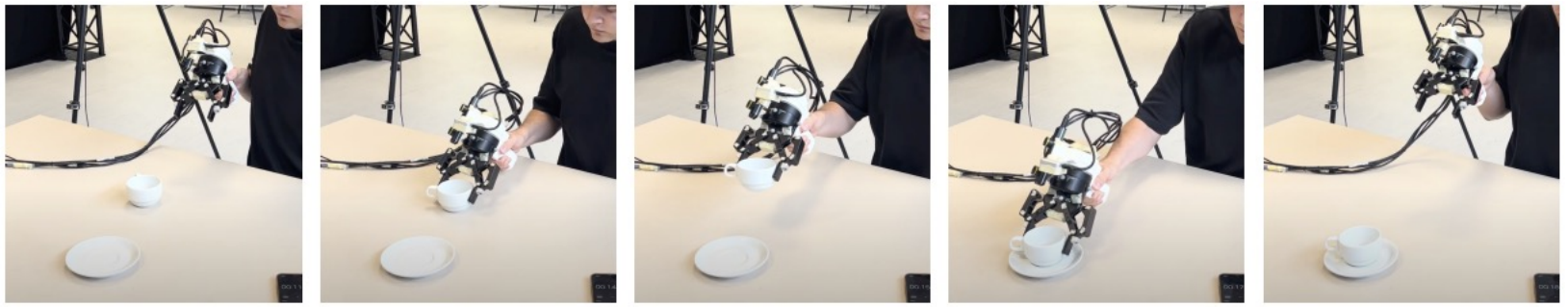
Metric is **Throughput (Demos / Time)**

Experiment:
Quantify Throughput with a User Study

Cup Arrangement

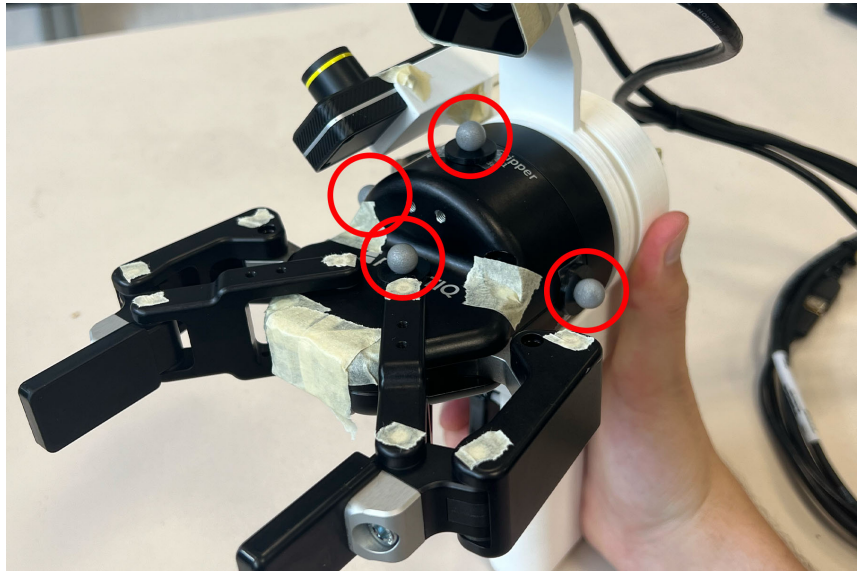
Complex actions:
Pick and place,
Pushing to reorientate

Throughput reported by UMI



Start *Pick* *Move* *Place* *End*

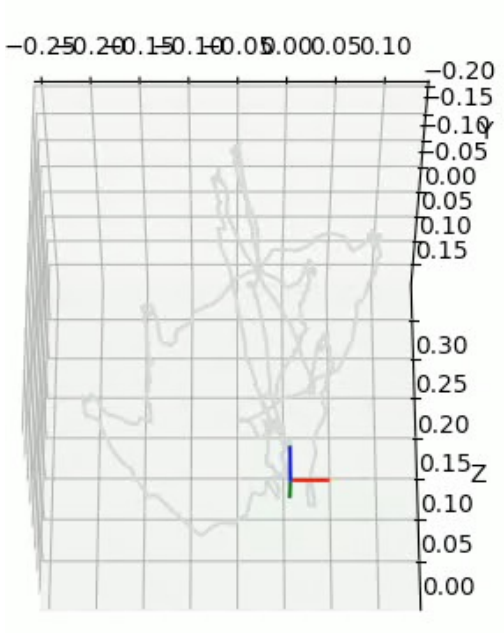
First Accuracy Analysis: Vicon Motion Capture System



Protocol: Do *Cup Arrangement* 24 times with random initial states.

