Learning Dexterous Manipulation from Human-Object Interaction

Li Yi

Nov, 2024

Self Introduction

Li Yi (弋力)

- 2009 - 2013, B.E. @ Tsinghua University
- 2013 - 2019, Ph.D. @ Stanford University
- 2019 - 2021, Research Scientist @ Google Research
- 2021 - now, Assistant Professor @ Tsinghua University
- Research: 3D Visual Computing and Embodied Perception
- Homepage: https://ericyi.github.io/
- Email: ericyi0124@gmail.com

Embodied AI

Embodied Perception and Interaction

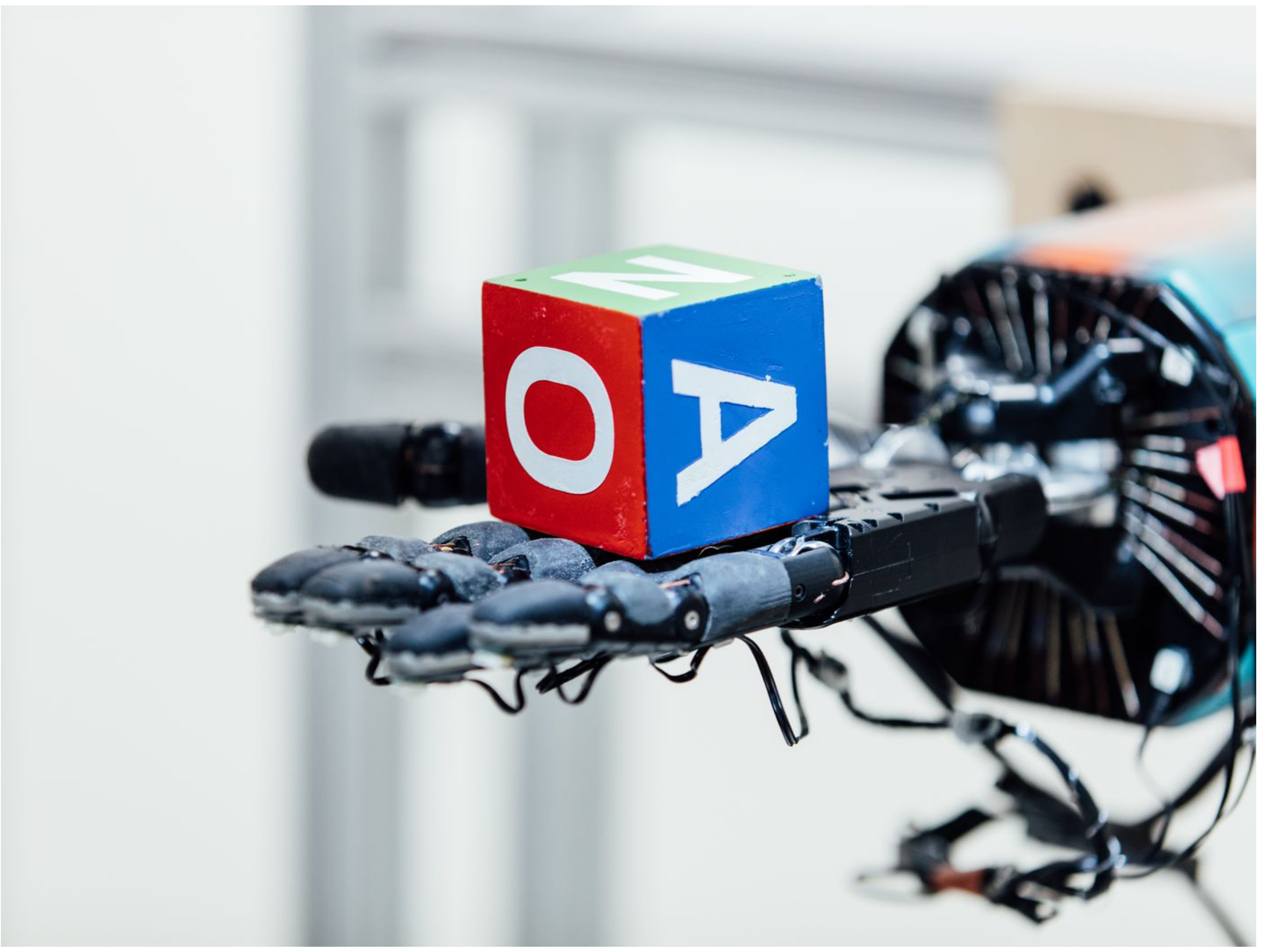
Embodied Agent

image credits: Matterport3D

Environment

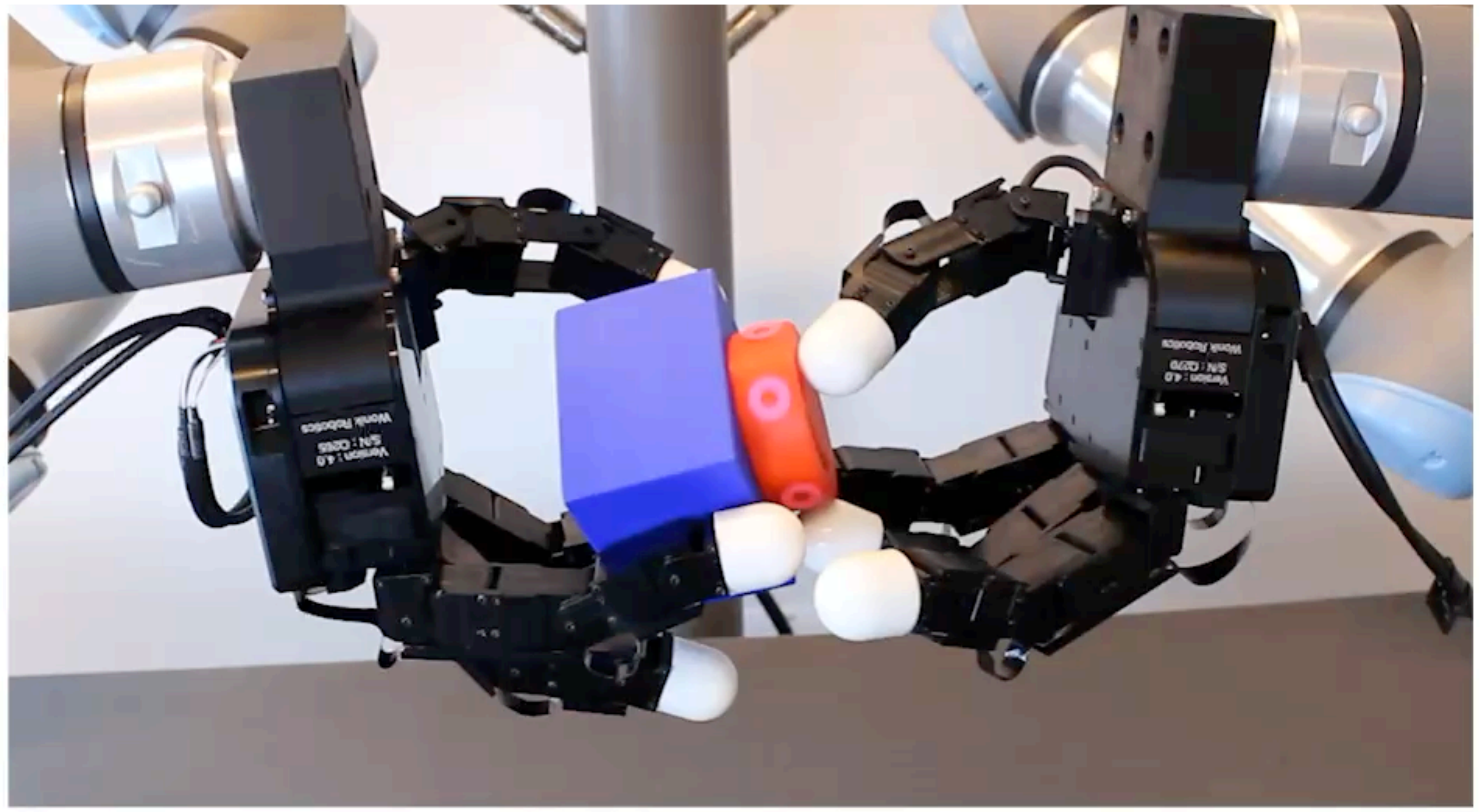
Goal: General-Purpose Dexterous Manipulation

Reality: Specialized Dexterous Manipulation



OpenAI, 2018

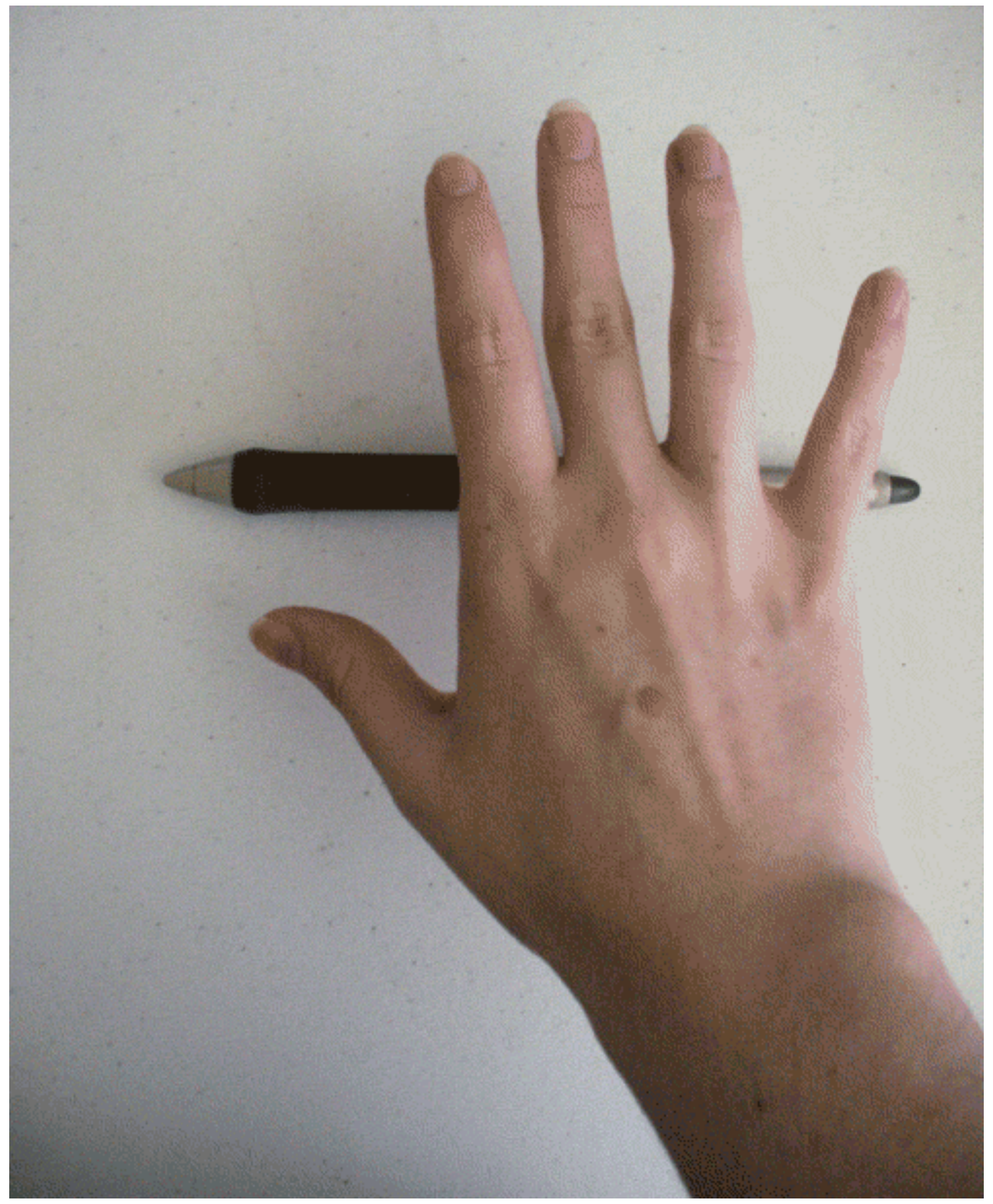
Wang et al., 2024



Lin et al., 2024

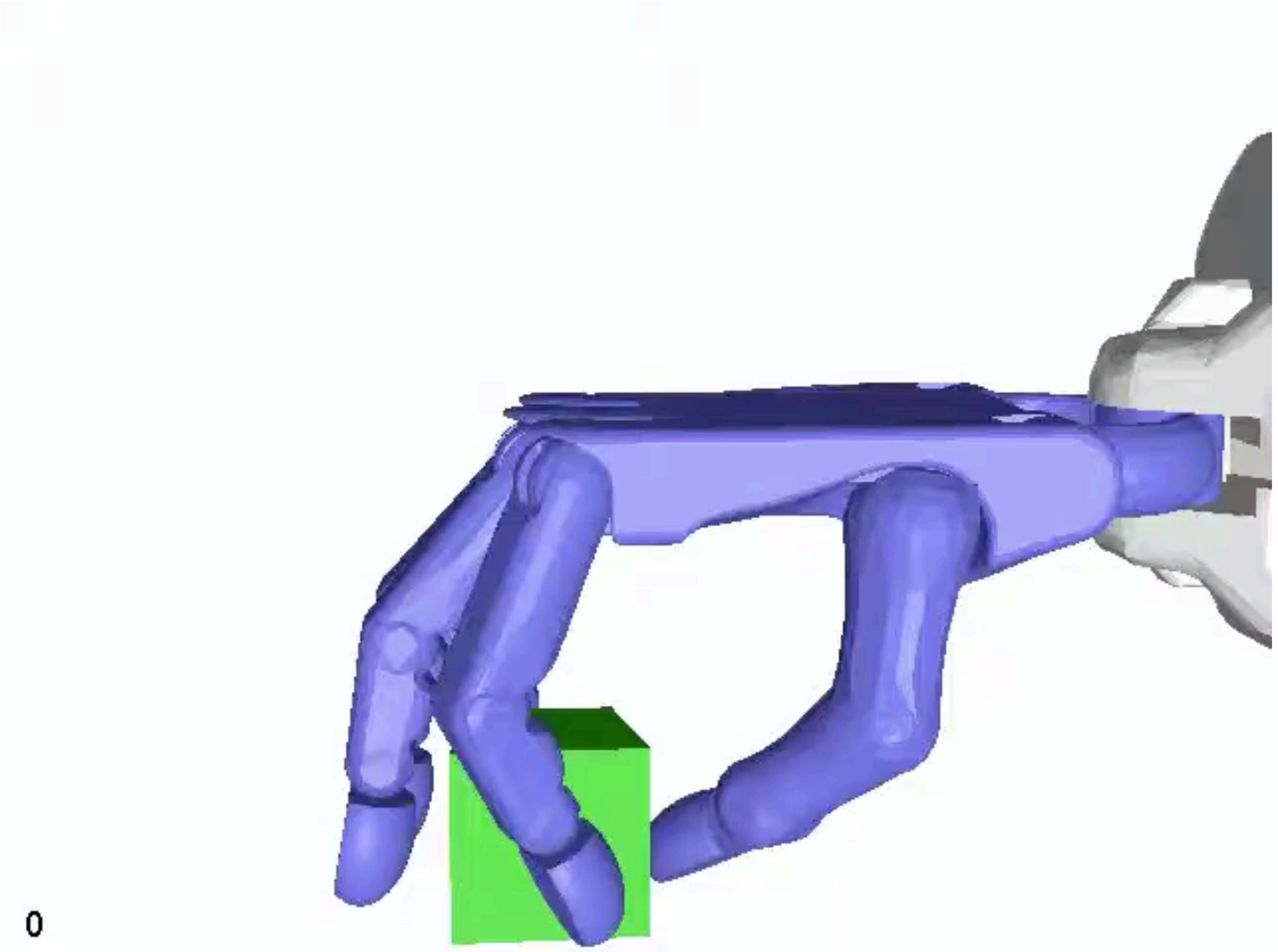
Why is There a Huge Gap?

- Problem complexity:
  - Rich and slipping contacts with complex dynamics
  - Underactuation during in-hand re-orientation or nonprehensile manipulation
  - High dimensional state and action spaces
  - Dynamic perception during heavy occlusion

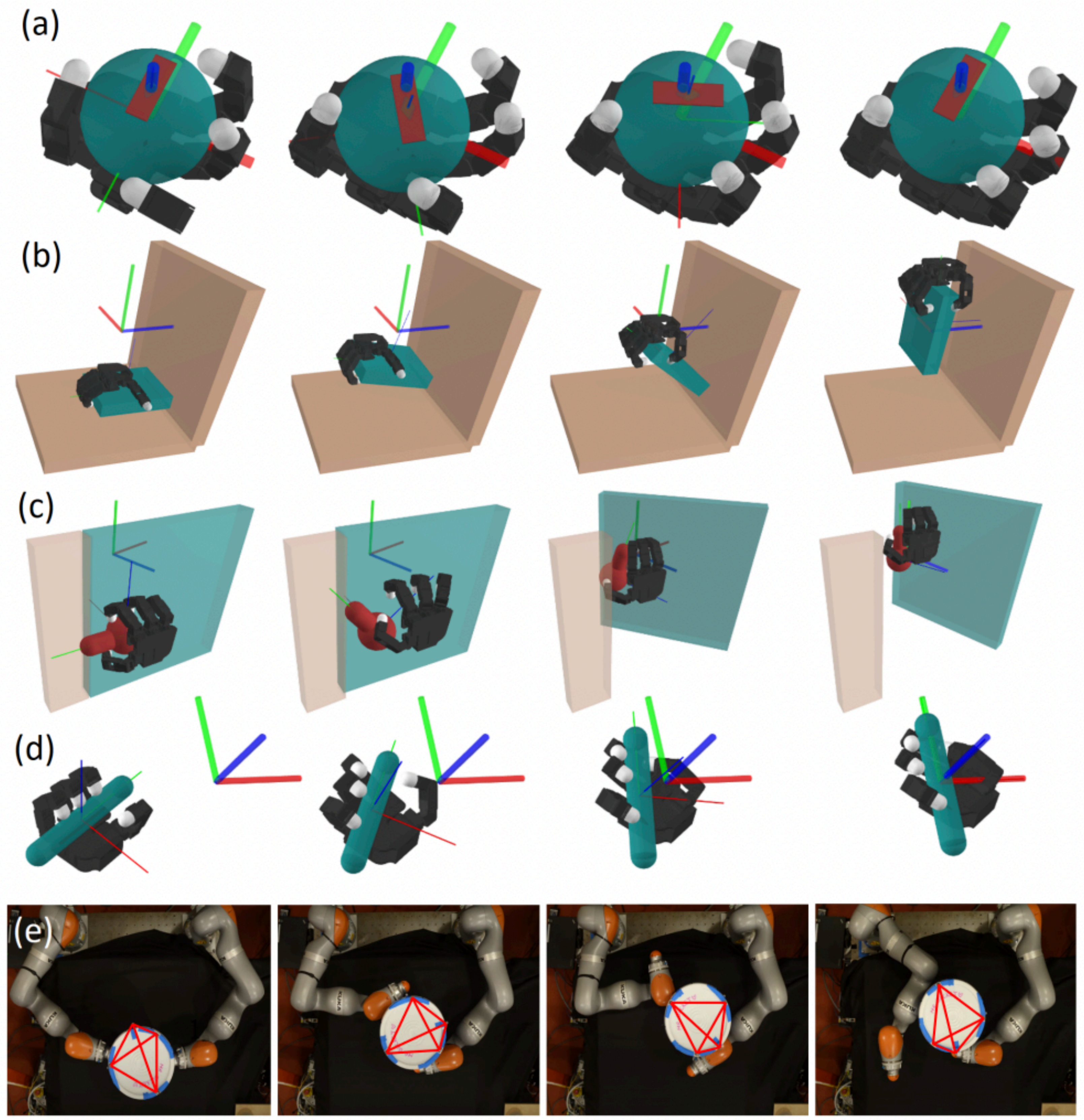


Why is There a Huge Gap?

- Popular paradigms: model-based trajectory optimization



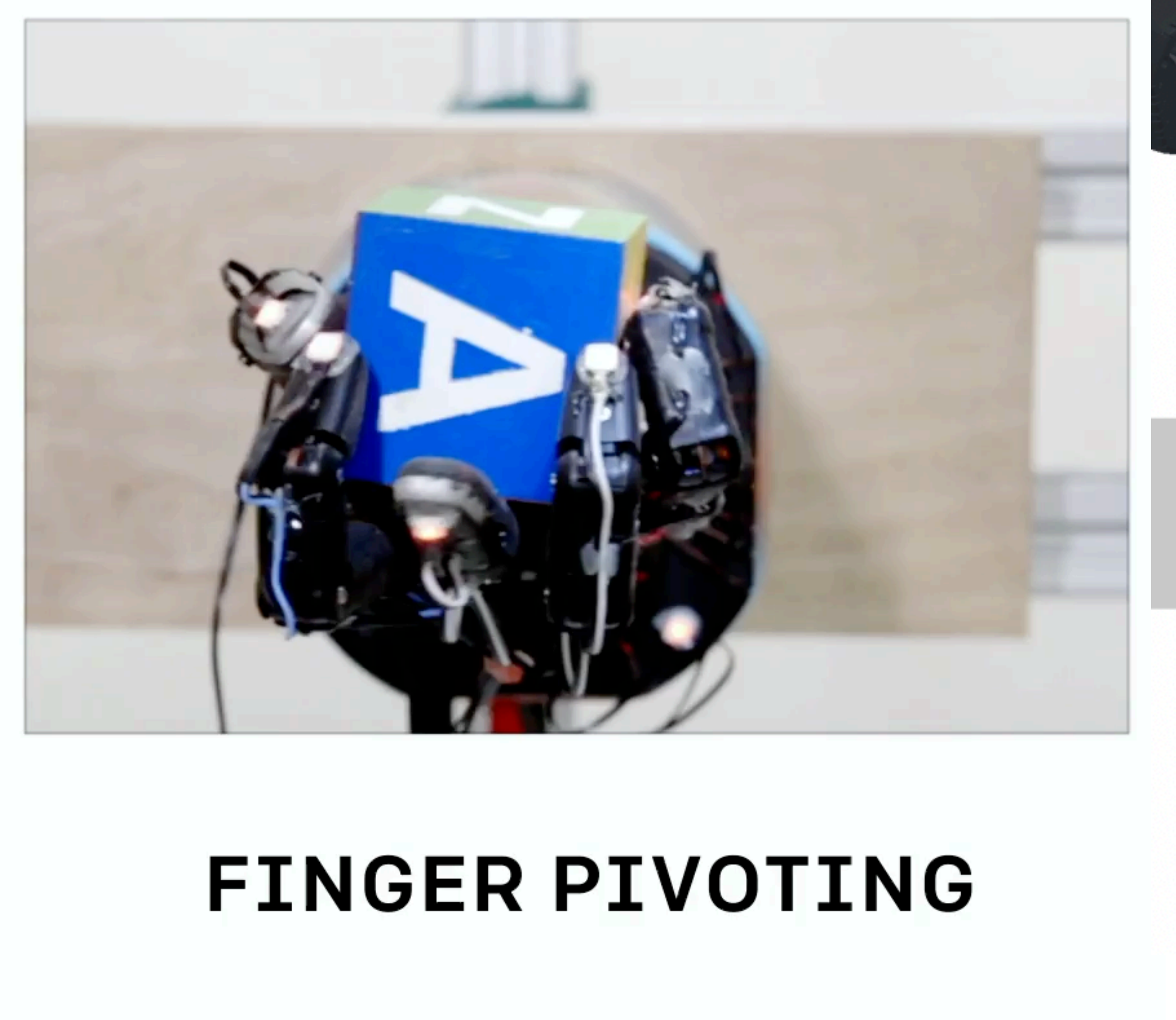
Bai et al., 2014

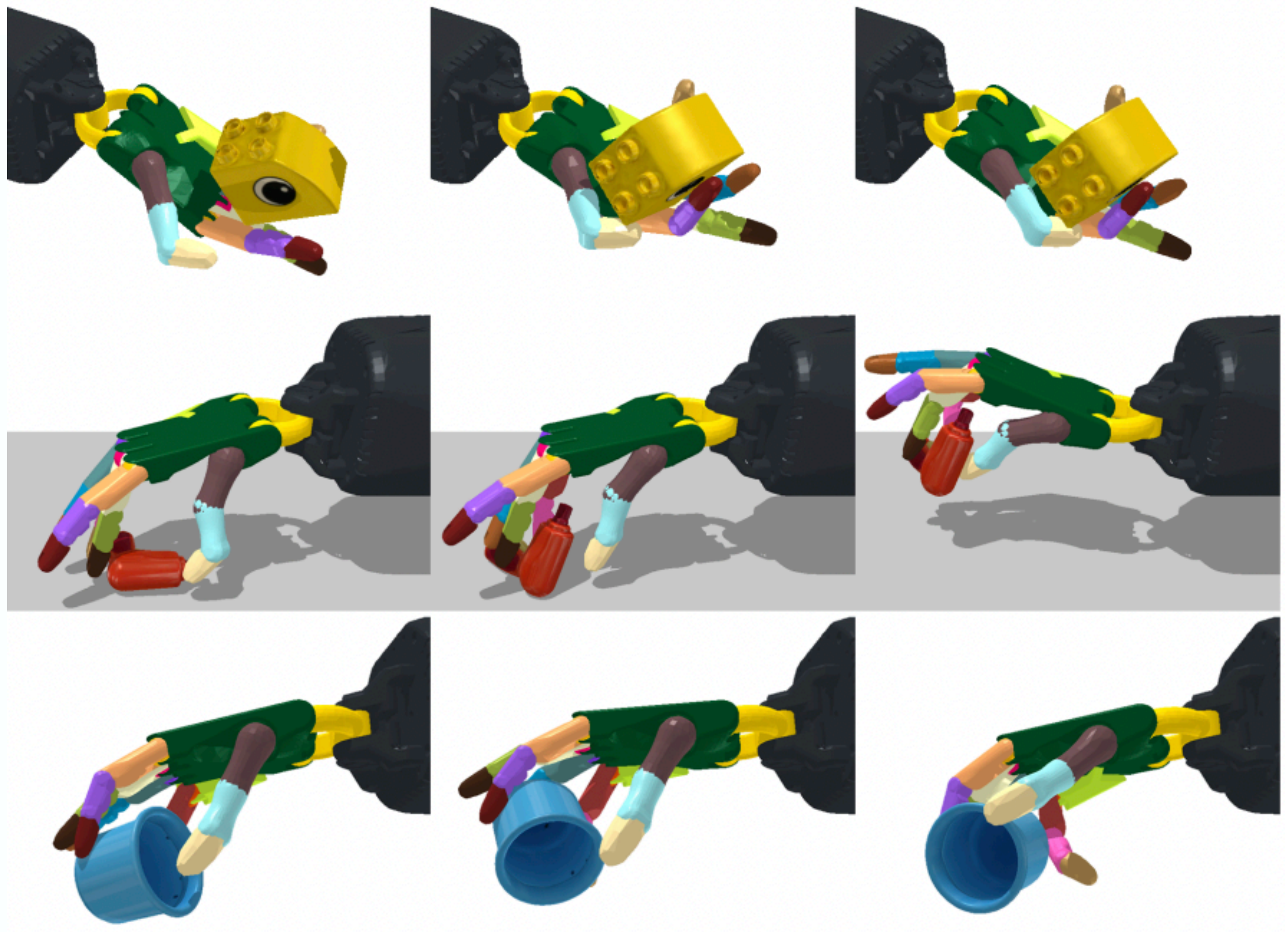


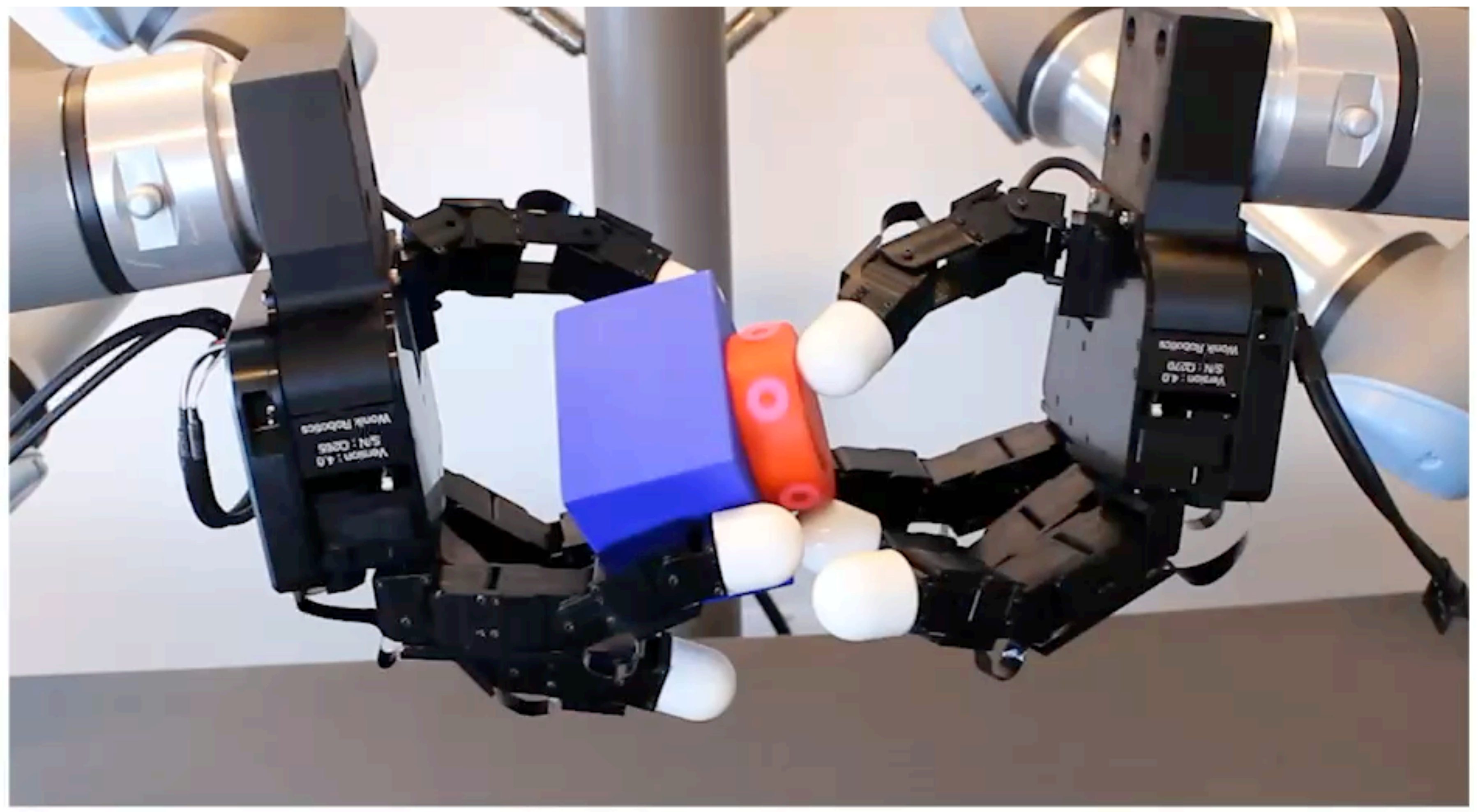
Pang et al., 2023

Why is There a Huge Gap?

- Popular paradigms: reinforcement learning







OpenAI, 2018

Chen et al., 2021

Lin et al., 2024

What about Learning from Human Motion?

- The progress in the vision community

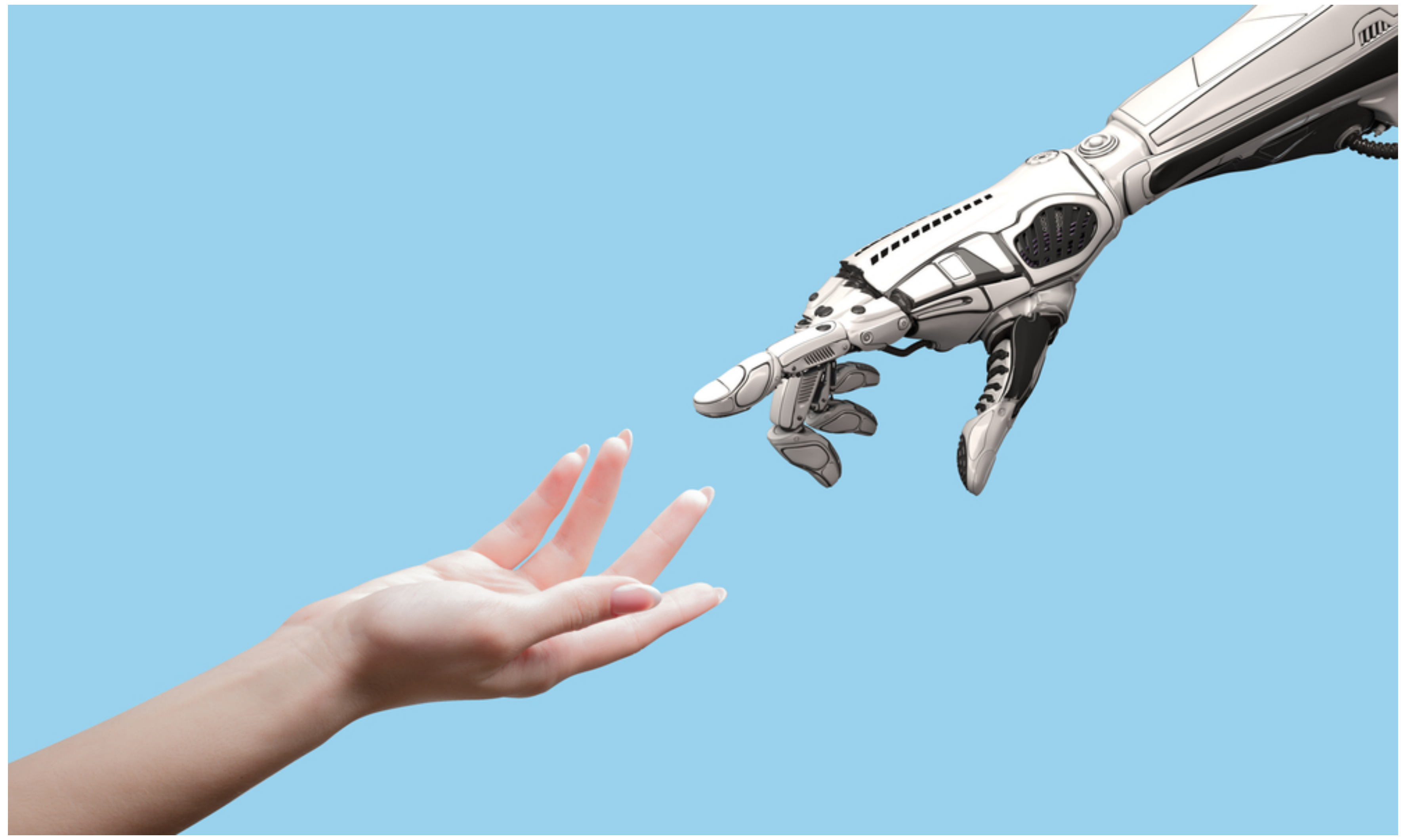
HOI4D, Liu et al., 2022

HaMeR, Pavlakos et al., 2024

MCC-HO, Wu et al., 2024

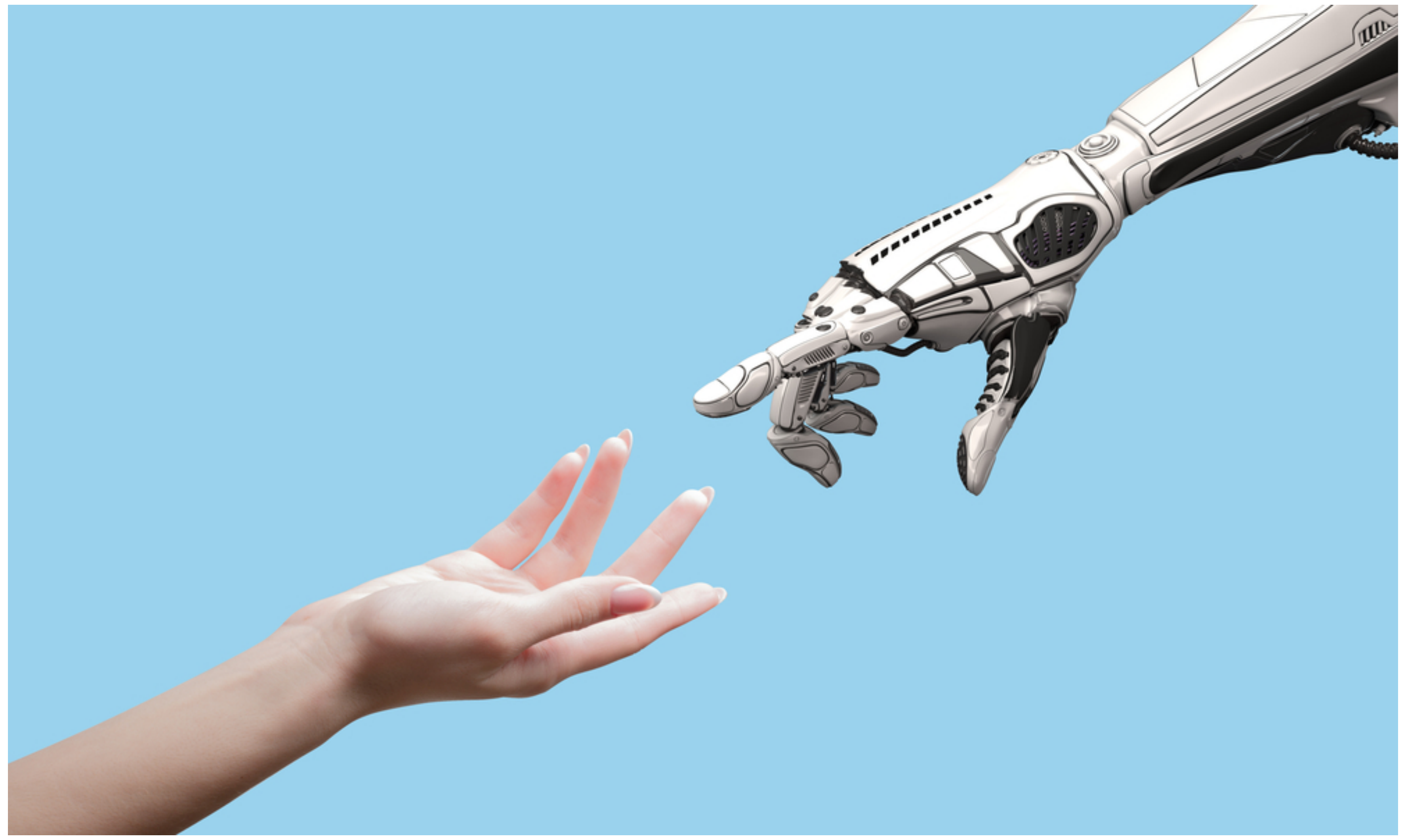
What about Learning from Human Motion?

- Challenges:
  - Embodiment gap
  - Missing of “actions”
  - Heterogenous operation targets and tasks



What about Learning from Human Motion?

