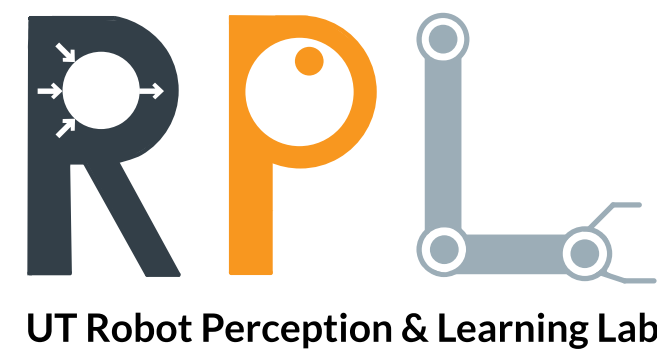


Toward Generalist Humanoid Robots:

Recent Advances, Opportunities, and Challenges

Yuke Zhu

October 31st, 2025



GTC 2024



Why Humanoids?

Why Humanoids Now?

The Humanoid Robot Timeline

“The Robot Dreams”



Rossum's Universal Robots
(1920)

“The Visionary Prototypes”

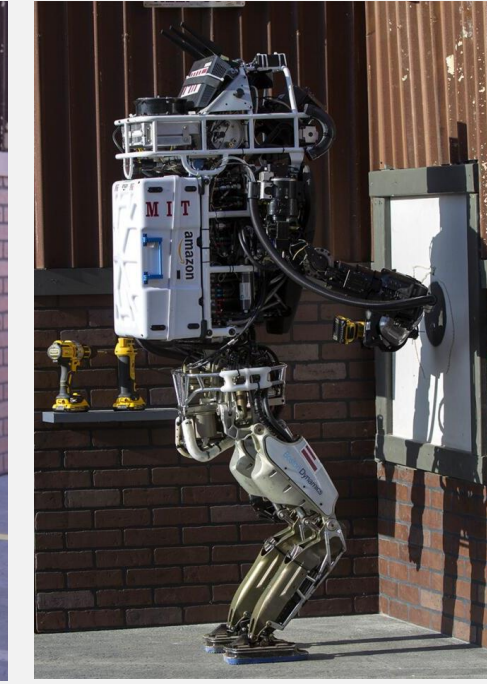


Honda Asimo
(2000)



Sony QRIO
(2003)

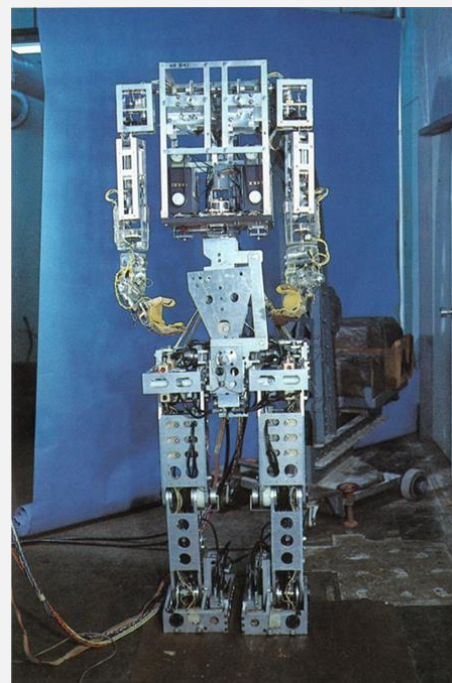
“The Reality Check”



DARPA Robotics Challenge
(2013-2015)



“The Initial Attempts”



WABOT-1
(1970-1973)



WABOT-2
(1980-1984)

“The Social Companions”



NAO
(2008)



Pepper
(2014)

“The Cambrian Explosion”

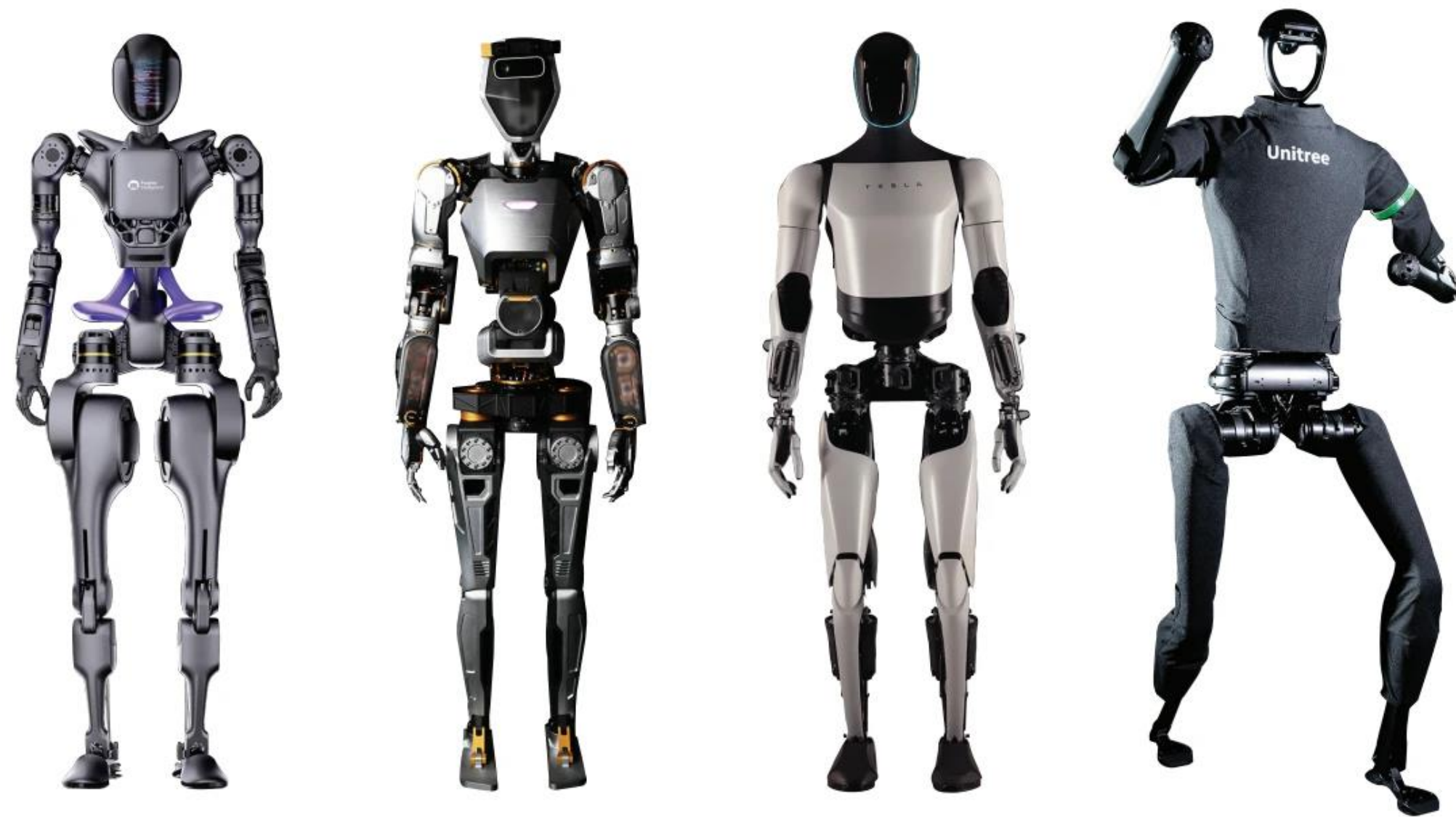


Recent Humanoid Wave
(circa 2022-Now)

Building Generalist Humanoid Robots

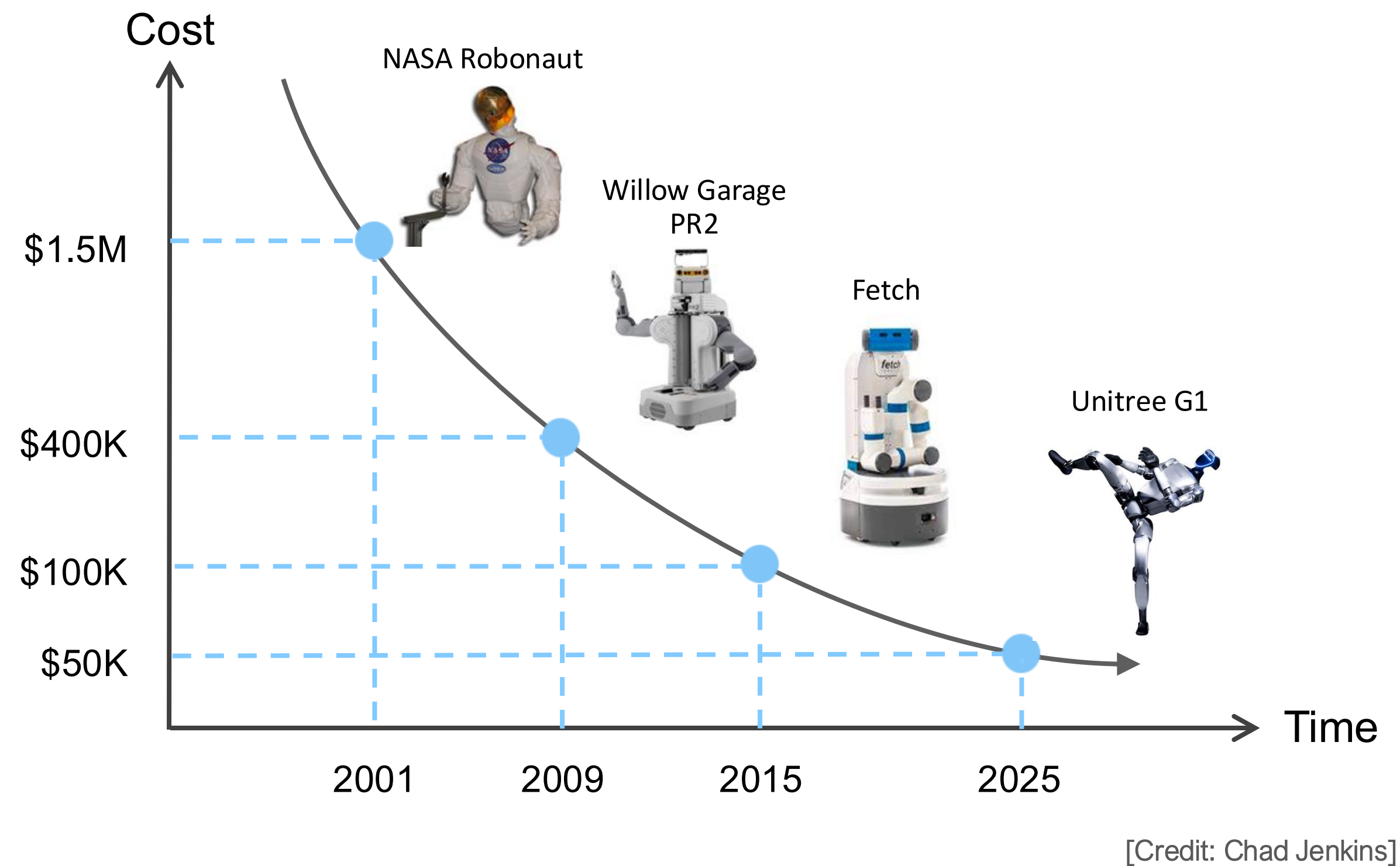
Why Humanoids? A Researcher's Perspective

- ❖ **Versatility:** General-purpose robot autonomy needs a versatile body.



Building Generalist Humanoid Robots

Why Humanoids? A Researcher's Perspective



- ❖ **Versatility:** General-purpose robot autonomy needs a versatile body.
- ❖ **Hardware:** Robot hardware gets cheaper and more robust, democratizing transformative research.



[ASAP, He*, Gao*, Xiao*, Zhang*, et al. 2025]

Building Generalist Humanoid Robots

Why Humanoids? A Researcher's Perspective



- ❖ **Versatility:** General-purpose robot autonomy needs a versatile body.
- ❖ **Hardware:** Robot hardware gets cheaper and more robust, democratizing transformative research.
- ❖ **Brownfield:** Humanoids can seamlessly integrate into the human world infrastructure.

[Source: Future Care Lab in Japan]

Building Generalist Humanoid Robots

Why Humanoids? A Researcher's Perspective



[Source: Getty Images]

- ❖ **Versatility:** General-purpose robot autonomy needs a versatile body.
- ❖ **Hardware:** Robot hardware gets cheaper and more robust, democratizing transformative research.
- ❖ **Brownfield:** Humanoids can seamlessly integrate into the human world infrastructure.
- ❖ **Impact:** Aging workforces, shrinking labor pools, and over \$50 trillion global GDP attributed to labor.

Building Generalist Humanoid Robots

Why Humanoids? A Researcher's Perspective

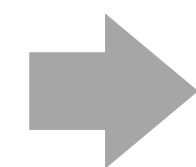
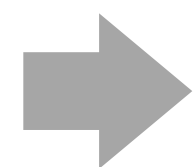


- ❖ **Versatility:** General-purpose robot autonomy needs a versatile body.
- ❖ **Hardware:** Robot hardware gets cheaper and more robust, democratizing transformative research.
- ❖ **Brownfield:** Humanoids can seamlessly integrate into the human world infrastructure.
- ❖ **Impact:** Aging workforces, shrinking labor pools, and over \$50 trillion global GDP attributed to labor.
- ❖ **Data:** Their similar physique unlocks Internet-scale, human-centered data sources.

OKAMI: Learning from Human Videos



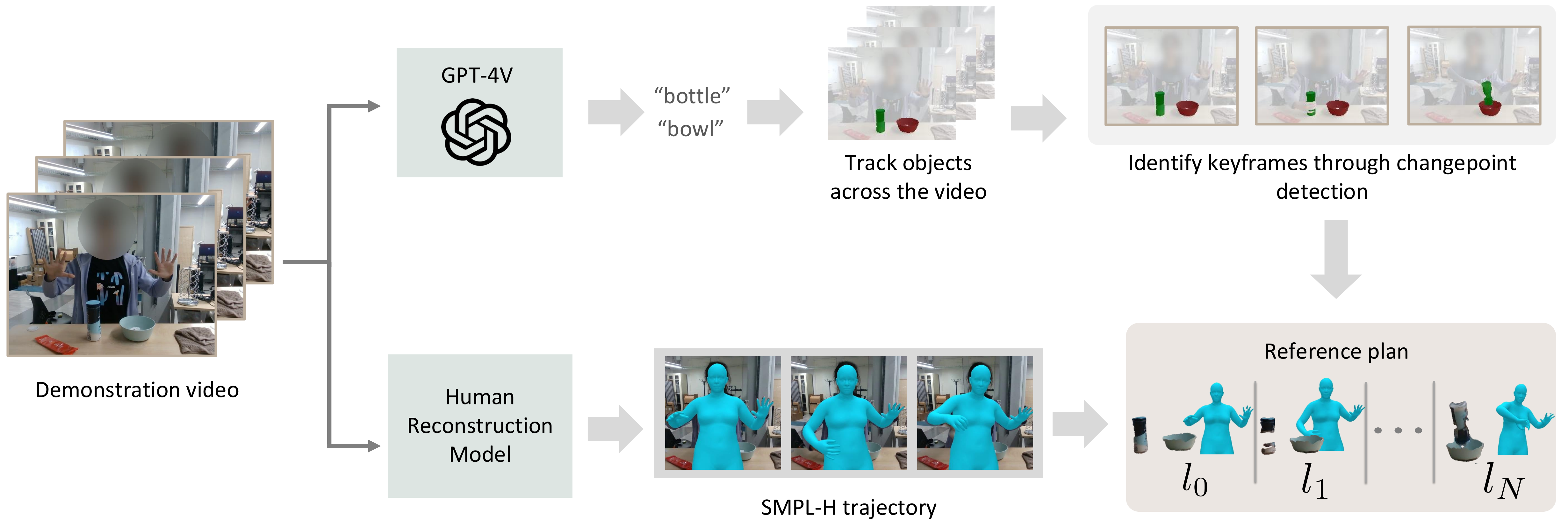
single video
demonstration



trajectory rollouts
in diverse scenes

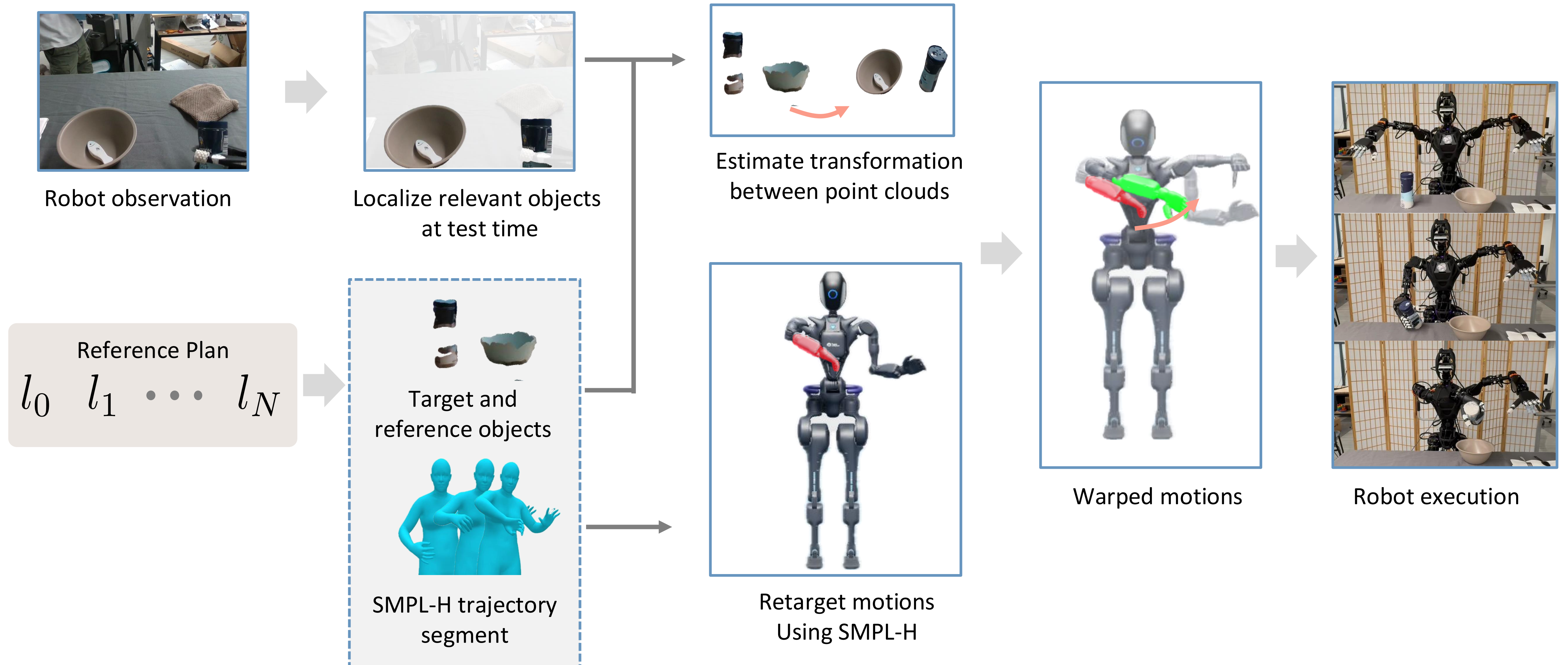
OKAMI: Learning from Human Videos

Stage I: Reference Plan Generation



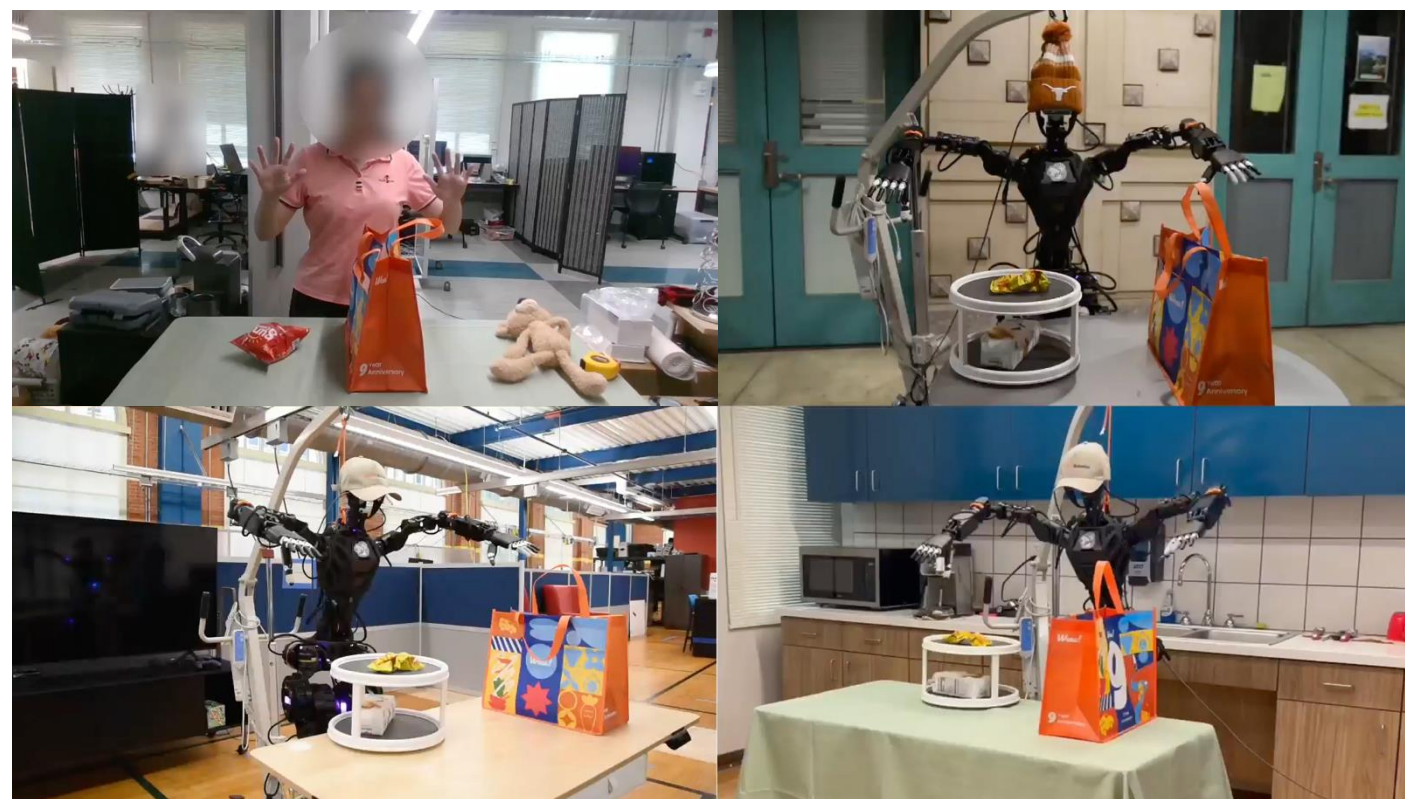
OKAMI: Learning from Human Videos

Stage II: Object-Aware Retargeting

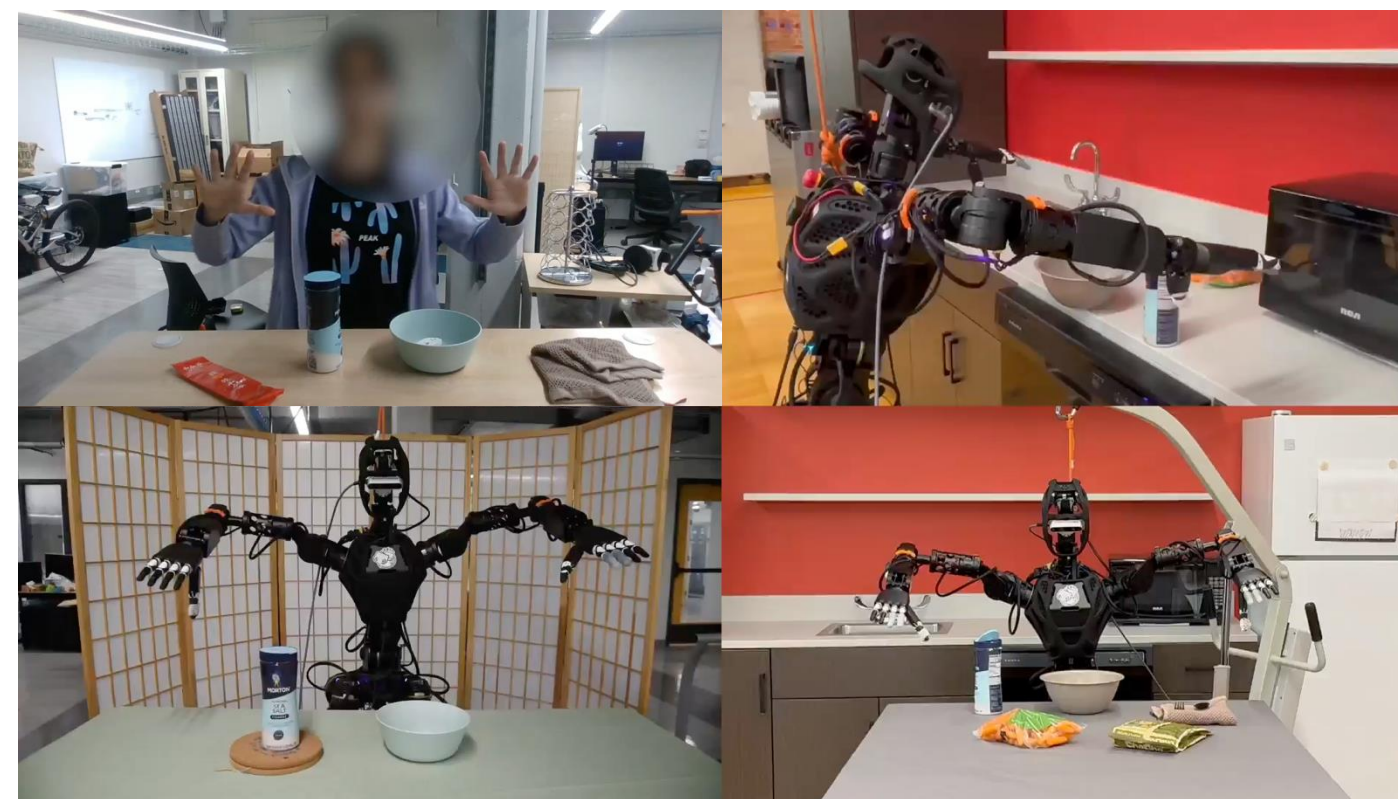


OKAMI: Learning from Human Videos

bagging (58.3%)



