Robot Perception and Learning

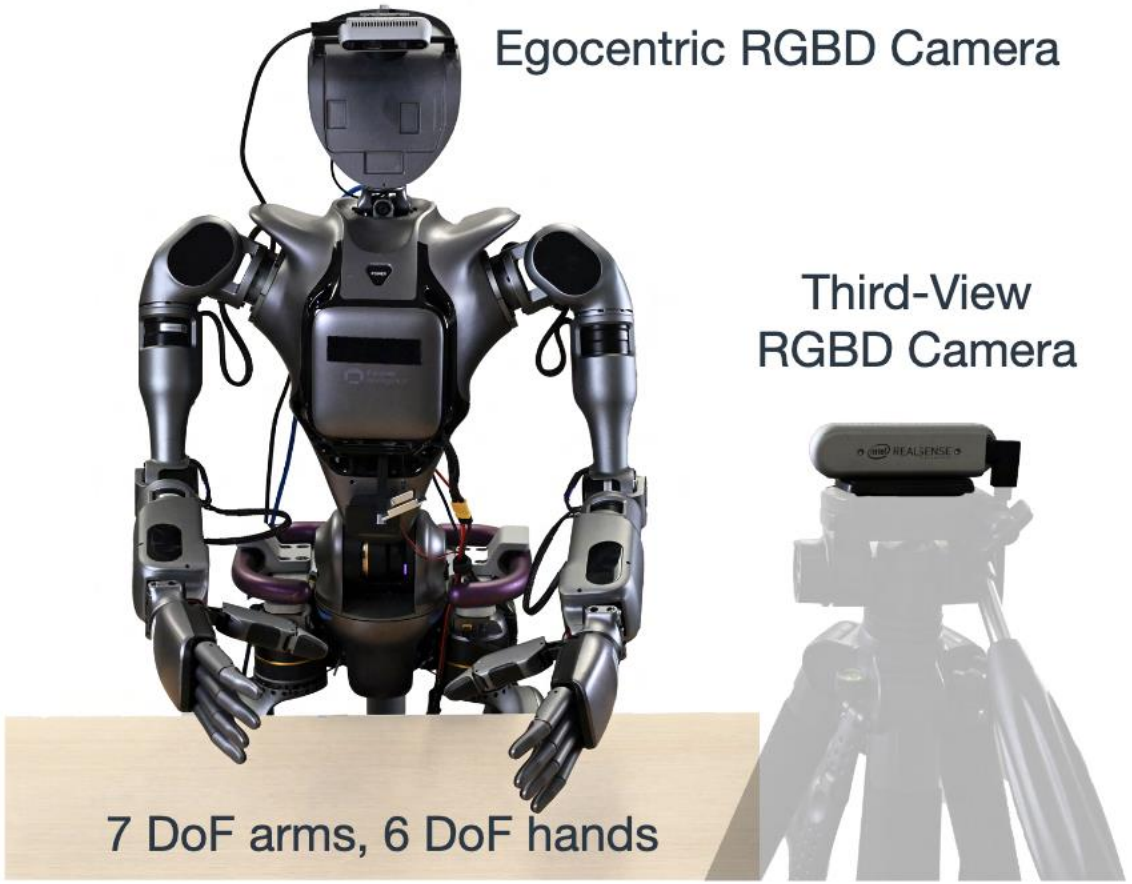
Humanoid Robot: Locomotion and Mobile manipulation

Tsung-Wei Ke

Fall 2025

Learning Robot Control Policies Has Achieved Great Success

The Single-arm Jaw-based Embodiment is Limited We Need Multi-Arm Multi-Fingered Robots…

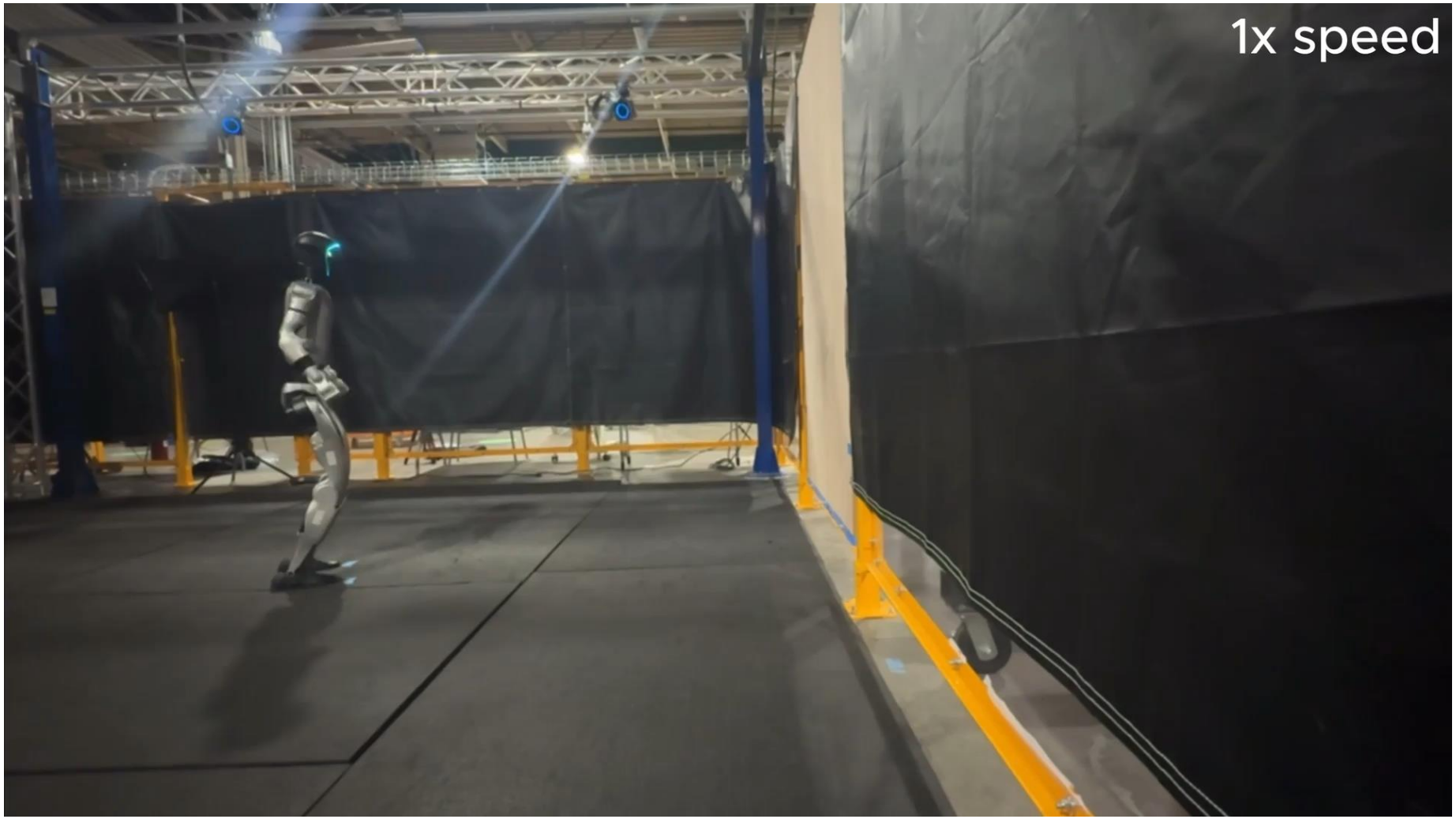


Sim-to-Real Reinforcement Learning for Vision-Based Dexterous Manipulation on Humanoids. Lin et al.

Fixed Robot is Limited. We Need Mobile Robots…

OmniRetarget: Interaction-Preserving Data Generation for Humanoid Whole-Body Loco-Manipulation and Scene Interaction. Yang et al.

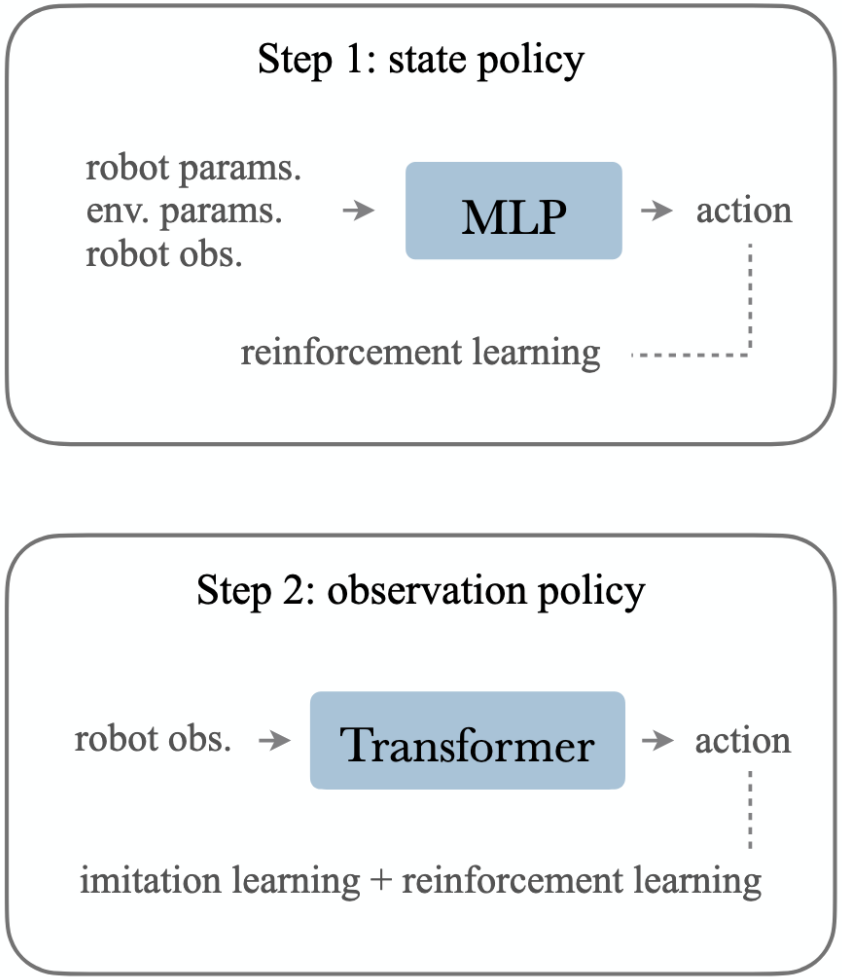
Fixed Robot is Limited. We Need Mobile Robots…



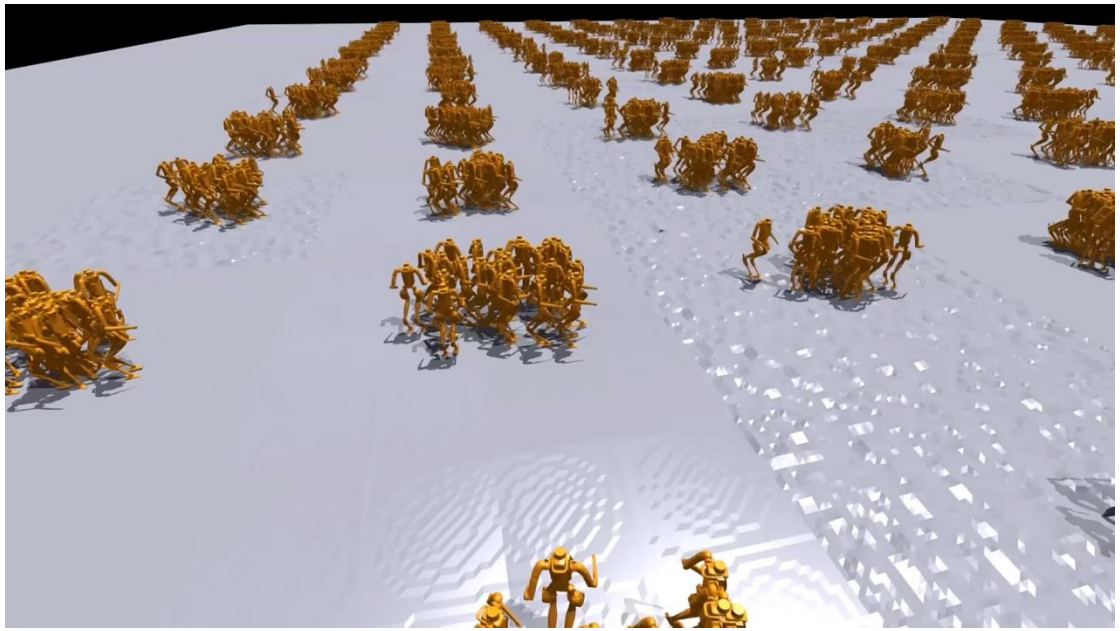
OmniRetarget: Interaction-Preserving Data Generation for Humanoid Whole-Body Loco-Manipulation and Scene Interaction. Yang et al.

How to Learn Policies for Locomotion?

Learning Locomotion with Reinforcement Learning



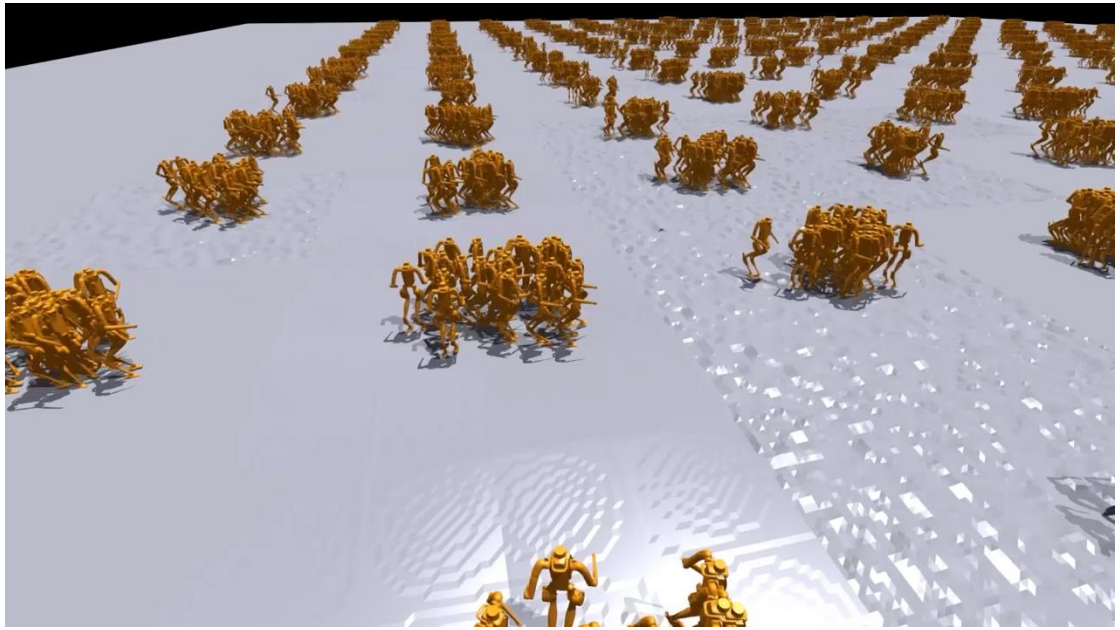
Train in simulation



Learning Locomotion with Reinforcement Learning

Trest in the real world

Train in simulation



Real-World Humanoid Locomotion with Reinforcement Learning. Ilija et al.

However, We Need More Diverse, Coordinated Lomotion

Coordinated locomotion

Video source:
https://www.reddit.com/r/gifs/comments/8gr87a/hold_my_beer/

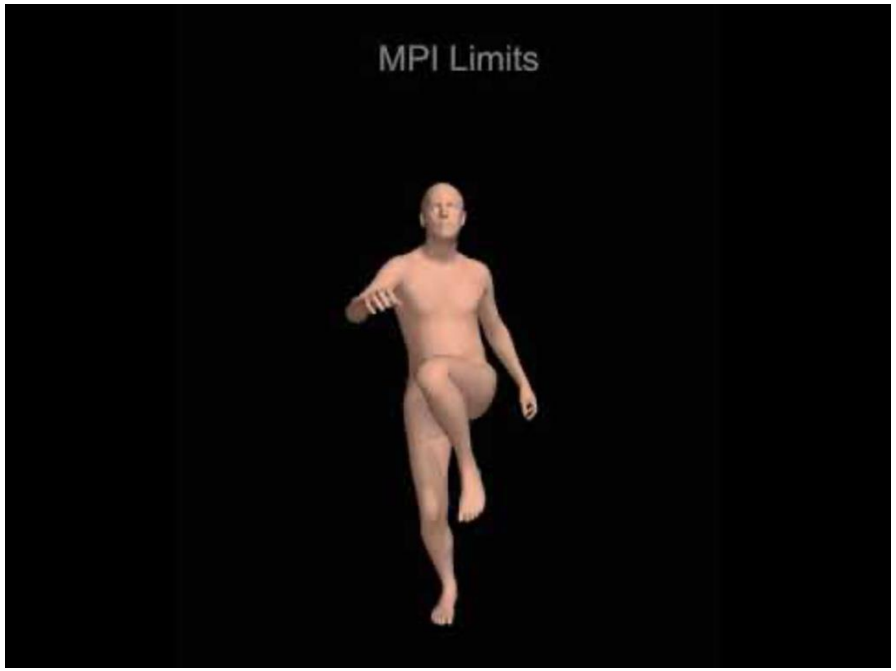
Diverse locomotion skills

Video by Tag Chases

How to design the reward…

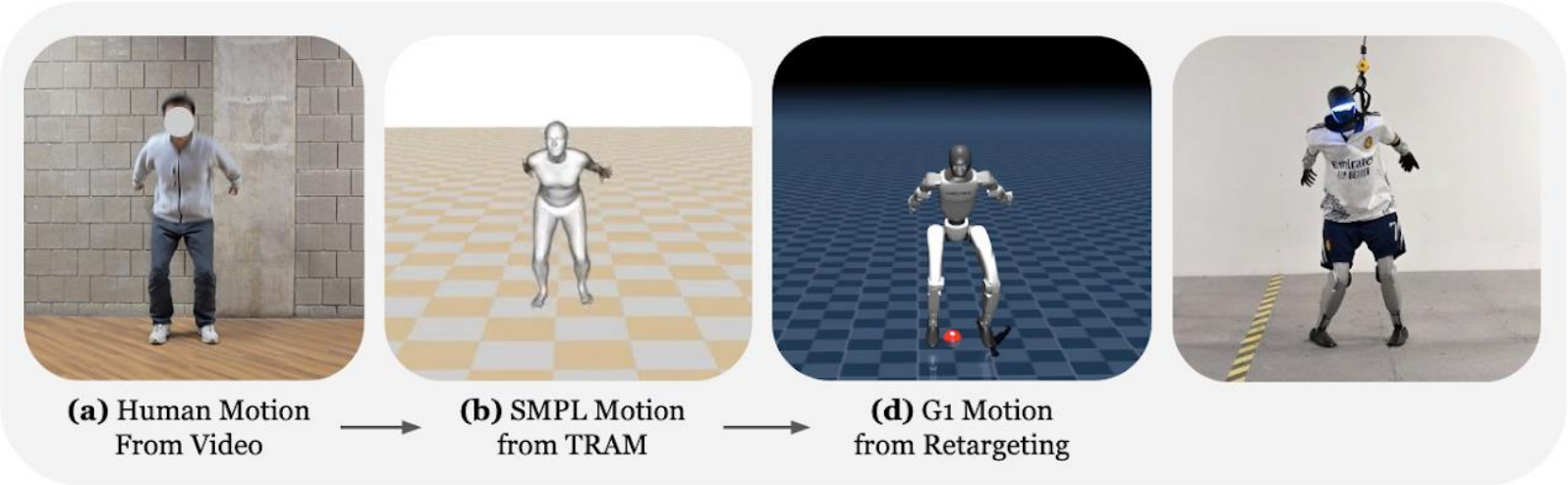
Idea: Learning with Human Motion Priors

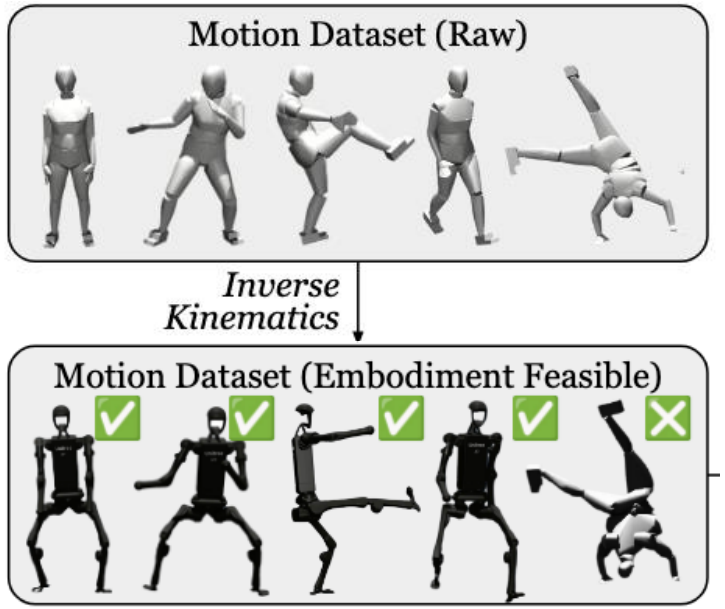
Object Motion Guided Human Motion Synthesis. Li et al.