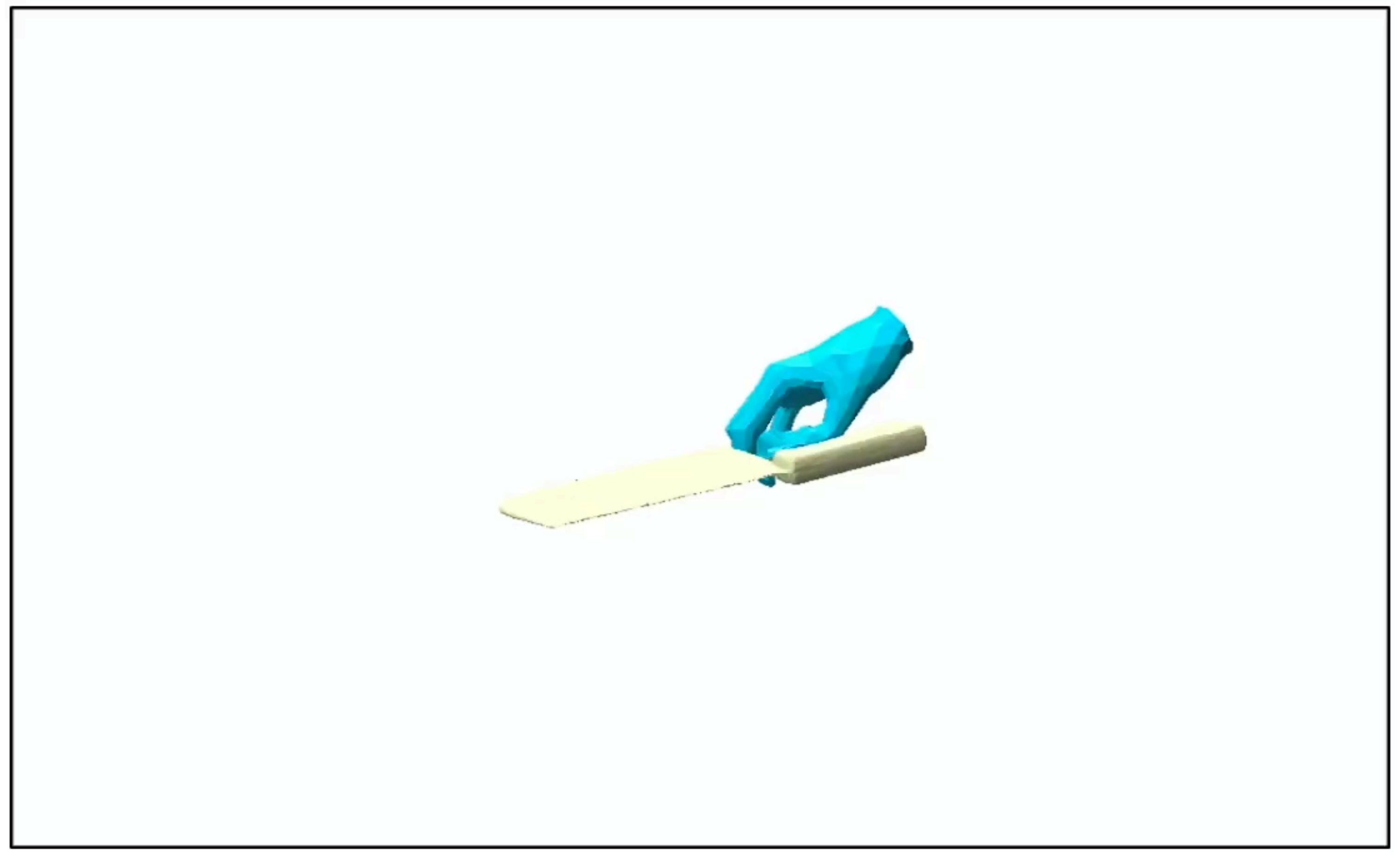
- Challenges:
  - Embodiment gap
  - Missing of “actions”
  - Heterogenous operation targets and tasks

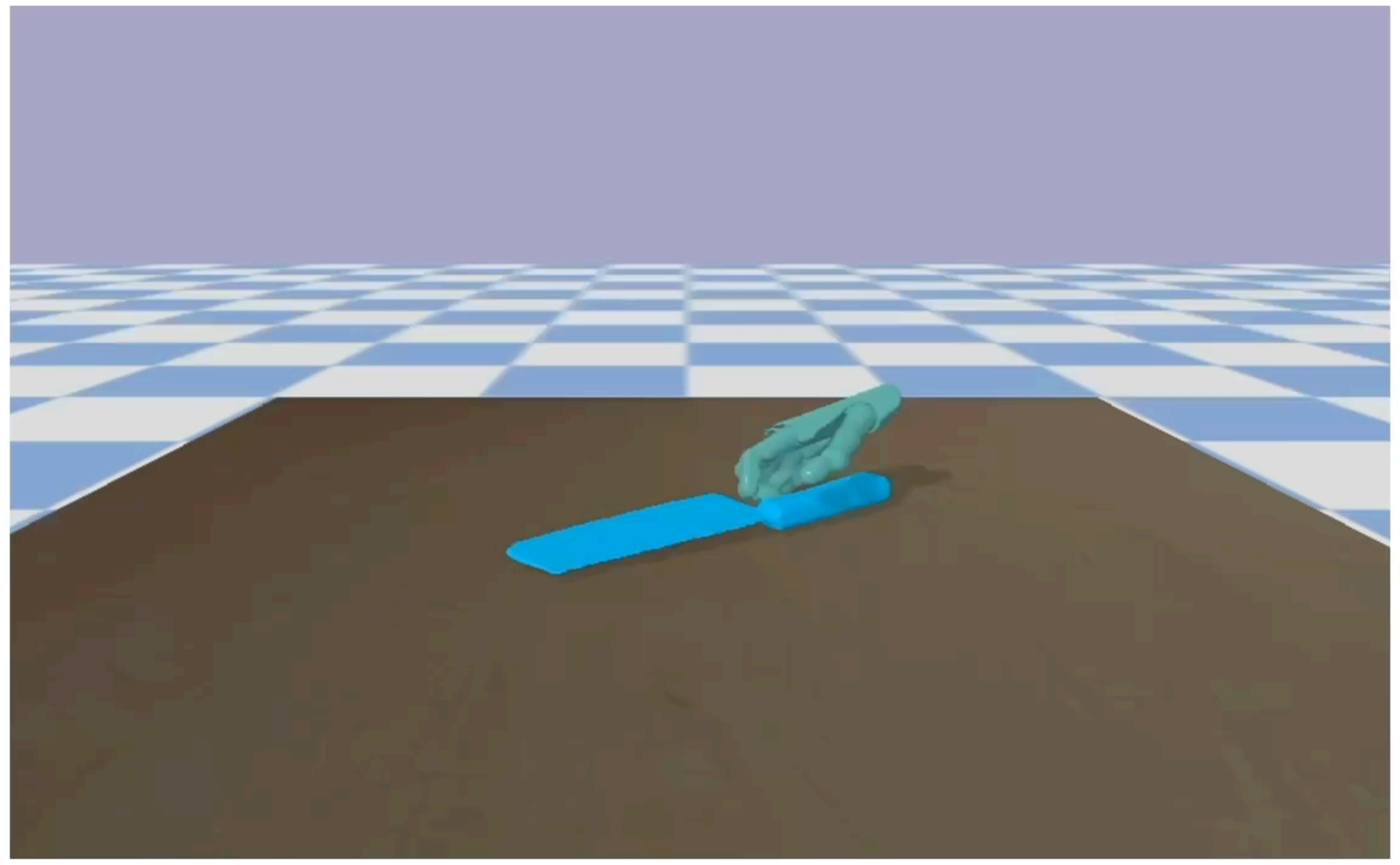


Only learn motion planning from human data and leave the rest to a general neural tracking controller

A Cross-Embodiment Tracking Control Paradigm

Using a knife to chop





Task Description

Generative Human Motion Planning

Cross-Embodiment
Neural Tracking Control

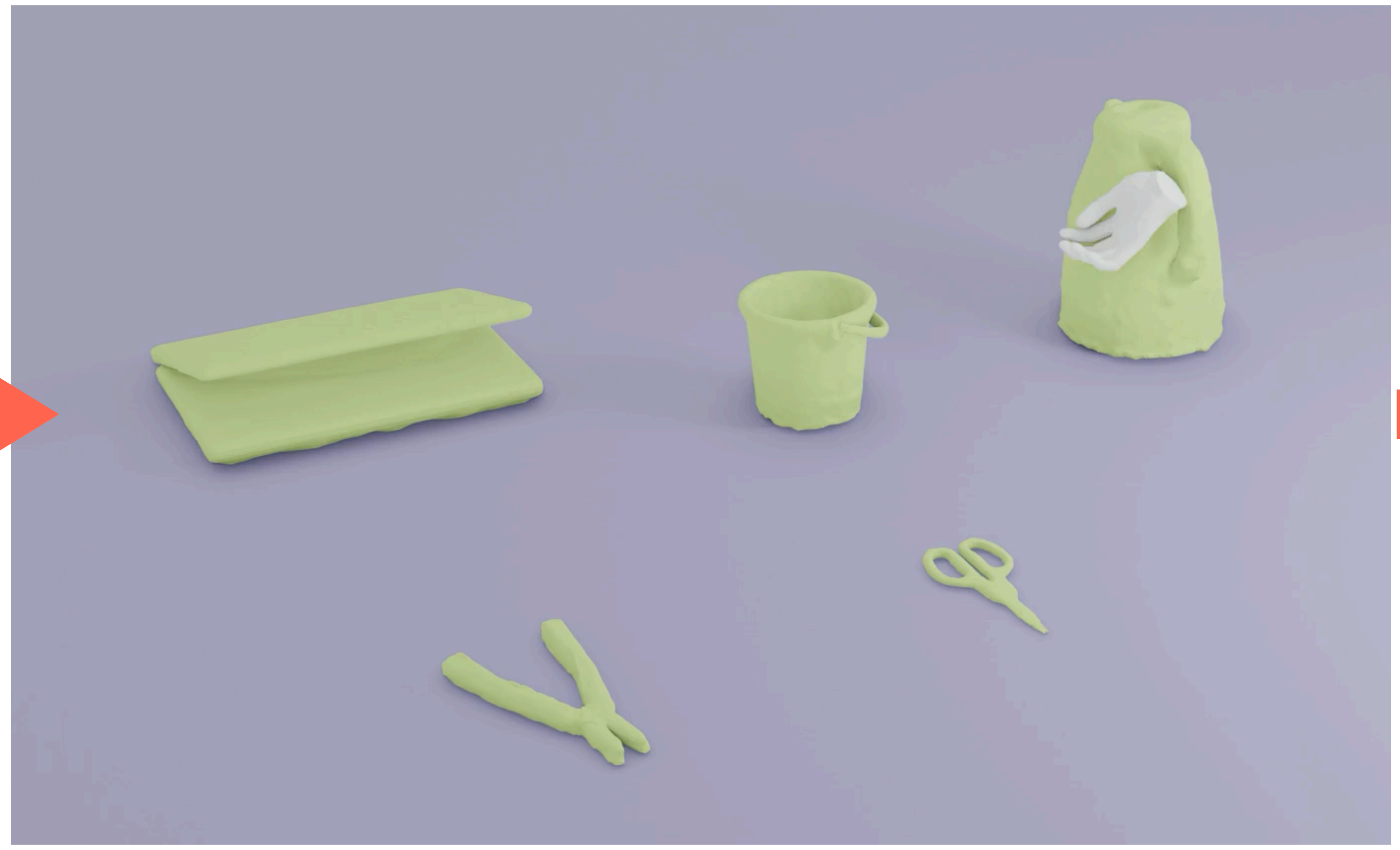
Advantages

- Separate planning and control similar to traditional paradigms
- Partially separate semantics from dynamics
- Neural planner and controller to harvest the power of data

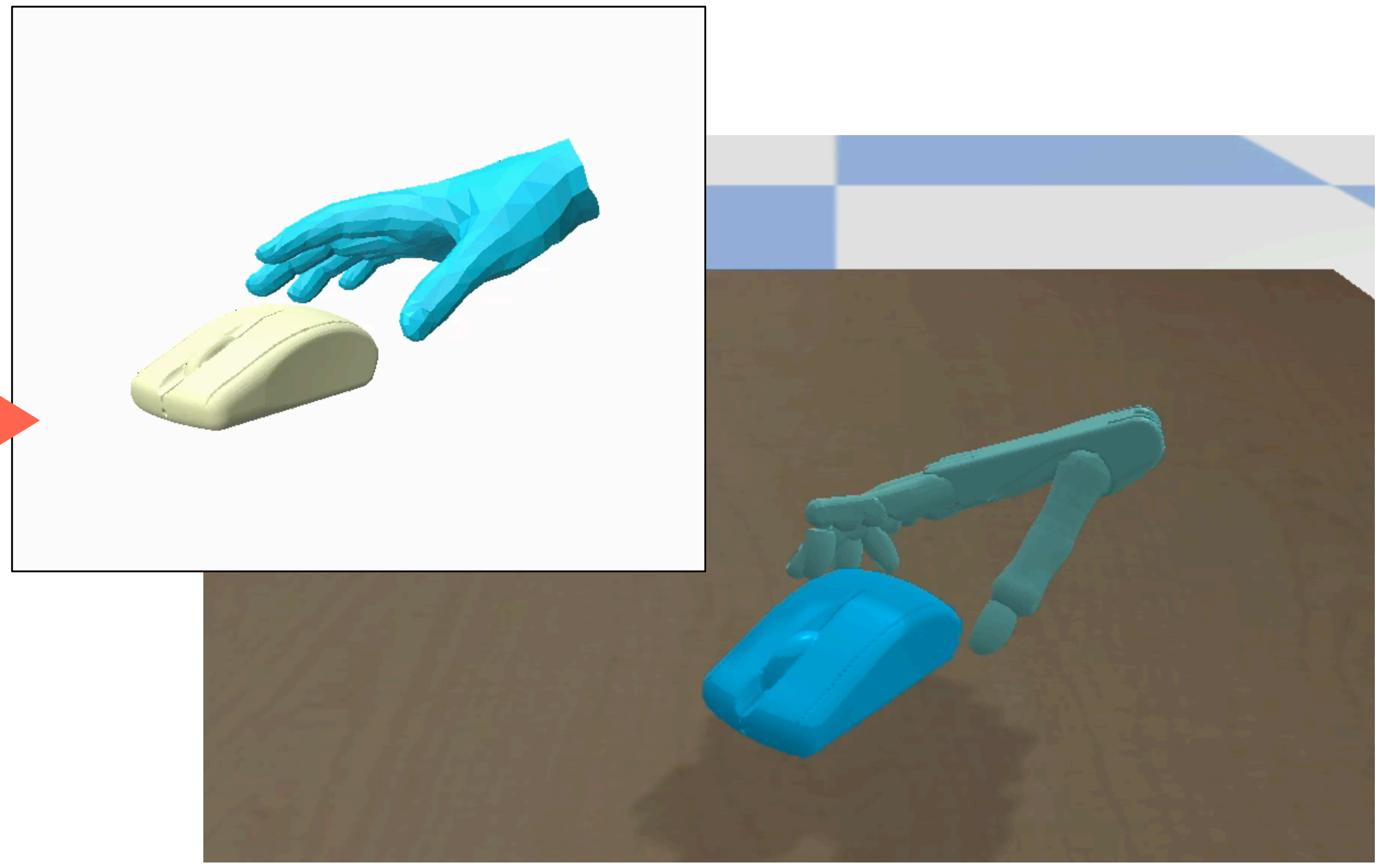
A Cross-Embodiment Tracking Control Paradigm

Capturing Human Manipulation Data

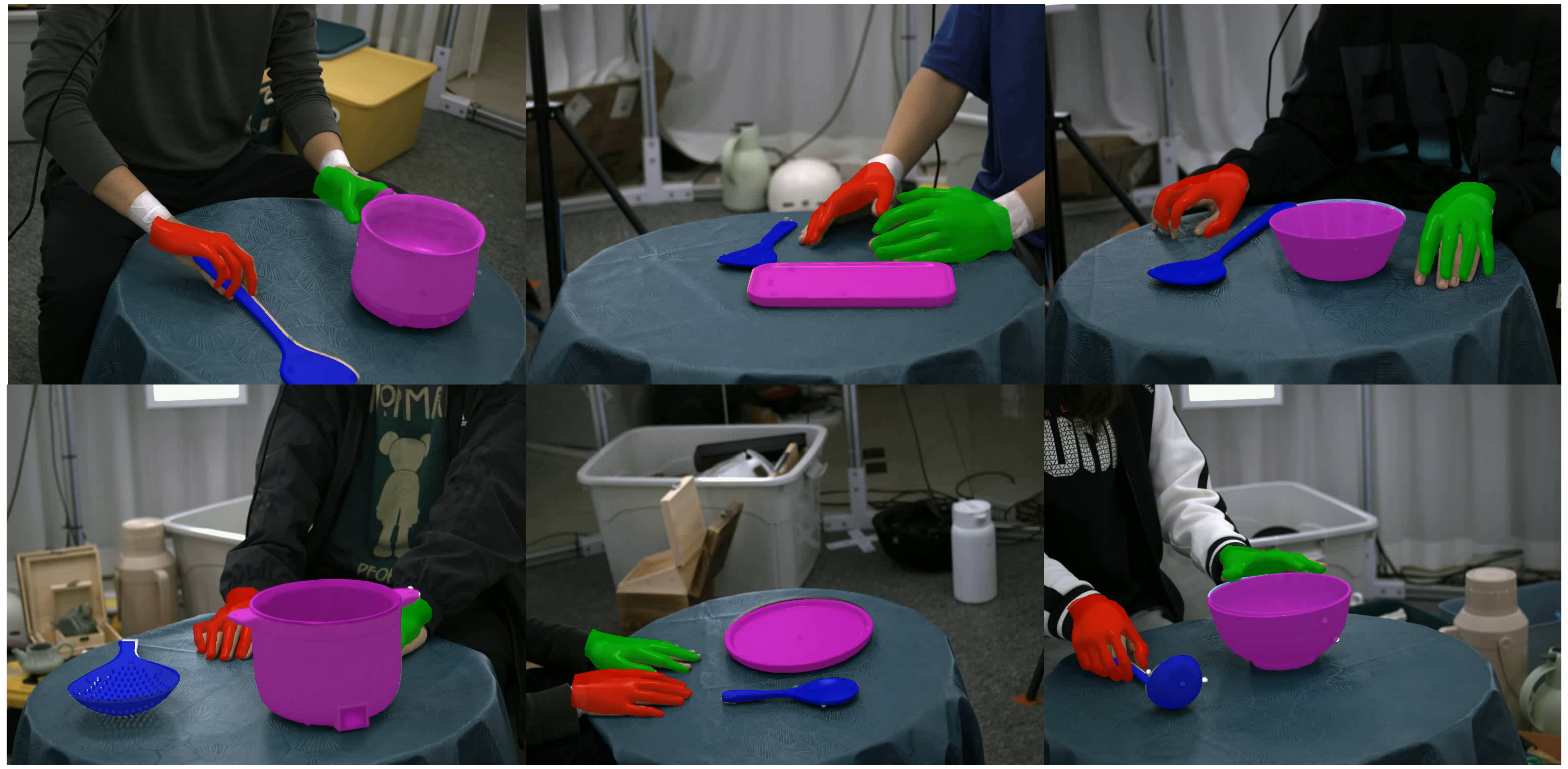
Generative Human Manipulation Planning



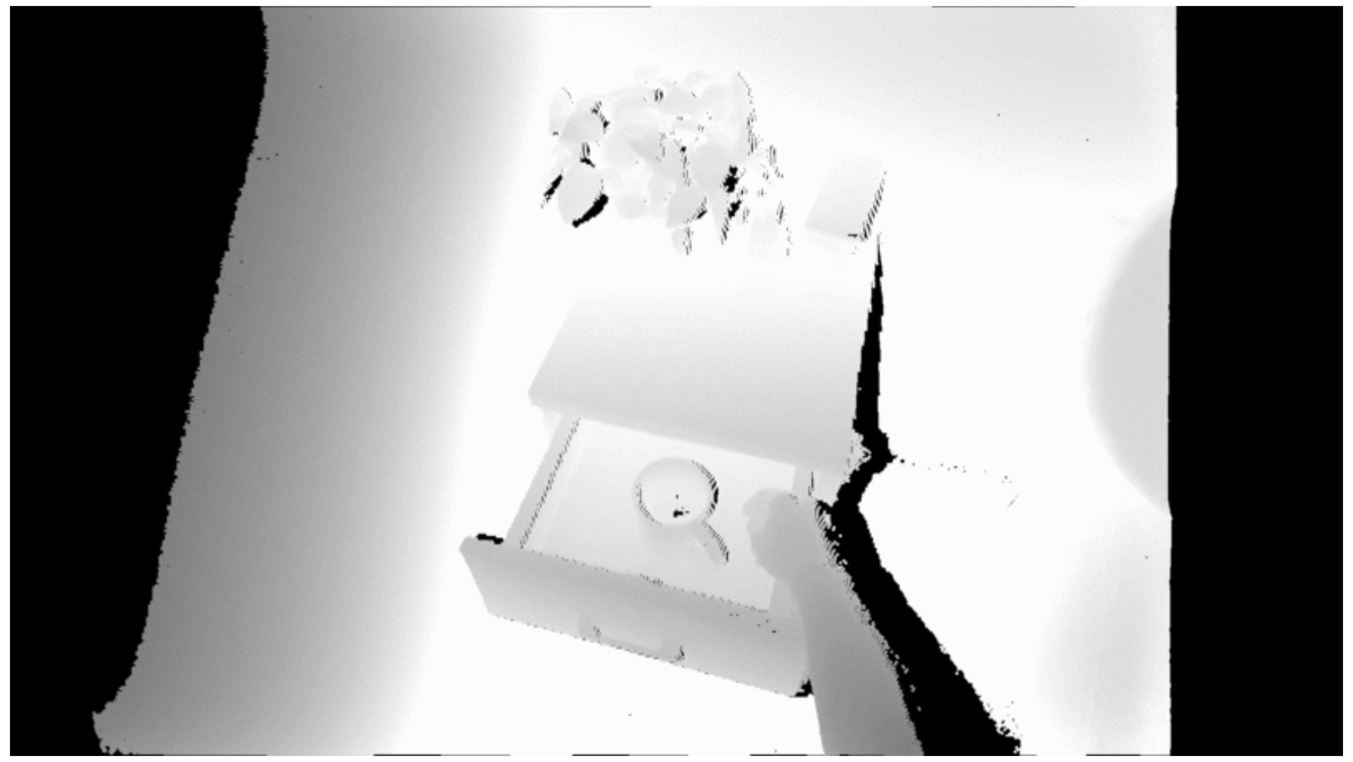
Cross-Embodiment Tracking Control



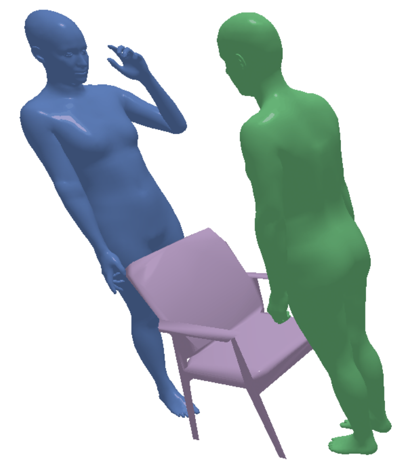
Capturing Human Manipulation Data



TACO: Benchmarking Generalizable Bimanual Tool-ACtion-Object Understanding
Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, Li Yi. CVPR 2024



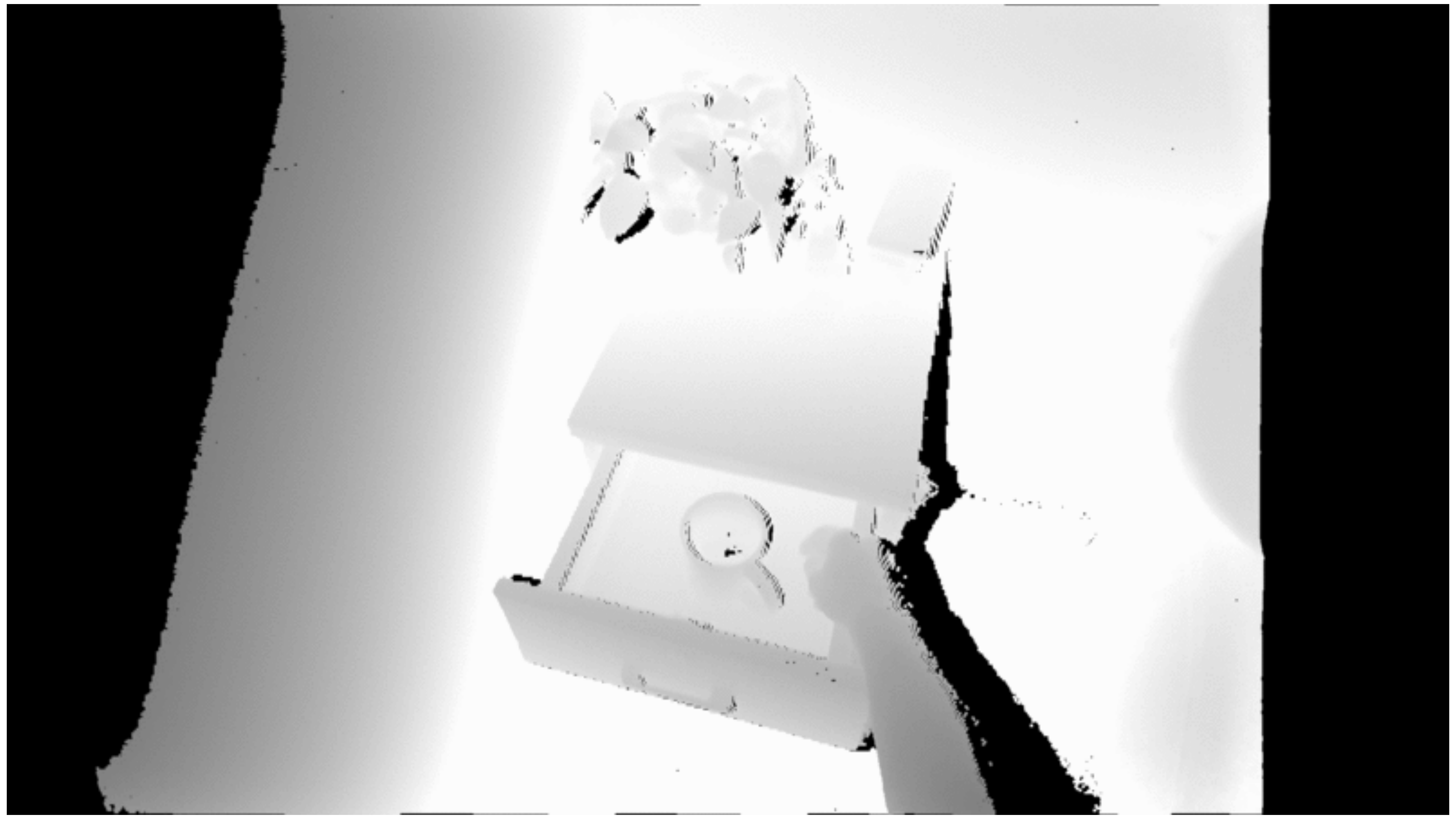
HOI4D: A 4D Egocentric Dataset for Category-Level Human-Object Interaction
Yunze Liu, Yun Liu*, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, Li Yi. CVPR 2022*



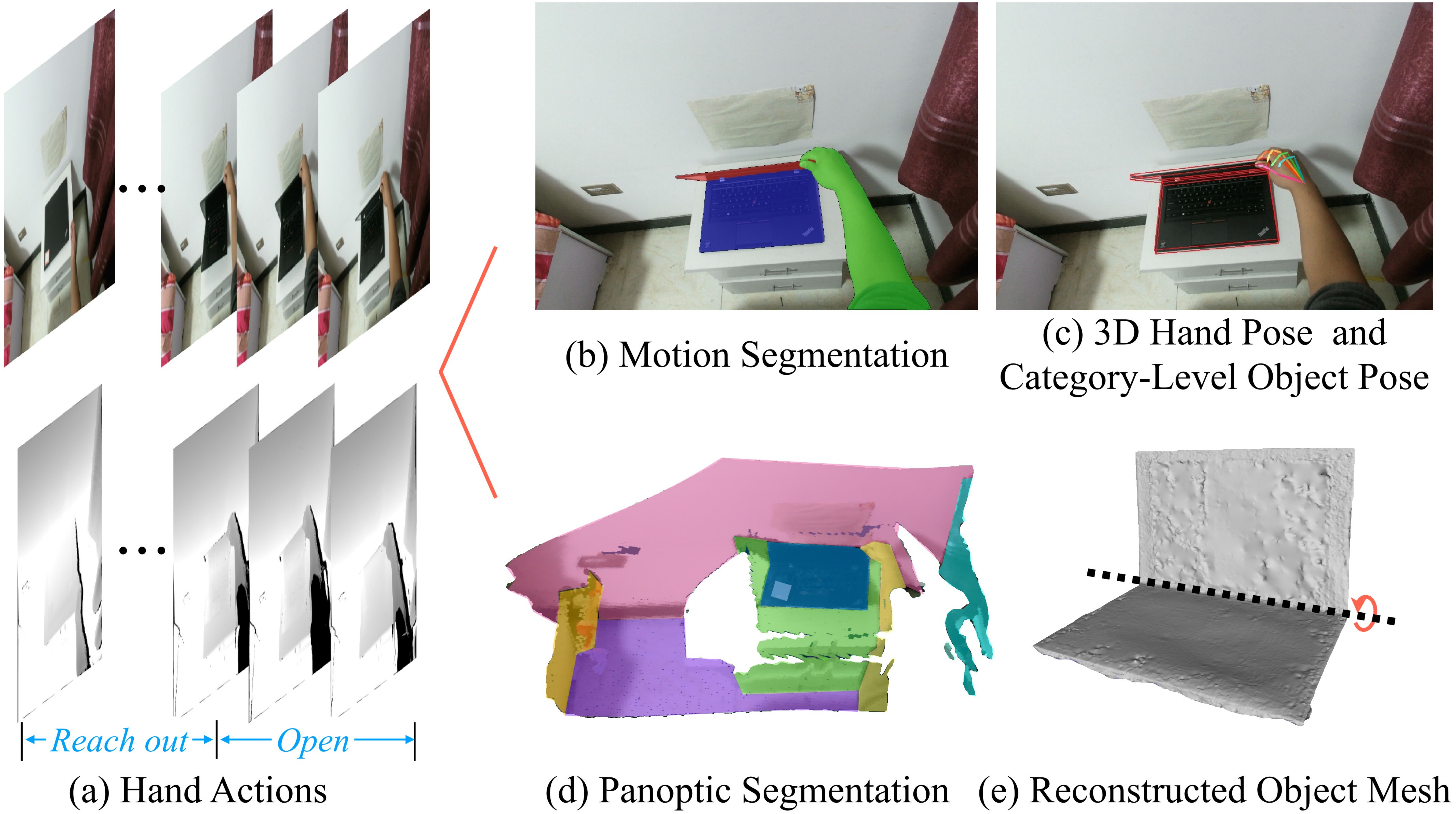
CORE4D: A 4D Human-Object-Human Interaction Dataset for Collaborative Object REarrangement
Chengwen Zhang, Yun Liu*, Ruofan Xing, Bingda Tang, Li Yi. In submission*

HOI4D Dataset

- The first dataset for 4D egocentric category-level human-object interaction

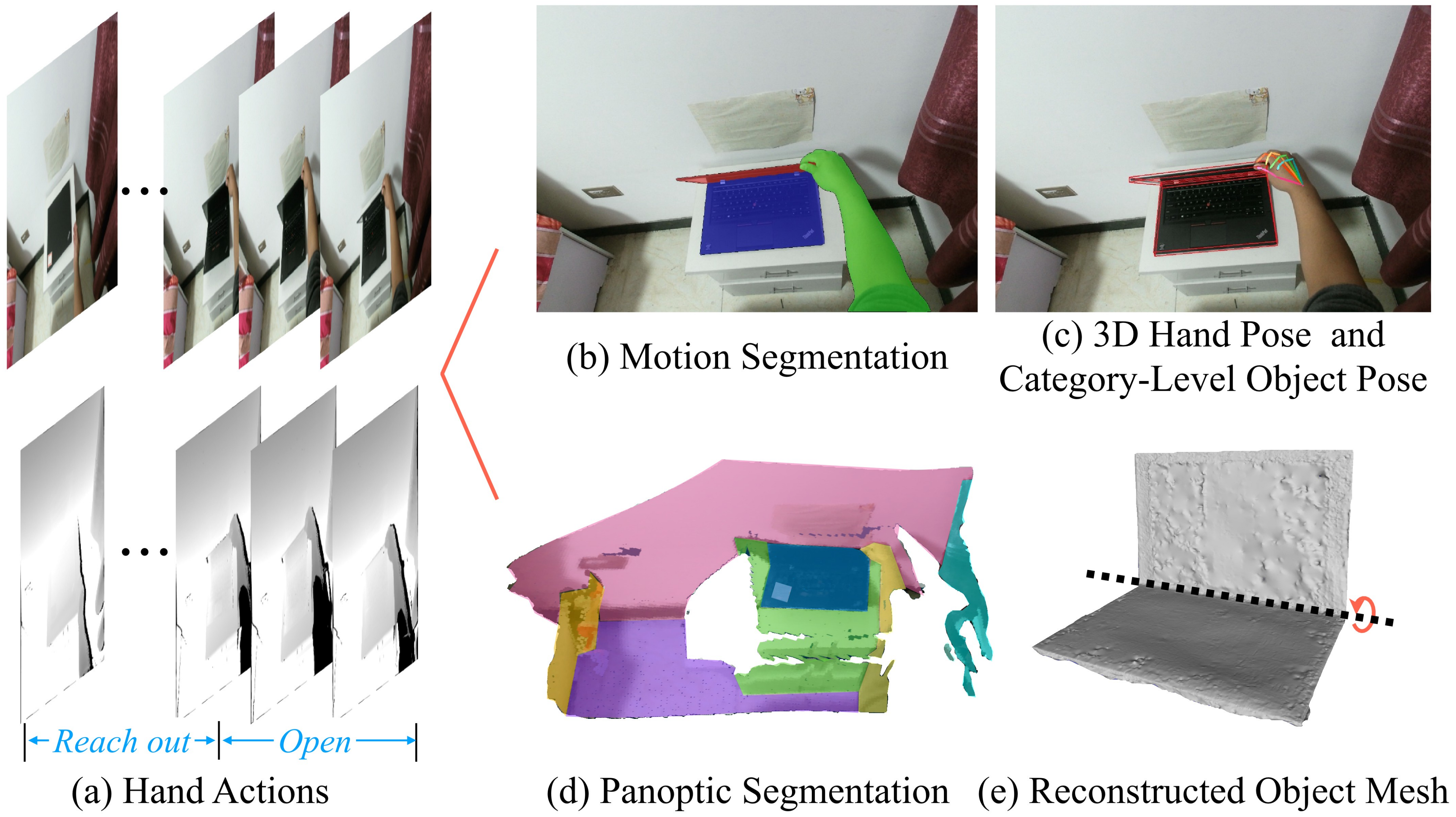


Rich Annotations

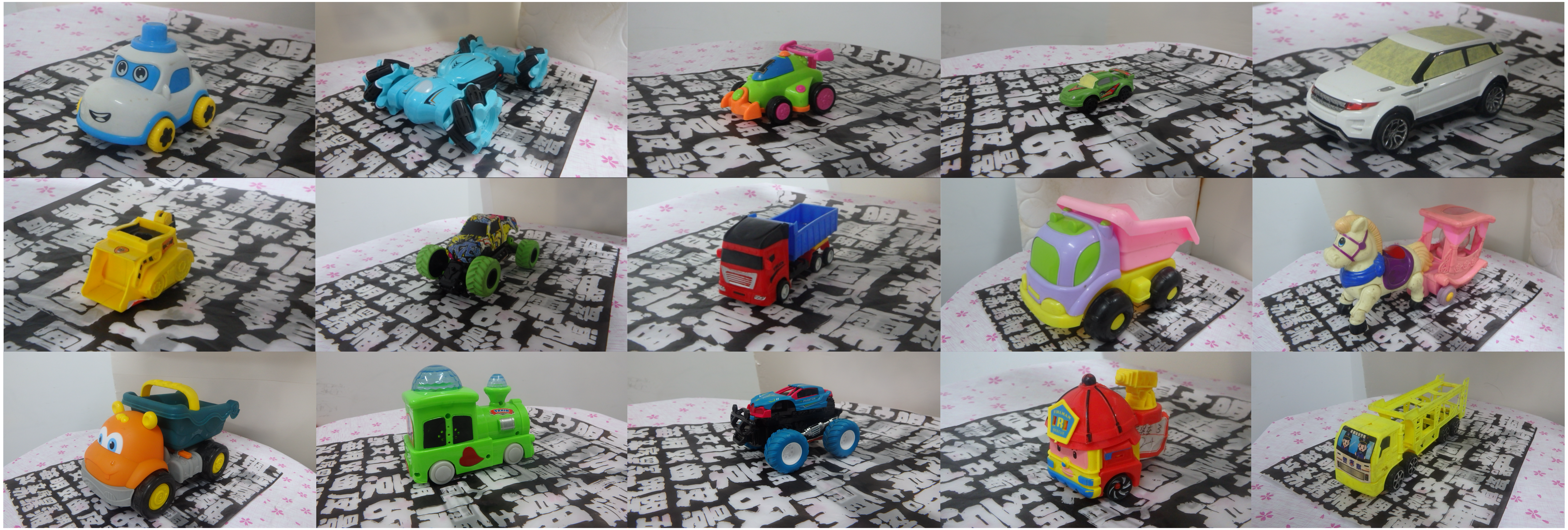


Rich Annotations

- 4D panoptic segmentation
- 3D hand pose
- Category-level object pose (rigid and articulated)
- Object mesh with mobility annotation
- Per-frame motion segmentation
- Camera pose
- Action segmentation



Feature I: Category Level

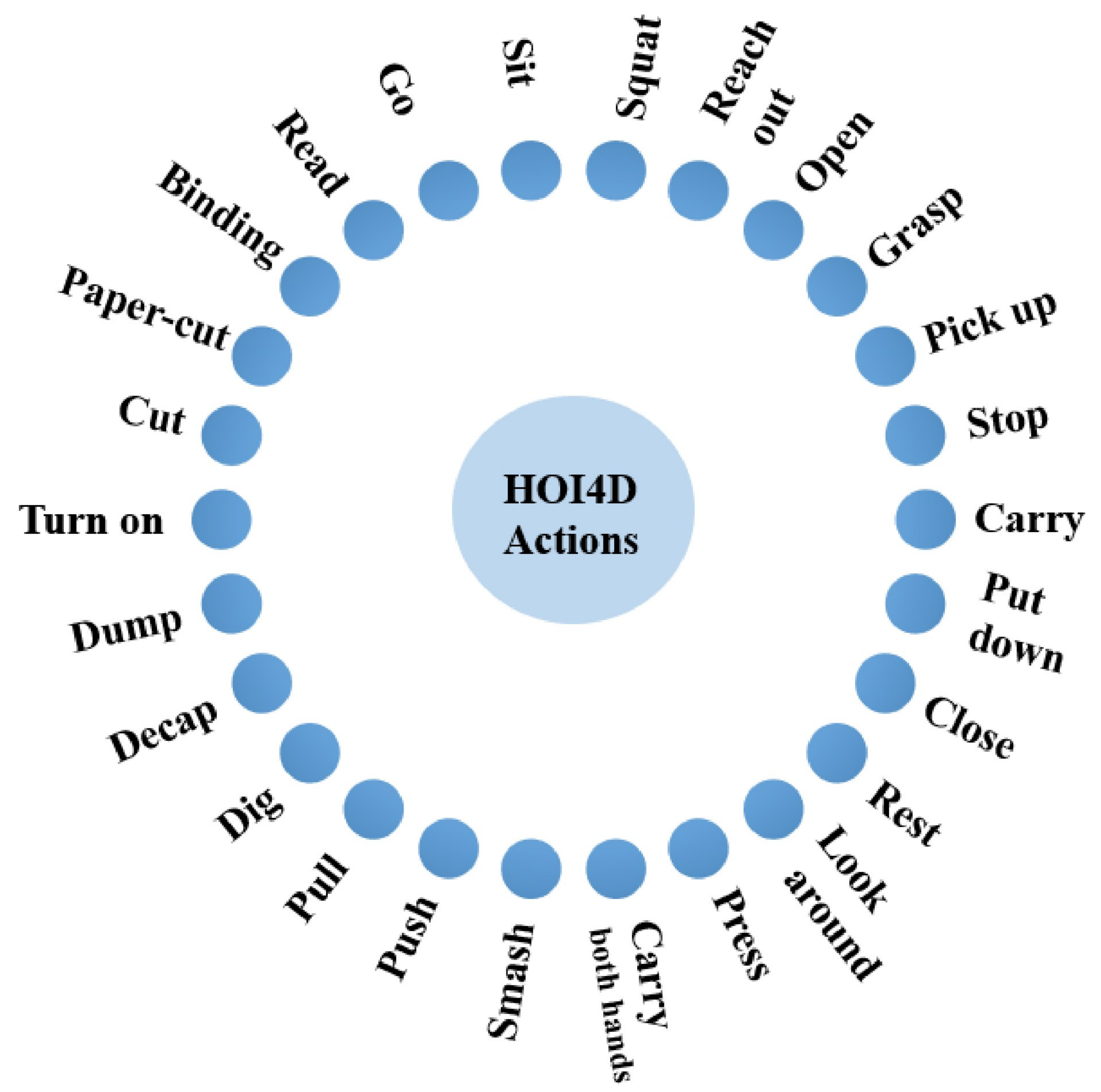


Feature II: Large Scale

- 2.4M RGB-D frames over 4,000 videos
- 800 object instances from 16 categories (7 rigid + 9 articulated)