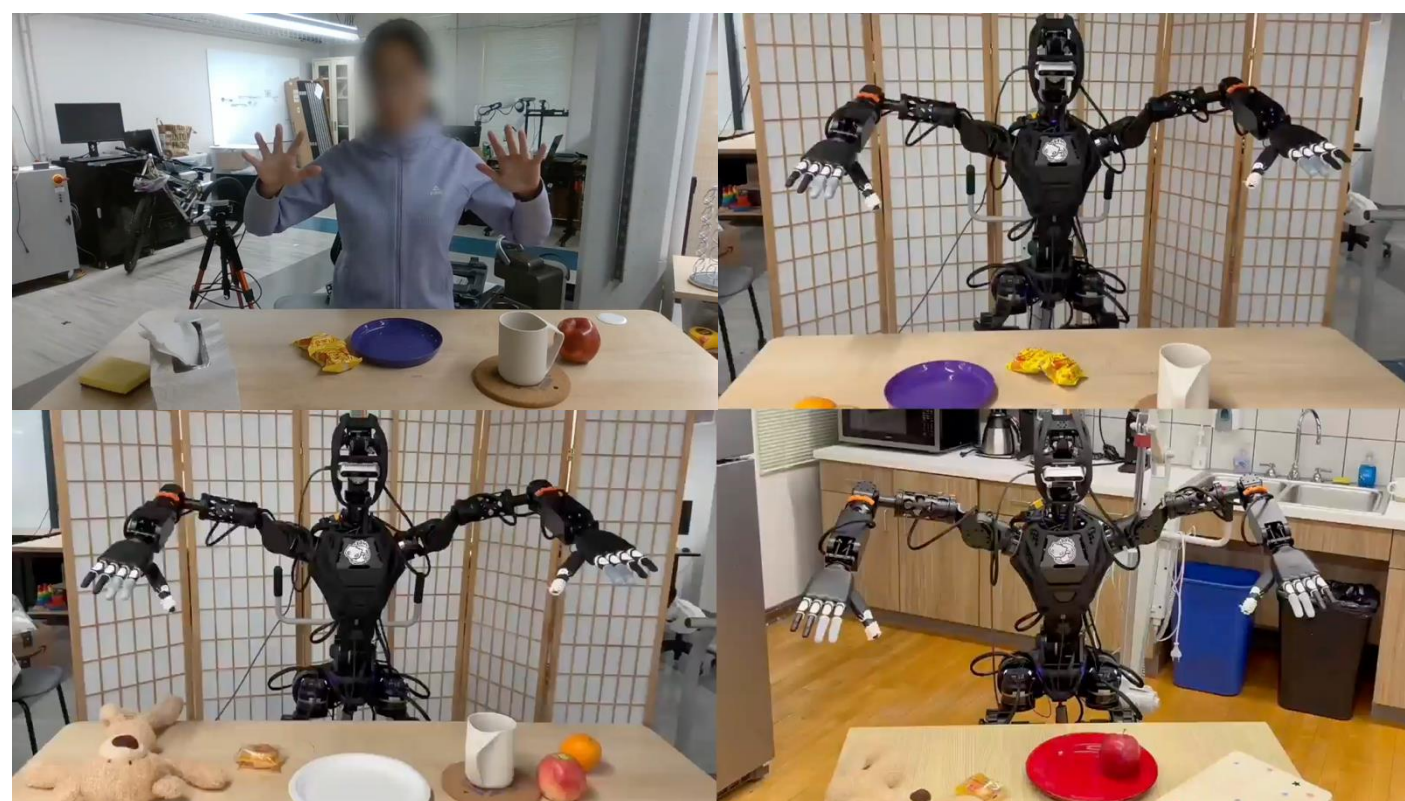
sprinkling salt (58.3%)



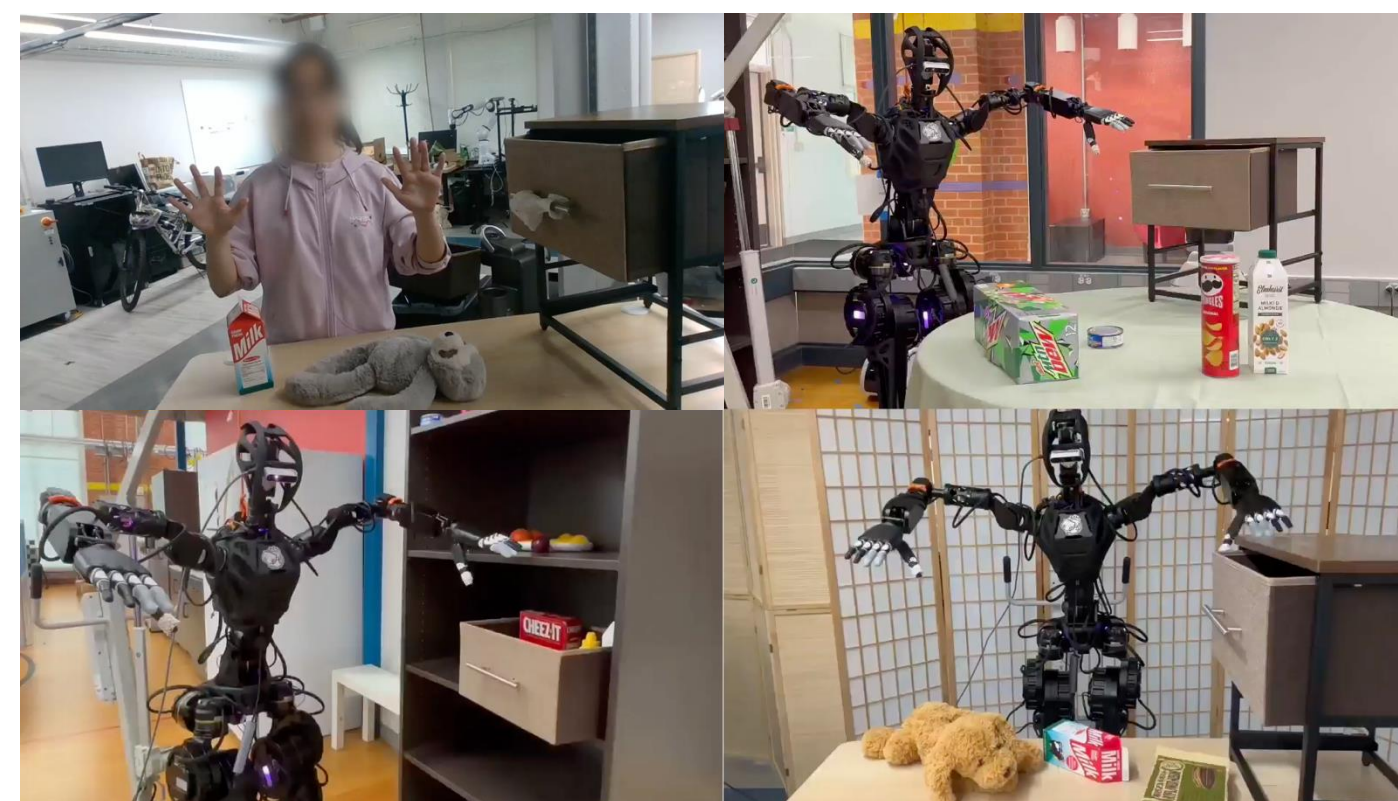
putting toy in basket (66.7%)



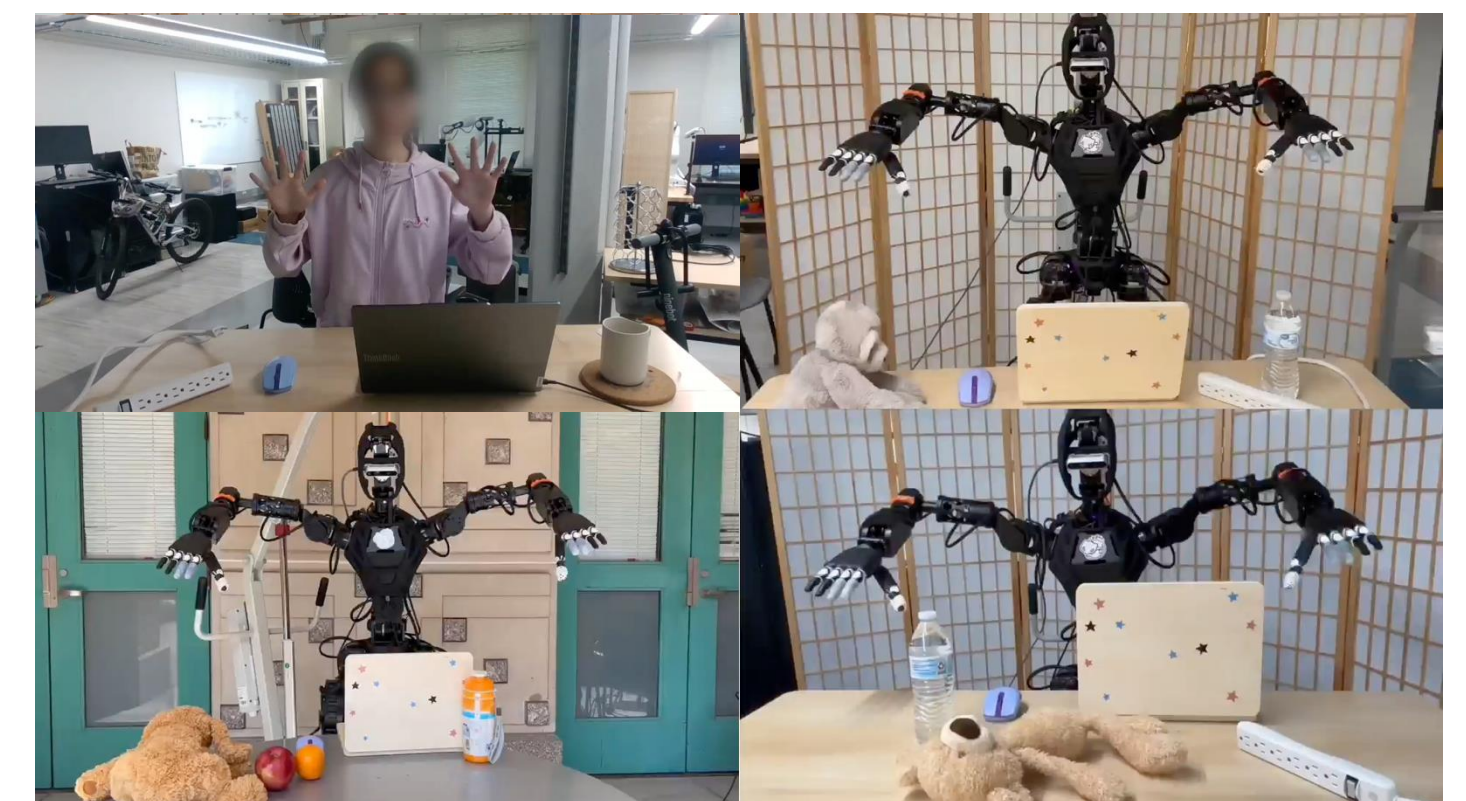
placing snacks on plate (75.0%)



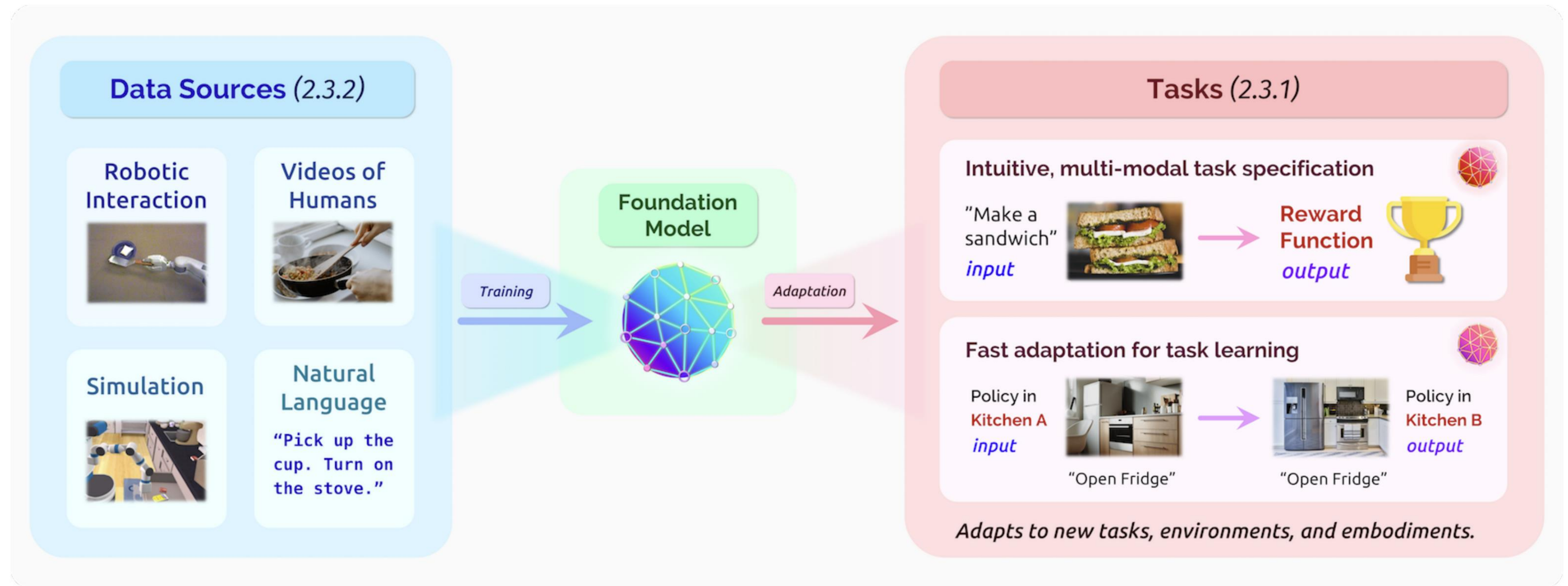
closing the drawer (75.0%)



closing the laptop (83.3%)



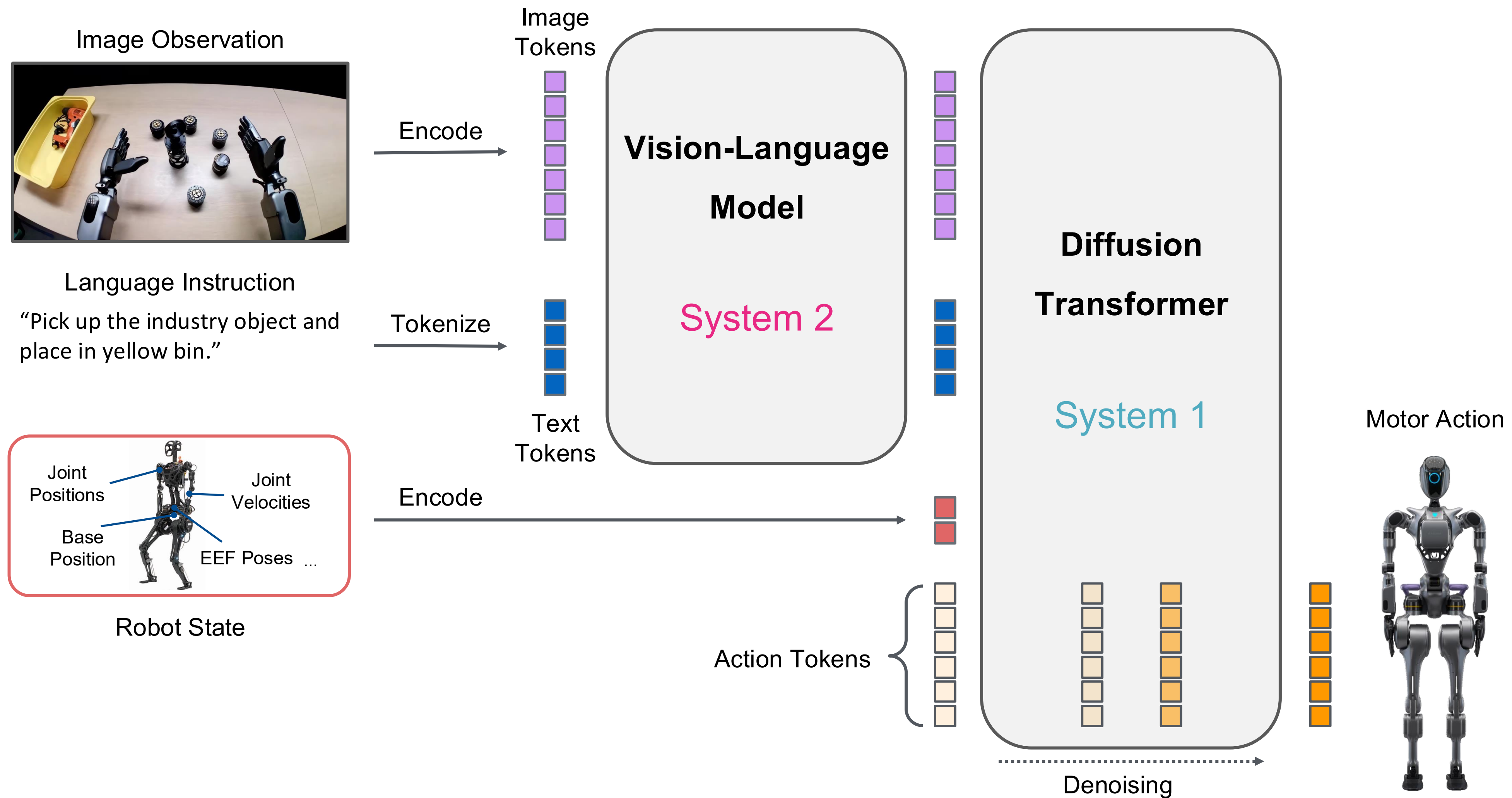
Robotic Foundation Models: The Data-Centric Trend

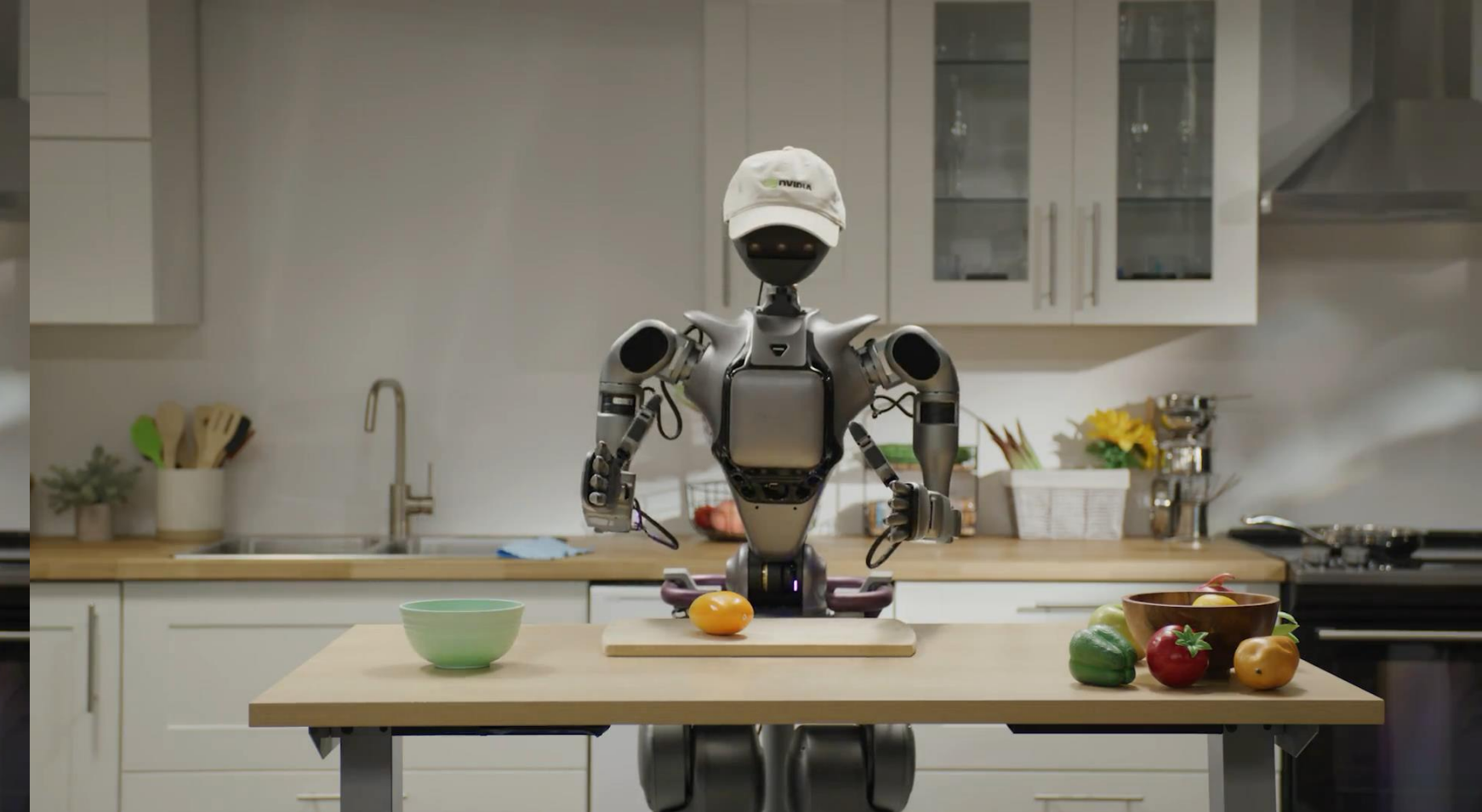


[Source: Bommasani et al. 2021]

1) scale large-scale training data, 2) train generalist transformer-based models, and 3) specialize the generalist model

GR00T N1: Open Humanoid Robot Foundation Model



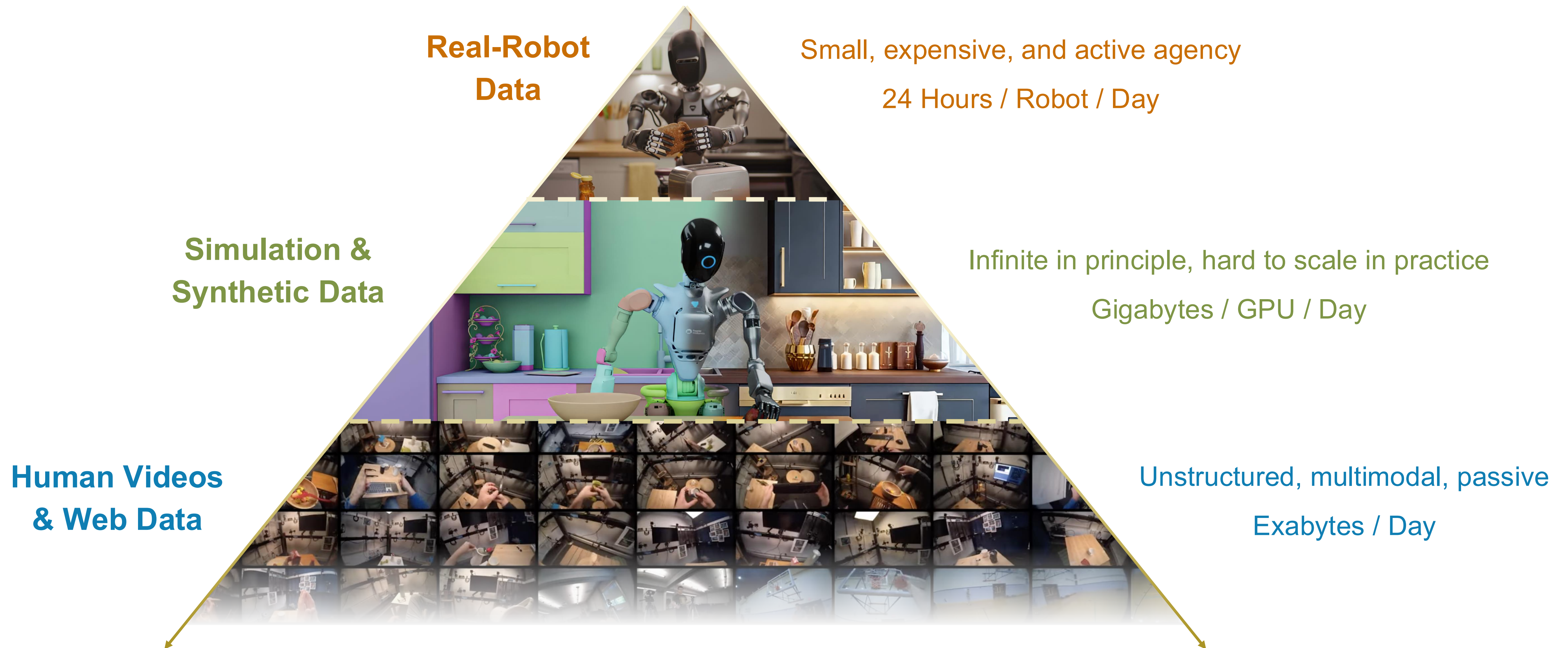






2x

Data Pyramid for Generalist Humanoid Robots



Data Pyramid for Generalist Humanoid Robots



Real-Robot Data

Data grows **linearly** with respect to time, money, human efforts, ...