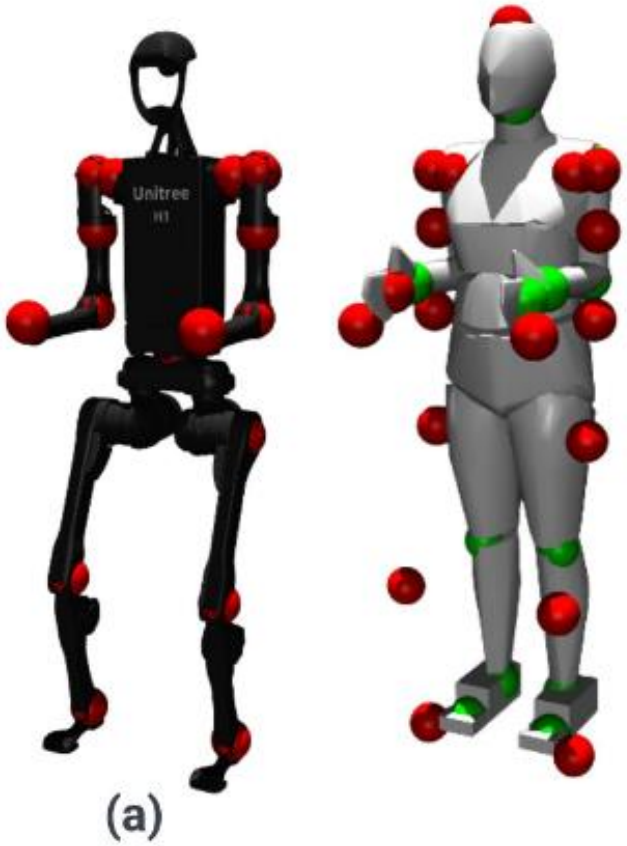
AMASS: Archive of Motion Capture As Surface Shapes. Mahmood et al.

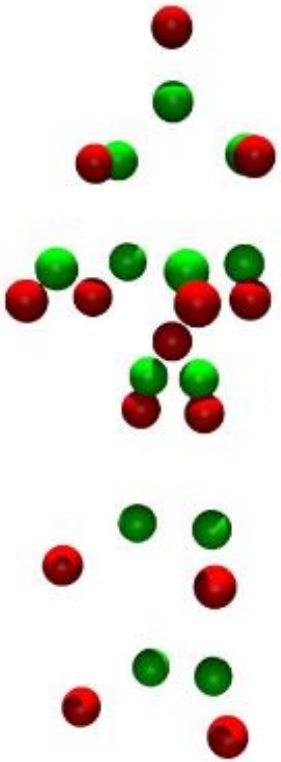
We Need to Retarget Human Motion to Robot Motion





Direct Human-to-Robot Motion Retargetting Fails due to Embodiment Gap





Robots Have Different:

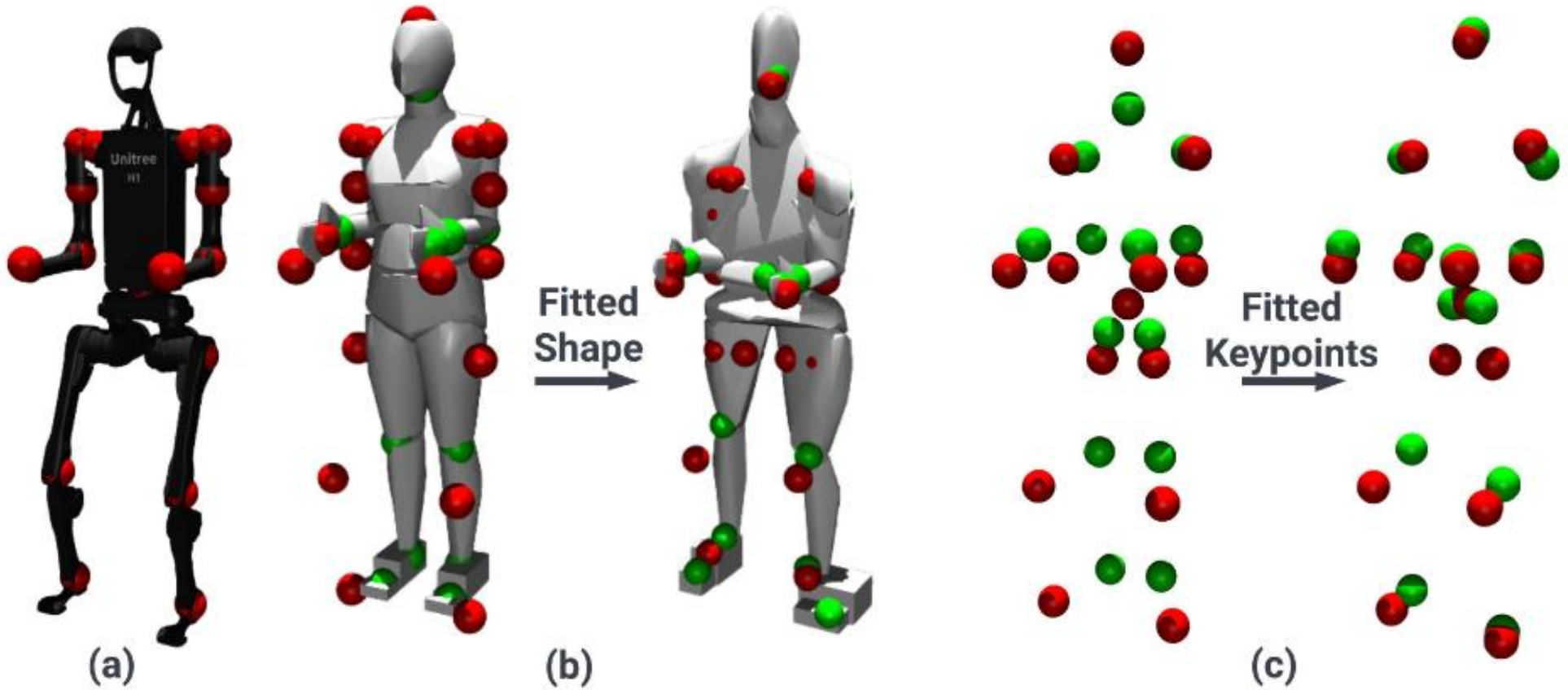
- Shape
- Size
- Geometry

Direct Human-to-Robot Motion Retargetting Fails due to Embodiment Gap

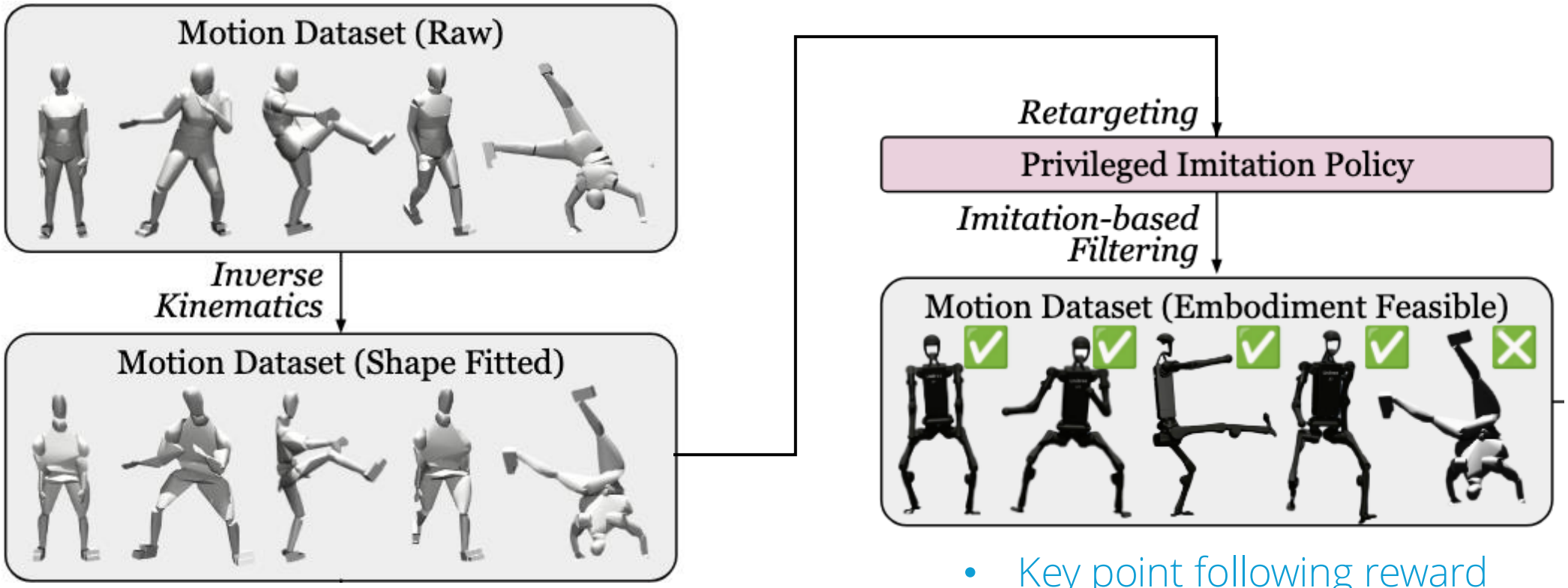
Video source: ENGINEAI/LinkedIn

Video source: Unitree Robotics

Step 1: Reshape Human 3D Model for Remapping



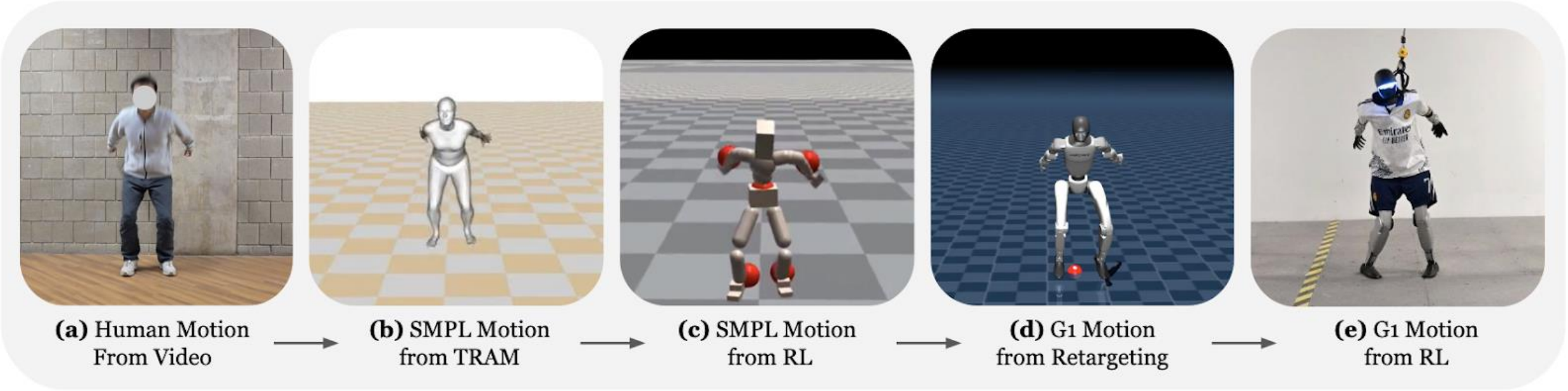
Step 2: Obtain Feasible Actions with RL

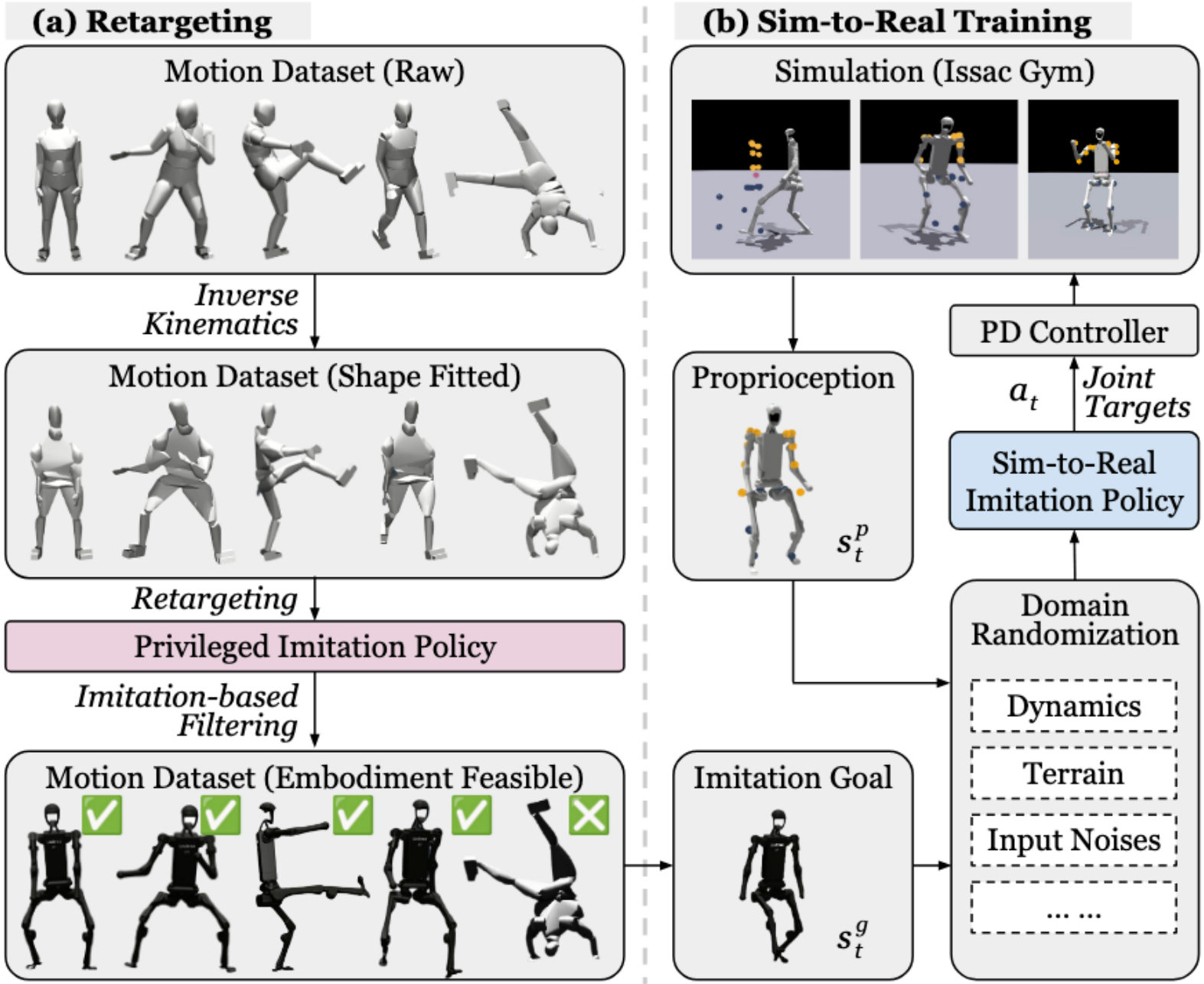


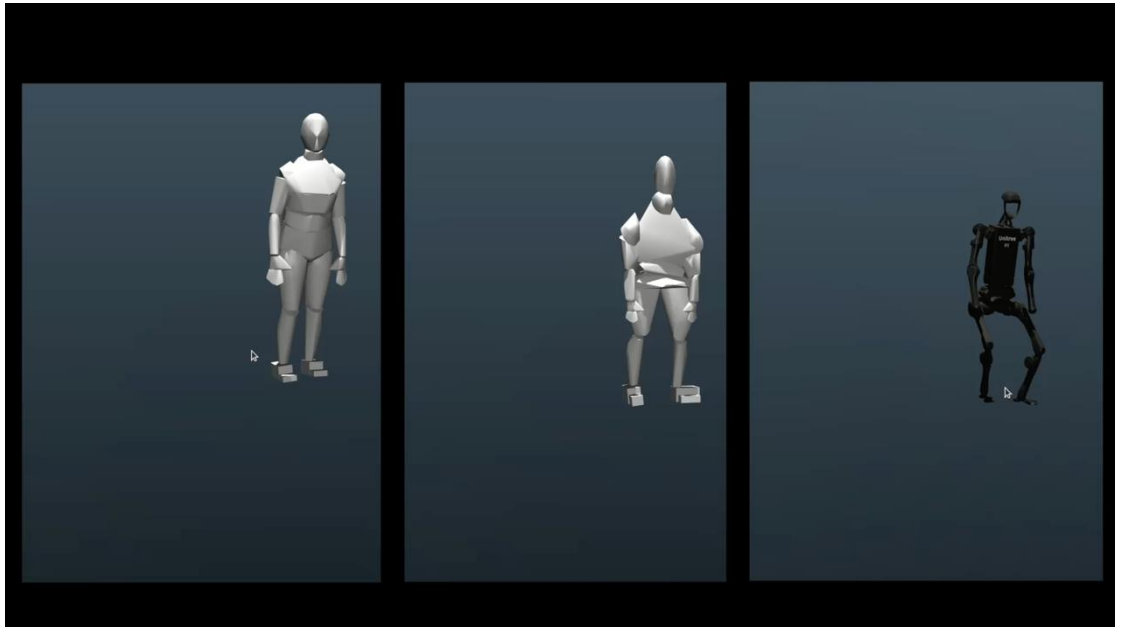
- Key point following reward
- Fall prevention reward
- Energy preservation reward
- …

Learning Human-to-Humanoid Real-Time Whole-Body Teleoperation. He et al

A General Human-to-Robot Motion Retargetting Pipeline




(a) Retargeting
Motion Dataset (Raw)
Inverse Kinematics
Motion Dataset (Shape Fitted)
Retargeting
Privileged Imitation Policy
Imitation-based Filtering
Motion Dataset (Embodiment Feasible)
(b) Sim-to-Real Training
Simulation (Issac Gym)
PD Controller
a_t
Joint Targets
Proprioception
s_t^p
Sim-to-Real Imitation Policy
Domain Randomization
Dynamics
Terrain
Input Noises
... ...
Imitation Goal
s_t^g



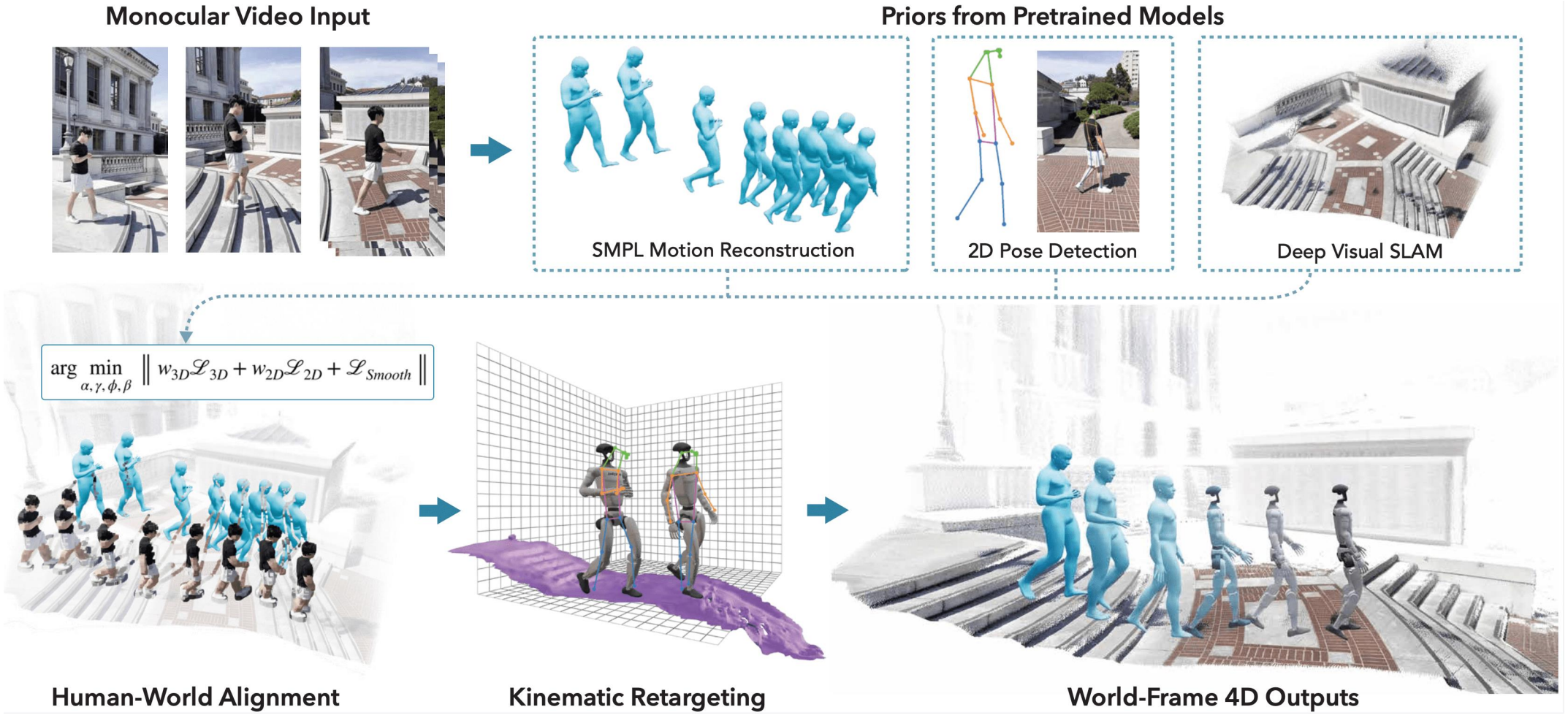
The Same Idea Applies to Video-to-Robot Locomotion

VideoMimic Visual imitation enables contextual humanoid control. Allshire et al.

RL in Digital Twins

VideoMimic Visual imitation enables contextual humanoid control. Allshire et al.