Feature II: Large Scale

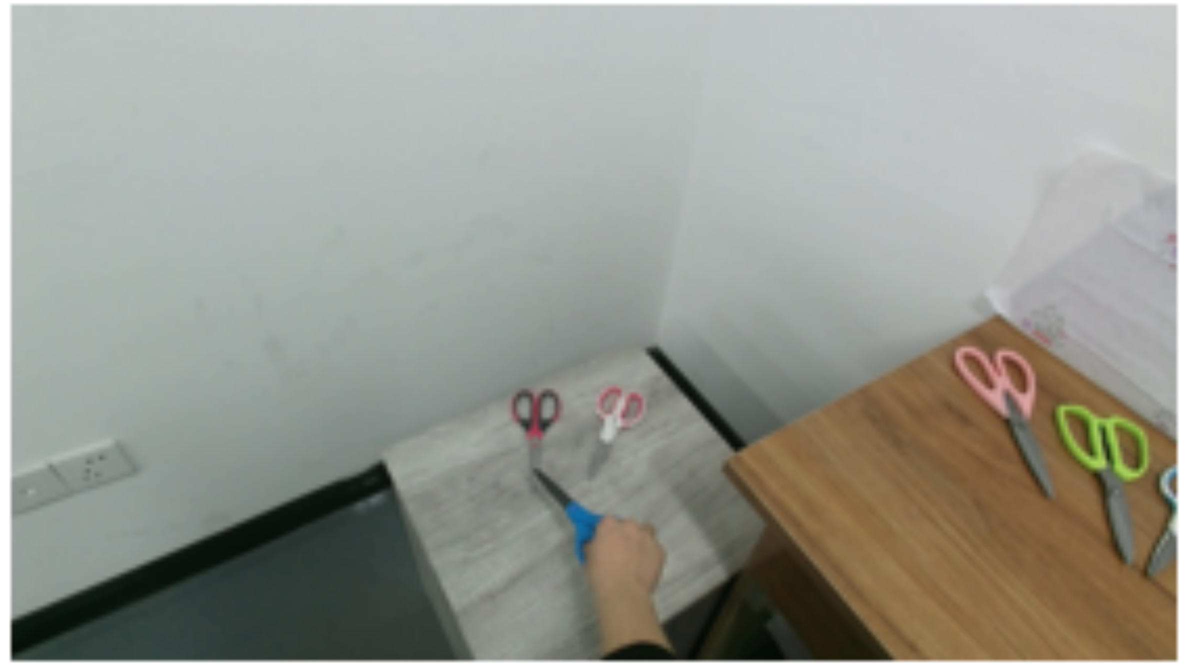
- 2.4M RGB-D frames over 4,000 videos
- 800 object instances from 16 categories (7 rigid + 9 articulated)
- 610 different indoor rooms
- 43 semantic category in 4D scenes
- 26 action categories
- 92 tasks including pick-and-place and functionality-b



Feature III: Functionality Driven

- Examples of interaction tasks

Safe: Open the door



Scissors : Pick up the scissors

Mug: Put it in the drawer

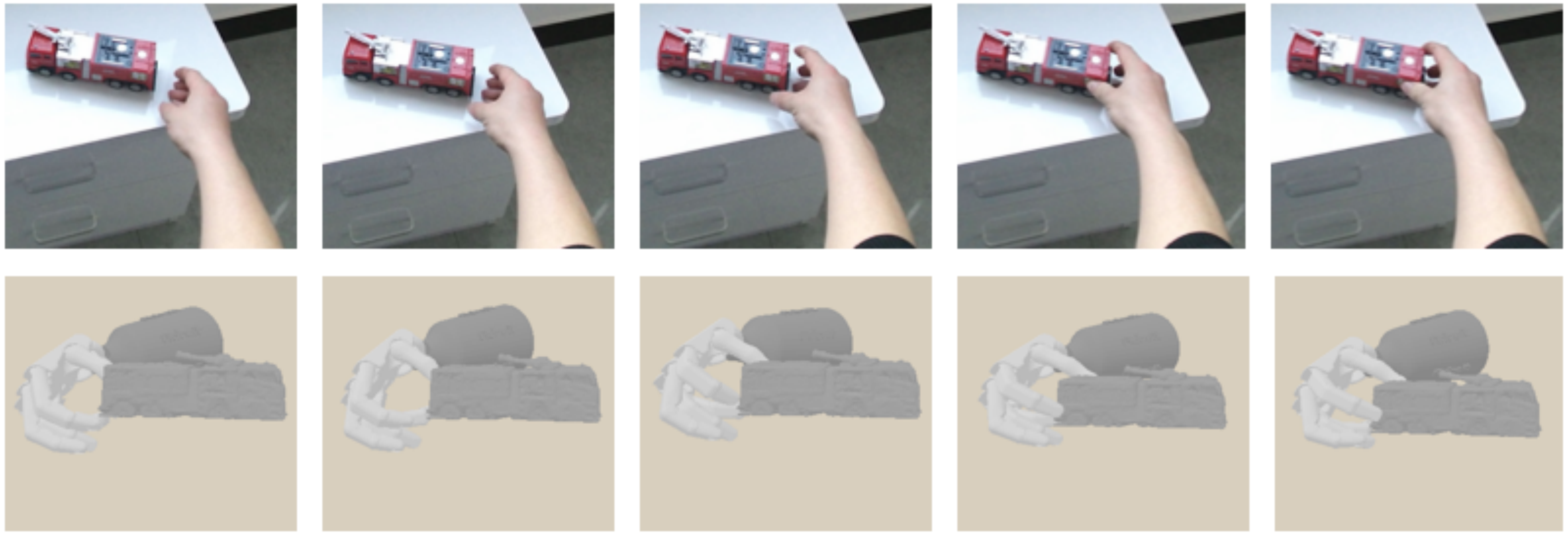
Bucket: Pour away the water

Laptop: Open your laptop

Hammer: Tap on the table

Application - Robot Learning from Human Demonstration

- Learning robotic dexterous manipulation from human demonstration

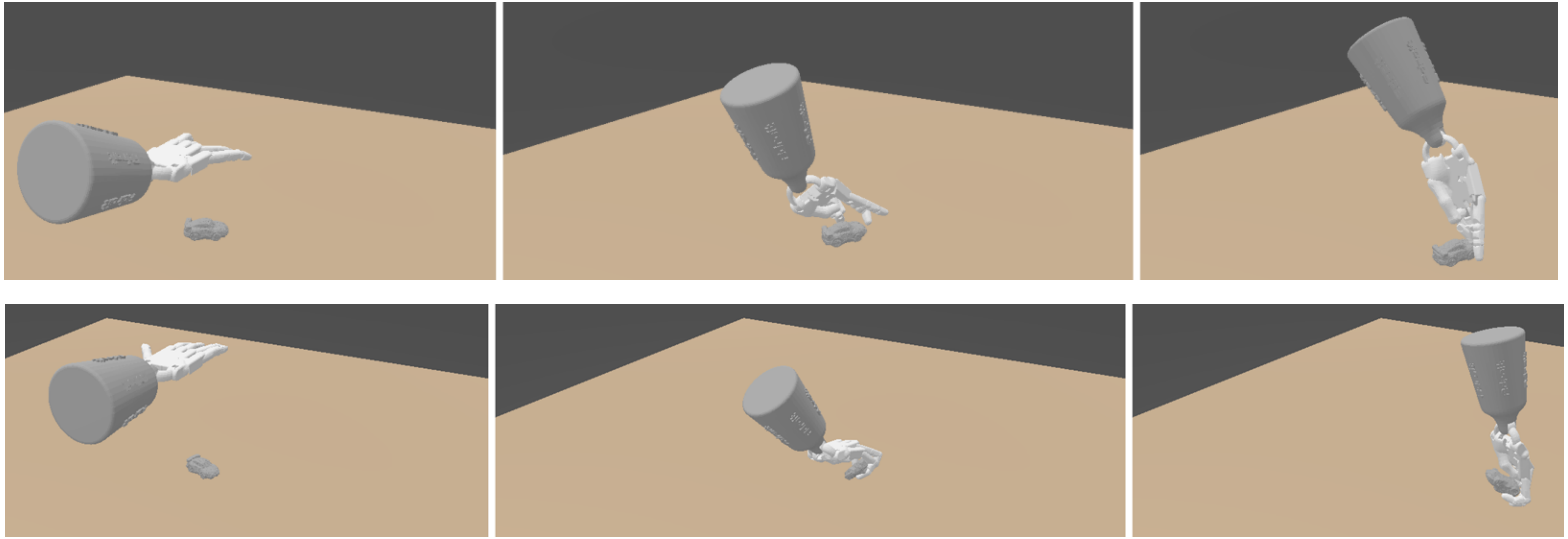


Application - Robot Learning from Human Demonstration

- Mixing imitation learning (IL) and reinforcement learning (RL)
- Task: Pick up the toy car and keep it a certain height from the table

RL only

RL + IL



More Applications

- Knowledge transfer across sensors
- Dynamic reconstruction
- Camera re-localization in dynamic scenes
- Action anticipation
- …

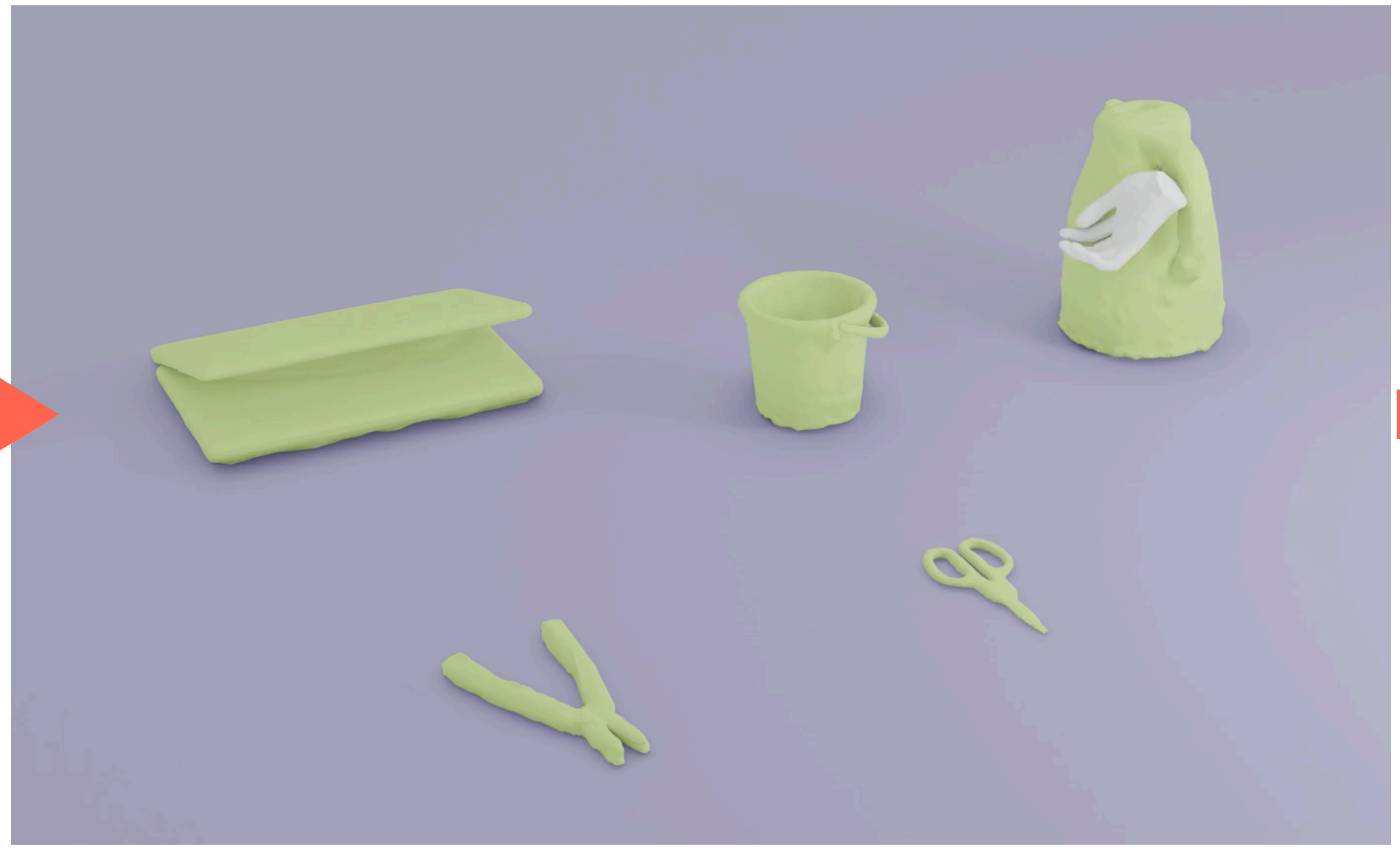
Summary of HOI4D

- The first dataset for 4D egocentric category-level human-object interaction
- An integrated data collection and annotation pipeline
- Various applications including 4D perception and robot learning

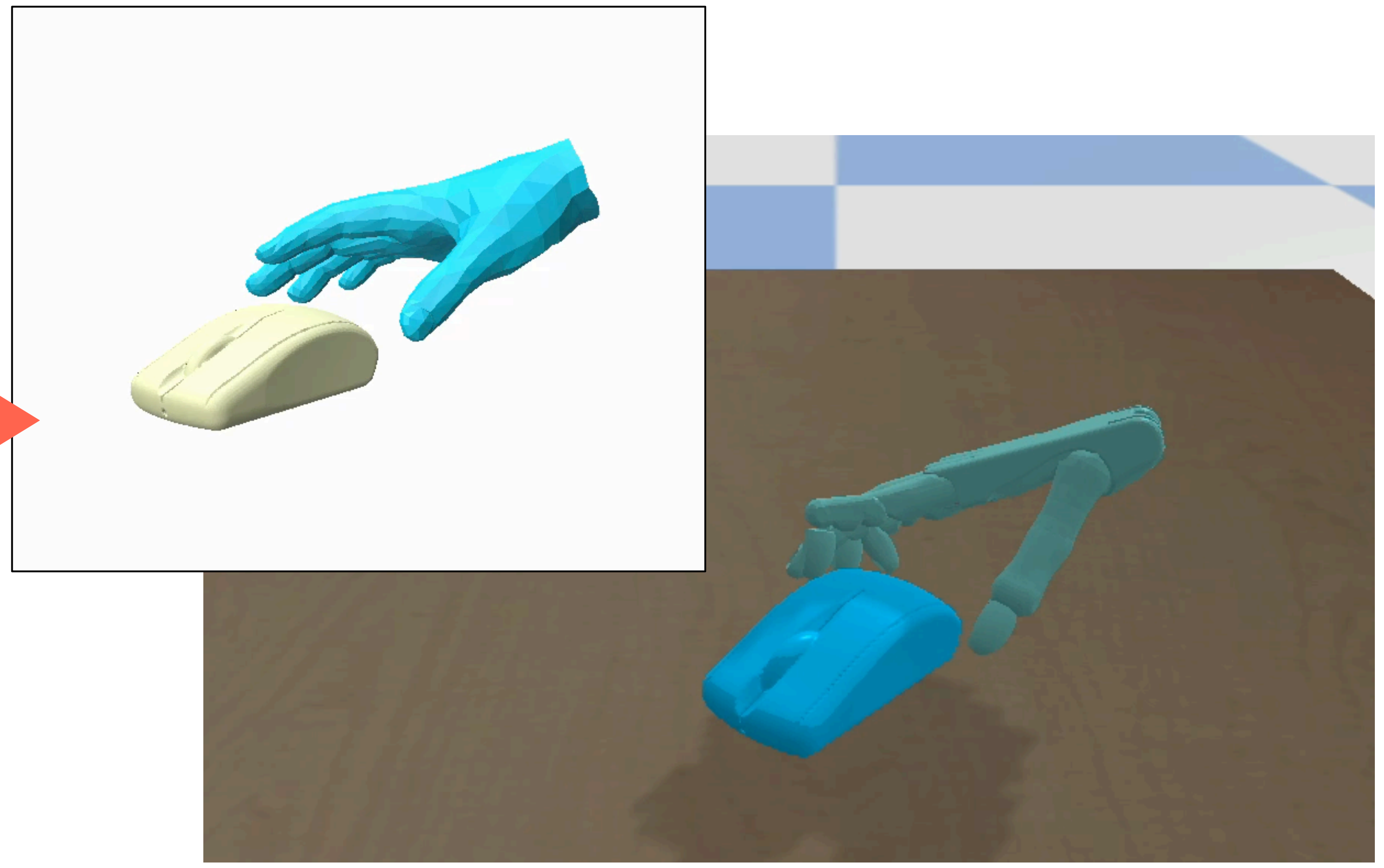
A Cross-Embodiment Tracking Control Paradigm

Capturing Human Manipulation Data

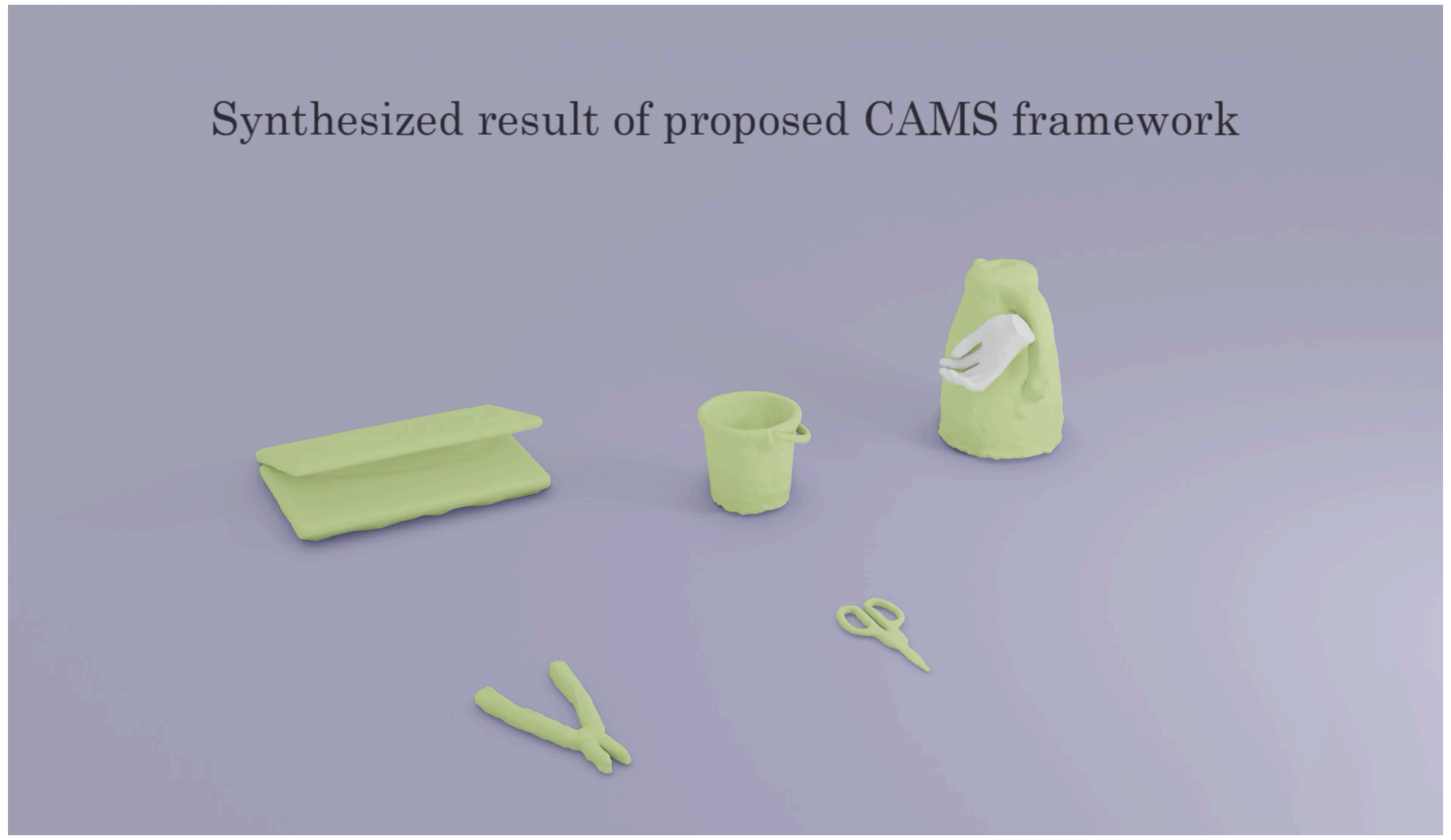
Generative Human Manipulation Planning



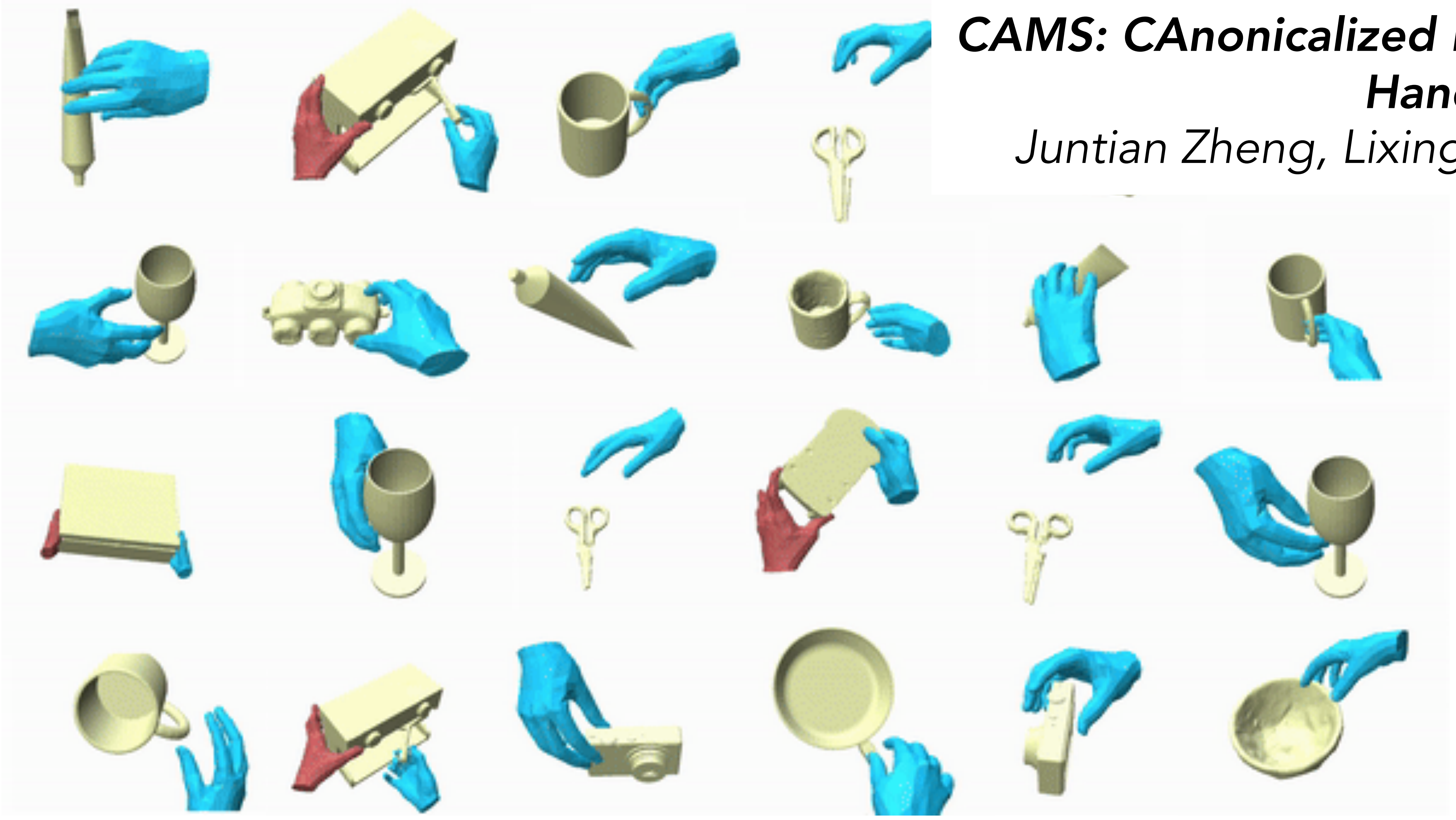
Cross-Embodiment Tracking Control



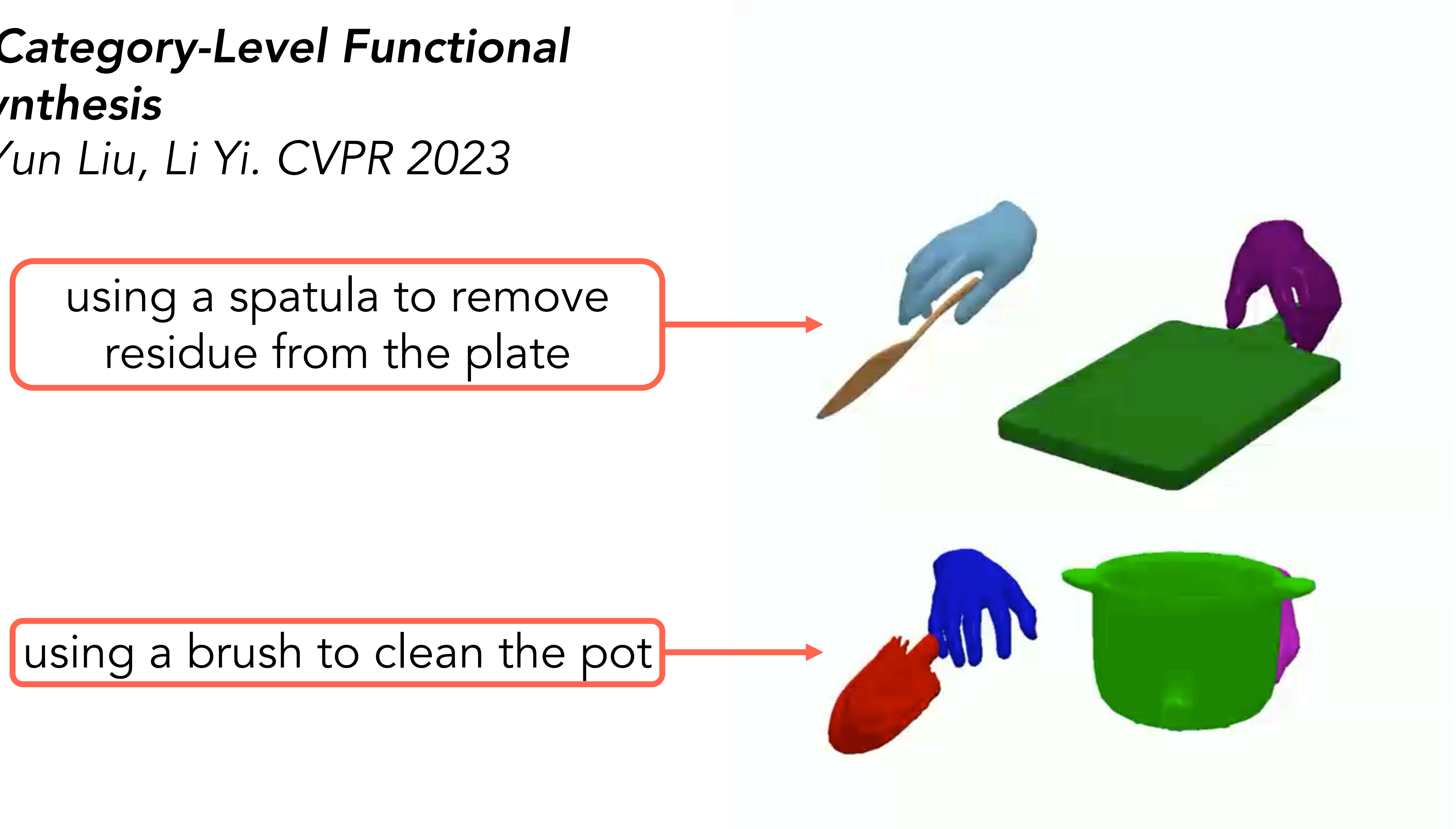
Generative Human Manipulation Planning



CAMS: CAnonicalized Manipulation Spaces for Category-Level Functional Hand-Object Manipulation Synthesis
Juntian Zheng, Lixing Fang, Qingyuan Zheng, Yun Liu, Li Yi. CVPR 2023



GeneOH Diffusion: Generalizable Hand-Object Interaction Denoising via Denoising Diffusion
Xueyi Liu, Li Yi. ICLR 2024



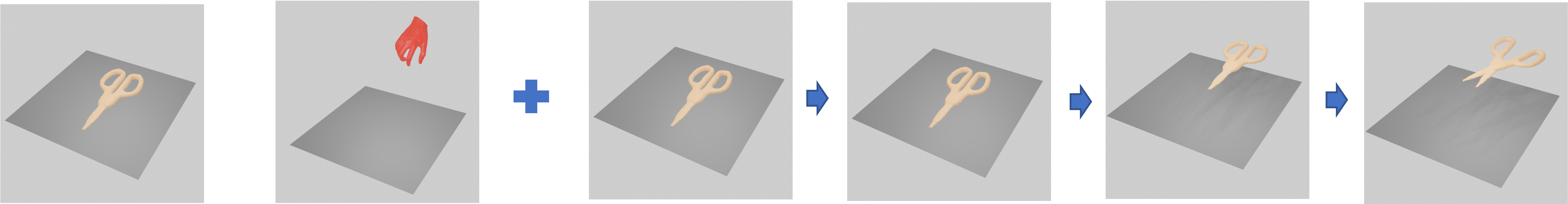
Multibody Human-Object Interaction Synthesis via Synchronized Motion Diffusion
Wenkun He, Yun Liu, Ruitao Liu, Li Yi. In submission

Synthesized result of proposed CAMS framework

CAMS: CAnonicalized Manipulation Spaces for Category-Level Functional Hand-Object Manipulation Synthesis

Juntian Zheng, Lixing Fang, Qingyuan Zheng, Yun Liu, Li Yi. CVPR 2023

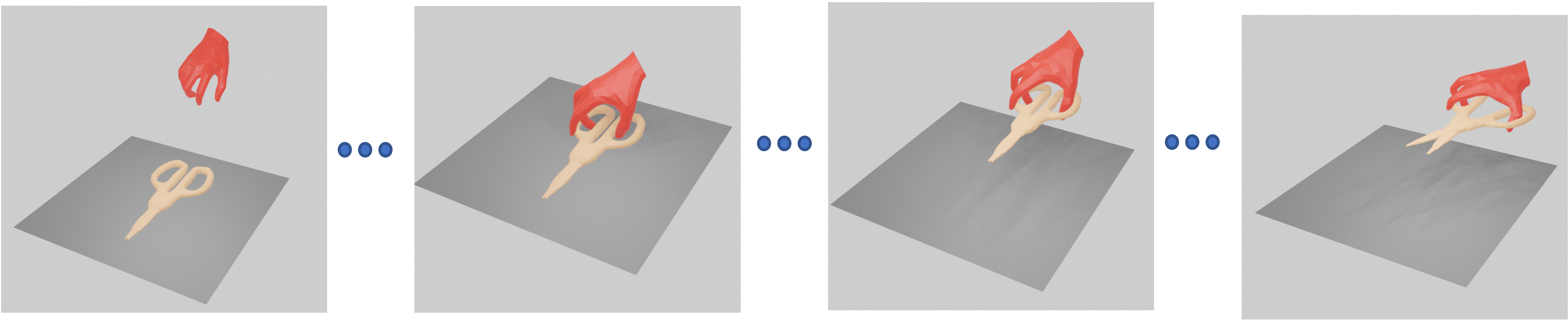
Task Definition & Challenges



Object Geometry　Initial Hand　Goal Sequence

Input

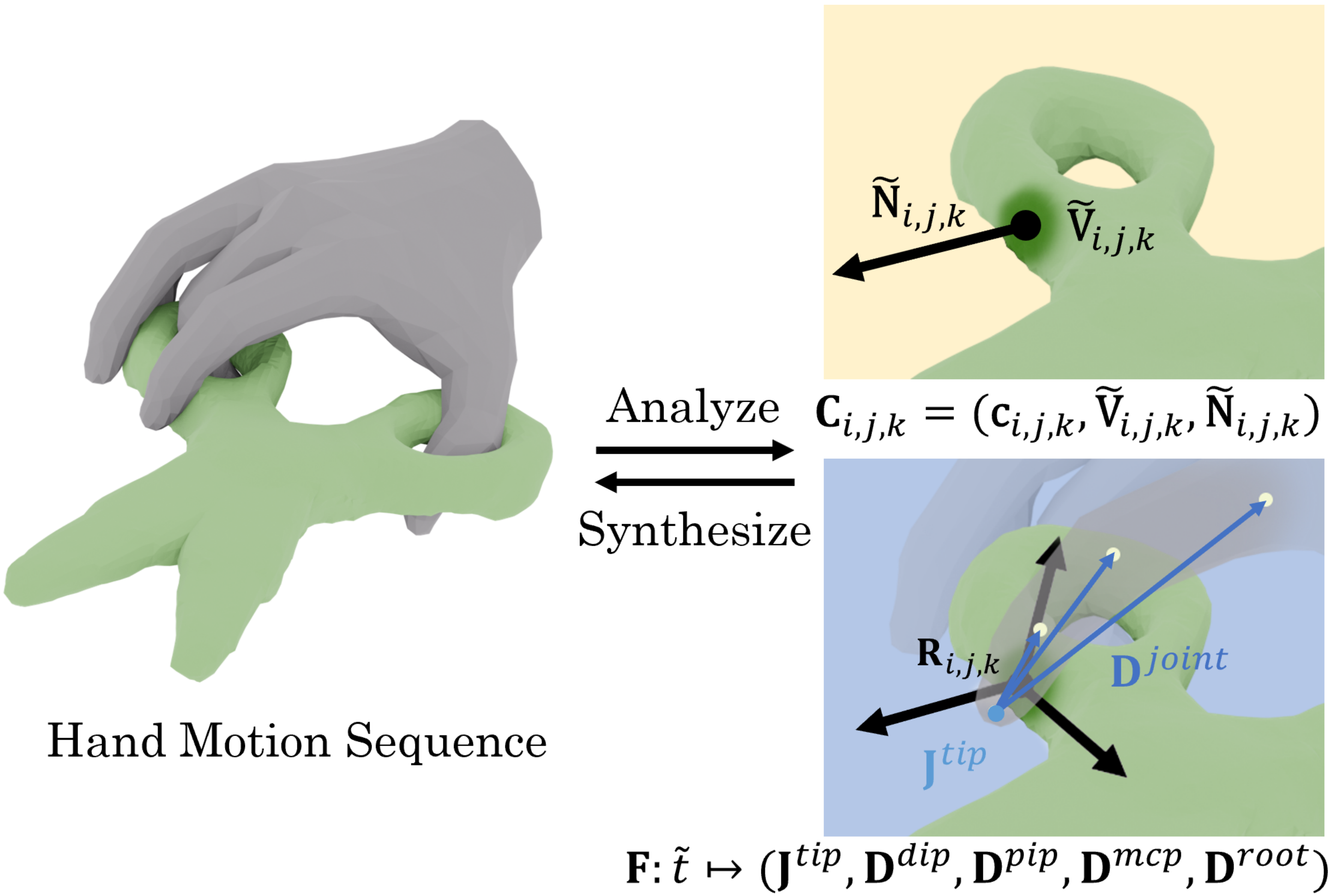
Output



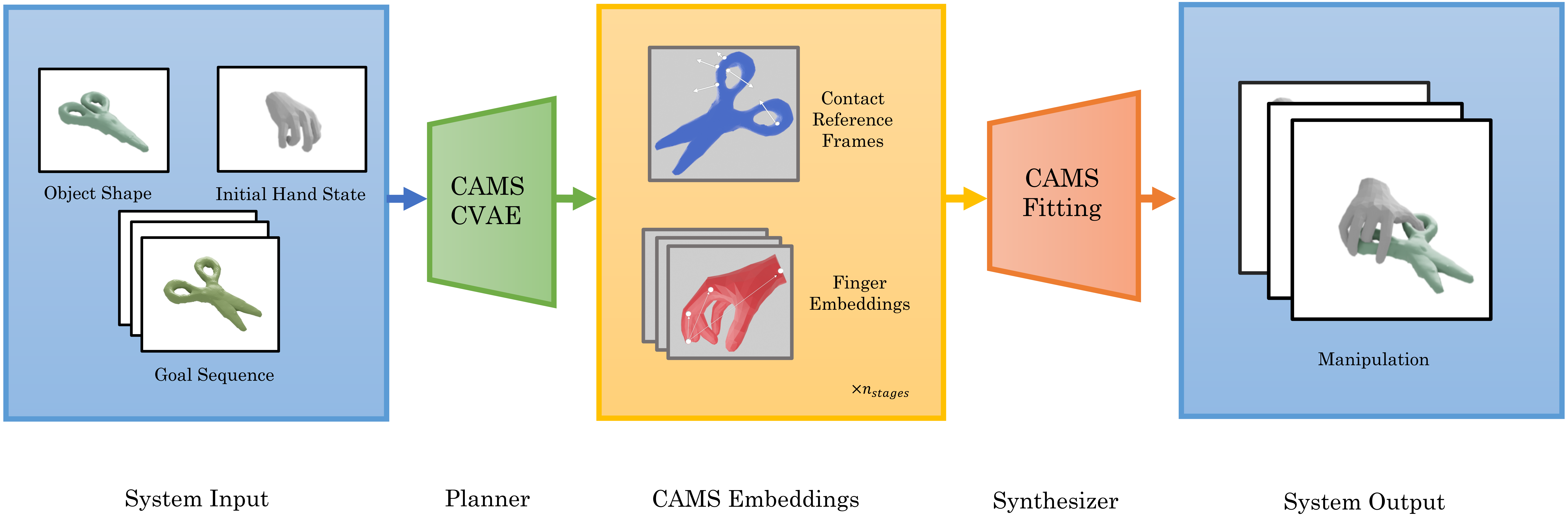
Challenges:
Shape Diversity
Manipulation Diversity

Functional Manipulation

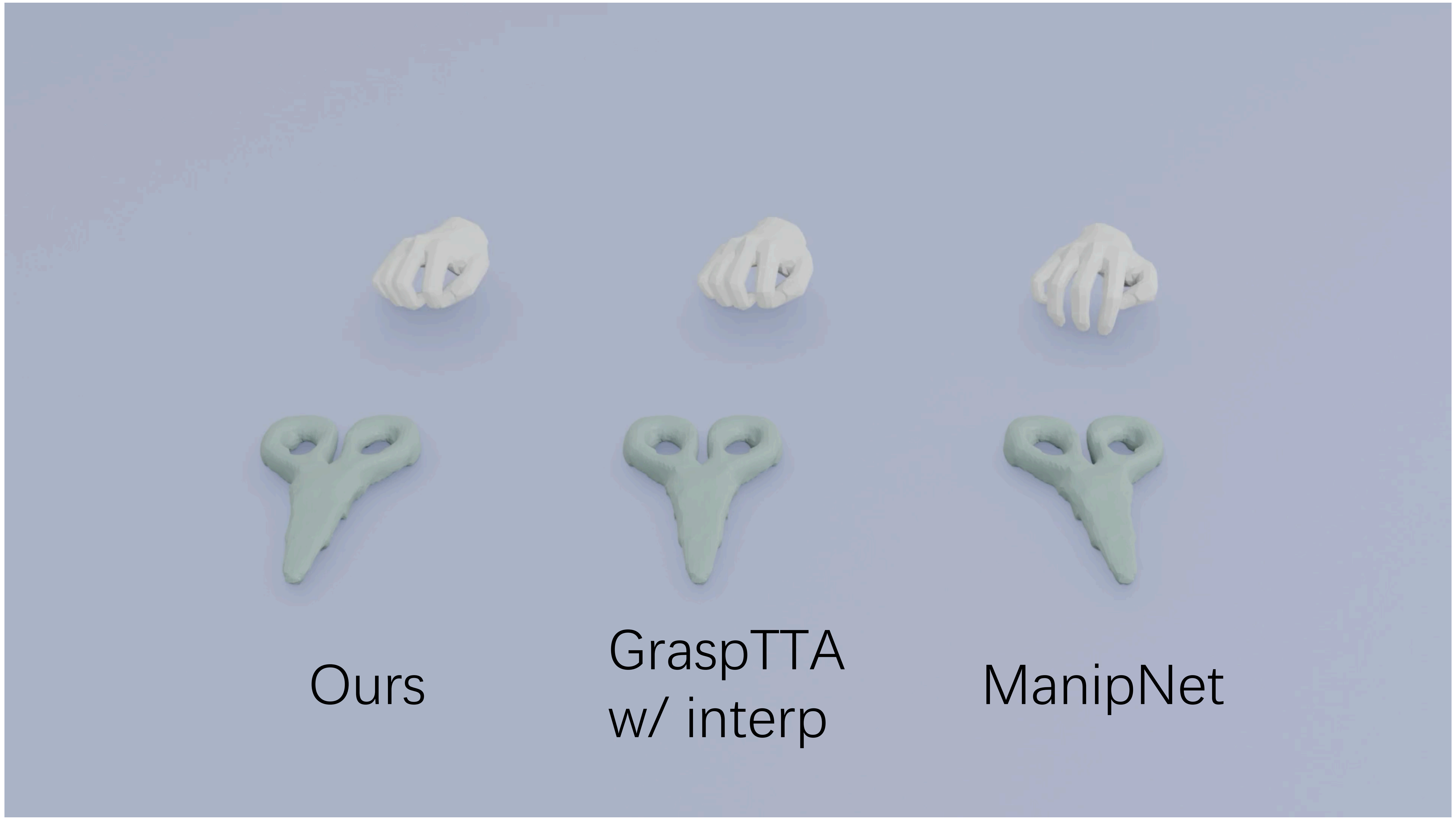
Contact-Centric Representation of Finger Motion