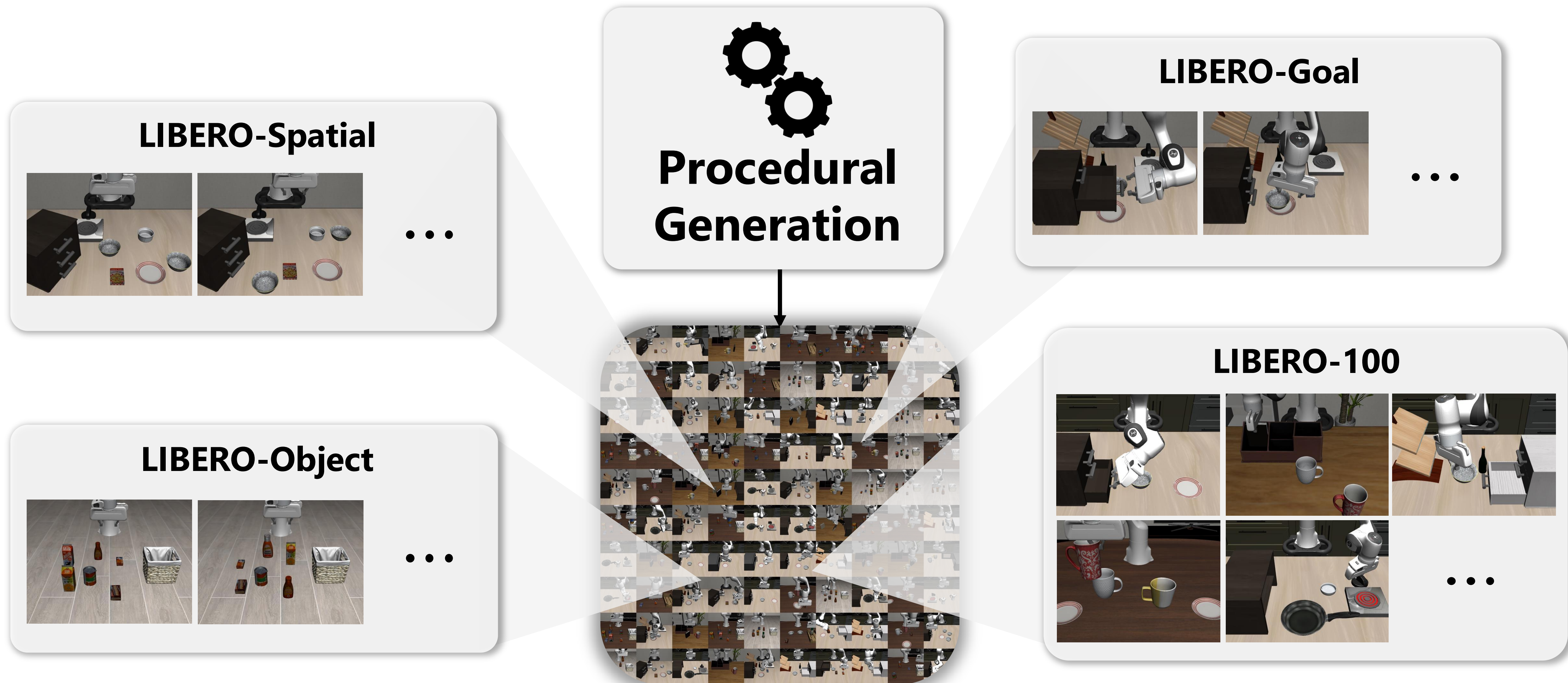
Data Pyramid for Generalist Humanoid Robots



Synthetic Data

Data grows **exponentially** with automated generation in simulation.

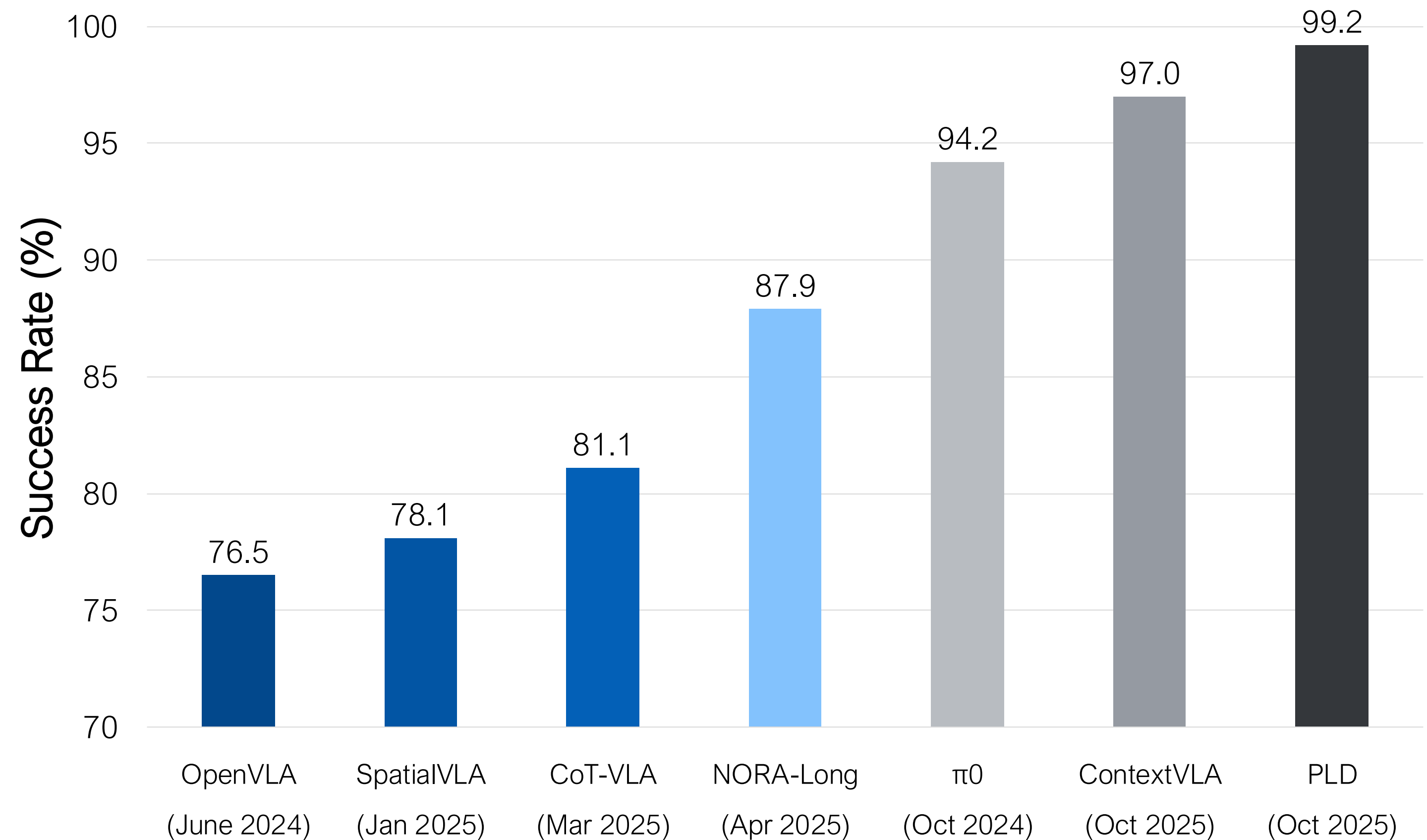
LIBERO: Simulation Benchmark for Lifelong Robot Learning

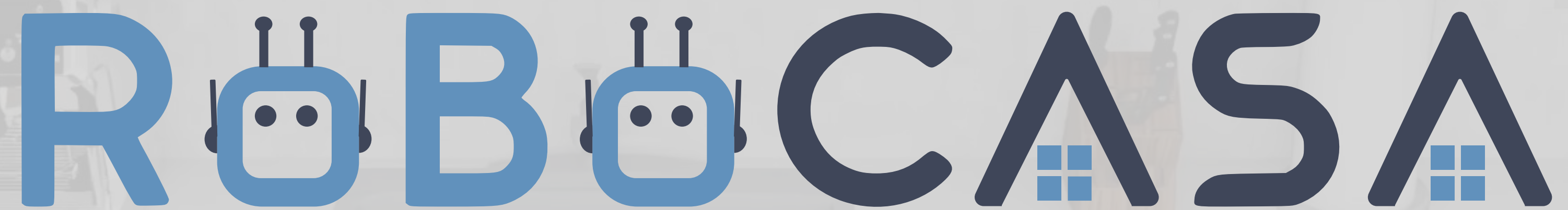


LIBERO: Simulation Benchmark for Lifelong Robot Learning



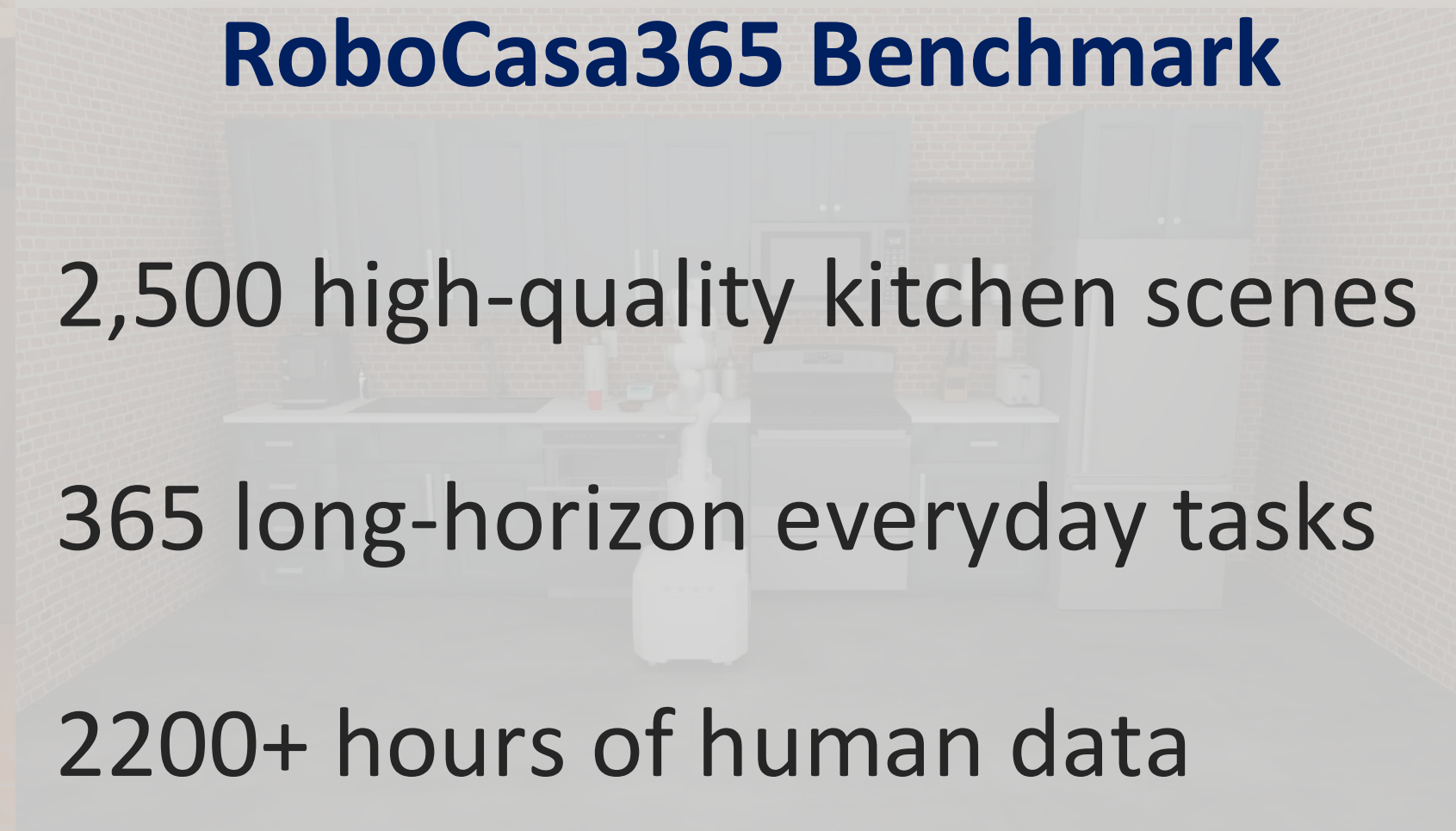
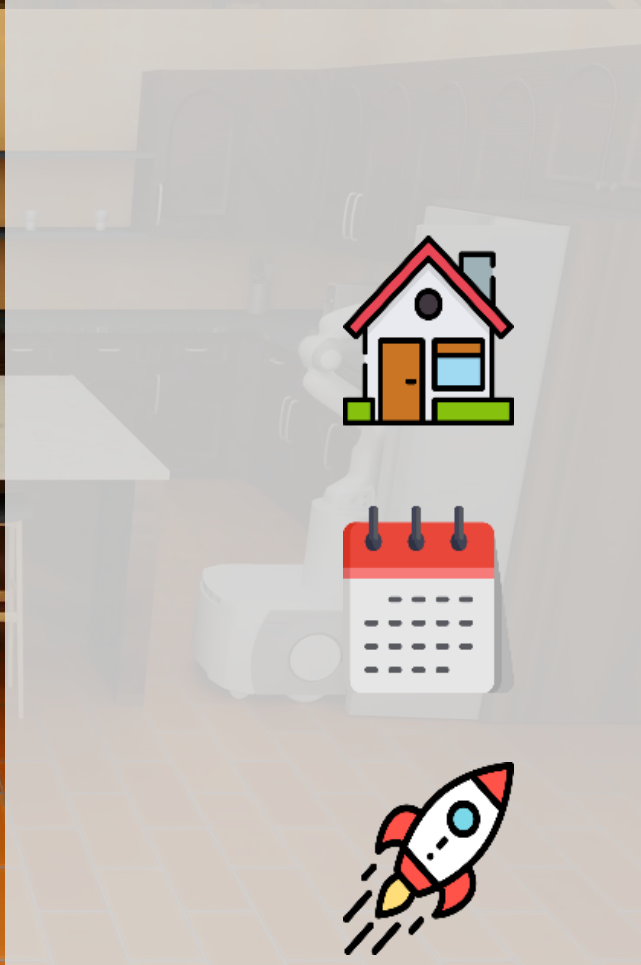
LIBERO Performance with VLA Models



The logo for RoboCasa is displayed in a large, semi-transparent white box. It features the word "RoboCasa" in a stylized font. The "R" and "B" are in a medium blue, while the "o"s are replaced by robot heads with two antennae. The "C" and "A" are in a dark blue, and the "S" and "A" are also in dark blue, with the "A"s featuring a small blue house icon with four windows.

RoboCasa

Large-Scale Simulation of Everyday Tasks for Generalist Robots



RoboCasa365 Benchmark

2,500 high-quality kitchen scenes

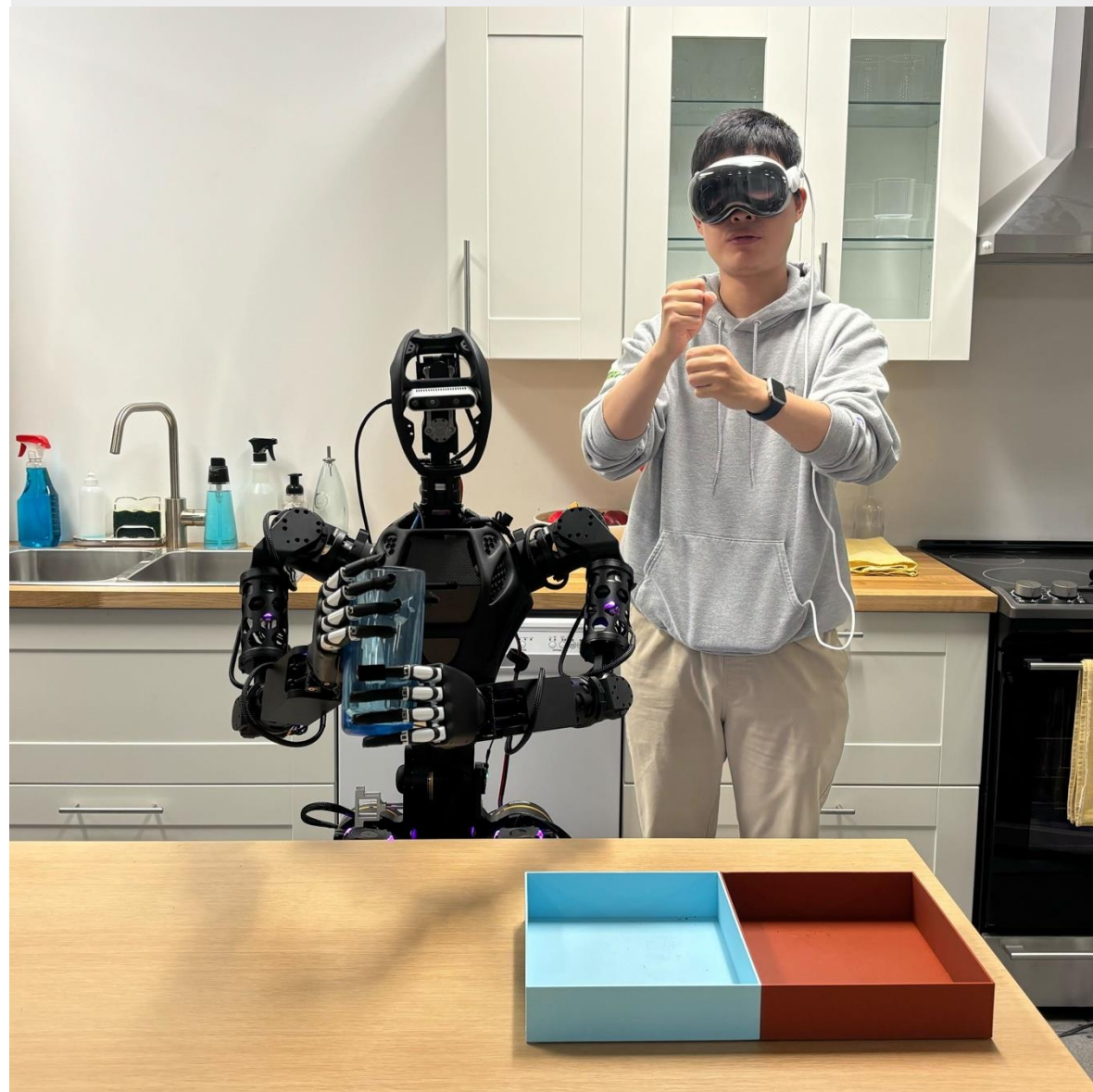
365 long-horizon everyday tasks

2200+ hours of human data

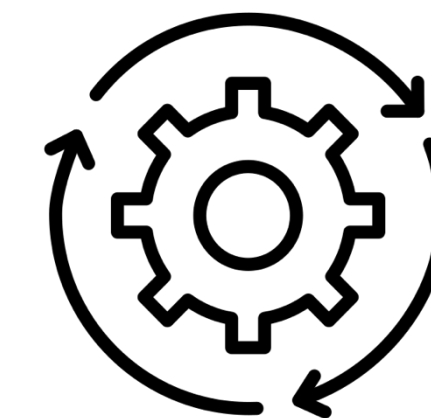
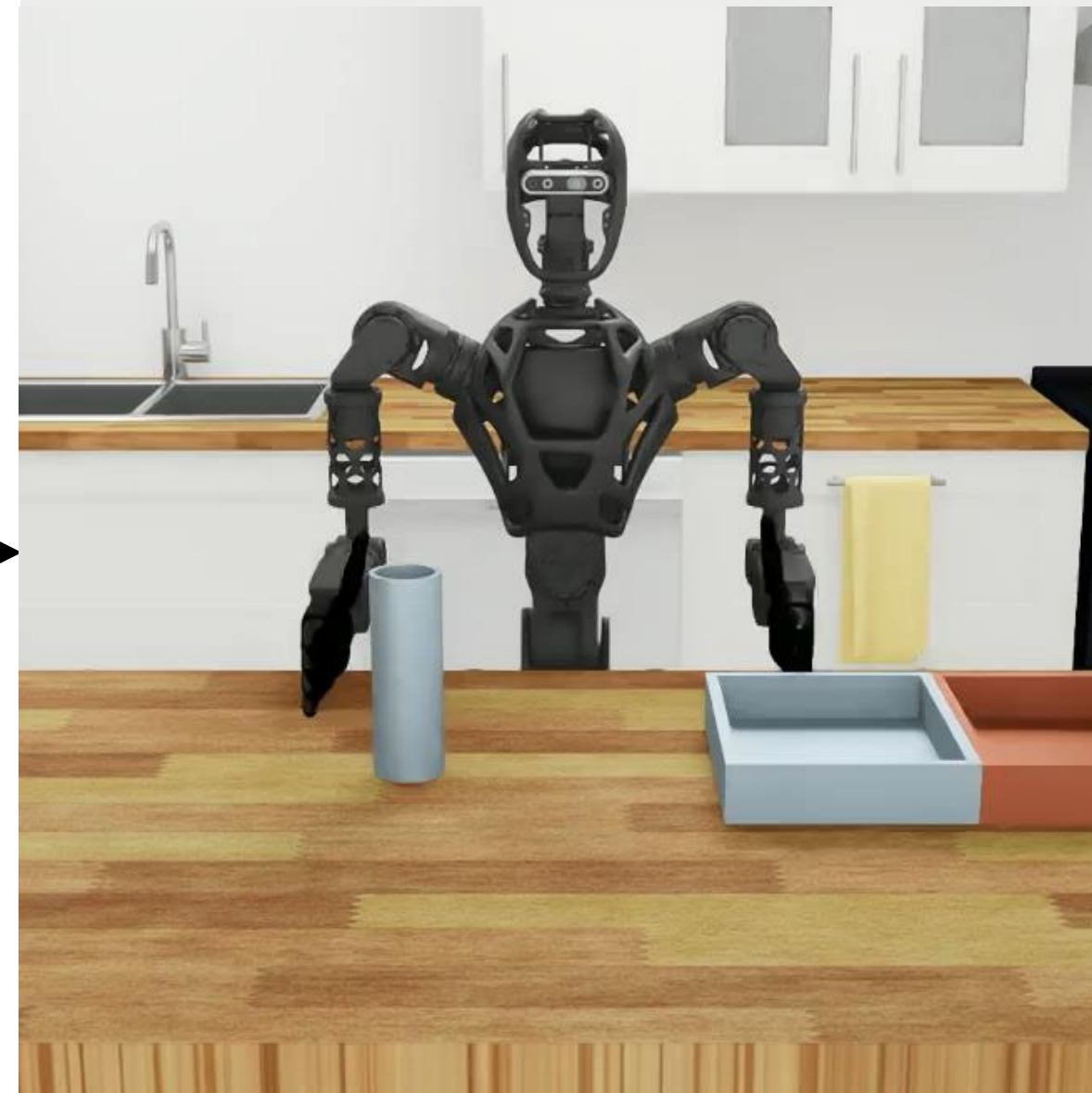