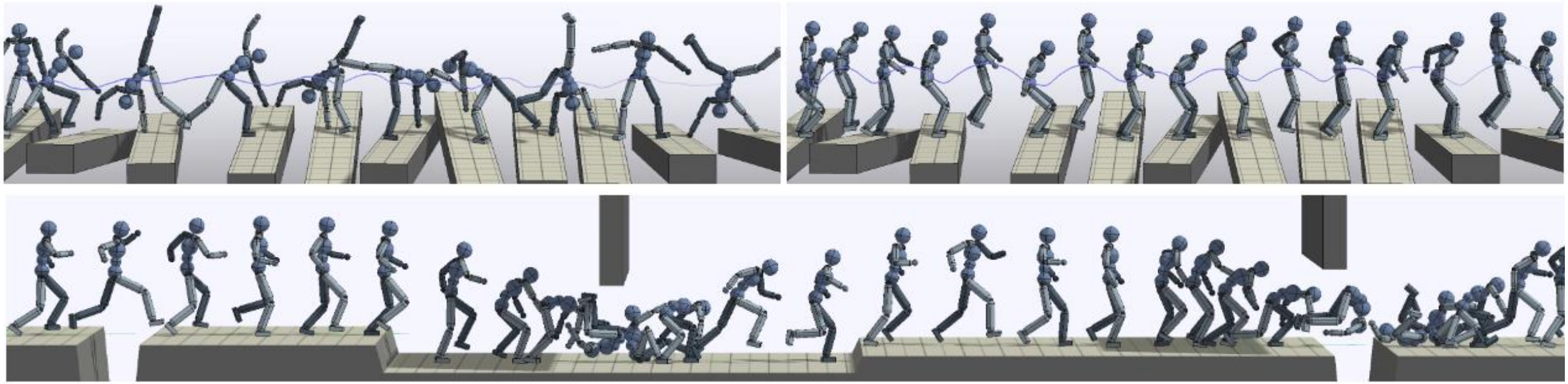
A Video-to-Robot Locomotion Pipeline



Have We Solved Humanoid Robot Locomotion?

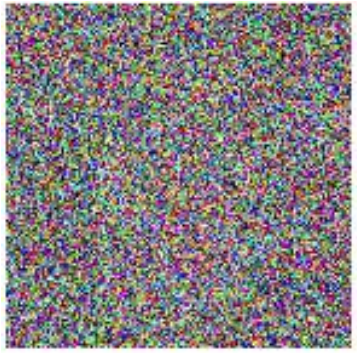
- Previous methods use an action space of joint angles, which has high degree of freedom
- How to adapt learned policies to new terrains / motions?
  - ➢ Sample-inefficient RL becomes challenging (again)

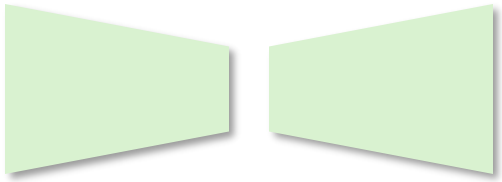
We need a better action space!



AMP: Adversarial Motion Priors for Stylized Physics-Based Character Control. Peng et al.

We Have the Same Problem in Visual Generation



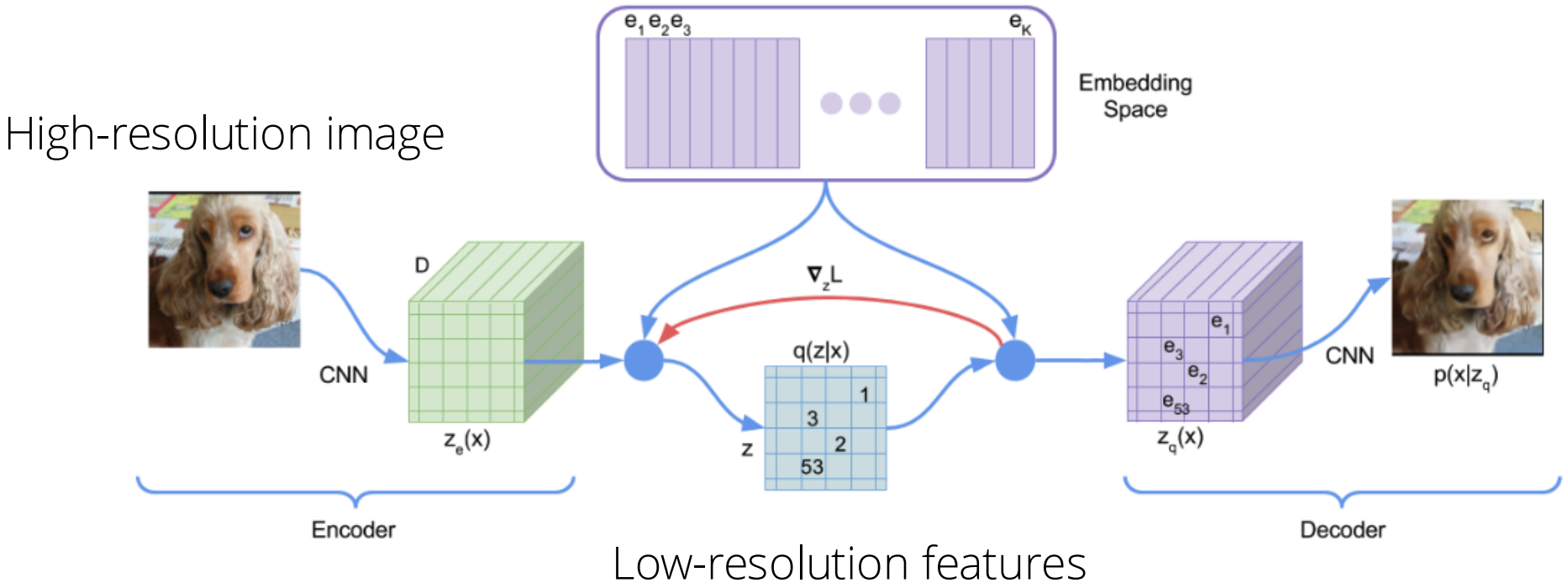


Generative model

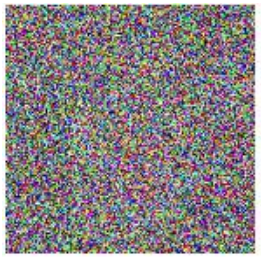
- Generating rgb pixels directly is super expensive!
  - ➢ No structural prior: a 16x16 red patch can be denoted by much simpler representations
  - ➢ High computational cost

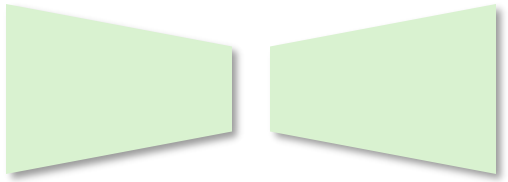
Idea: Encode Images in a Compact Feature Space

- Learn to encode images by auto-encoding
  - ➢ information bottleneck + reconstruction loss



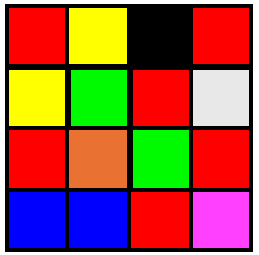
Idea: Generate Images in a Compact Feature Space





Generative model

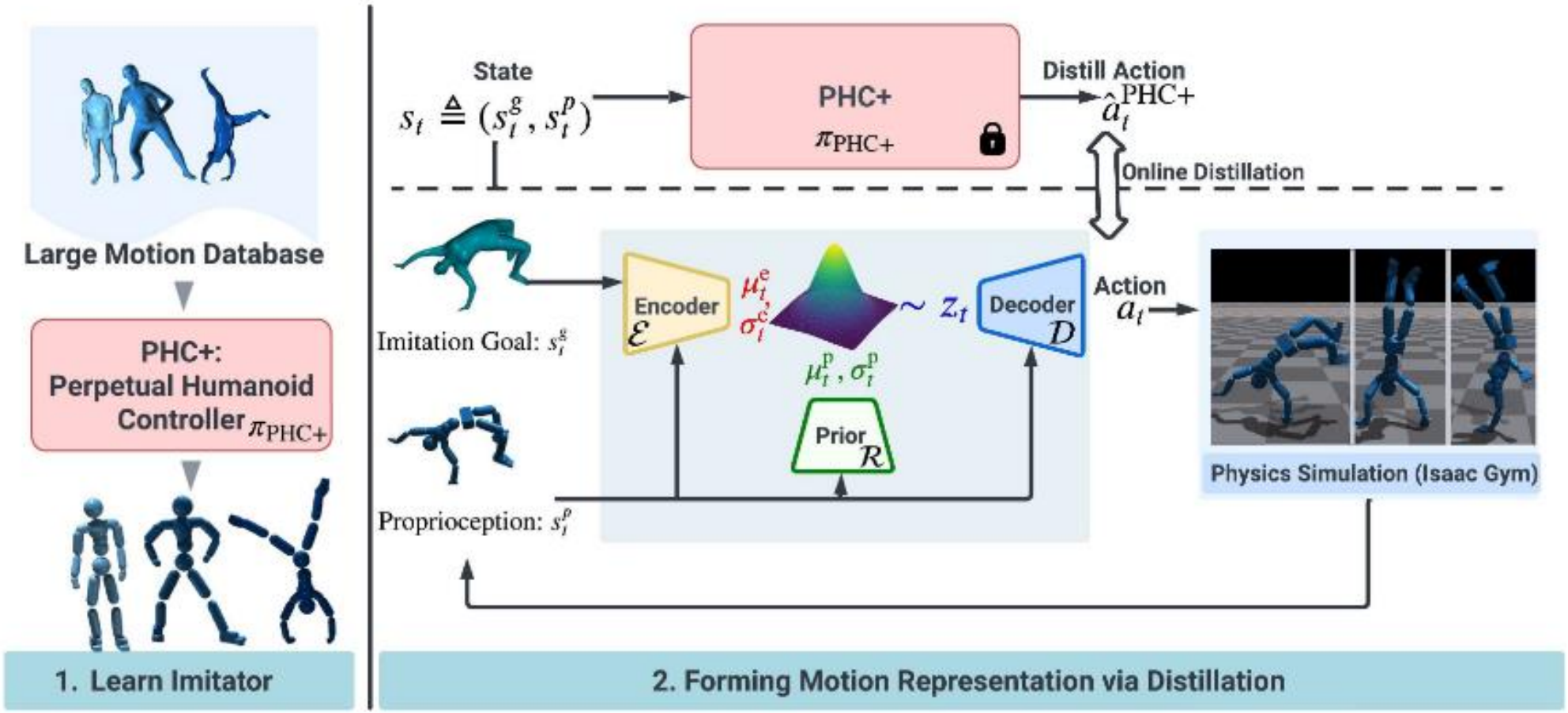
Low-resolution features



Decoder

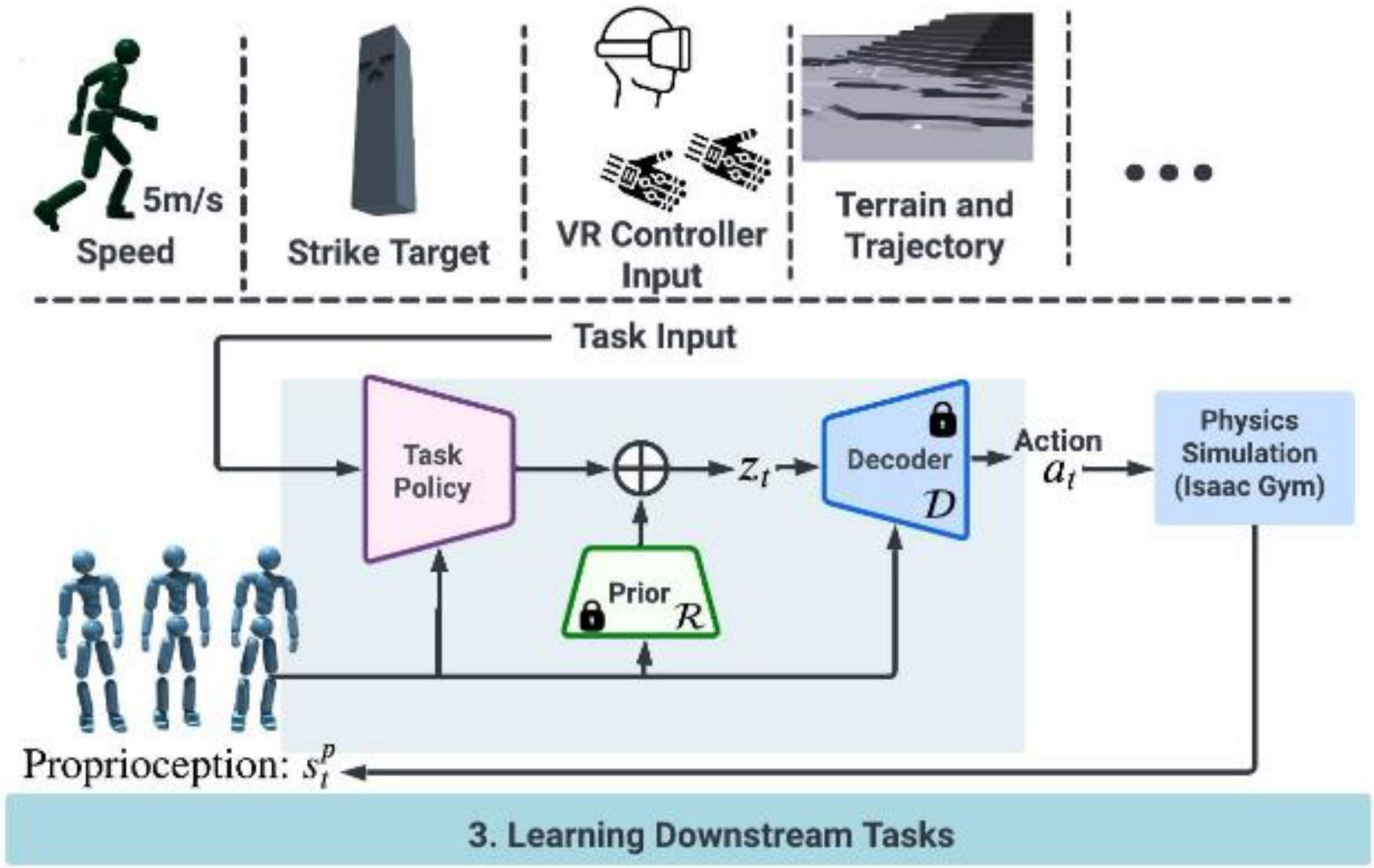
High-resolution image

Let's Apply the Same Idea for RL Policies



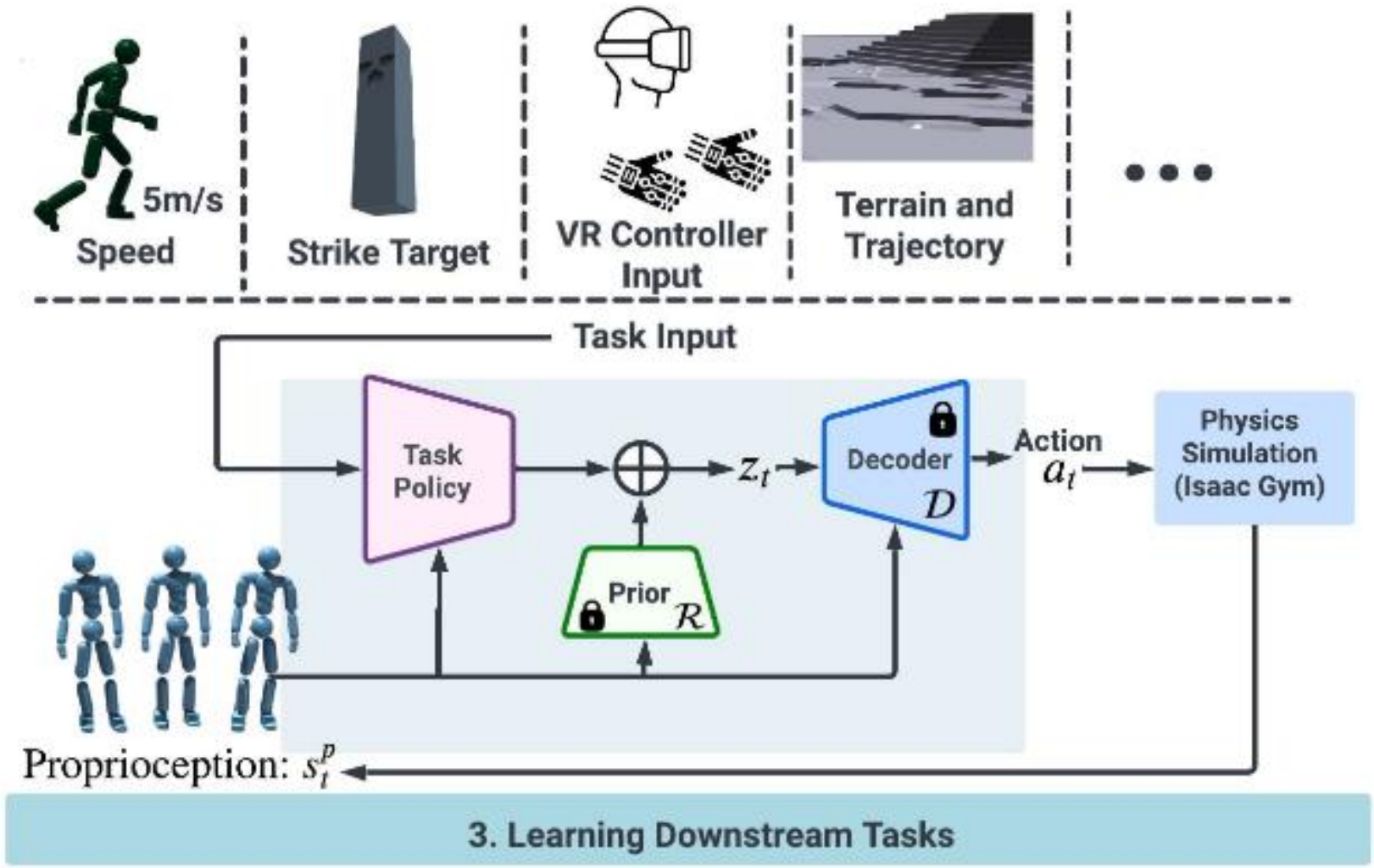
Universal Humanoid Motion Representations for Physics-Based Control. Luo et al.

Efficient RL Adaptation to Downstream Tasks

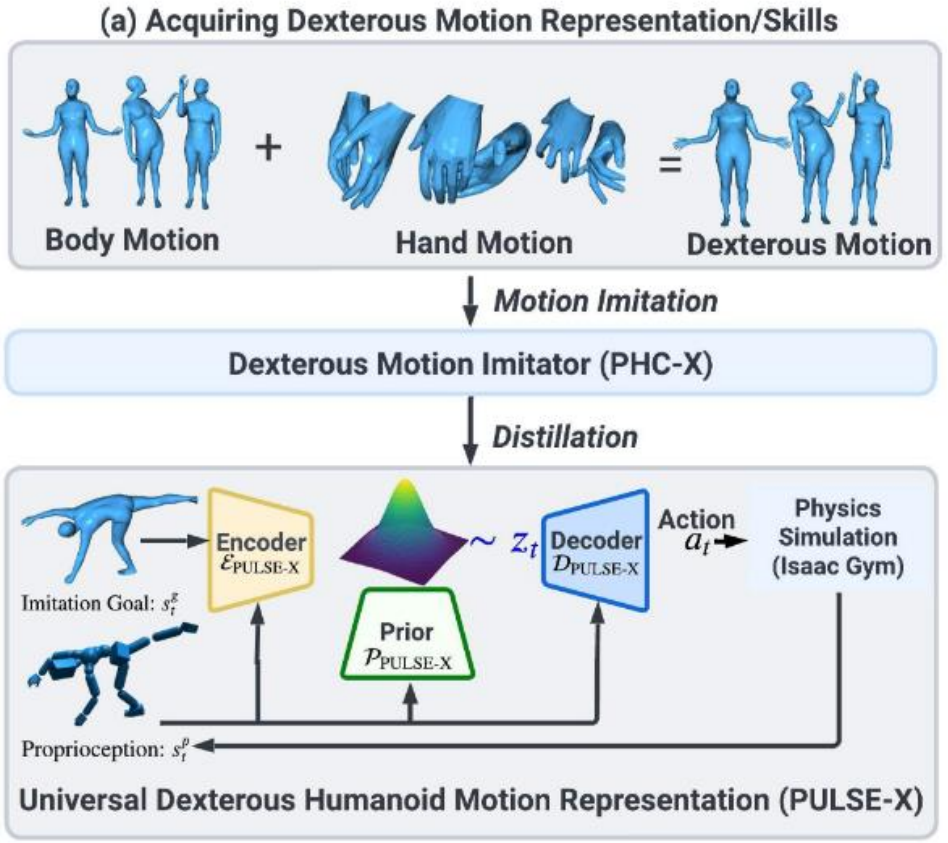


Universal Humanoid Motion Representations for Physics-Based Control. Luo et al.

