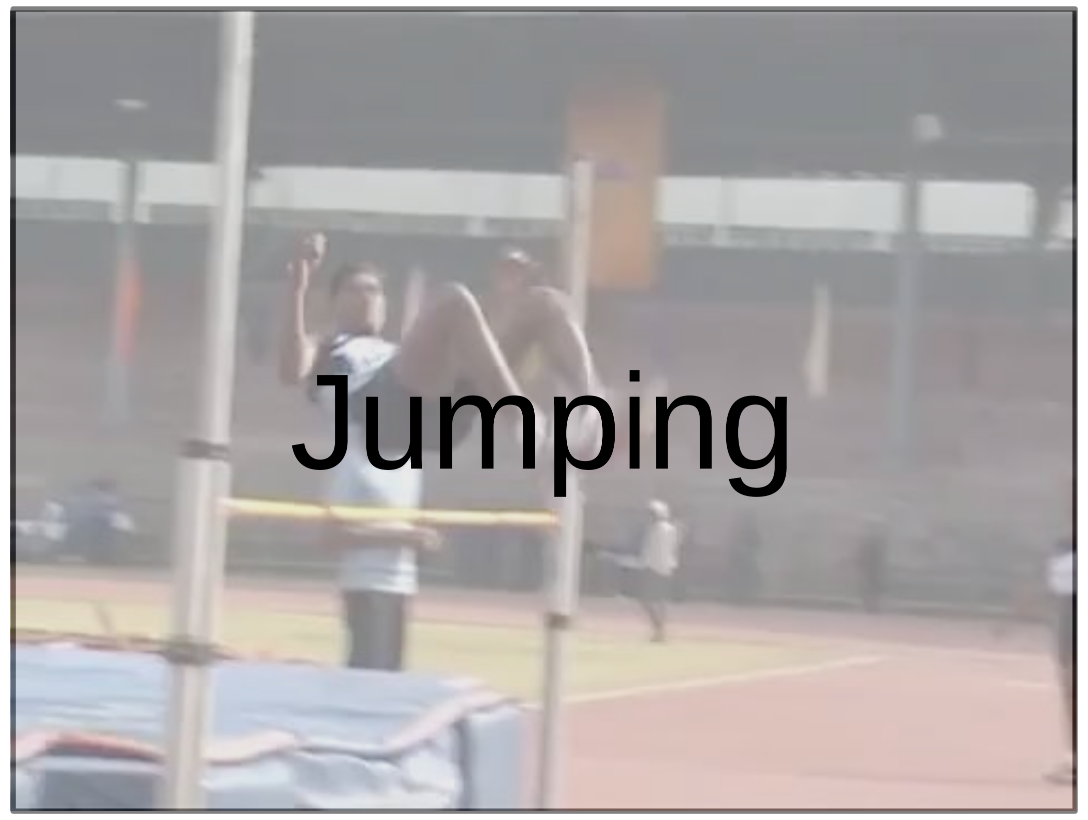
Video Understanding for Robotics

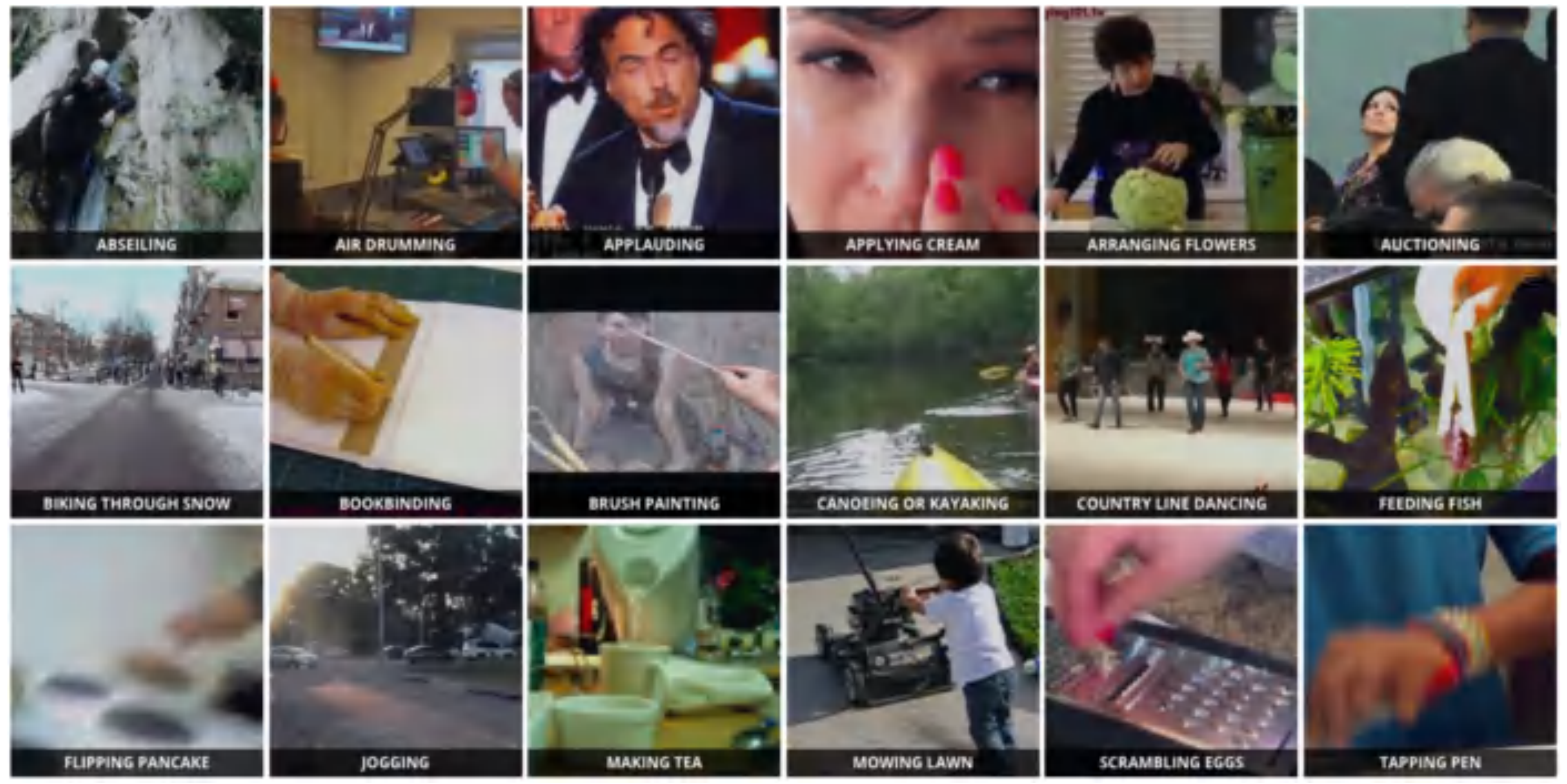
Xiaolong Wang
UC San Diego

An agent observes a dynamic world

Research in Videos: Activity Understanding

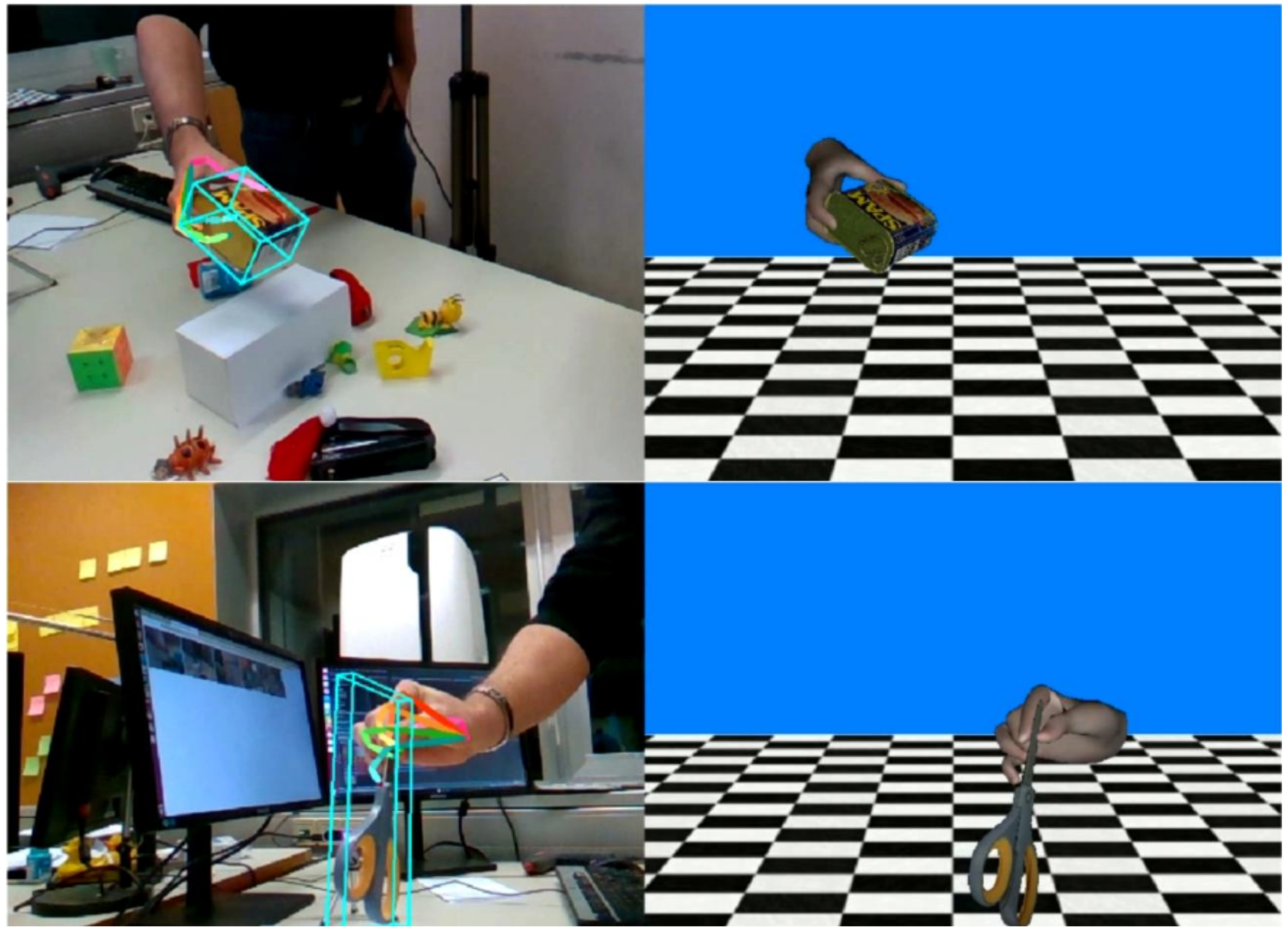


ActivityNet
300K videos

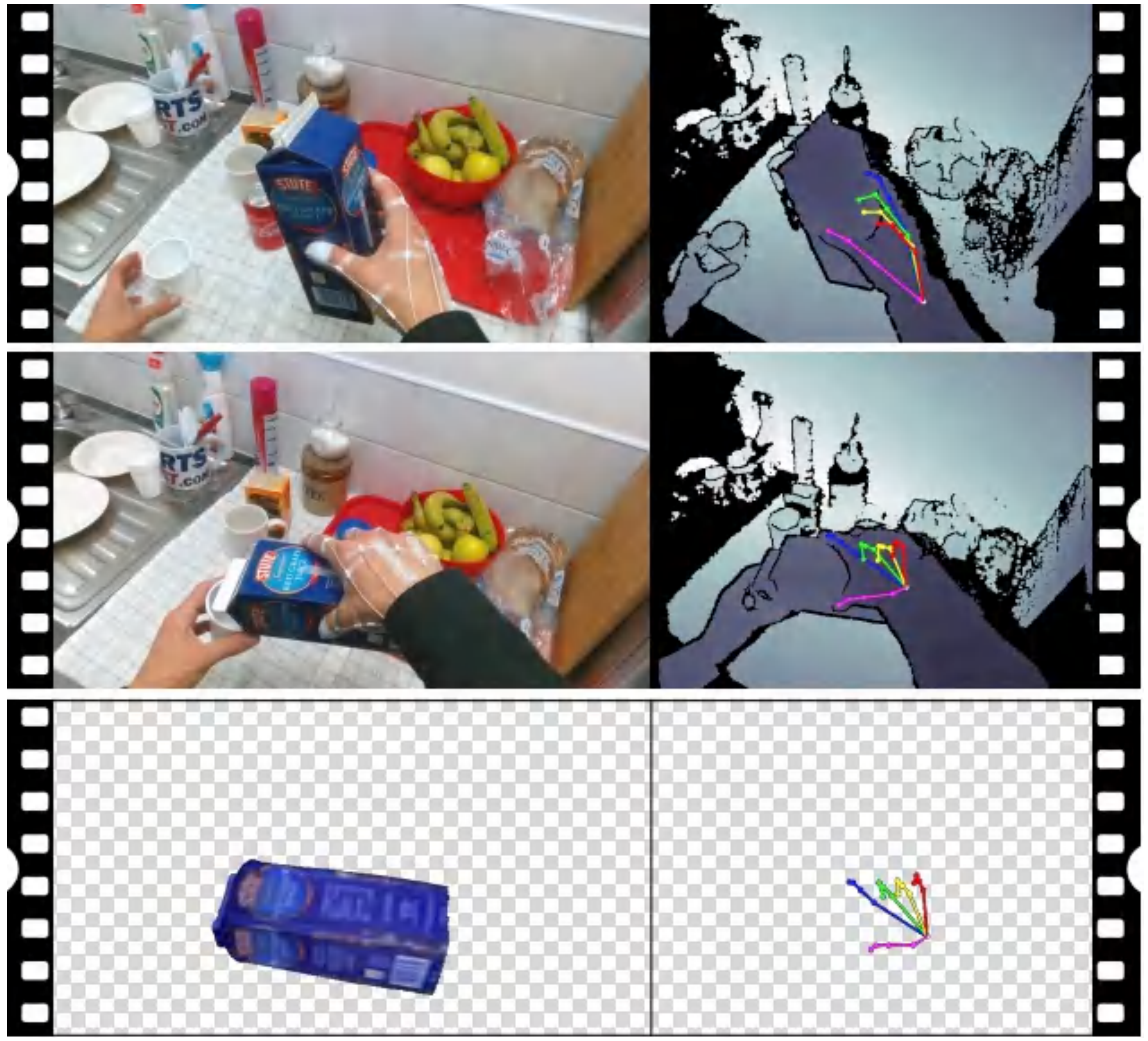


Kinetics
650K videos

Research in Videos: Perceiving 3D Structure



Hampali et al. 2019



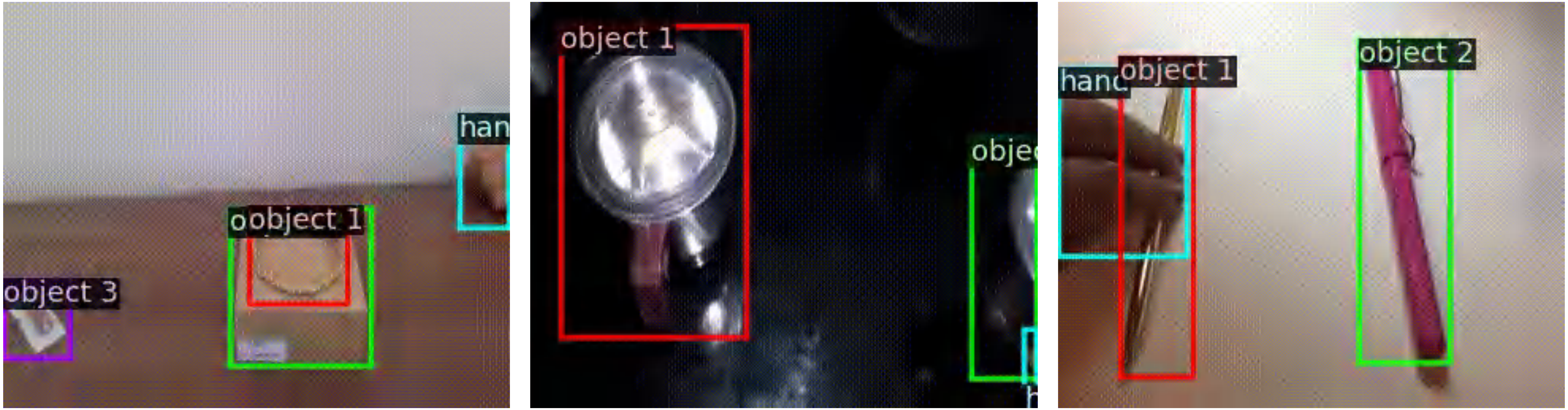
Garcia-Hernando et al. 2018

Video Understanding -> Imitation Learning

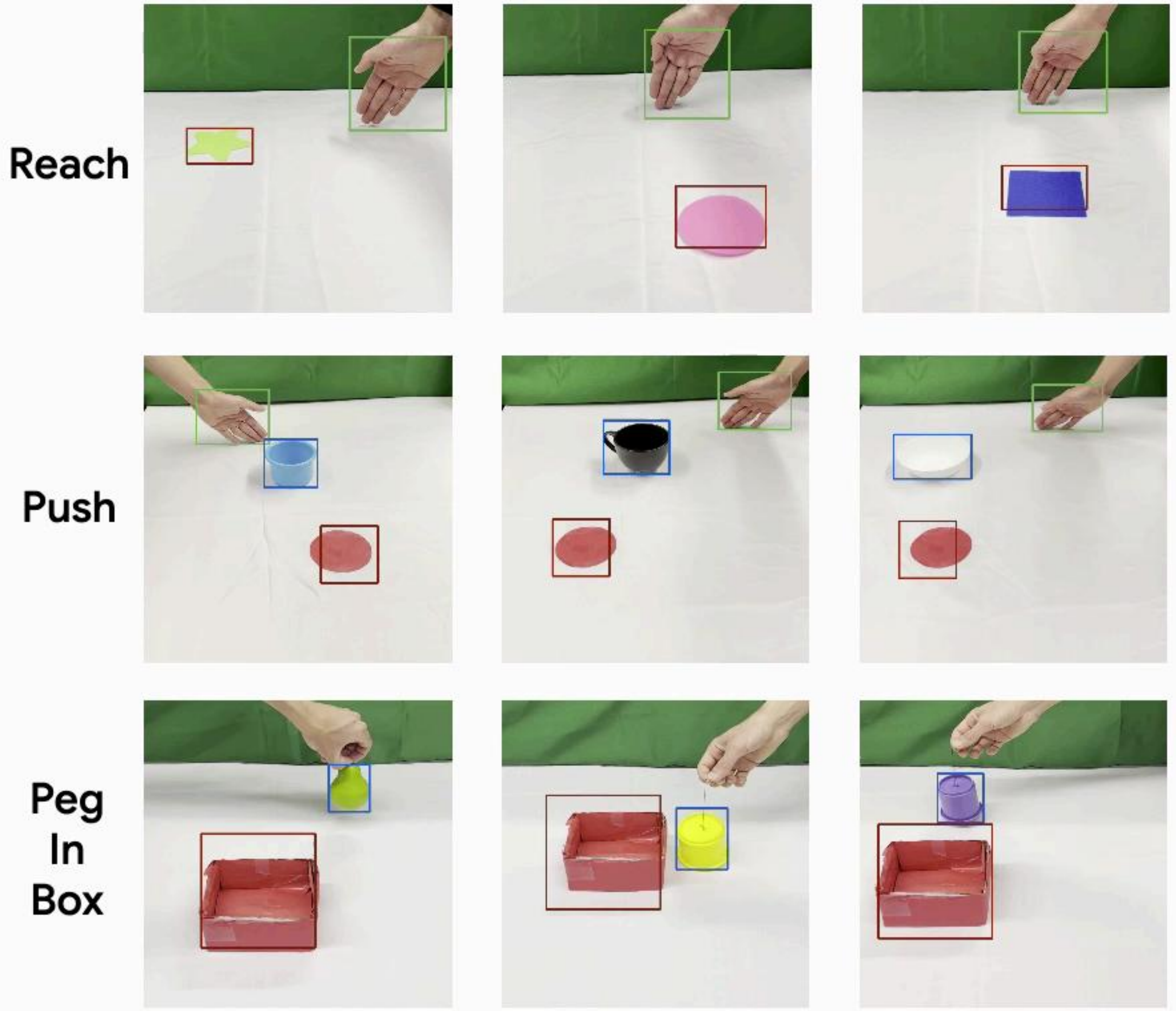
Space-Time and 3D Understanding

http://www.inclusivedesigntoolkit.com/UCdex/dex.html
Hasson, et al. "Learning joint reconstruction of hands and manipulated objects." CVPR 2019.
https://www.dlr.de/rm/en/desktopdefault.aspx/tabid-11886/#gallery/28915

Hand Object Interaction in Space-Time



Materzynska et al. Something-Else: Compositional Action Recognition with Spatial-Temporal Interaction Networks. CVPR 2020.