DexMimicGen: Automated Data Generation System

Human
teleoperation

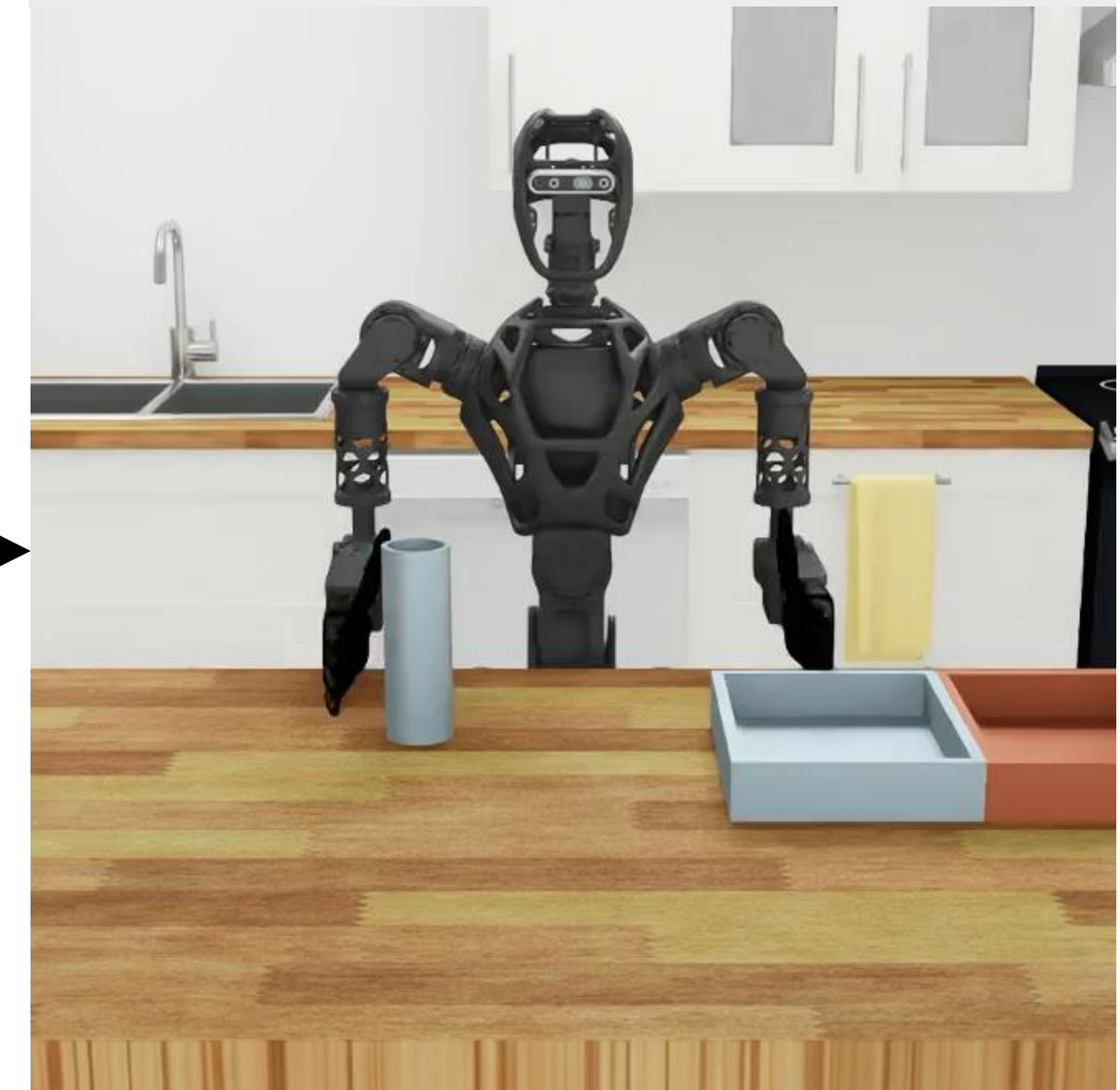


5 source
demos

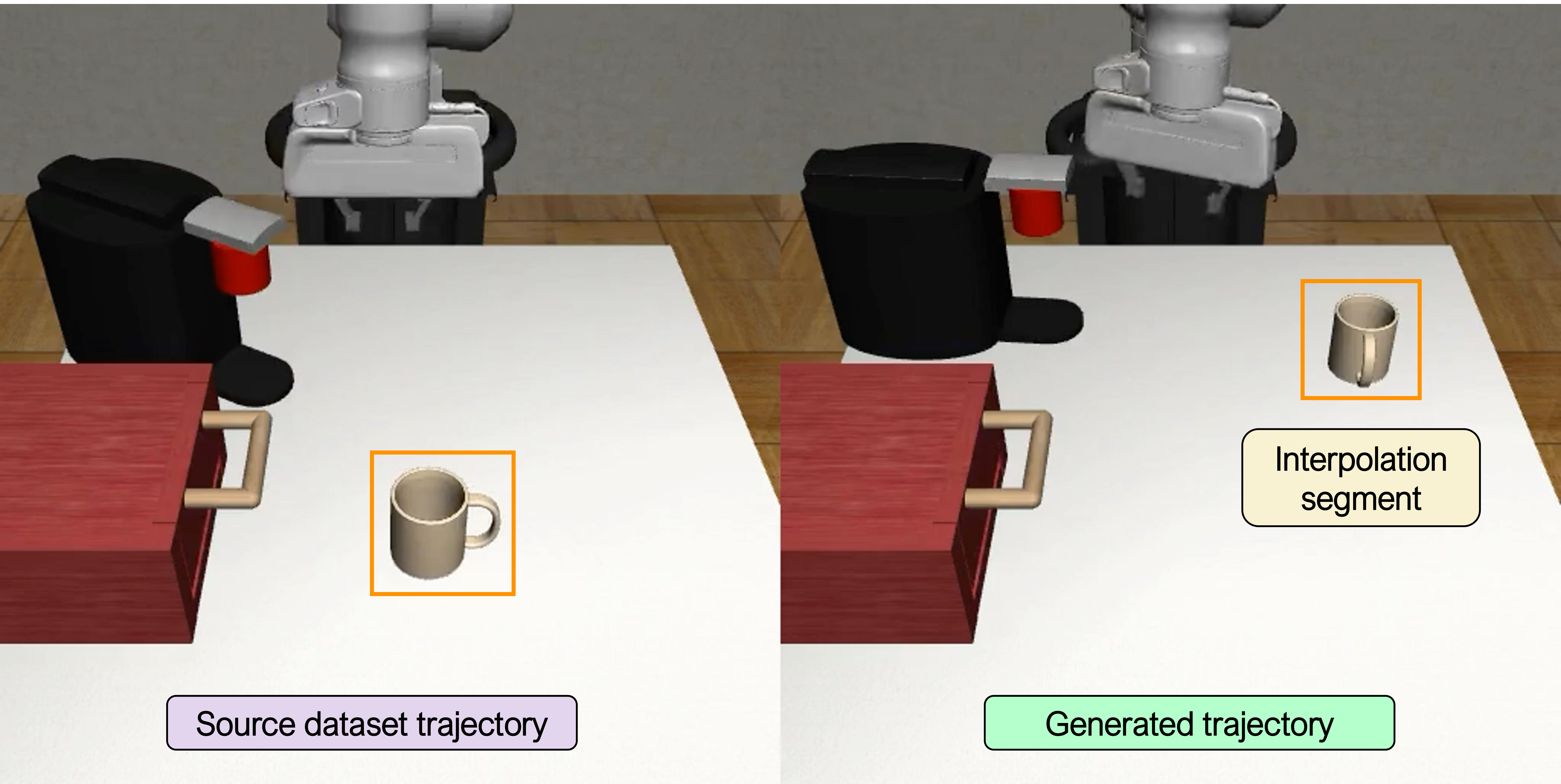


DexMimicGen

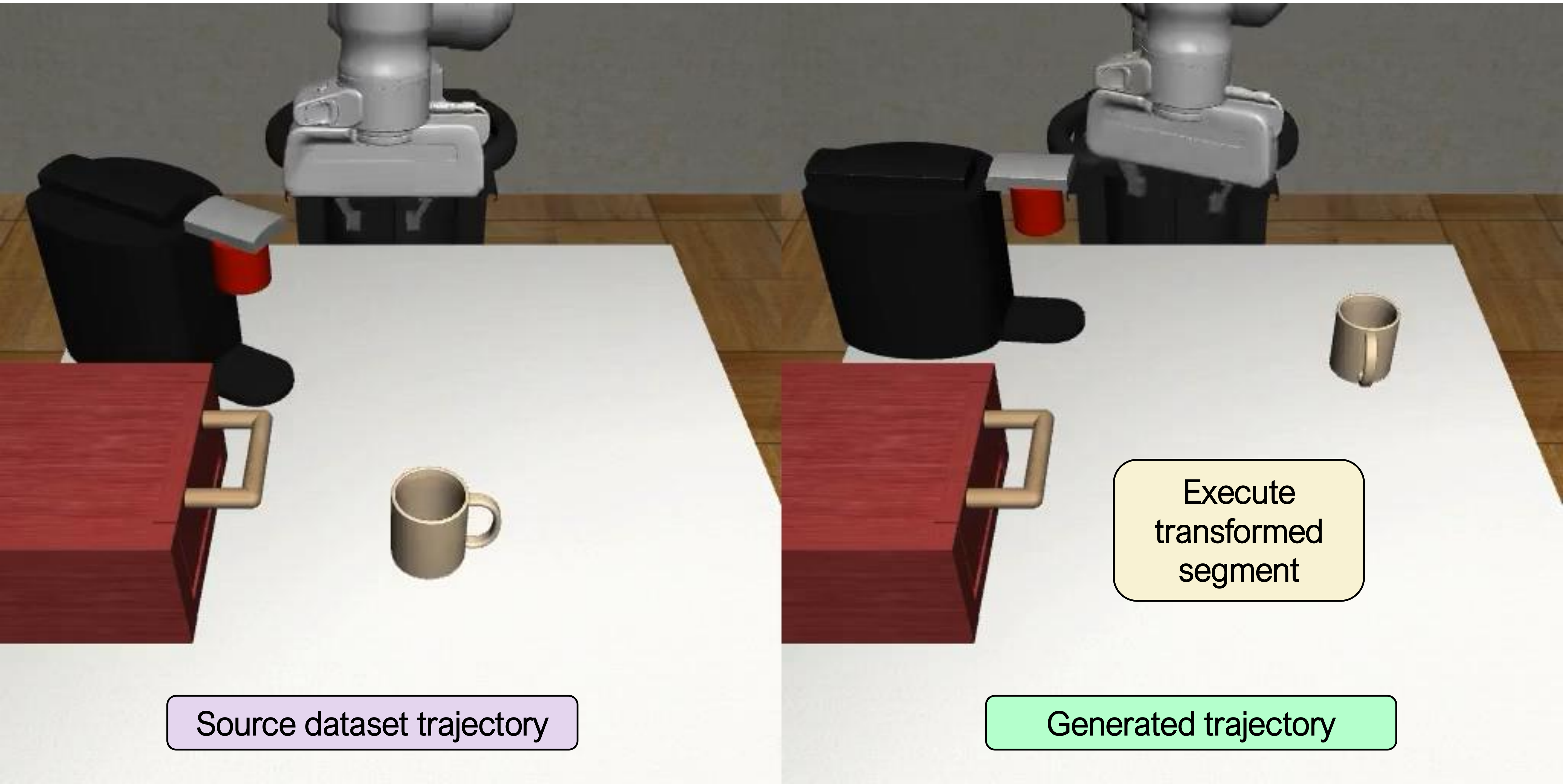
1000 generated demos



DexMimicGen: Data Generation Example



DexMimicGen: Data Generation Example



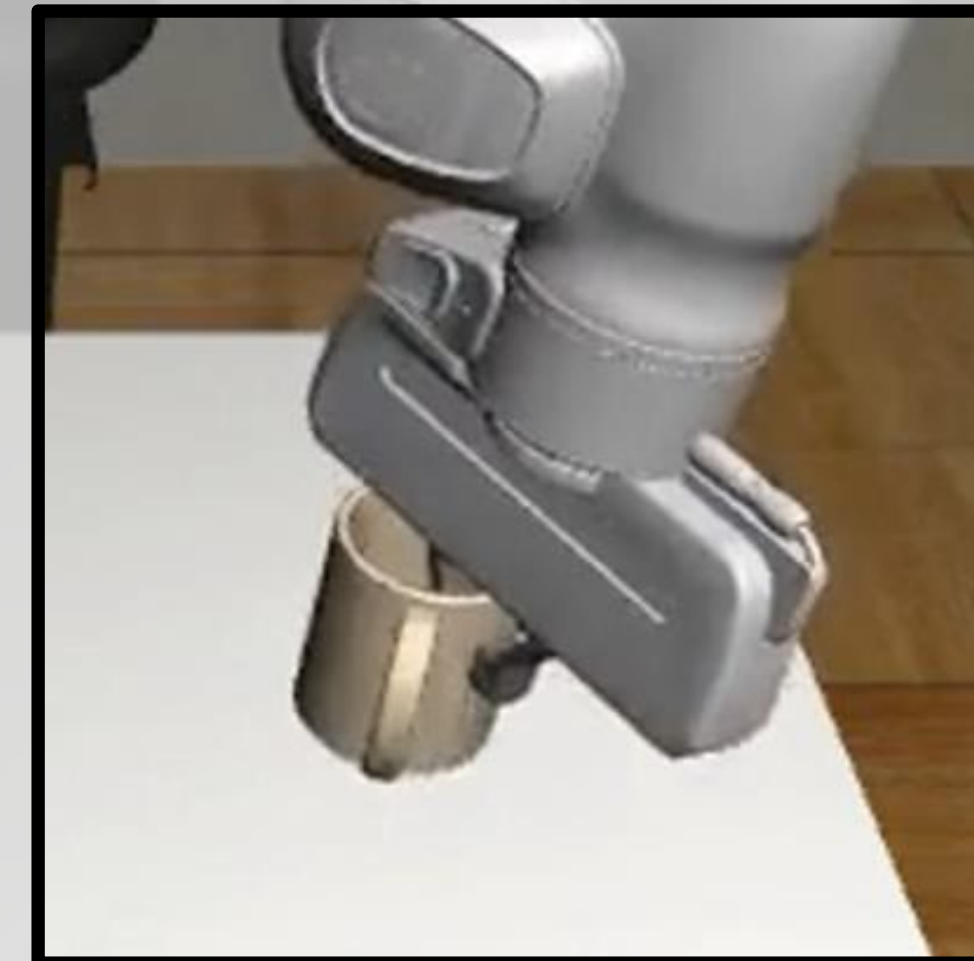
DexMimicGen: Data Generation Example

Functional equivariance: Preserving the relative poses of **robot end-effector** and **object** under SE(3) transformations



Source dataset trajectory

Mug grasp is consistent!



Generated trajectory

DexMimicGen: Automated Data Generation System

Parallel
Subtasks



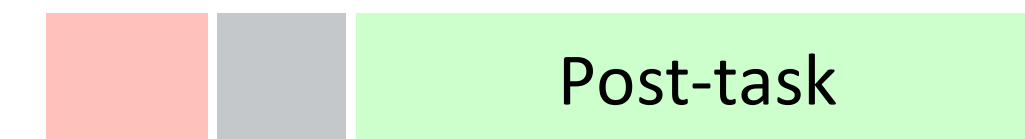
Coordination
Subtasks



Sequential
Subtasks



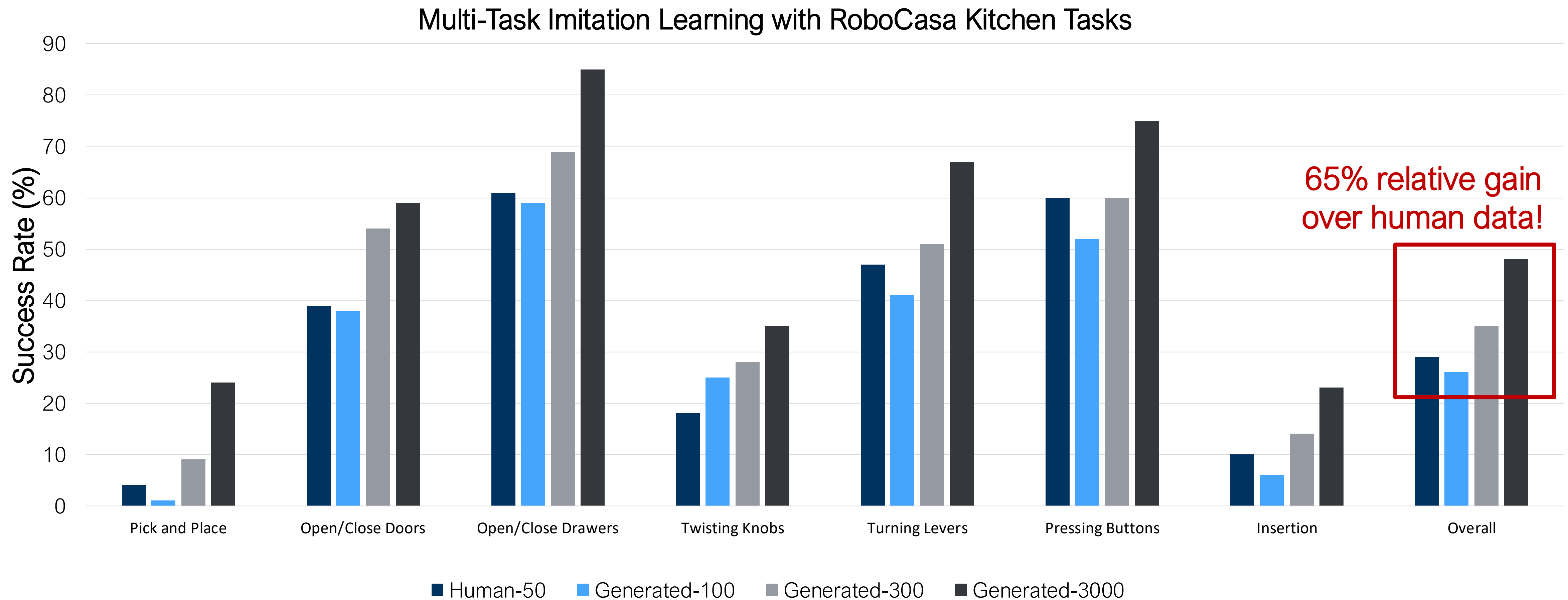
Right hand

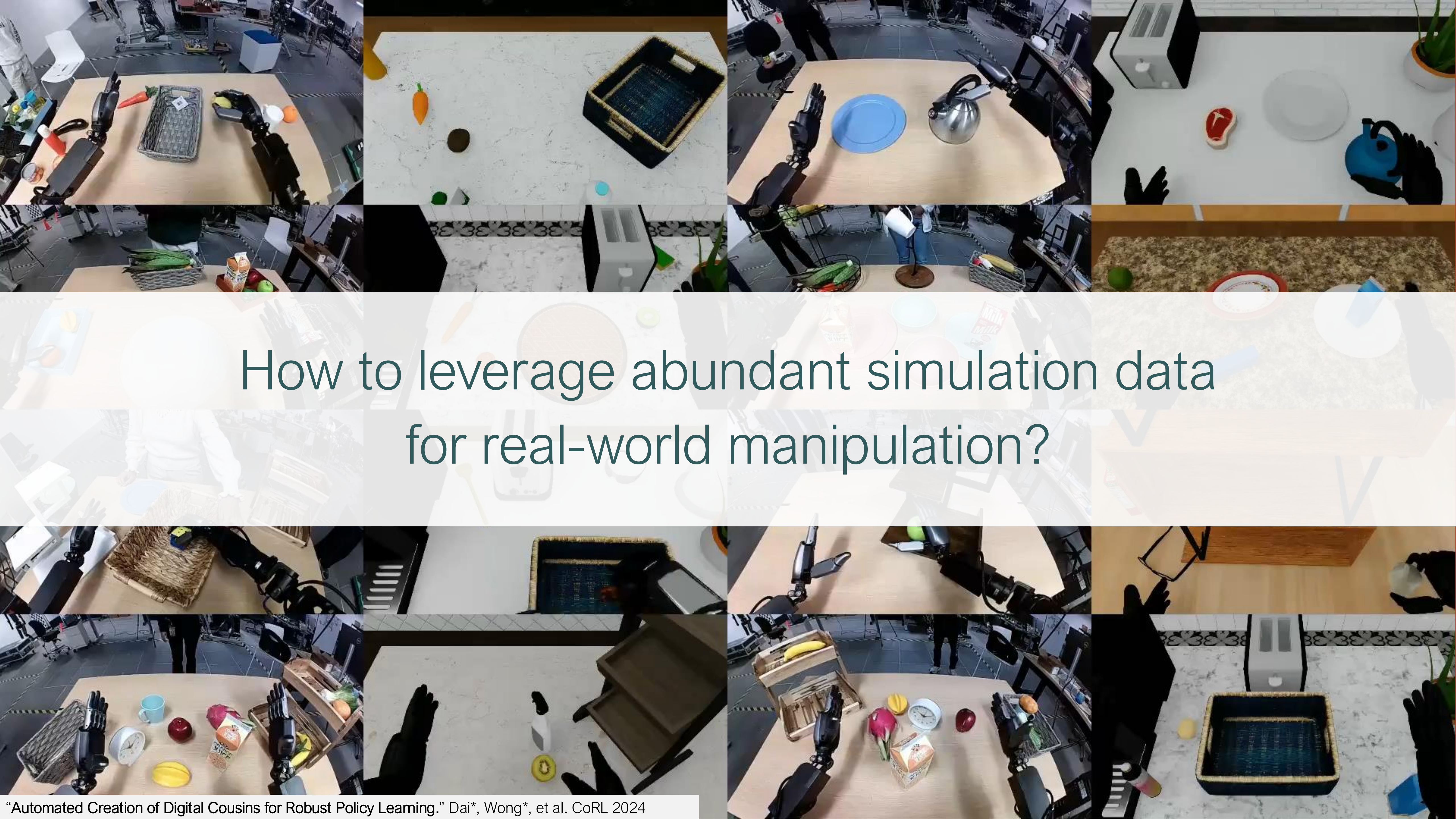


Left hand



DexMimicGen: Automated Data Generation System





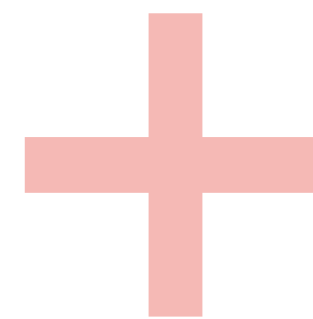
How to leverage abundant simulation data for real-world manipulation?

Sim-and-Real Co-Training: Learning from Hybrid Datasets

Large-scale sim data

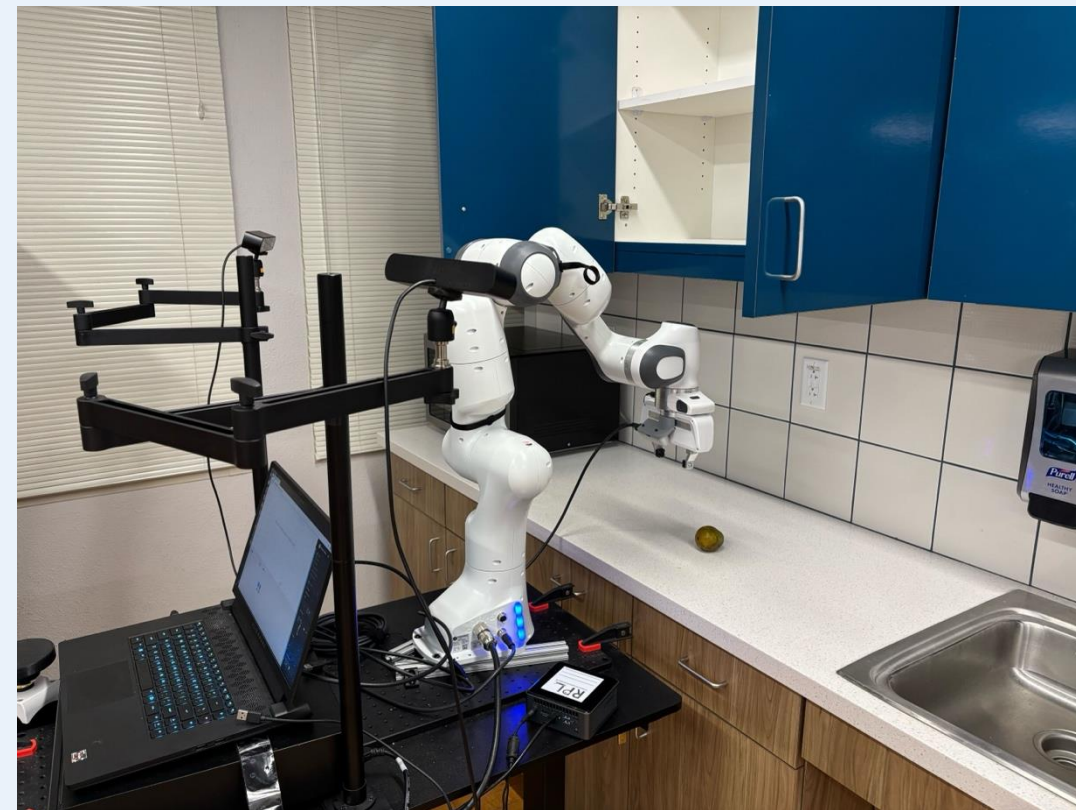


Small-scale real data



Co-training Pipeline

① Task Setup

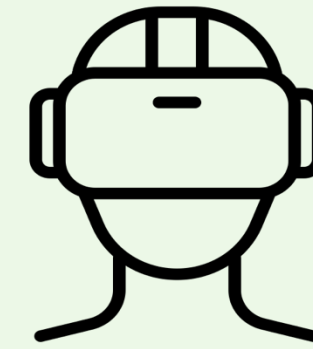


Real-world task

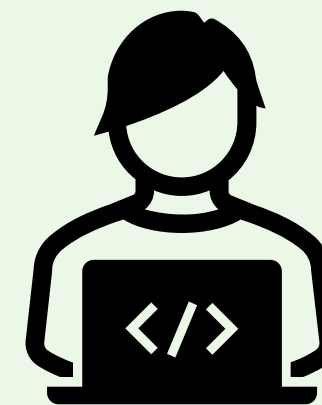


Prior sim datasets

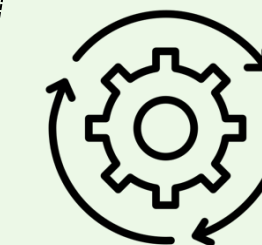
② Data Preparation



Collect dozens
of real-world demos



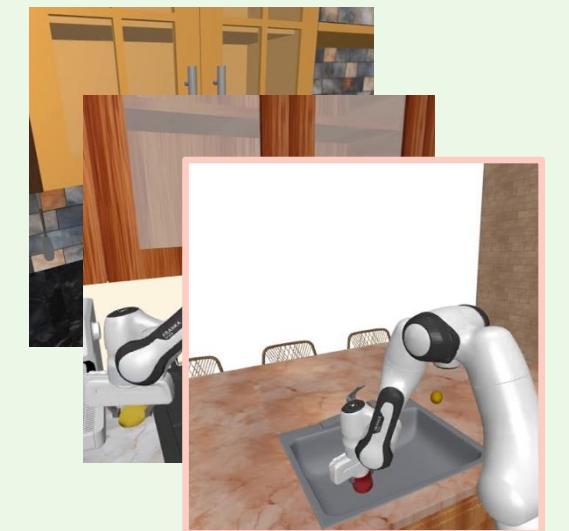
Create digital
cousin task



Collect dozens of sim demos,
100x via DexMimicGen



Real-world data (1x)



Digital cousin sim
data (100x)



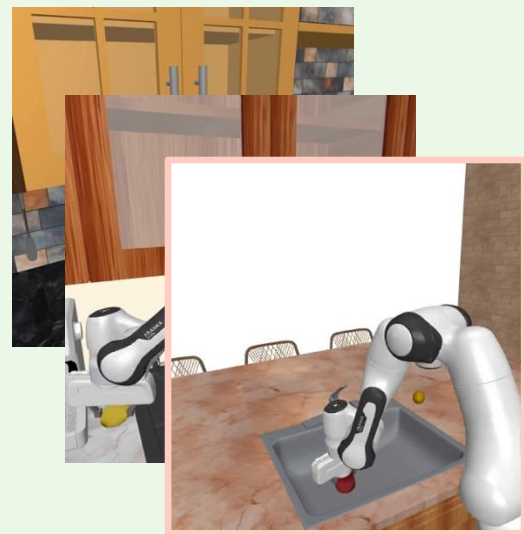
Prior sim data (1000x)

Co-training Pipeline

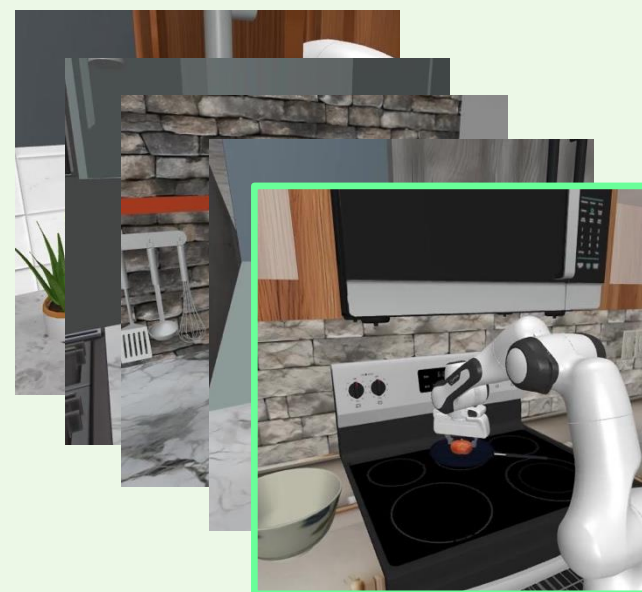
② Data Preparation



Real-world data (1x)



Digital cousin sim data (100x)



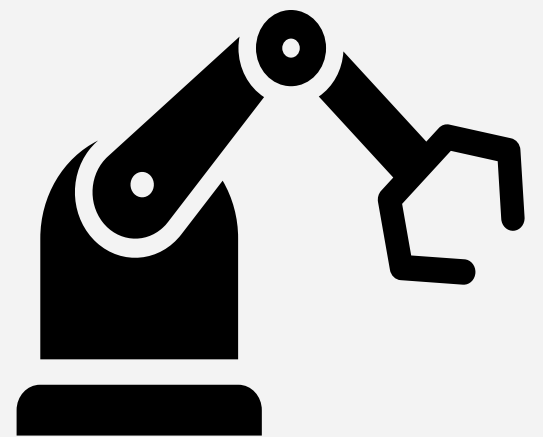
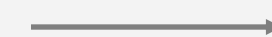
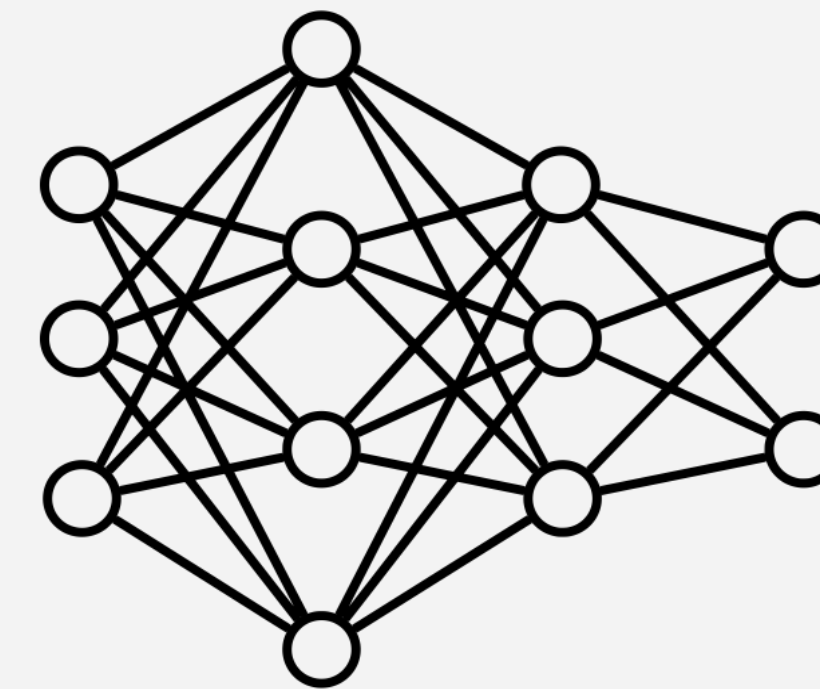
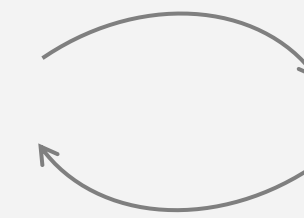
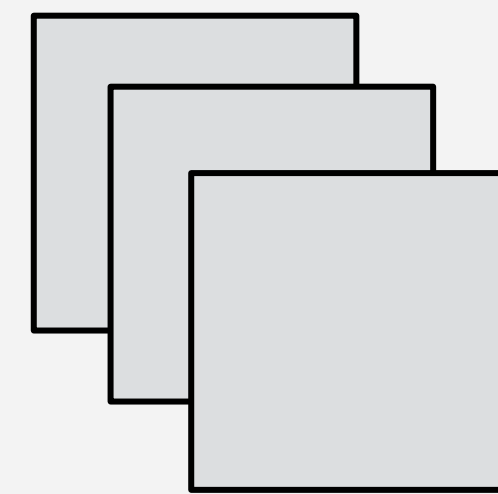
Prior sim data (1000x)

③ Co-Training and Deployment

Batch Sampling

Policy Co-Training

Deployment



Real Data



...