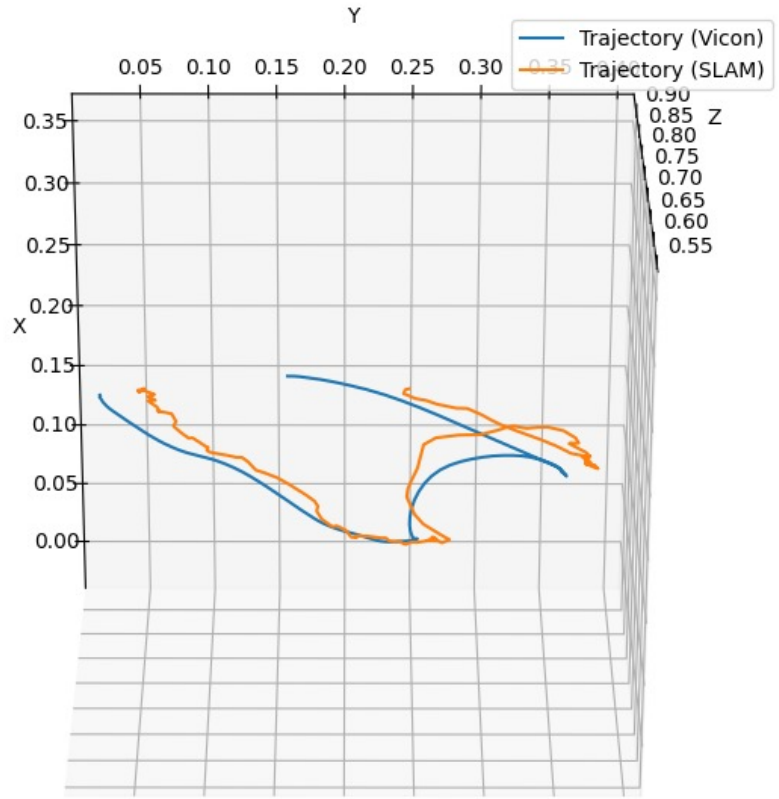
First Accuracy Analysis: Results

Example Demonstration Data



(Twitches can be observed)

When compared, the action data looks way better in Lab (Small) than in Lab (Vicon).



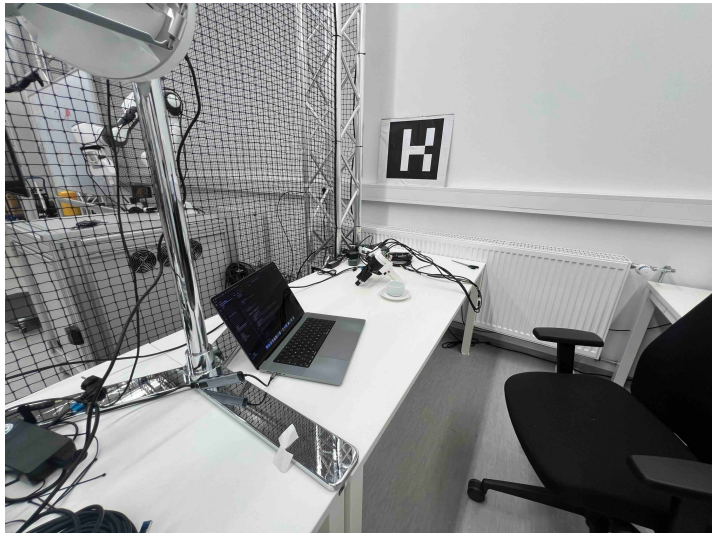
Estimated vs. Real Trajectory (Example)

Mean Tracking Error:

32.83 mm

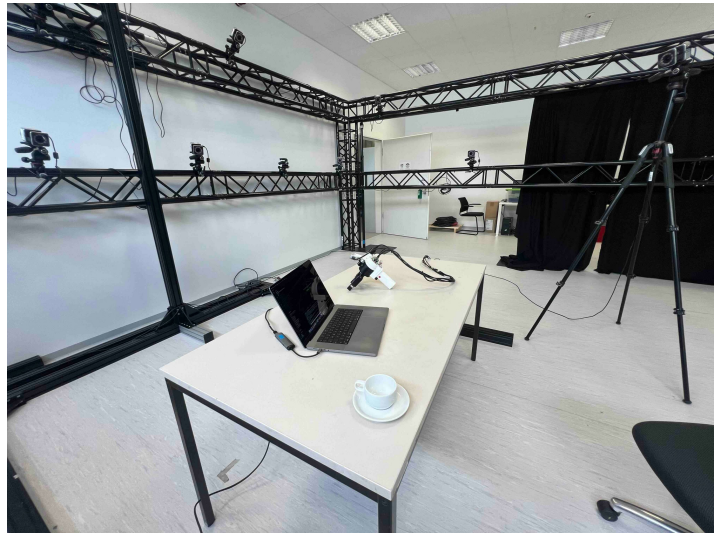
(UMI's error is **6.1 mm**)

Testing in 5 Environments

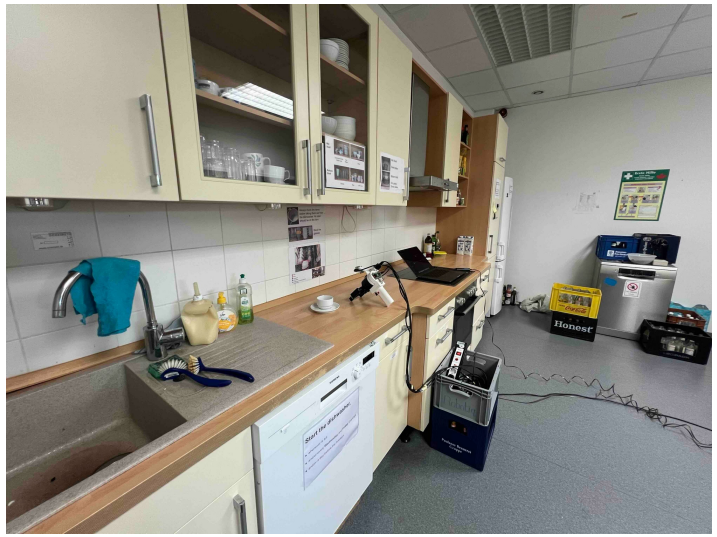


Lab (Small)

Lab (Large)



Lab (Vicon)



Kitchen

Outside

Differences:

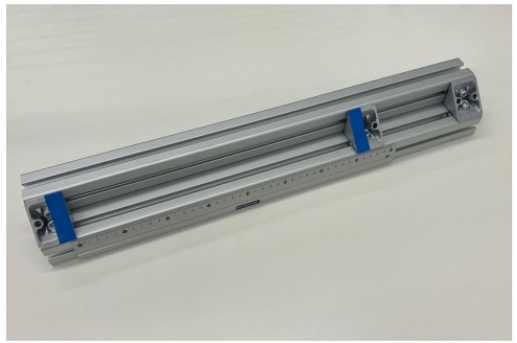
- Lighting,
- Visual features,
- Distance to environment

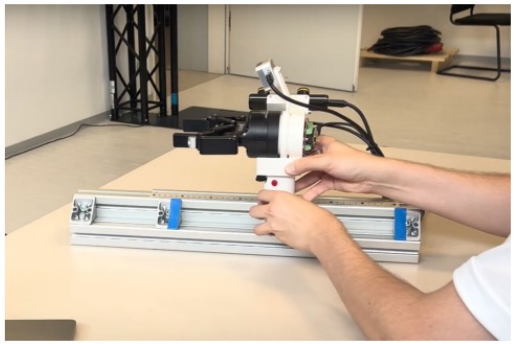
Second Accuracy Analysis: Strategy

How does the environment influence the device's tracking accuracy?