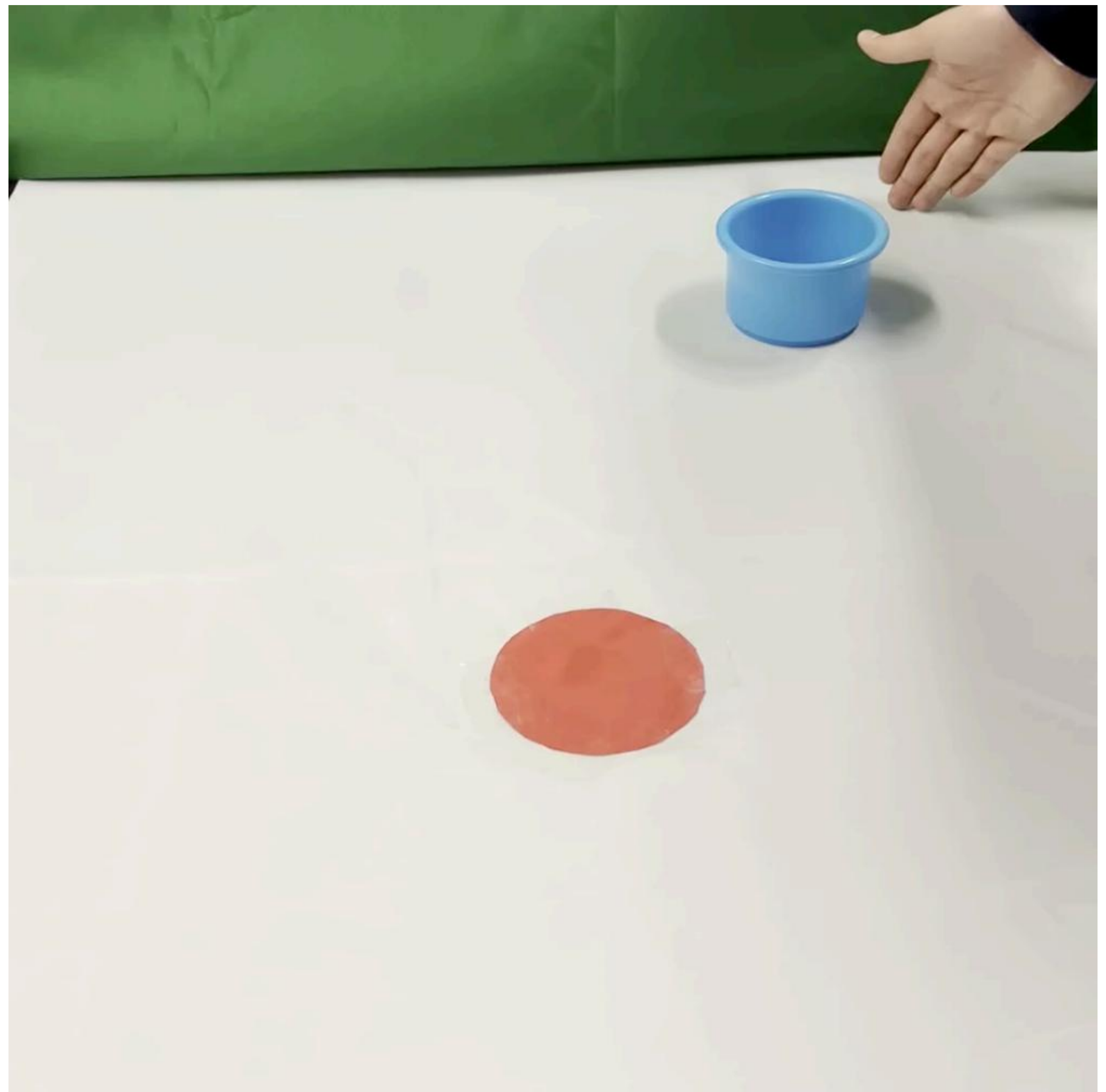
We learn a task reward with a graph abstraction from diverse videos.
No manual reward design is required for goal-conditioned RL.

Graph Inverse Reinforcement Learning from Diverse Videos
Sateesh Kumar, Jonathan Zamora, Nicklas Hansen*, Rishabh Jangir, Xiaolong Wang*
CoRL (Oral Presentation)

How are Rewards Obtained?

Computer Games

Real World



Directly obtained from environment

Often **manually** designed for each task separately

Can we learn rewards directly from Videos?

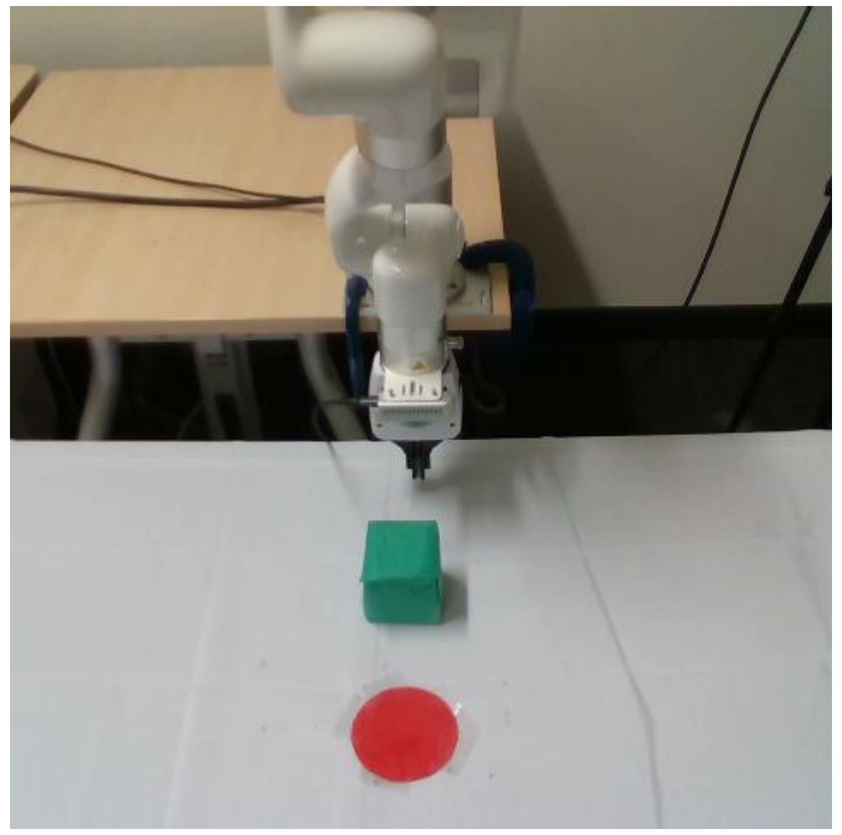
Collect Video Demonstrations

Learn Reward Function

Learn Reward from Videos

Reinforcement Learning

Learn Policy in Simulation

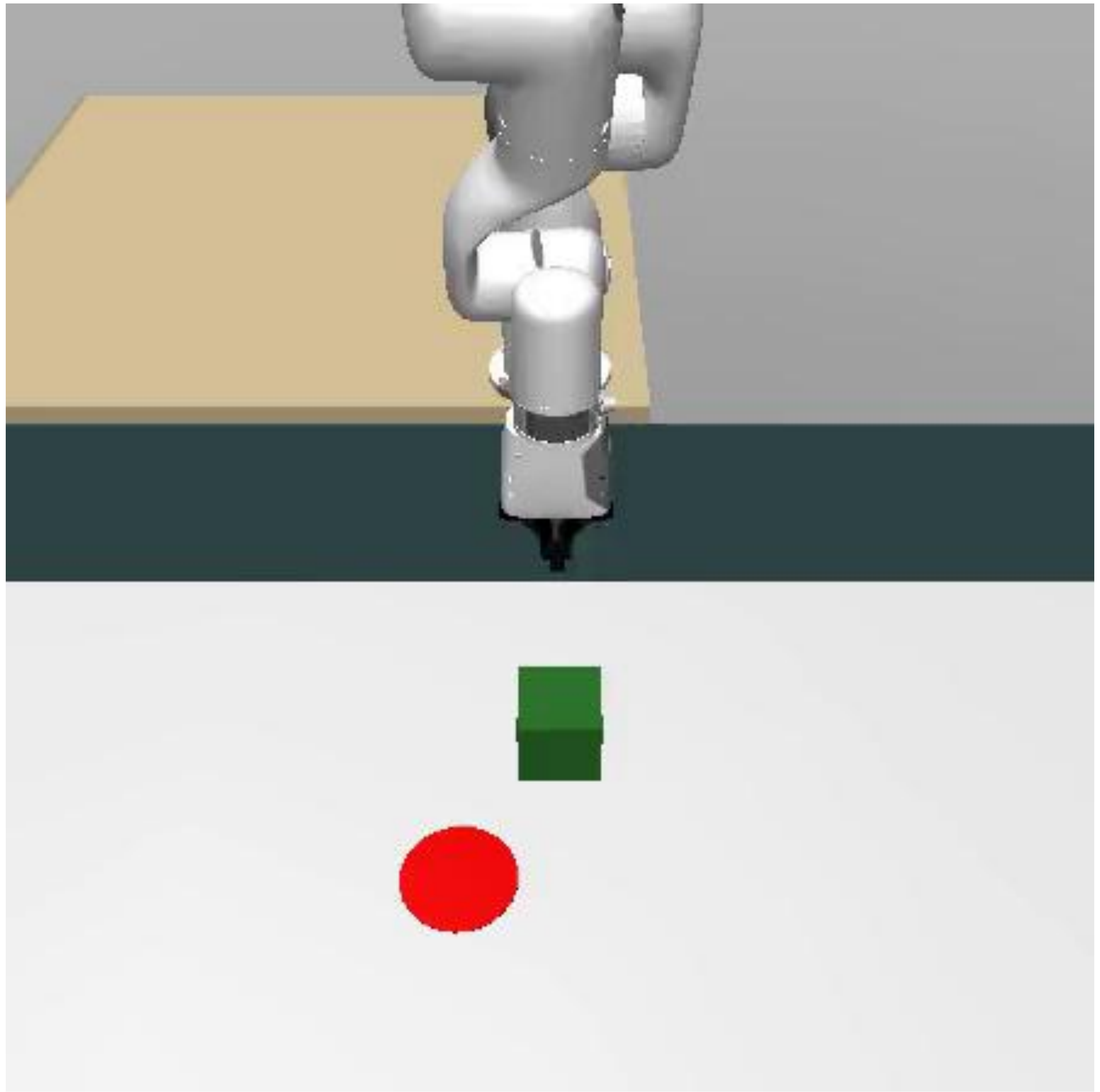


Deploy Learned Policy in Real

- Scalability
- Ease of data collection
- Unified pipeline

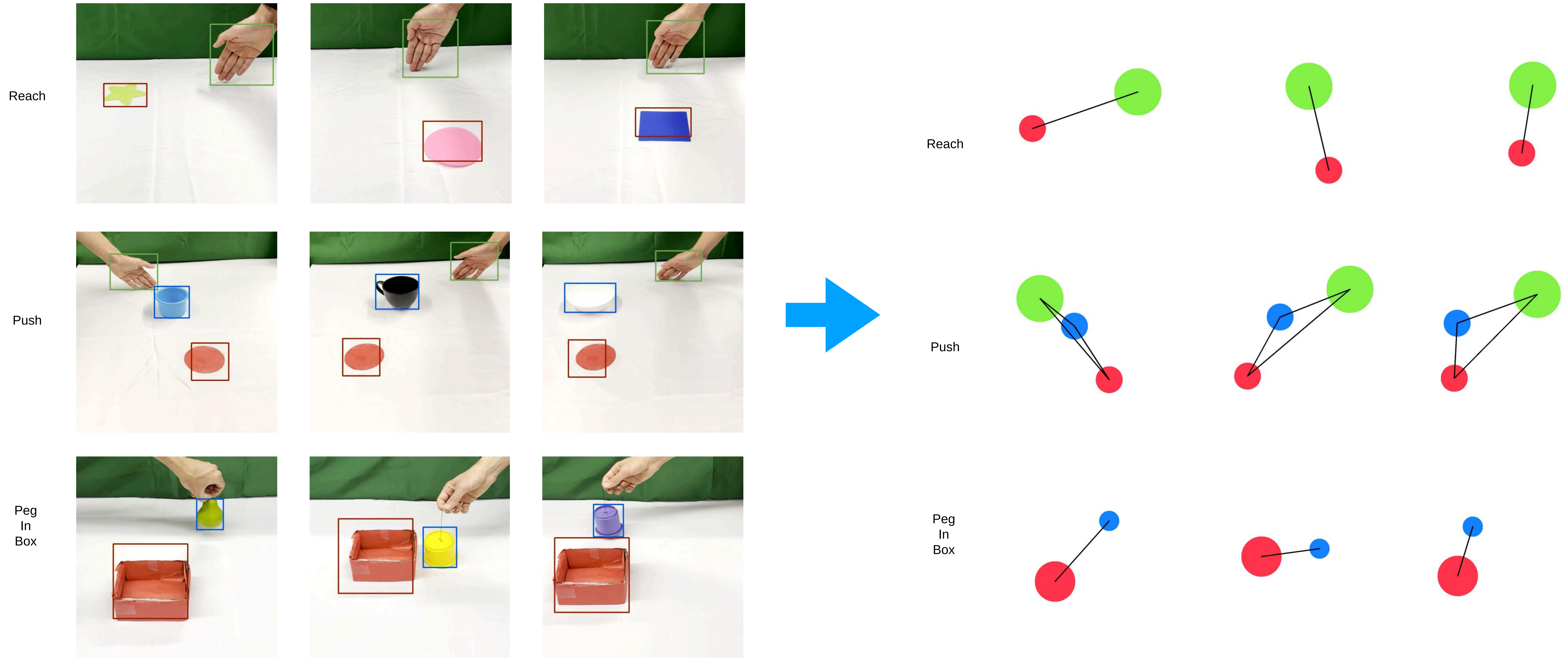
Domain Gap

Demonstrations



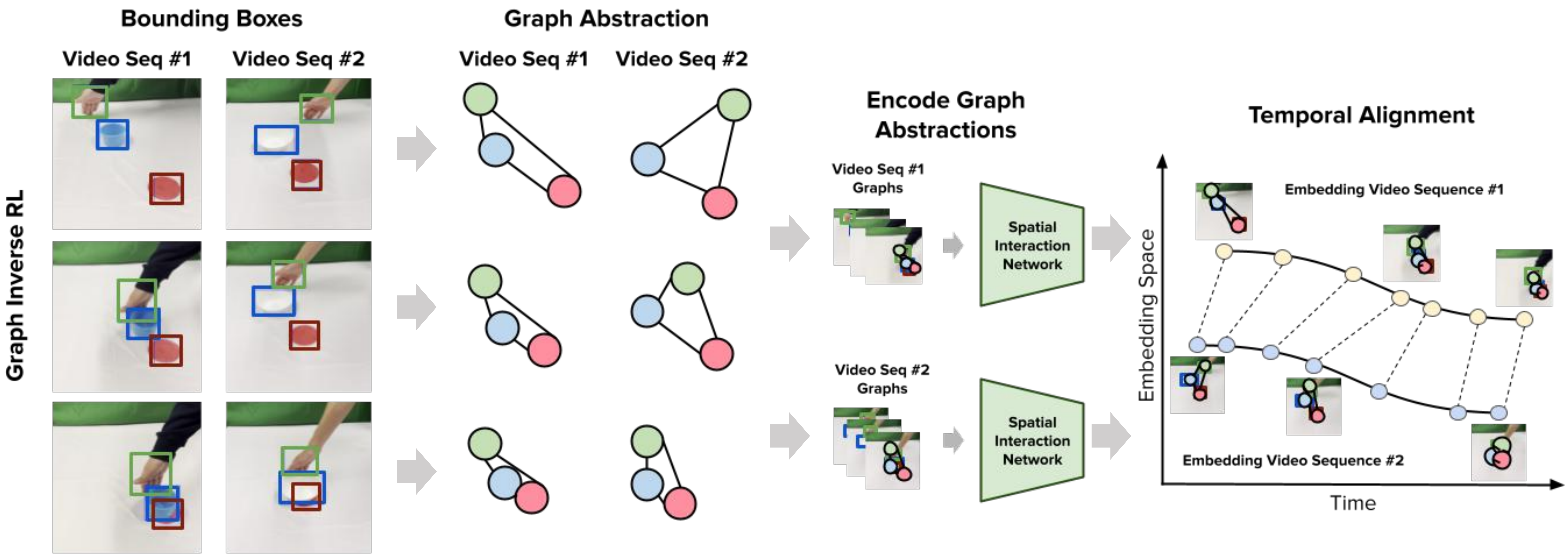
Simulation

Large variations in visual appearance, viewpoint, object shapes

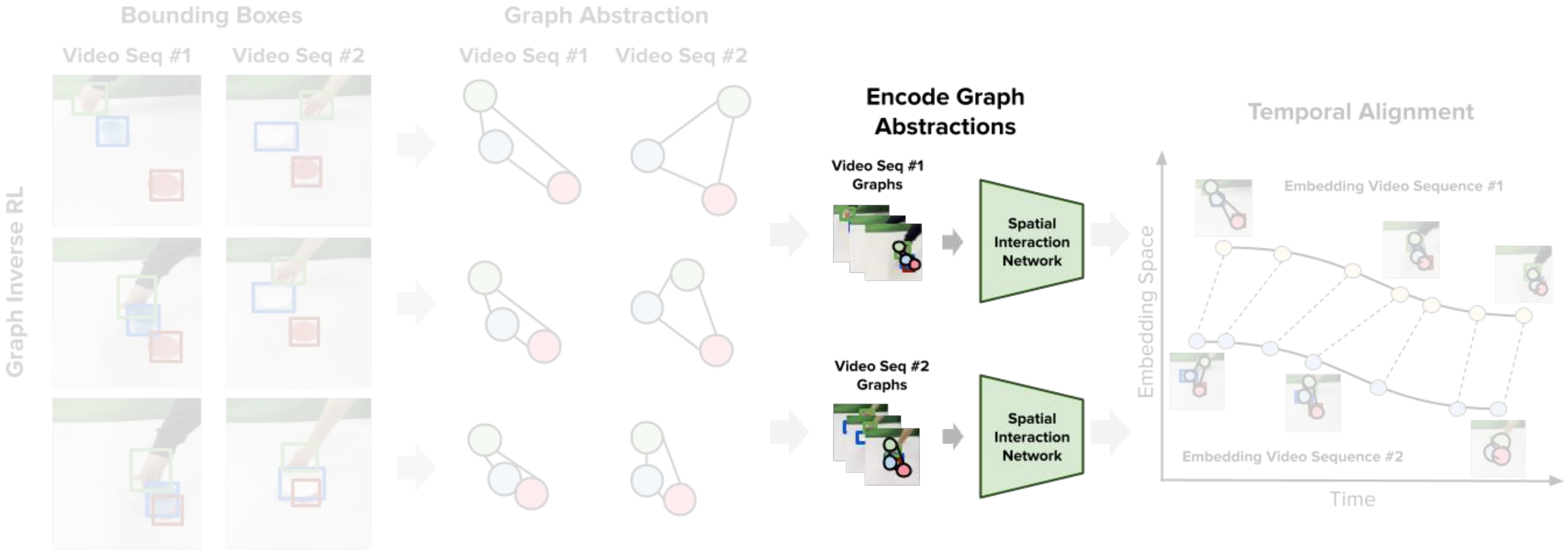


Despite large variance in videos, the underlying scene structure remains largely similar for manipulation tasks

The precise details of how the door is opened don’t matter, what matters is whether it is open



GraphIRL learns a task reward function via a graph abstraction through its 4 components


Graph Inverse RL
Bounding Boxes
Video Seq #1
Video Seq #2
Graph Abstraction
Video Seq #1
Video Seq #2
Encode Graph Abstractions
Video Seq #1 Graphs
Spatial Interaction Network
Video Seq #2 Graphs
Spatial Interaction Network
Temporal Alignment
Embedding Video Sequence #1
Embedding Space
Embedding Video Sequence #2
Time

Spatial Interaction Networks

The **self** representation of an object can be written as:

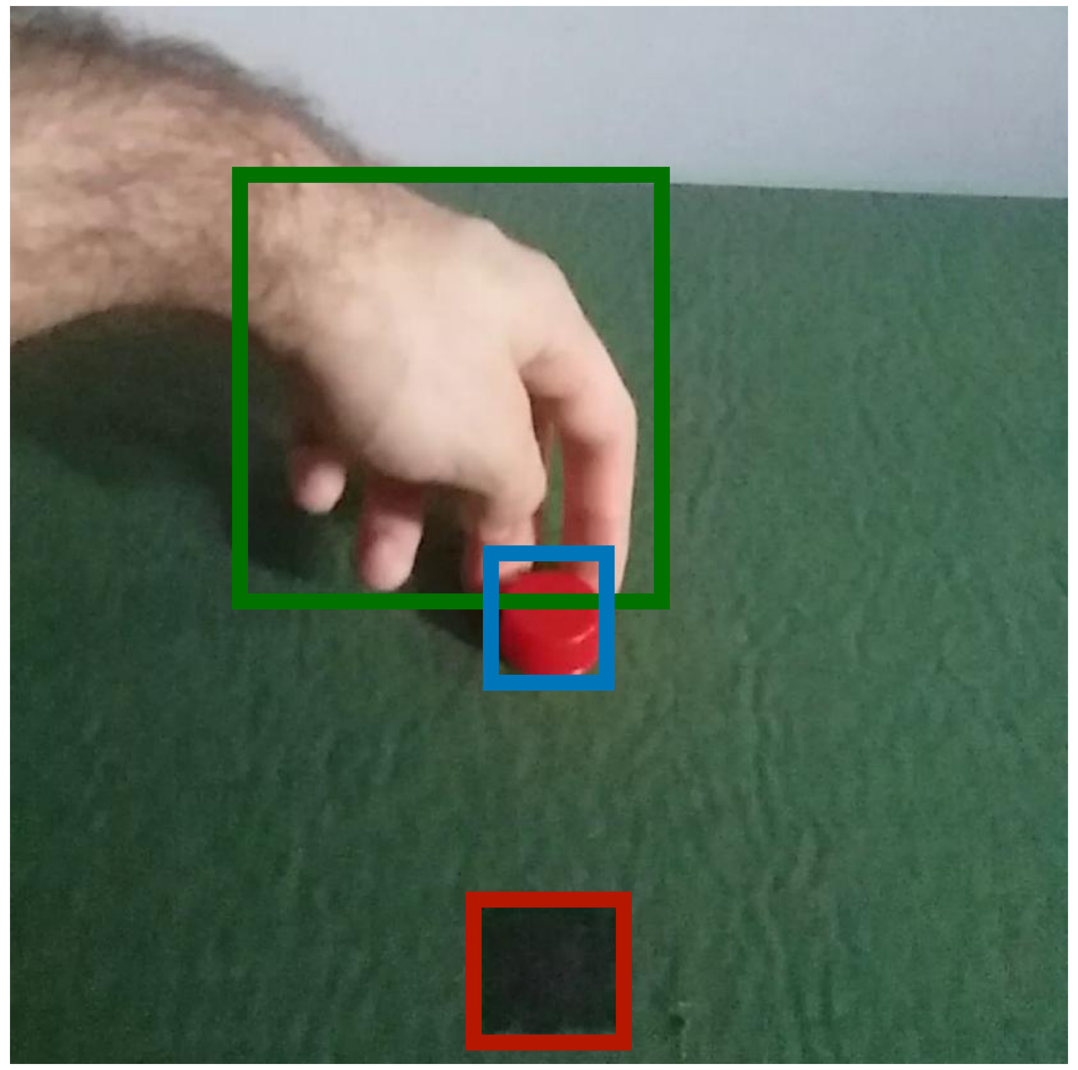
$$f_s(o_i) = \phi_s(o)$$

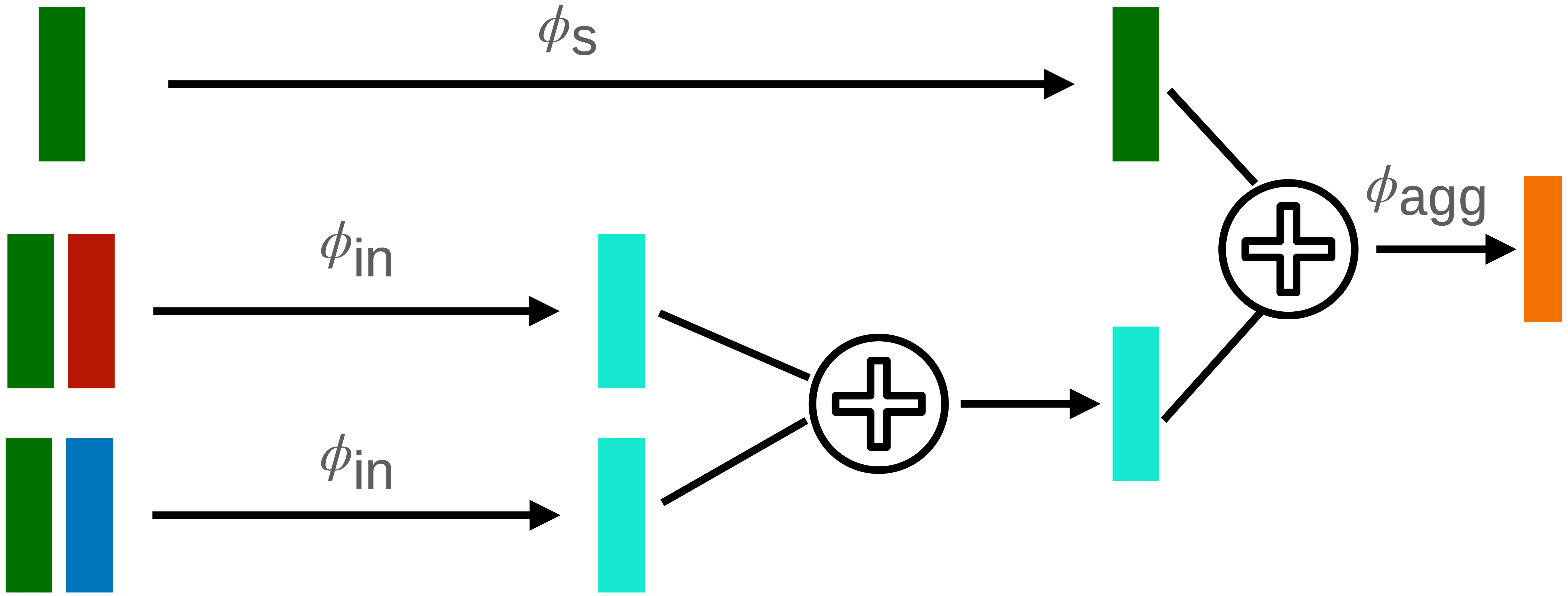
Similarly, the **interactional** representation of an object is:

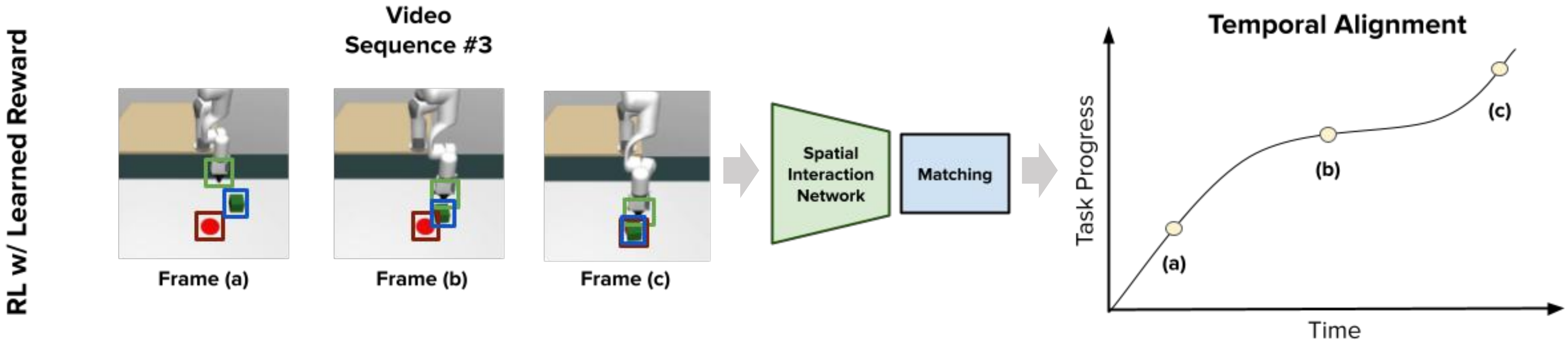
$$\sum_{j=1}^{m} \phi_{\text{in}}((o_i, o_j))$$

The final representation corresponding to a frame is:

$$f_o(o_i) = \phi_{\text{agg}}(f_{\text{s}} + f_{\text{in}})$$







The learned reward function is then used for Reinforcement Learning

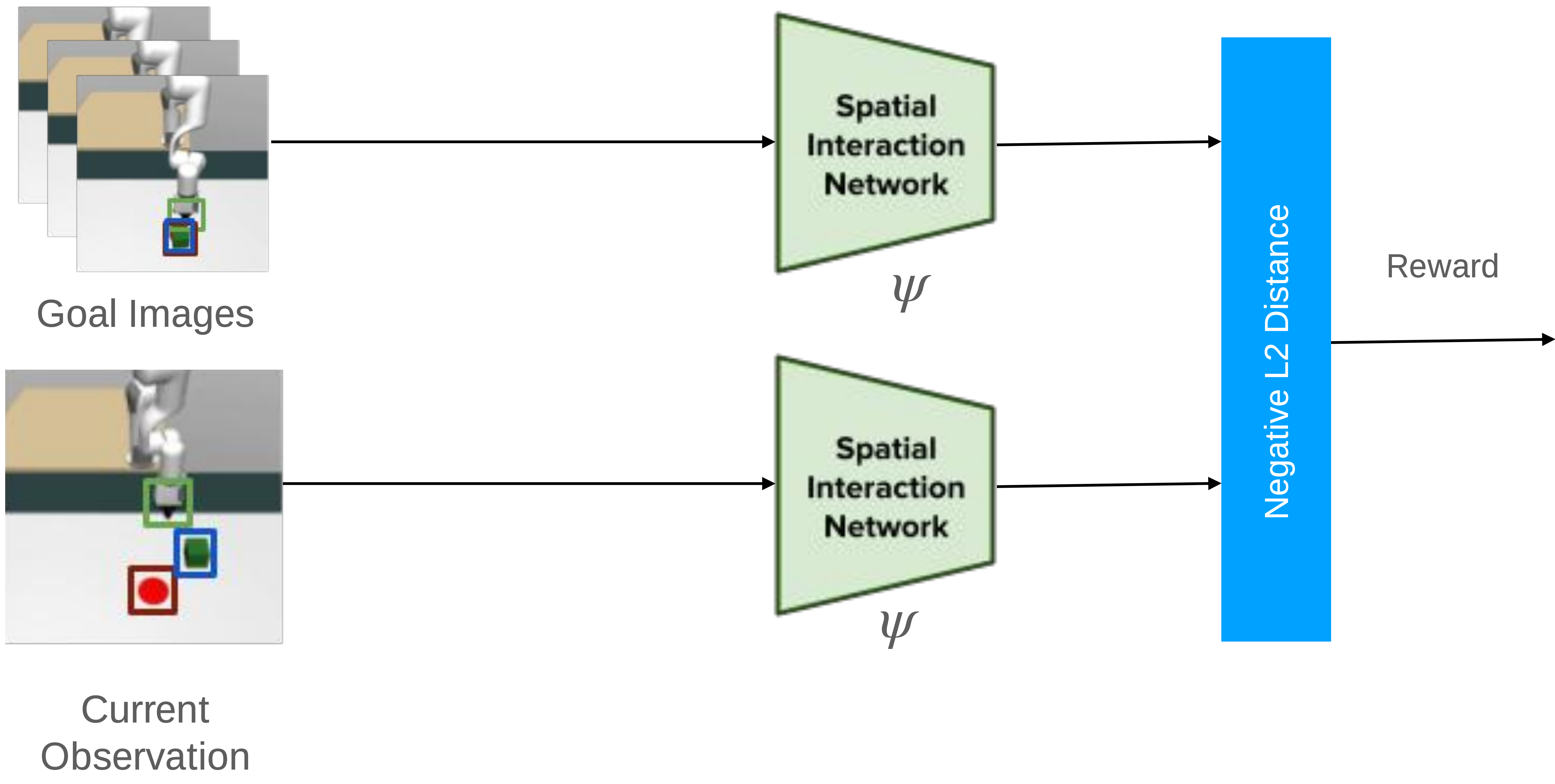
Learned Representations to Reward

- Representative goal-frame embedding:

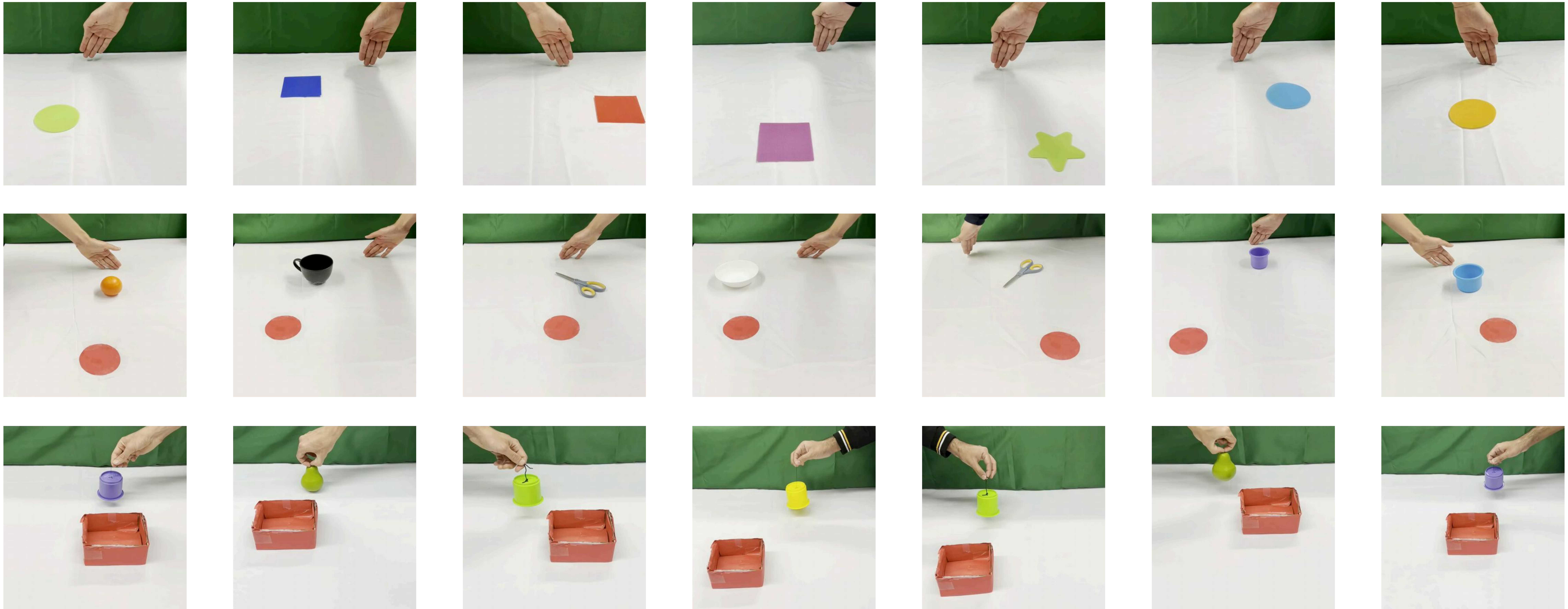
$$g = \sum_{i=1}^{n} \psi(I'^{i}_{m_i})$$

- The reward can be constructed as:

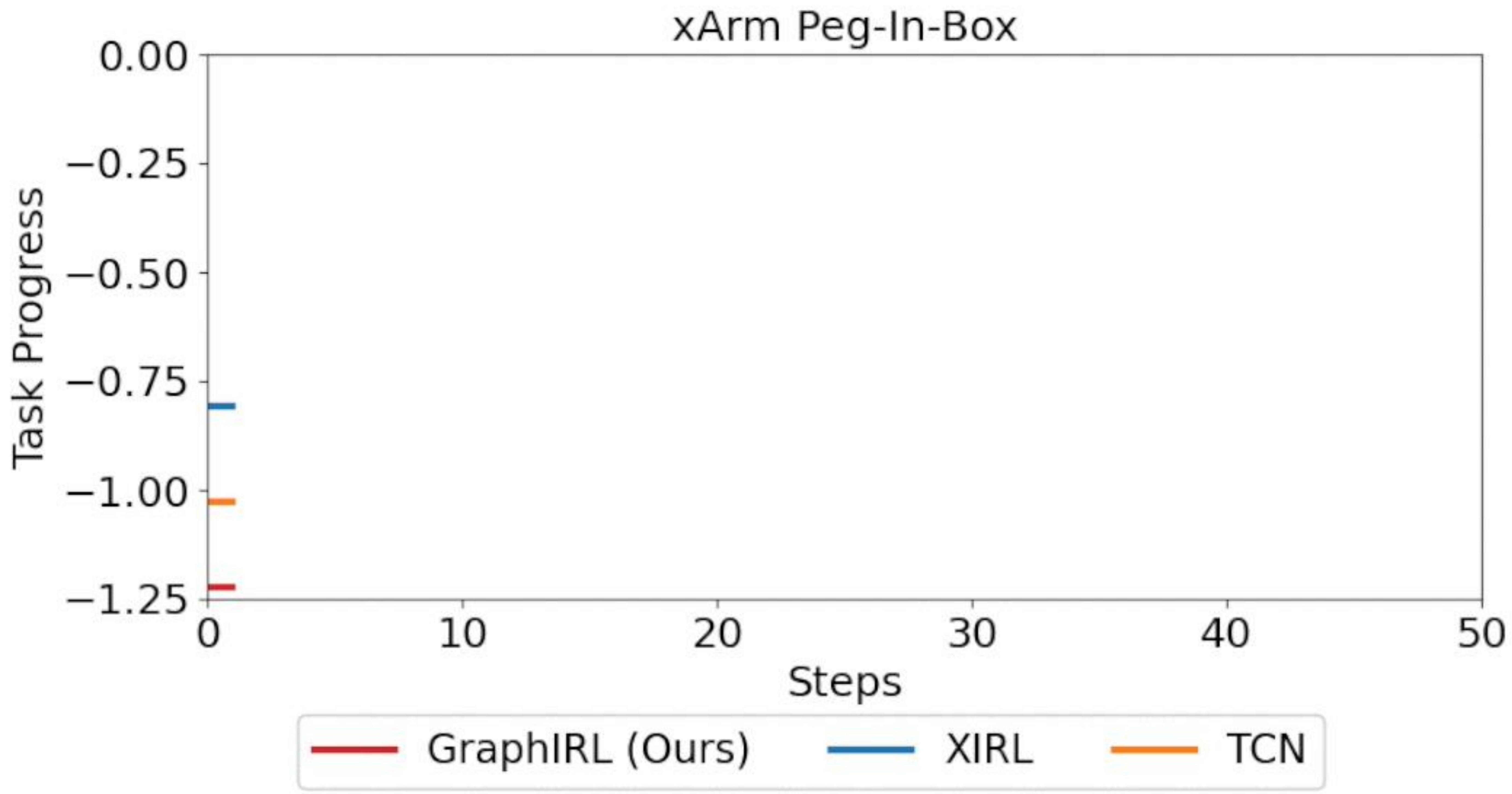
$$R = -\frac{1}{c} || \psi(o) - g ||^2$$