+

Sim Data



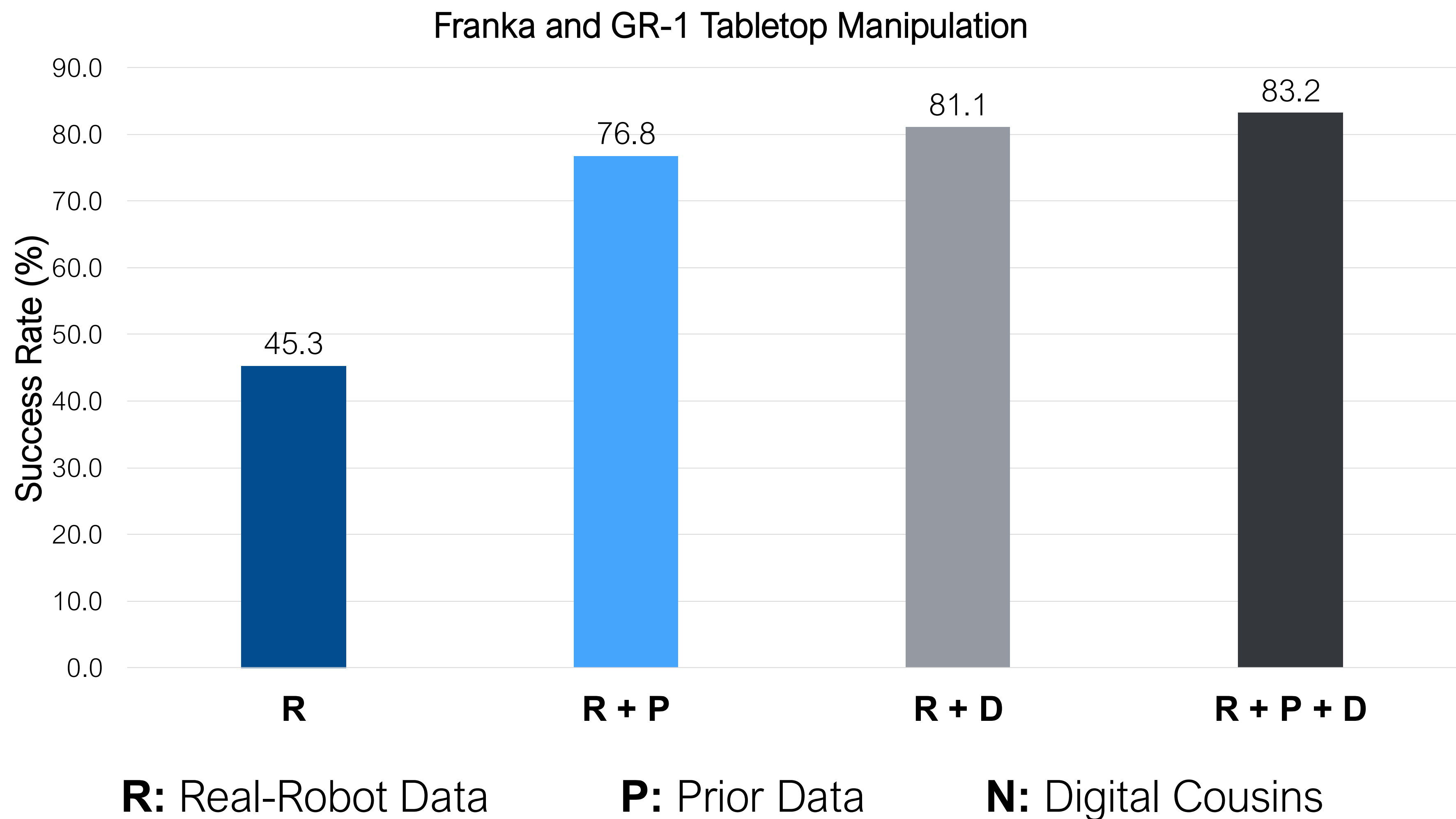
...



$(1 - \alpha) \cdot \text{batch_size}$

$\alpha \cdot \text{batch_size}$

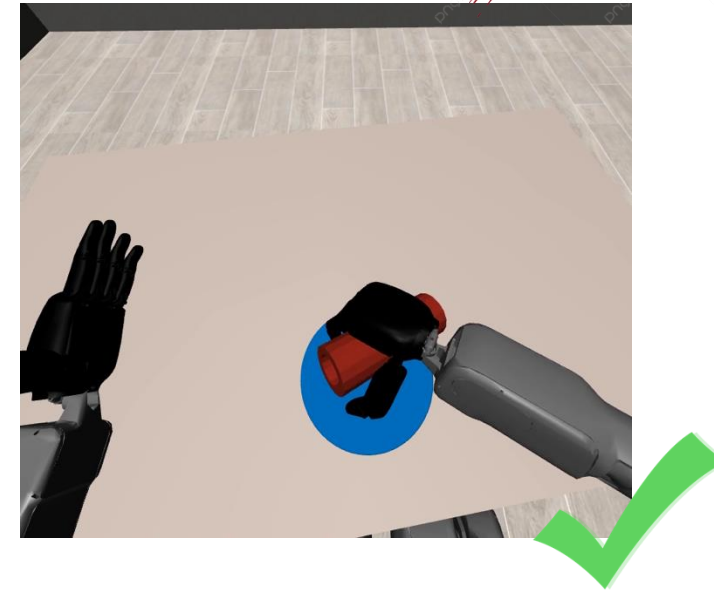
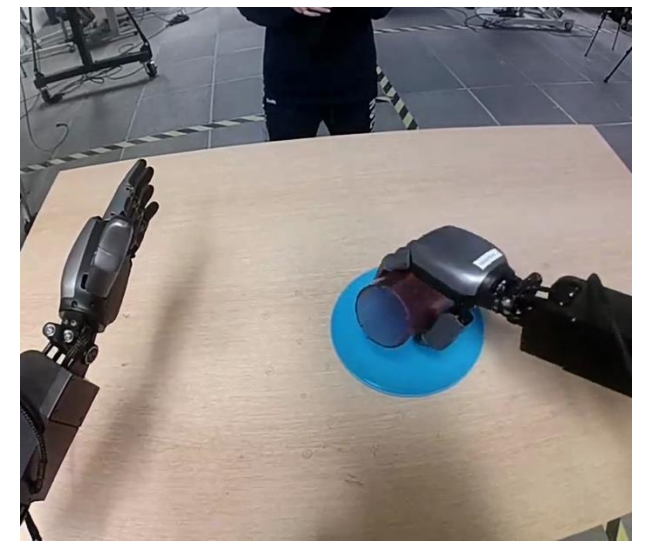
Sim-and-Real Co-Training: Learning from Hybrid Datasets



Sim-and-Real Co-Training: Learning from Hybrid Datasets

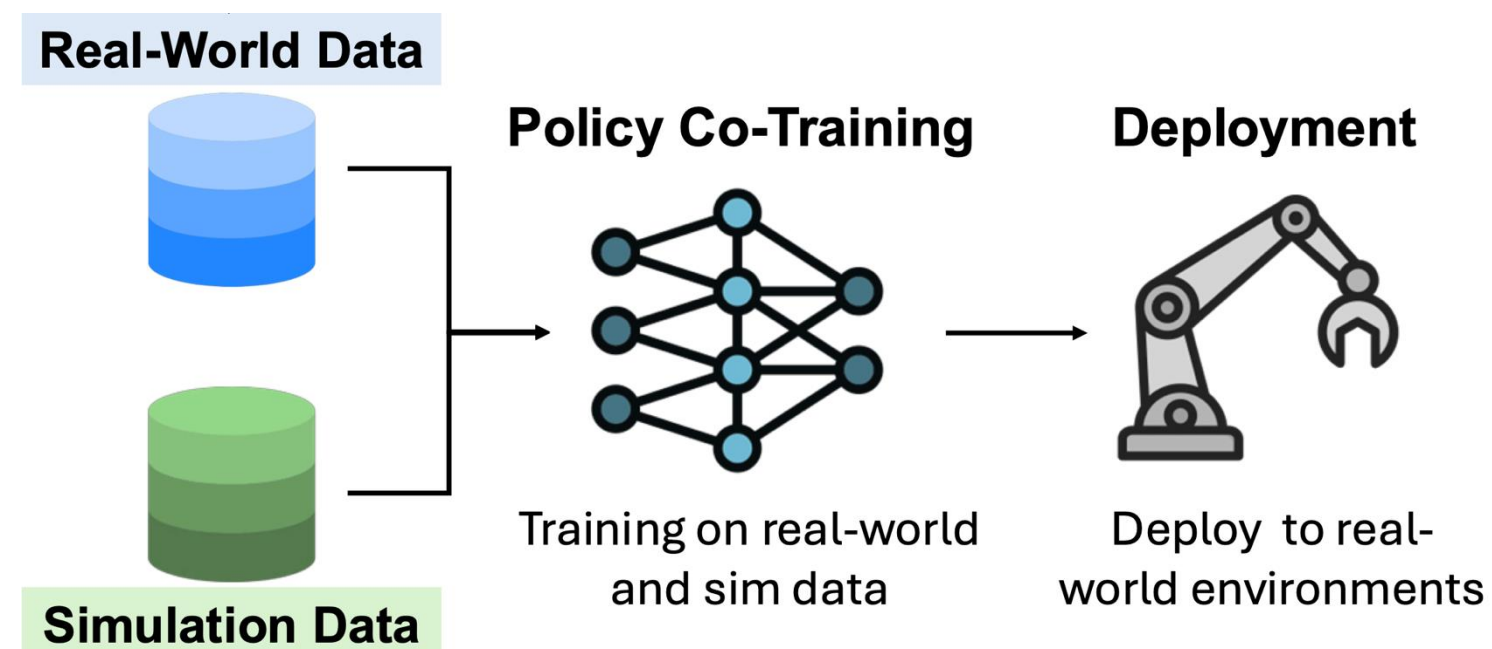
The devil is in the details for a successful co-training recipe.

- Enough of both **sim and real demonstrations** is required.
- Tuning the **co-training ratio** is important.
- **Camera alignment** is critical, but perfect alignment is not required.
- Co-training works in **large-scale, multi-task imitation learning** settings (more on this later).



Recent Work on **Co-Training** in Robotics

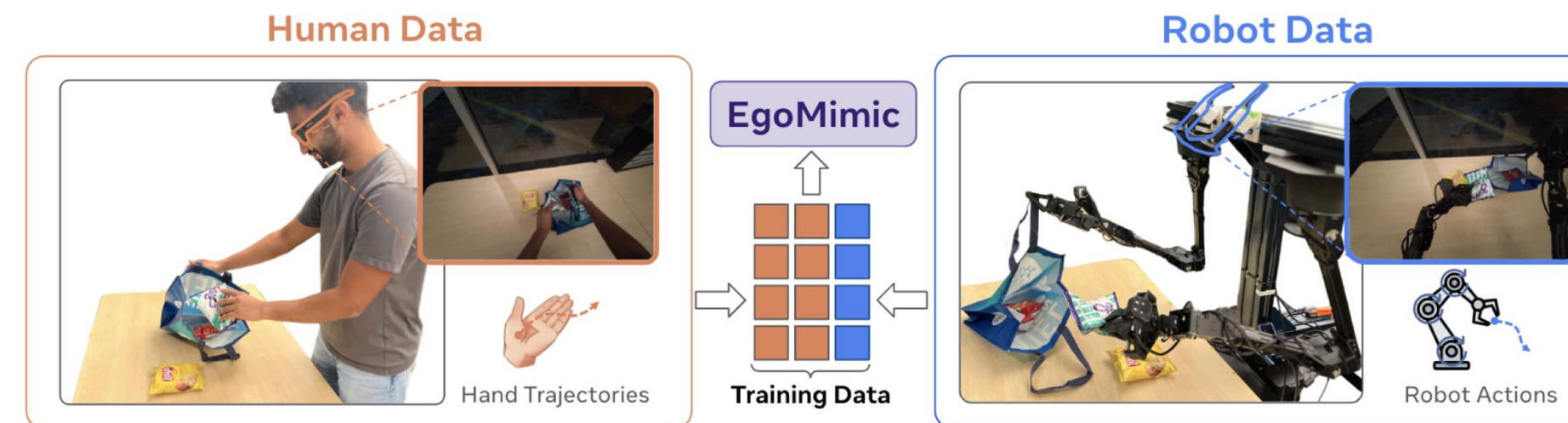
Sim-and-Real Co-Training



[Maddukuri et al. RSS'25; Wei et al. IROS'25]

- Small **real-world** data
- Large **simulation** data

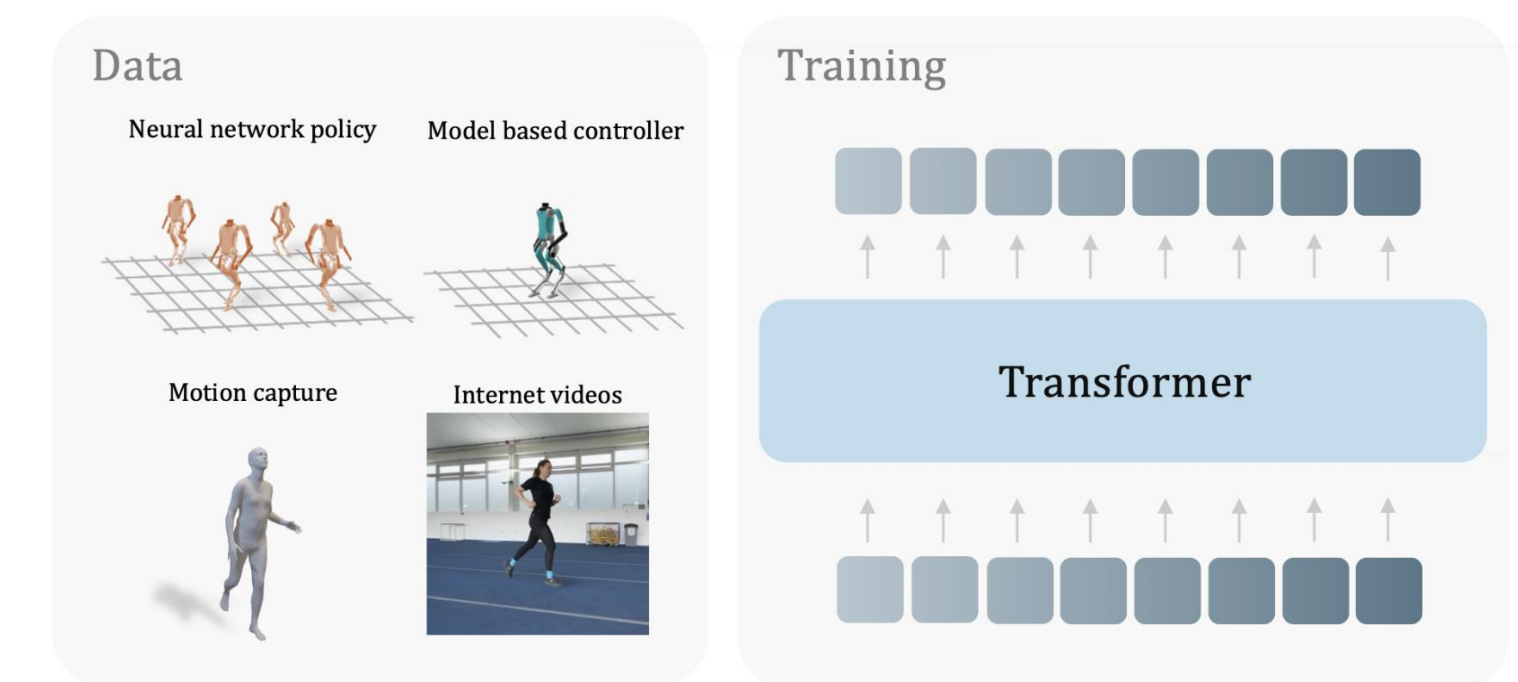
Human-Robot Co-Training



[Kareer et al. ICRA'25; Qiu et al. CoRL'25; Lepert et al. arXiv'25;
Liu et al. CoRL'25; Punamiya et al. NeurIPS 2025]

- Small **robot teleoperation** data
- Large **human manipulation** data

State-Action Co-Training

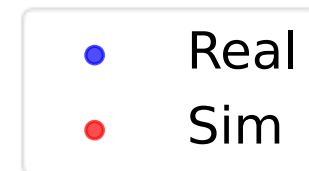


[Radosavovic et al. NeurIPS'24; Li et al. RSS'25]