Efficient RL Adaptation to Downstream Tasks

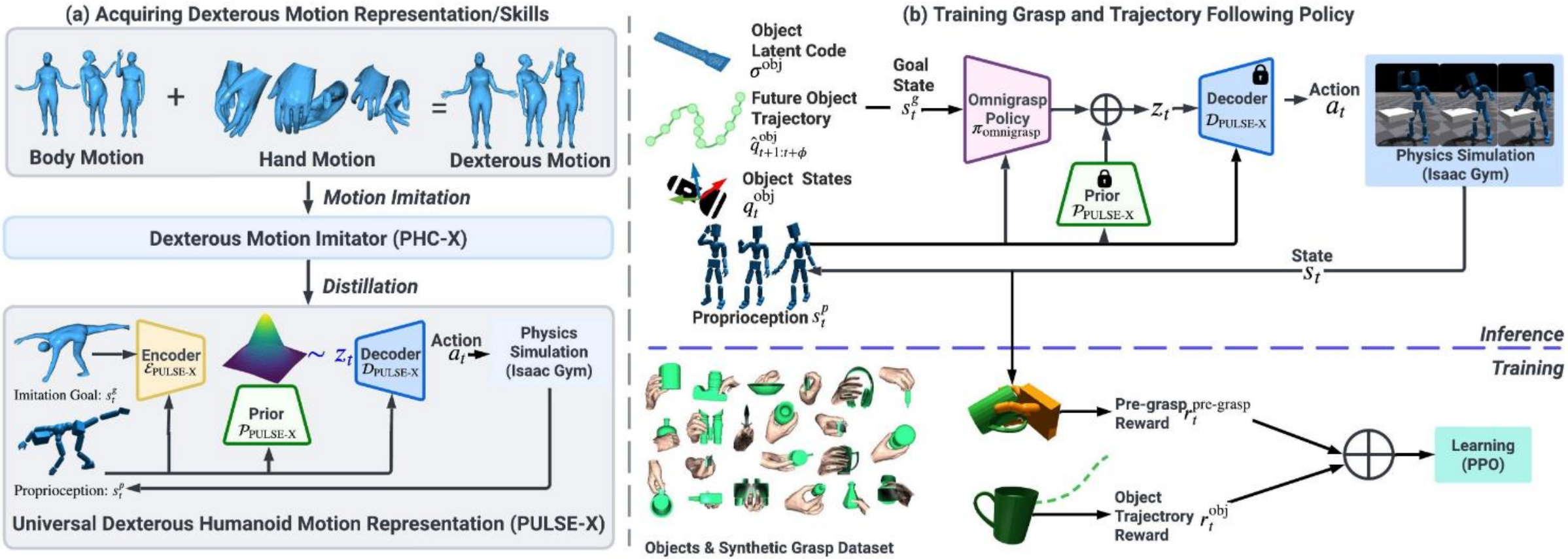


The Same Idea Extends to Mobile Manipulation



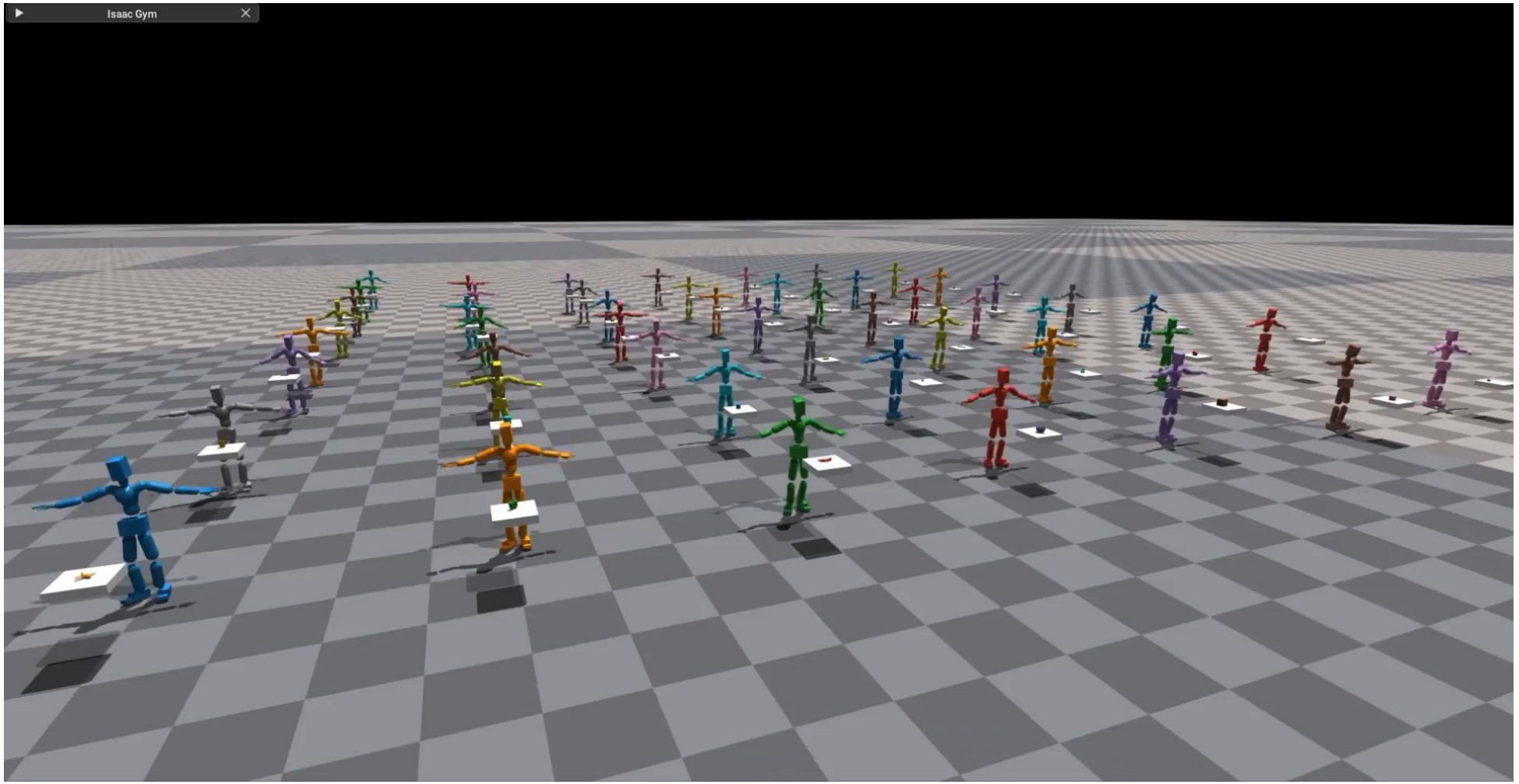
Omnigrasp: Grasping Diverse Objects with Simulated Humanoids. Luo et al.

The Same Idea Extends to Mobile Manipulation



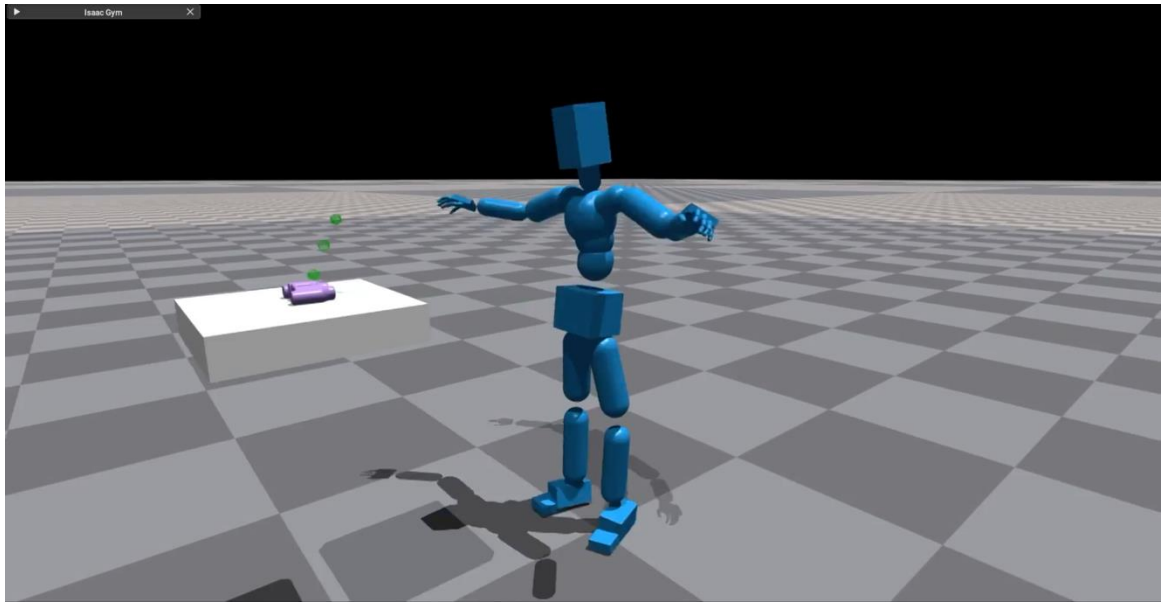
Omnigrasp: Grasping Diverse Objects with Simulated Humanoids. Luo et al.

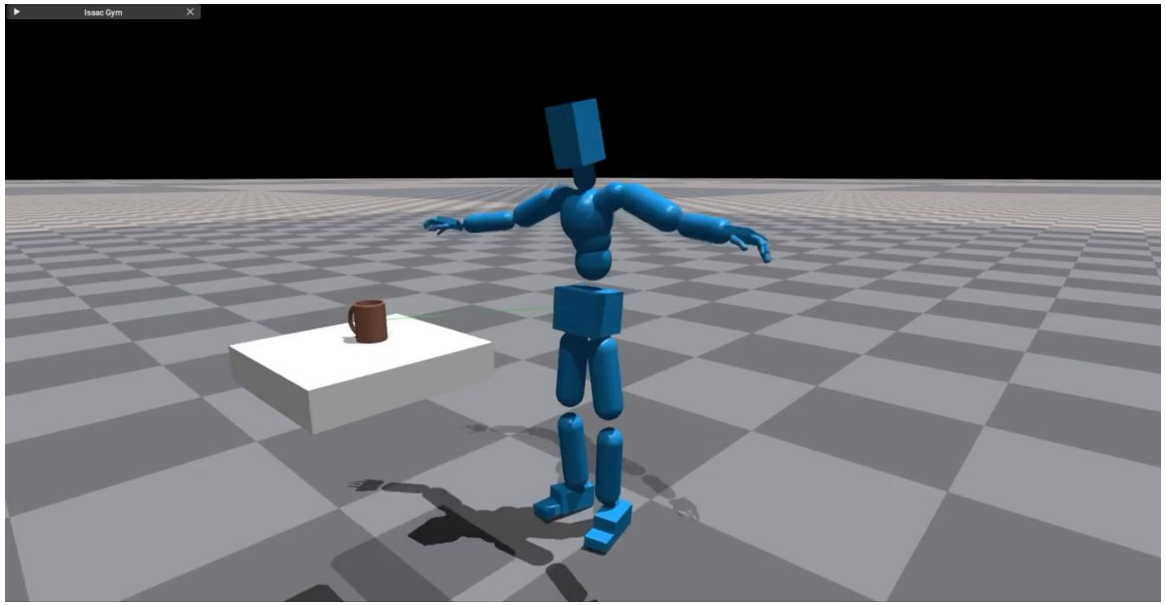
In-Domain Evaluation



Omnigrasp: Grasping Diverse Objects with Simulated Humanoids. Luo et al.

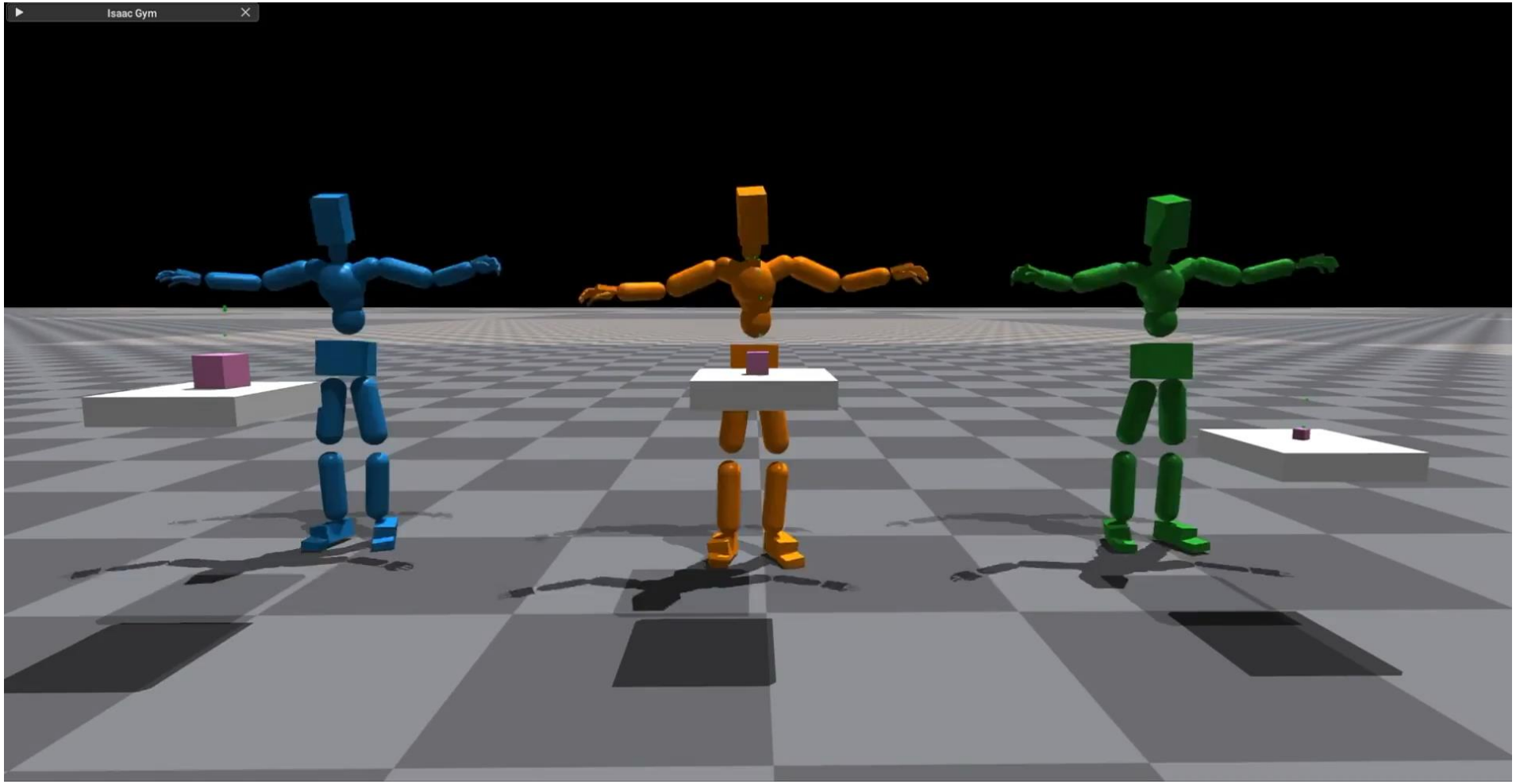
Out-of-Domain Evaluation: Unseen Object Instances





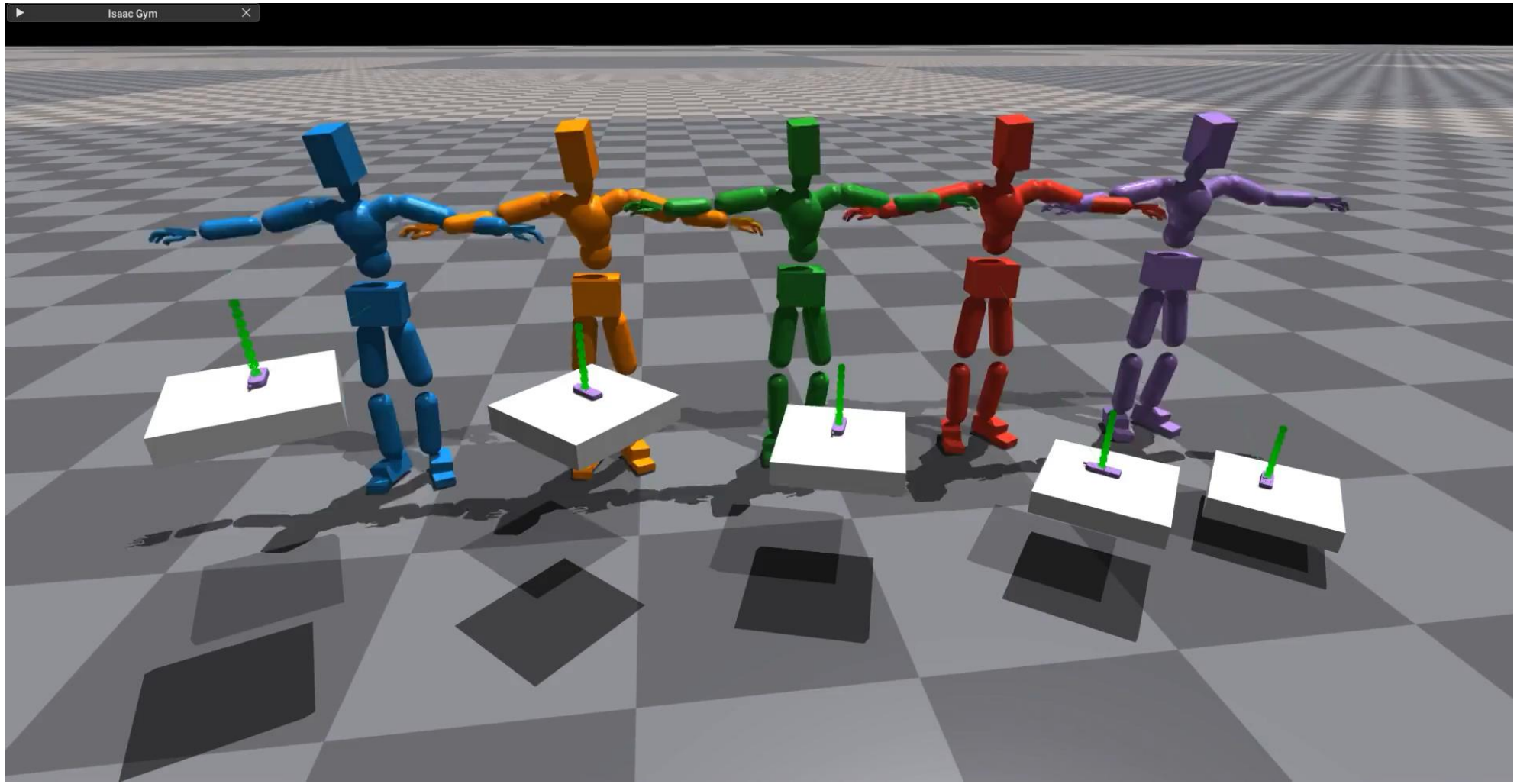
Omnigrasp: Grasping Diverse Objects with Simulated Humanoids. Luo et al.

Out-of-Domain Evaluation: Unseen Object Scales



Omnigrasp: Grasping Diverse Objects with Simulated Humanoids. Luo et al.

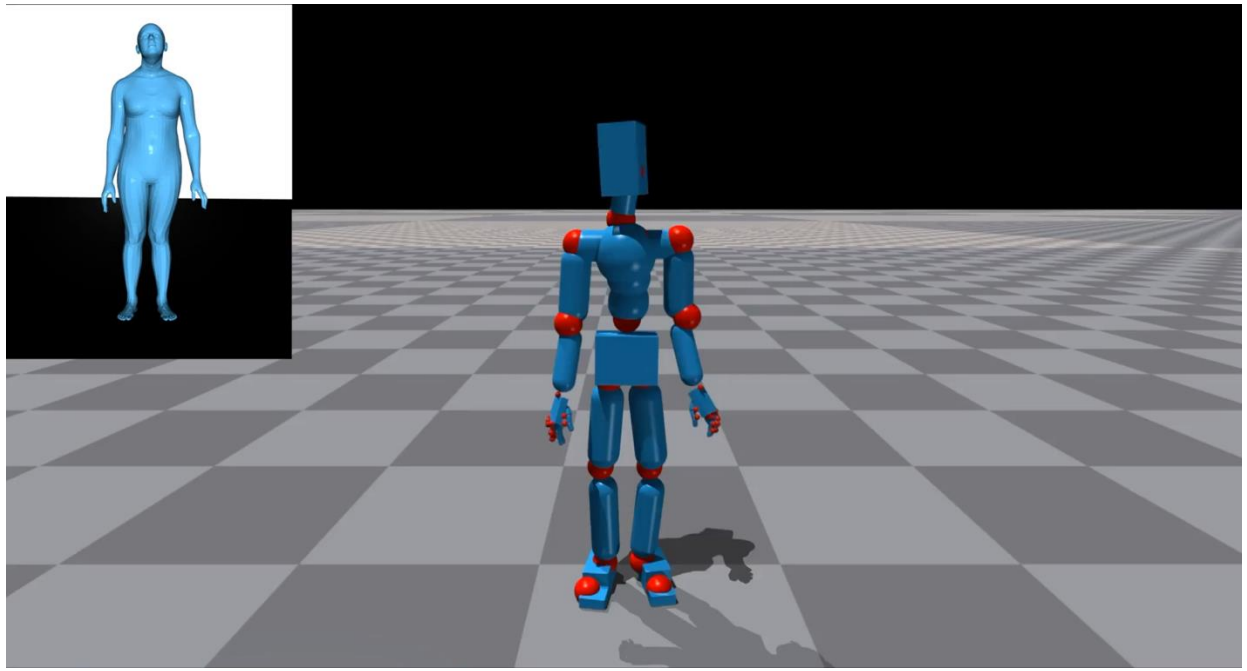
Out-of-Domain Evaluation: Unseen Object Poses

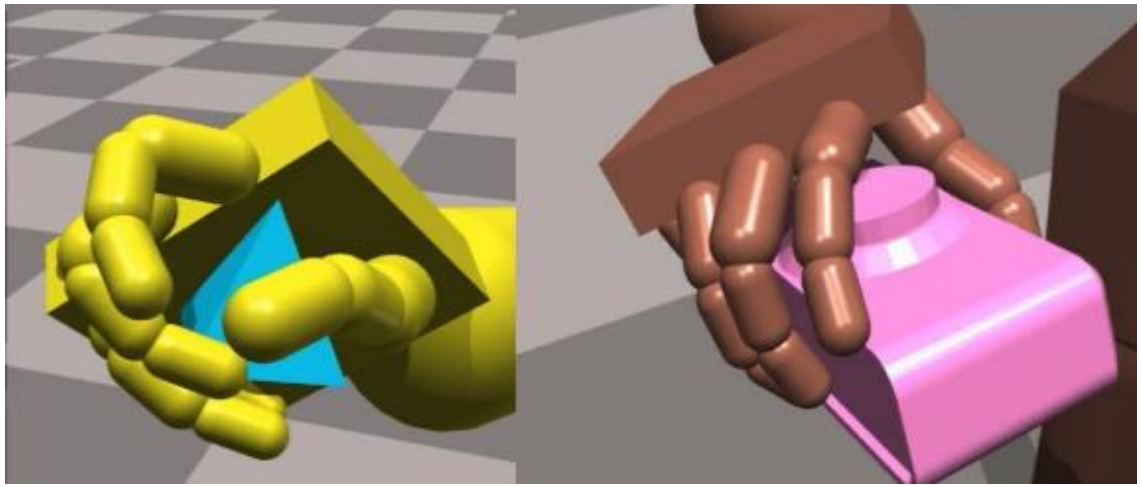


Omnigrasp: Grasping Diverse Objects with Simulated Humanoids. Luo et al.