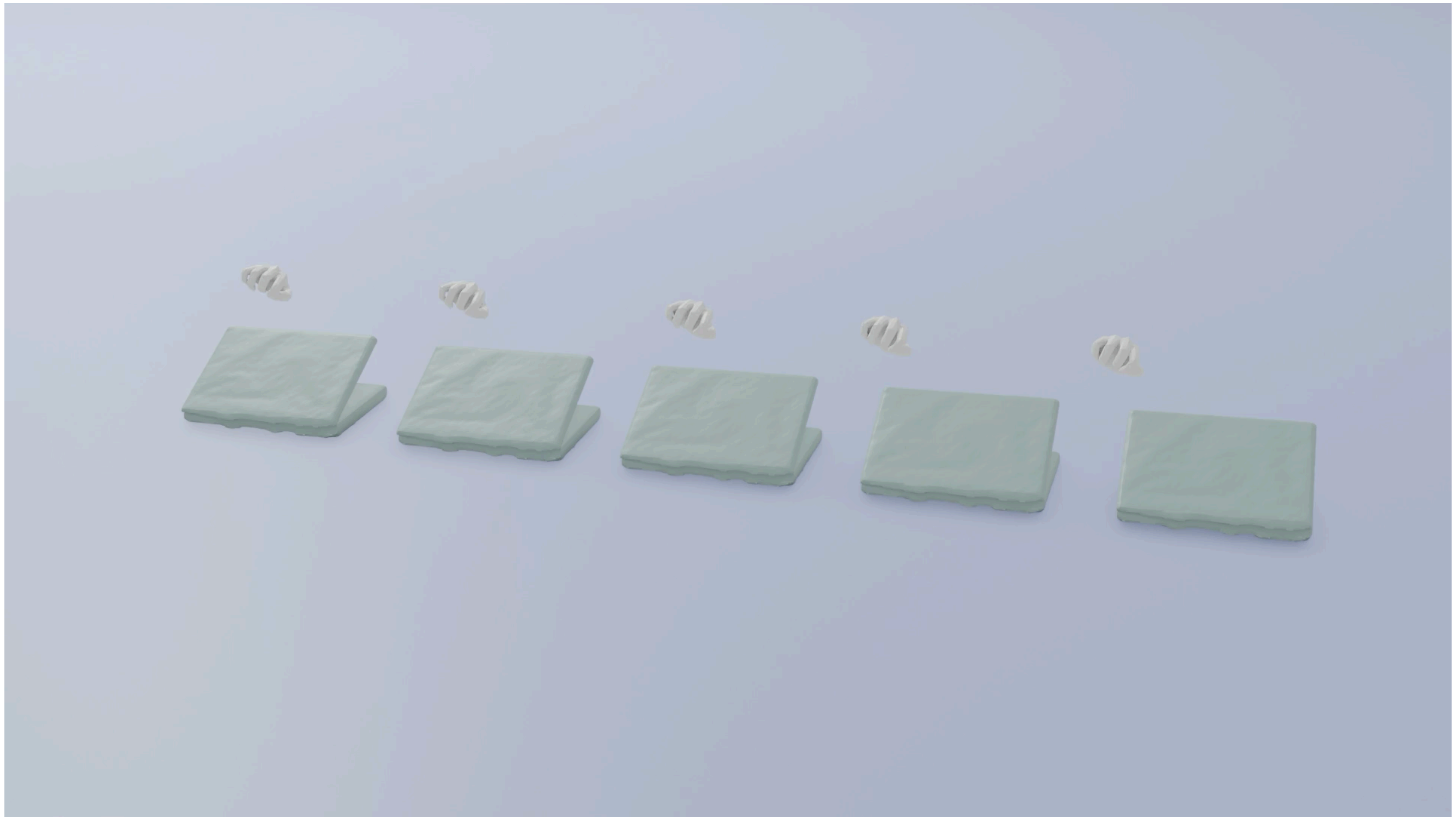
Overview of Motion Generation Framework



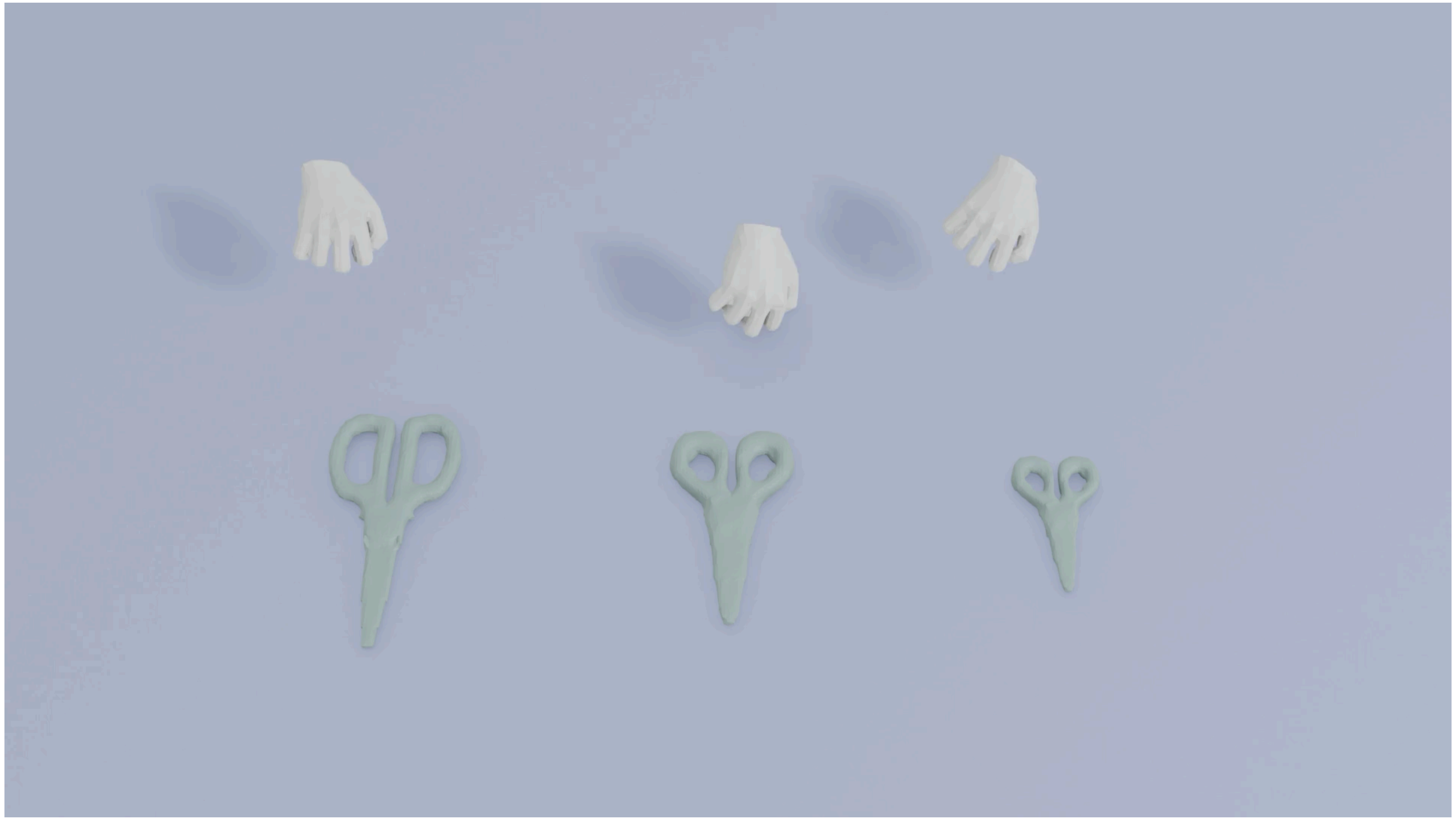
Comparison



Manipulation Diversity



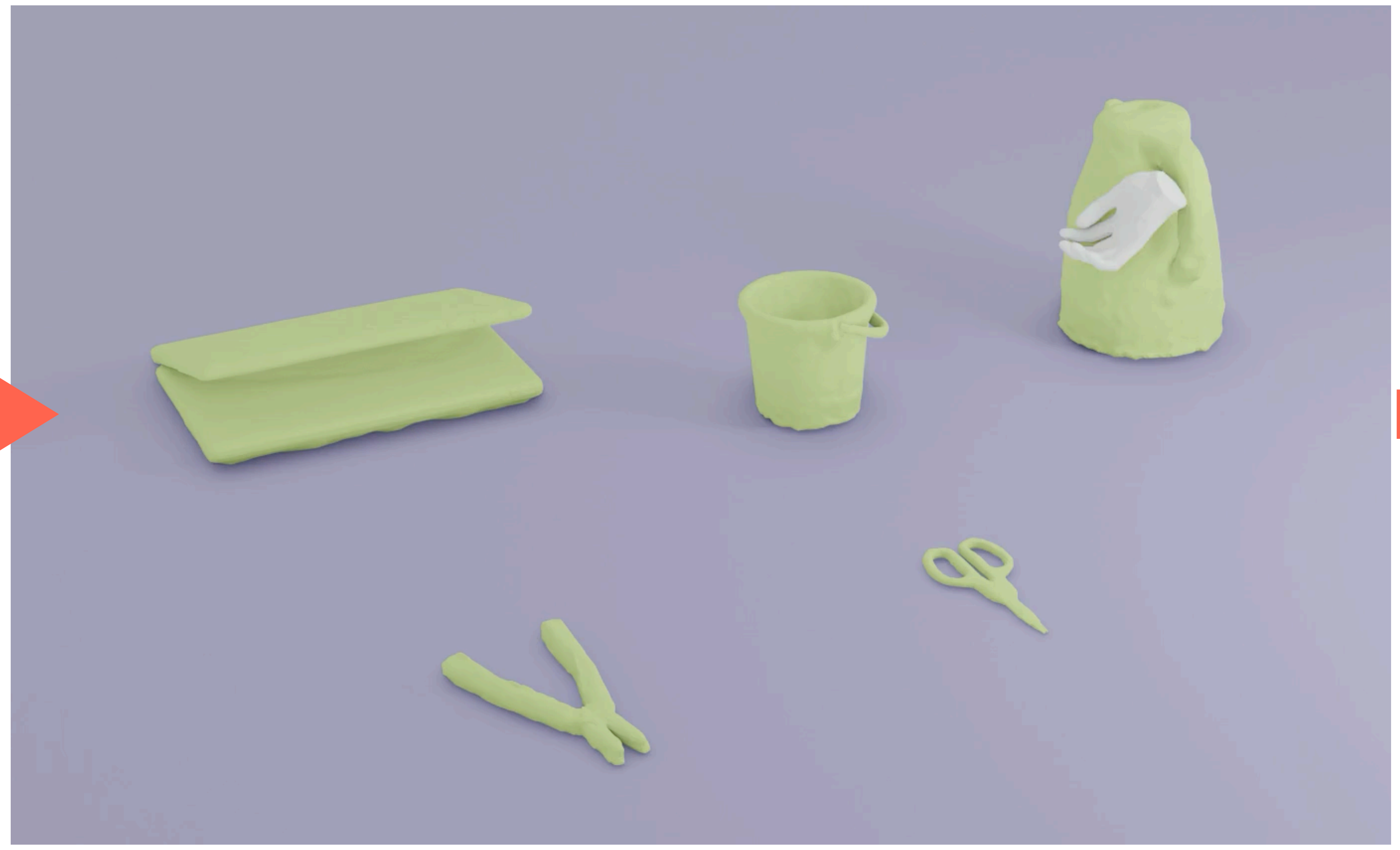
Robustness to Diverse Shapes



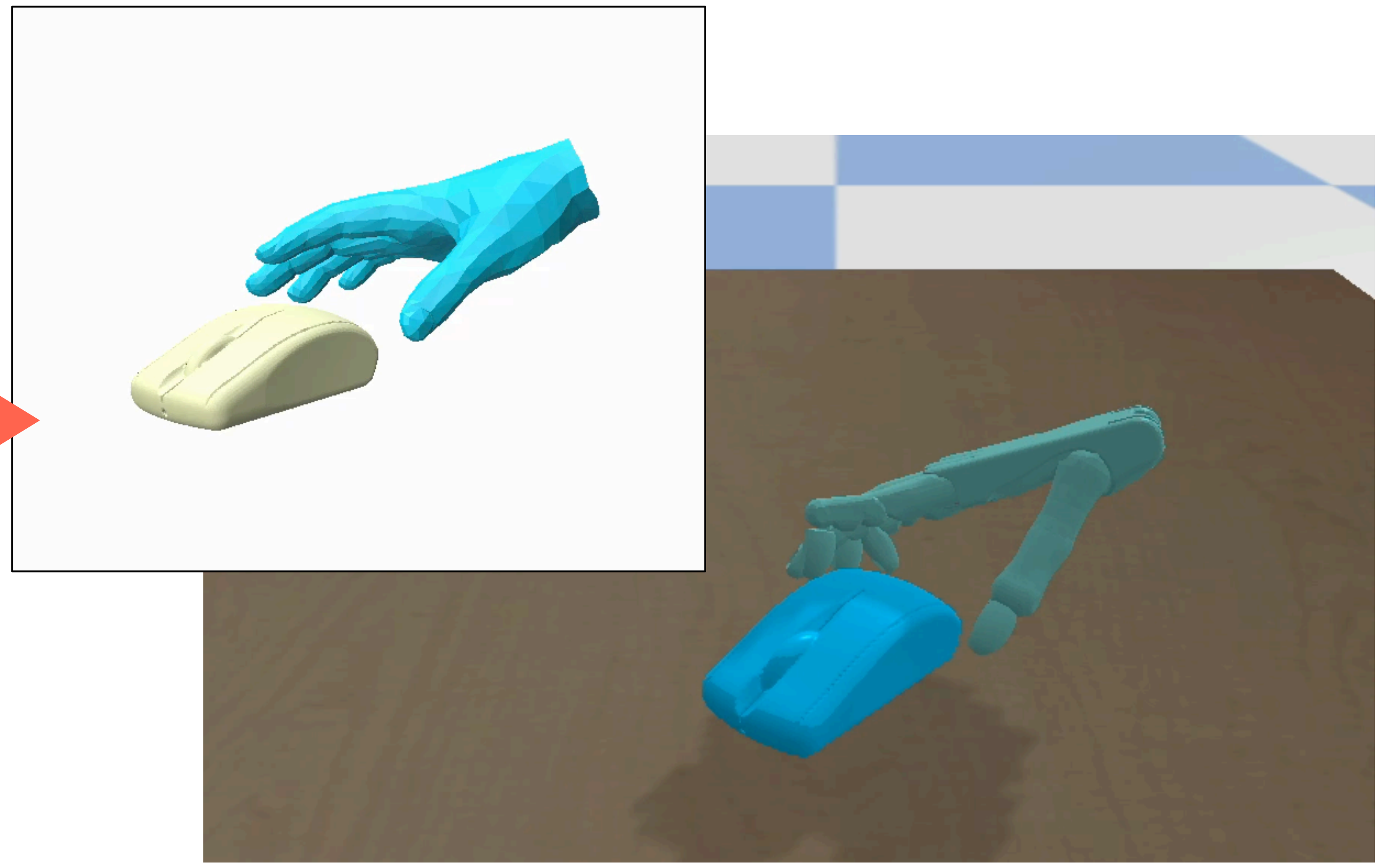
A Cross-Embodiment Tracking Control Paradigm

Capturing Human Manipulation Data

Generative Human Manipulation Planning

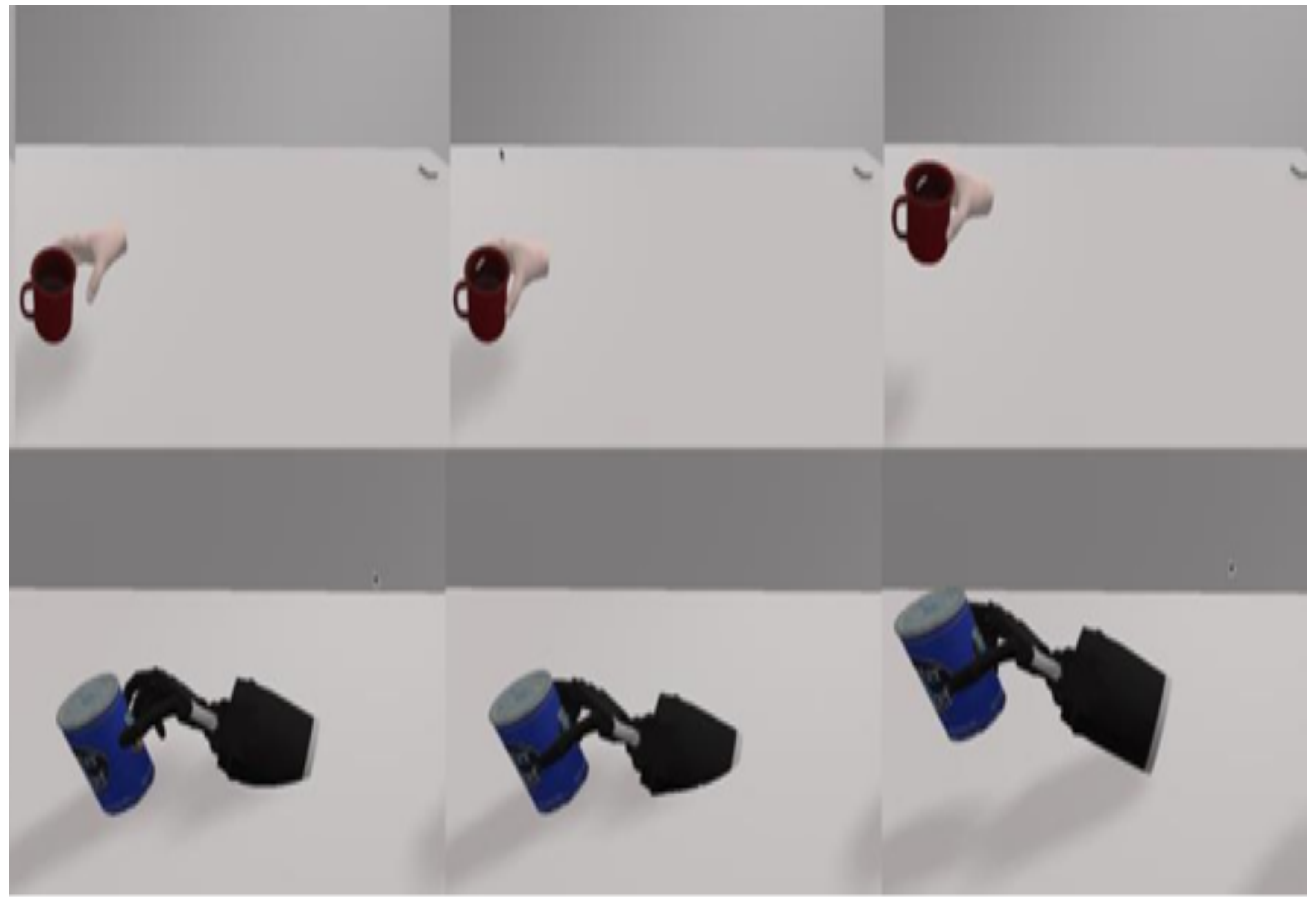


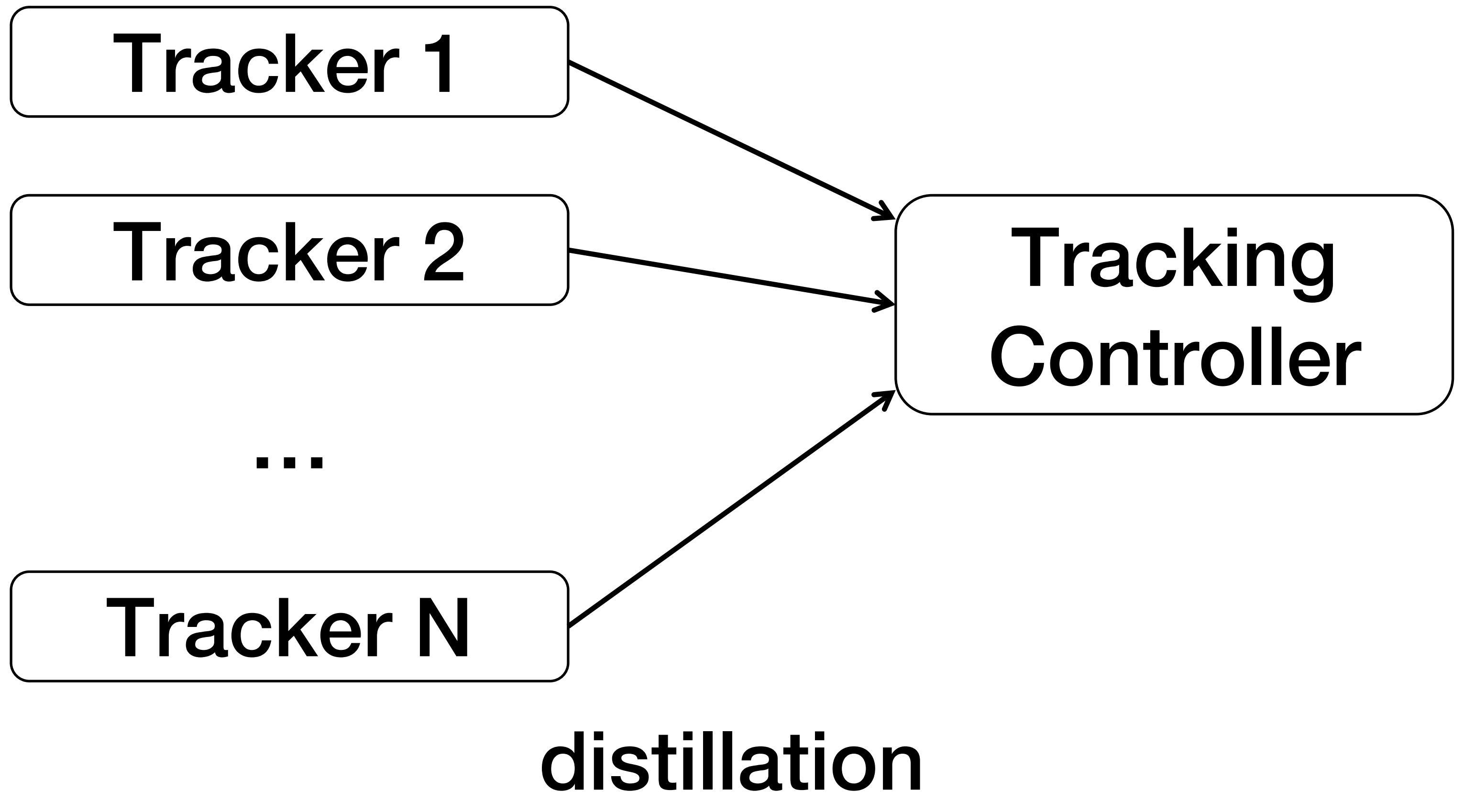
Cross-Embodiment Tracking Control



Cross-Embodiment Tracking Control

Motion Retargeting → Learning a Per-Trajectory Tracker → Learning a Tracking Controller

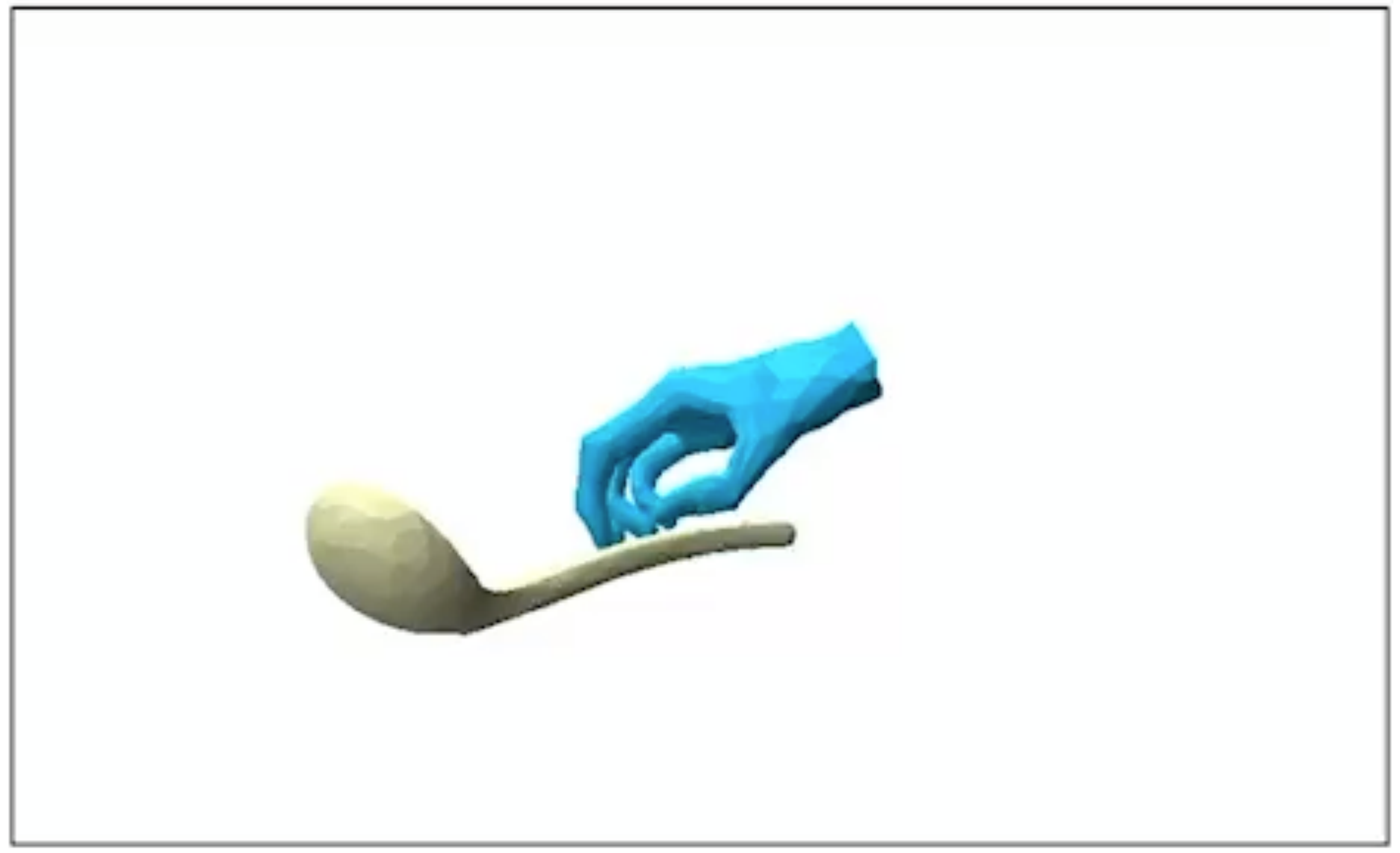




QuasiSim: Parameterized Quasi-Physical Simulators for Dexterous Manipulations Transfer
Xueyi Liu, Kangbo Lyu, Jieqiong Zhang, Tao Du, Li Yi. ECCV 2024
Kinematics-Only Human Demonstration
Dexterous Manipulations Transferred to a Simulated Robot Hand by Our Method

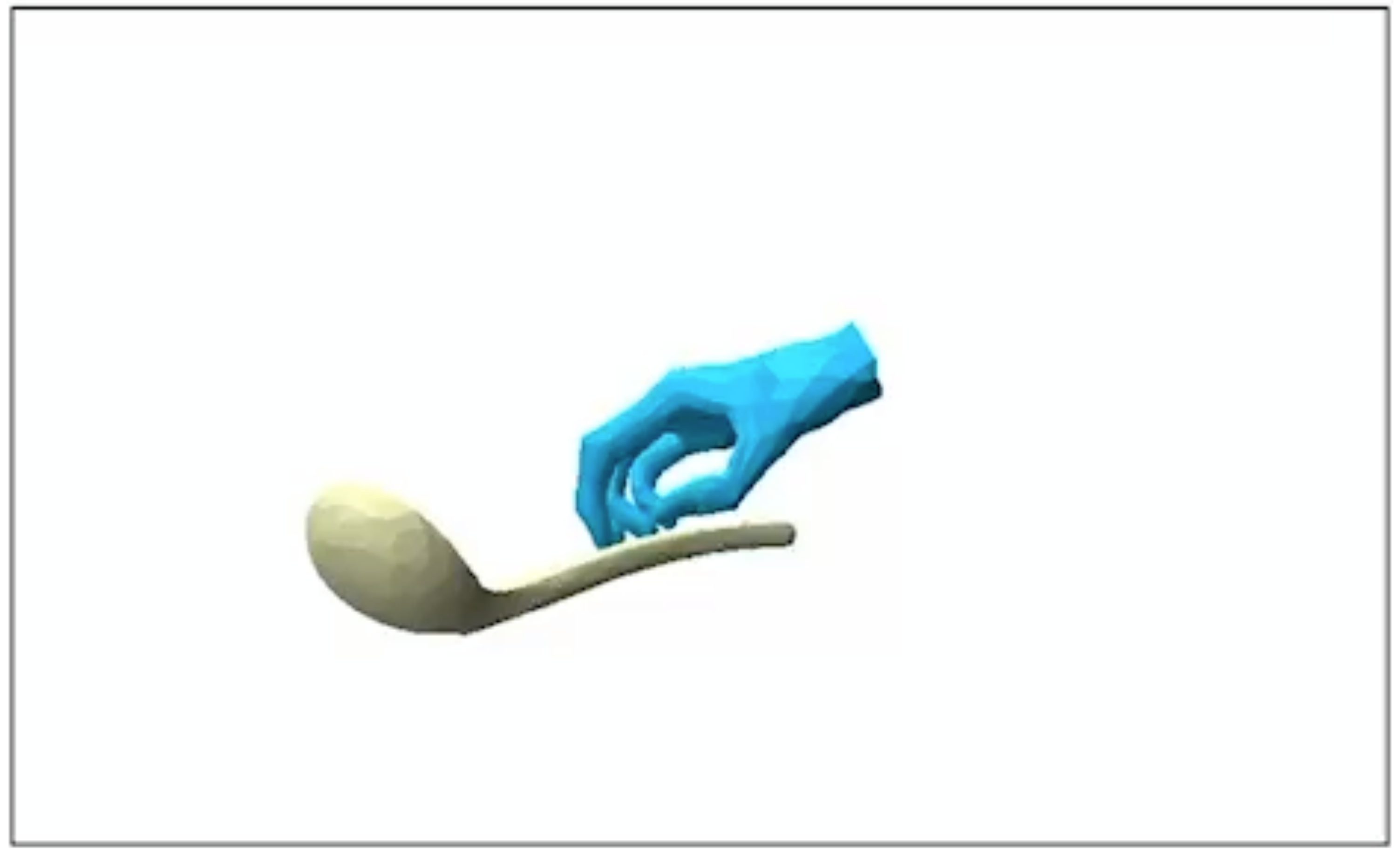
Tracking a Single Trajectory

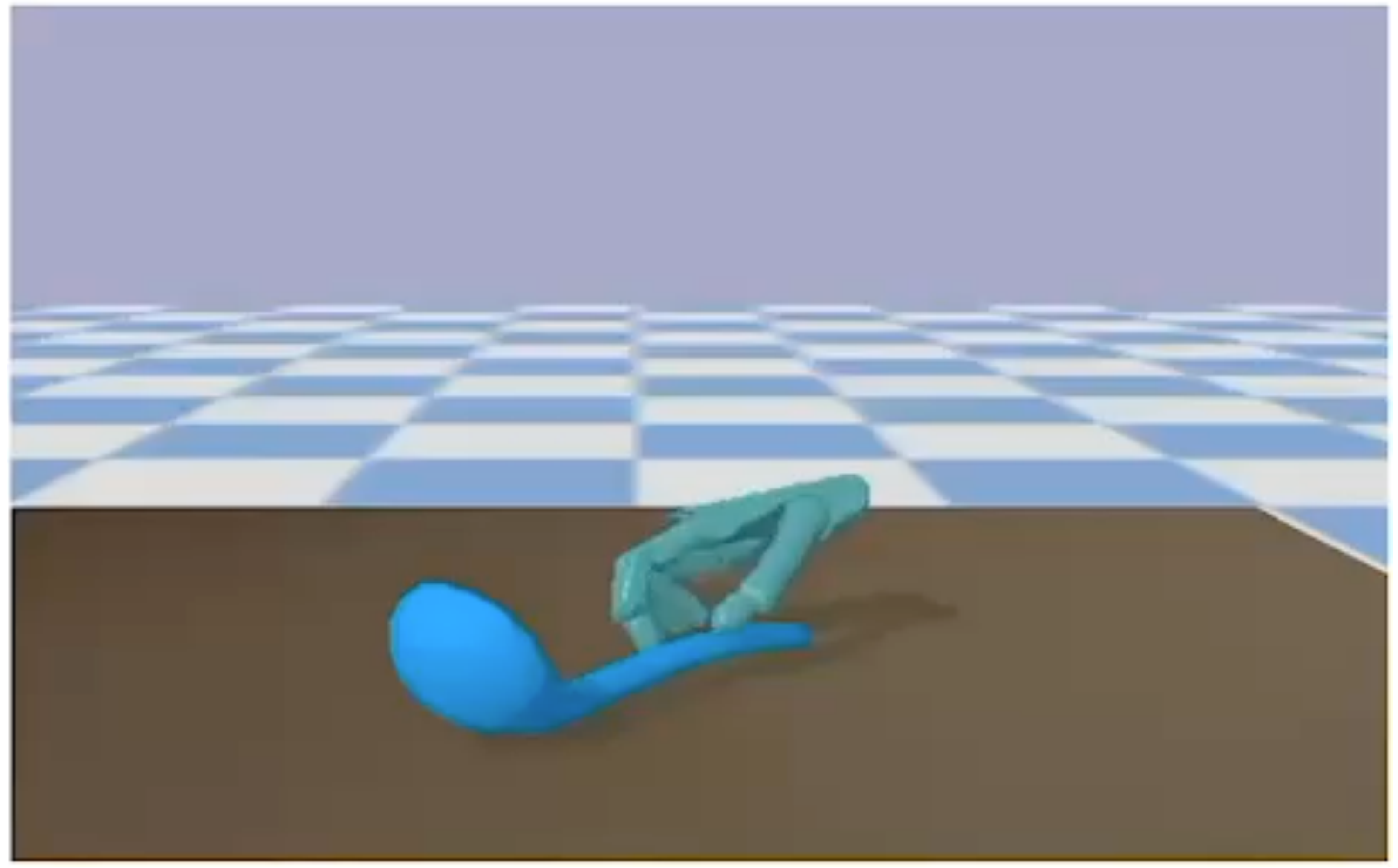
- Problem setup:
  - Input: a motion reference $\{s_0, s_1, \ldots, s_n\}$ describing a human hand manipulating an object



Tracking a Single Trajectory

- Problem setup:
  - Input: a motion reference $\{s_0, s_1, \dots, s_n\}$ describing a human hand manipulating an object
  - Output: a dynamic sequence $\{\hat{s}_0, \hat{a}_0, \hat{s}_1, \hat{a}_1, \dots, \hat{s}_n\}$ transferring the skill to a robotic dexterous hand





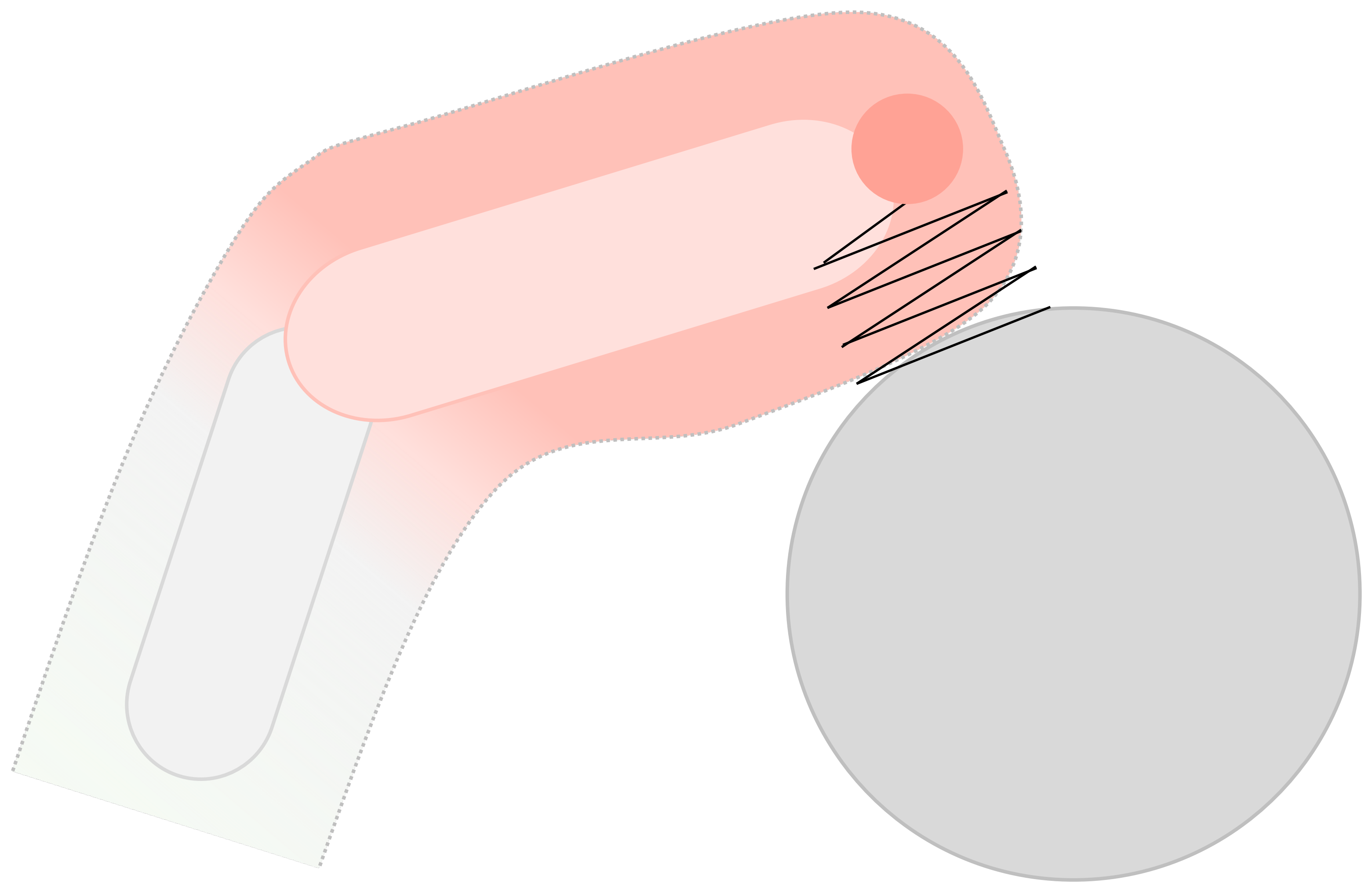
Tracking a Single Trajectory

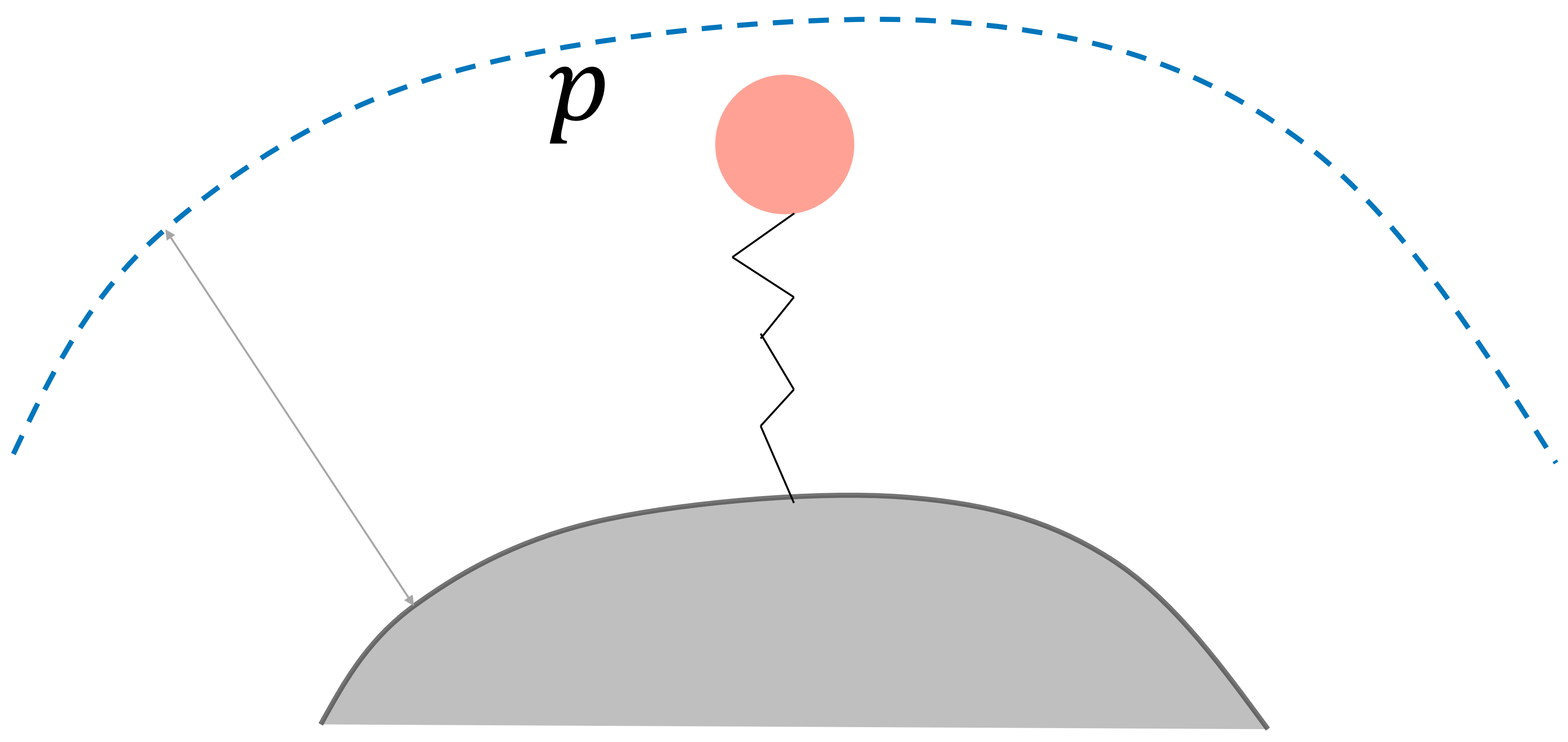
- Tough dynamics challenge trajectory optimization or RL
- Instead of focusing on optimization algorithms, can we optimize the simulator design?
- More generally: how to optimize simulation strategy for robot learning?