- Small **state-action paired** data
- Large **actionless** data

Why Co-Training Works? Local Alignment in Feature Space

Embeddings from Visual Encoder



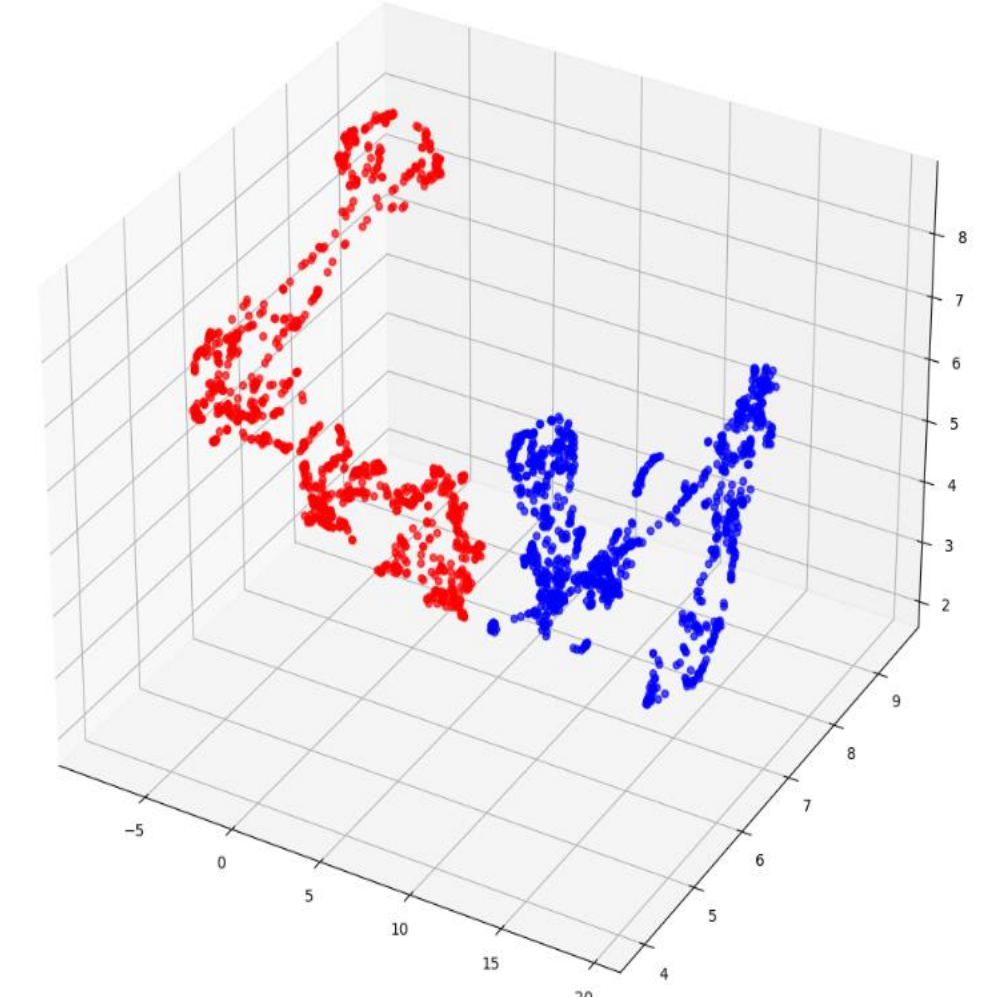
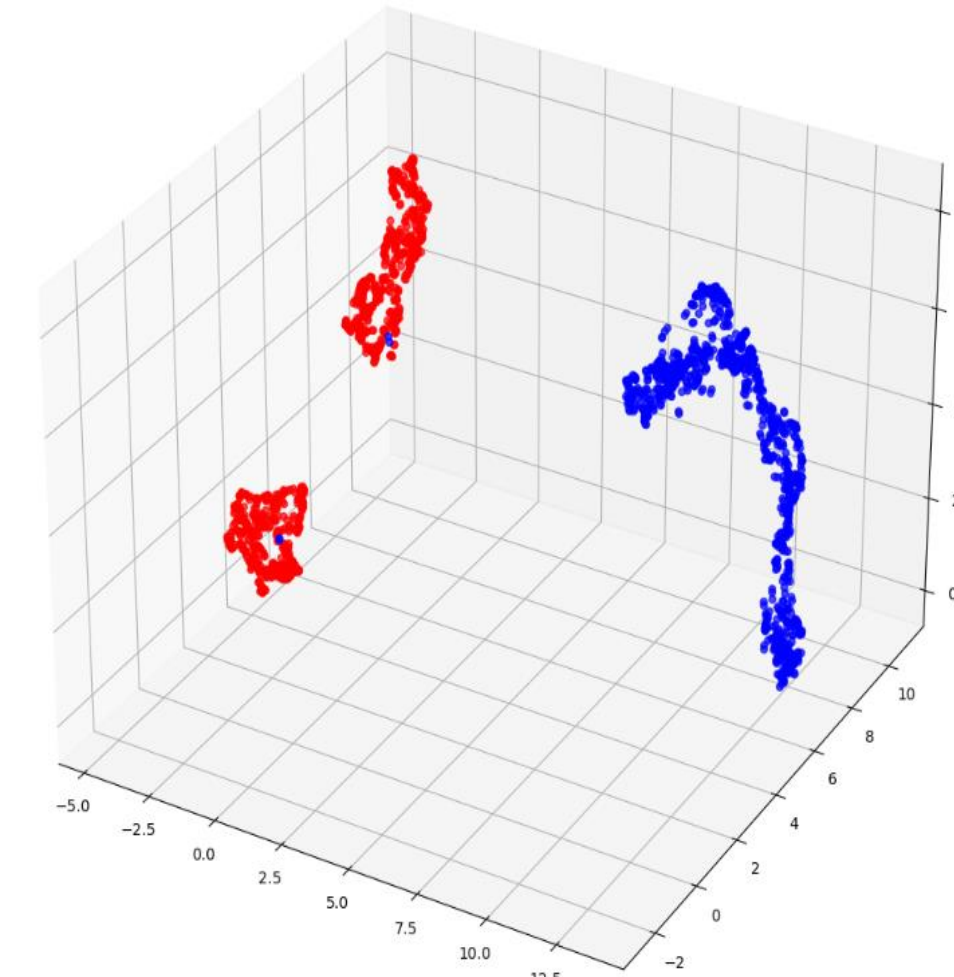
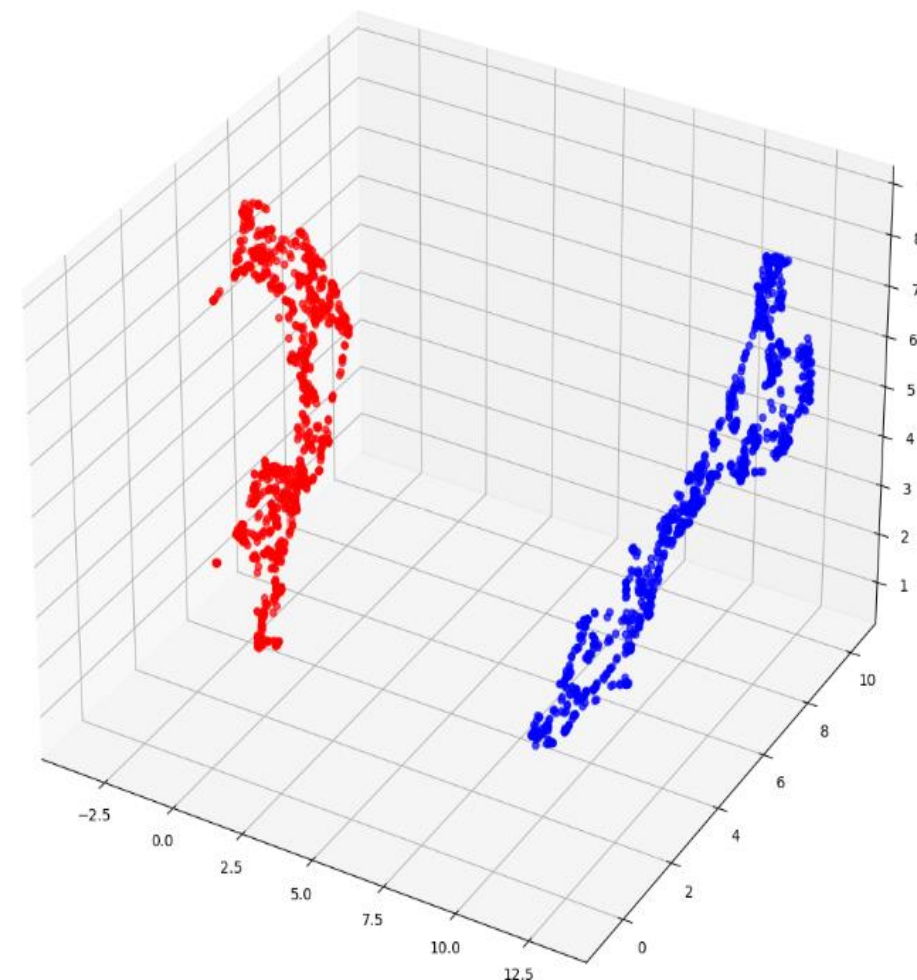
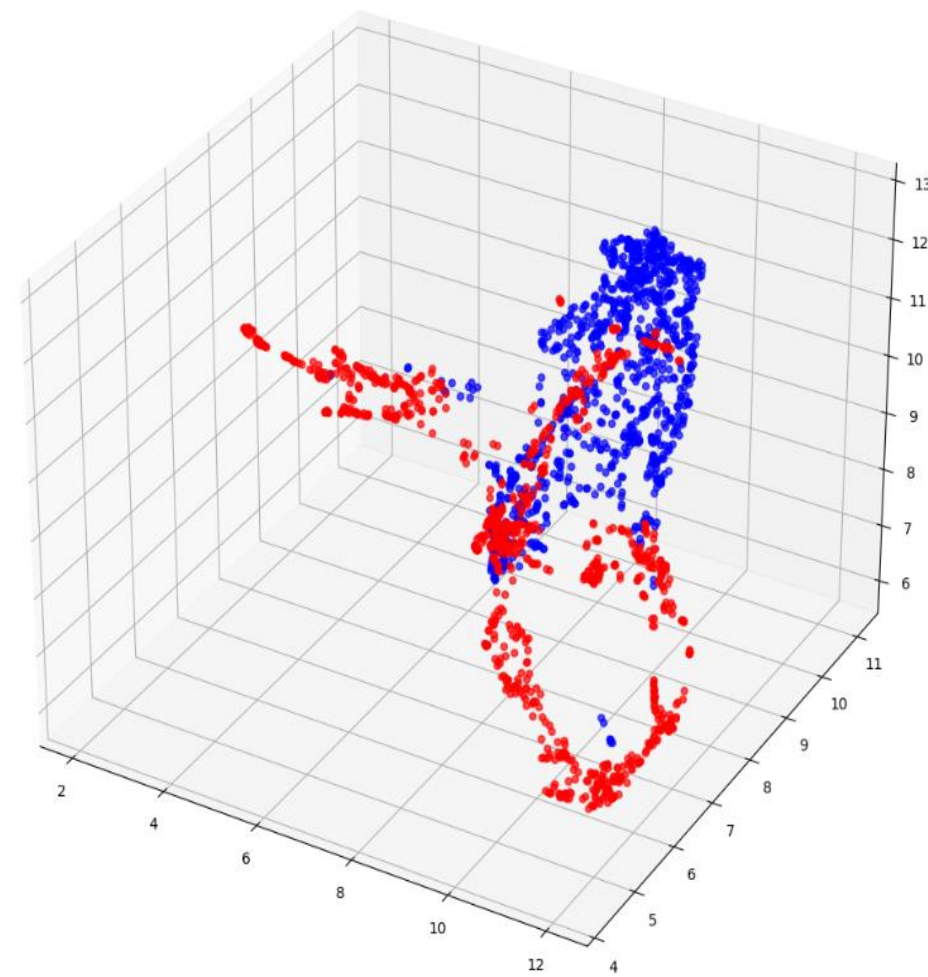
Real

Real + DC

Real + Prior

Real + DC + Prior

CounterTo
Cab



Sim and real visual features share very similar structures!

Why Co-Training Works? Local Alignment in Feature Space

Embeddings from Visual Encoder

● Real
● Sim

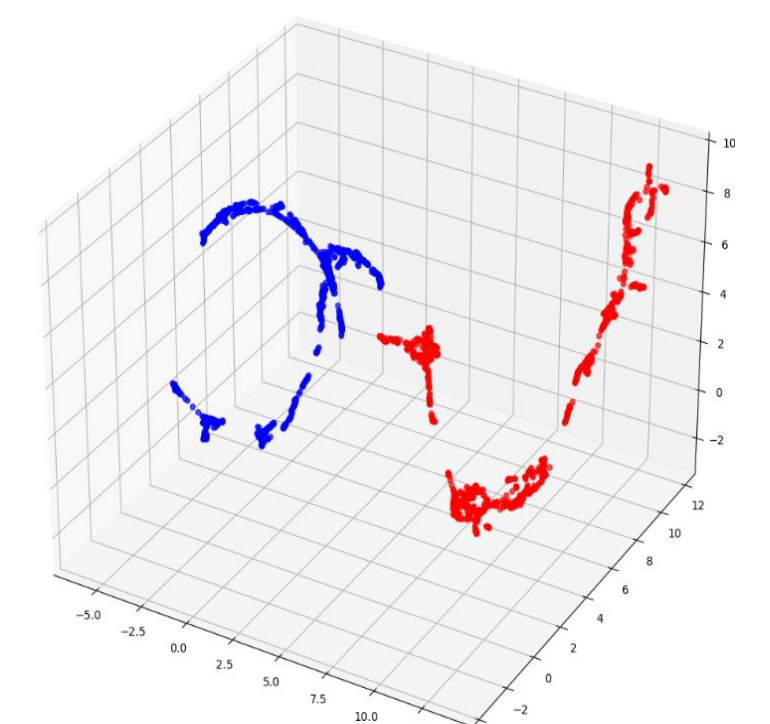
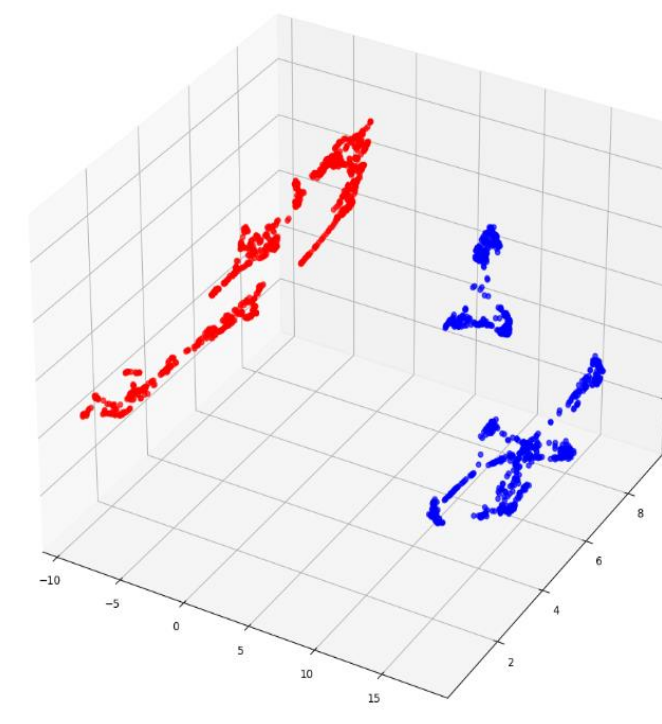
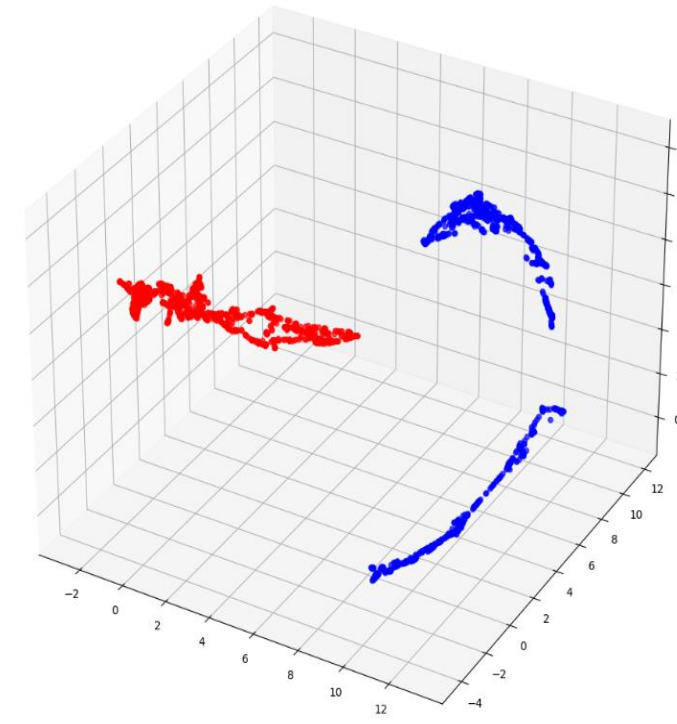
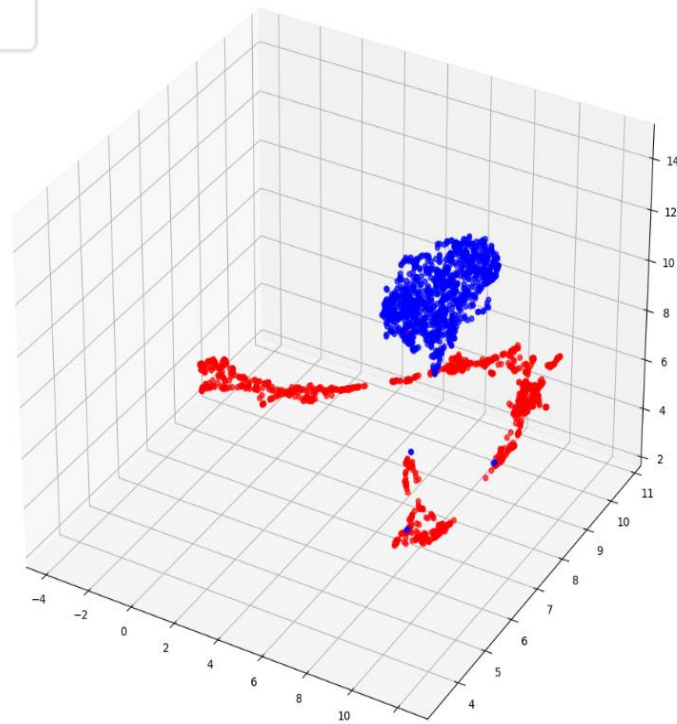
Real

Real + DC

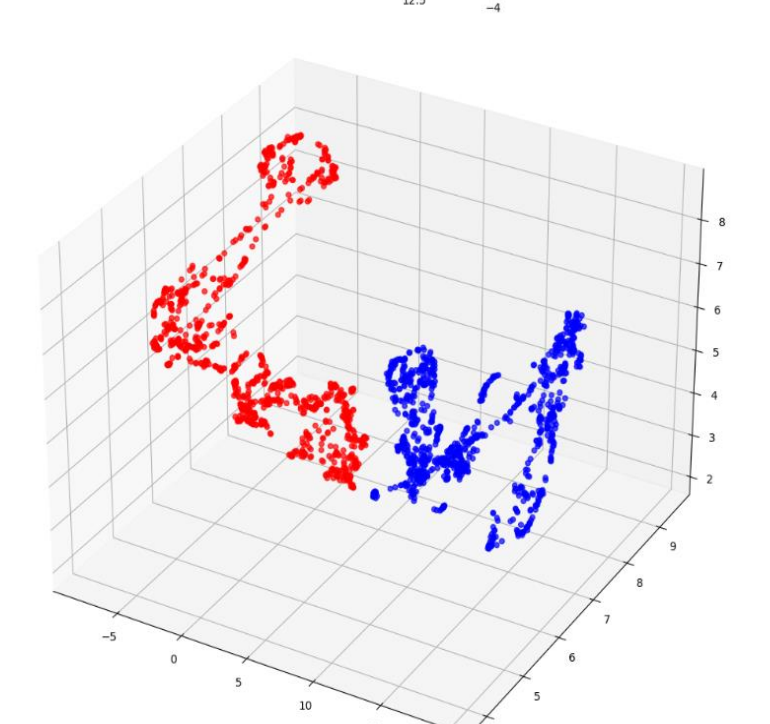
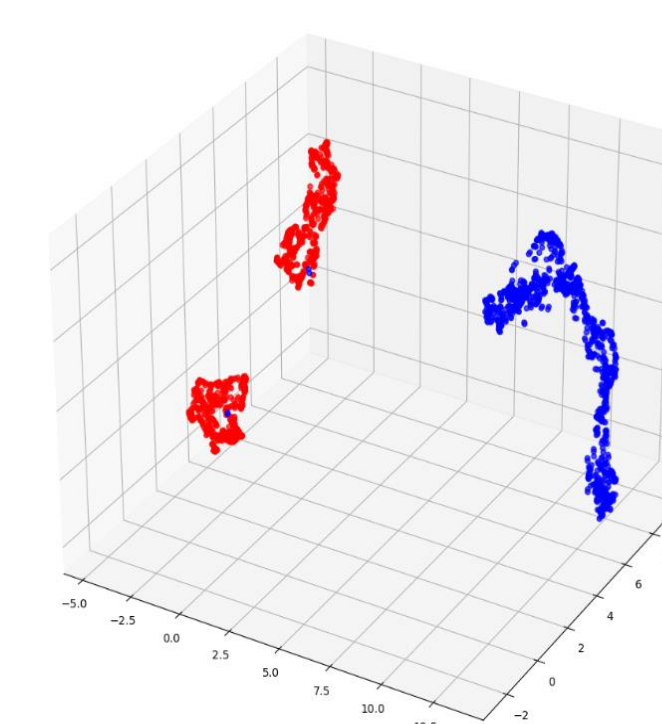
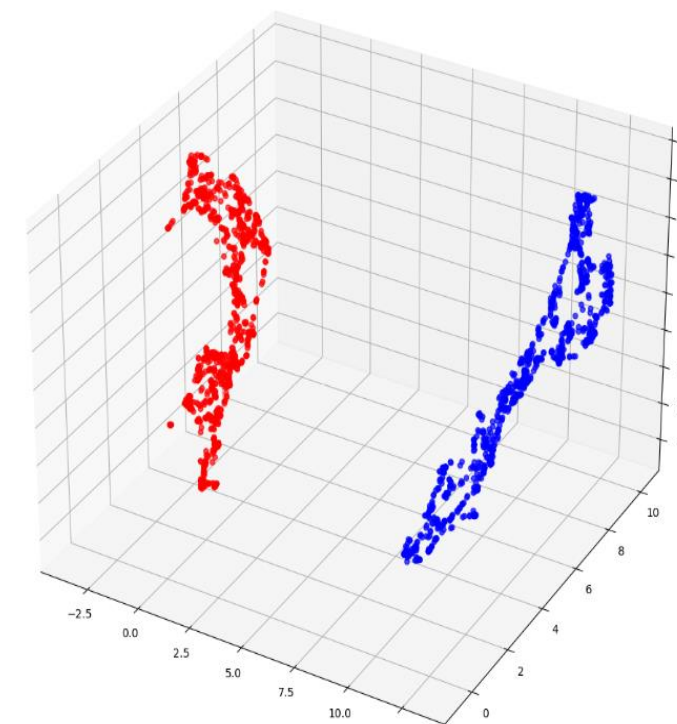
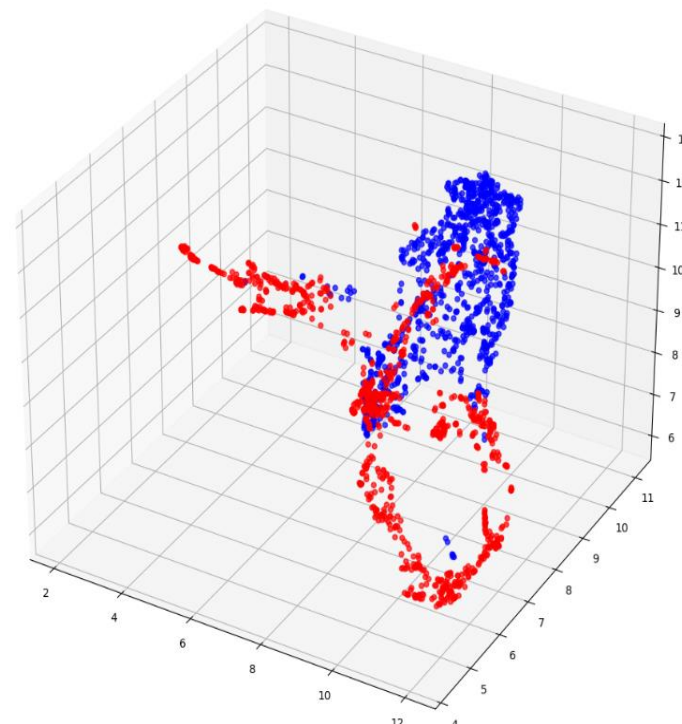
Real + Prior

Real + DC + Prior

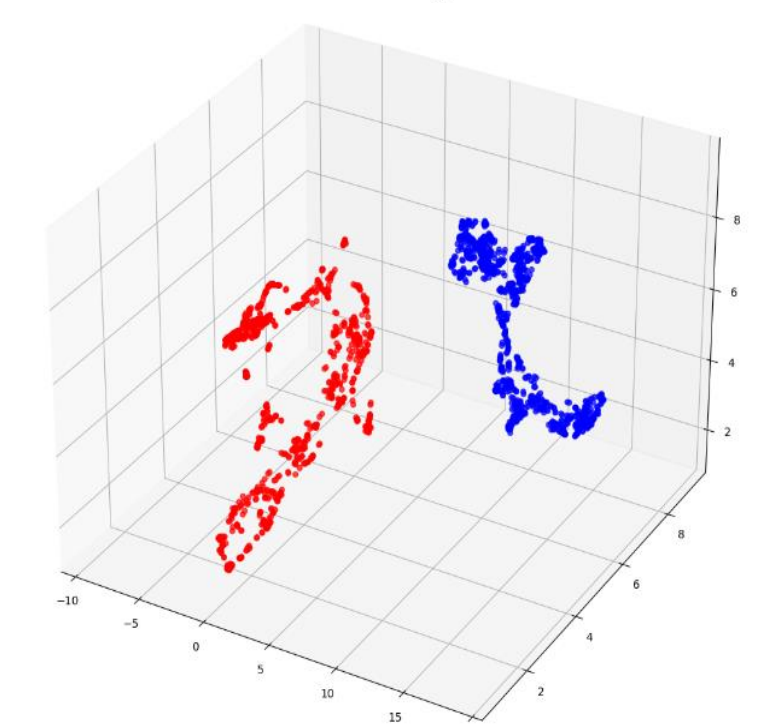
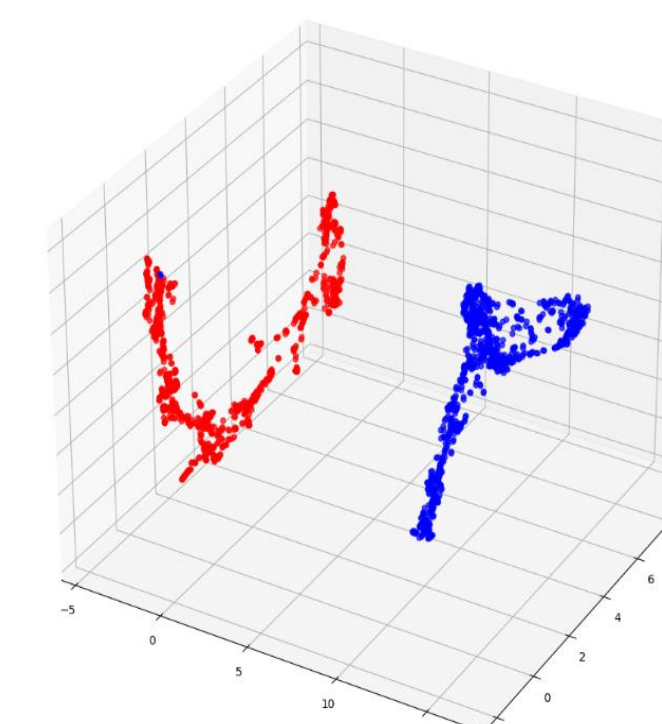
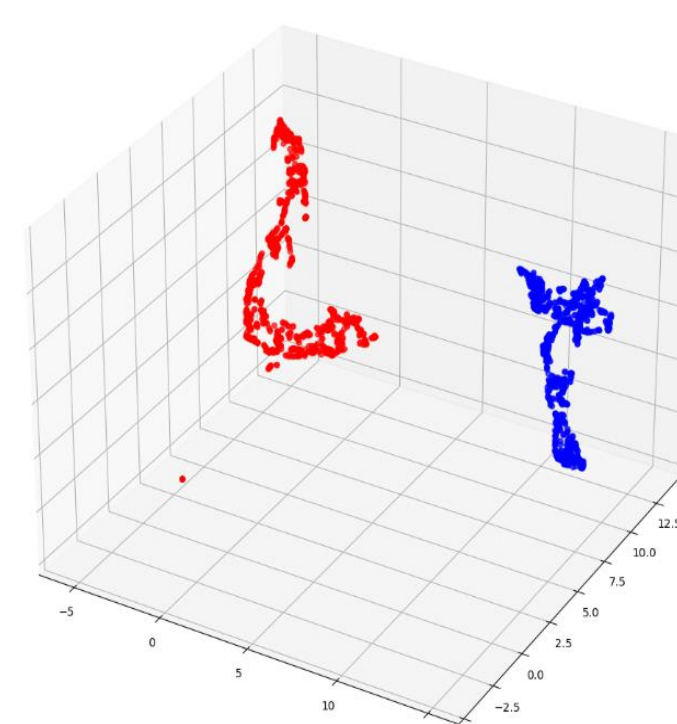
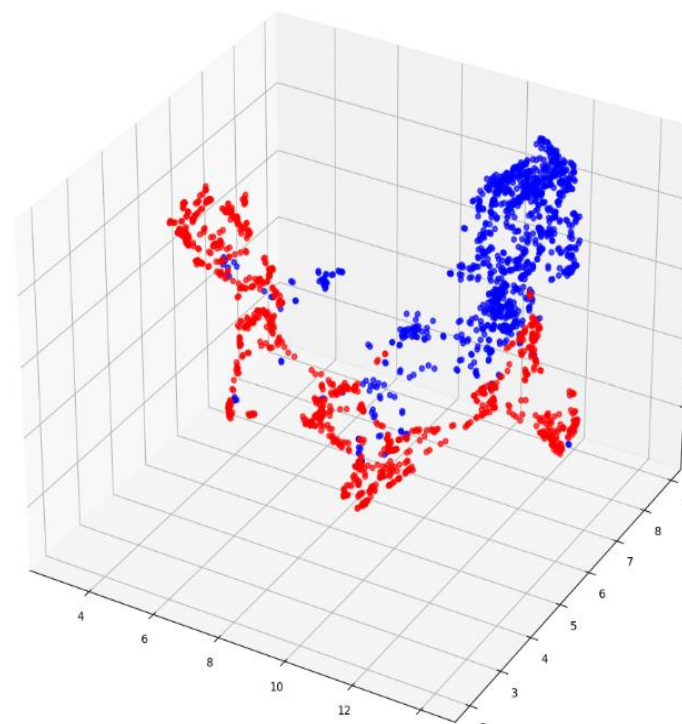
CloseDoor