Have We Solved Mobile Manipulation?

- No! Oversimplified humanoid robots, overlooking embodiment gap





Omnigrasp: Grasping Diverse Objects with Simulated Humanoids. Luo et al.

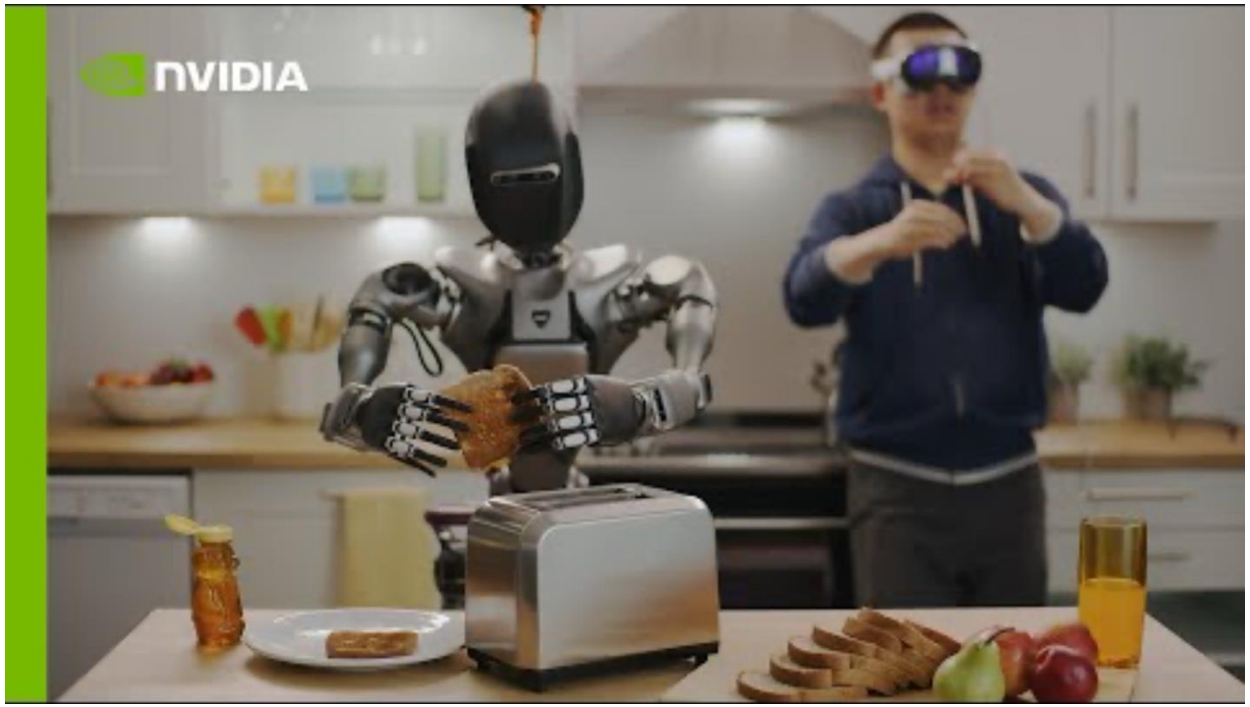
How to Learn Policies for Dexterous Manipulation?

Key to Success: Cloning Human Expert Behaviors

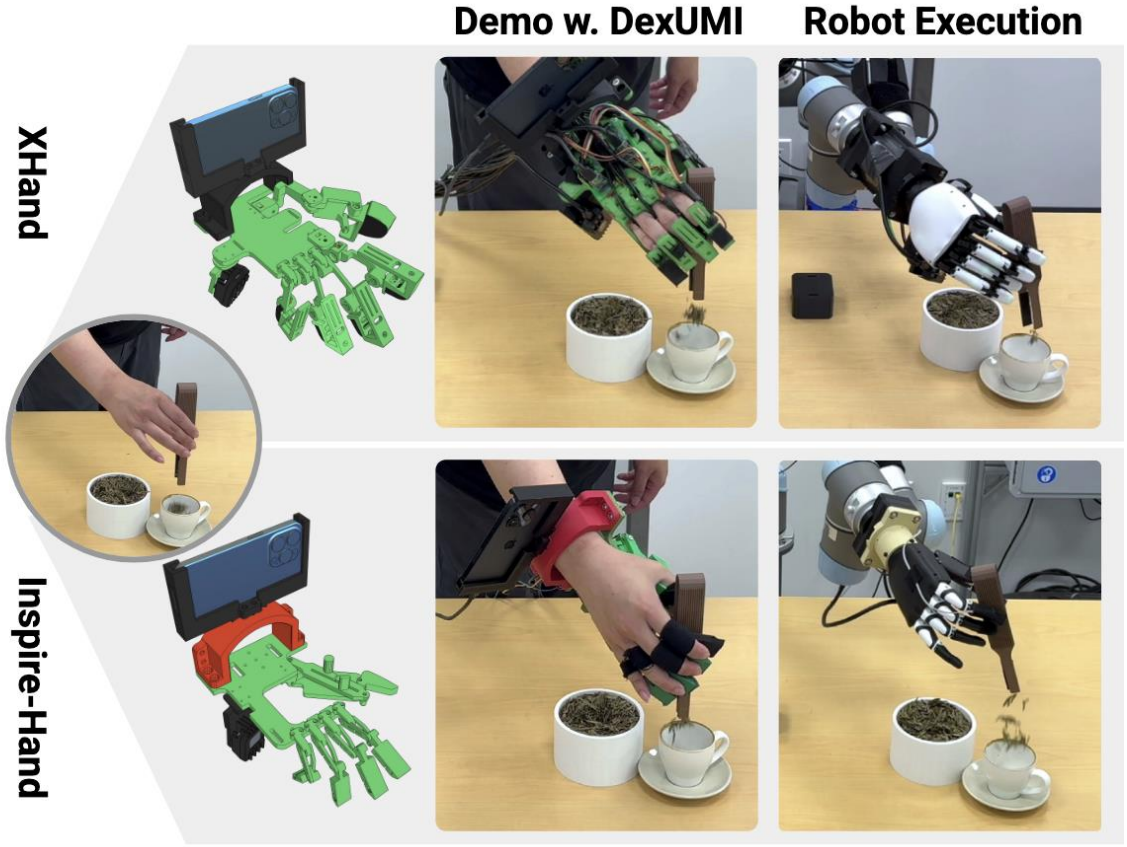
https://blog.ohiohealth.com/simple-ways-teach-family-heritage/

ALOHA 2: An Enhanced Low-Cost Hardware for Bimanual Teleoperation

Collecting Expert Demonstrations for Multi-arm Multi-Fingered Robots is Expensive…

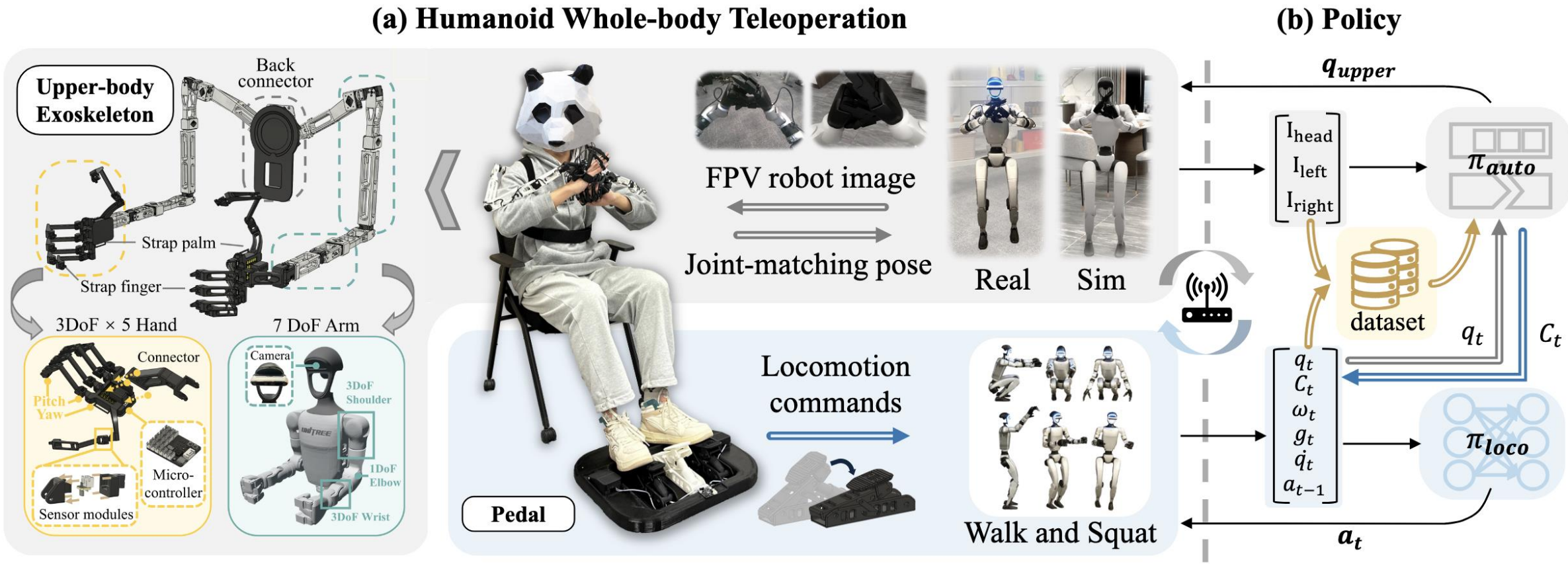


https://youtu.be/Bhg3uOx9ZPw?si=et7L0endzGvGPJz-



DexUMI: Using Human Hand as the Universal Manipulation Interface for Dexterous Manipulation

Collecting Expert Demonstrations for Multi-arm Multi-Fingered Mobile Robots is Even More Expensive…



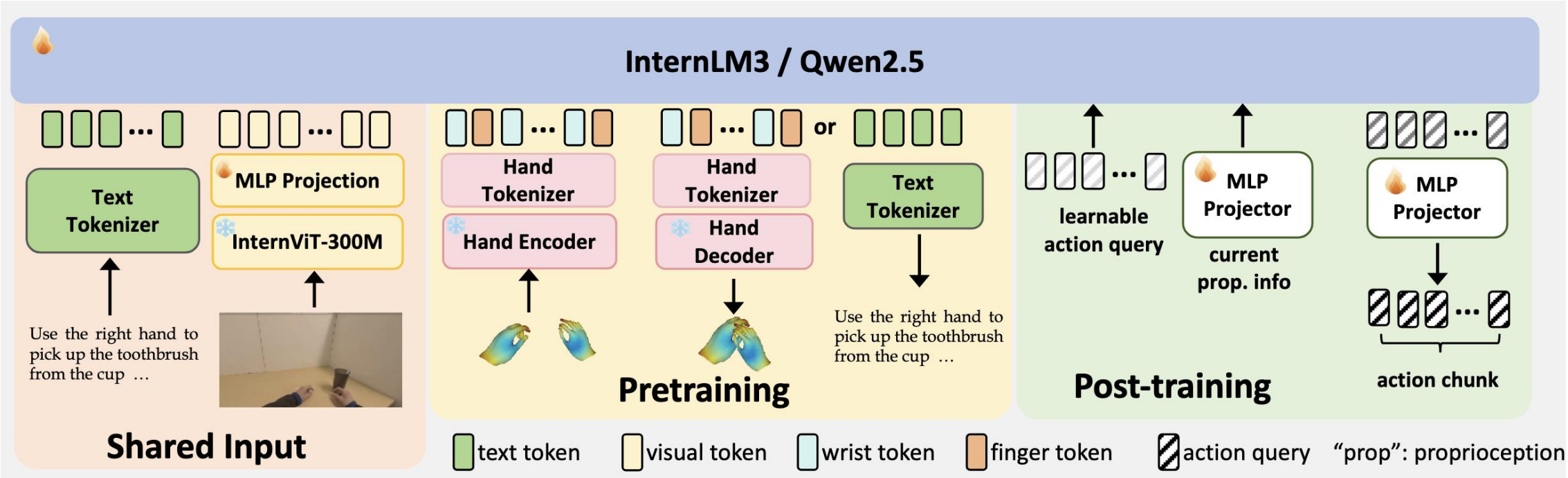
How to Solve the Data Collection Problem?

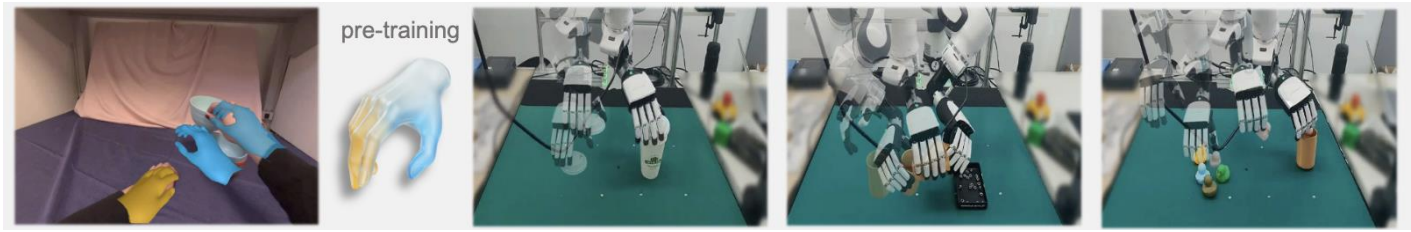
HOT3D by Meta

Veo3 by Google Deepmind

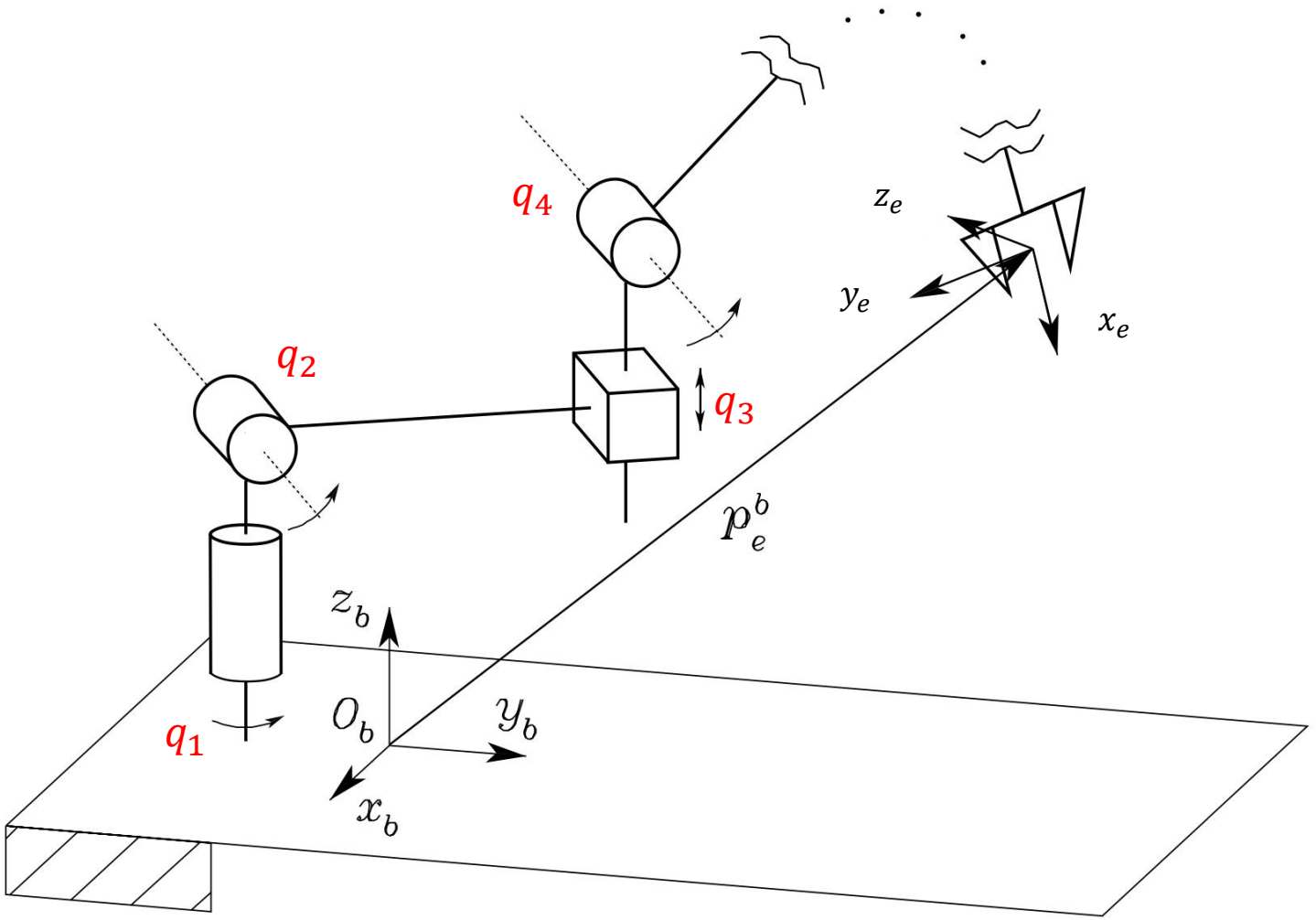
Idea 1: Imitate Human Hand Motions

Idea 1: Imitate Human Hand Motions





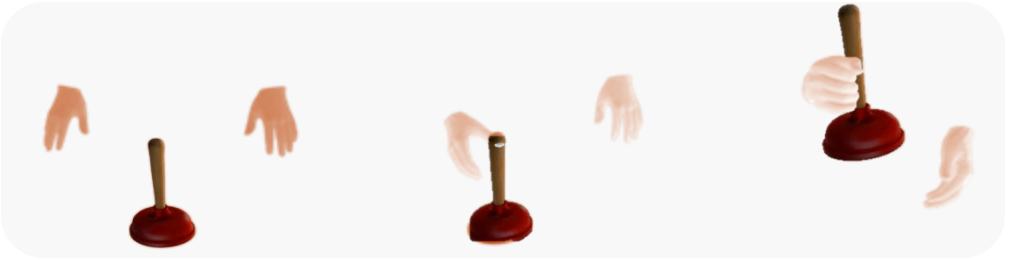
Inverse Kinematic Obtains Robotic Configurations from End-Effector Poses



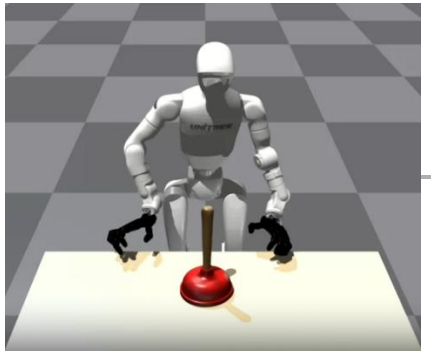
Issue: Inverse Kinematic Fails due to Embodiment Gap

Hand video

Hand pose estimation



Hand pose annotations



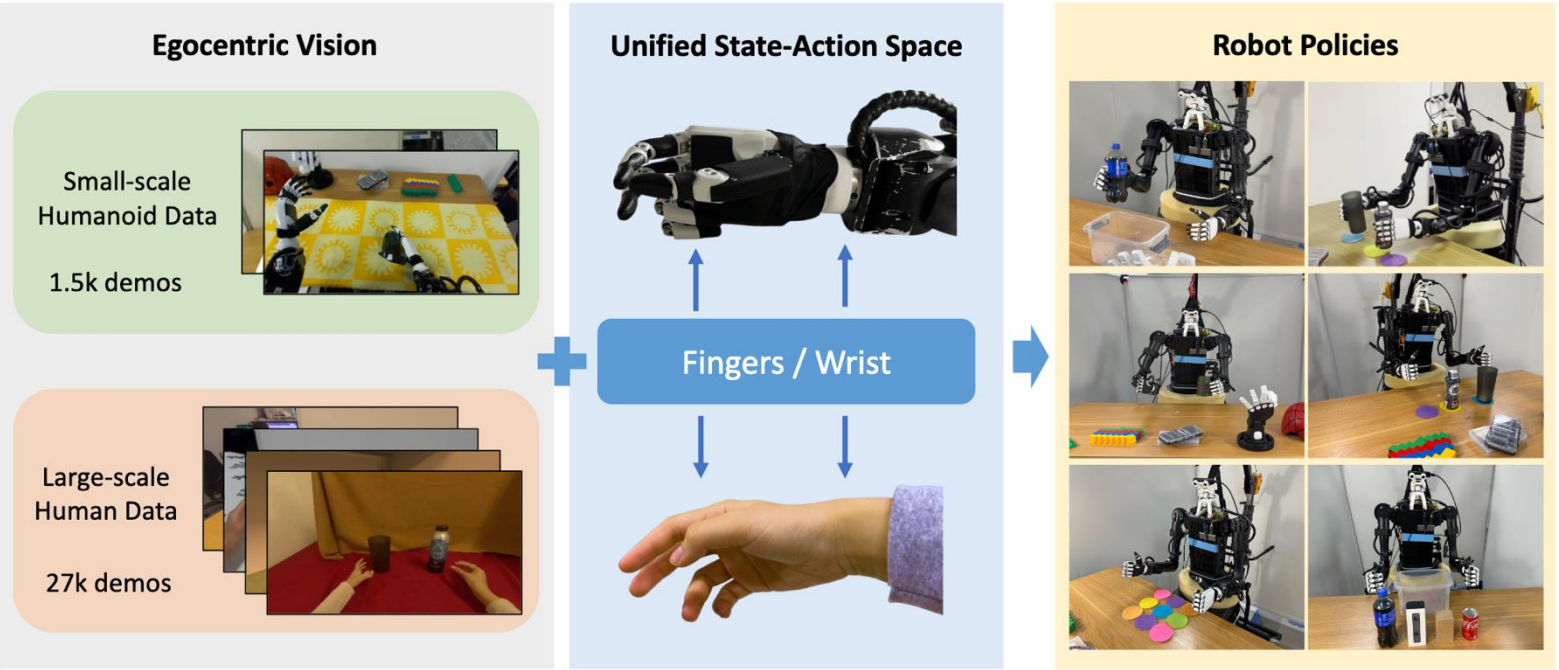
Inverse kinematic generates
failed robot skills

Embodiment Gap: Human Hand vs. Robot Hand

Video by Shadow Robot: https://youtu.be/2ggpvigfEZE?si=3R3k00W_pA1pnInV

Idea 1.5: Joint training Human Actions and Robot Actions

Problem: A small set of Paired human-robot action annotations is needed

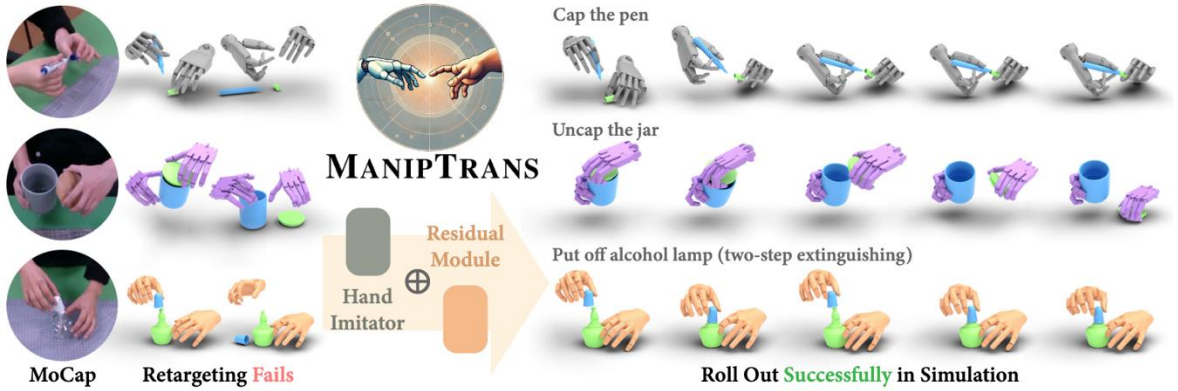


Humanoid Policy ~ Human Policy. Qiu et al.