Optimizing Physical Simulation

- Relaxed physical constraints help optimization via smoothing out the optimization objective

- High fidelity physics is critical for sim-to-sim or sim-to-real transfer





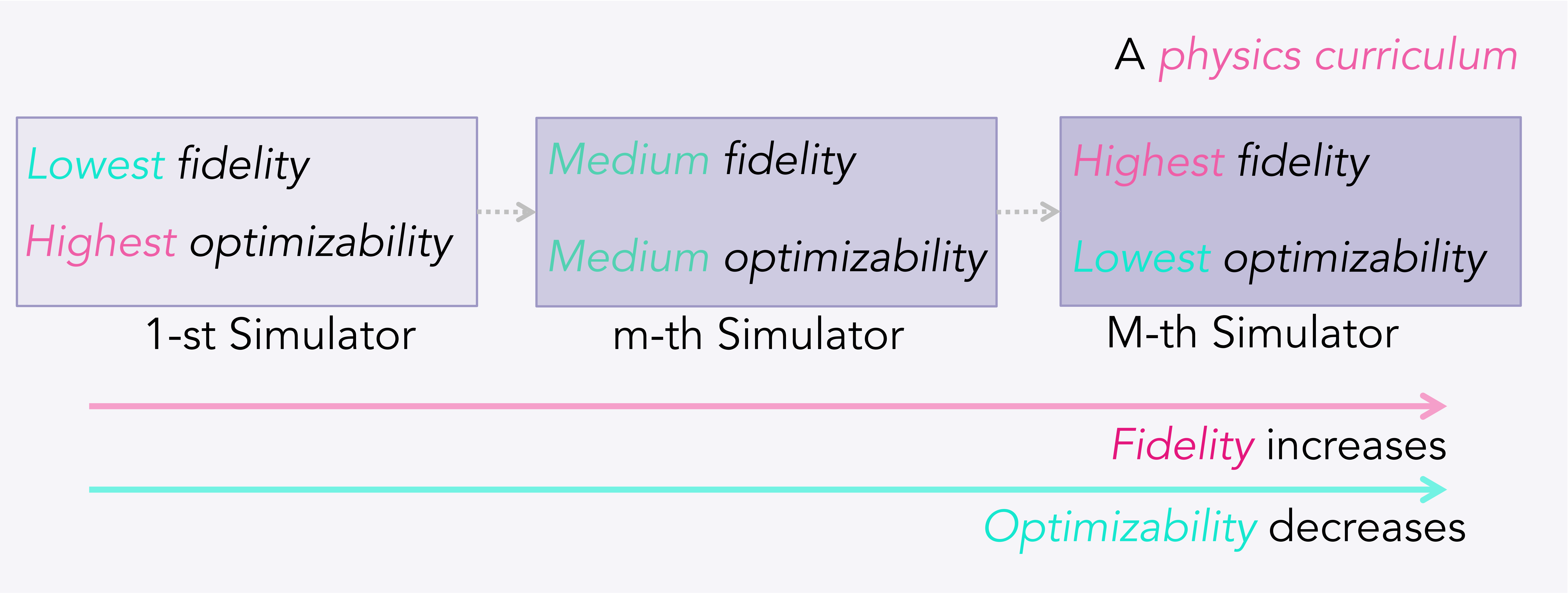
How to Benefit from Both?

- Relaxed physical constraints help optimization via smoothing out the optimization objective
- High fidelity physics is critical for sim-to-sim or sim-to-real transfer

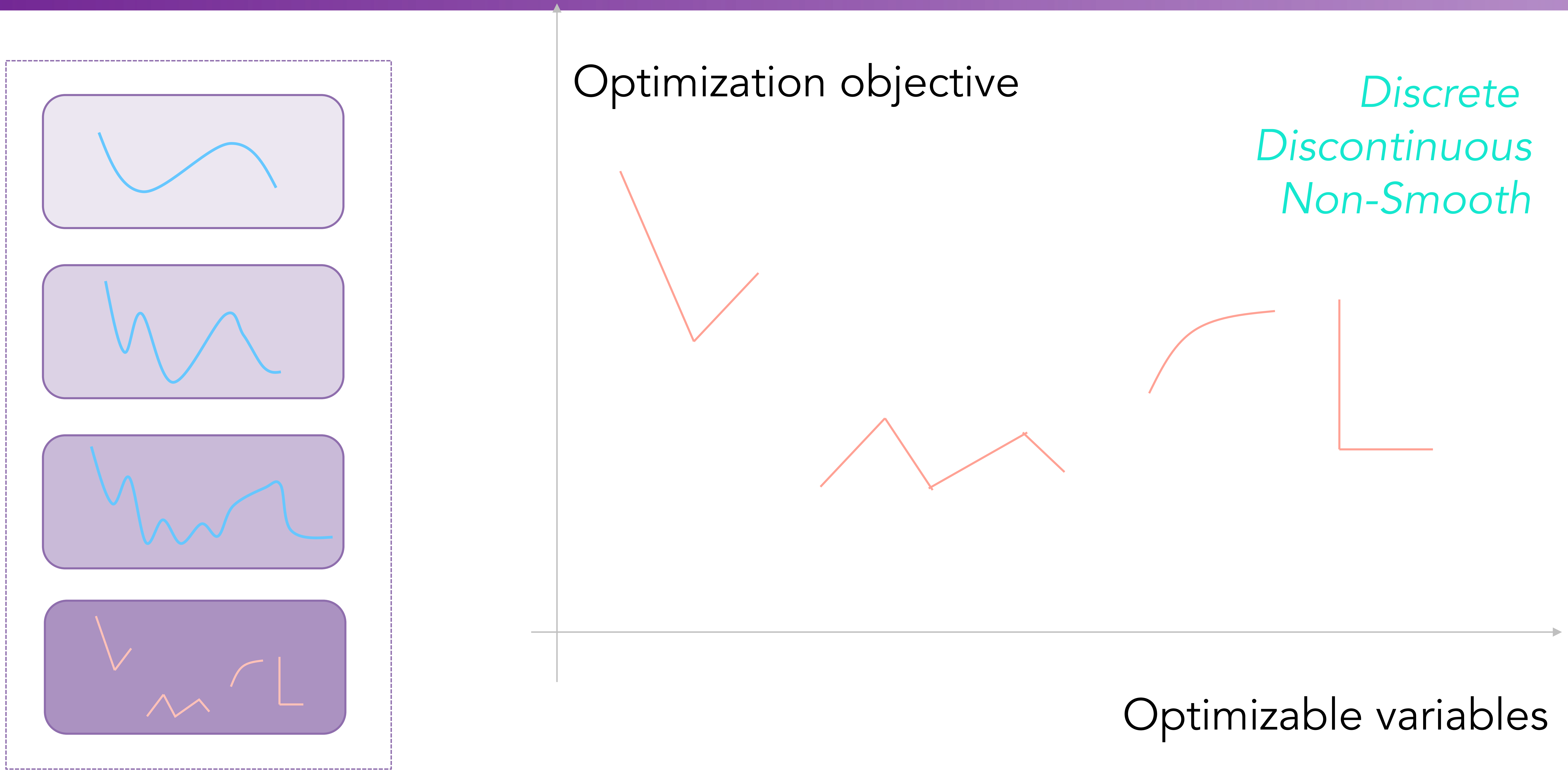
↓

Using both in a physics curriculum!

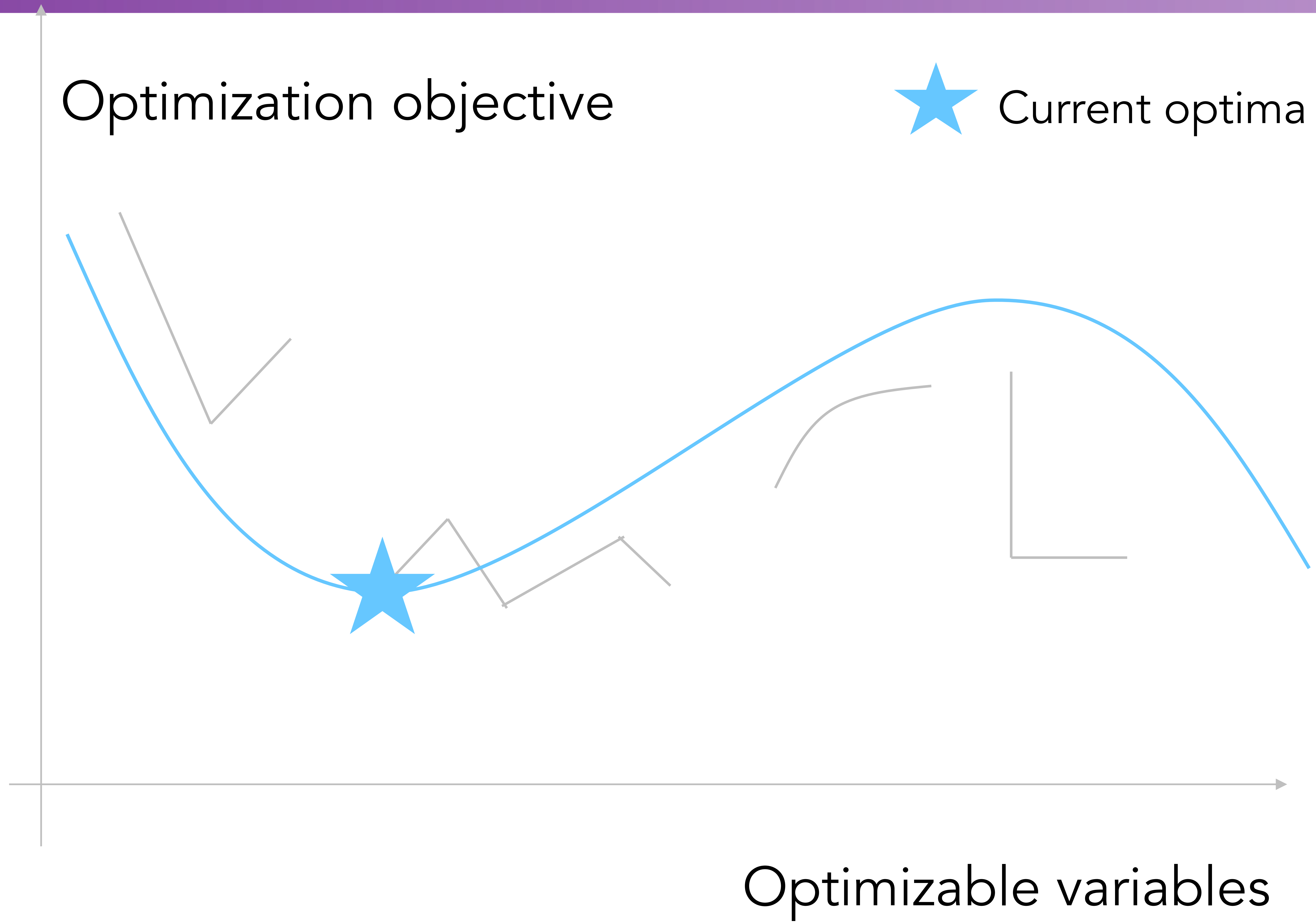
A Physics Curriculum



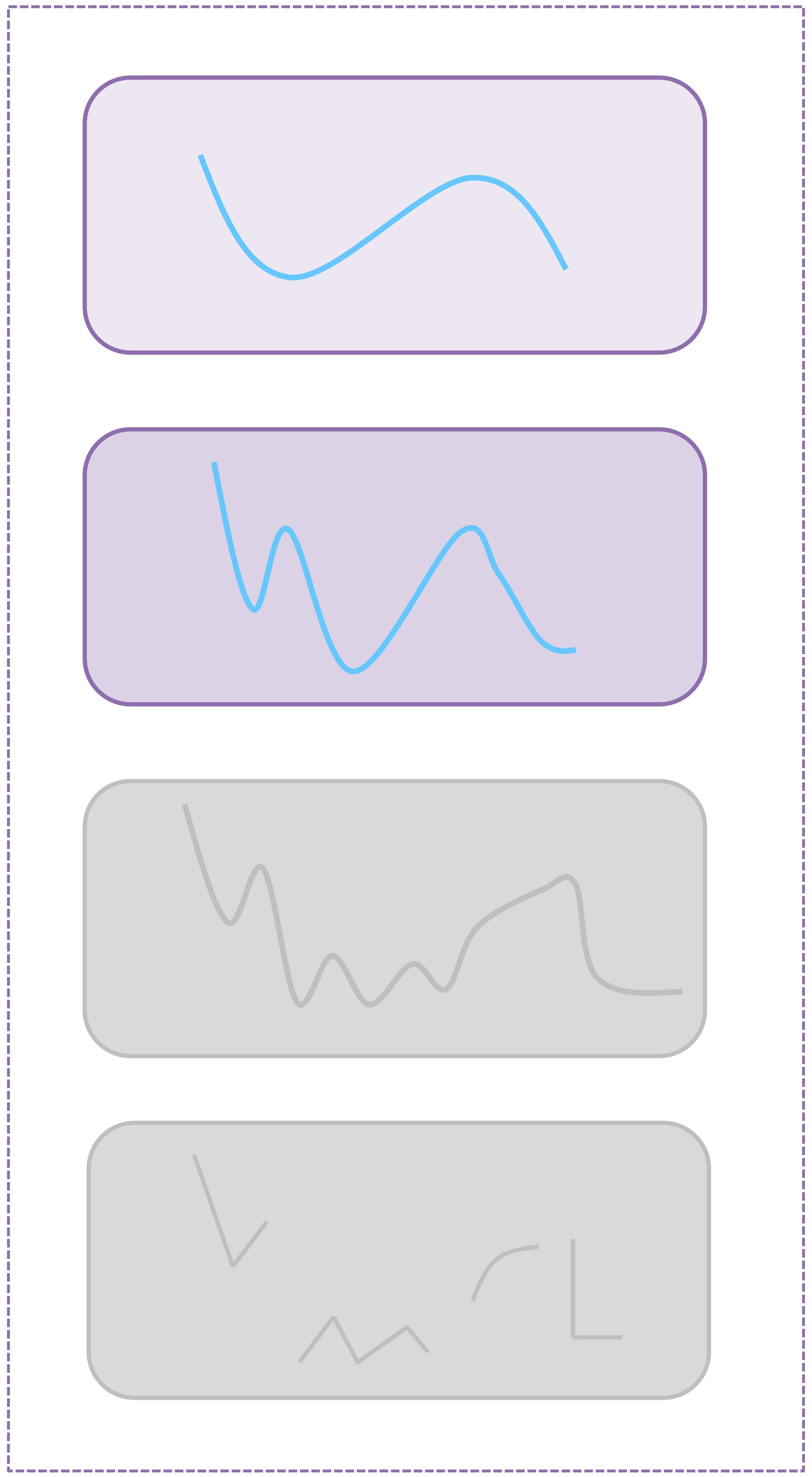
Key Idea: Optimizing through a physics curriculum

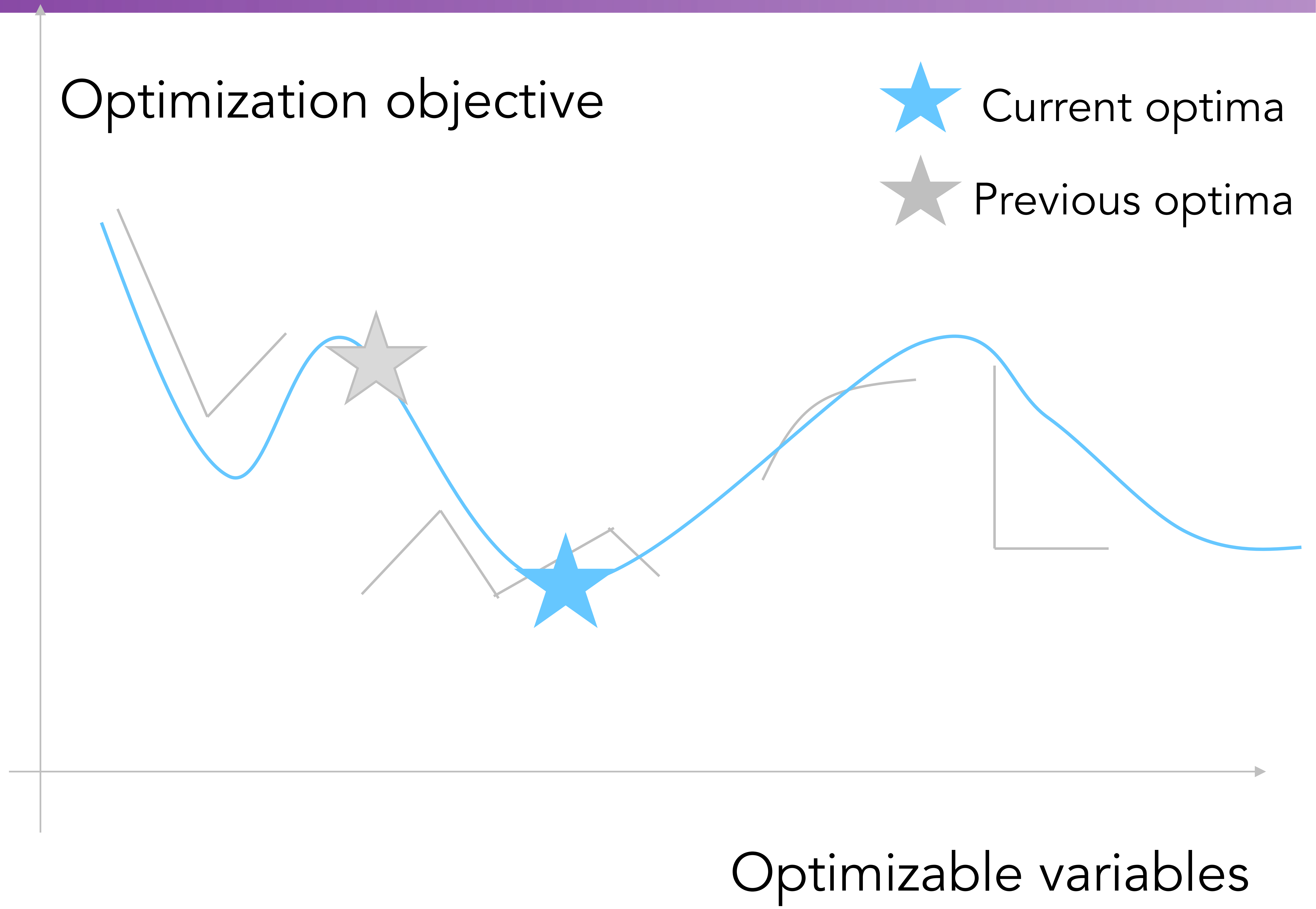


Key Idea: Optimizing through a physics curriculum

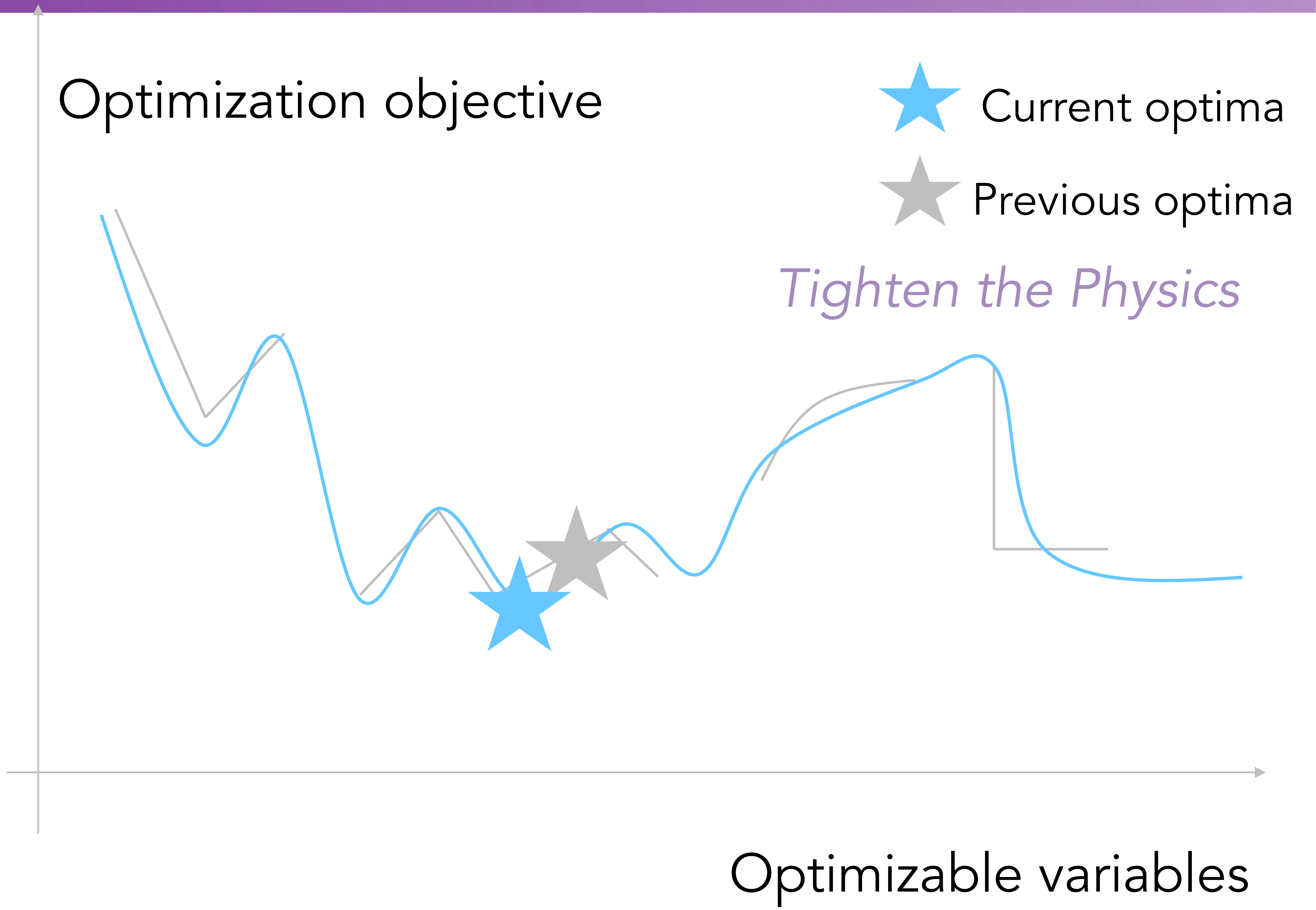


Key Idea: Optimizing through a physics curriculum

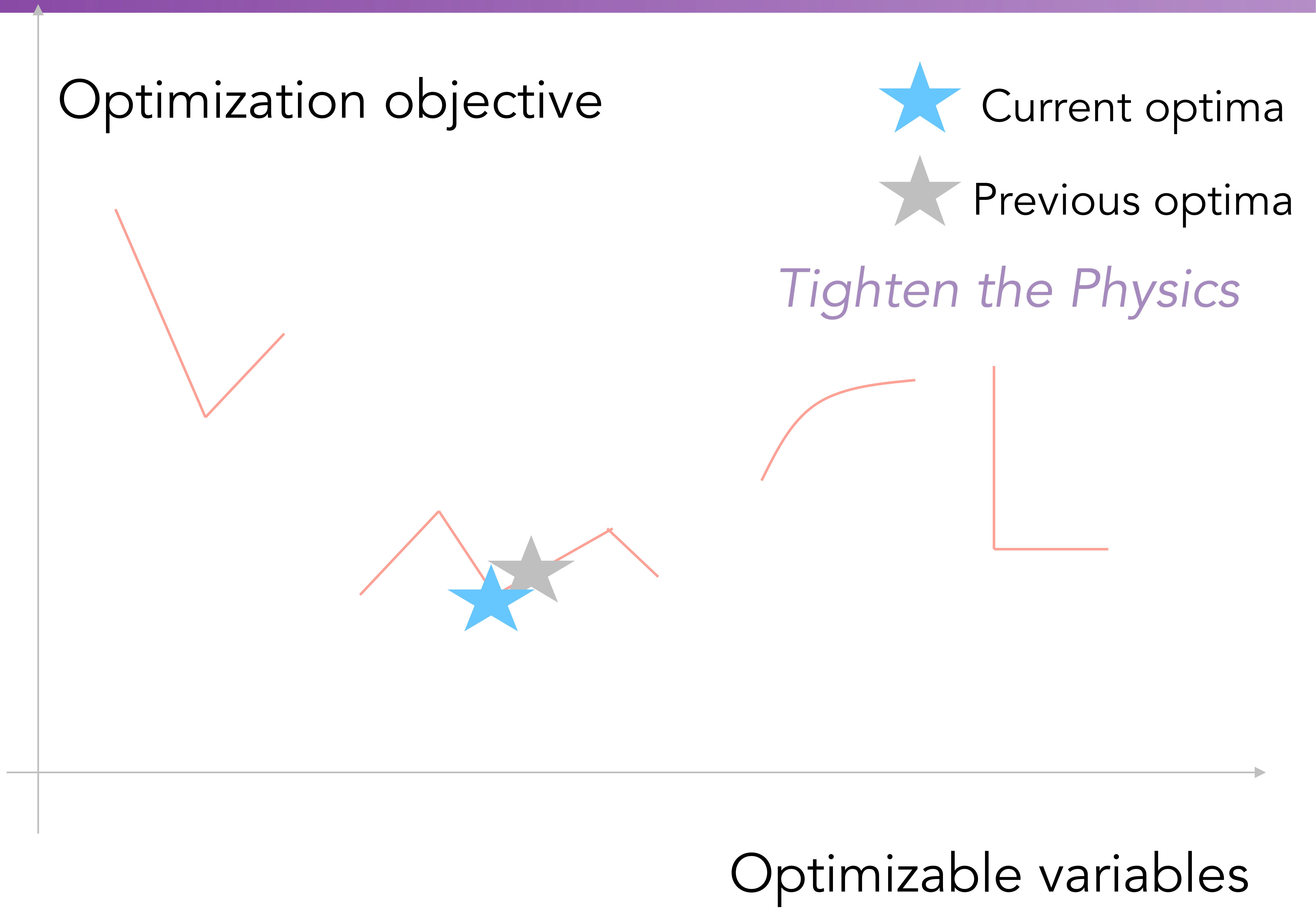




Key Idea: Optimizing through a physics curriculum



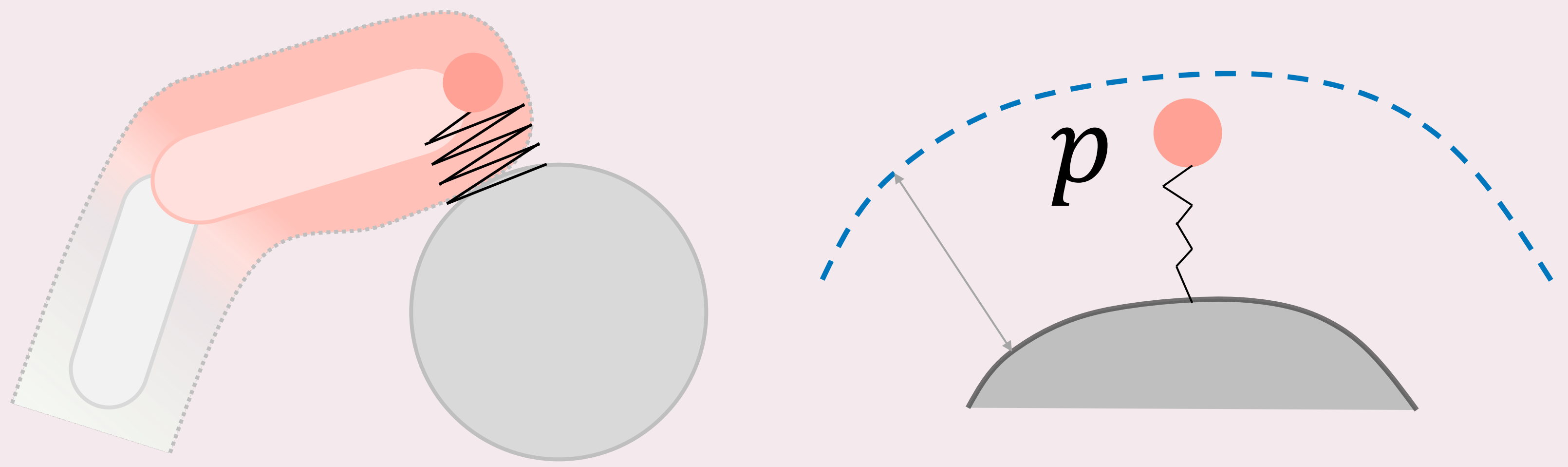
Key Idea: Optimizing through a physics curriculum



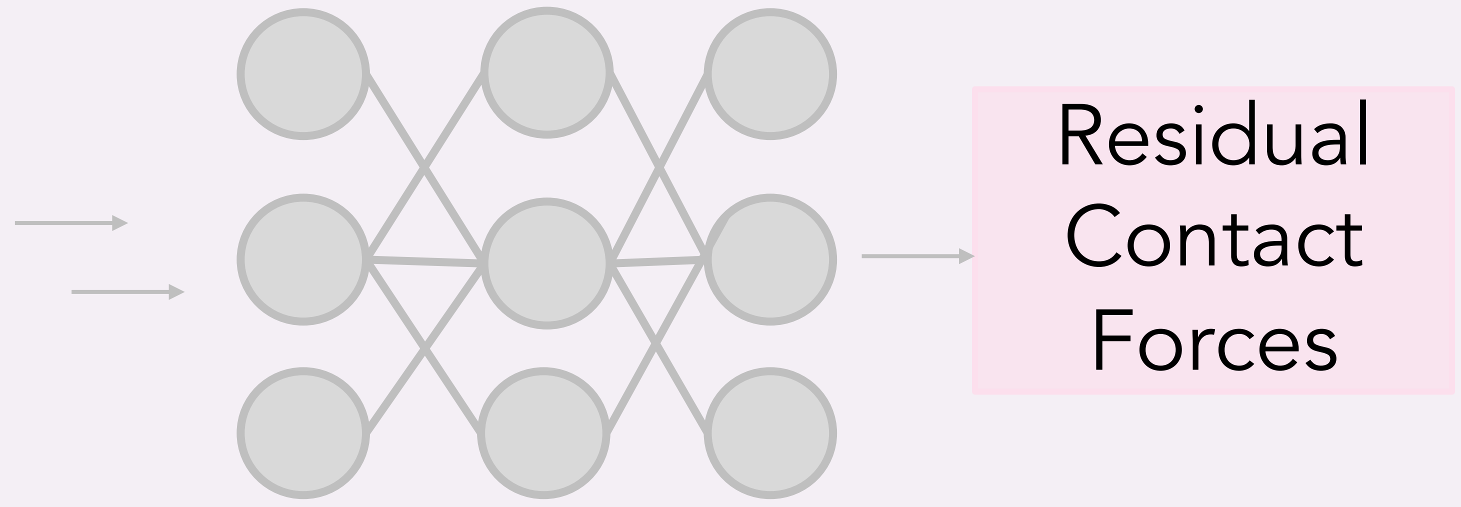
Parameterized Quasi-Physical Simulator

Parameterized *Quasi-Physical Simulator*

Parameterized
Analytical Simulator



Parameterized
Residual Physics

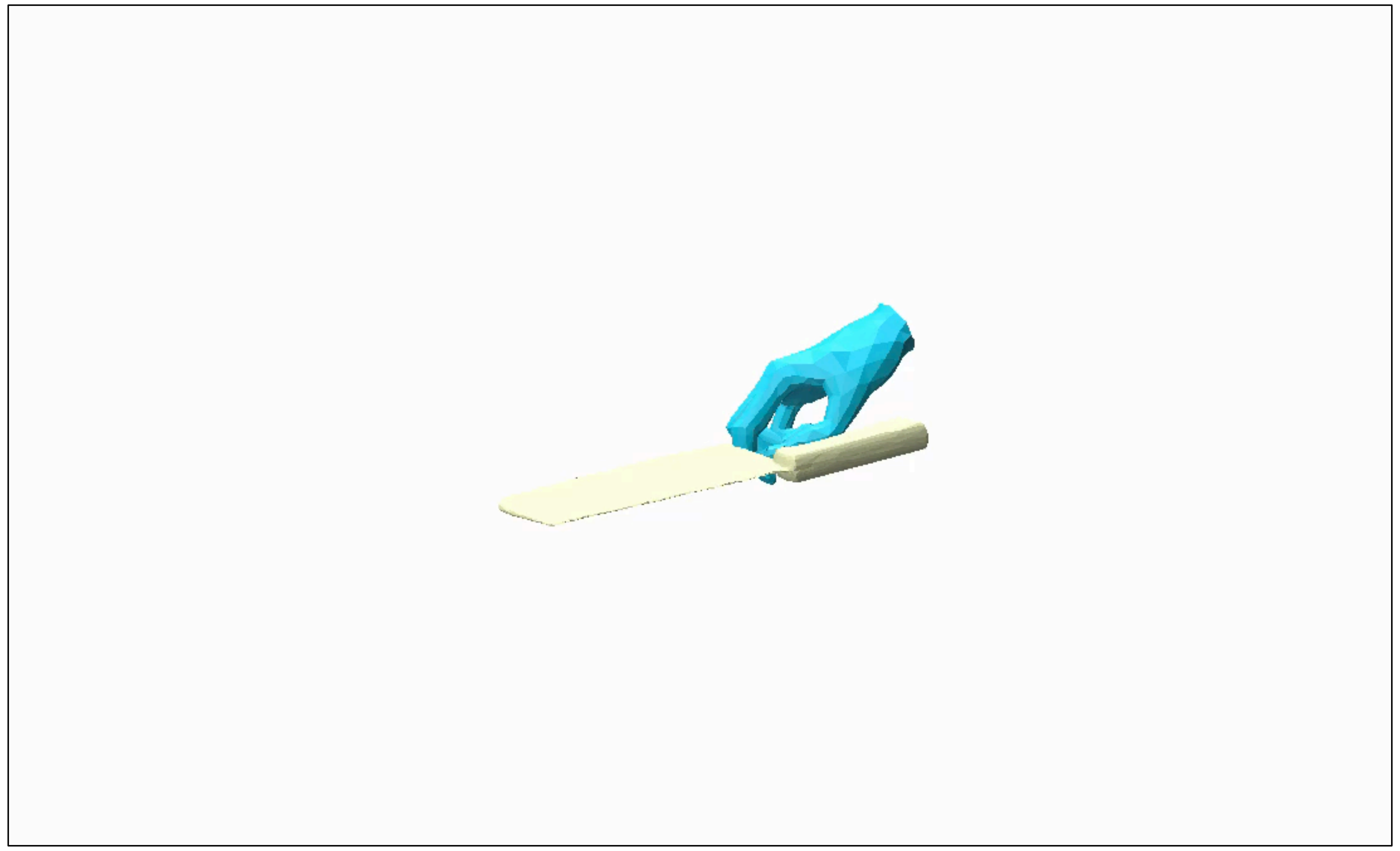


Residual Force Network

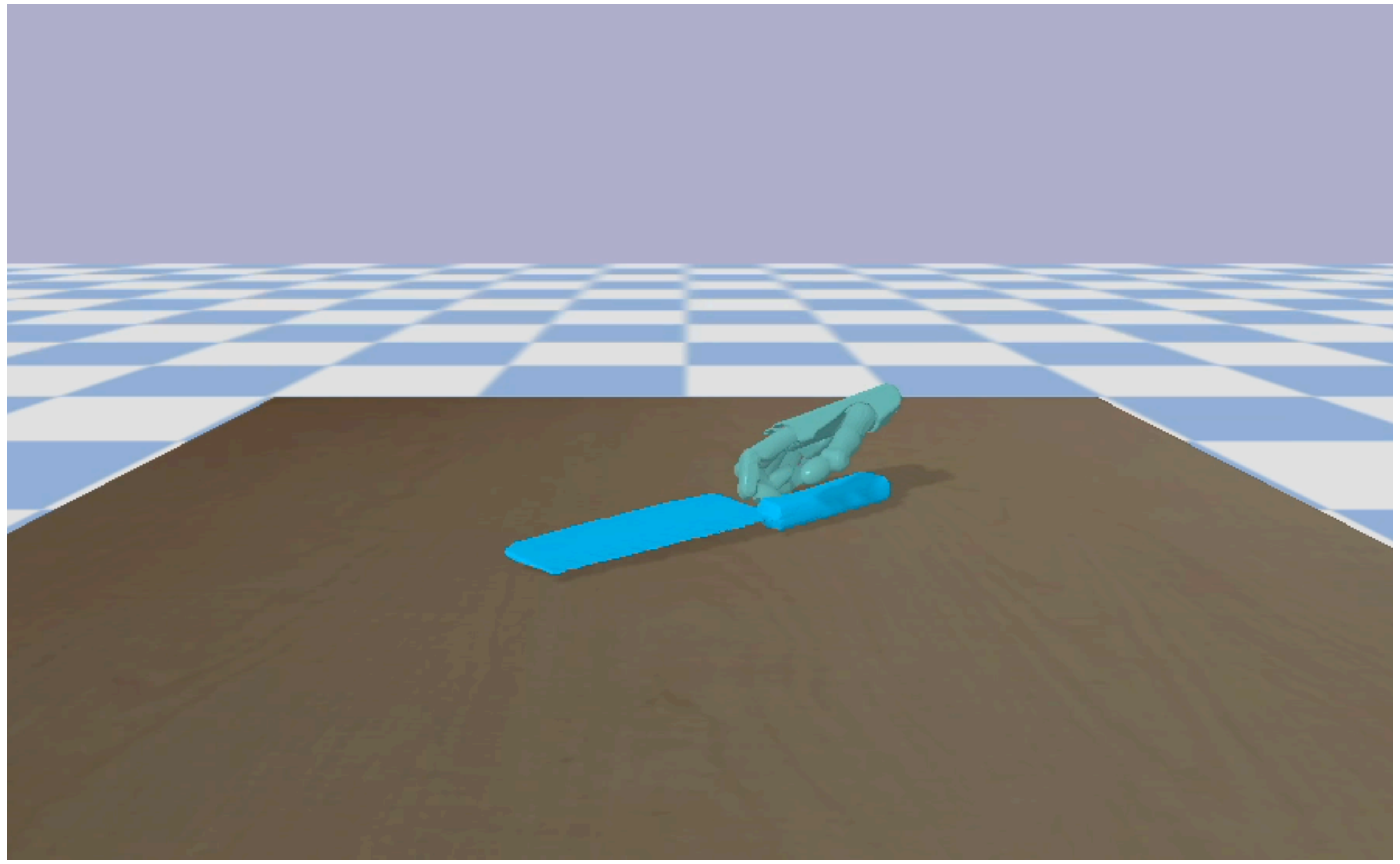
Controllable analytical relaxations on
Articulated multi-rigid constraints
Contact constraints