CounterTo
Cab

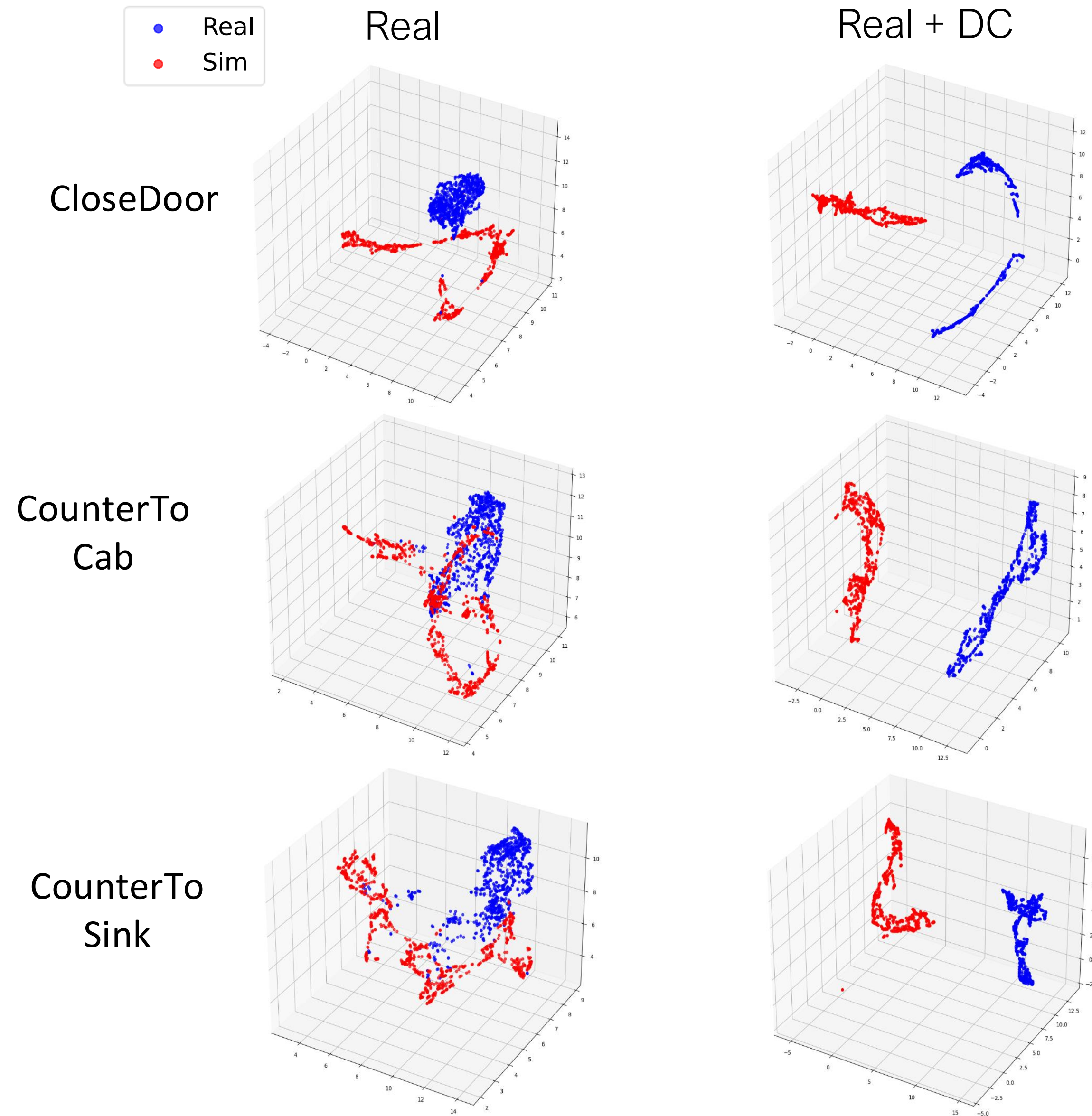


CounterTo
Sink



Why Co-Training Works? Local Alignment in Feature Space

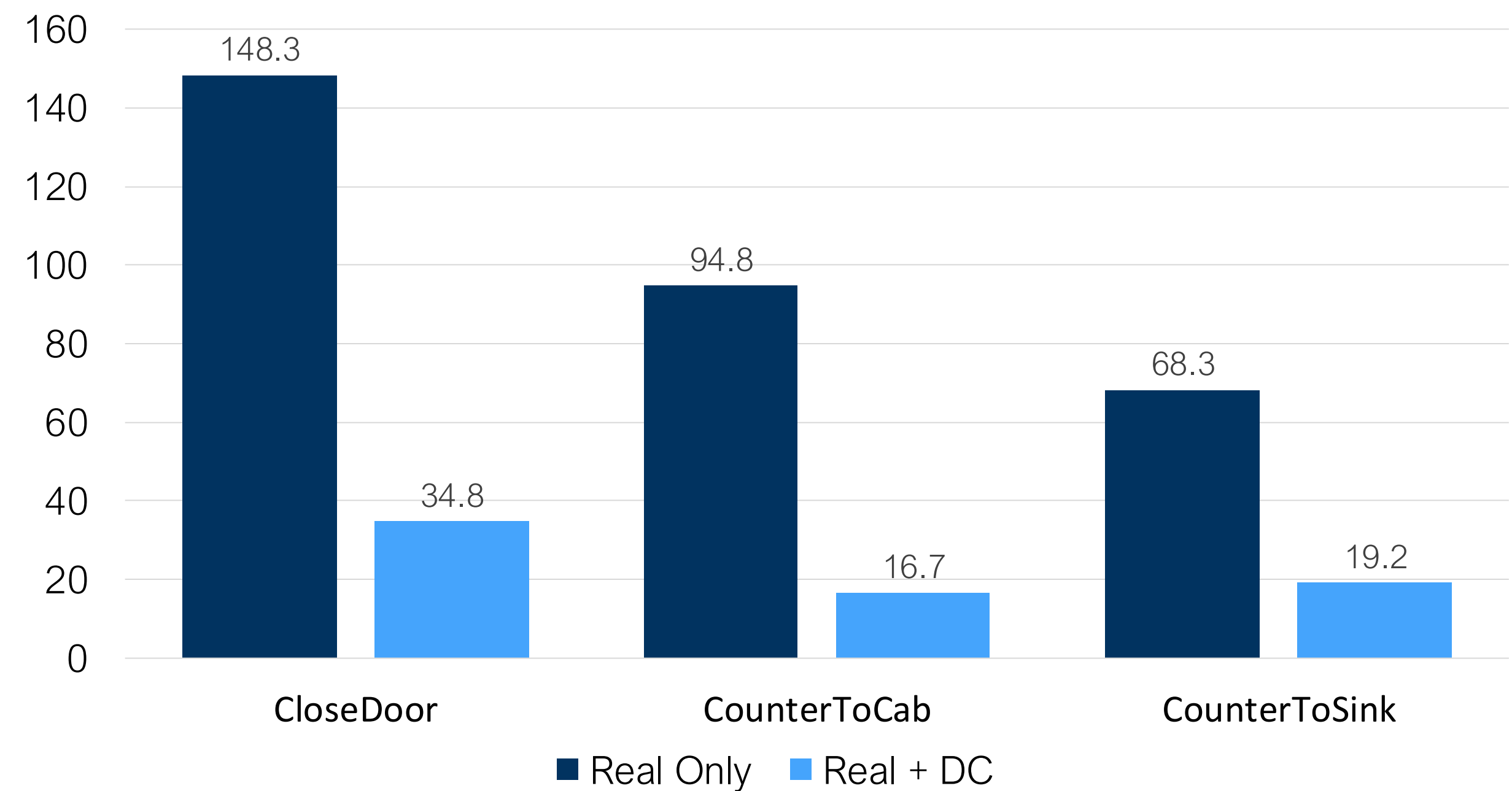
Embeddings from Visual Encoder



d_{GW} : Similarity metric for intrinsic geometry of neural features

- $d_{GW} = 0$: represents exactly the same intrinsic geometry
- The smaller d_{GW} is, the more similar geometry two distributions share

Gromov-Wasserstein Distance



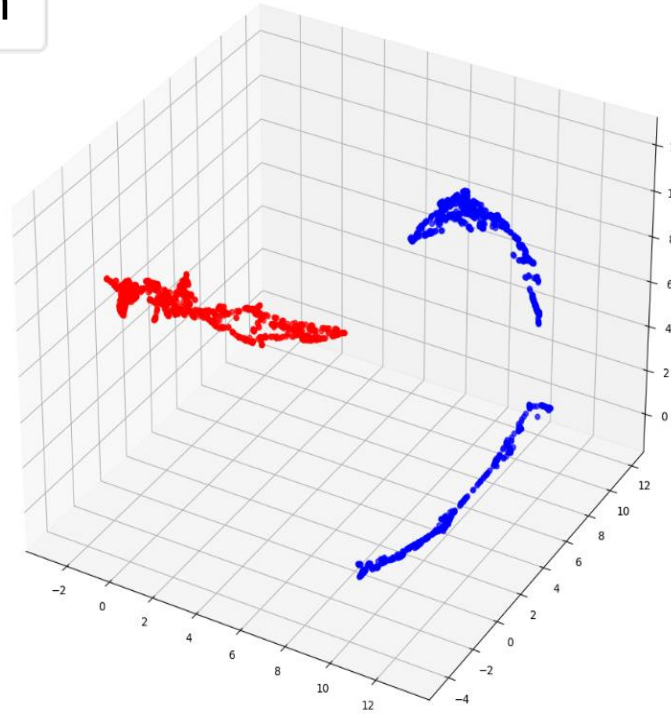
Why Co-Training Works? Global Alignment in Feature Space

Features Evolve through the Networks

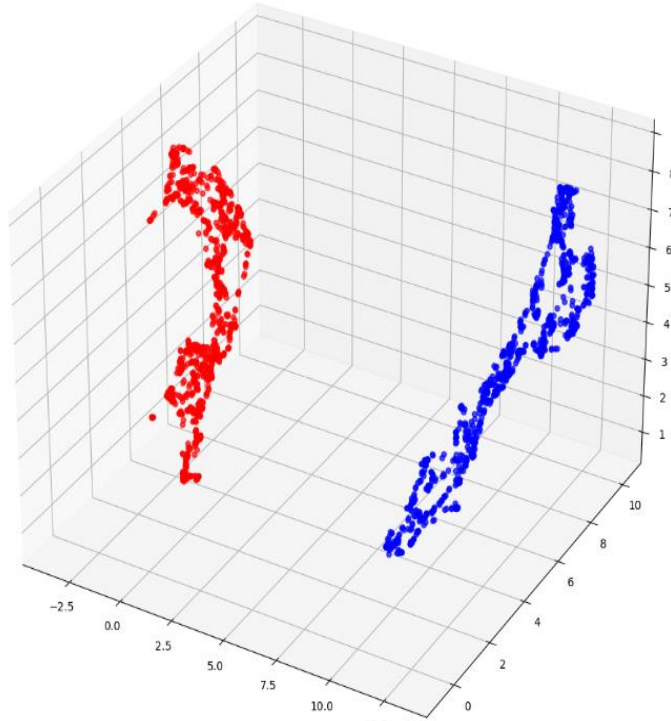
Visual Feature

● Real
● Sim

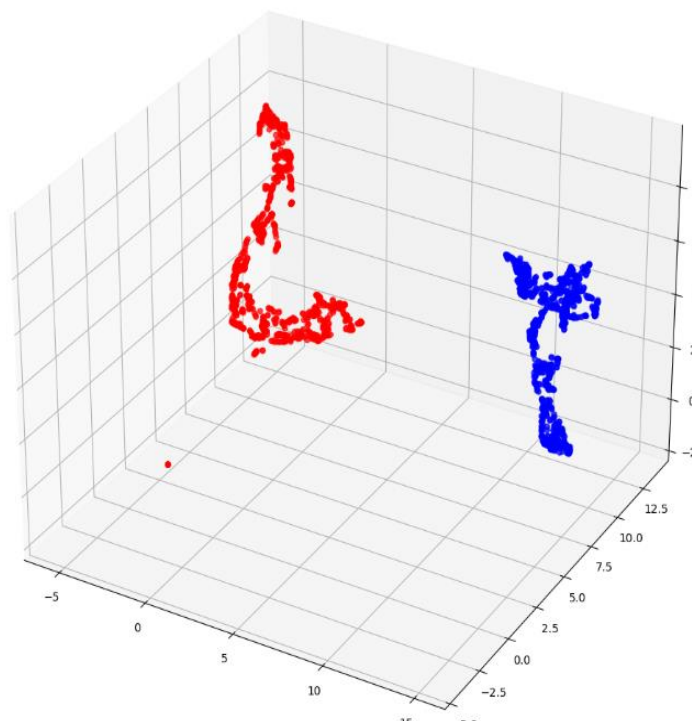
CloseDoor



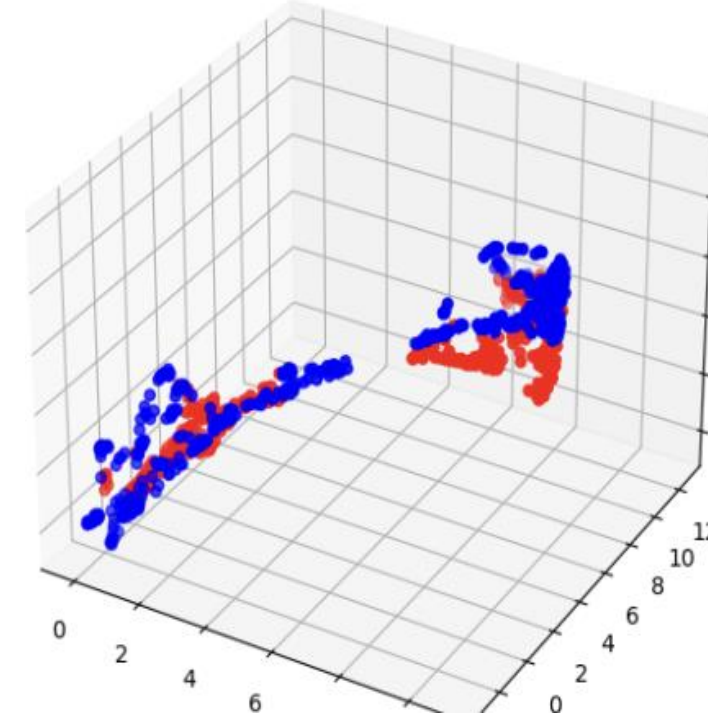
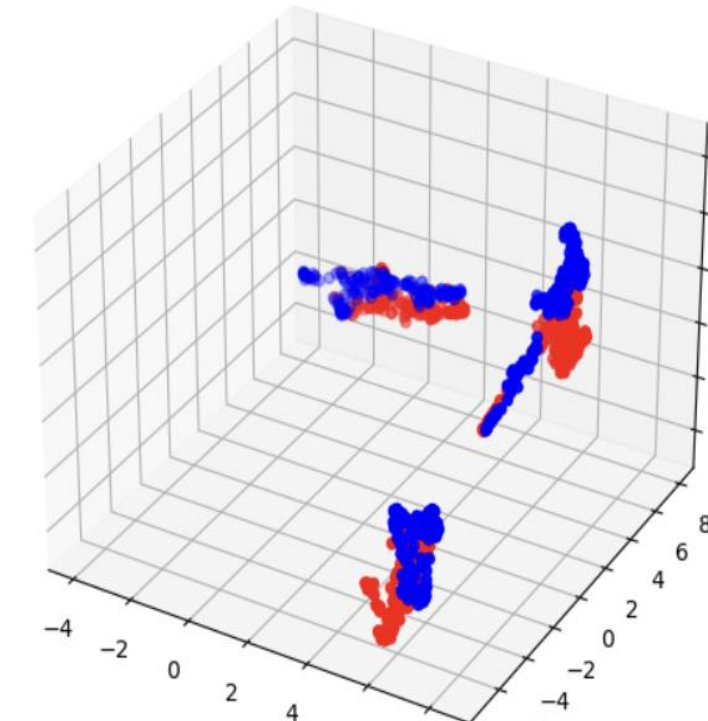
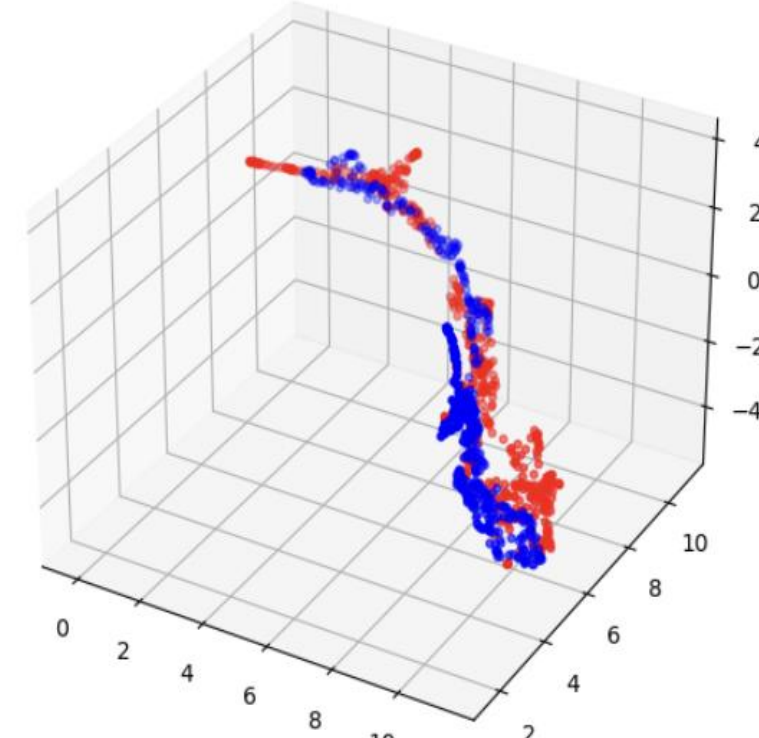
CounterTo
Cab



CounterTo
Sink



Transformer Feature

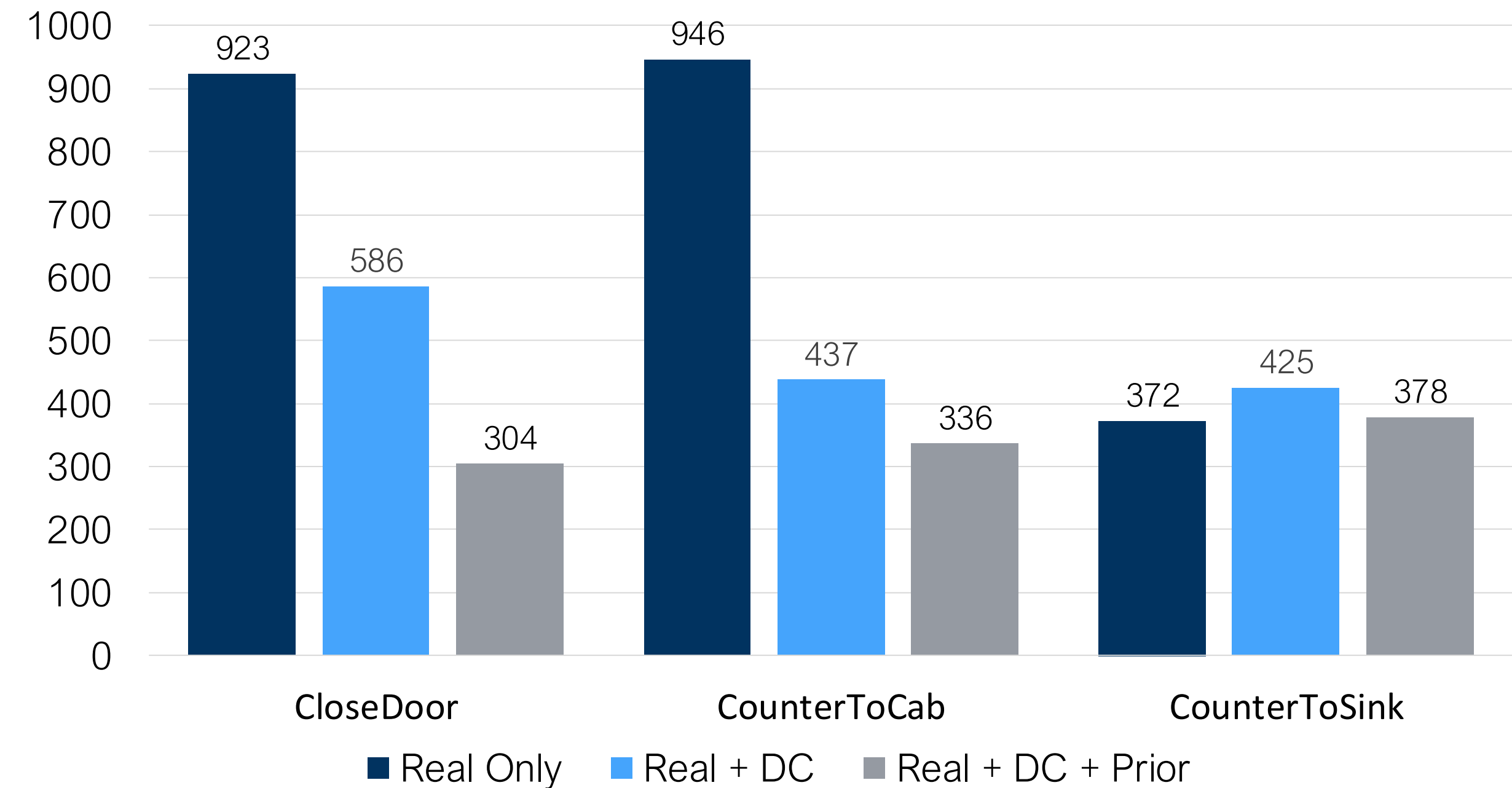


$$d_{Fused-GW} = \beta \cdot d_{WD} + (1 - \beta) \cdot d_{GW}^{[1]}$$

Global Distance Alignment + Local Structure Alignment

Scale Balance Coefficient

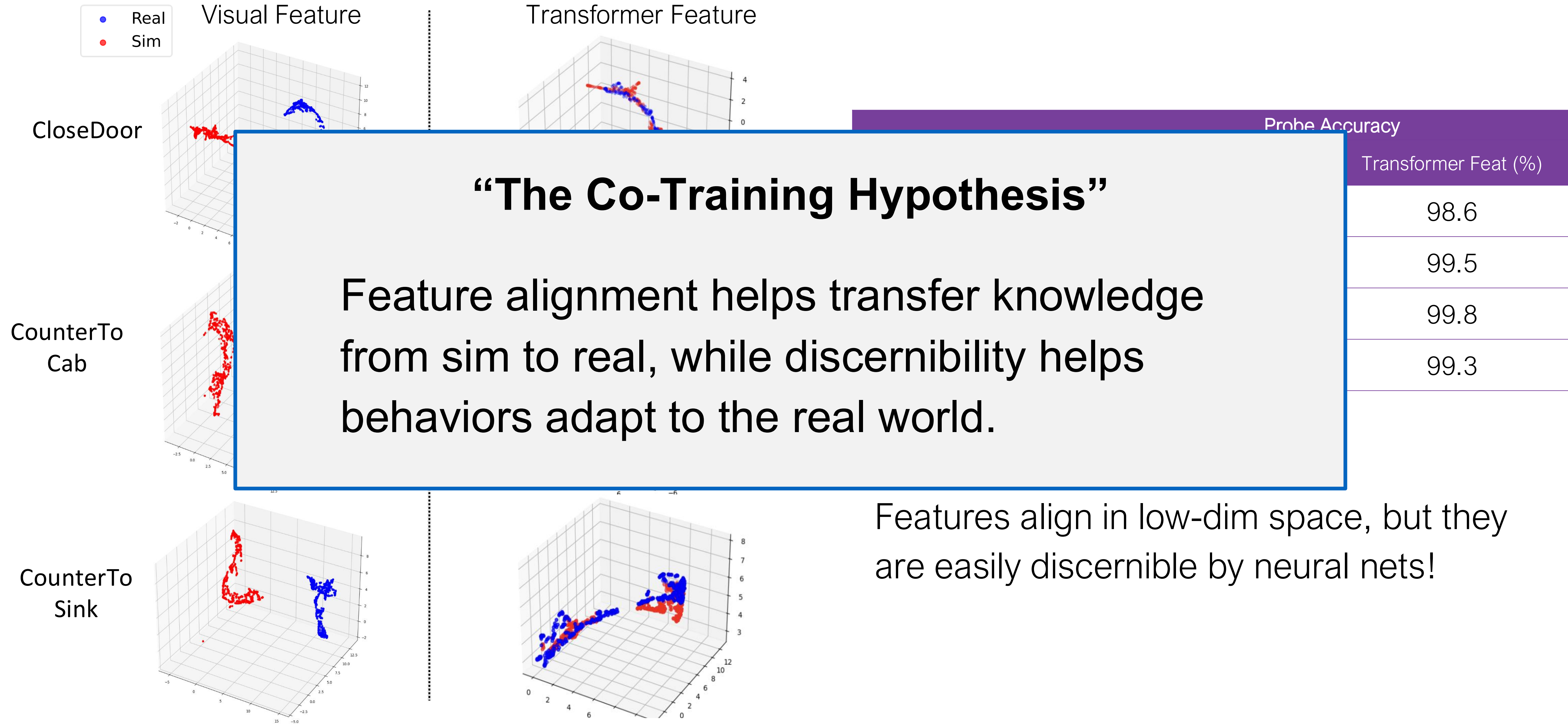
Transformer Encoder Feature Fused-GW Distance



[1] Vayer, Titouan, et al. "Fused Gromov-Wasserstein distance for structured objects: theoretical foundations and mathematical properties." *arXiv preprint arXiv:1811.02834* (2018).

Why Co-Training Works? Feature Discernibility

Linear Probes for Real/Sim Binary Classification on Transformer Features



Video Generation Models: Towards Neural Simulators

 **runway**
GEN-3

March 2023



May 2025



GOOGLE
VEO 3

 **Wan**


Sora 2


Dream
MACHINE
LUMA AI


Hunyuan Video

[Source: "Will Smith Eating Spaghetti test", Wikipedia]

DreamGen: Data Synthesis with Video Generation Models

Image Observation



Language Instruction

“Pick up the object and
place it on the pan.”