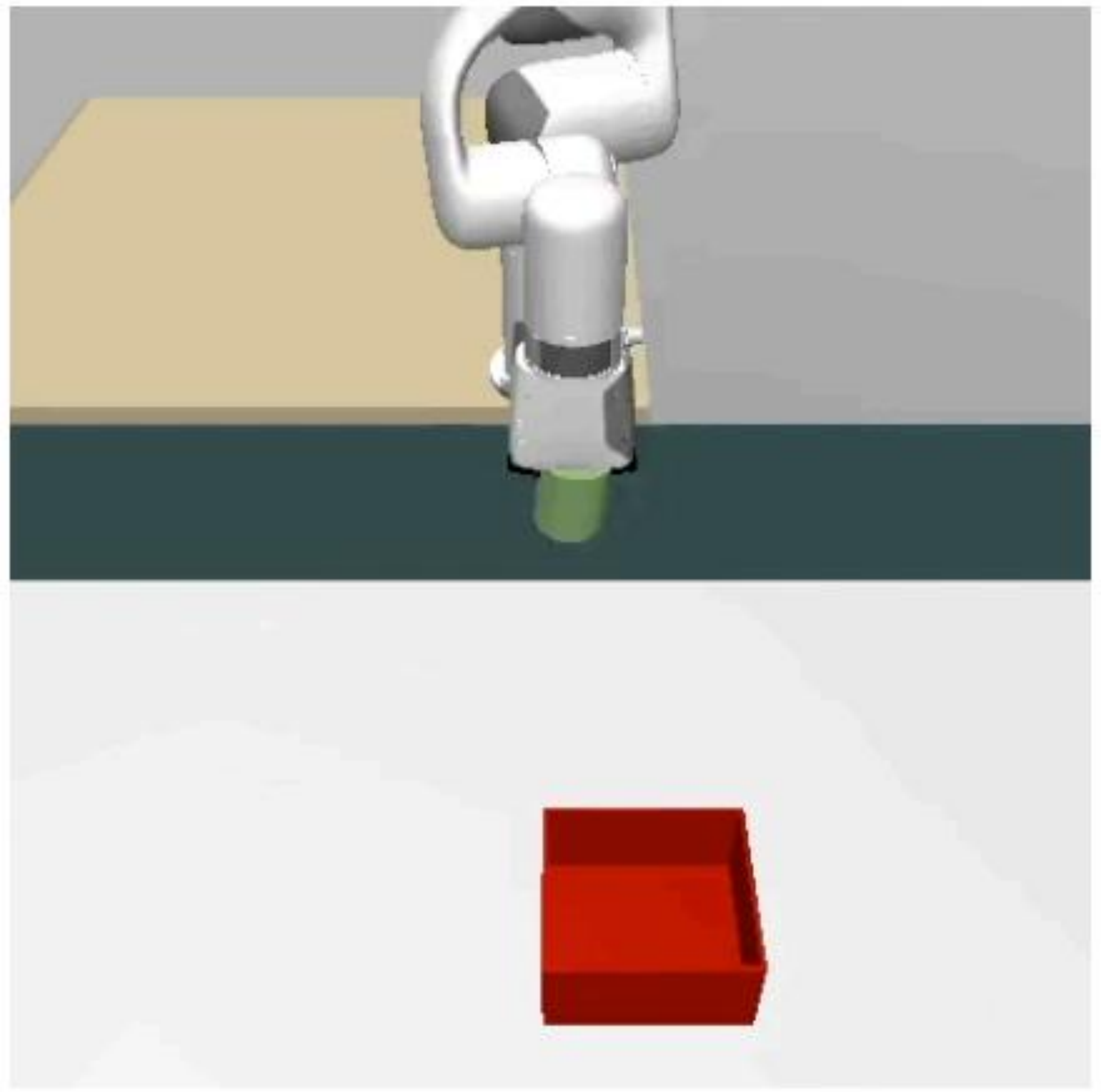


Diverse Demonstrations for Reward Learning



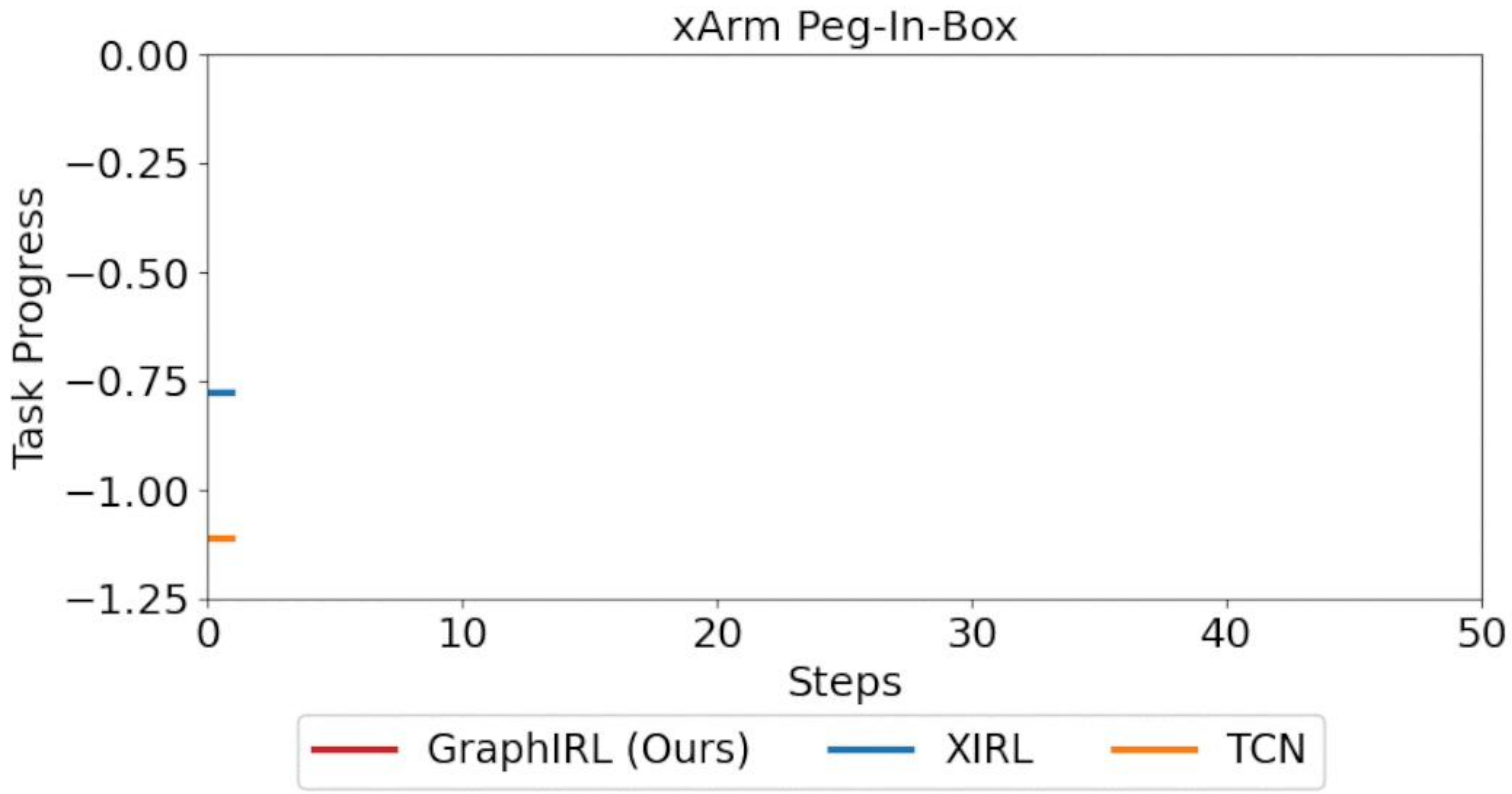
Successful Trial #1 ✅

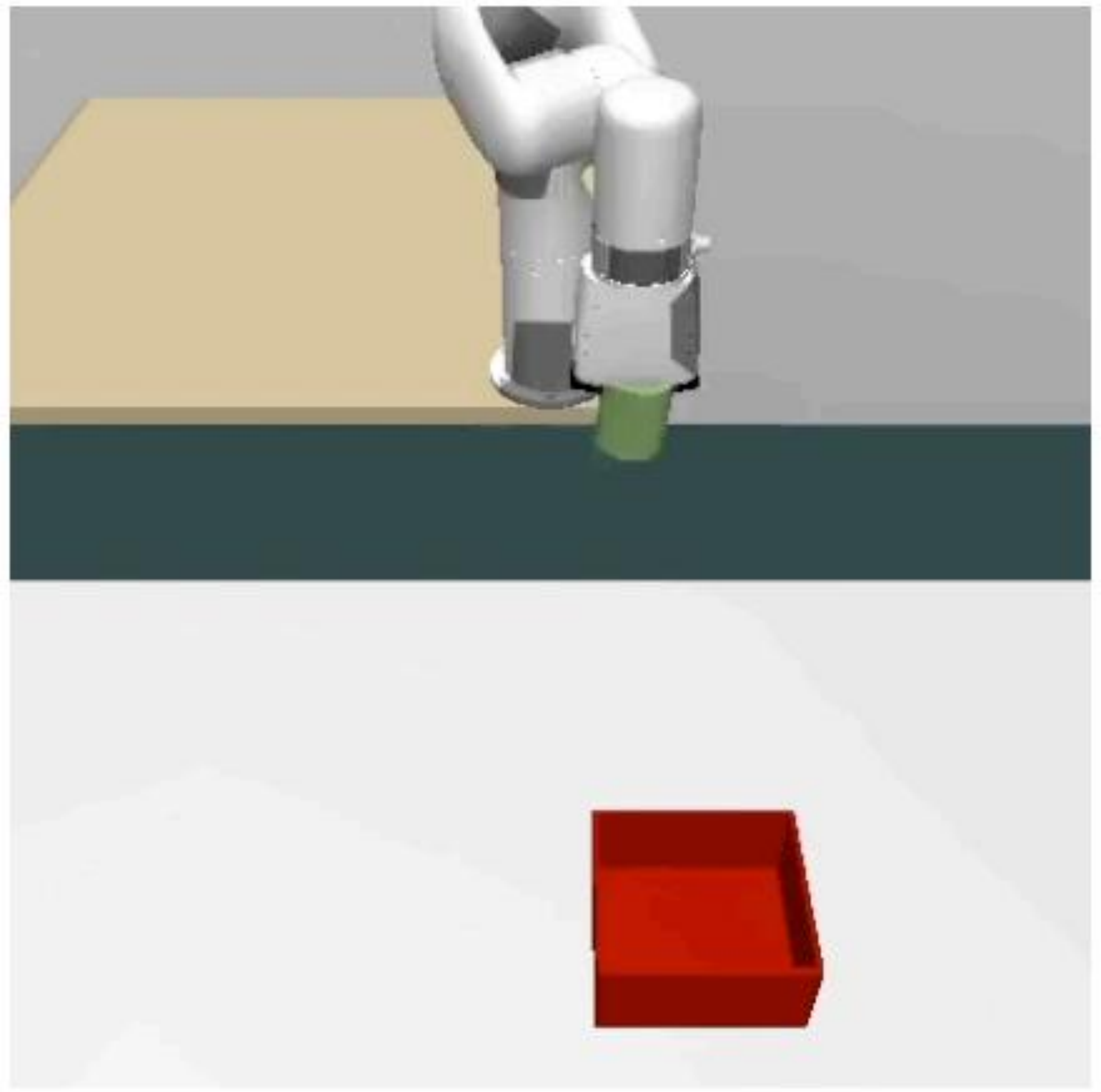




Peg
In
Box

Successful Trial #2 ✅

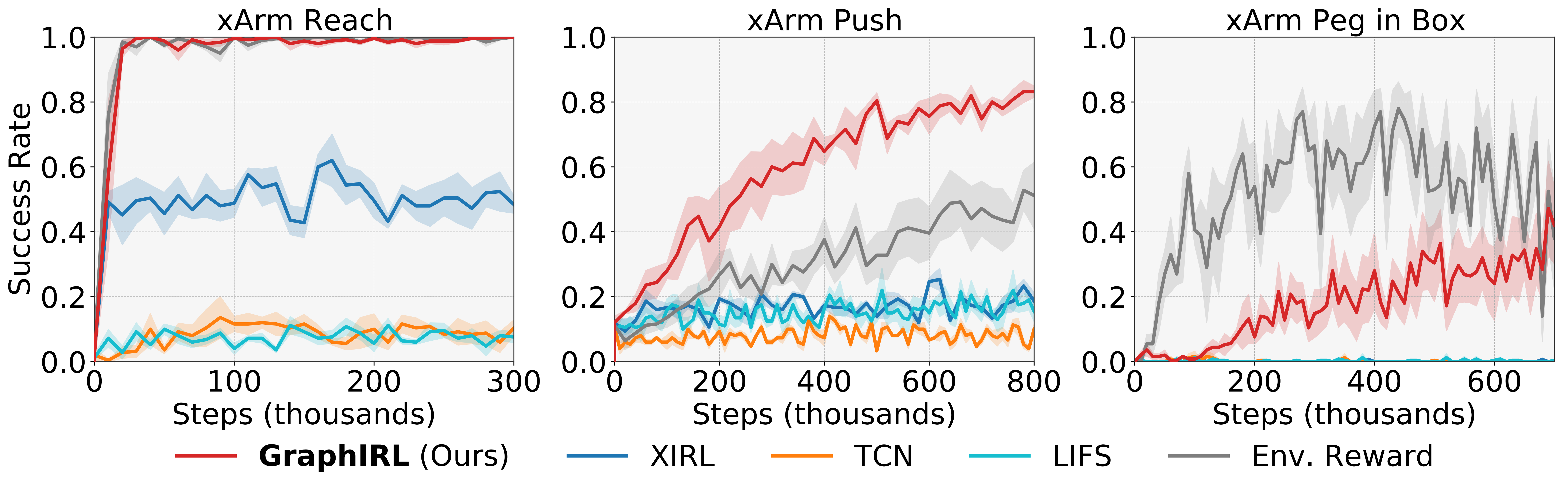




Peg In Box

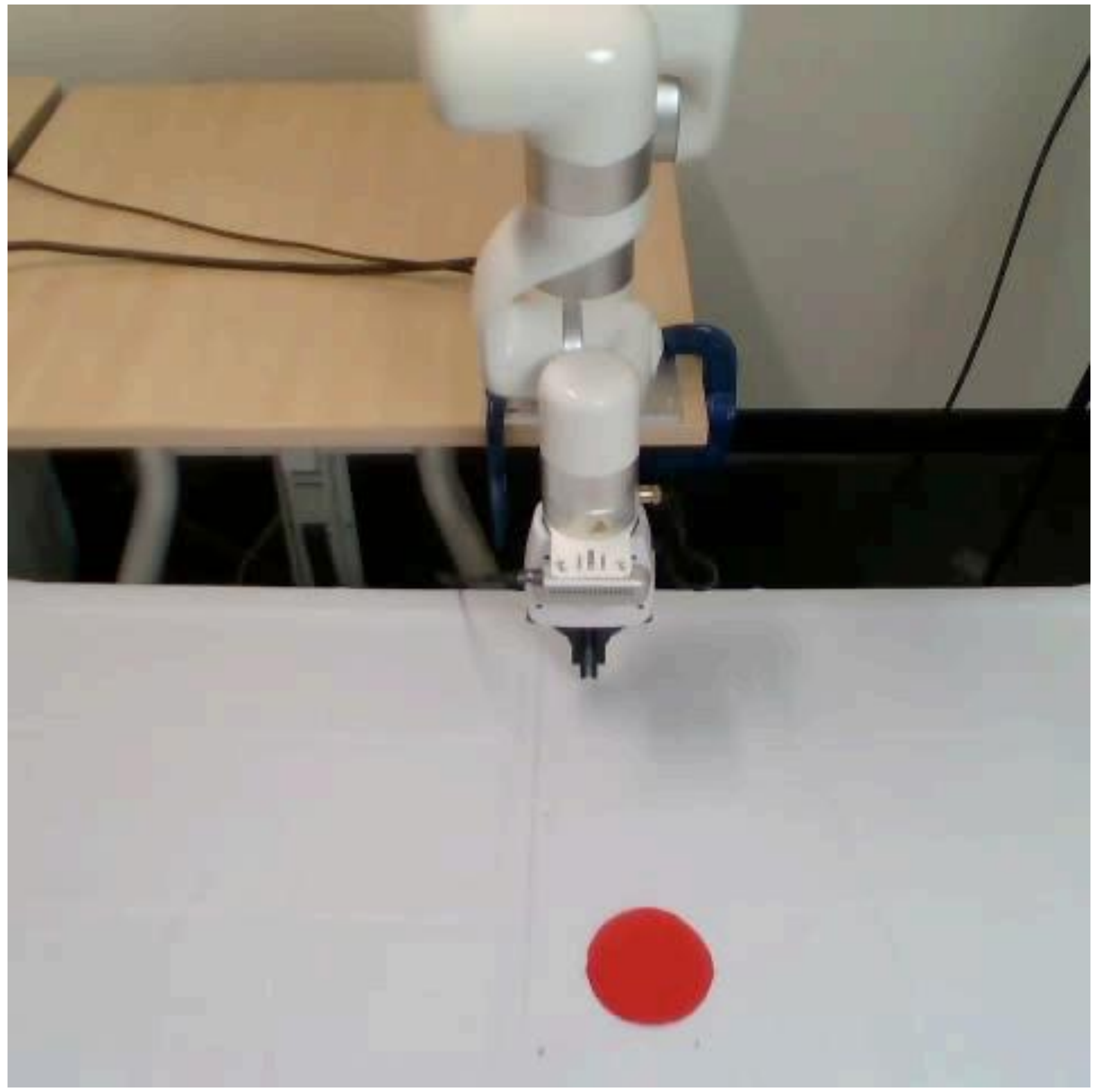
Robot Manipulation in Simulation

- GraphIRL outperforms Vision-based baselines by upto 40%
- GraphIRL solves all tasks without using any task-specific task reward



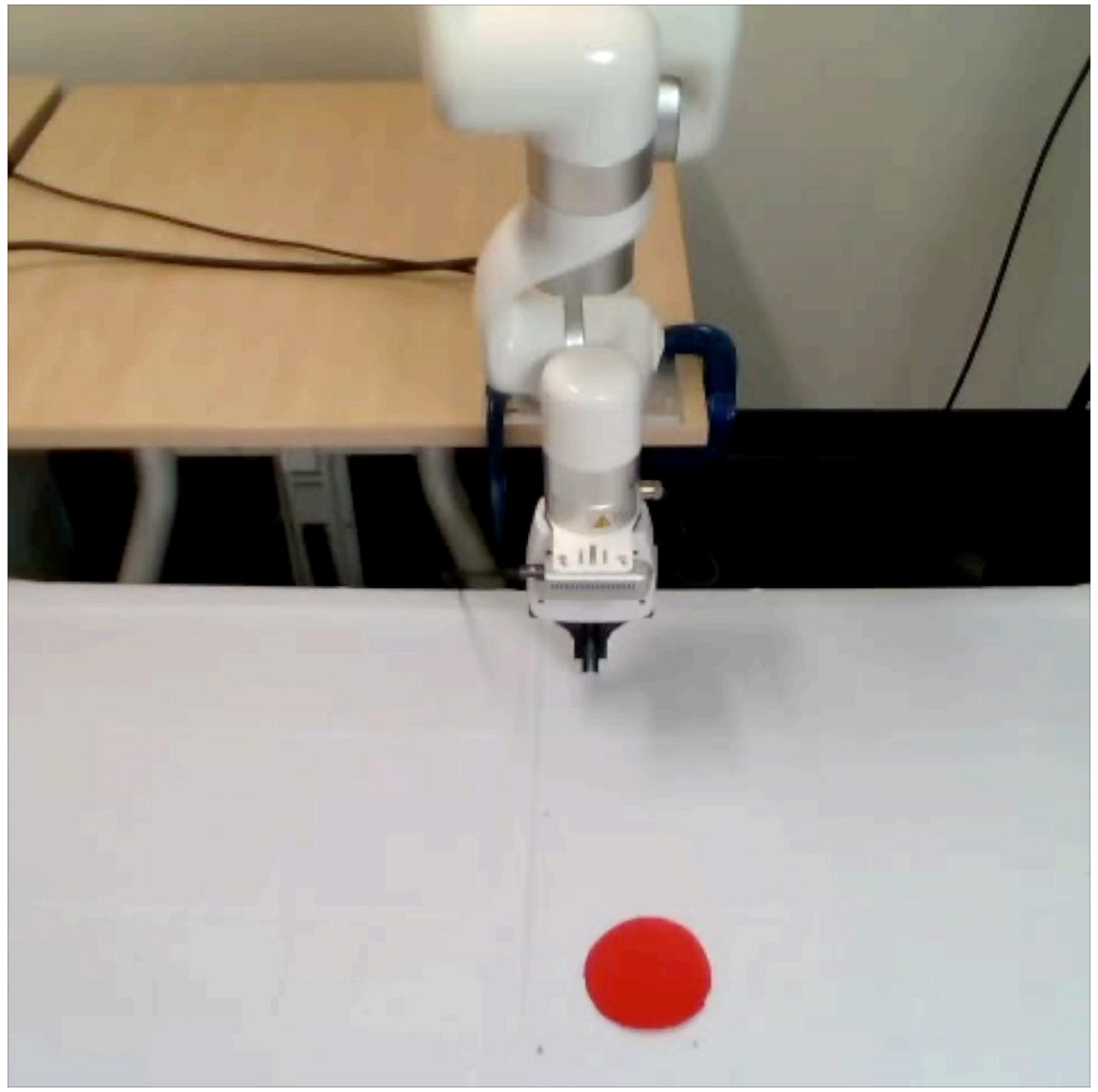
Task: Reach

XIRL
[Zakka et al., 2022]



Success Rate: 26%

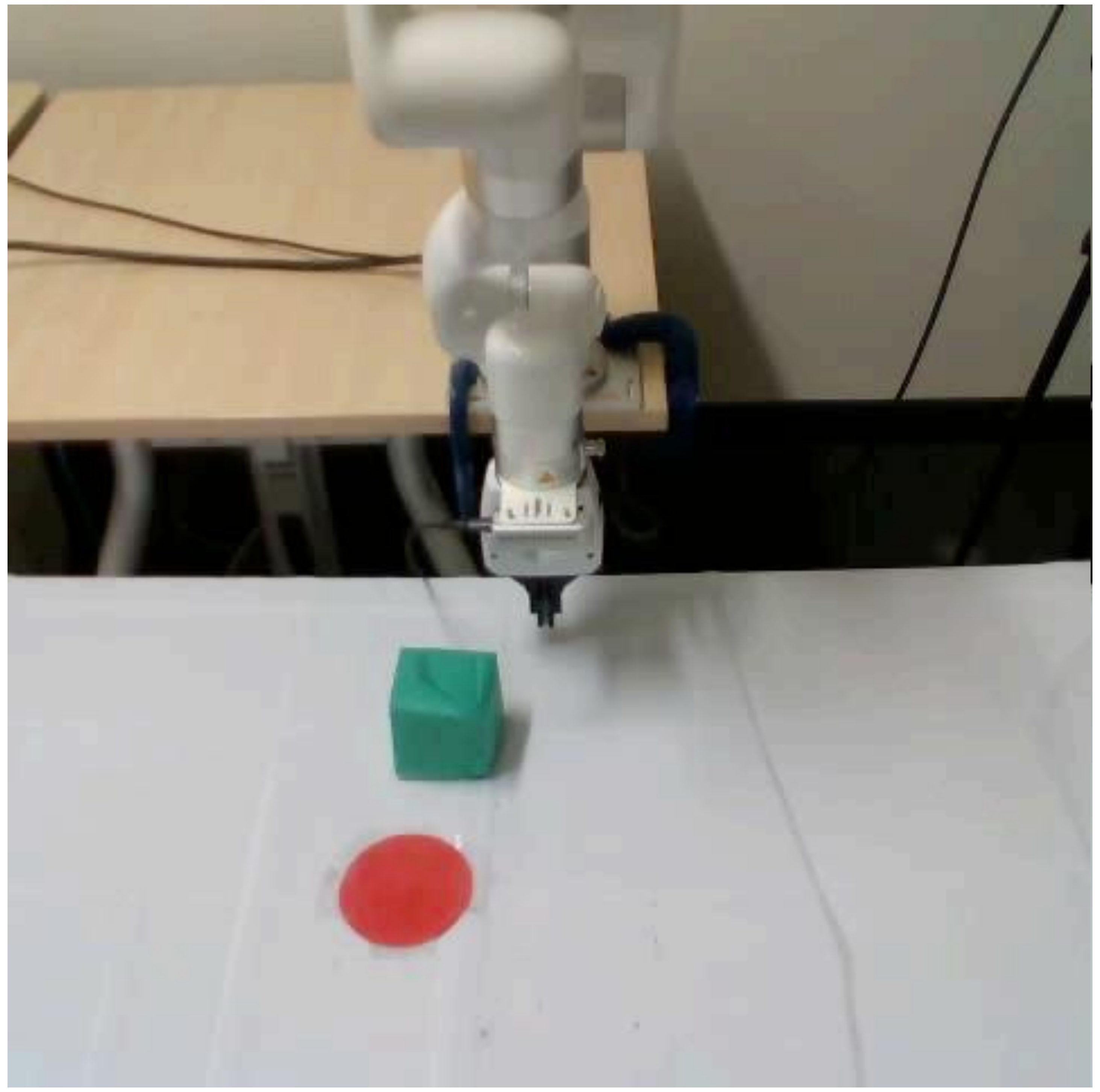
GraphIRL
[Ours]