Flexible Neural networks for
Approximating ***high-fidelity physics***

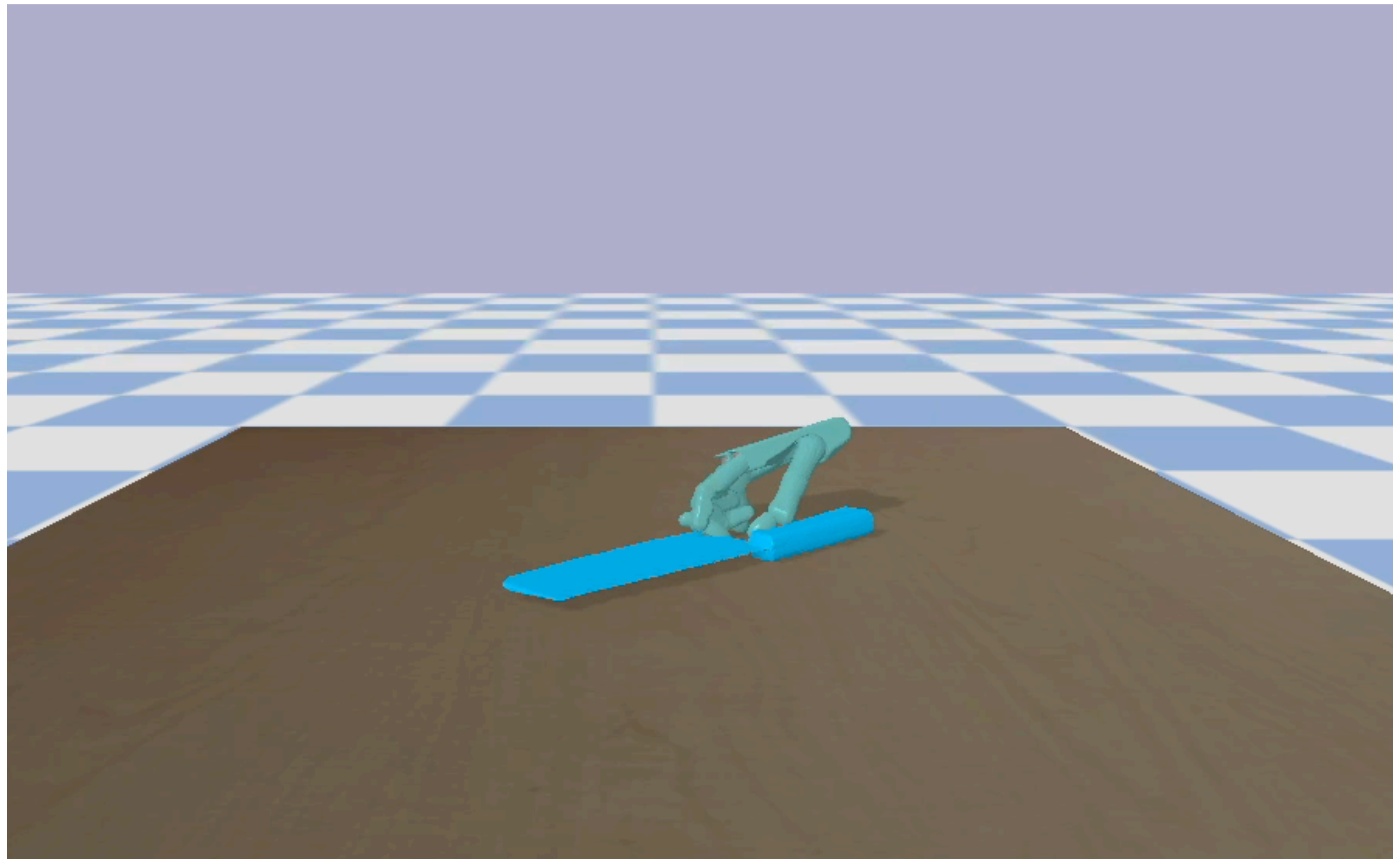
Experimental Results



Human Demonstration



Ours

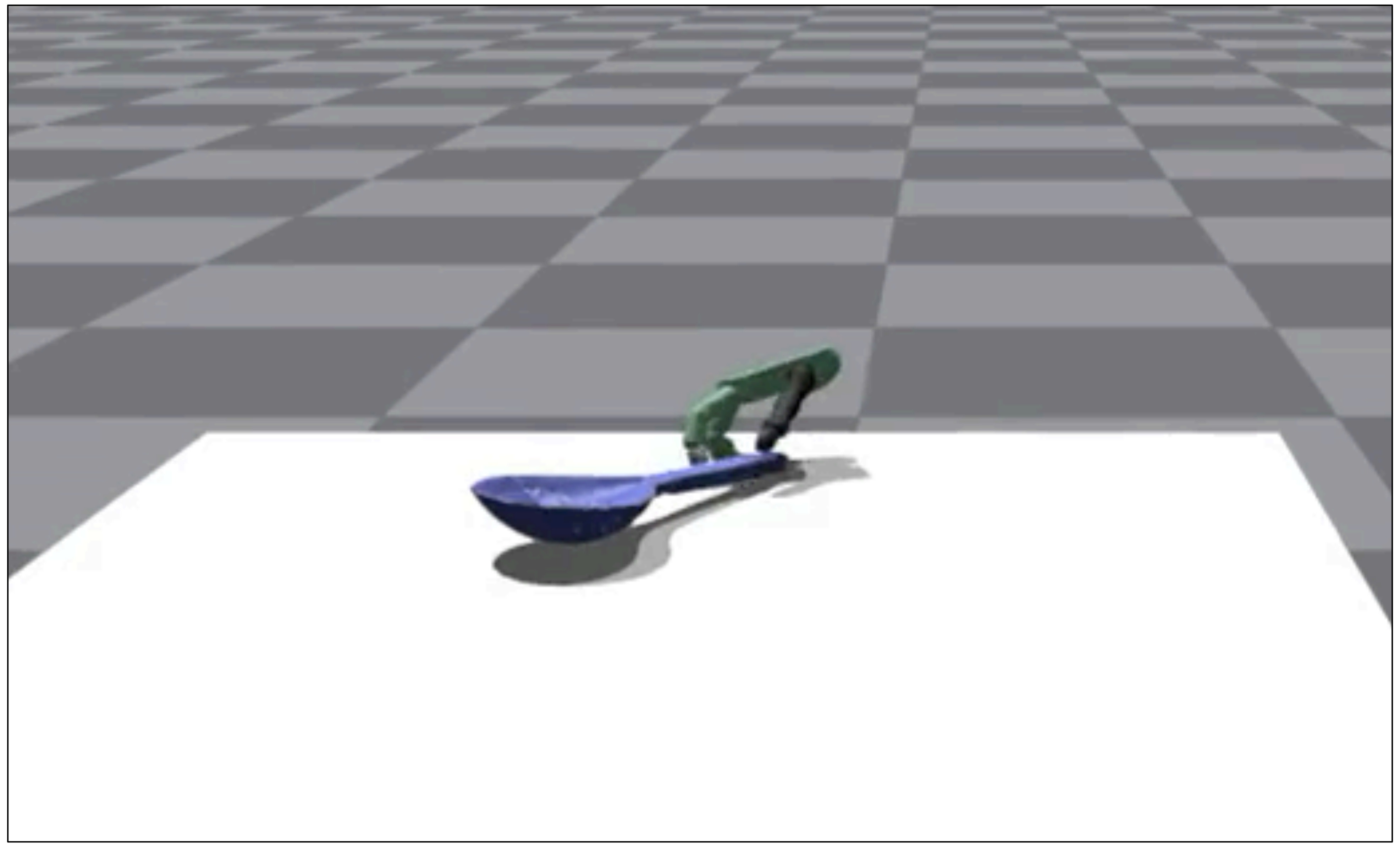


Baseline

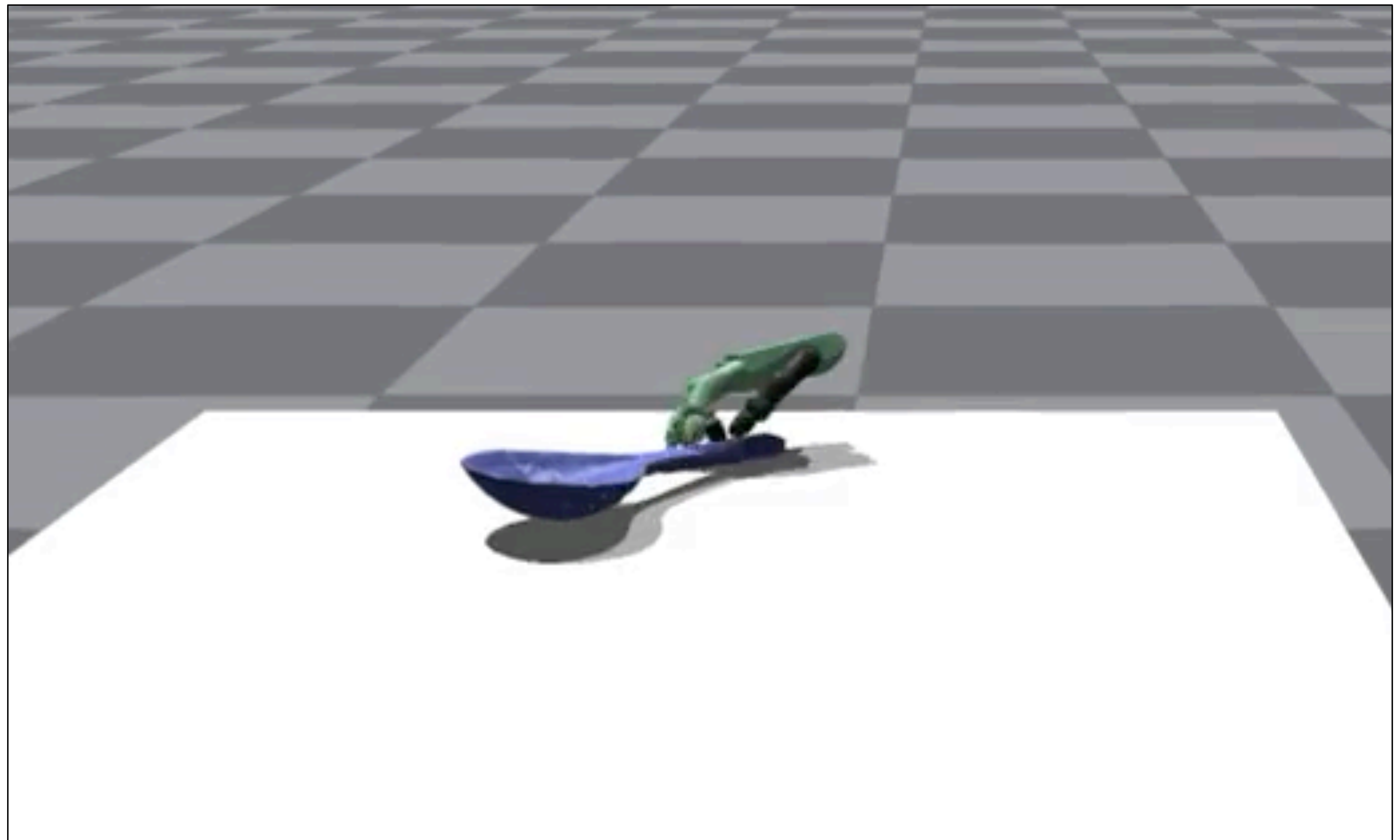
[Bullet]

Experimental Results

Human Demonstration



Ours

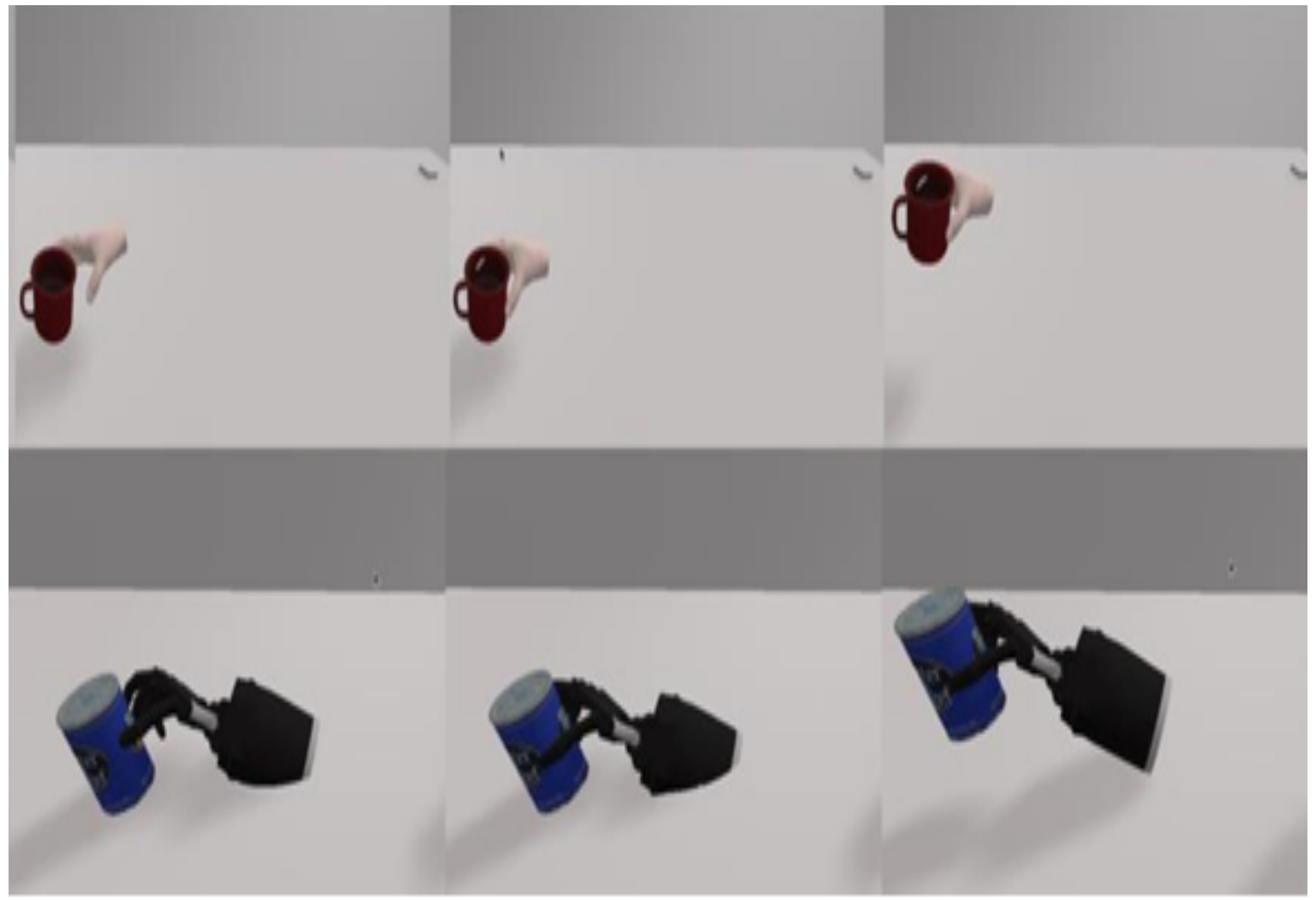


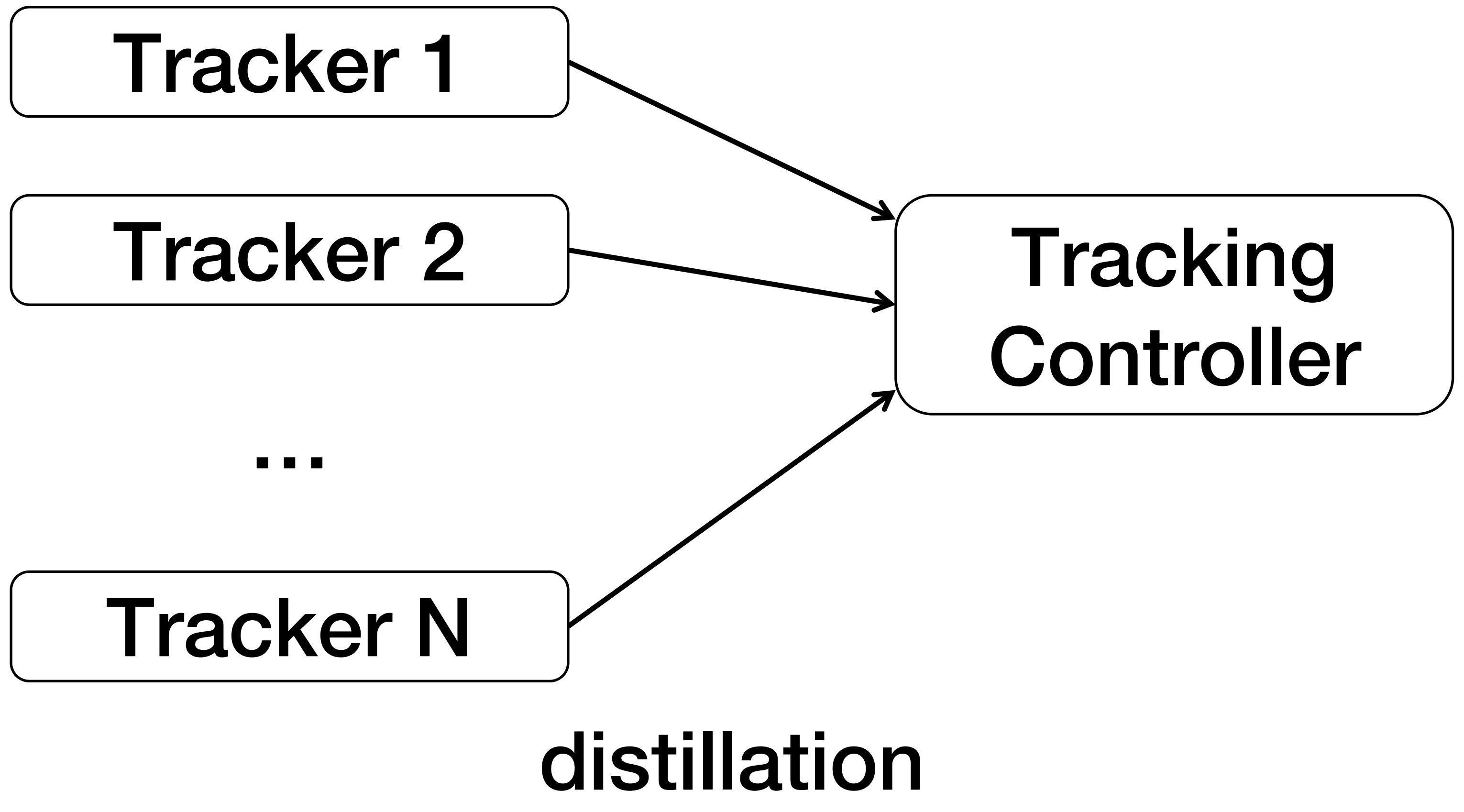
Baseline

[Isaac Gym]

Cross-Embodiment Tracking Control

Motion Retargeting → Learning a Per-Trajectory Tracker → Learning a Tracking Controller





Towards Generalizable Neural Tracking Control for Dexterous Manipulation from Human References

Xueyi Liu, Jianibieke Adalibieke, Qianwei Han, Yuzhe Qin, Li Yi. In submission.

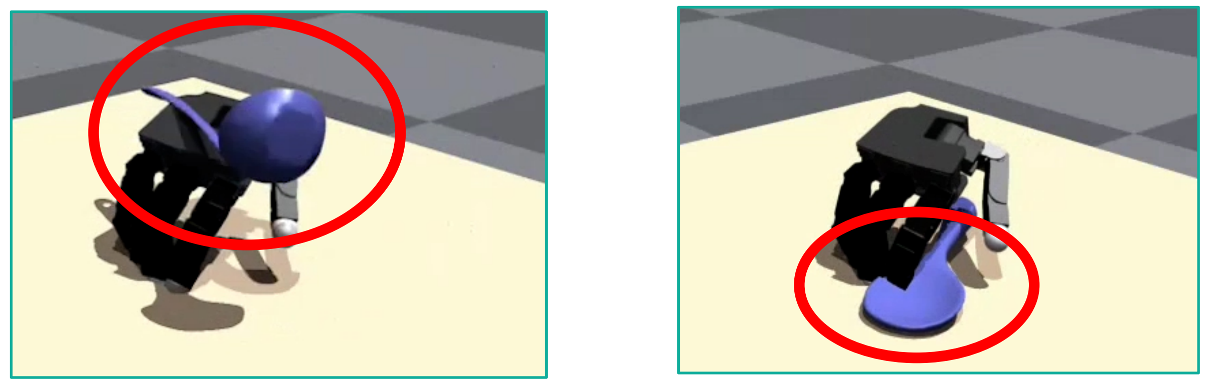
A *Generalizable* Neural Tracking Controller

Tracking Controller

Generalize

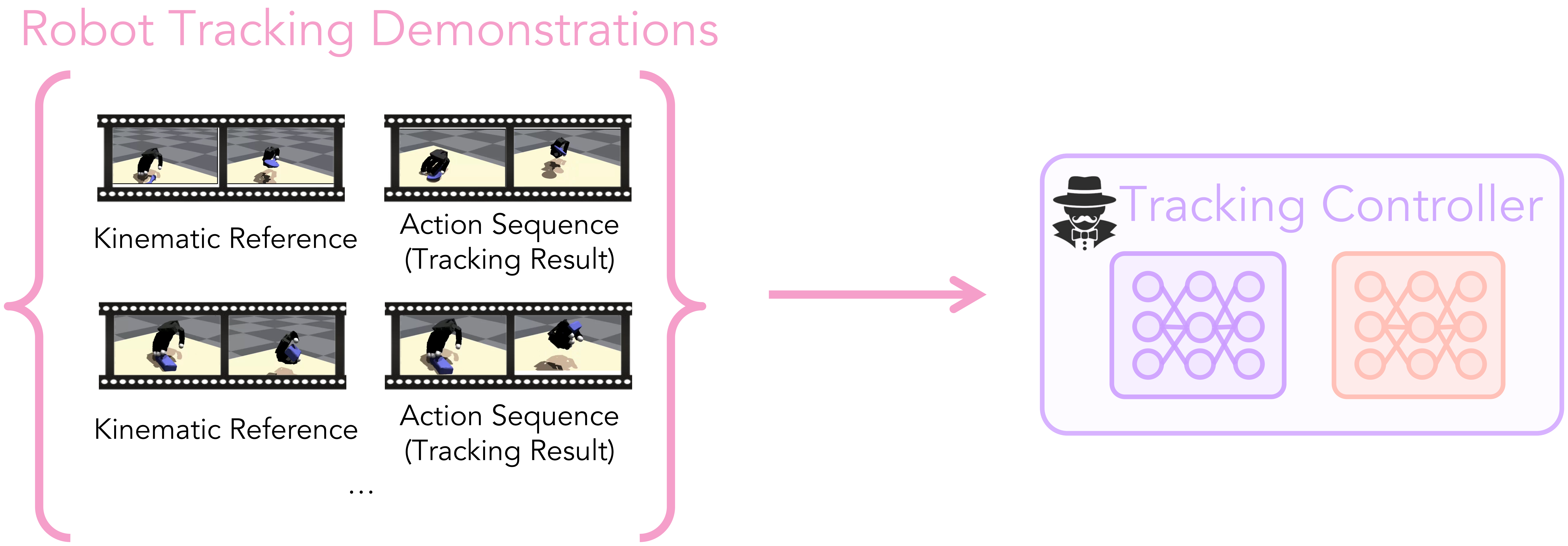
Robust to

Unseen Reference Trajectories with Novel Objects



Large Noise in Reference Motions (unreasonable state, penetration, etc.)

Large Scale Imitation

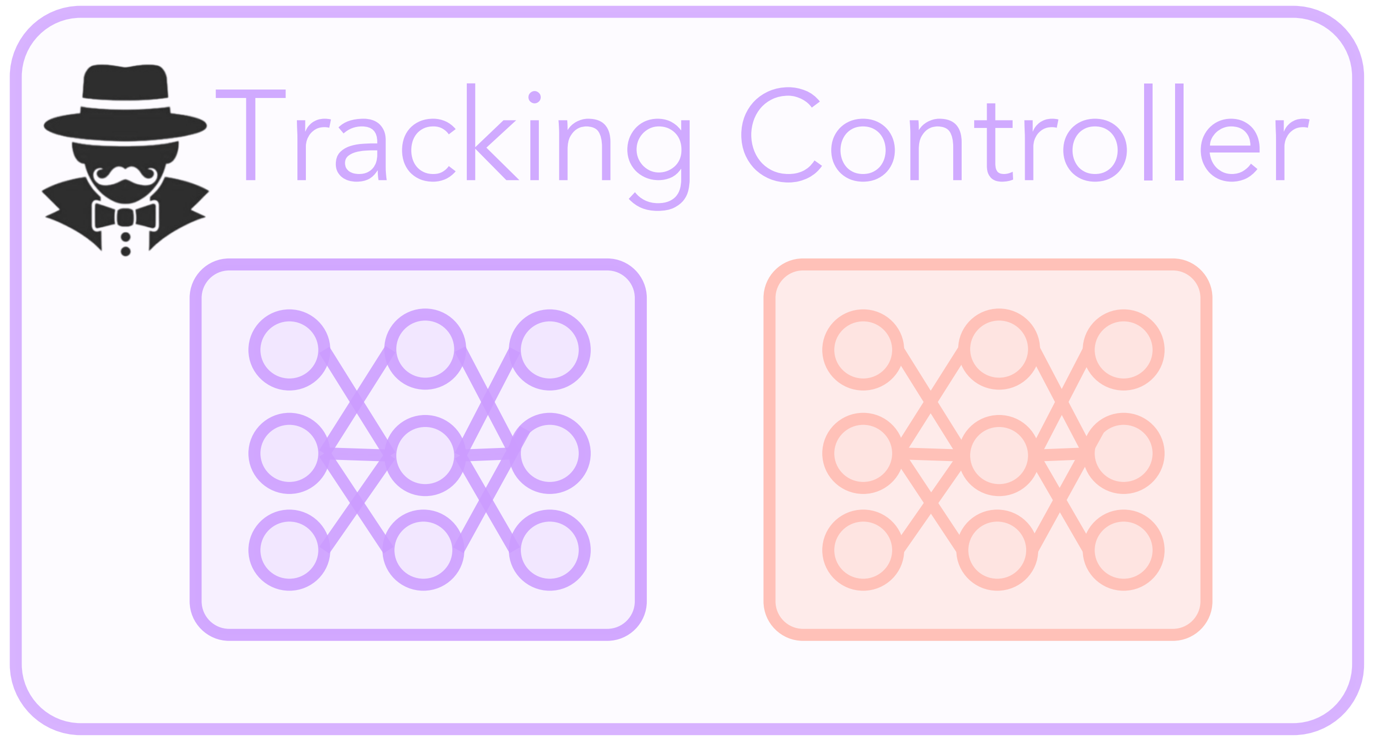


Challenges

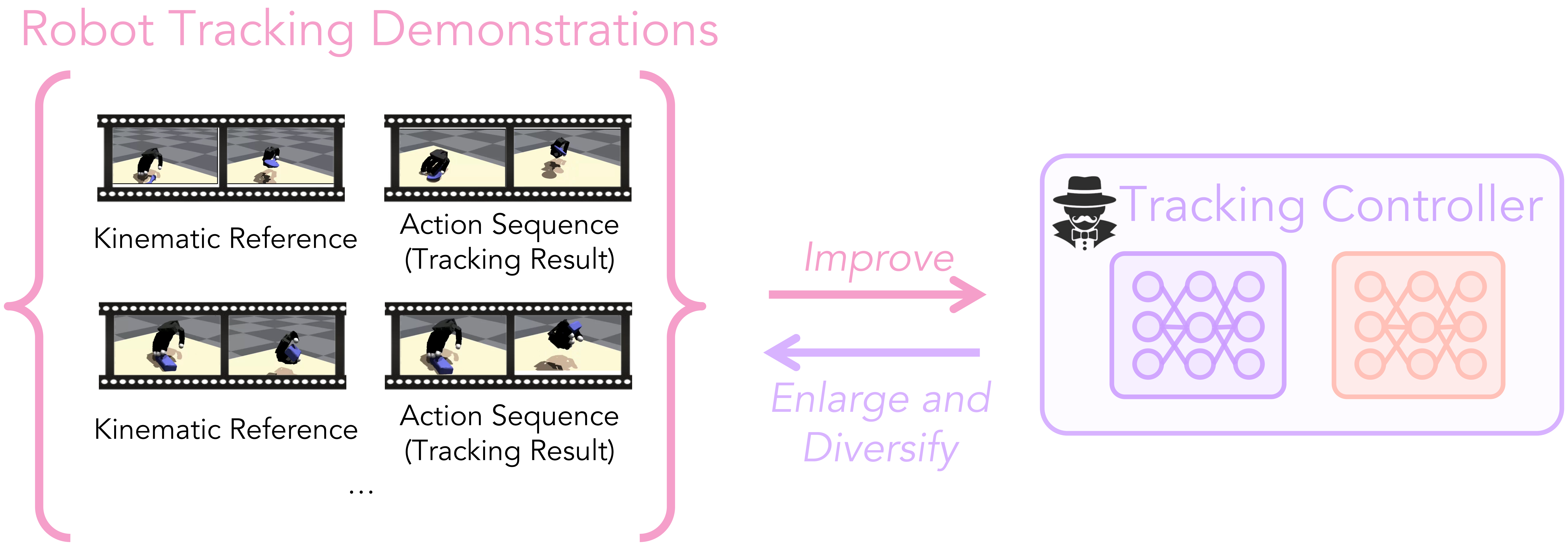
Complex dynamics

Tracking complexity varies

Very biased tracking results!
Diversity is important!



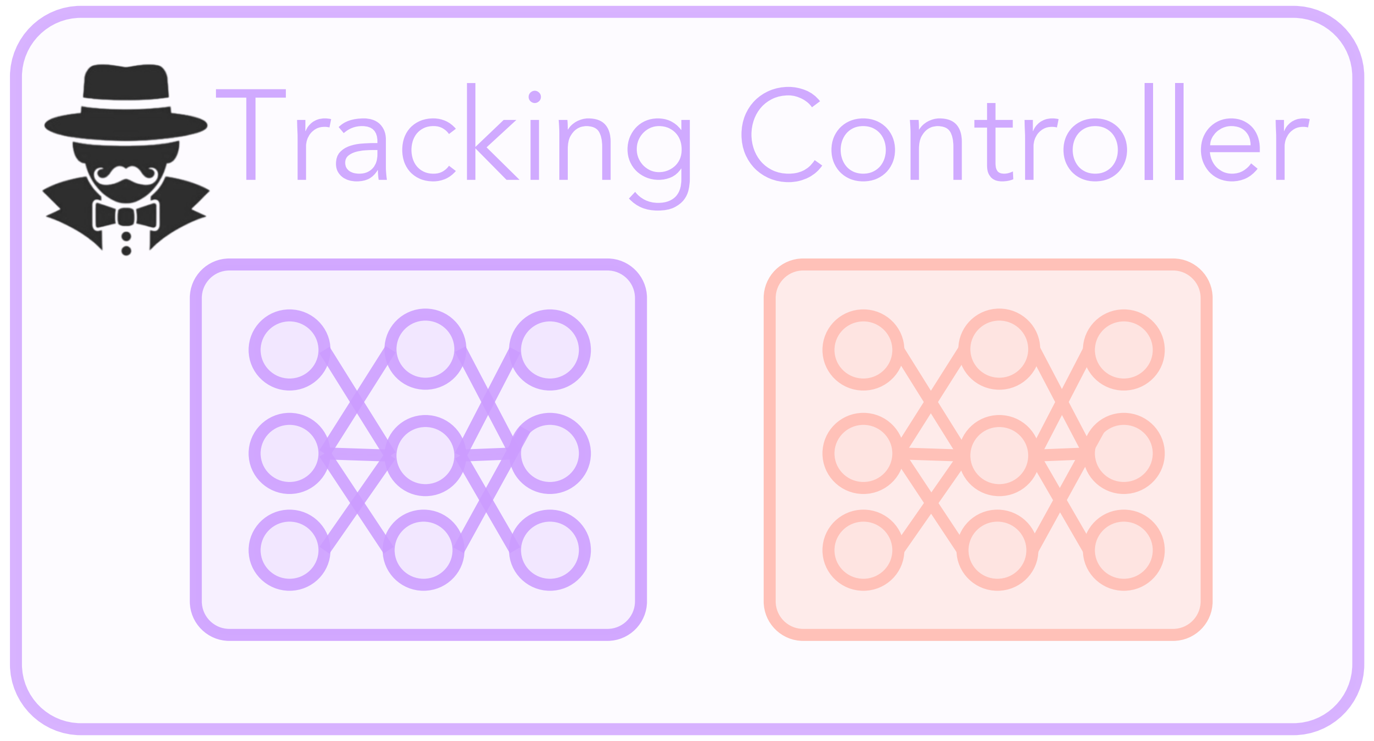
Key Idea: Building a Data Flywheel



Learning a Neural Tracking Controller from Demonstrations

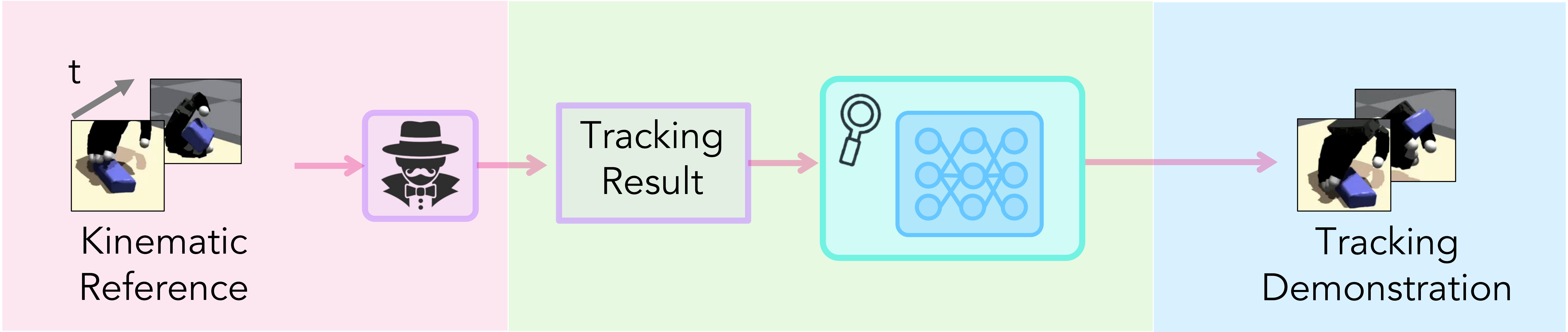
Improving Per-Trajectory Tracking via Data Prior

{ Robot Tracking Demonstrations }



Can we leverage the *tracking controller* to improve tracking demonstrations?

Tracking controller could provide the base for further optimization!