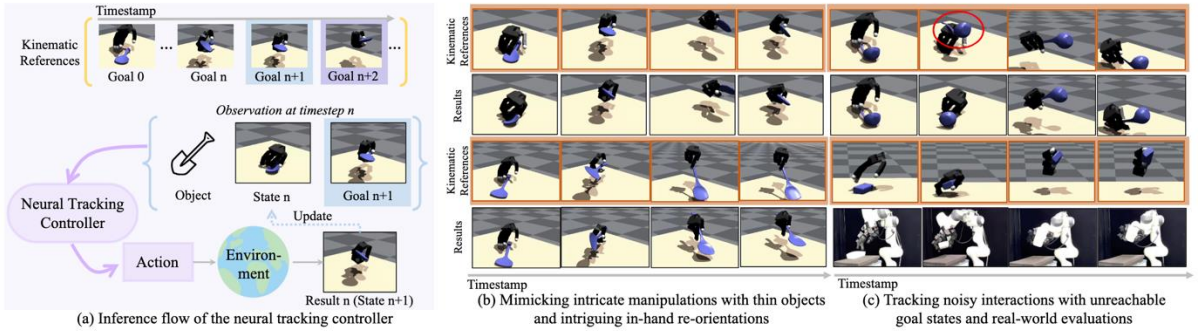
Idea 2: Reinforcement Learning to Imitate Human Hand Motions and Reproduce Object Motions

https://explorer.projectaria.com/hot3d-aria/P0001_95eabc0a?st=%220%22

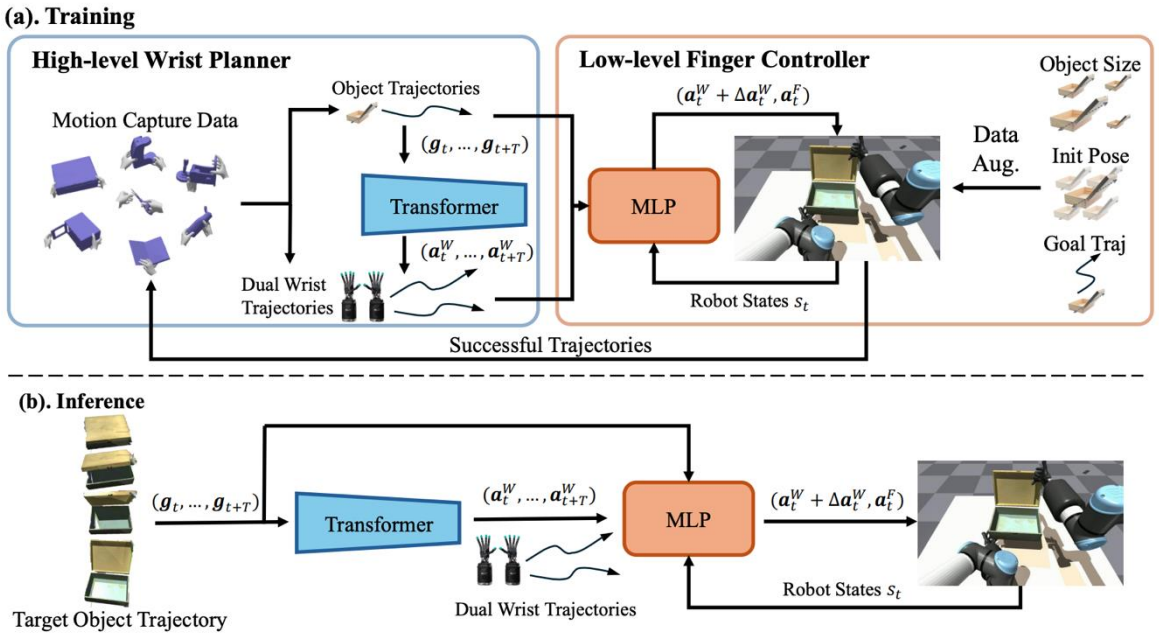
Idea 2: Reinforcement Learning to Imitate Human Hand Motions and Reproduce Object Motions



MANIPTRANS: Efficient Dexterous Bimanual Manipulation Transfer via Residual Learning. Li et al.

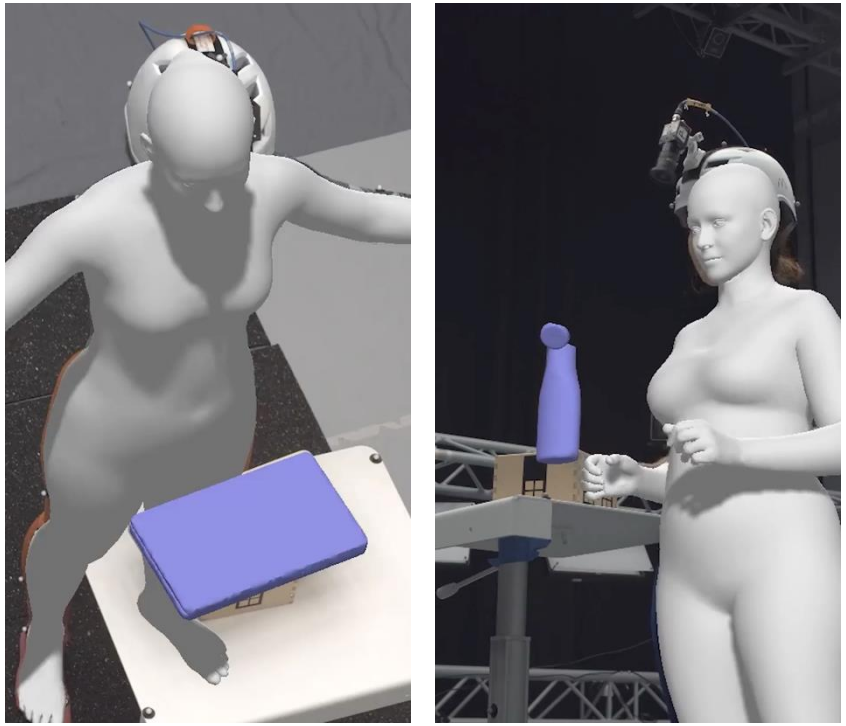


DexTrack: Towards Generalizable Neural Tracking Control for Dexterous Manipulation from Human References. Liu et al.

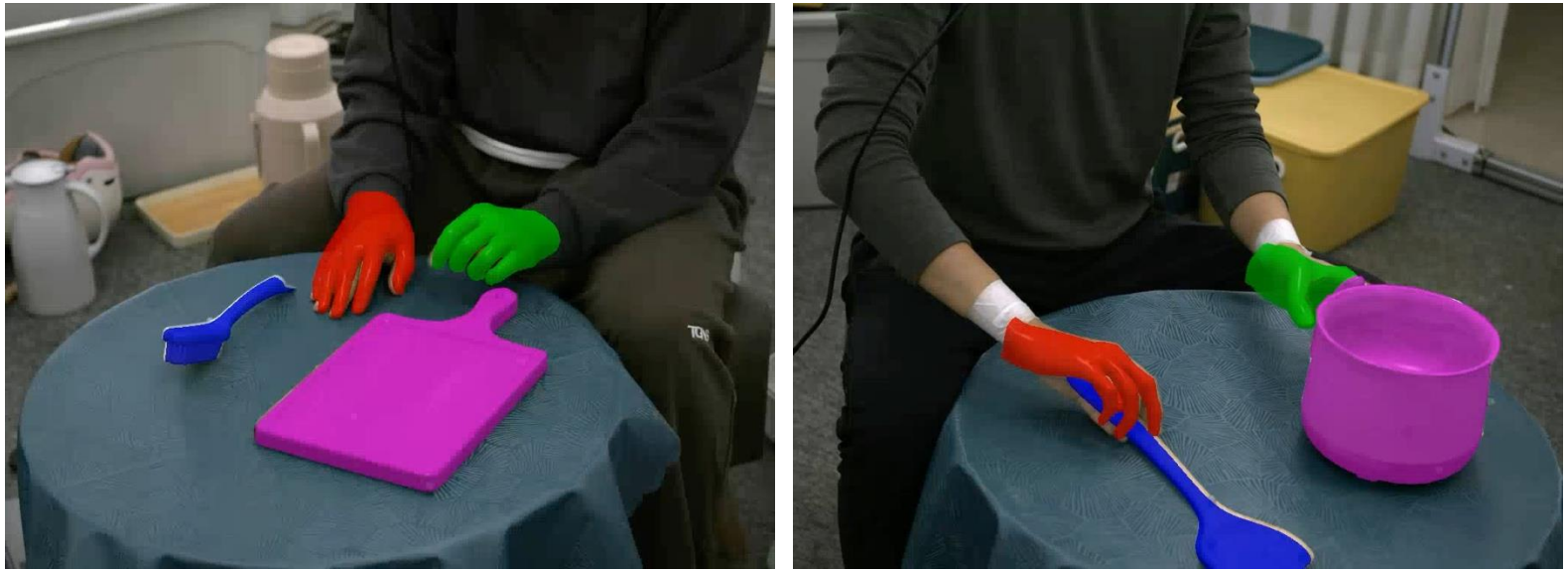


Object-Centric Dexterous Manipulation from Human Motion Data. Chen et al.

Recipe 1: A Motion-Capture Dataset with Human and Object Motion Annotations

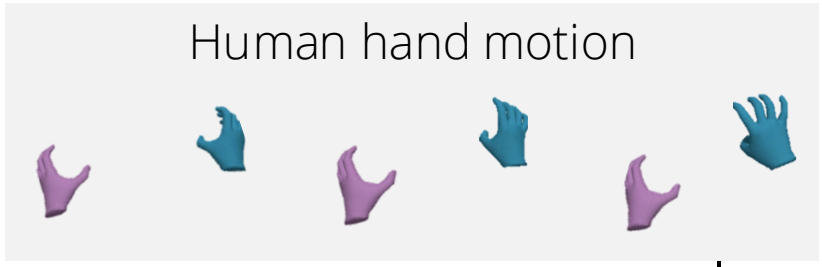


ARCTIC: A Dataset for Dexterous Bimanual Hand-Object Manipulation. Fan et al.



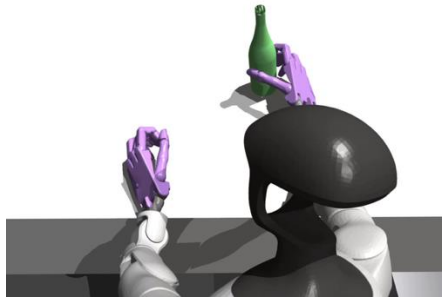
TACO: Benchmarking Generalizable Bimanual Tool-ACtion-Object Understanding. Liu et al.

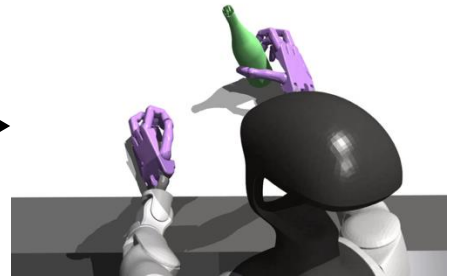
Recipe 2: Two Objectives of Reinforcement Learning

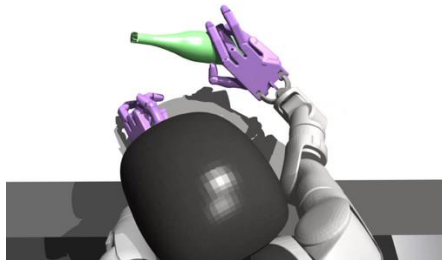


Robot Proprioception

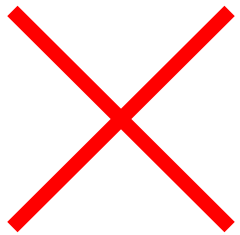
Inverse Kinematic



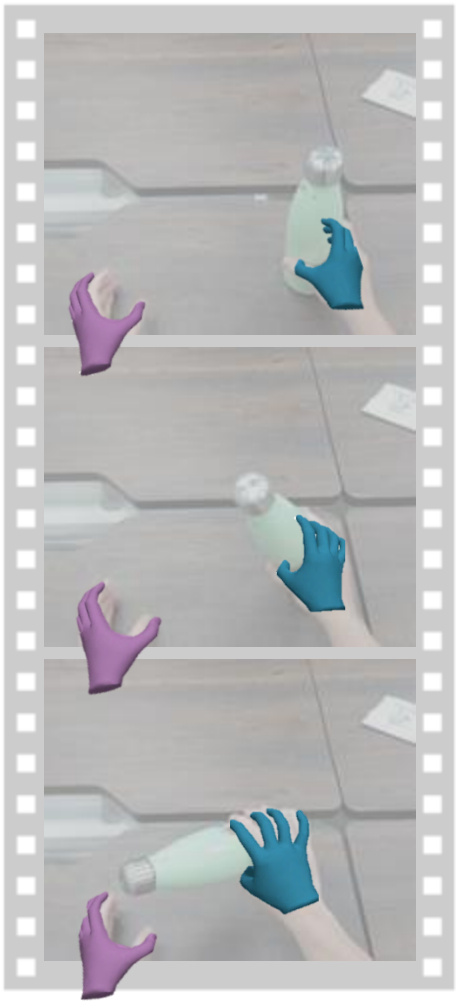




Robot joint angles



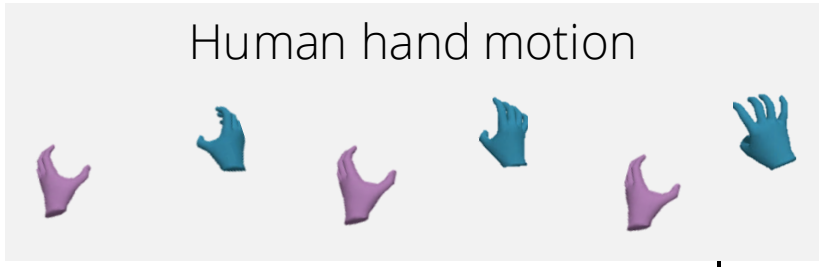
Inverse kinematic doesn't work due to embodiment gap



Goal: Track human hand motions

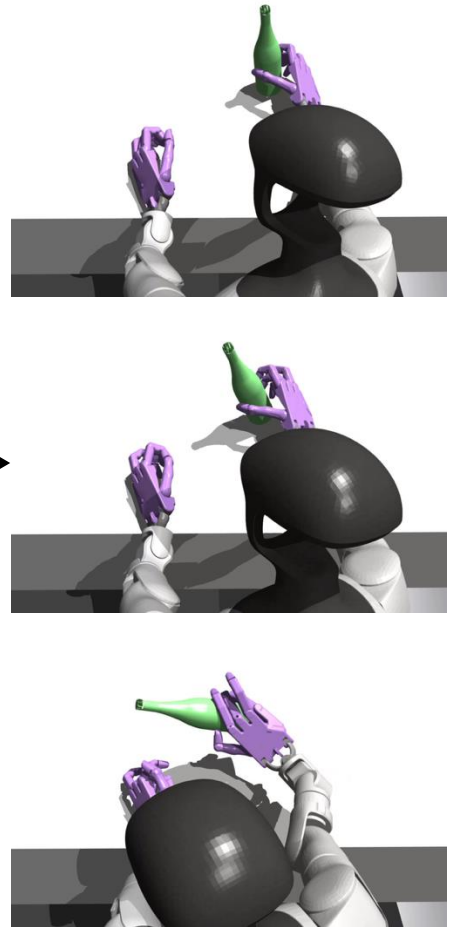
Goal: Reproduce object motions

Recipe 2: Two Objectives of Reinforcement Learning

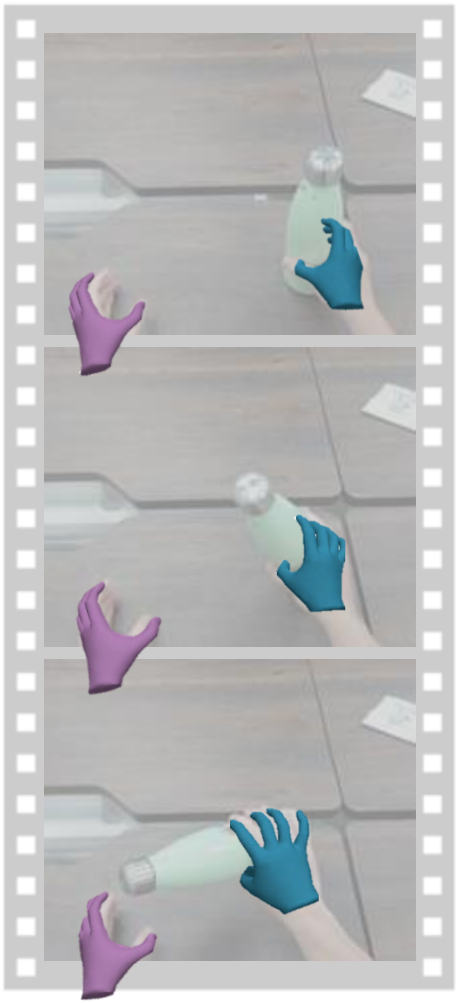


Robot Proprioception

Residual RL



Robot joint angles



Goal: Track human hand motions

Goal: Reproduce object motions

Results: Human-to-Robot Motion Retargeting

Shake the flask	Scoop something	Stir

Issue 1: Previous methods only consider floating hands, overly simplifying the motion retargeting problem
Issue 2: Expensive motion-capture data is required