Quantitative Analysis

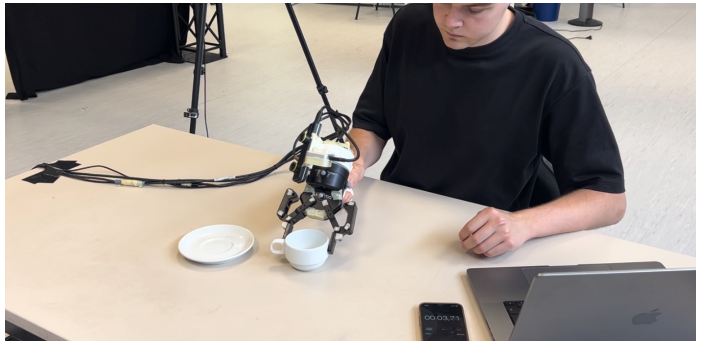
Experiment:
Perform a defined motion in all environments.





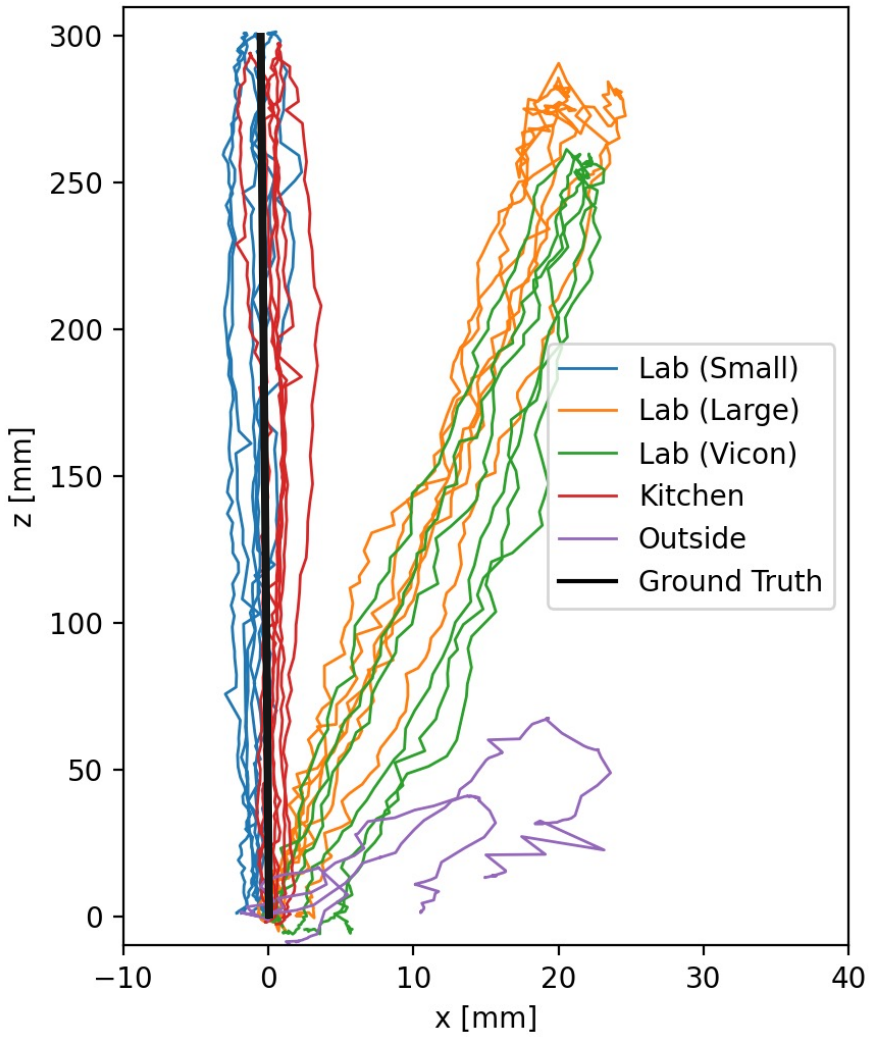
Qualitative Analysis

Experiment: Compare *Cup Arrangement* Trajectories from all environments.