Success Rate: 86%

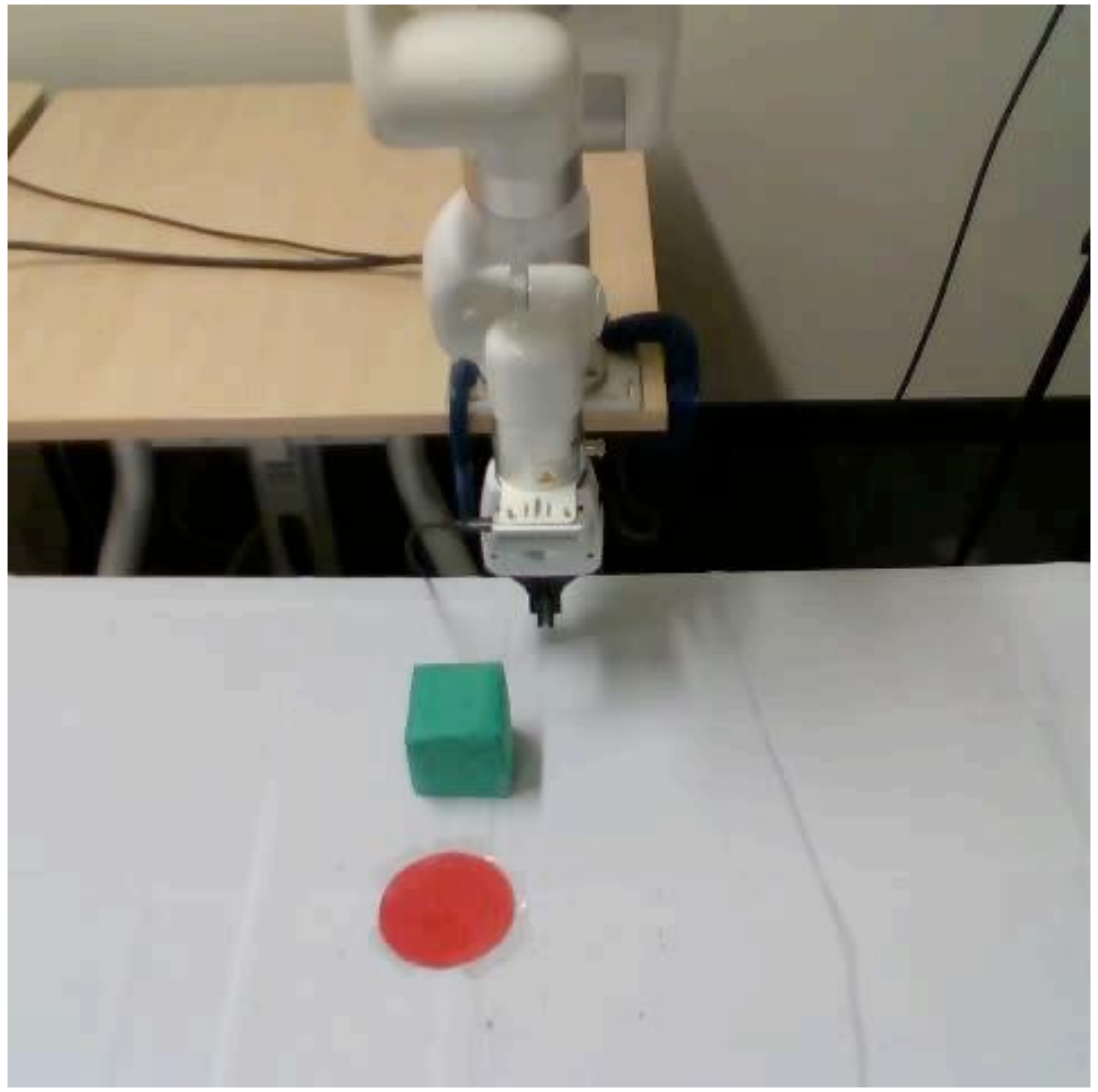
Task: Push

XIRL
[Zakka et al., 2022]



Success Rate: 27%

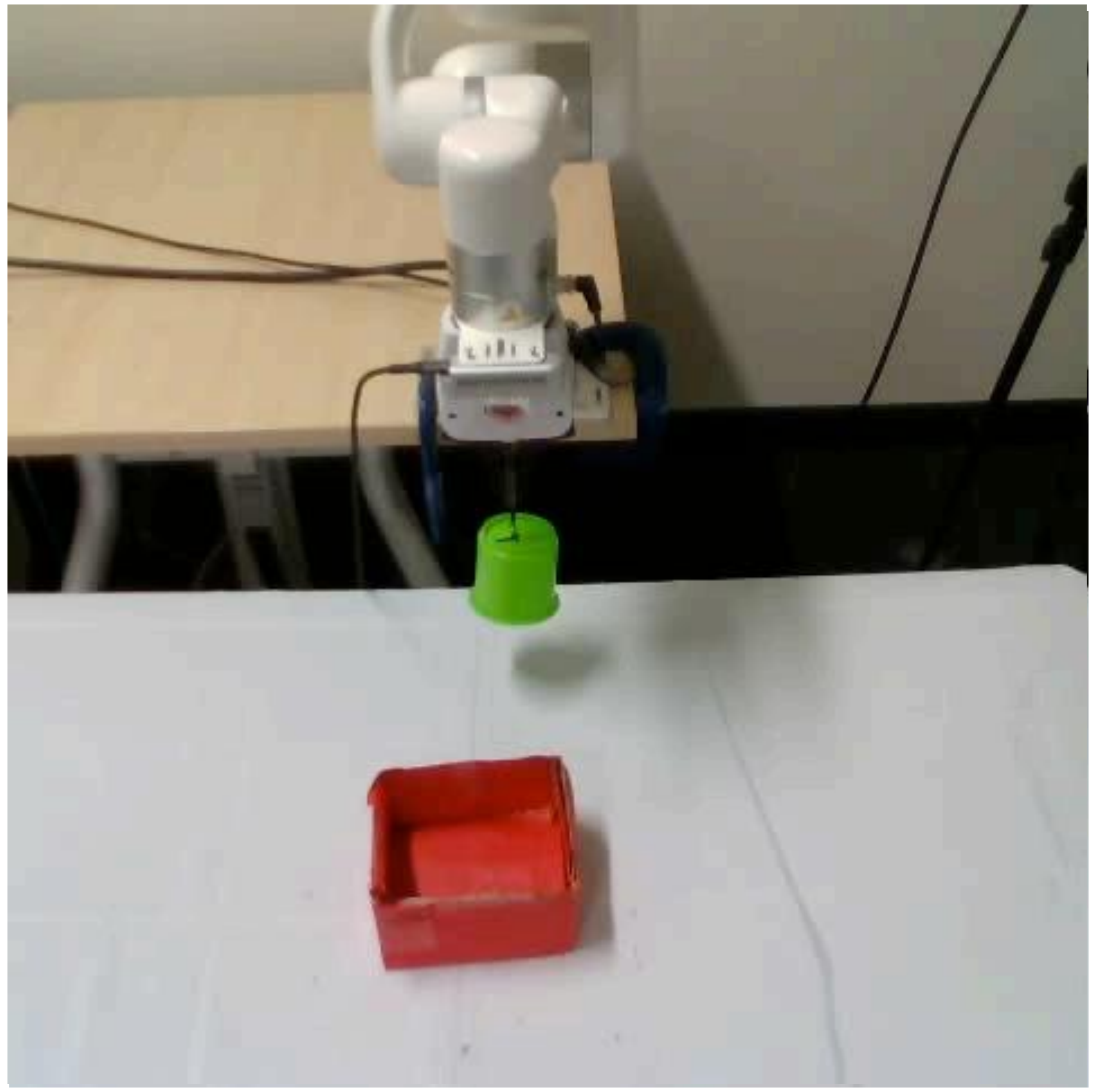
GraphIRL
[Ours]



Success Rate: 60%

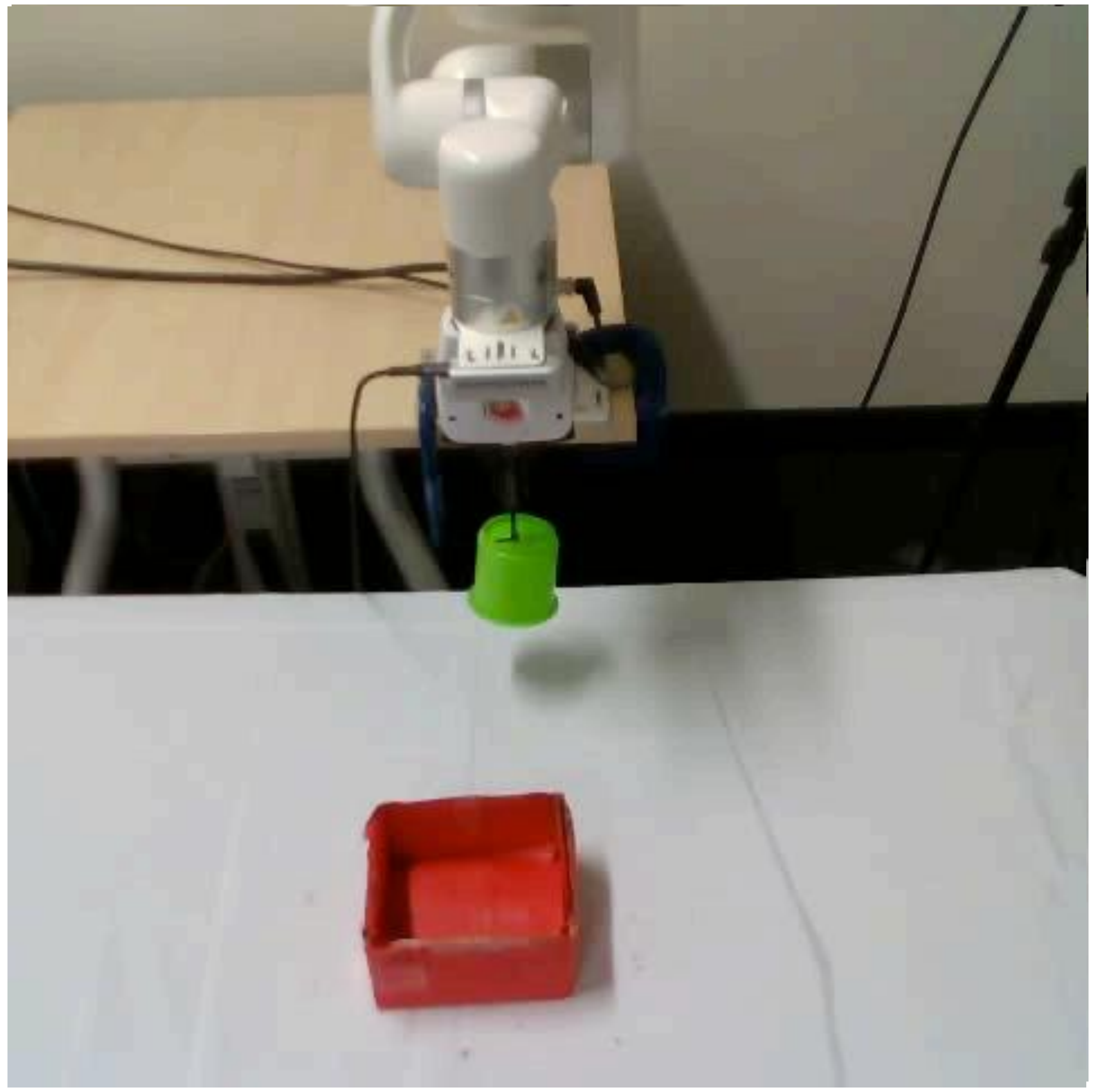
Task: Peg in Box

XIRL
[Zakka et al., 2022]



Success Rate: 6%

GraphIRL
[Ours]

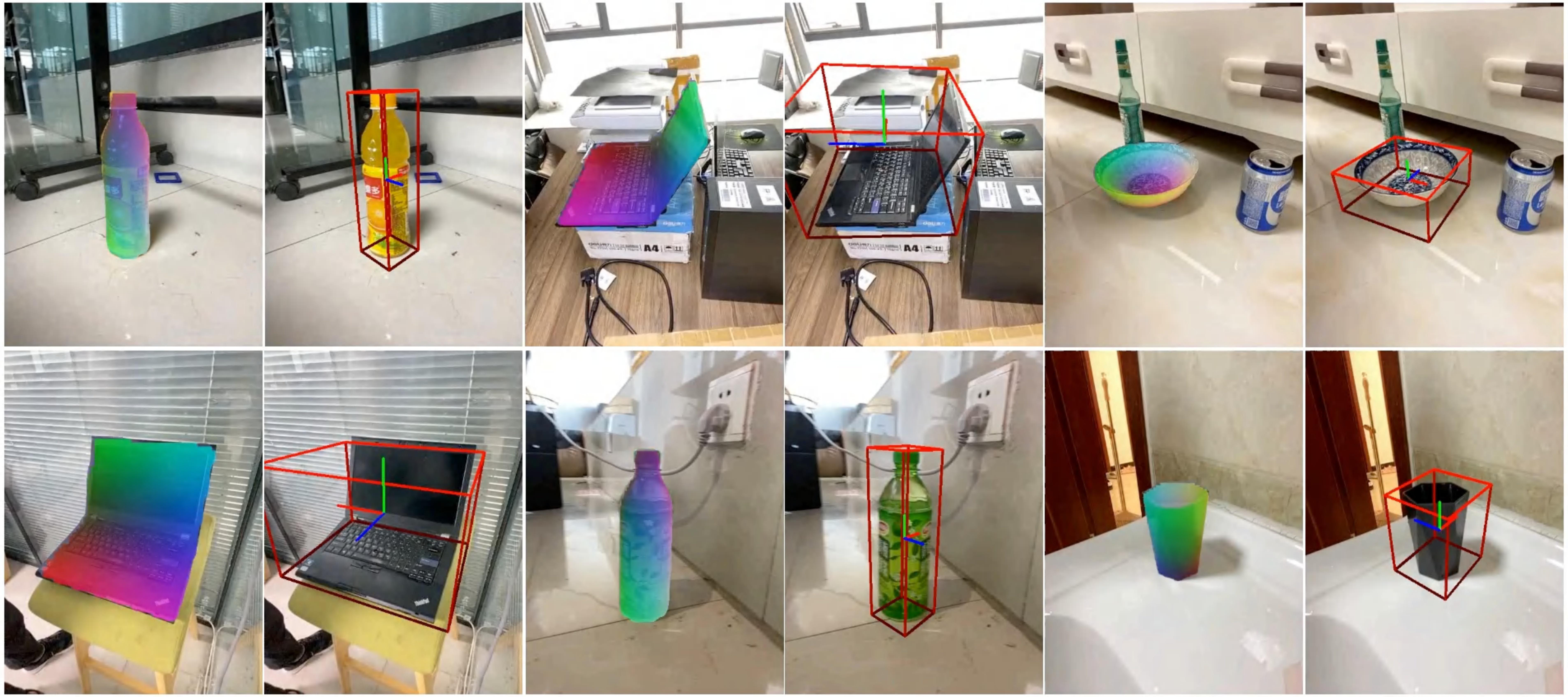


Success Rate: 53%

Video Understanding -> Imitation Learning

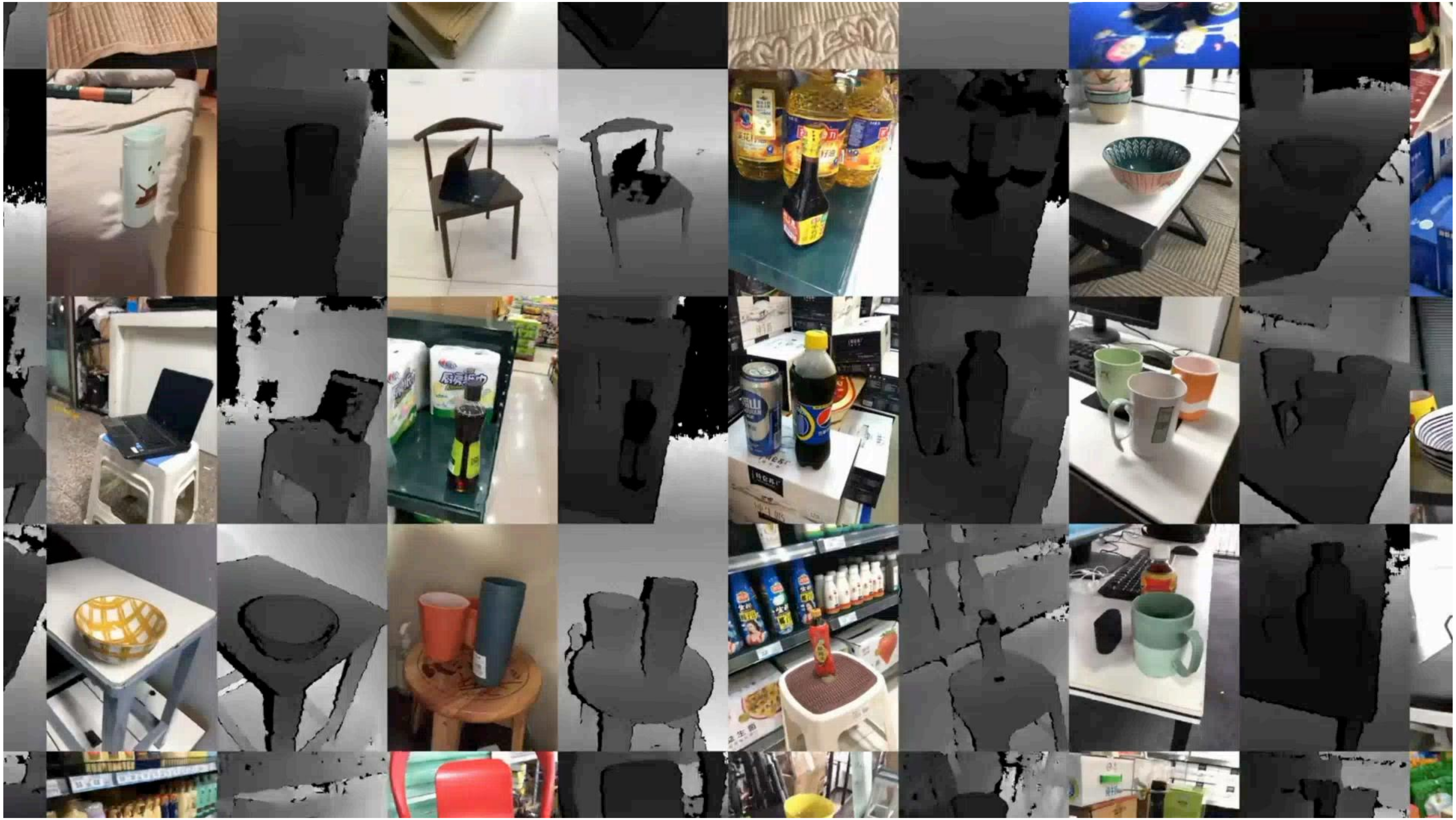
→ 3D Structure?