DexMan: Learning Bimanual Dexterous Manipulation from Human and Generated Videos

Jhen Hsieh

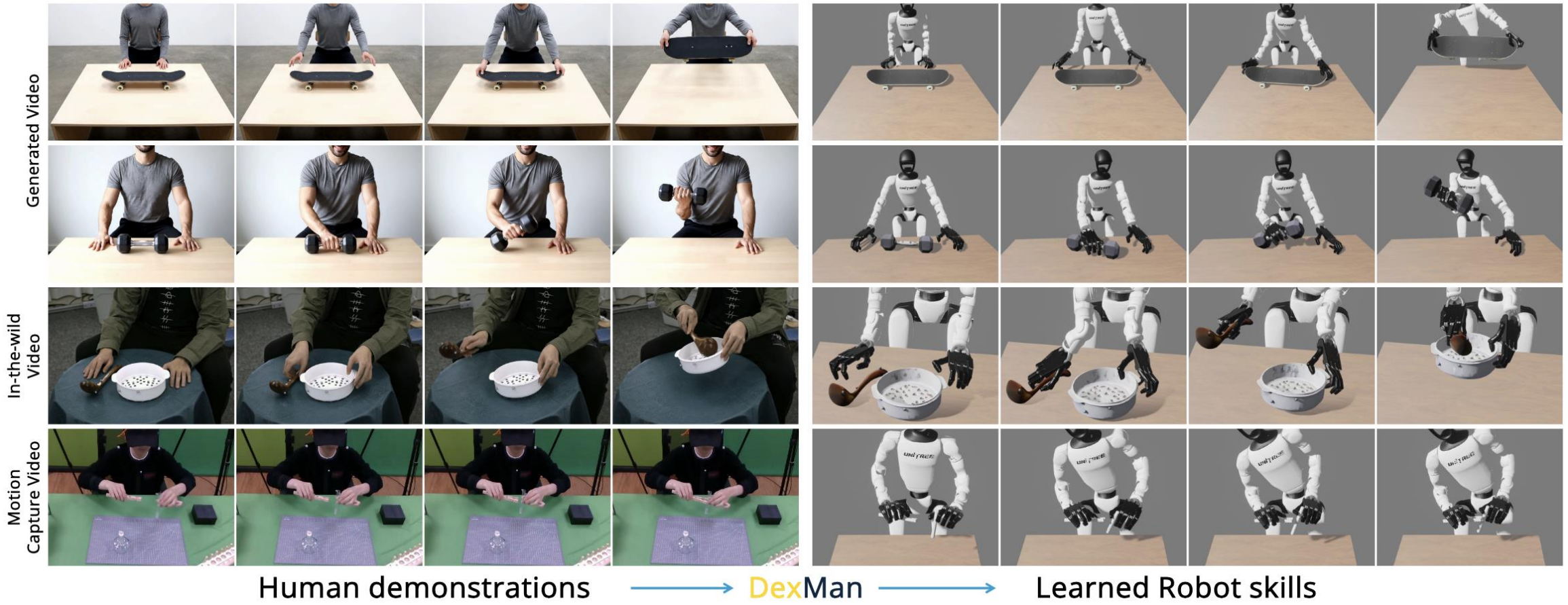
Kuan-Hsun Tu

Kuo-Han Hung

Tsung-Wei Ke

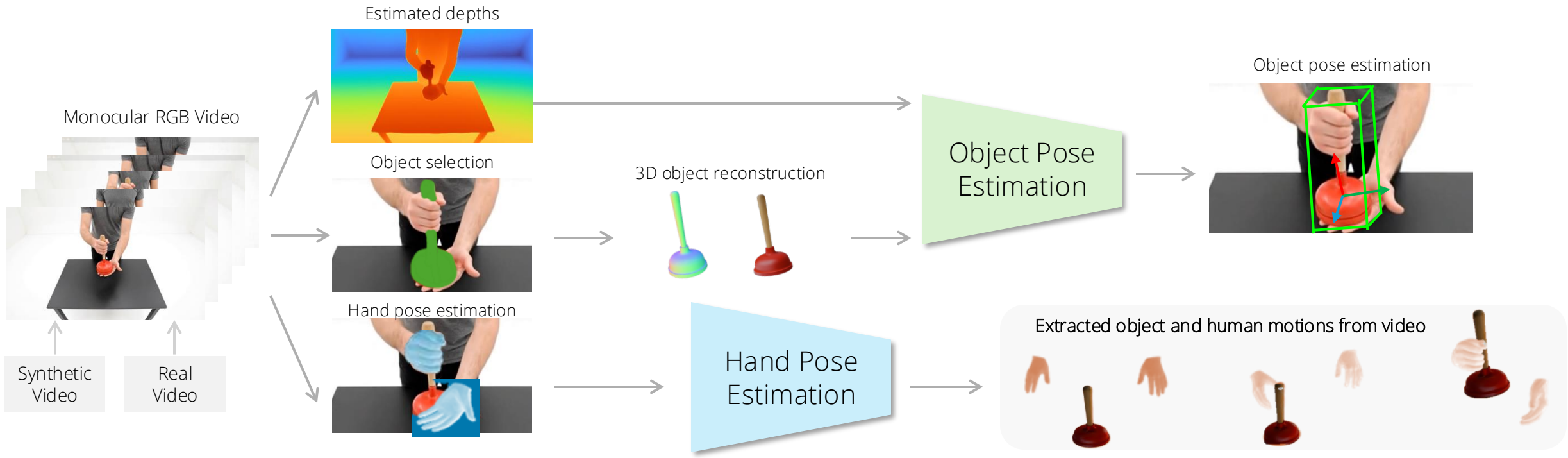
In submission

Our Goal: From RGB Video to Plausible Humanoid Robot Actions for Bimanual Dexterous Manipulation

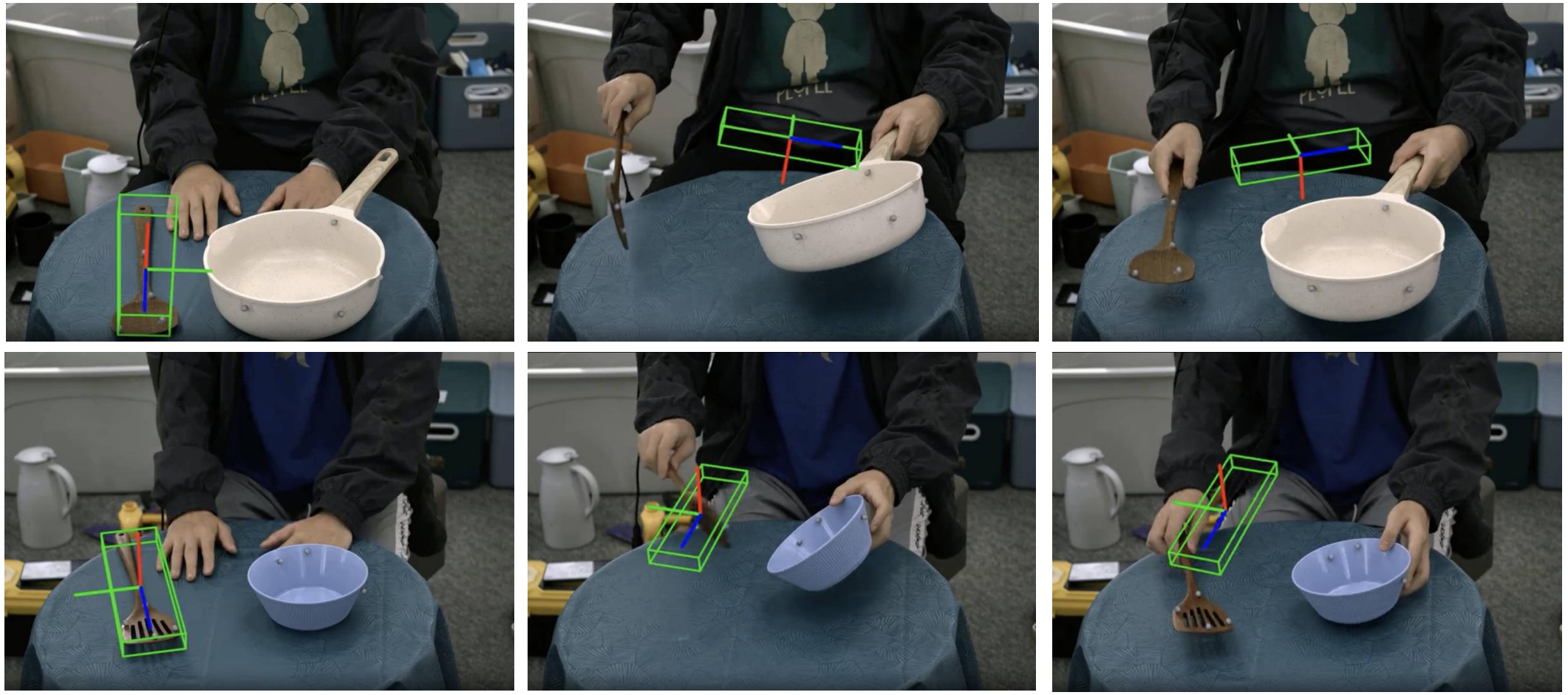


Issue: Expensive motion-capture data is required

- What we need: (1) 3D human hand motion, and (2) 3D object motion
- Solution: existing computer vision methods **kind of** work

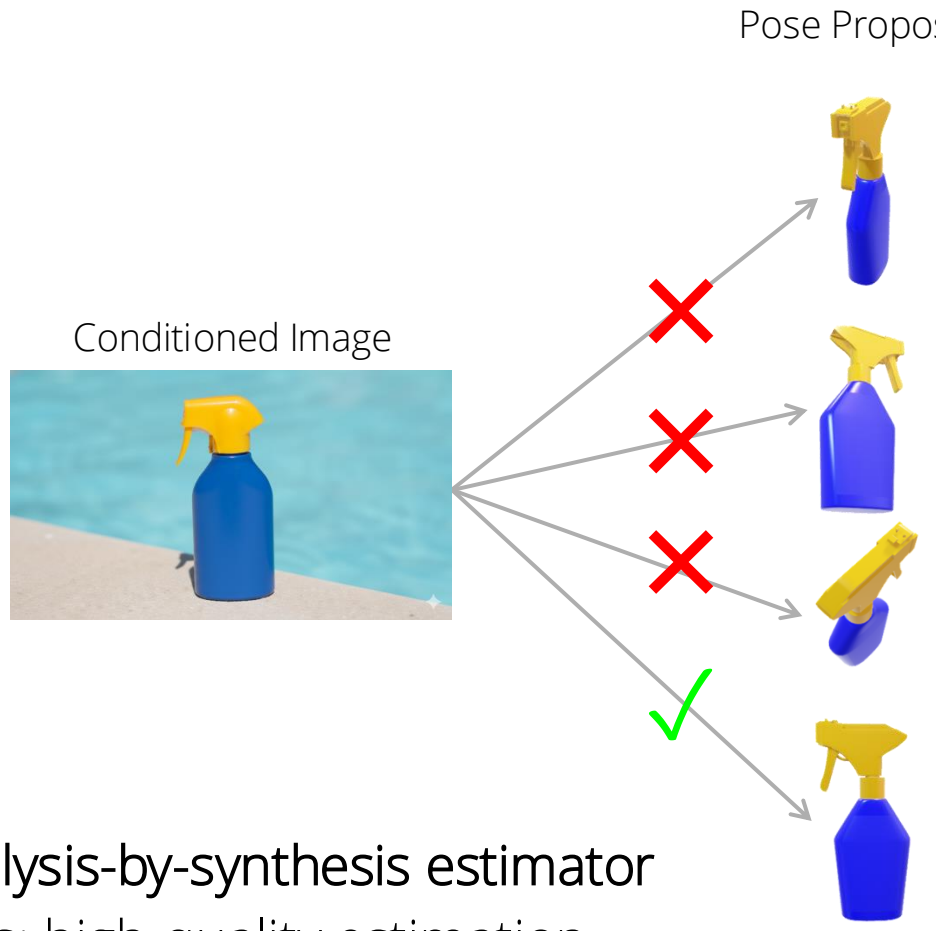


Object Pose Estimation Still Needs Improvement



time

Unstable Analysis-by-Synthesis Object Pose Estimation

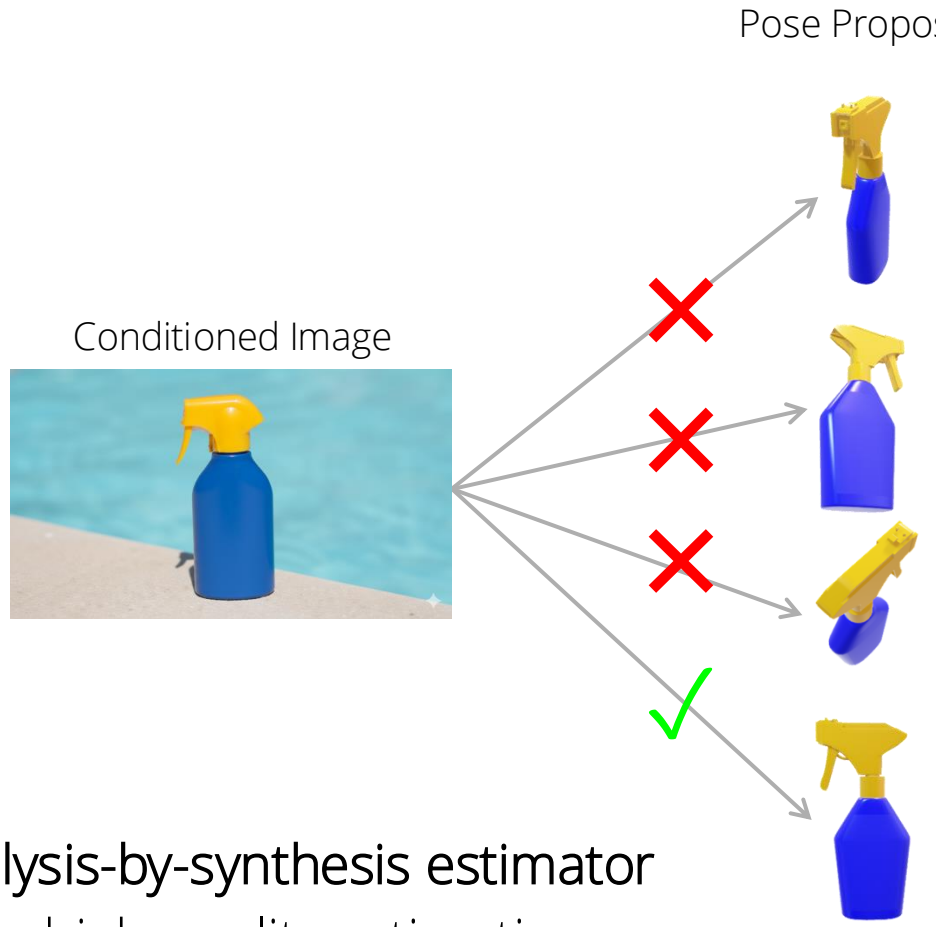


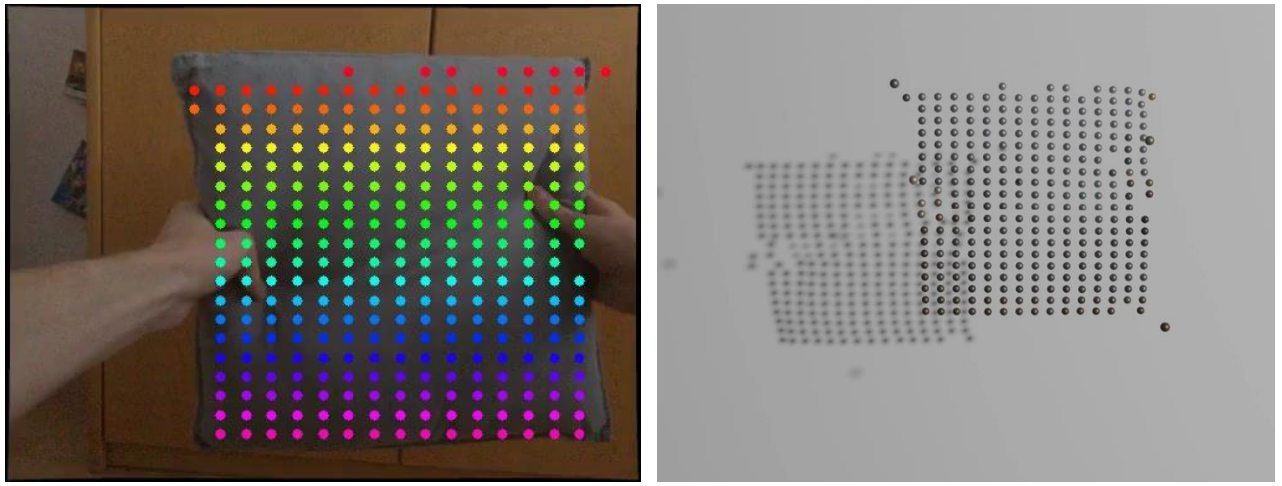
Analysis-by-synthesis estimator

Pros: high-quality estimation

Cons: no temporal consistency

Enhance Temporal Consistency with 3D Point Tracking





SpatialTracker: Tracking Any 2D Pixels in 3D Space. Hsiao et al.

Analysis-by-synthesis estimator

Pros: high-quality estimation

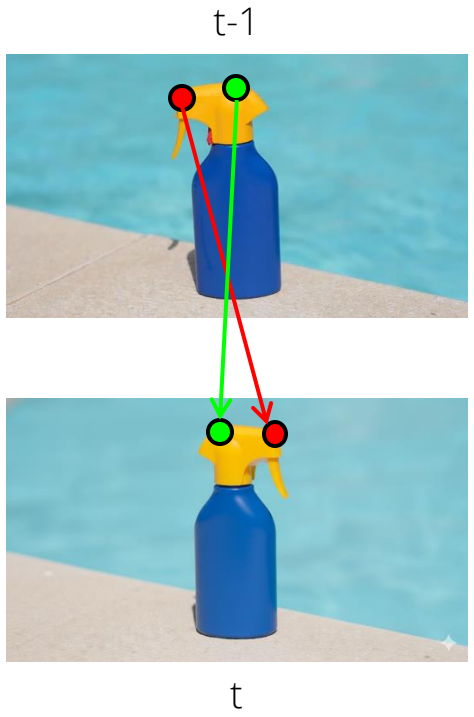
Cons: no temporal consistency

3D Pixel motion estimator

Pros: Spatio-temporal consistency

Cons: no direct rigid-body pose estimation

Enhance Temporal Consistency with 3D Point Tracking



Relative rigid transformation:

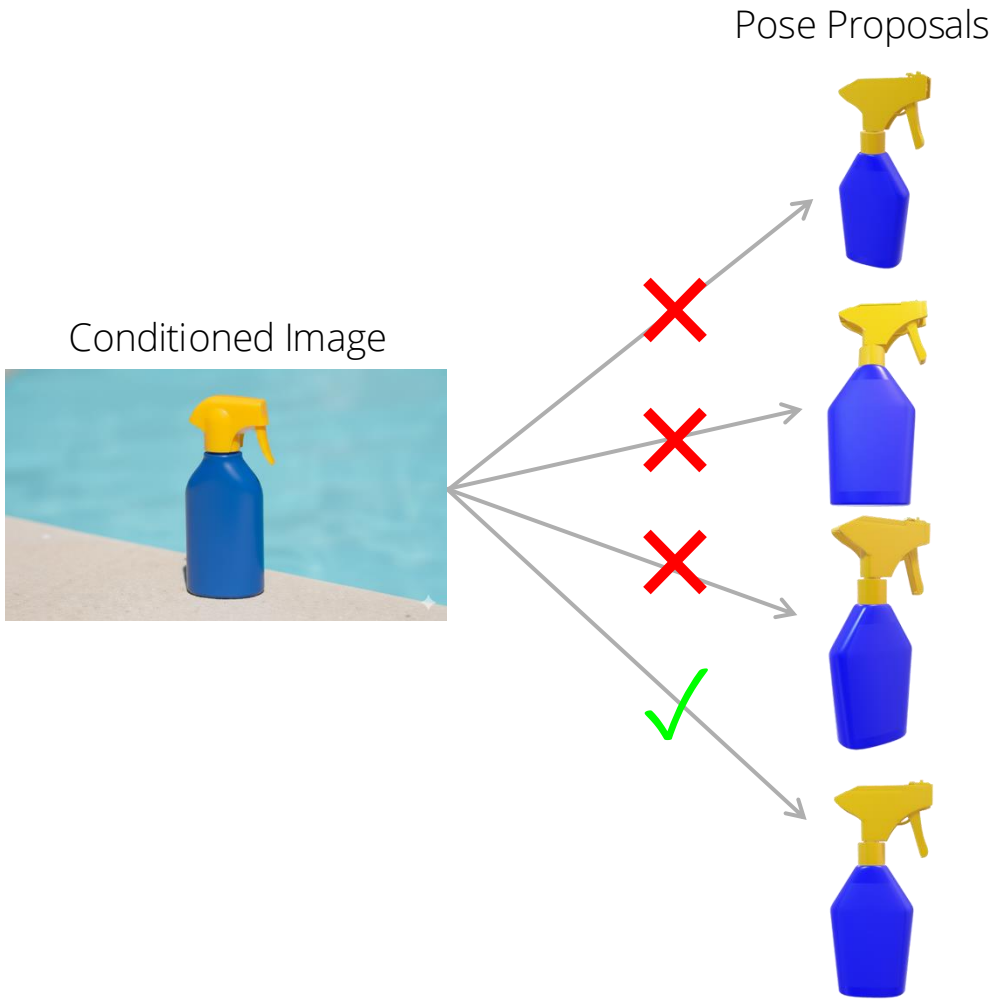
$$\Delta T_{t-1}^{t} = \mathbf{Kabsh}\,(X_{t-1}, X_t)$$

Current pose estimation from 3D point tracks:

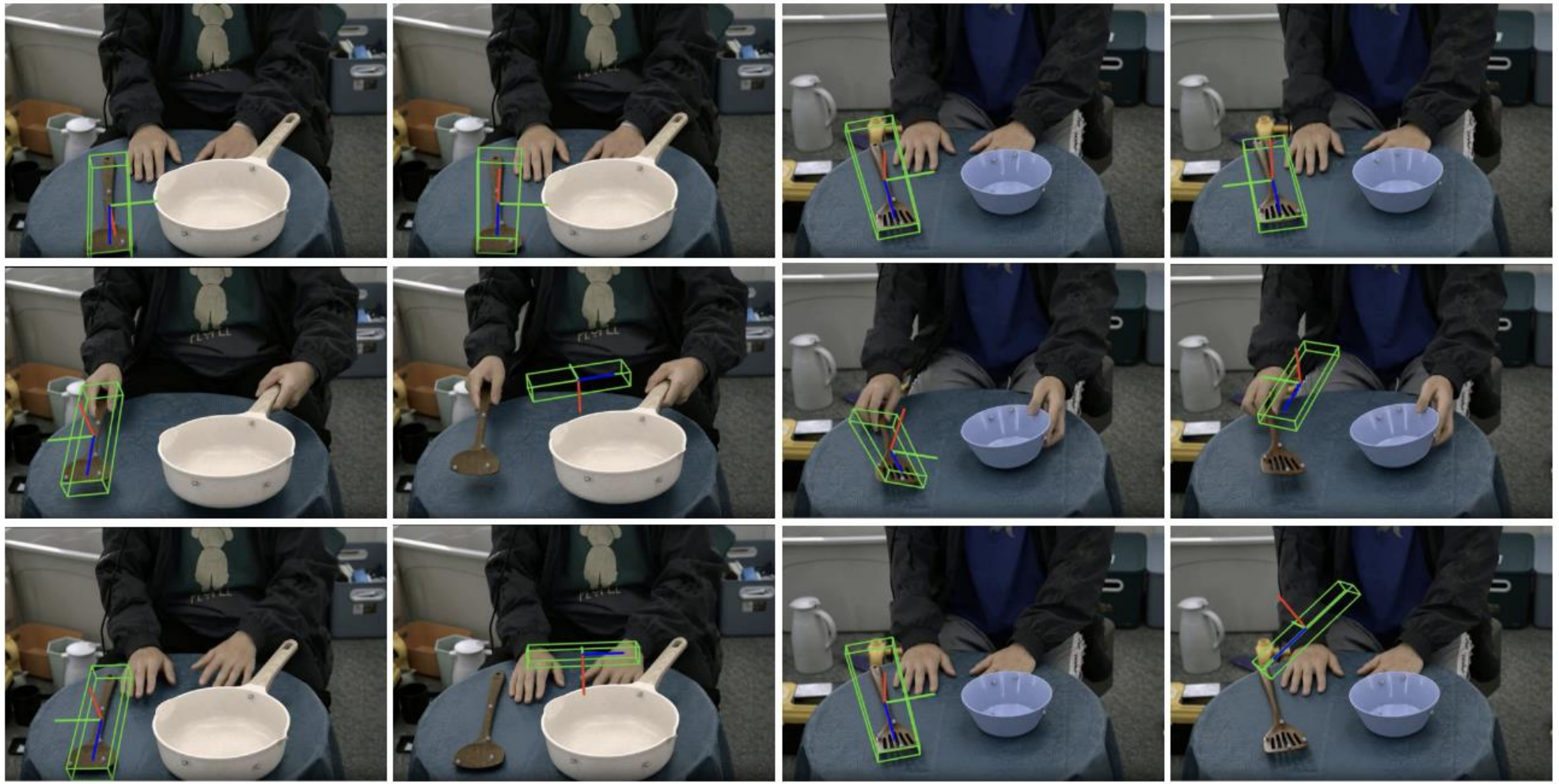
$$P_t = P_{t-1}\Delta T_{t-1}^{t}$$

New pose proposal:

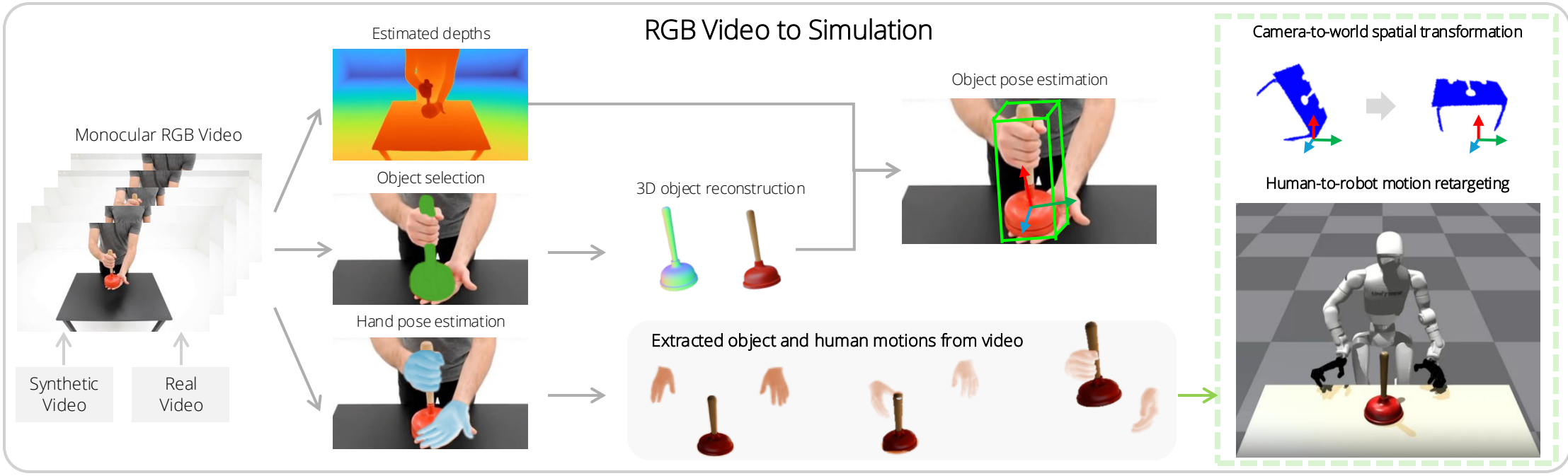
$$P_{t,\mathbf{candidate}} = \{P_t + \epsilon\}$$

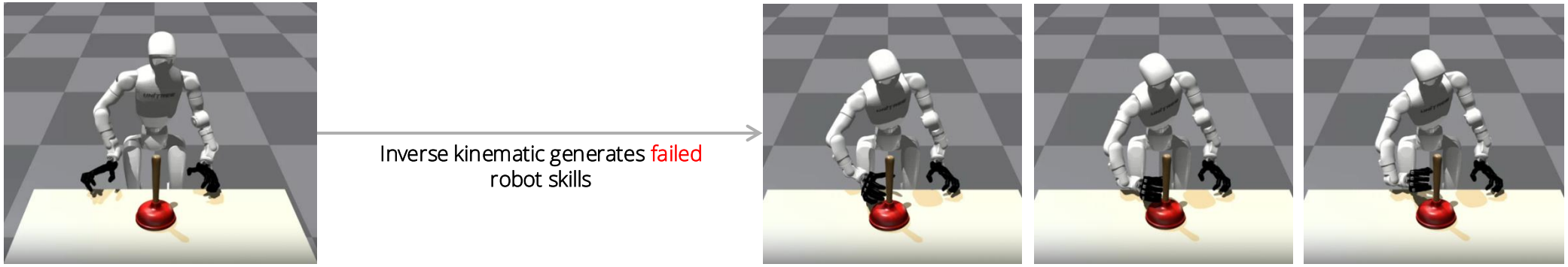


Enhanced Object Pose Estimation

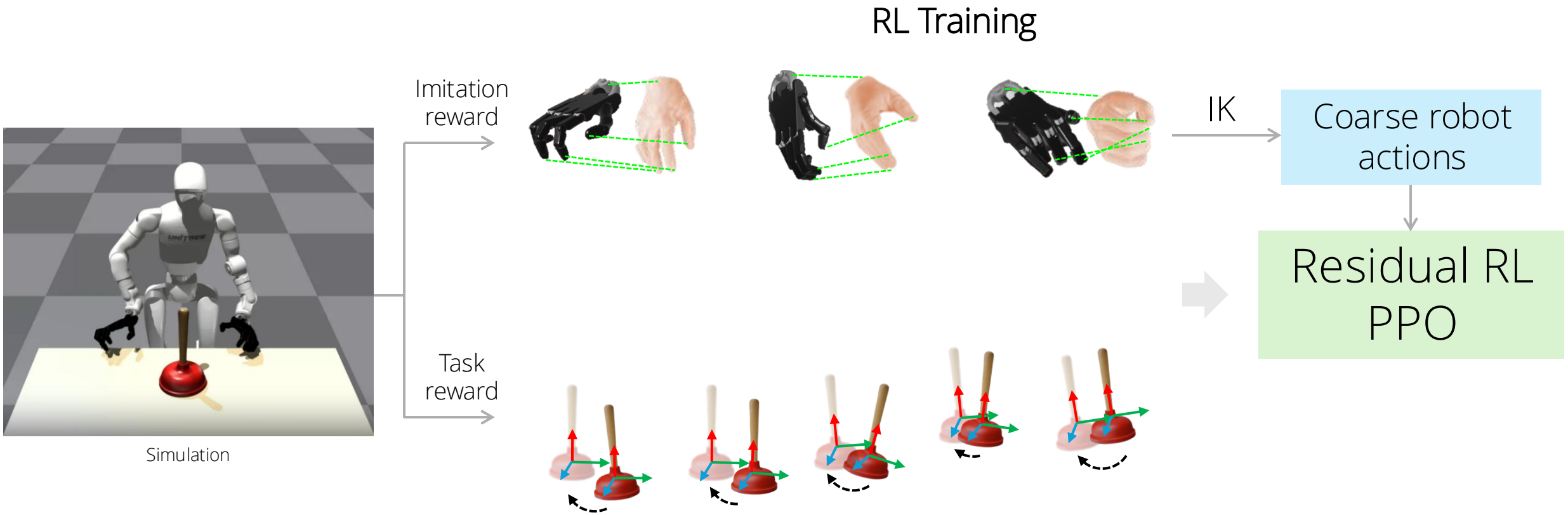


Issue: No Humanoid Robot Action Annotations

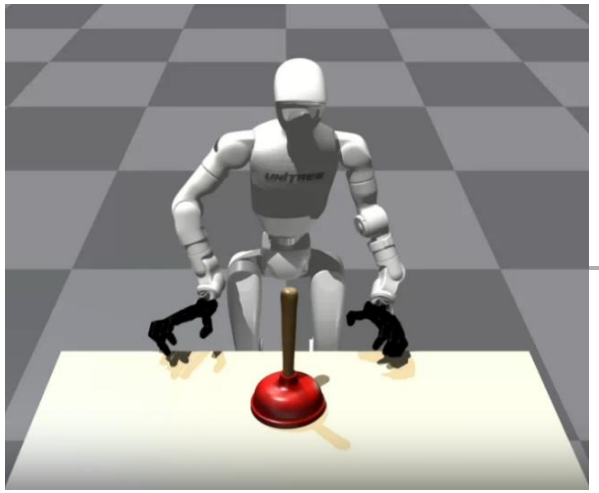




Generate Humanoid Robot Actions with Residual RL



Generate Humanoid Robot Actions with Residual RL

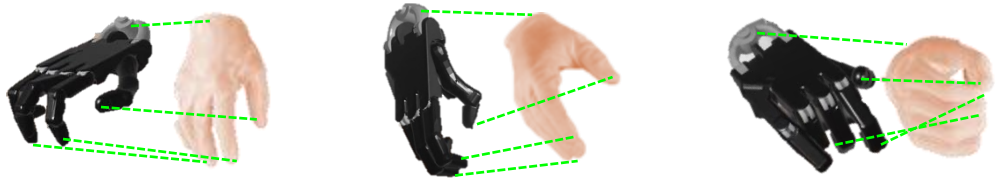


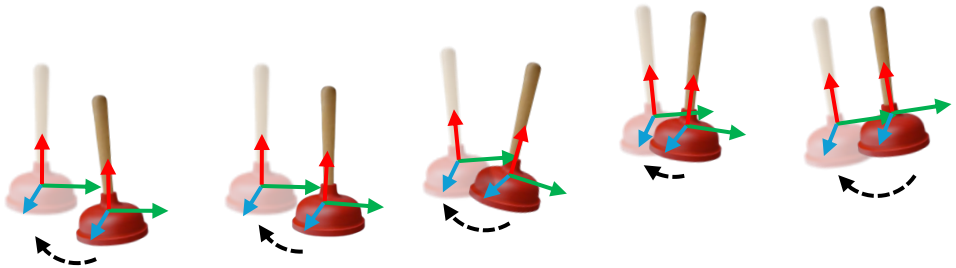
Simulation

Imitation reward

Task reward

RL Training





IK

Coarse robot actions

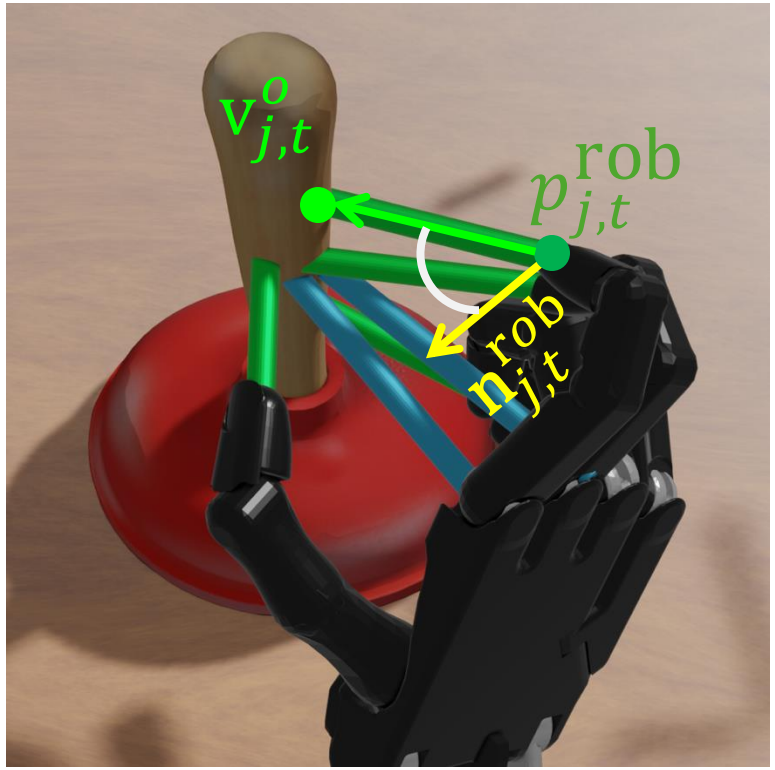
Residual RL PPO

RL is hard...

Can We Leverage More Human Motion Priors?

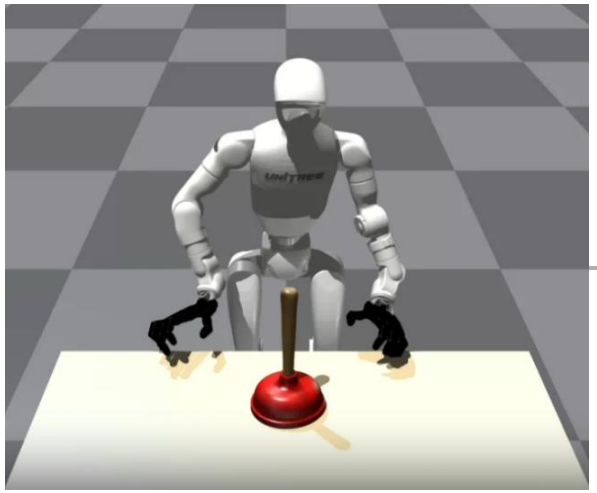
- What other motion priors can we extract from human videos?
- Idea: Contact priors!

Human demonstrations



Contact reward

Involve Physical Feedback as RL Objectives



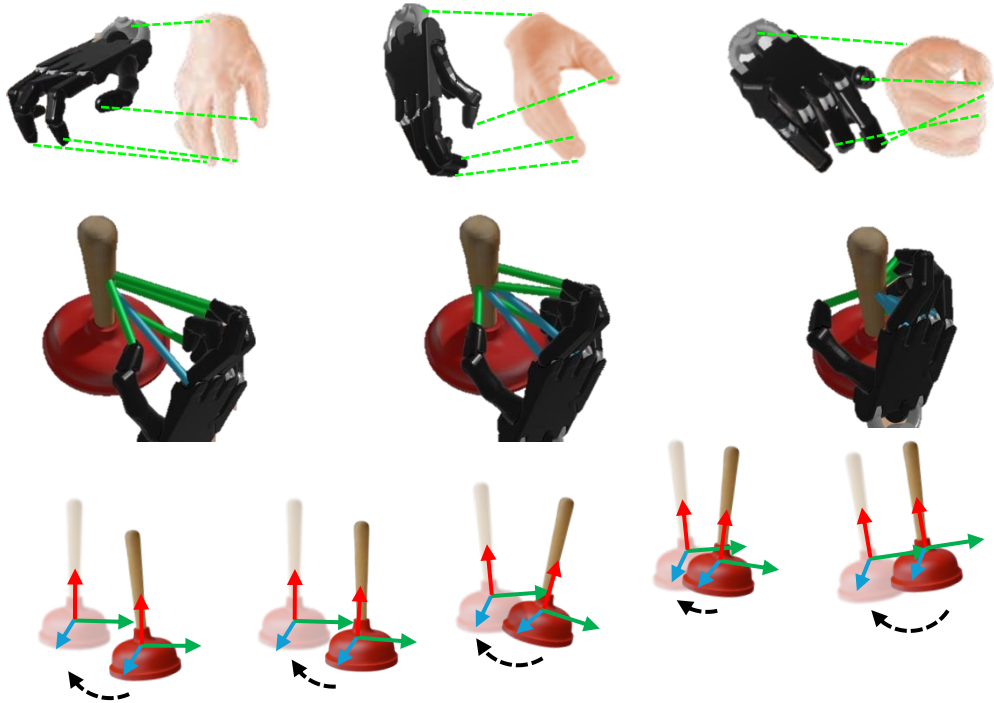
Simulation

Imitation reward

Contact reward

Task reward

RL Training



IK

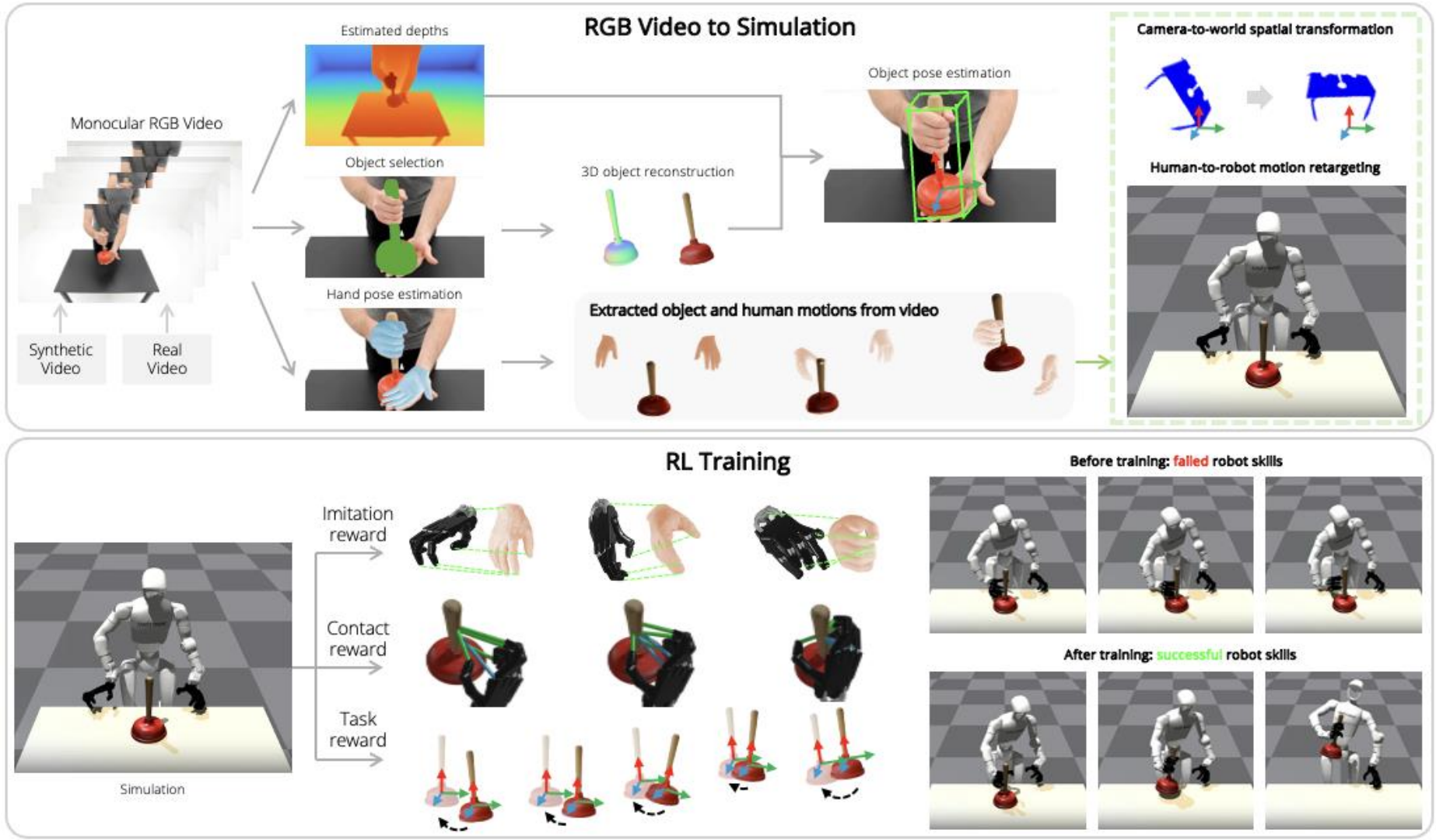
Coarse robot actions

Residual RL PPO

✓

Kind of work...

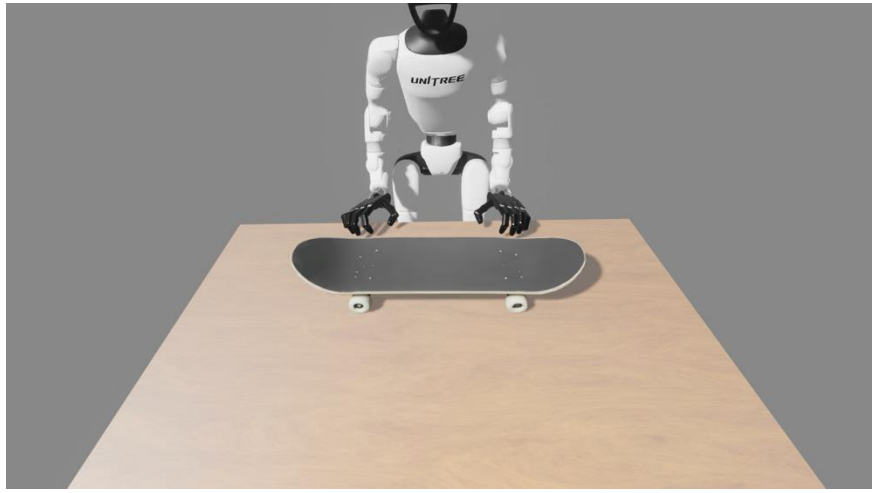
Putting Everything Together



Results: Strong Humanoid Robot Controller for Bimanual Dexterous Manipulation

	Success Rate ↑	$E_r \downarrow$	$E_t \downarrow$
MANIPTRANS	25.3	0.180	0.00646
DexMan (ours)	**44.3**	0.178	0.00688

Results: Automated Video to Robot Action Acquisition Pipeline



Results: Automated Video to Robot Action Acquisition Pipeline