Second Accuracy Analysis (Quantitative): Results

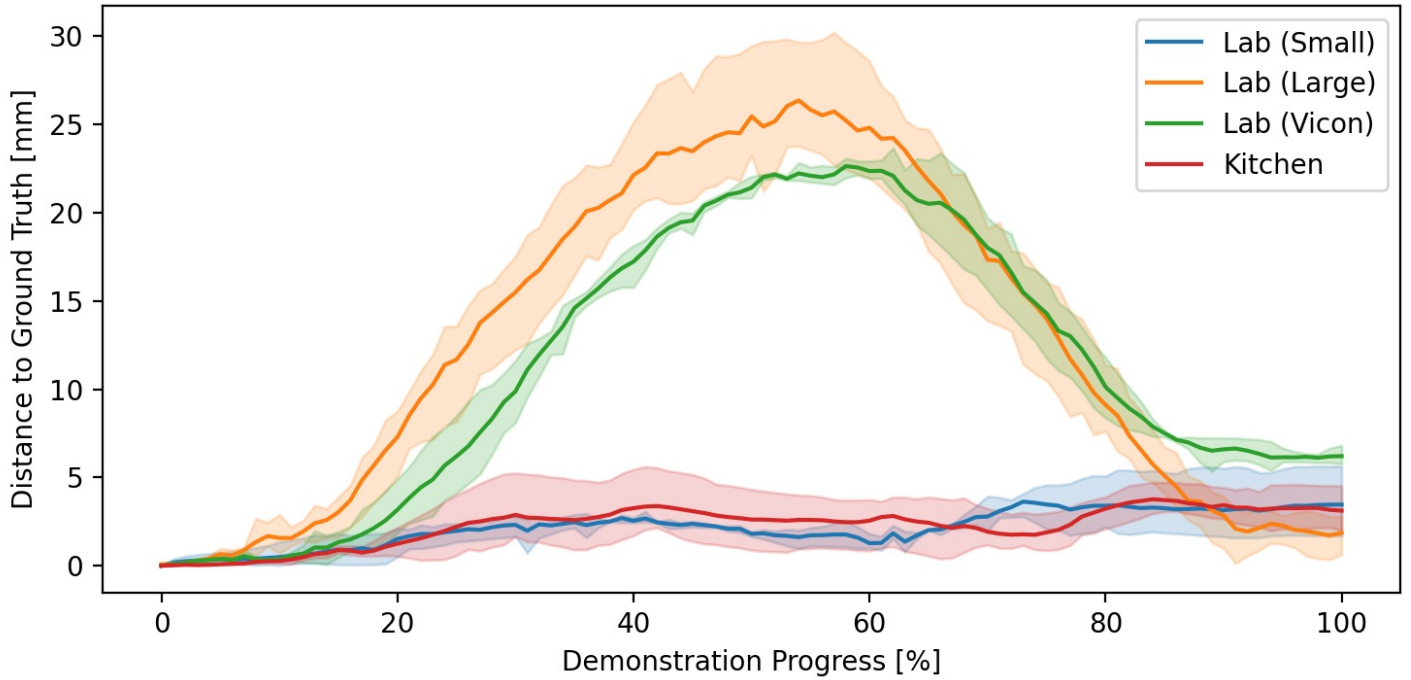
Paths in the x-y-Plane



Protocol:

- place the guide-rail in the workspace
- move the device in a straight line 300 mm back and forth
- repeat three times