Self-Supervised Geometric Correspondence for Category-Level 6D Object Pose Estimation in the Wild



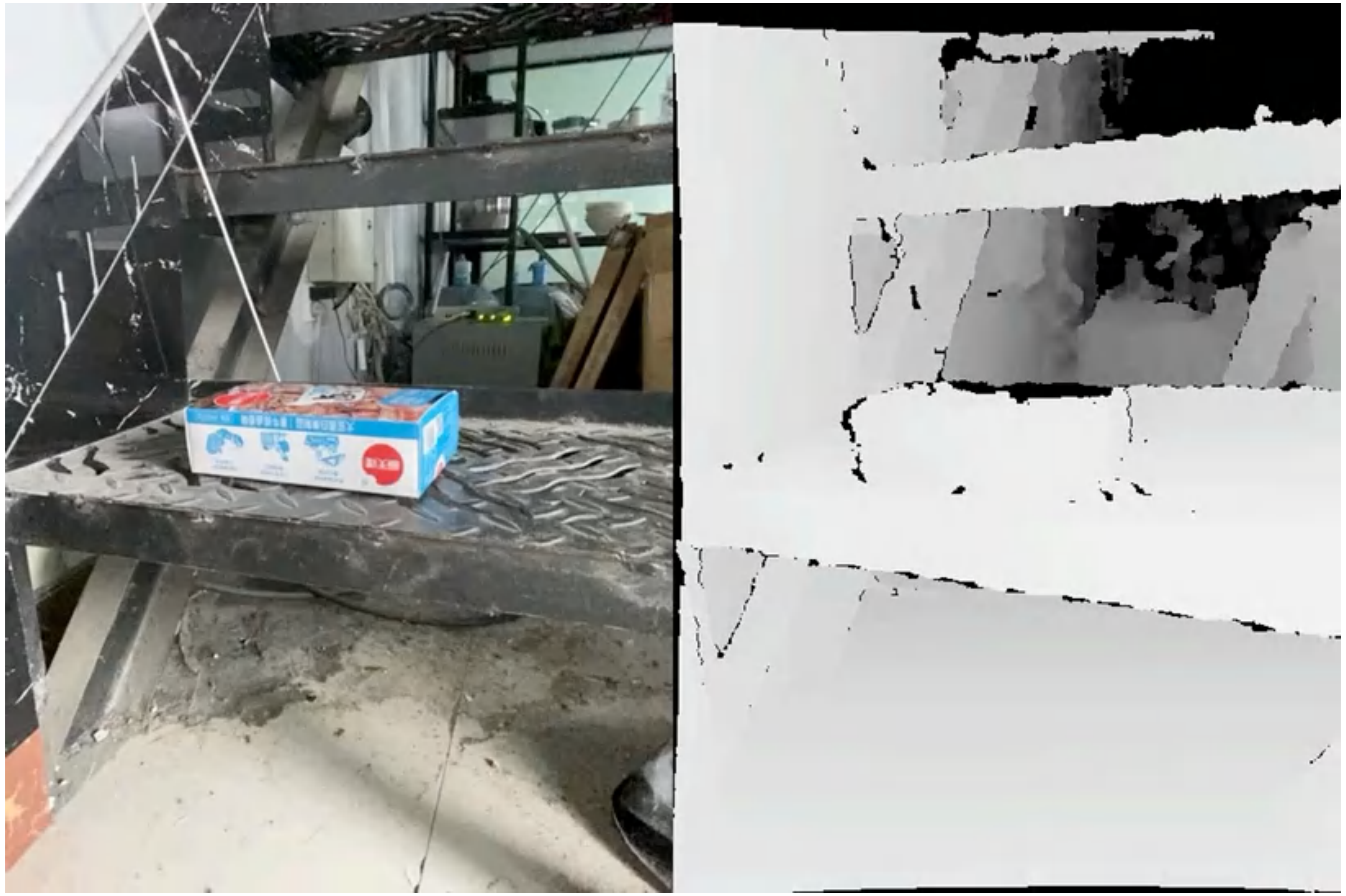
Kaifeng Zhang[1], Yang Fu[2], Shubhankar Borse[3], Hong Cai[3], Fatih Porikli[3], Xiaolong Wang[2]

[1] Tsinghua University, [2] UC San Diego, [3] Qualcomm AI Research



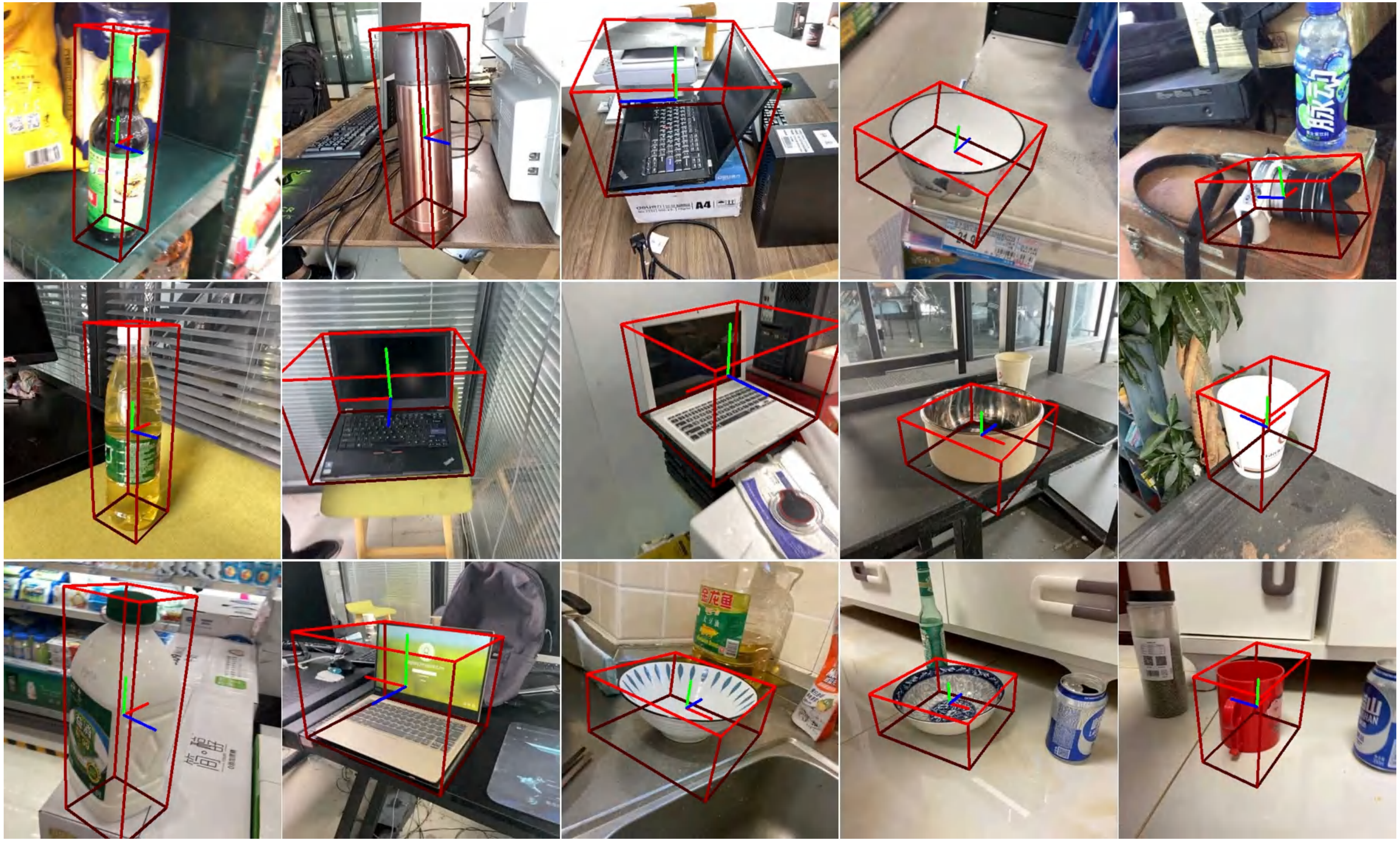
Wild6D dataset.
Yang Fu and Xiaolong Wang. NeurIPS 2022.

Wild6D Examples



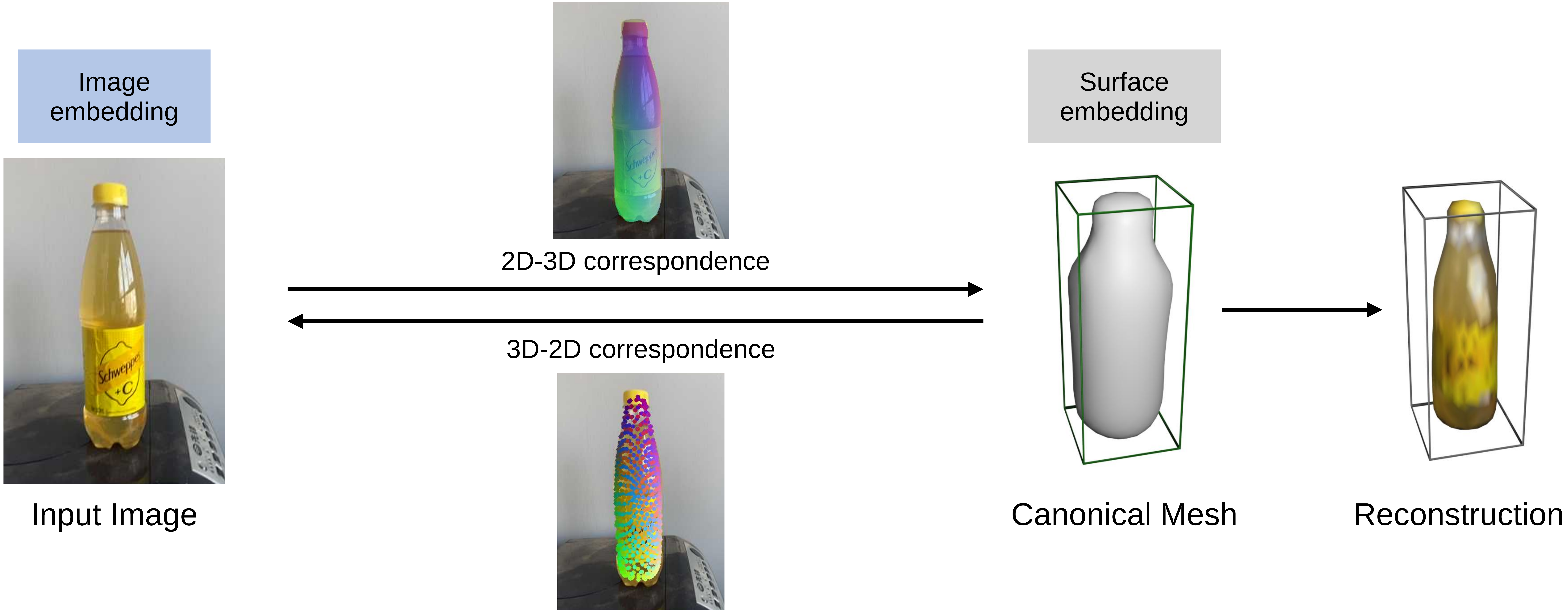
- Recording with iPhone or iPad.
- More than 5,000 RGBD videos across 1,700 objects (>1.1 million images).
- We provide annotations for 486 videos over 162 instances as a test set

Our goal: learning 2D-3D **dense correspondences** for self-supervised category-level 6D pose estimation on large-scale in-the-wild images.



Method

Overview



We build dense correspondences between pixels and mesh vertices via feature similarity in a shared embedding space.

Method

Overview

Different object instances correspond to the same canonical space.

Method

Overview

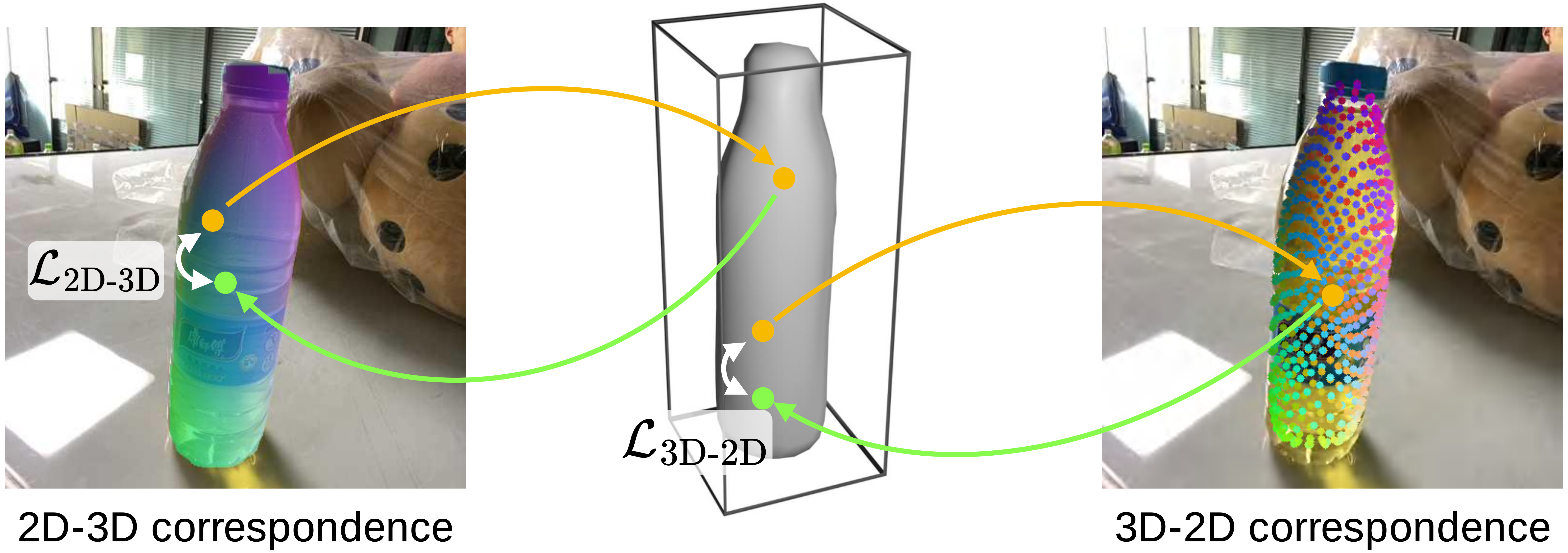
Pose fitting

We apply pose fitting to get the estimated pose from correspondence.

Method

Cycle consistency loss

(a) Instance cycle consistency



We propose novel cycle consistency losses for training correspondence.

The instance cycle consistency penalizes over correspondence-projection disparity within an image-mesh pair.

Method

Cycle consistency loss

(b) Cross-instance and cross-time cycle consistency

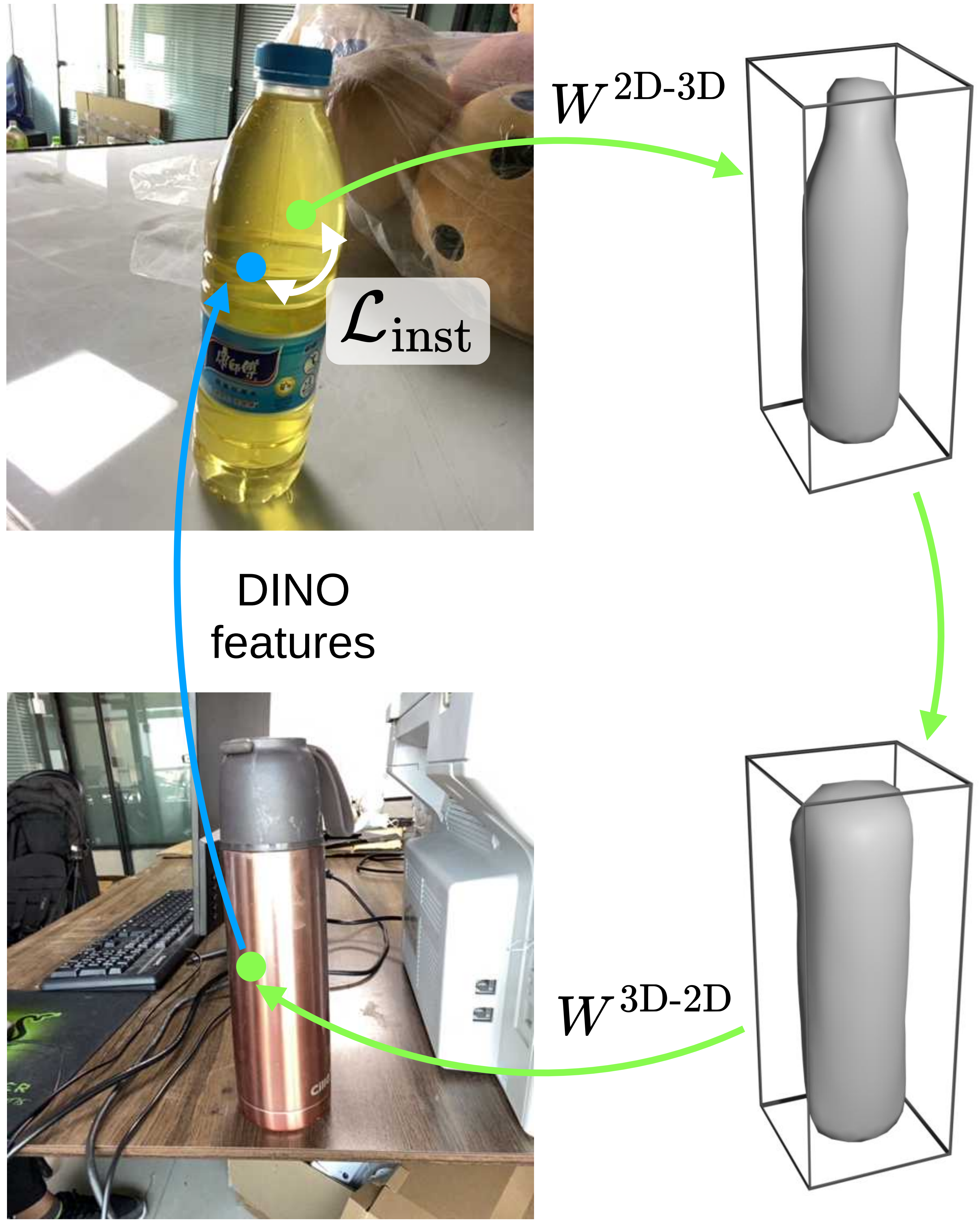
Image i Mesh i Cross-instance image j Mesh j Cross-time image j' Mesh j'

We also go beyond a single image to cross-instance and cross-time images.

Method

Cycle consistency loss

(b) Cross-instance and cross-time cycle consistency

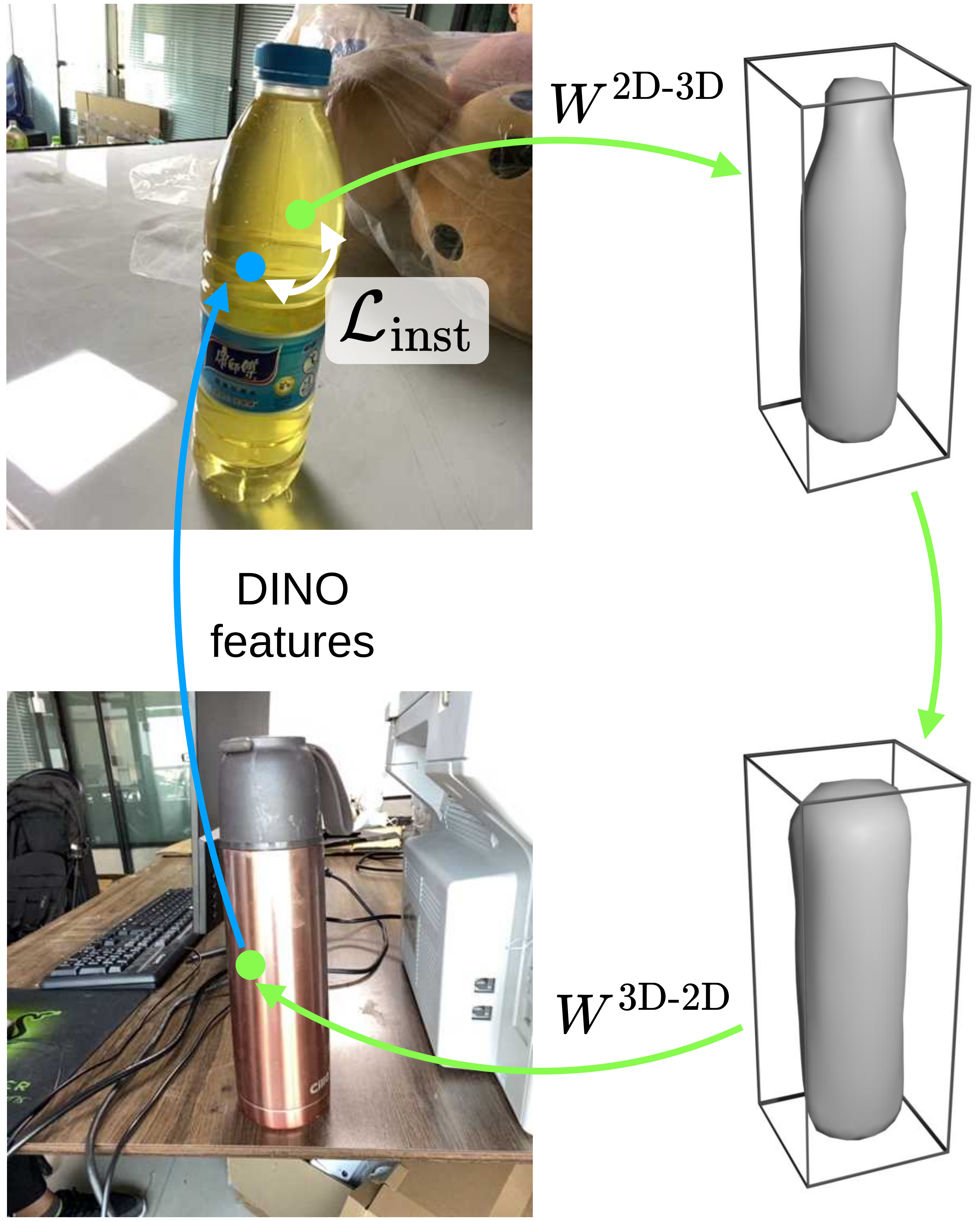


By building a 4-step cycle, we encourage different images to consistently correspond to the shared canonical space.