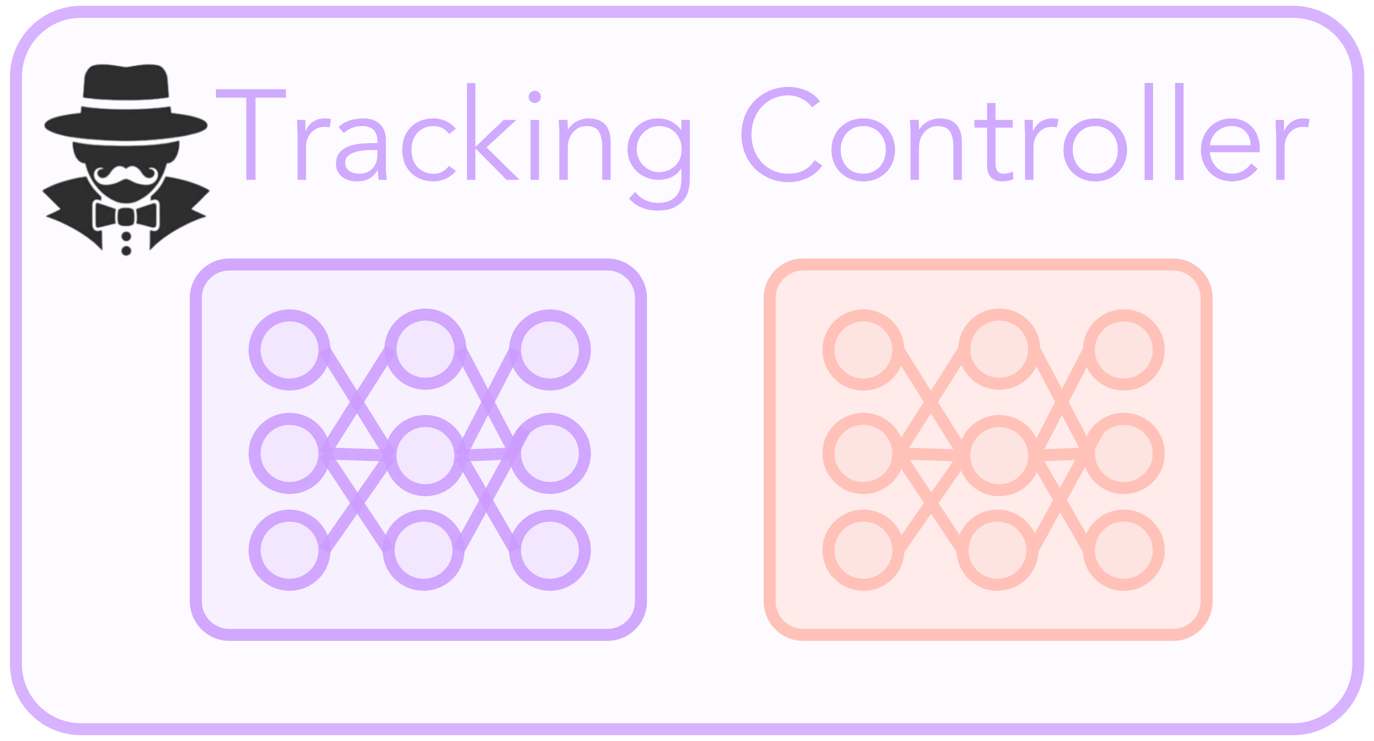
Track the trajectory via the tracking controller

Initialize the tracking policy via the tracking result

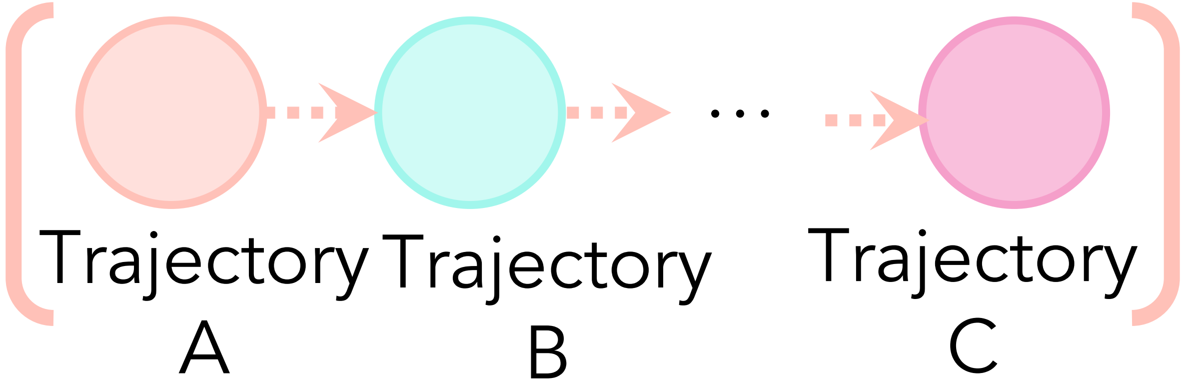
Solving per-trajectory tracking problem

Improving Per-Trajectory Tracking via Homotopy Optimization

Robot Tracking Demonstrations

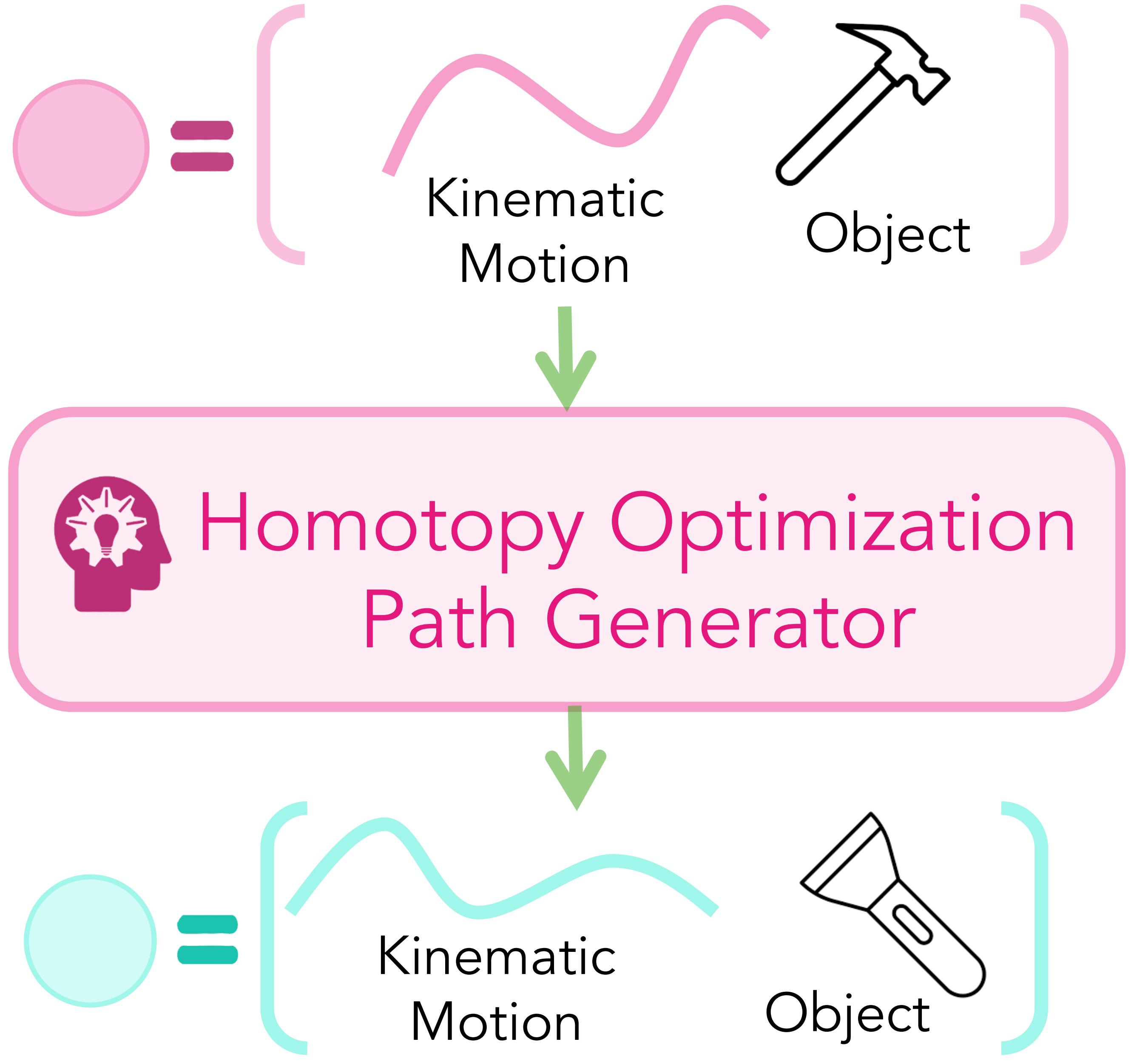


Can we explore *trajectory relations* to improve the per-trajectory tracking?



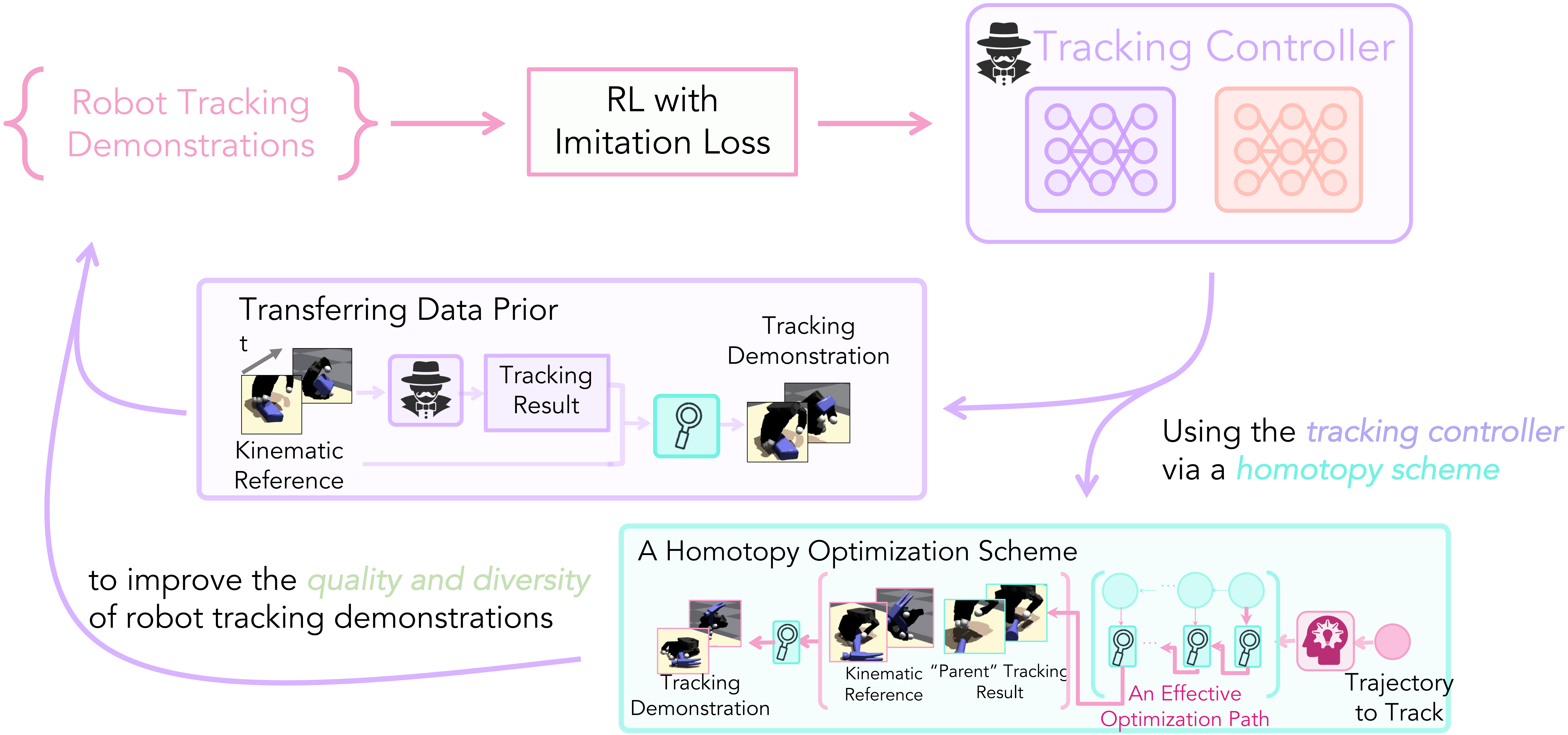
An effective optimization path can improve the per-trajectory tracking result of C

How to model and utilize such trajectory relations to improve the tracking demonstrations?



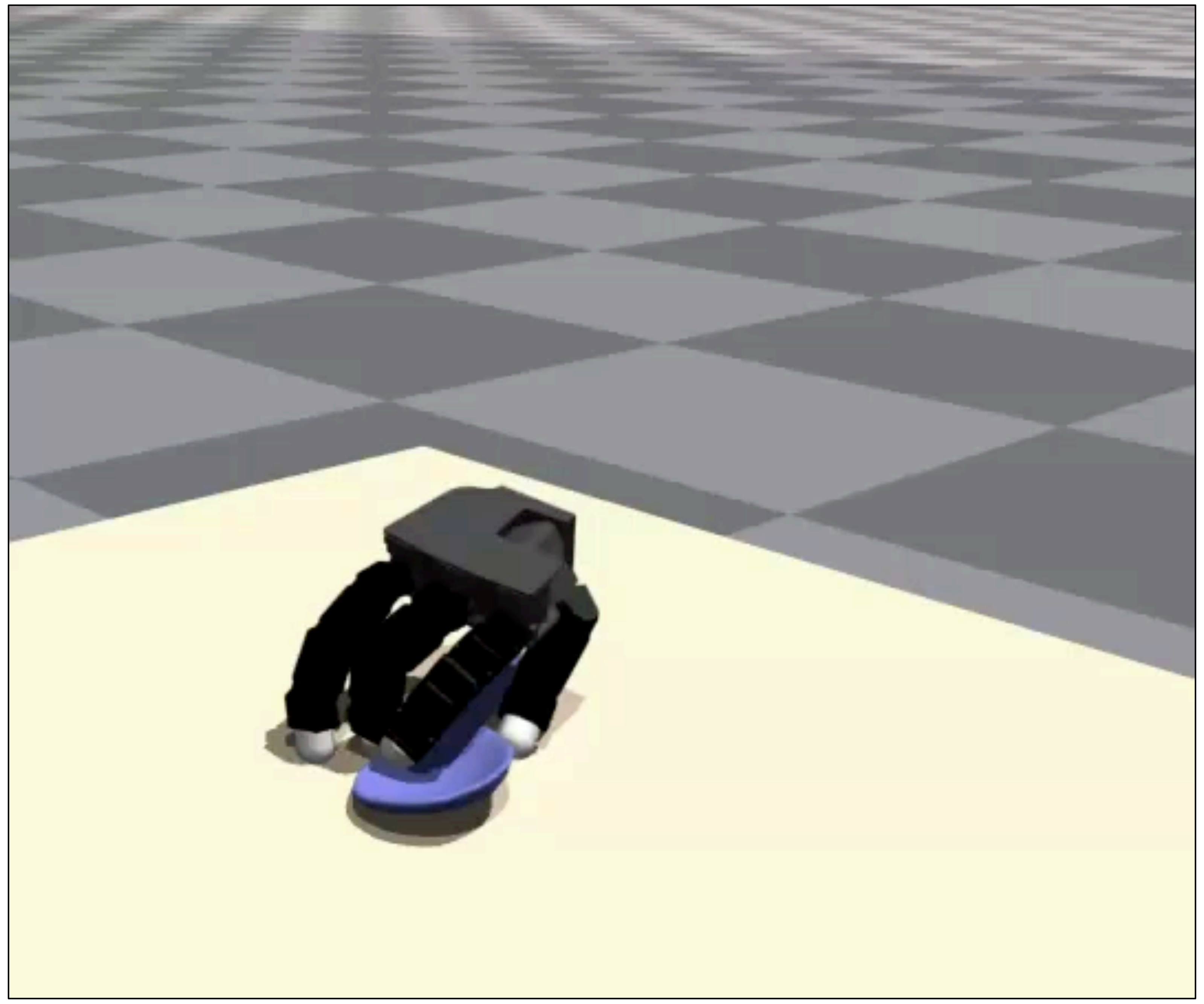
Modeling "parent-child" relations

Learning a Neural Tracking Controller from Demonstrations

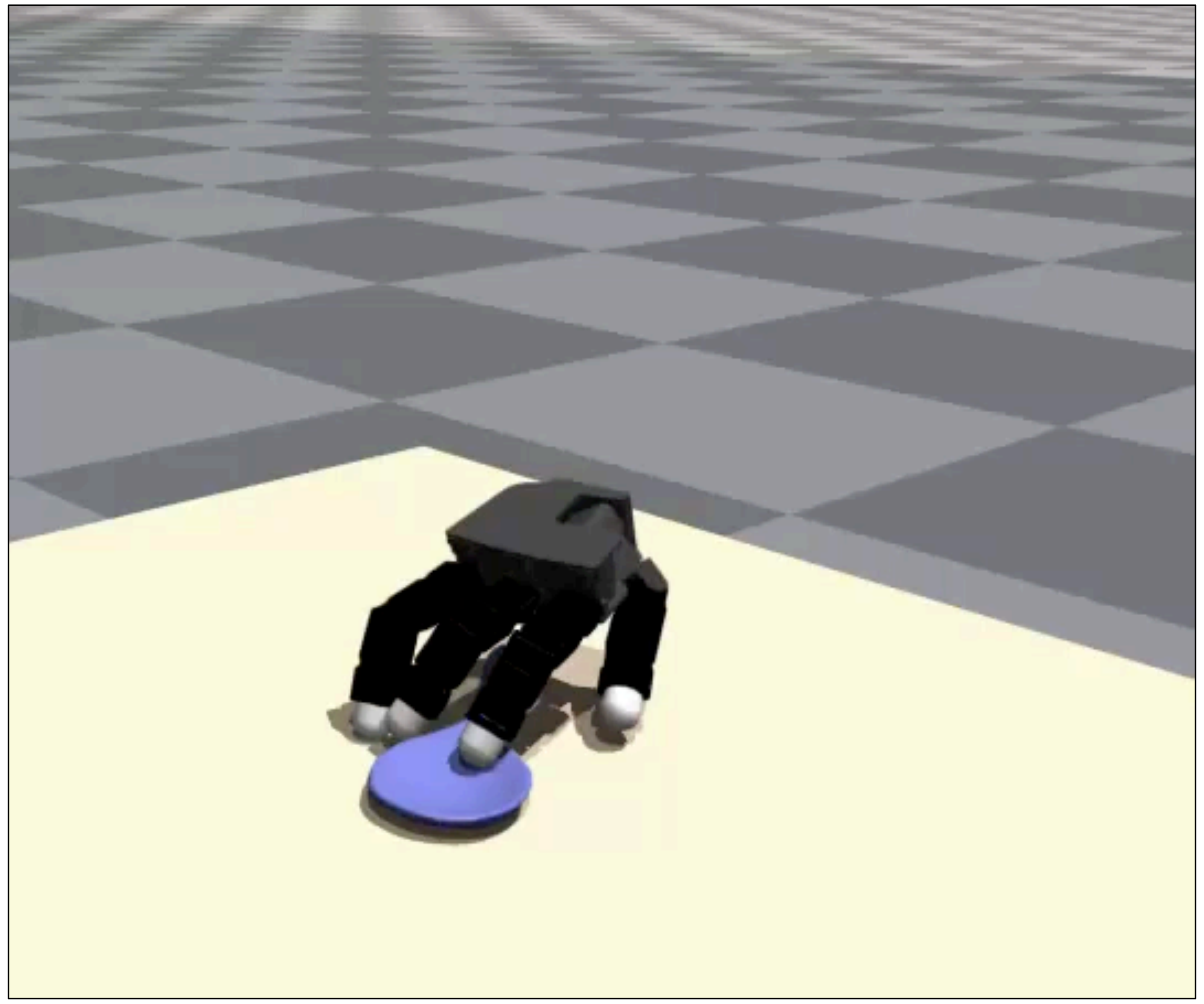


Experimental Results

Retargeted
Kinematic Reference



Ours



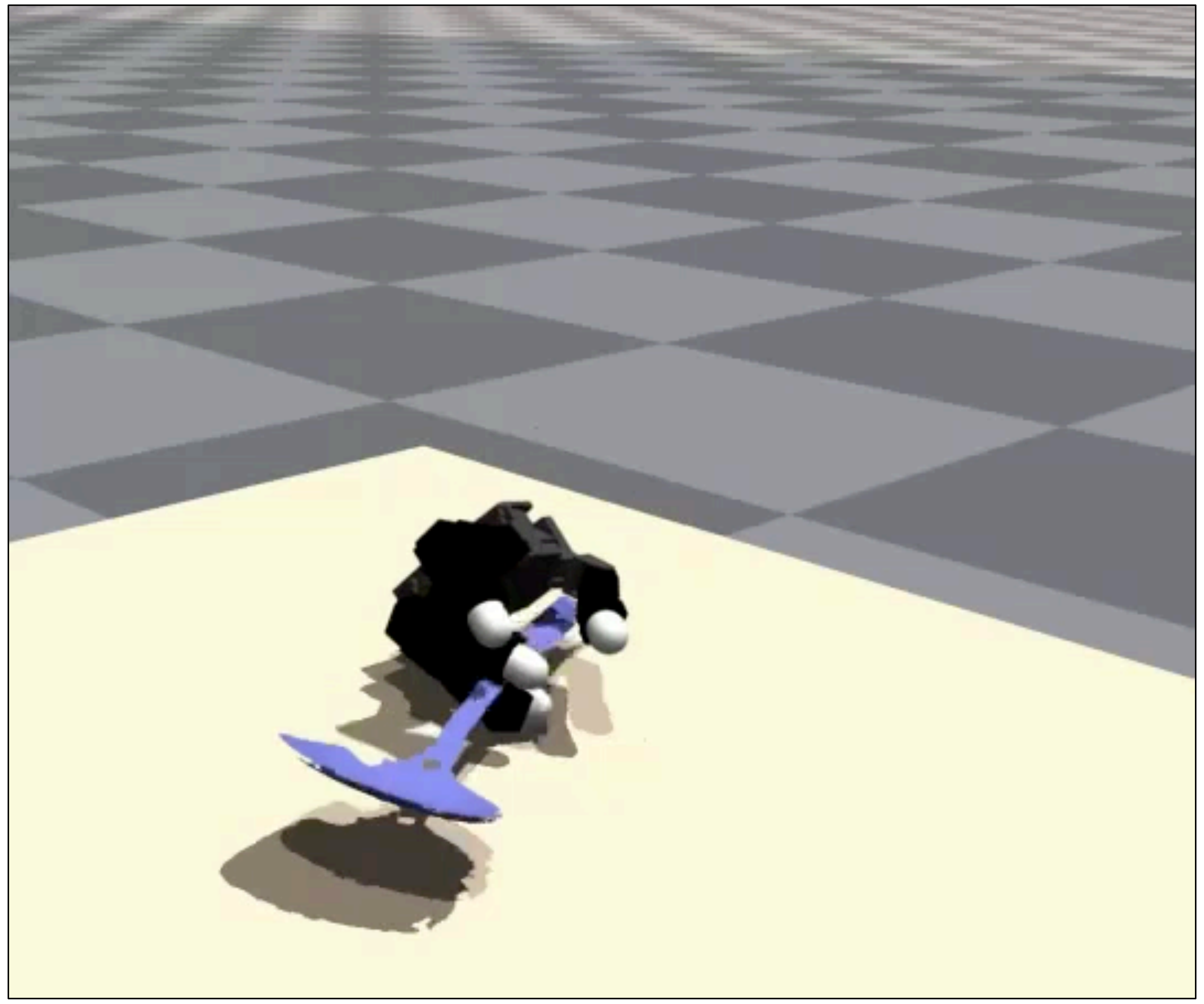
Baseline

Difficulties:

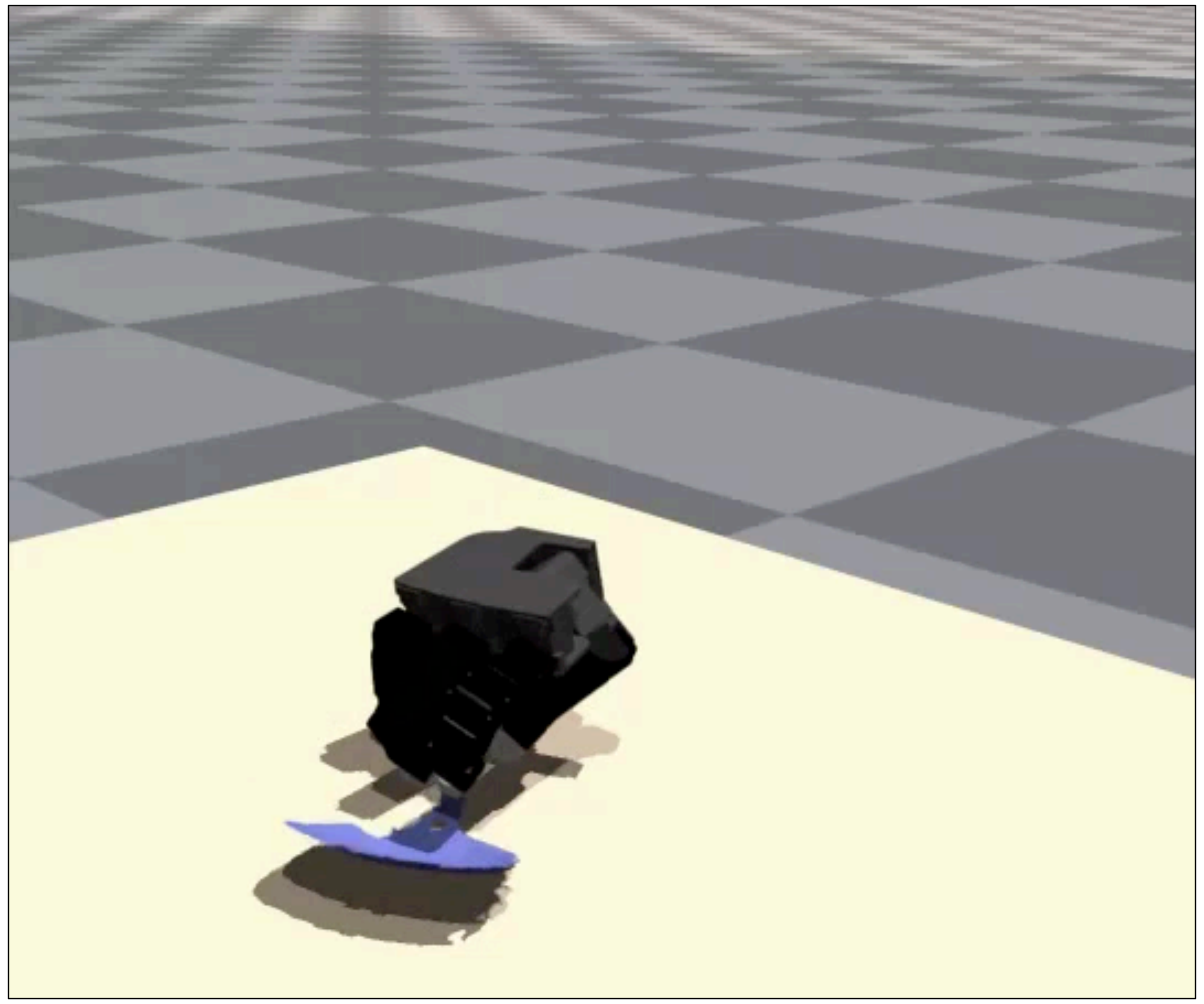
1) Small and thin shovel
2) Complex object movements with subtle in-hand re-orientation

Experimental Results

Retargeted
Kinematic Reference



Ours



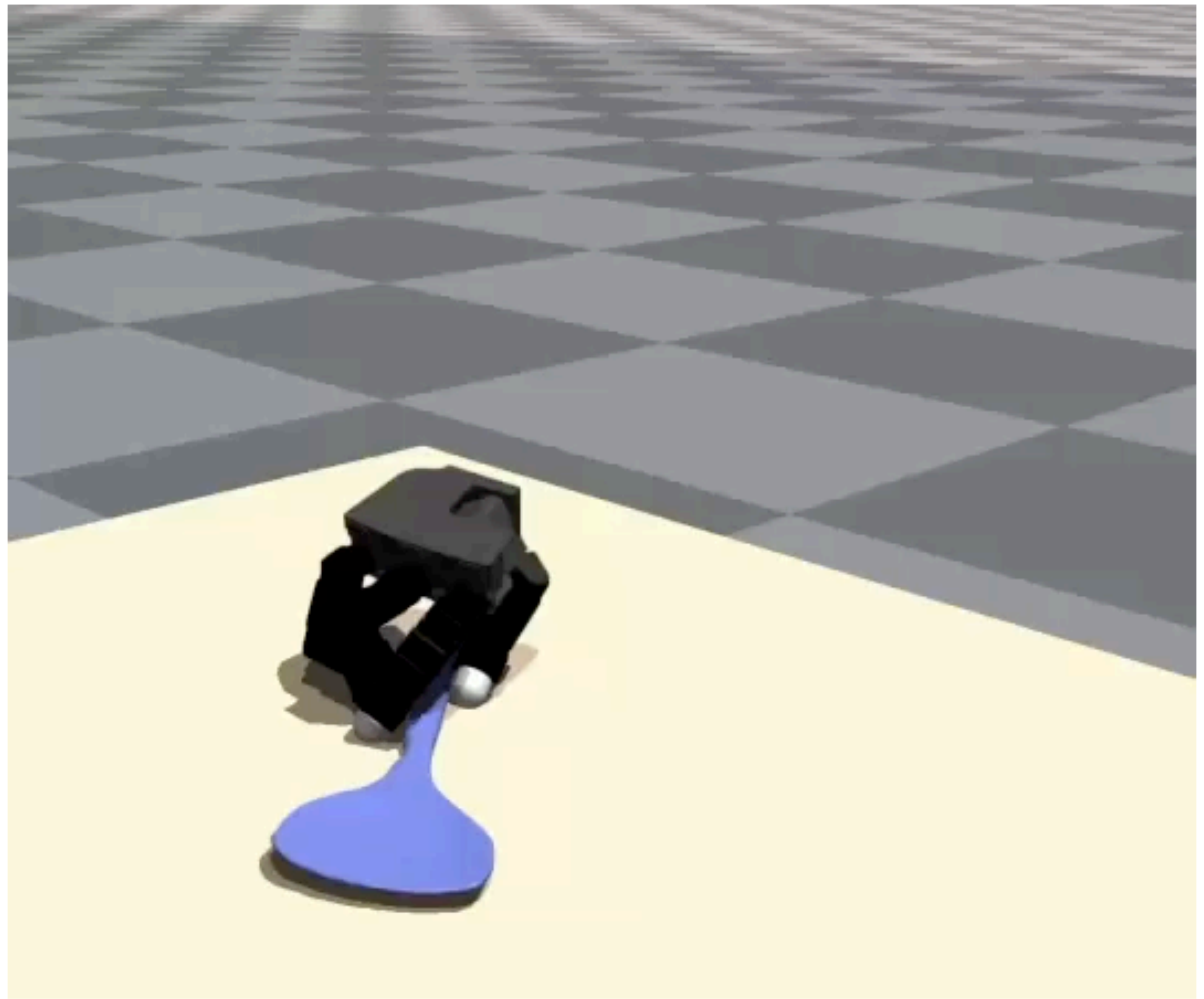
Baseline

Difficulties:
1) Thin shovel with missing faces
2) Complex object movements (lifting – waving stage 1 – waving stage 2)

Experimental Results

Retargeted Kinematic Reference

Ours

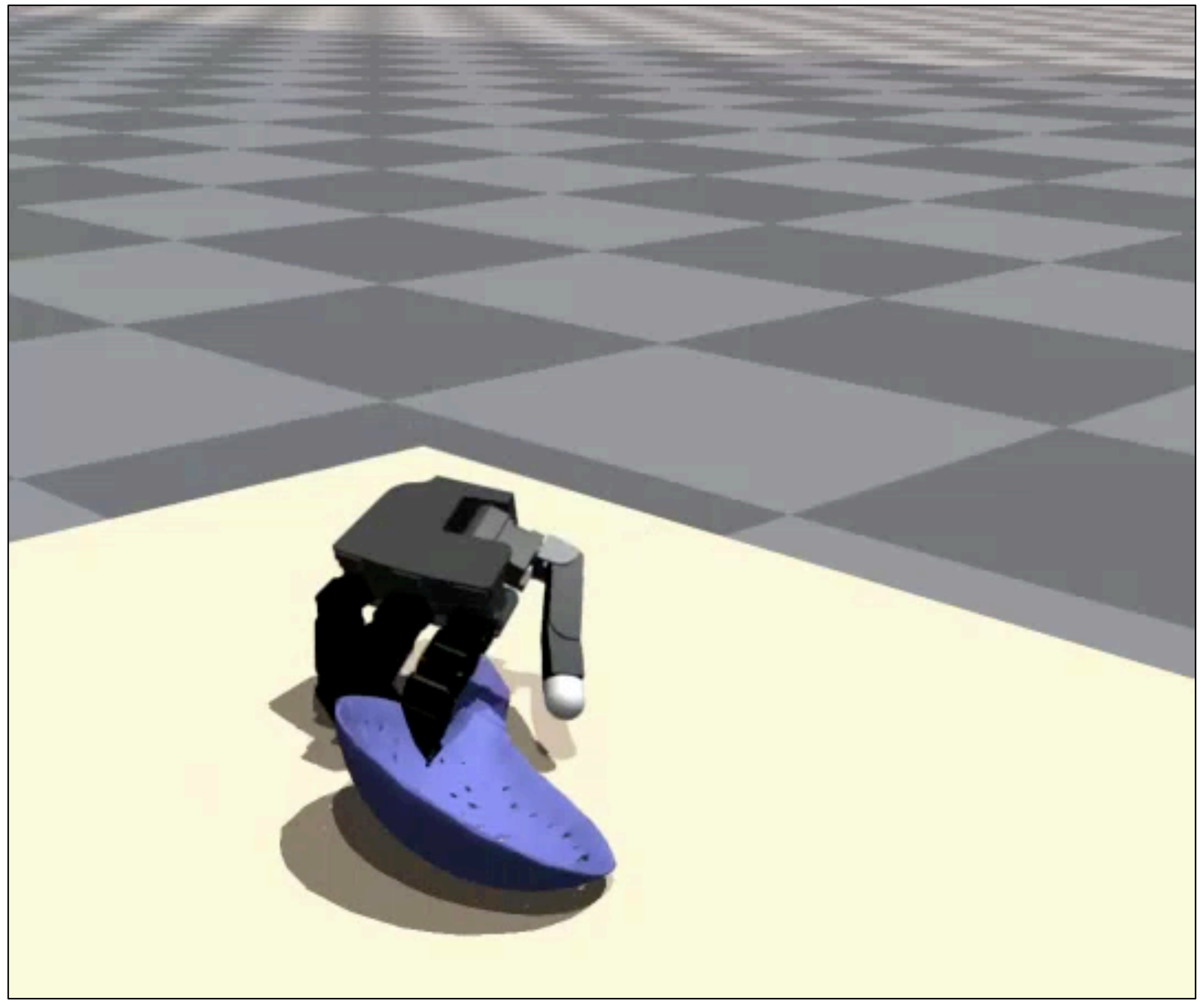


Baseline

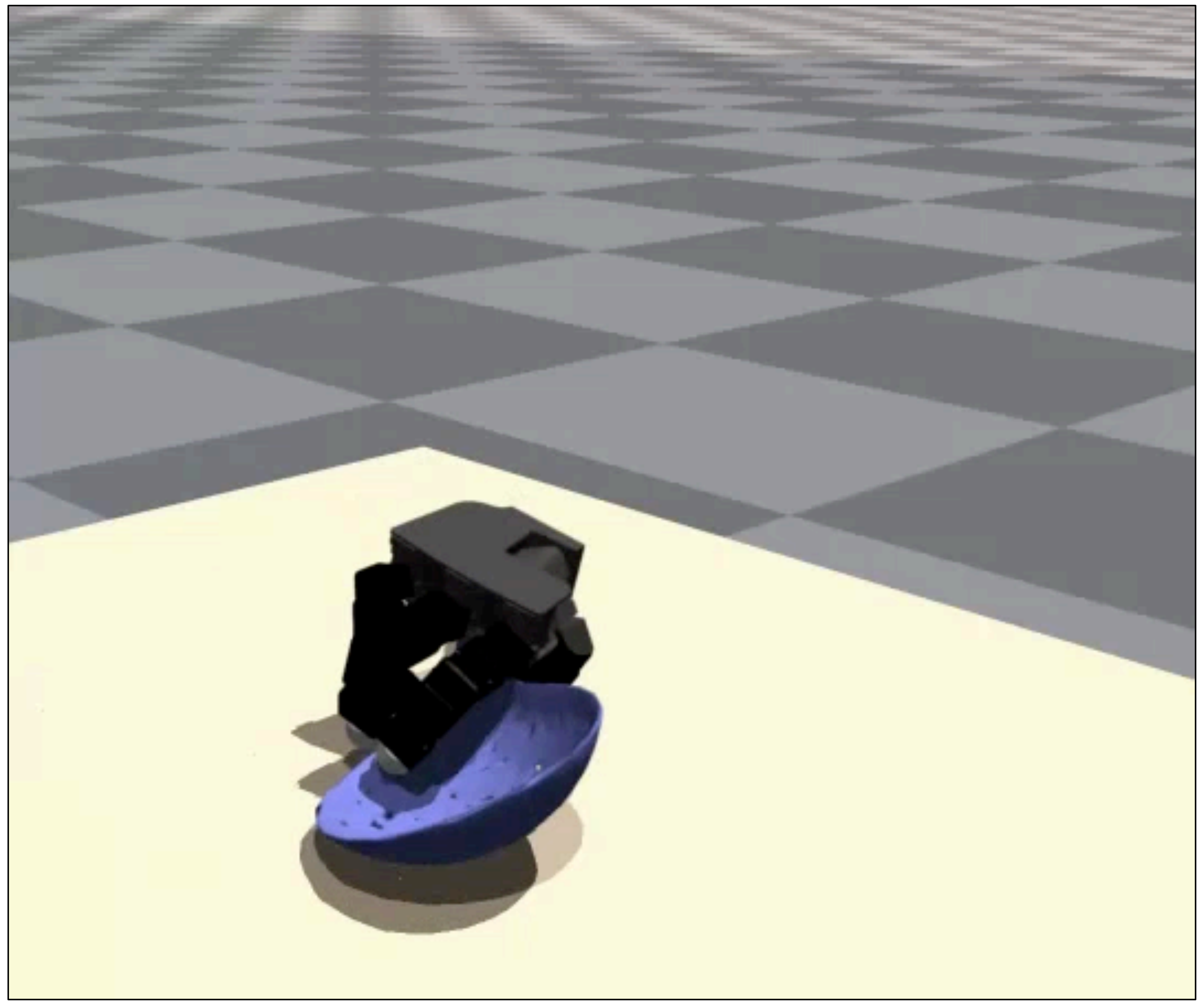
Difficulties:

1) Thin (*hard to grasp*) and long (*difficult to hold firmly and control*) shovel
2) Complex object movements (lifting -- the challenging waving stage)