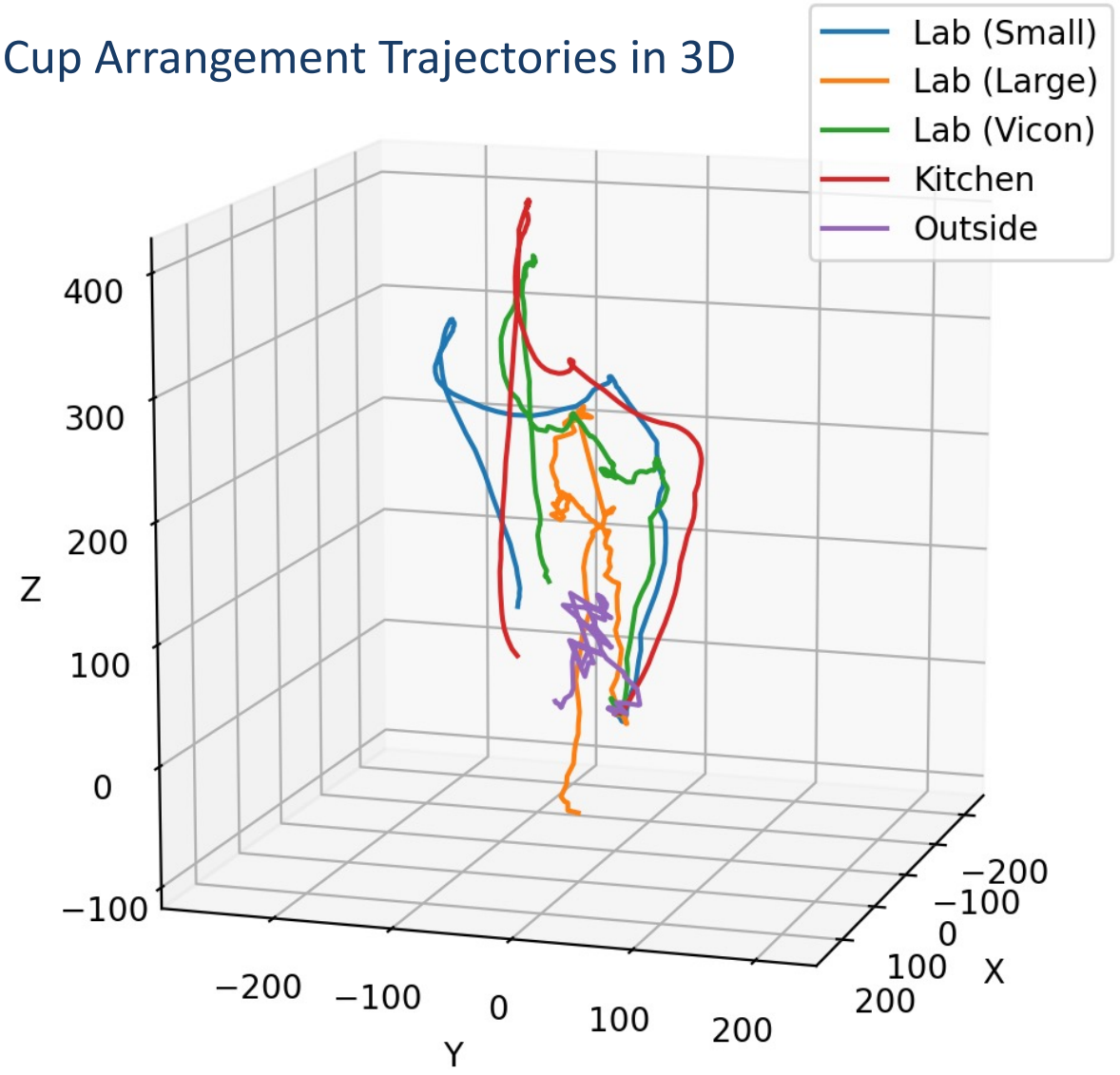
Point-to-Curve Euclidean Distance



Tracking error of **<5 mm** for Lab (Small) and Kitchen!

Second Accuracy Analysis (Qualitative): Results

Cup Arrangement Trajectories in 3D



Observations:

- *Lab (Small)* and *Kitchen* show smooth trajectories
- *Lab (Large)* and *Lab (Vicon)* show twitches
- *Outside* does not represent reasonable motion

Conclusion:

A strong correlation between *distance to environment* and data quality can be observed.