Method

Cycle consistency loss

(b) Cross-instance and cross-time cycle consistency



By building a 4-step cycle, we encourage different images to consistently correspond to the shared canonical space.

Result

Category-level 6D pose estimation on Wild6D

Result

Category-level 6D pose estimation on Wild6D

Result

Category-level 6D pose estimation on Wild6D

Result

Category-level 6D pose estimation on Wild6D

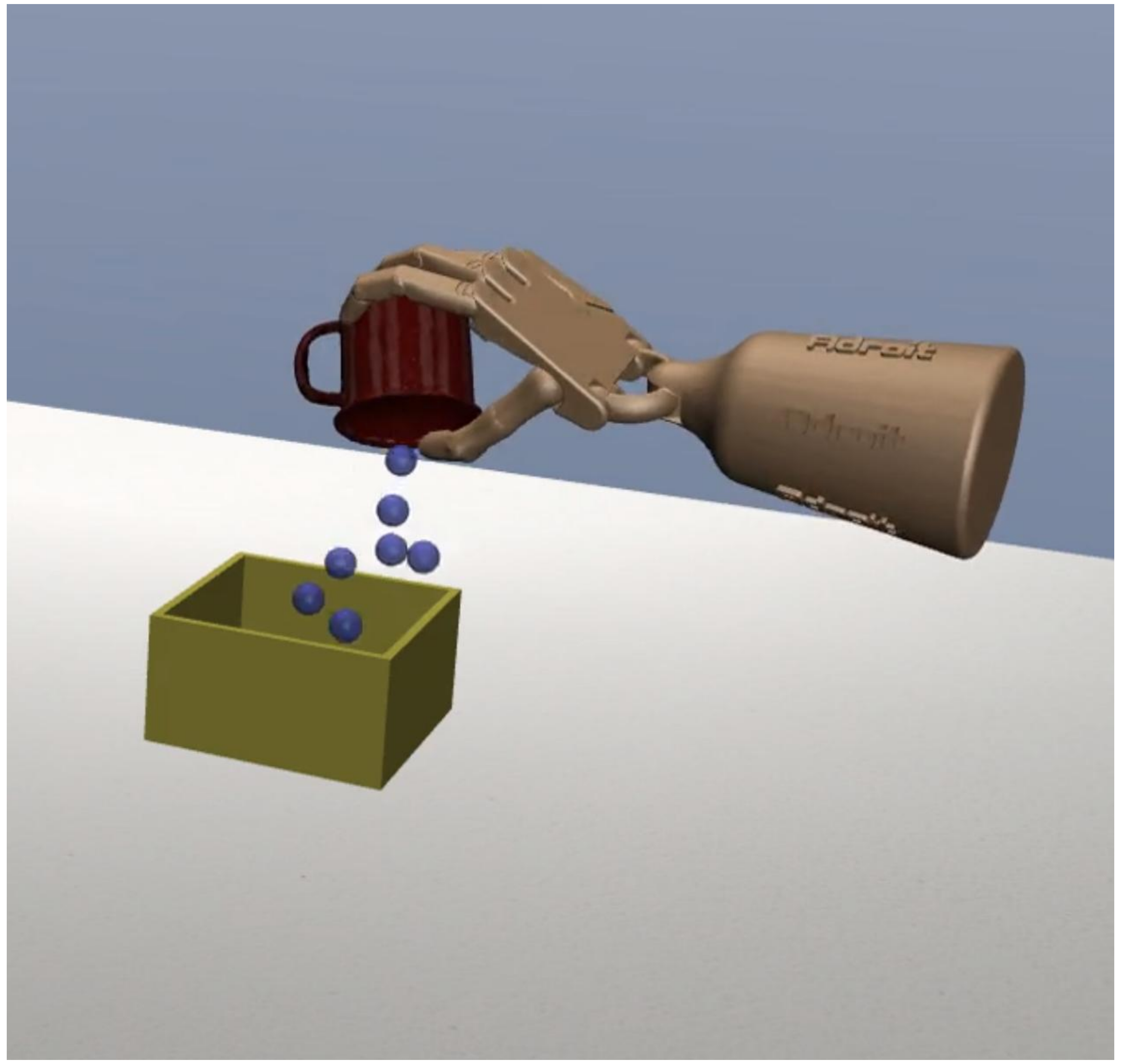
Result

Category-level 6D pose estimation on Wild6D

Result

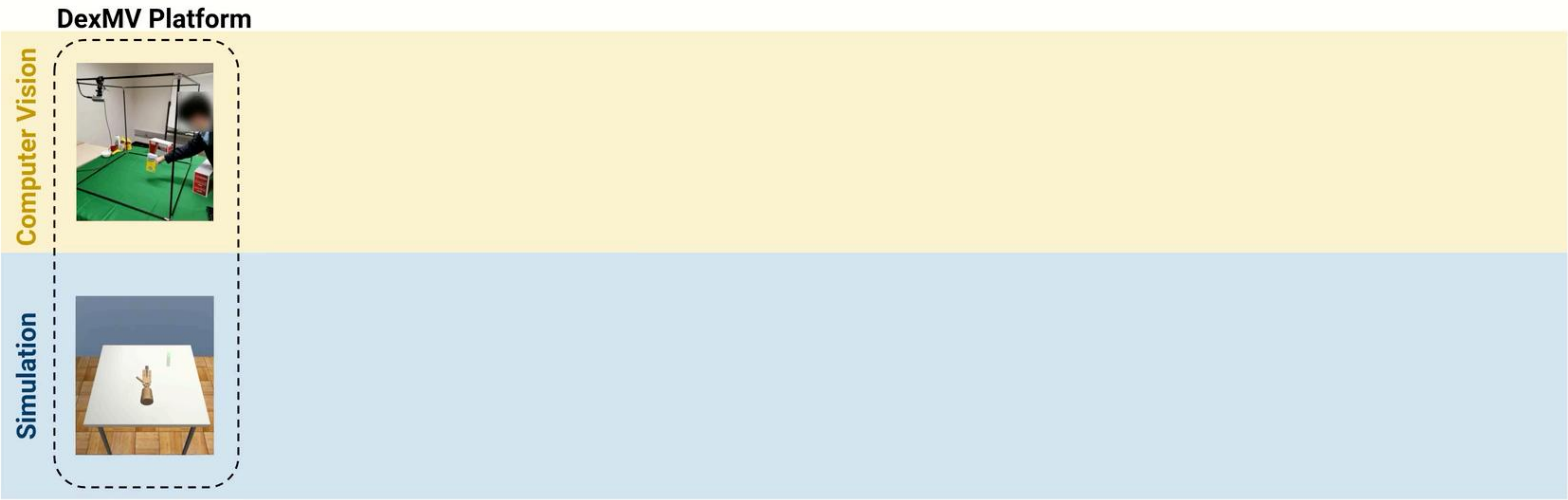
Category-level 6D pose estimation on Wild6D

DexMV Platform for Imitation Learning



DexMV: Imitation Learning for Dexterous Manipulation from Human Videos.
Yuzhe Qin*, Yueh-Hua Wu*, Shaowei Liu*, Hanwen Jiang*, Ruihan Yang, Yang Fu, Xiaolong Wang
ECCV 2022

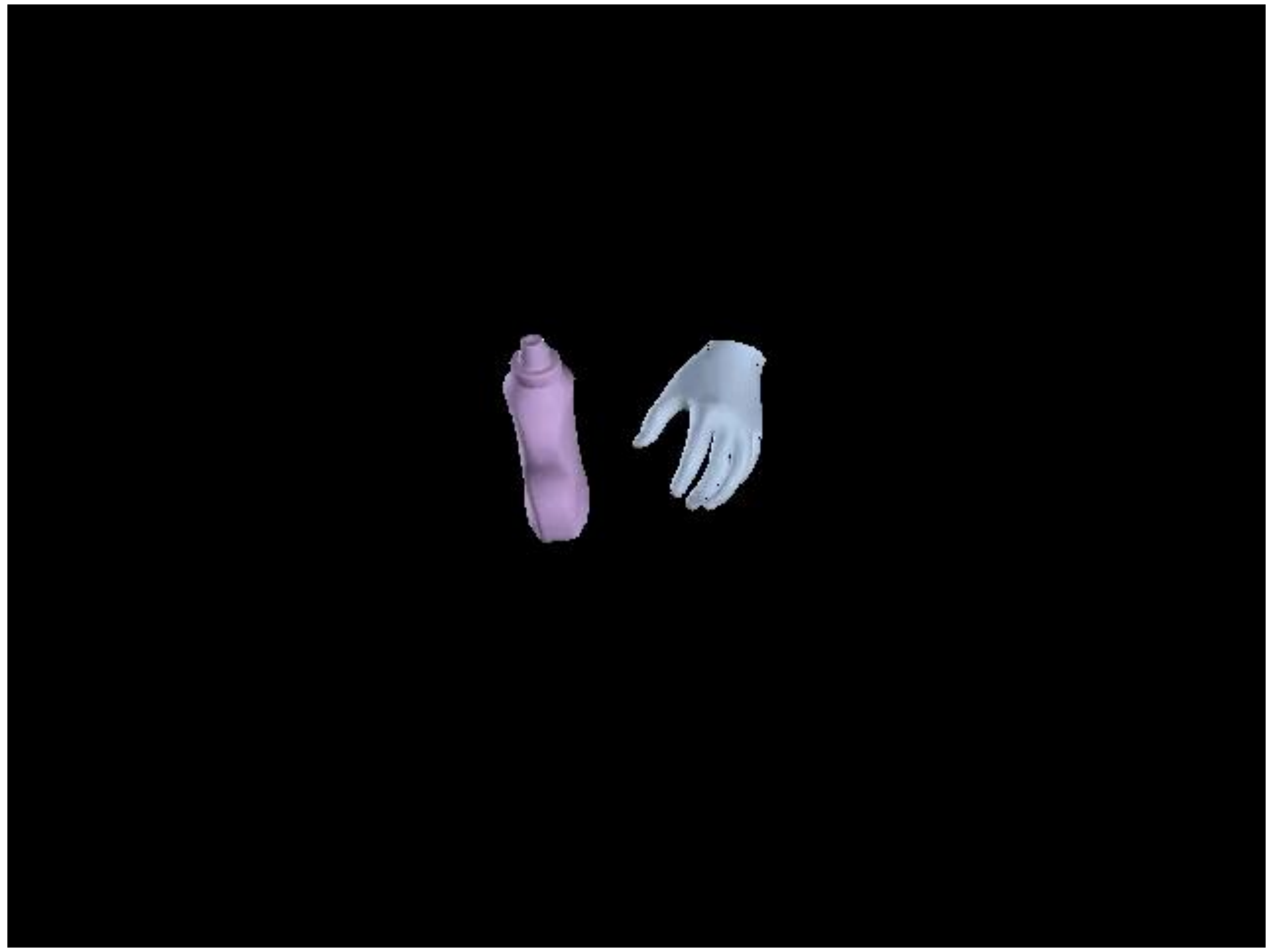
DexMV Platform for Imitation Learning


DexMV Platform
Computer Vision
Simulation

The Computer Vision System

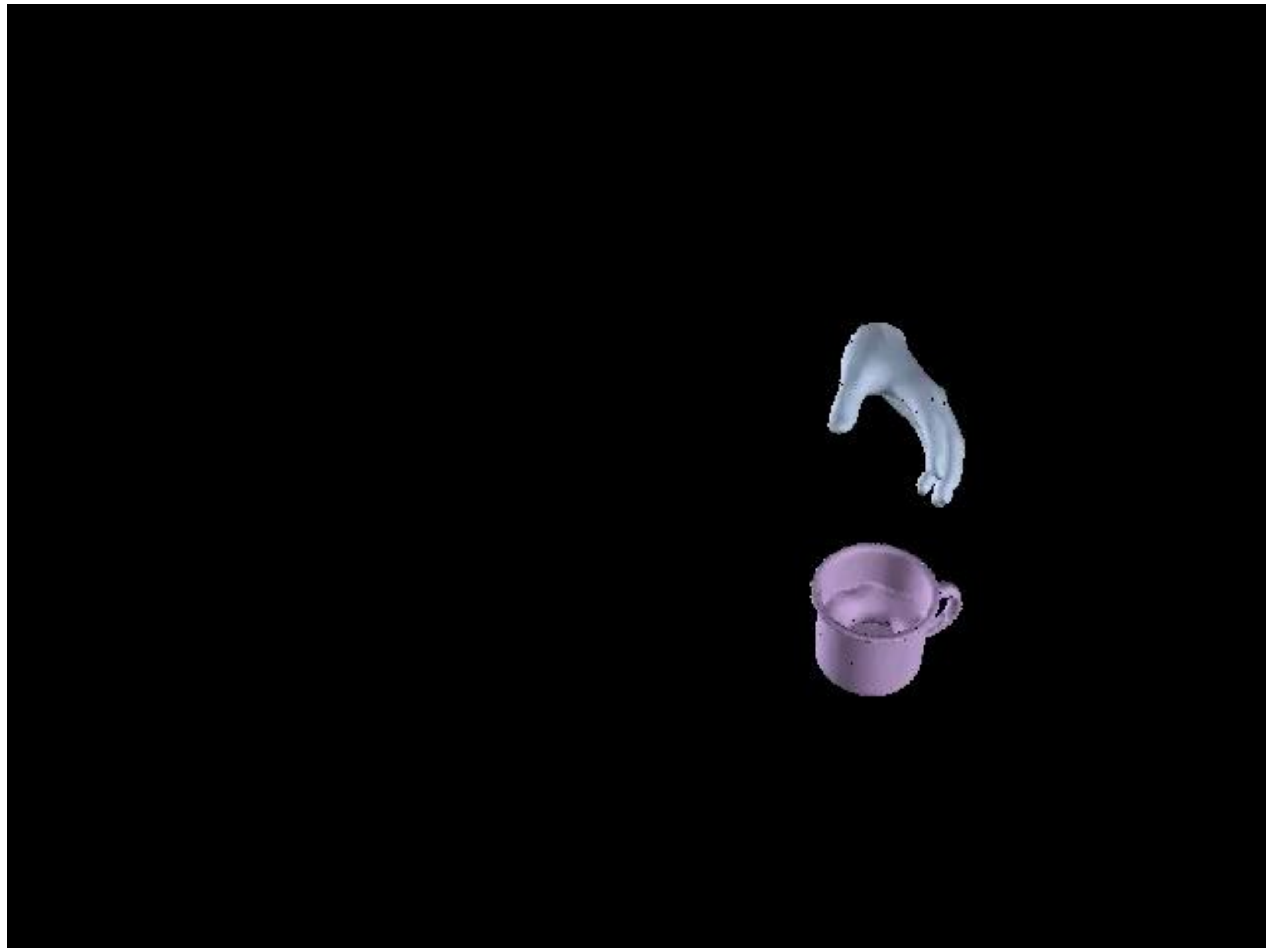
- In computer vision system, we collect human demonstrations, perform 3D Pose Estimation, and motion retargeting to generate demonstrations.

Examples for Mustard Bottle



We can collect 100 demonstrations in 1 hour

Examples for Pour



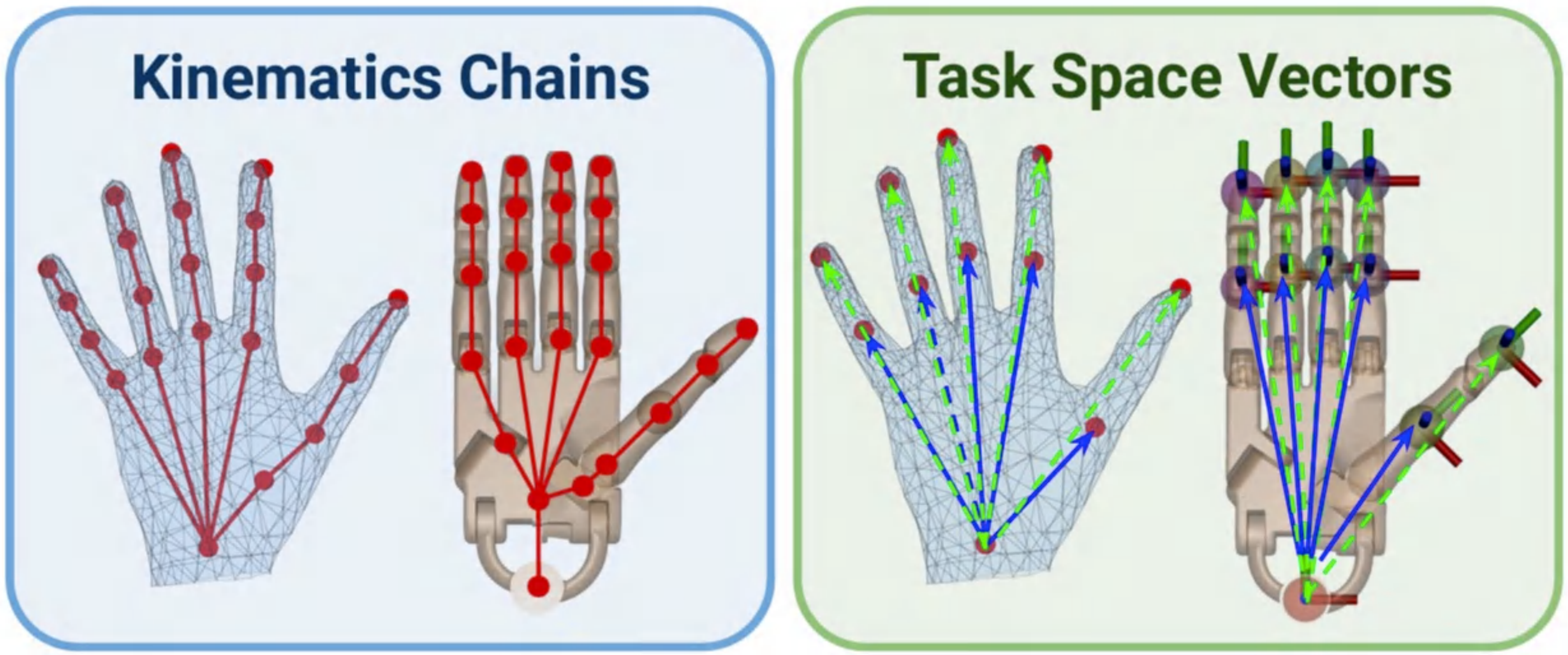
We can collect 100 demonstrations in 1 hour

Hand Motion Retargeting

- We collect demonstration on human hand manipulating objects, but we need to perform imitation learning on a robot hand.

- Human and robot hand are different in both geometry and kinematics.

- We match the task space vectors (green dot arrows).