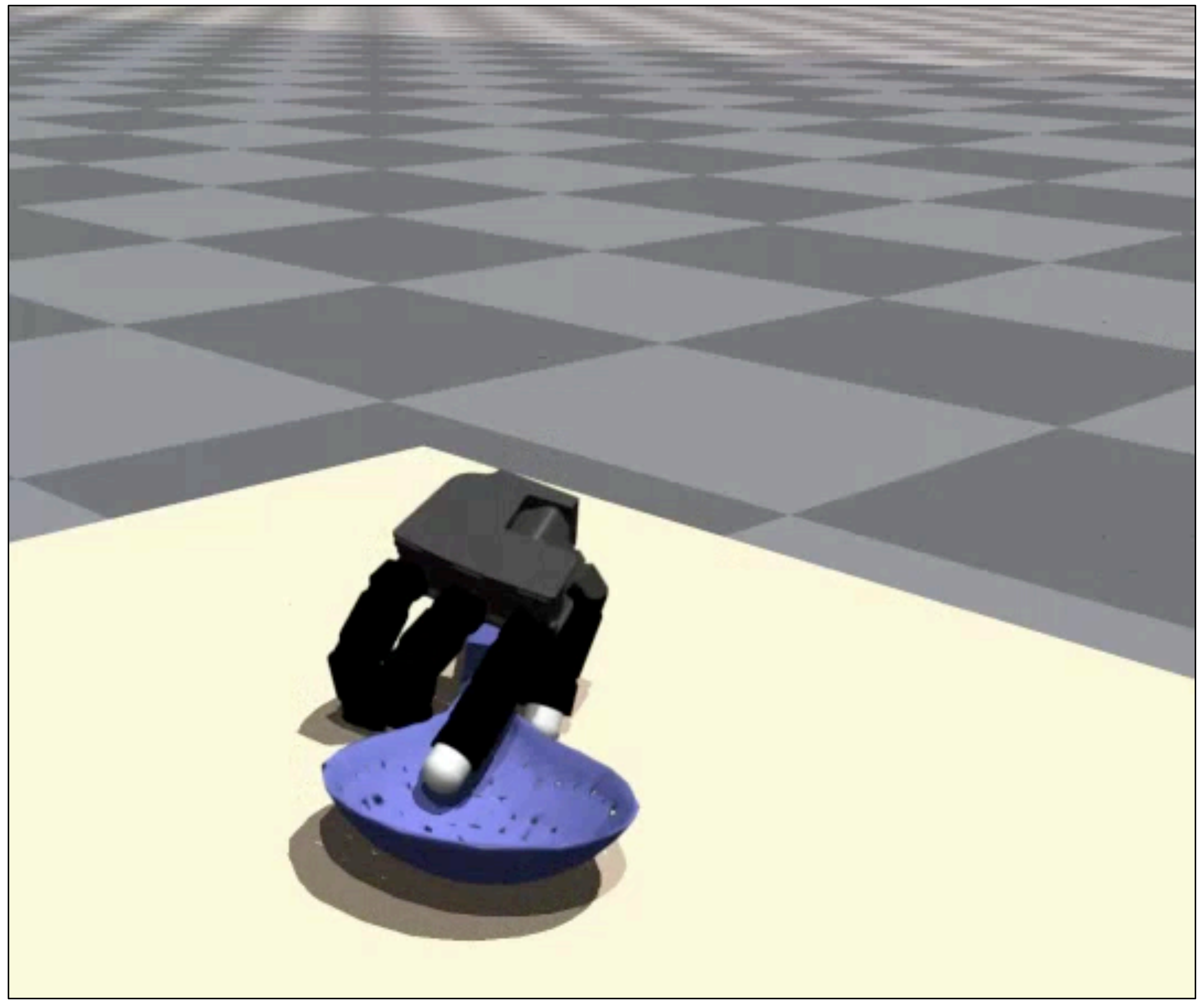
Experimental Results



Retargeted
Kinematic Reference



Ours

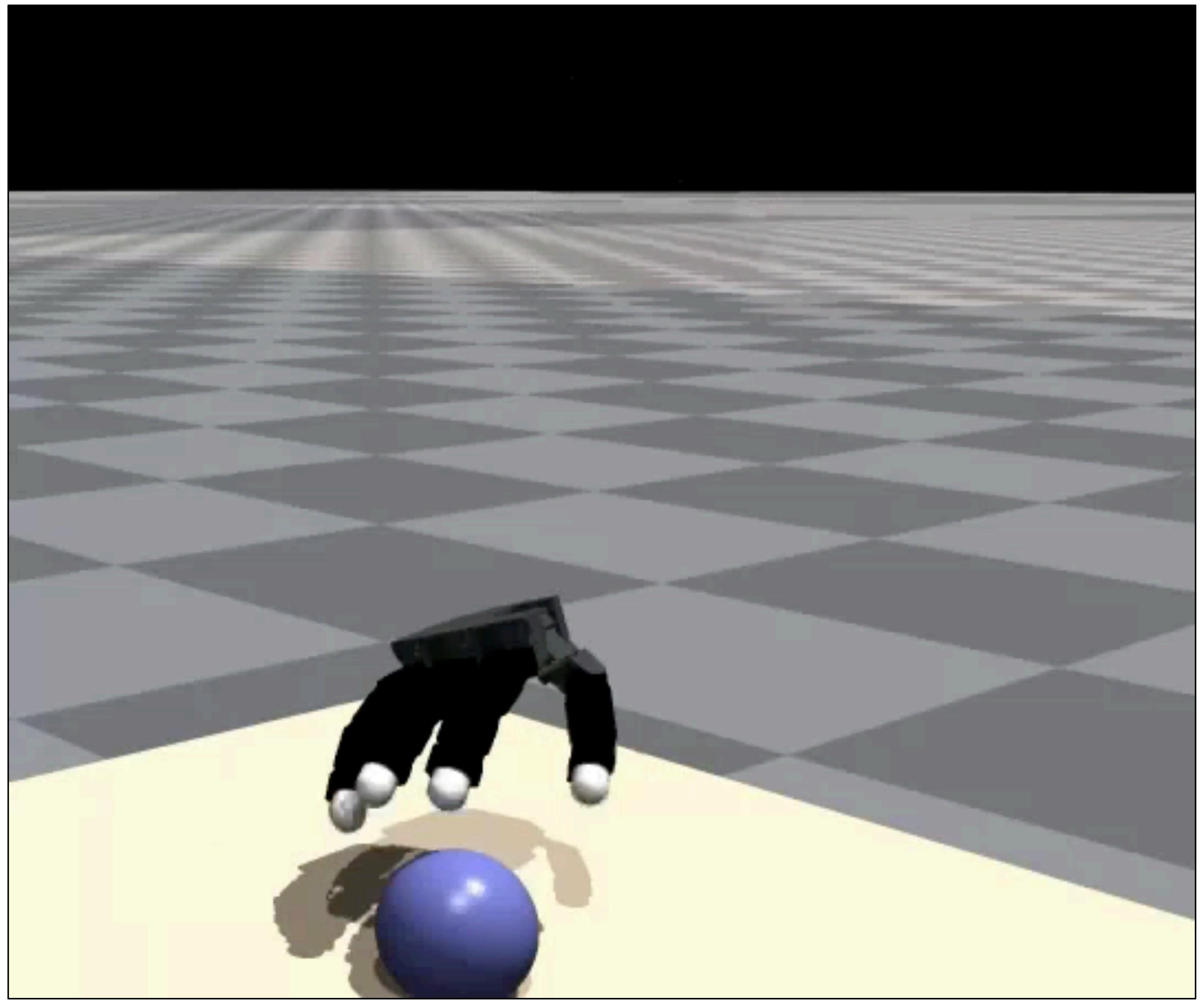


Baseline

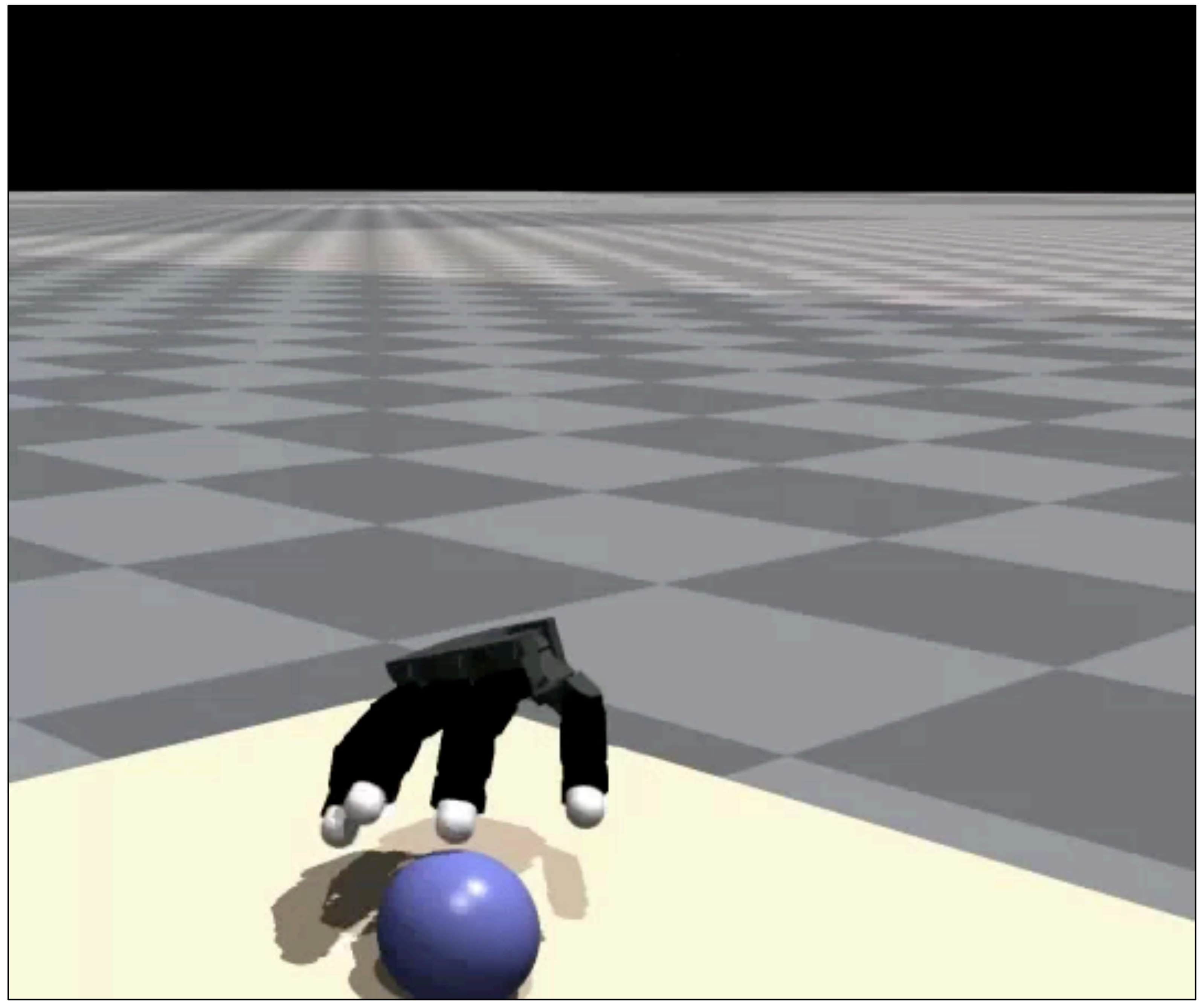
Difficulties:

1) Large and long (*difficult to hold firmly and control*) object with *a challenging gravity center*
2) Complex object movements (lifting -- the waving stage)

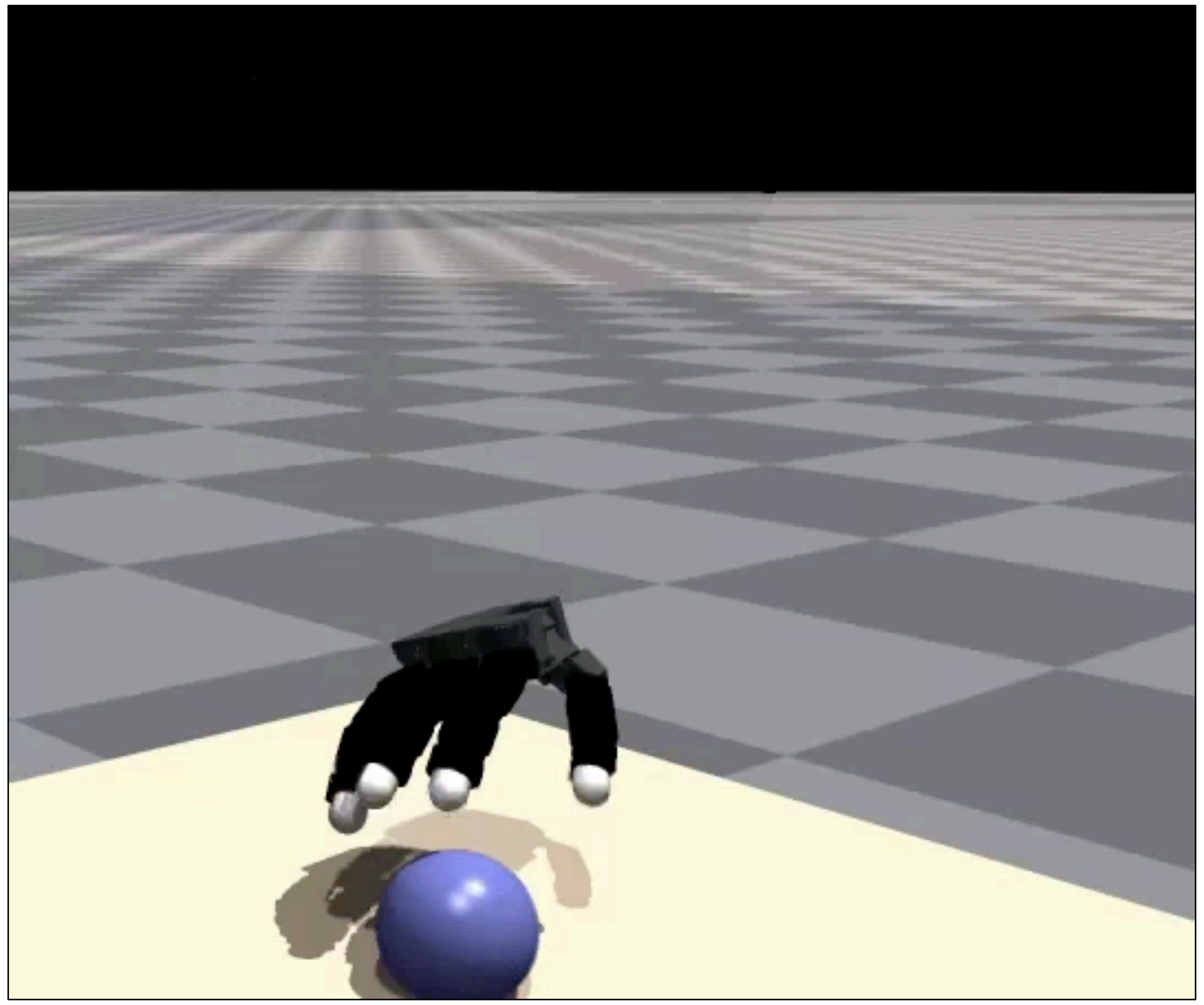
Experimental Results



Retargeted
Kinematic Reference



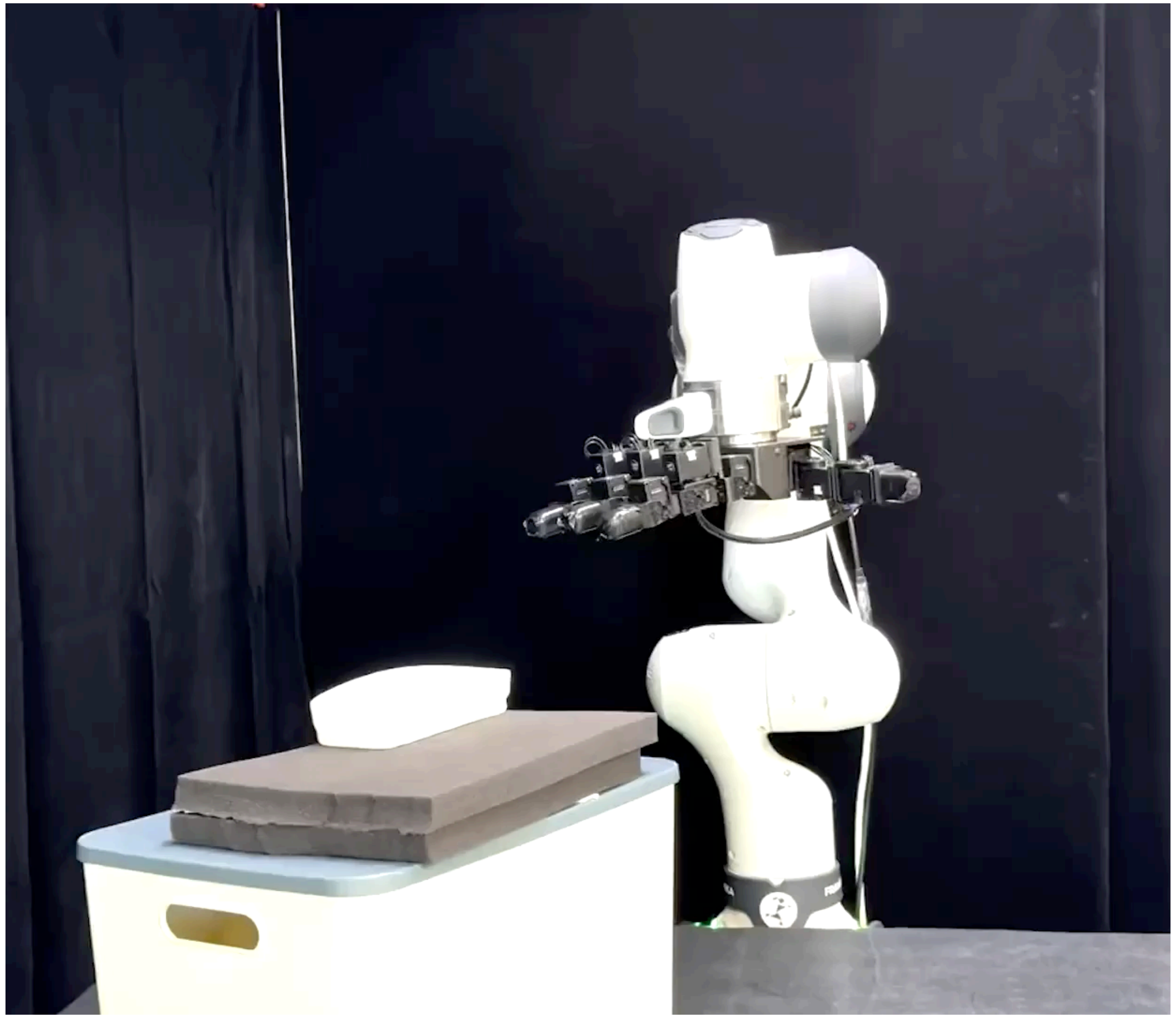
Ours



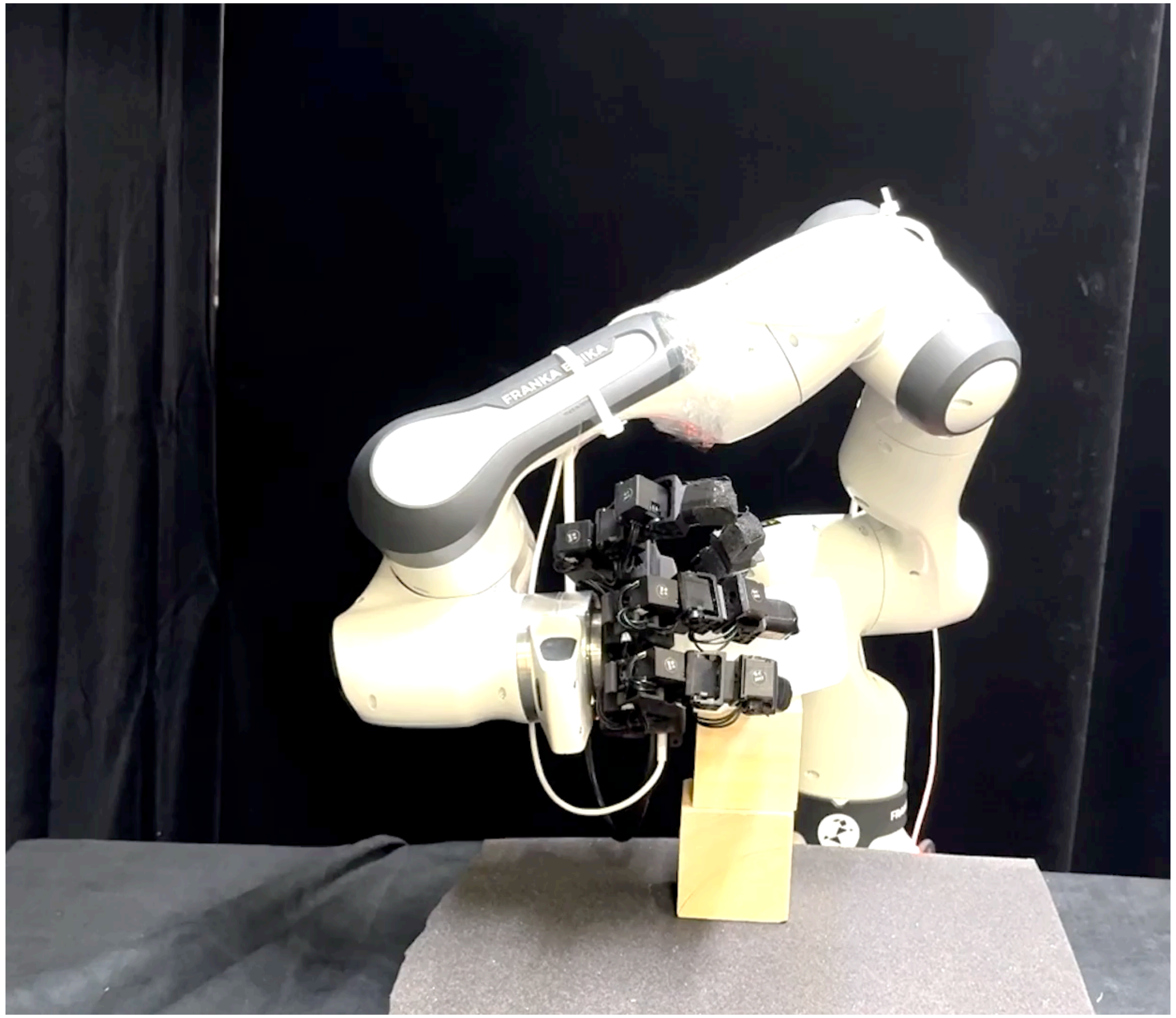
Baseline

Difficulties:
Round sphere that is hard to grasp

Experimental Results

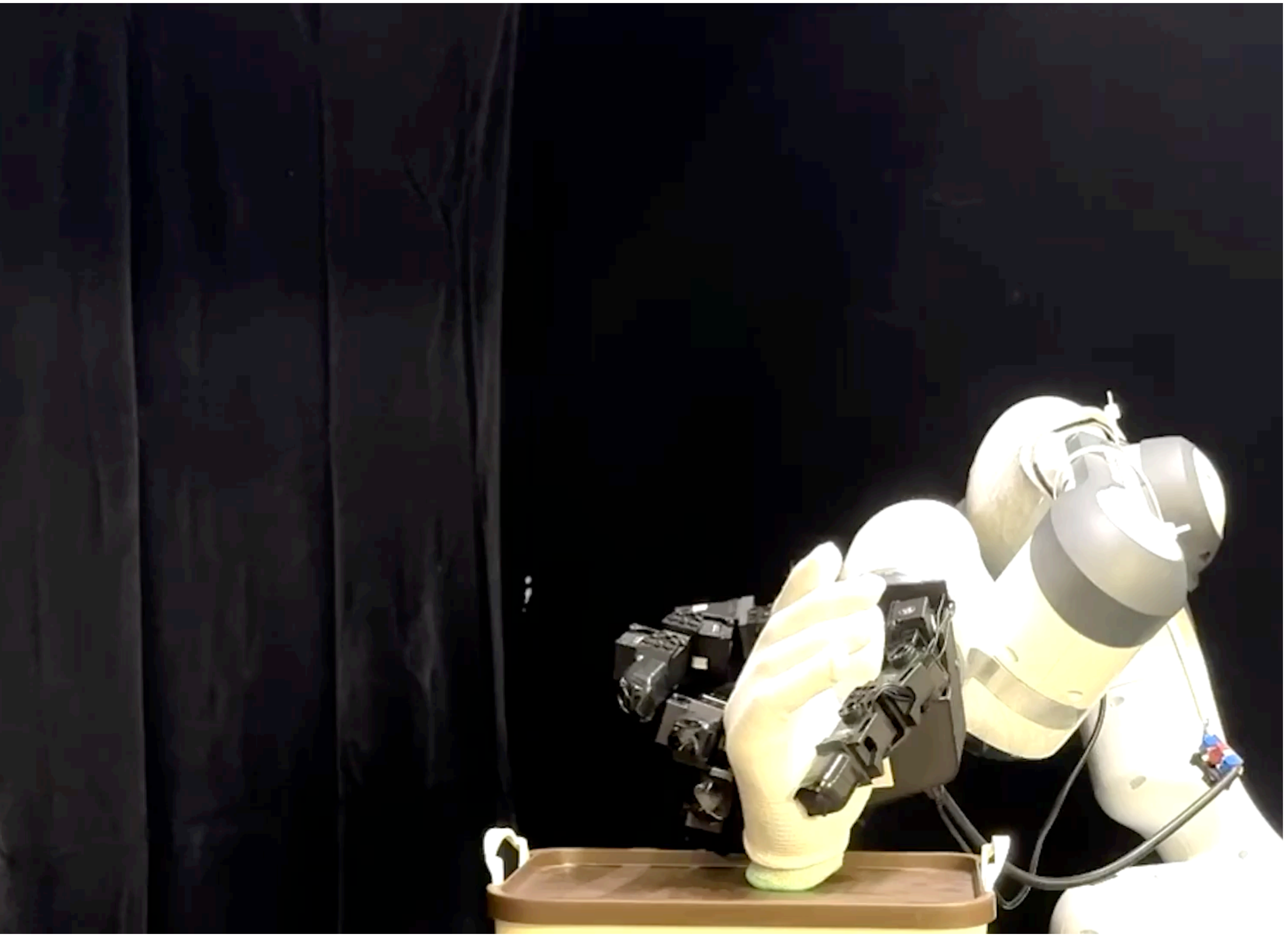


Ours

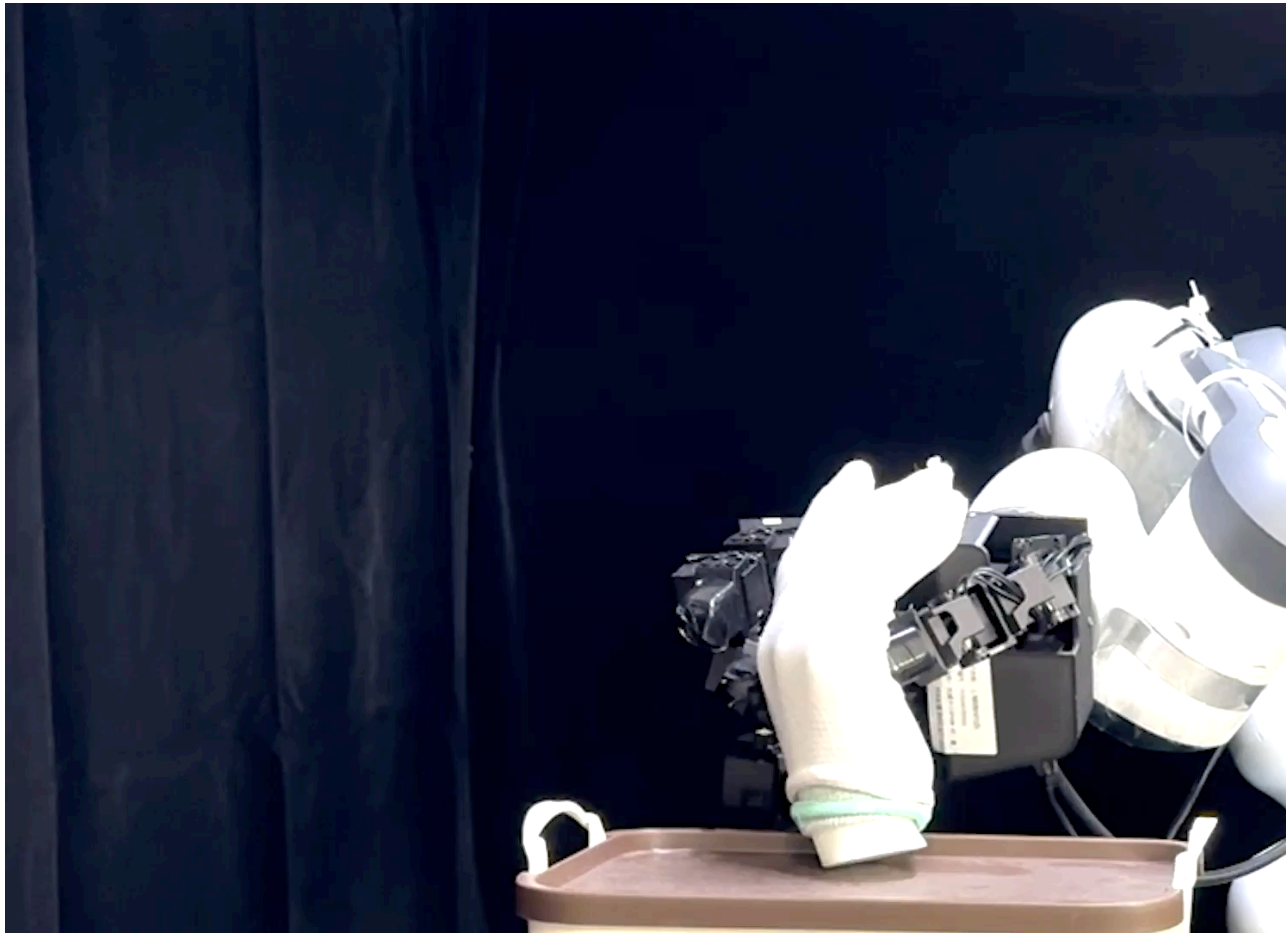


Baseline

Experimental Results

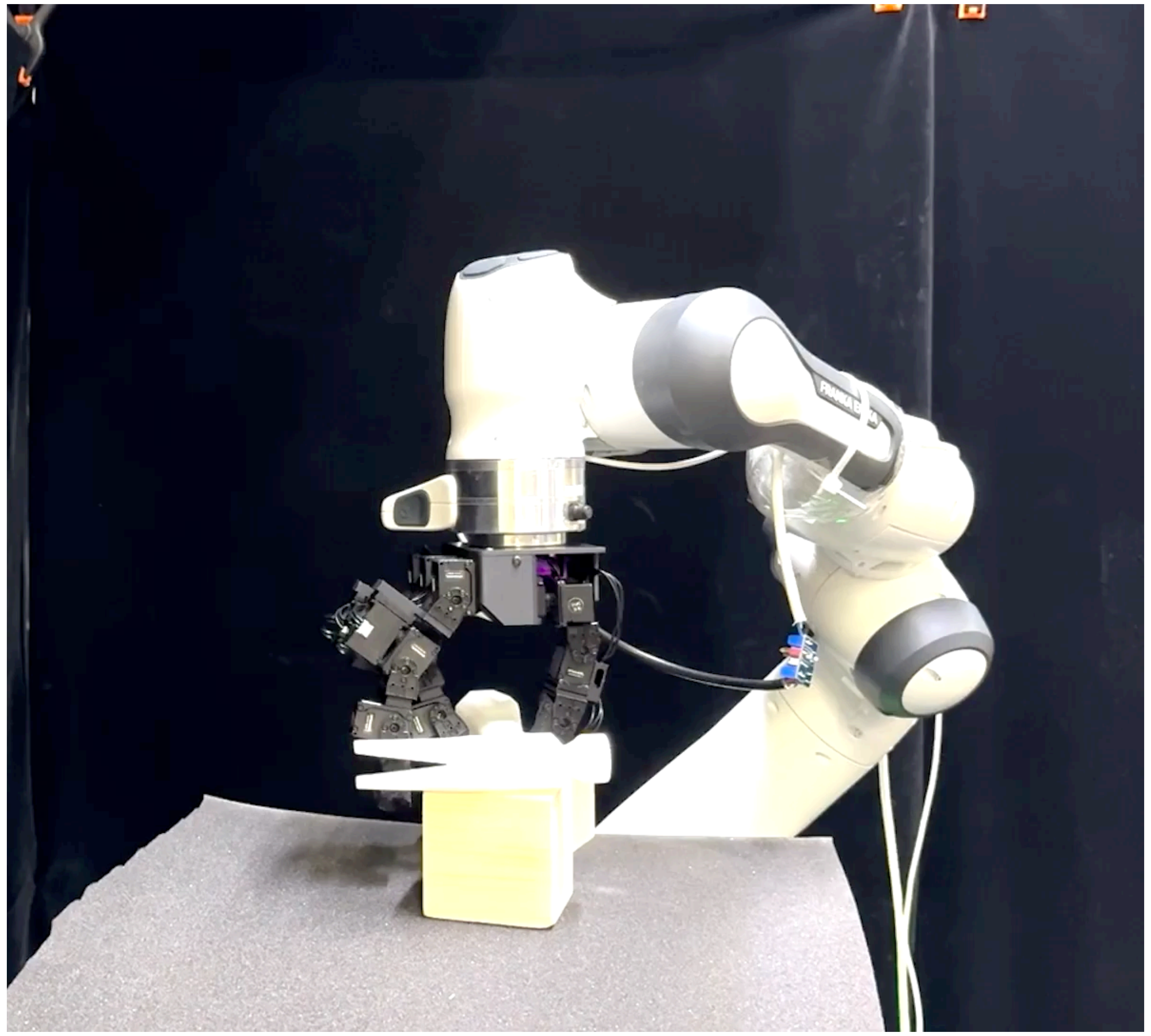


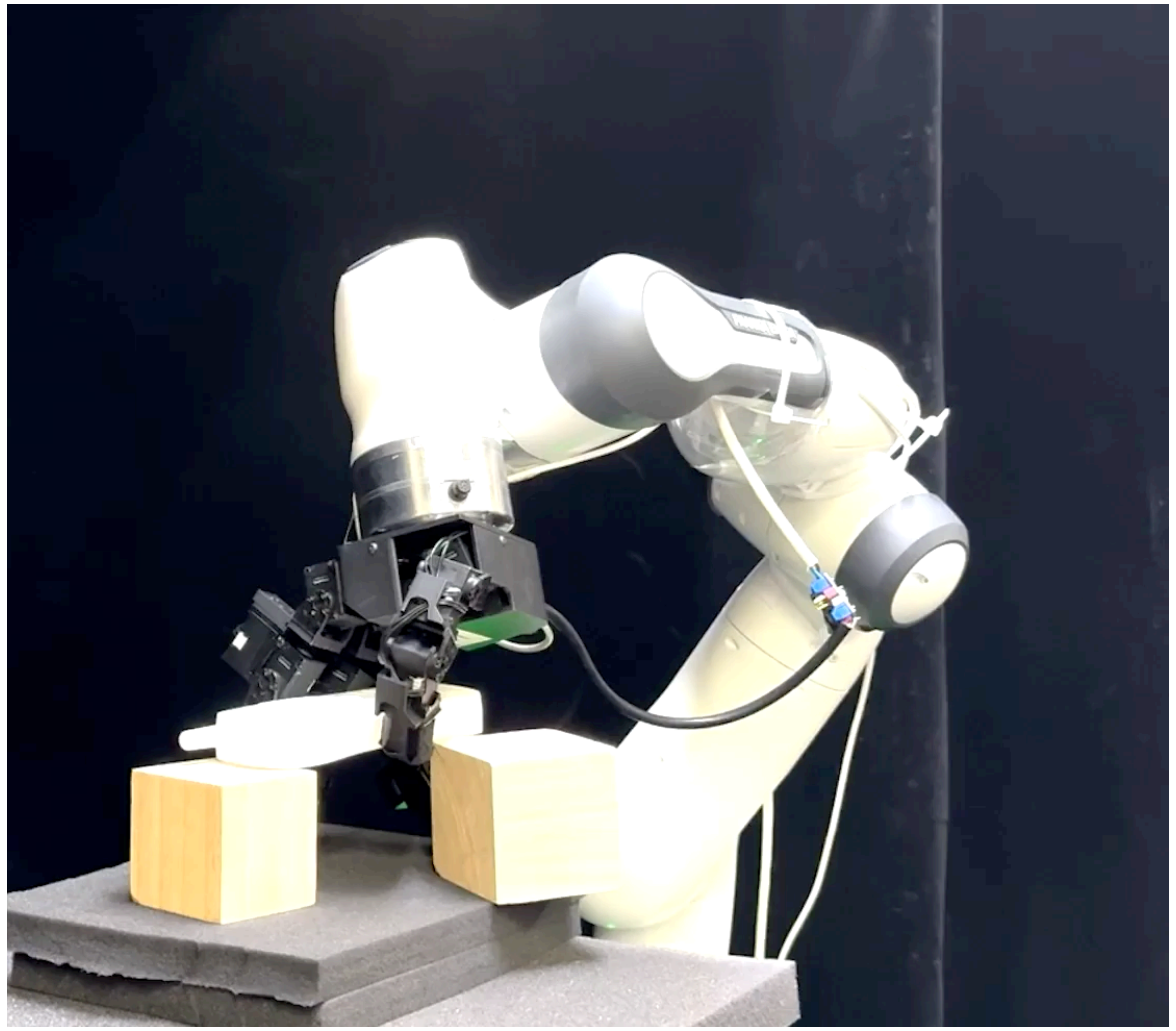
Ours

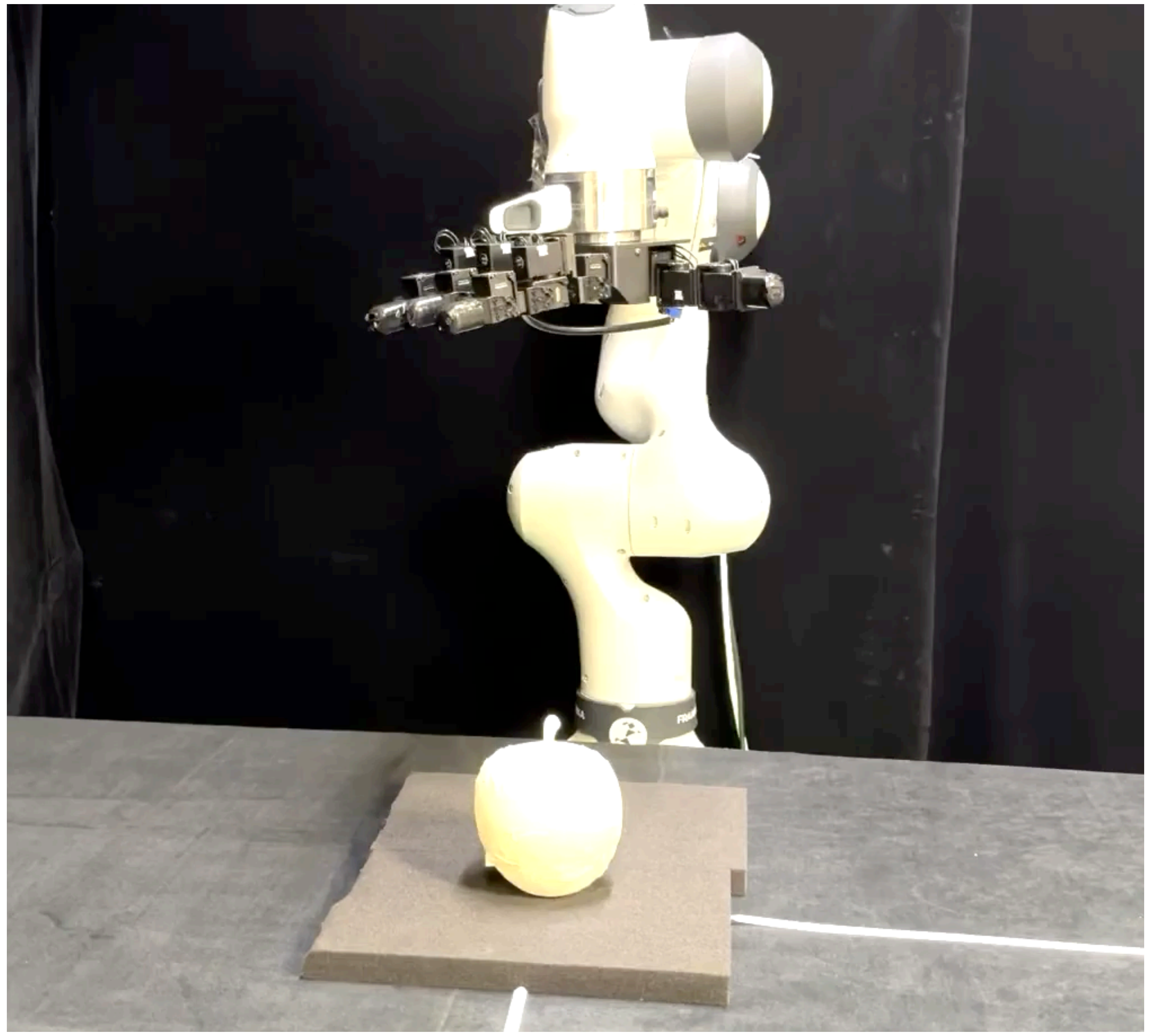


Baseline

Experimental Results







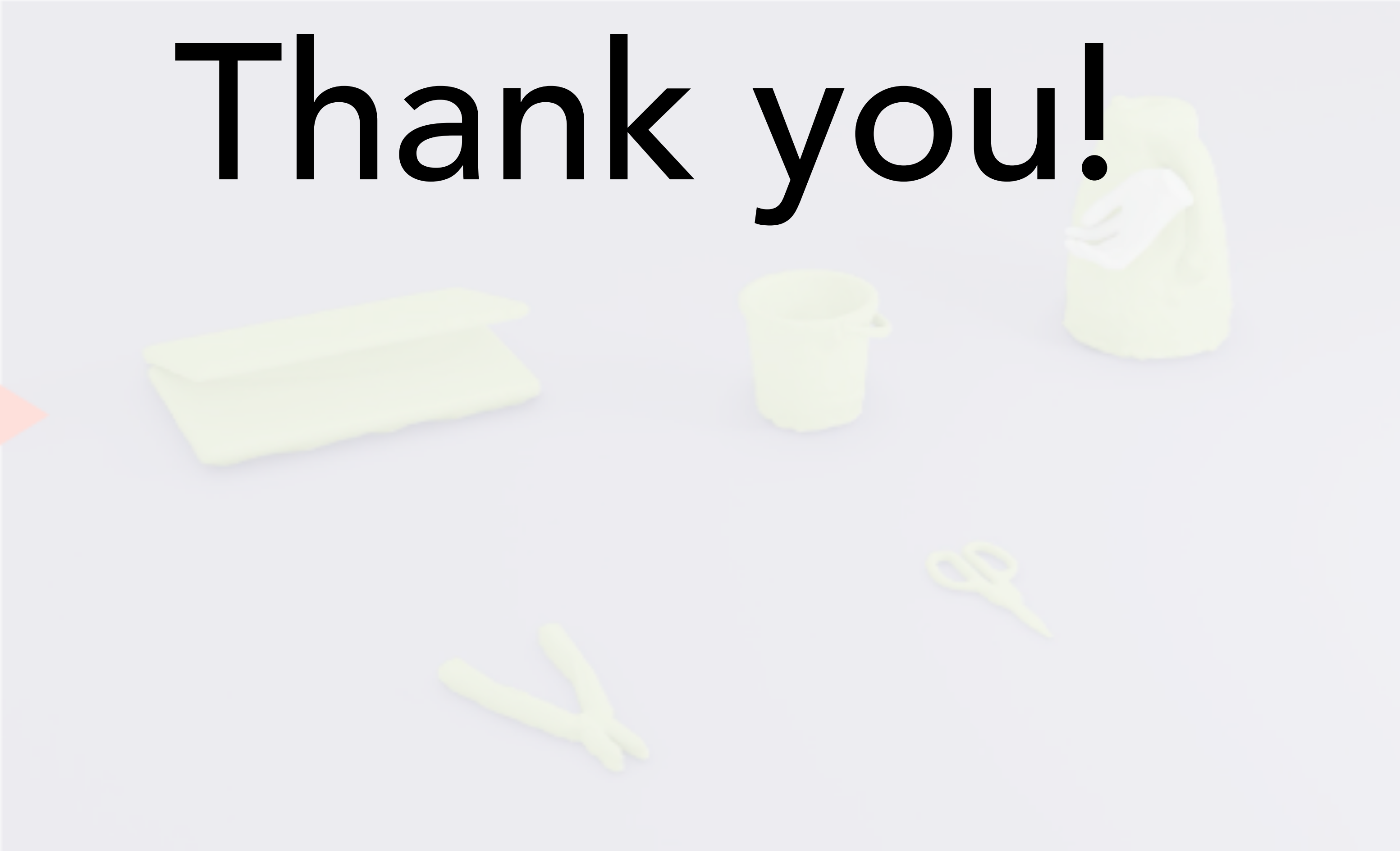
Conclusion

- Human videos are ubiquitous online containing huge amount of manipulation data
- Learning to plan semantic manipulation from human data is possible as the AIGC technology progresses
- Cross-embodiment tracking control can physically control a dexterous hand to follow the planned trajectory for general purpose dexterous manipulation

Thank you!

Acquiring Human Manipulation Data

Generative Human Manipulation Planning



Cross-Embodiment Tracking Control