Efficiency Analysis: User Study

Person A - Lab (Vicon)
13 Demos / 3 min

Person B - Lab (Vicon)
16 Demos / 3 min

Person C - Kitchen
11 Demos / 3 min

Baseline Recording
34 Demos / 3 min with Human Hand

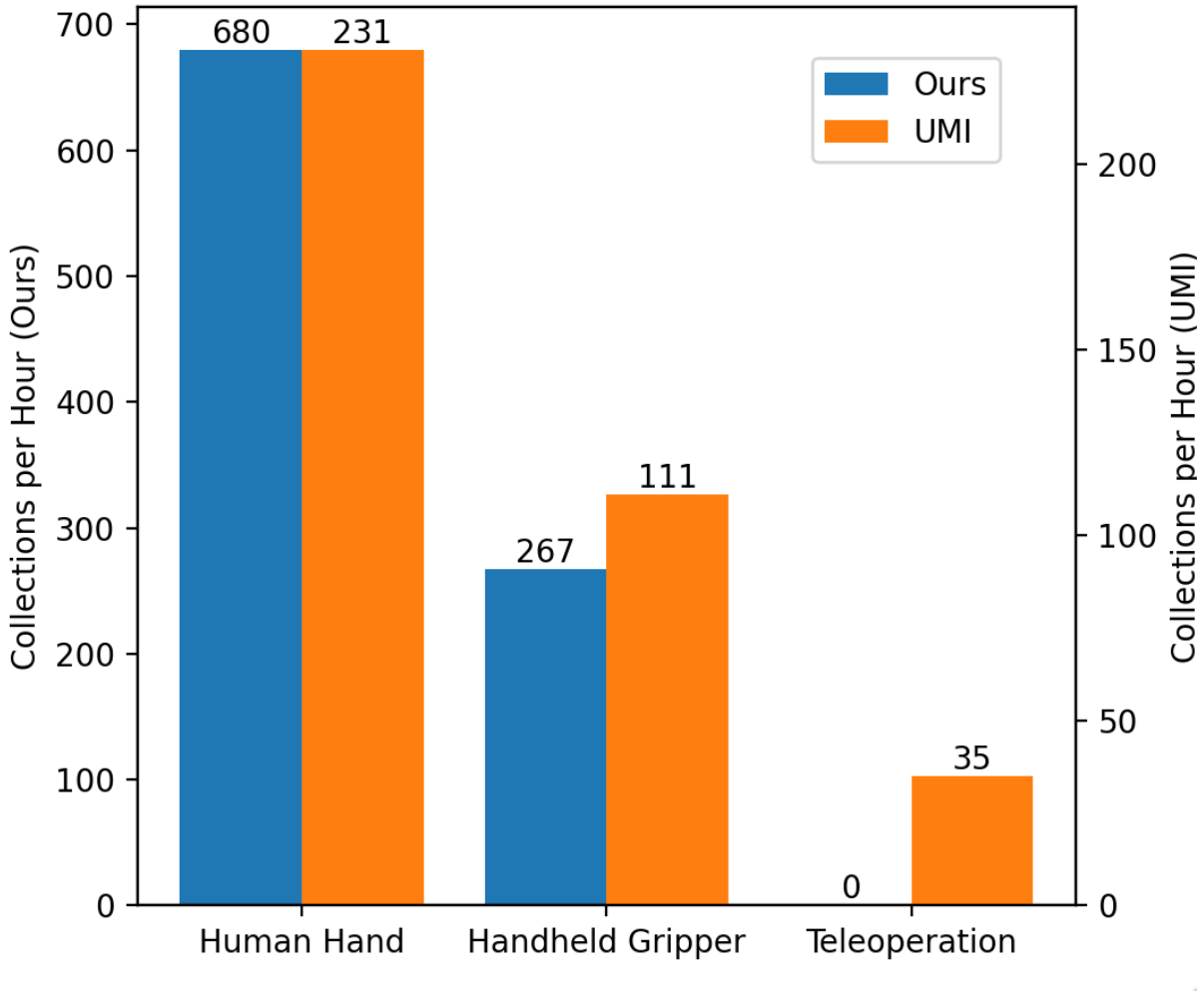
Protocol:
- Orientation Session (1 min)
- Practice Phase (2 min)
- Execution Phase (3 min)

Cup Arrangement & Re-orientation

Average Throughput:
276 Collections/Hour

(UMI's throughput is **111 C/h**)

Efficiency Analysis: Results and Comparison



Notes:

- small sample size of participants and number of demonstrations
- task speed can vary drastically

Relative Speed:

39% Speed of the Human Hand

(UMI's reports **49%**)

Conclusion

Summary:

A hand-held demonstration interface was developed, that:
- enables real-time data processing
- optimizes workflow control
- supports the use of different end-effectors

Future Work:

- transferring recorded data to a robot policy
- ablation studies on camera positioning, FoV, depth data usefulness etc.
- improvement of SLAM robustness / use of other localization approaches
- in-depth user studies on ergonomics and data collection efficiency
- real-time data processing & feedback (e.g. simulating kinematic constrains)