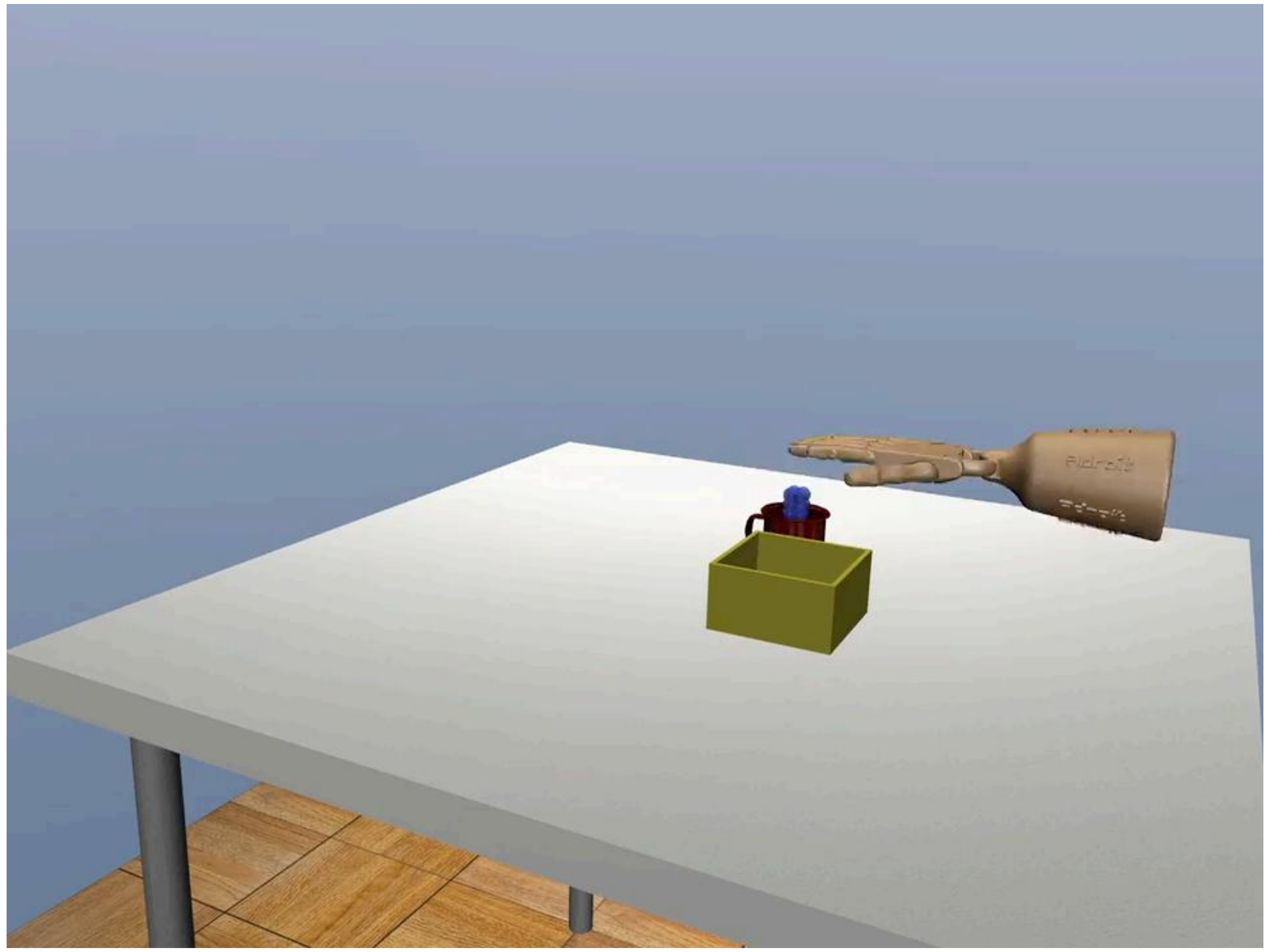
Examples for Hand Motion Retargeting

Examples for Hand Motion Retargeting

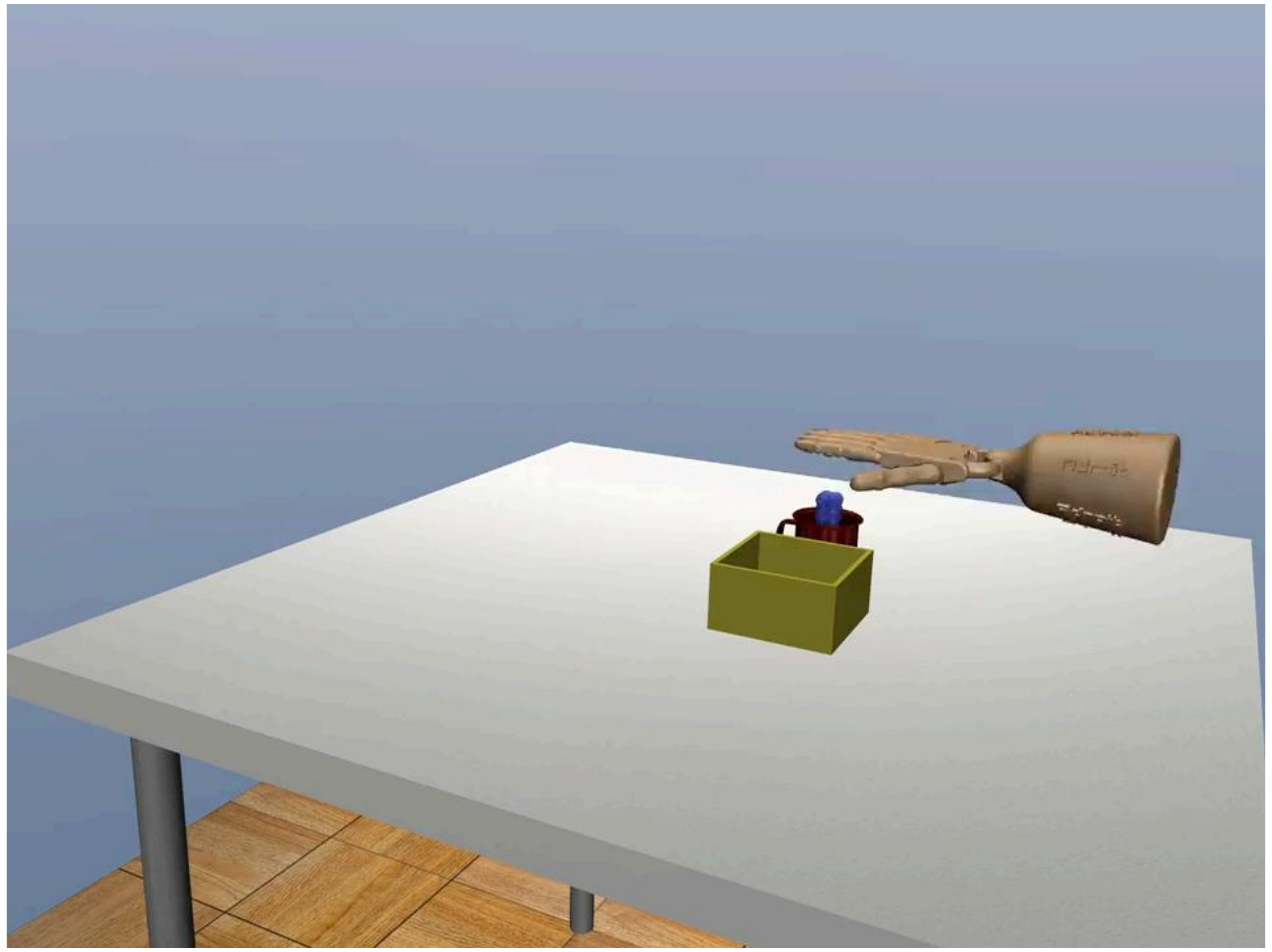
The Simulation System

- In the simulation system, we perform imitation learning by augmenting the RL objective with the demonstrations from the computer vision system

Example for Pour with Trained Policy

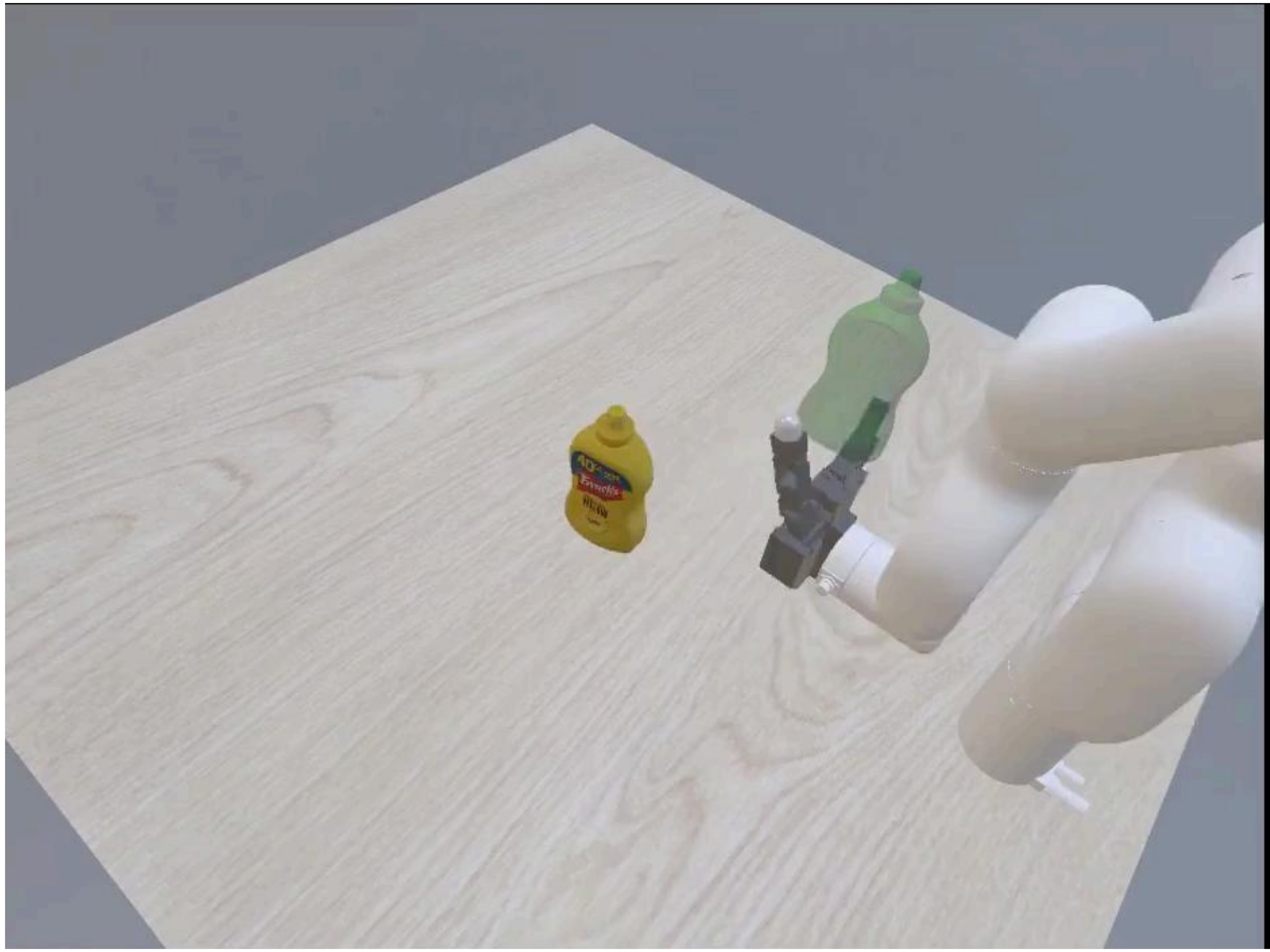


Pure Reinforcement Learning

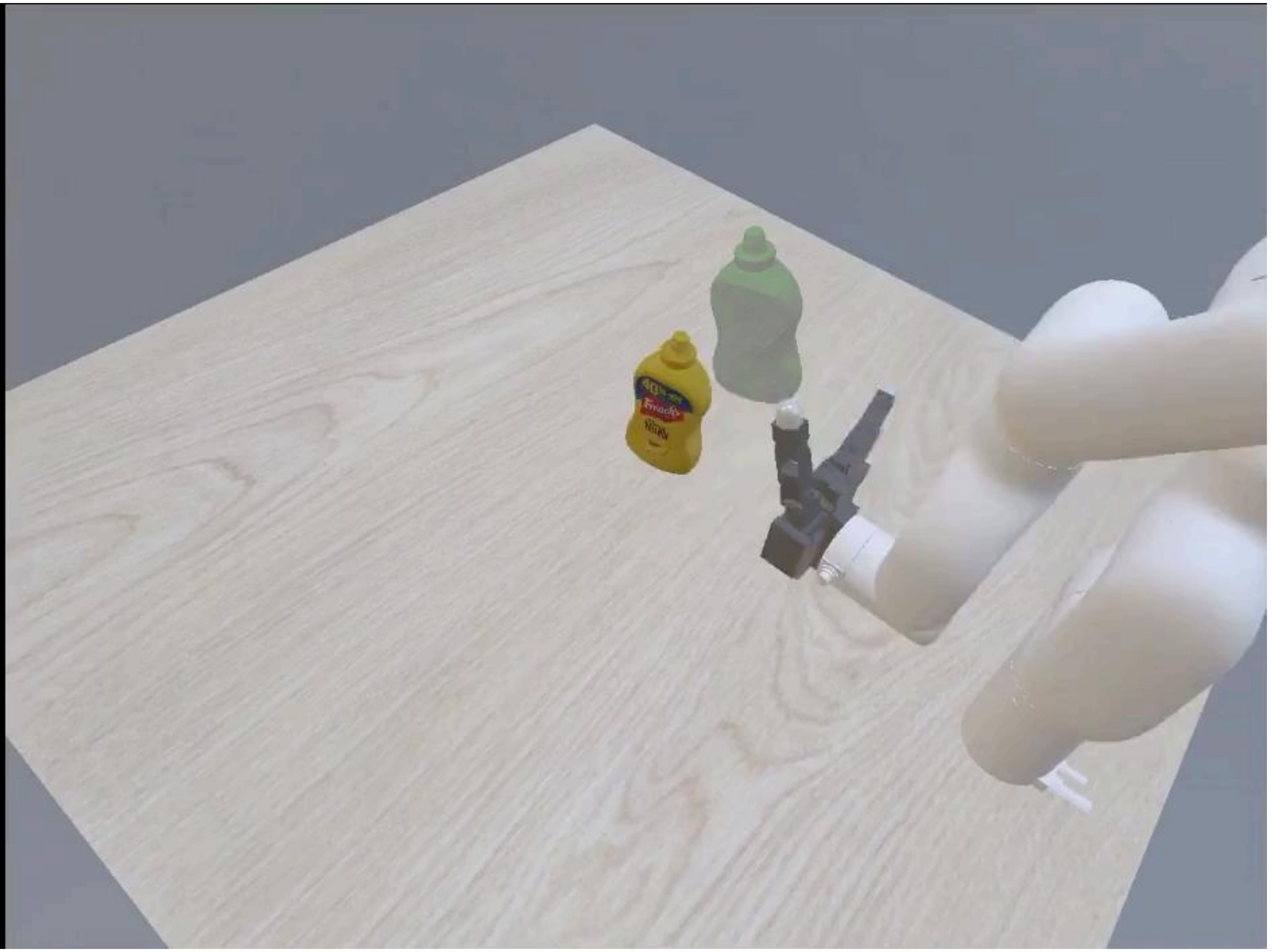


Imitation with Demonstration

Sim2Real with Xarm + Allegro Hand



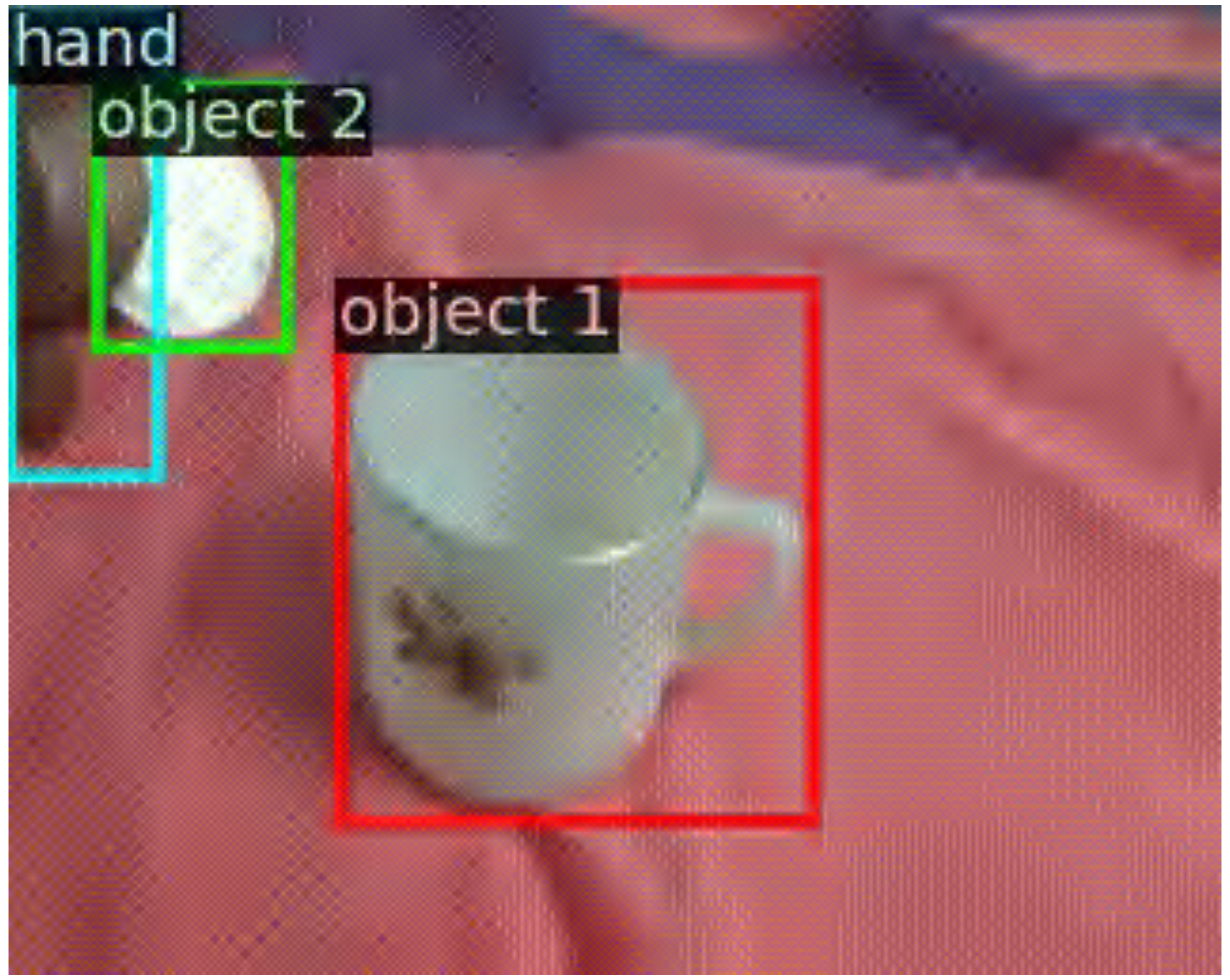
Reinforcement Learning
without Demonstrations

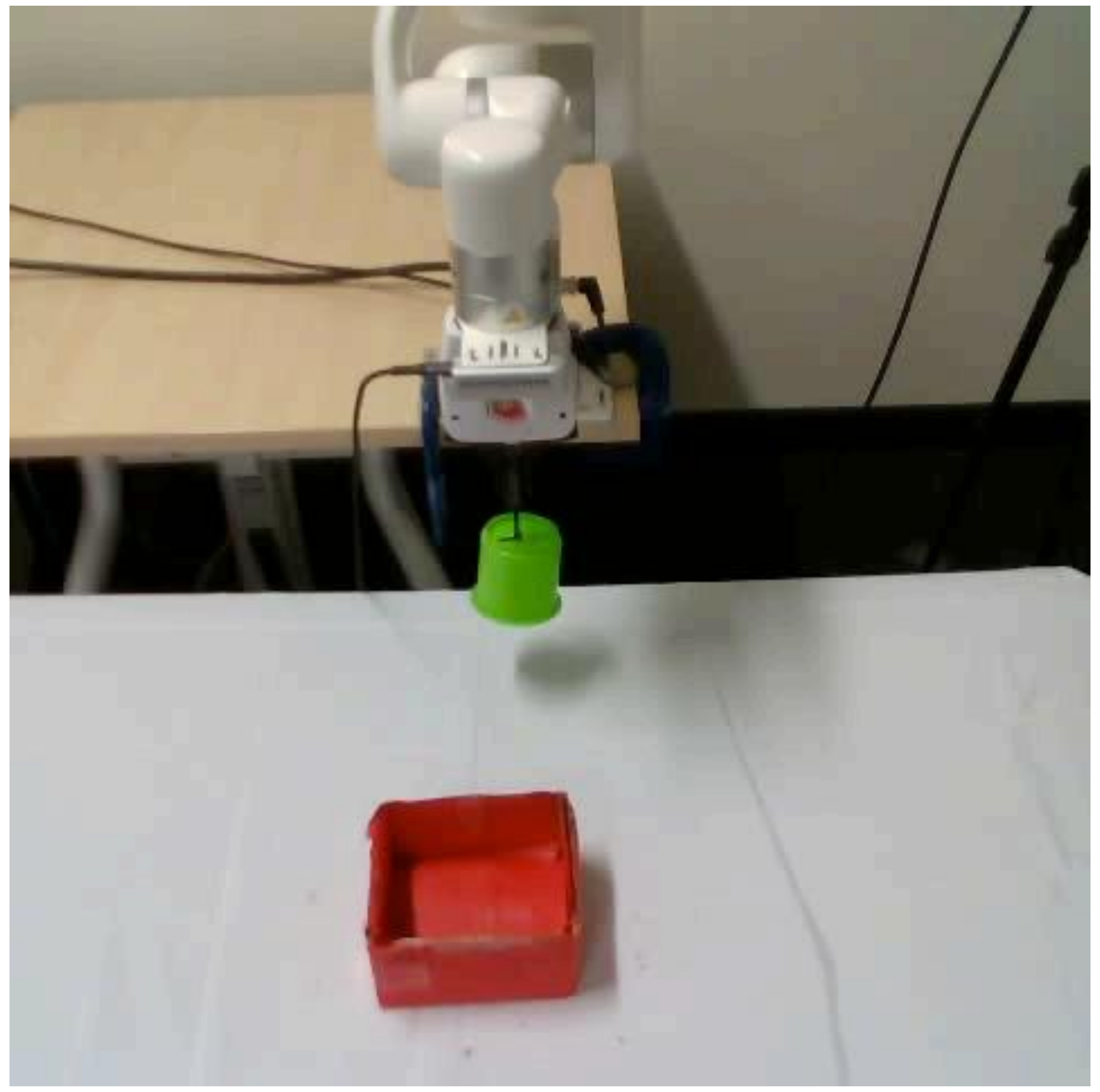


Imitation Learning with
Demonstrations

Yuzhe Qin, Hao Su, Xiaolong Wang. IROS 2022

Video Understanding -> Imitation Learning





- Accurate
- Efficient
- Robust
- Safe