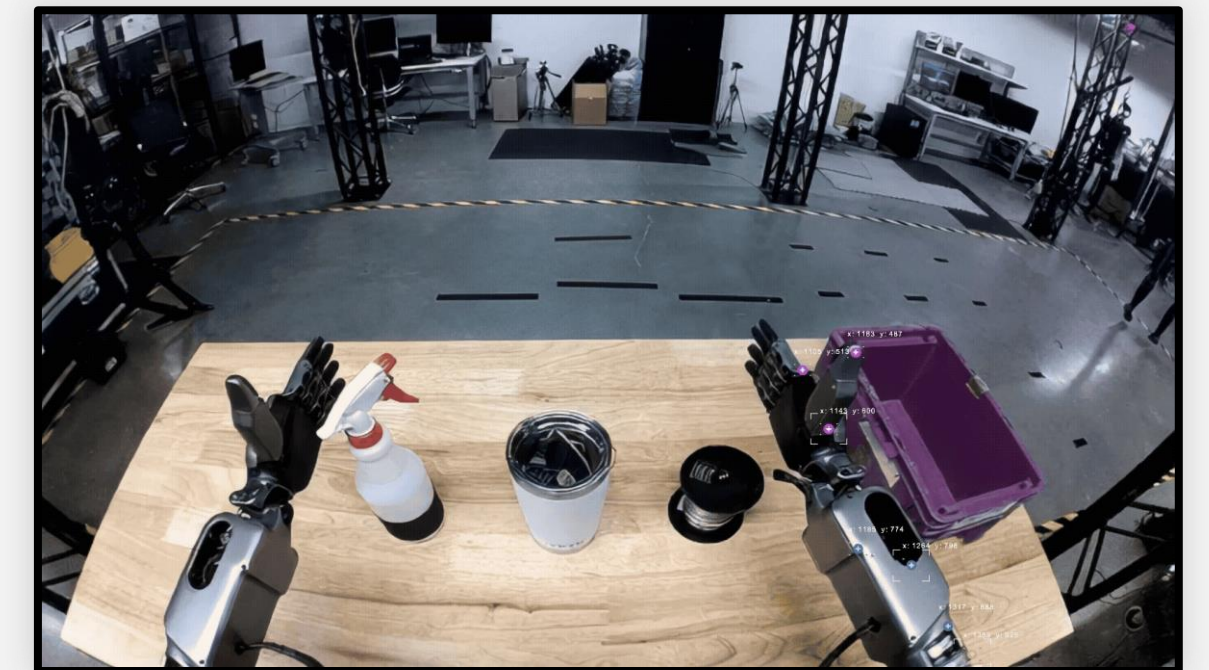
DreamGen: Data Synthesis with Video Generation Models

Image Observation

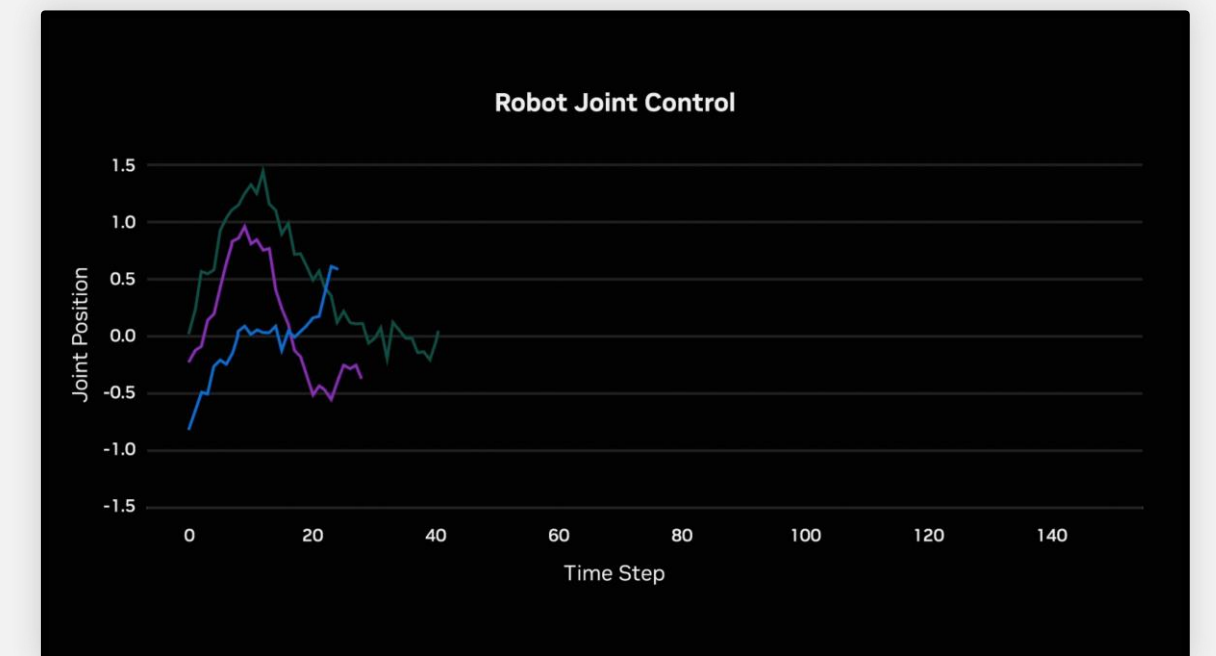


Language Instruction

“Pick up the object and place it on the pan.”



Generated Video



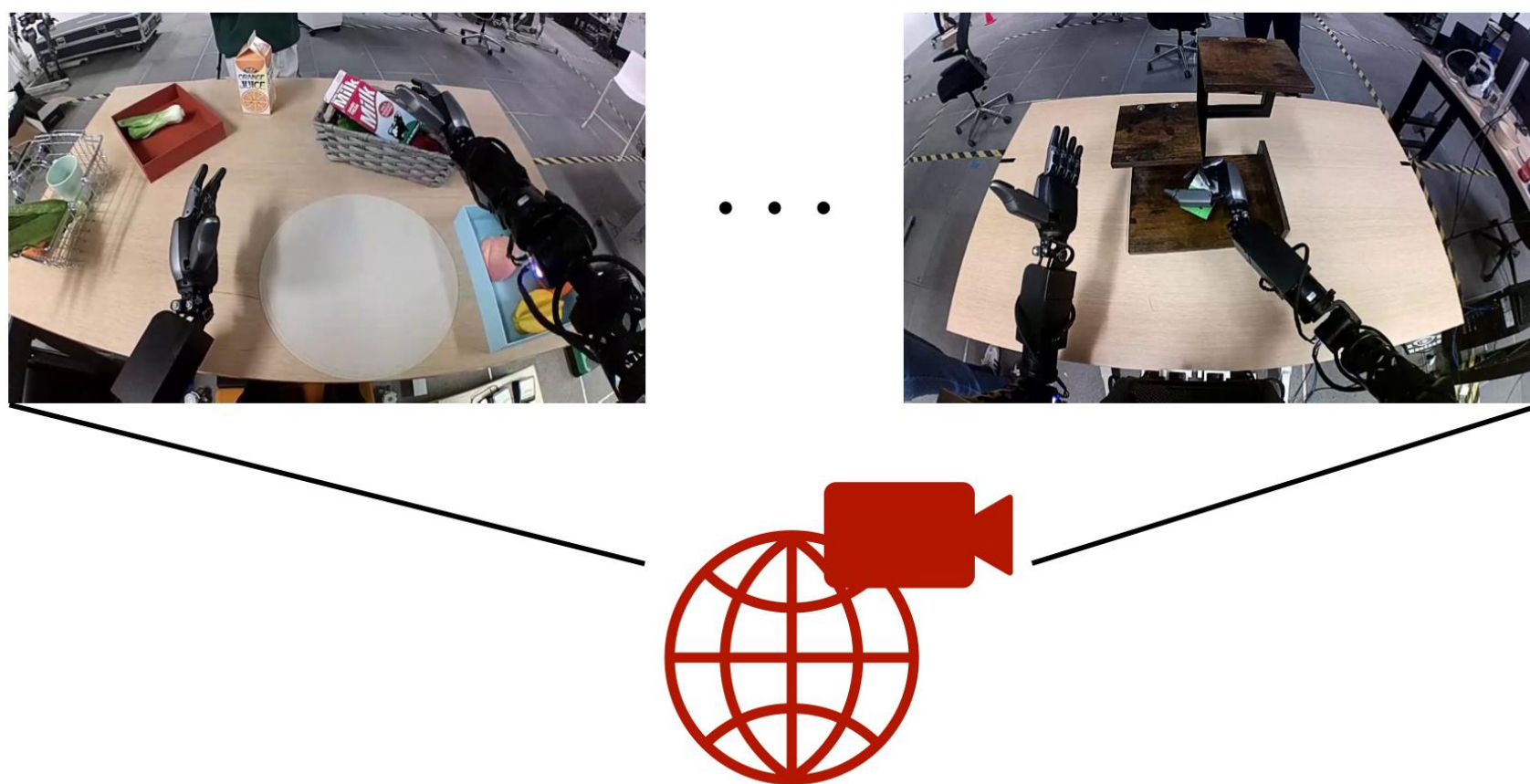
Predicted Robot Action

Neural Trajectories

DreamGen: Data Synthesis with Video Generation Models

Step 1. Finetune Video World Model

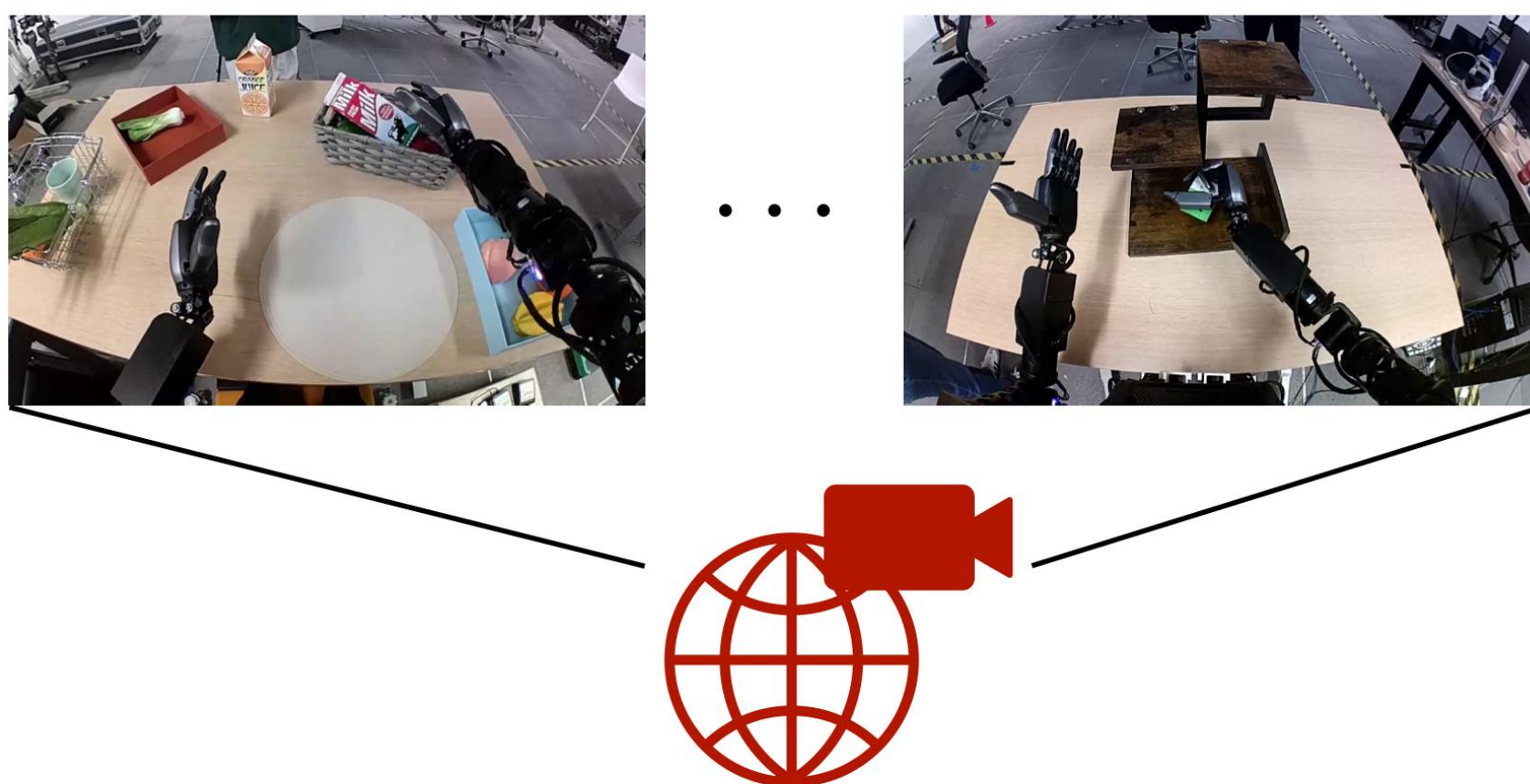
Human teleoperation data



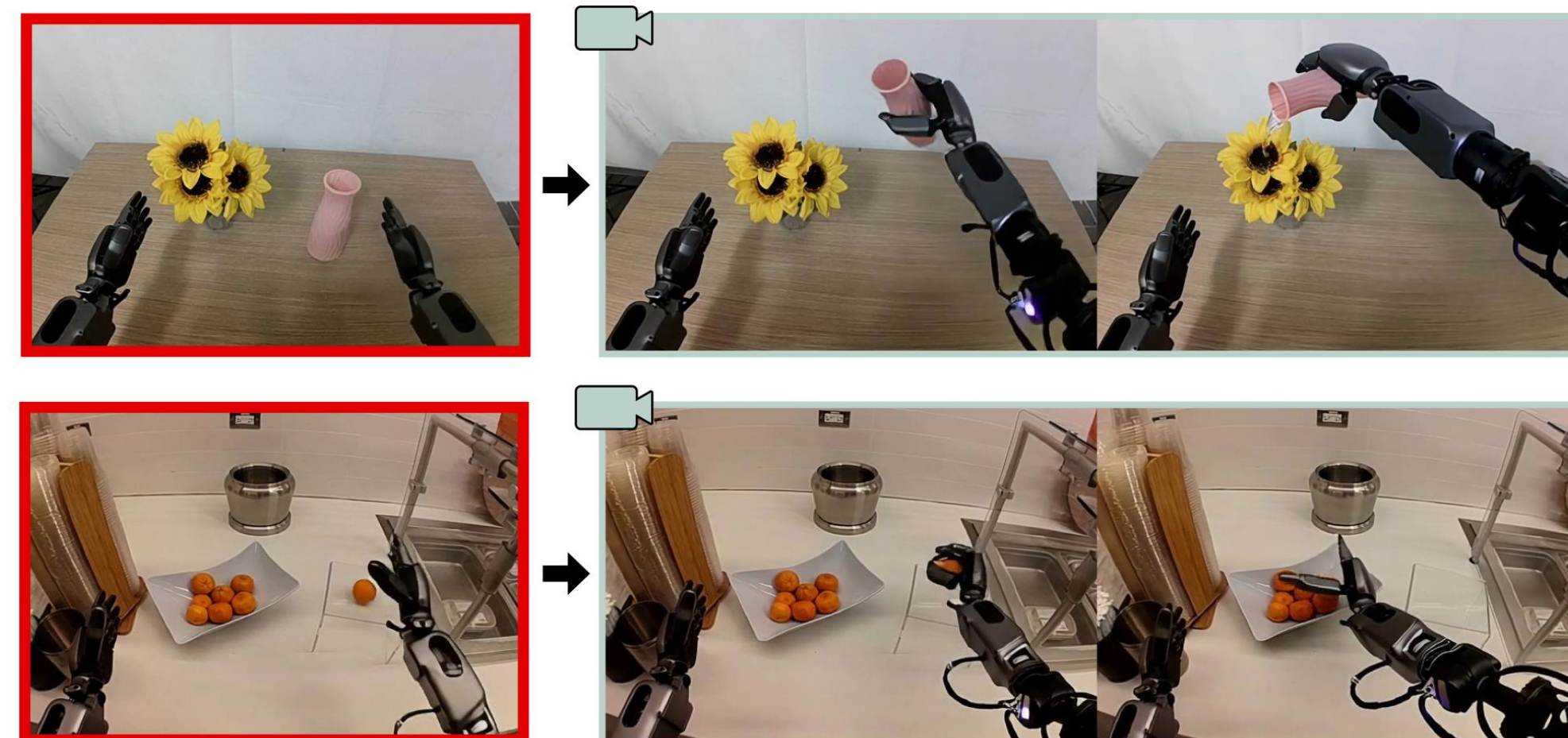
DreamGen: Data Synthesis with Video Generation Models

Step 1. Finetune Video World Model

Human teleoperation data



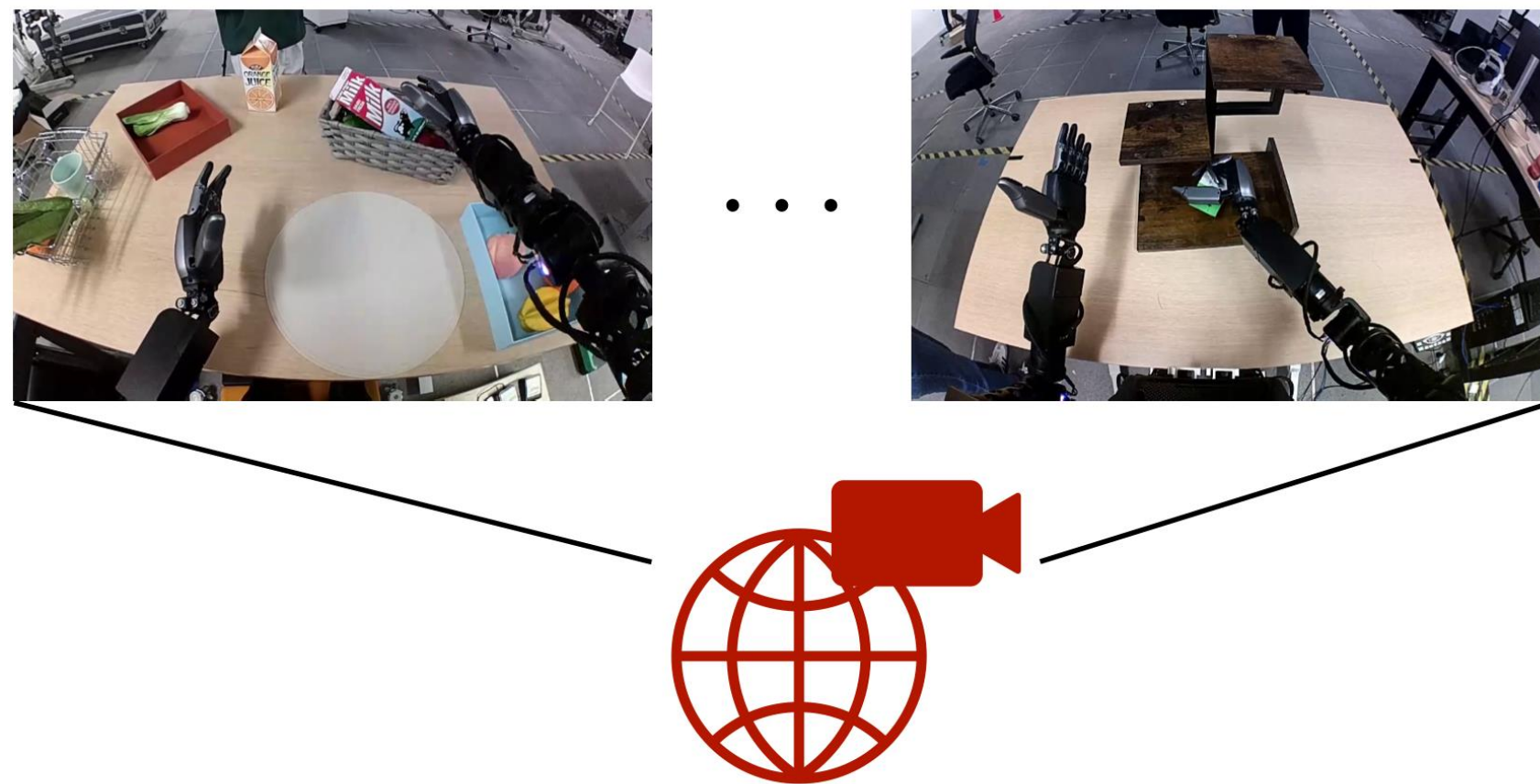
Step 2. Rollout Video World Model



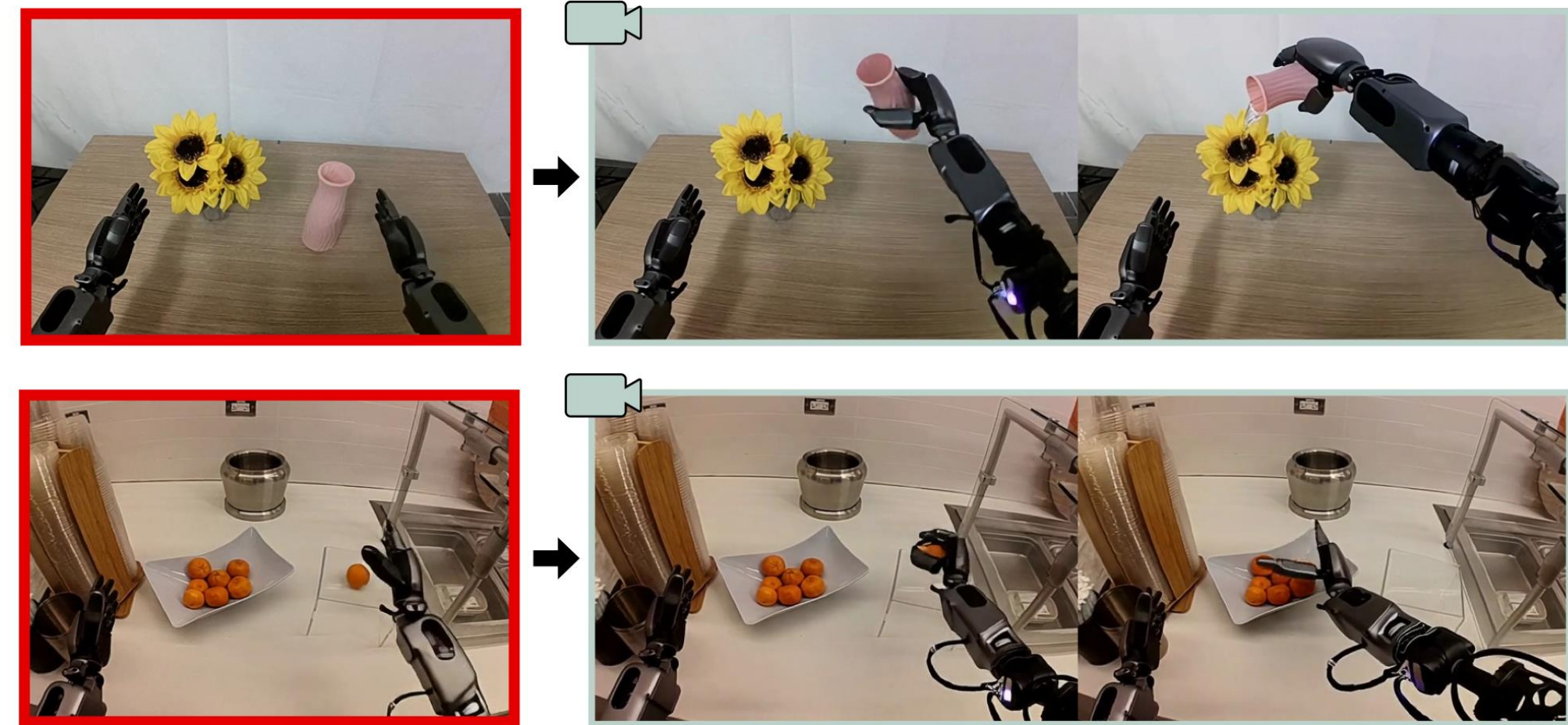
DreamGen: Data Synthesis with Video Generation Models

Step 1. Finetune Video World Model

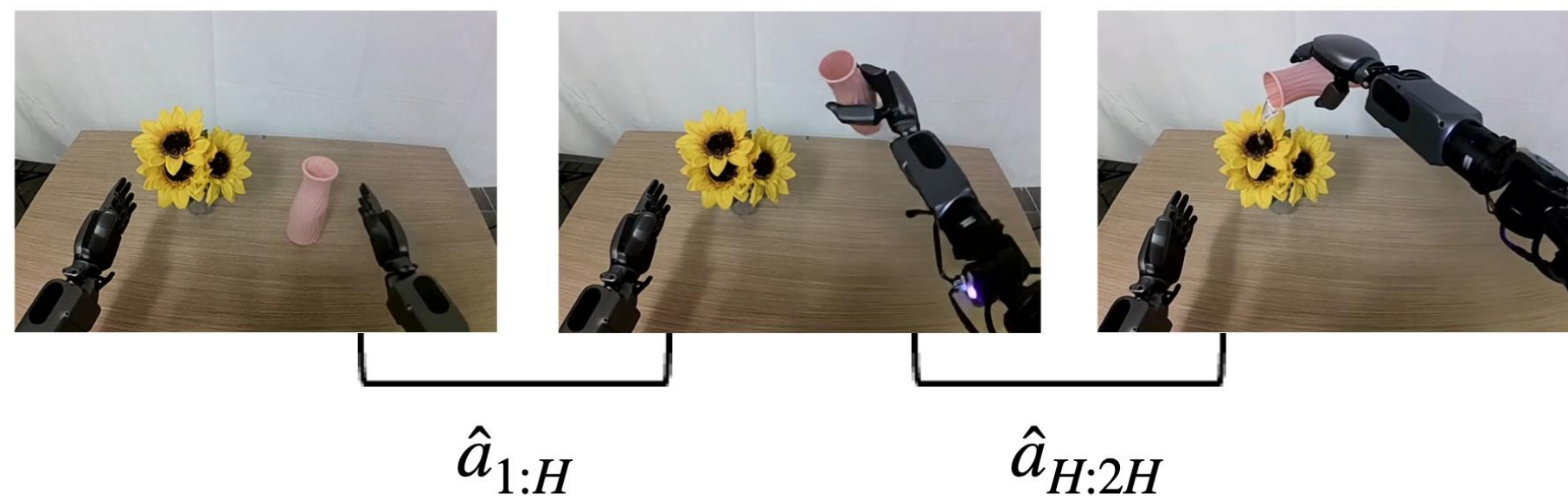
Human teleoperation data



Step 2. Rollout Video World Model



Step 3. Label Pseudo Actions

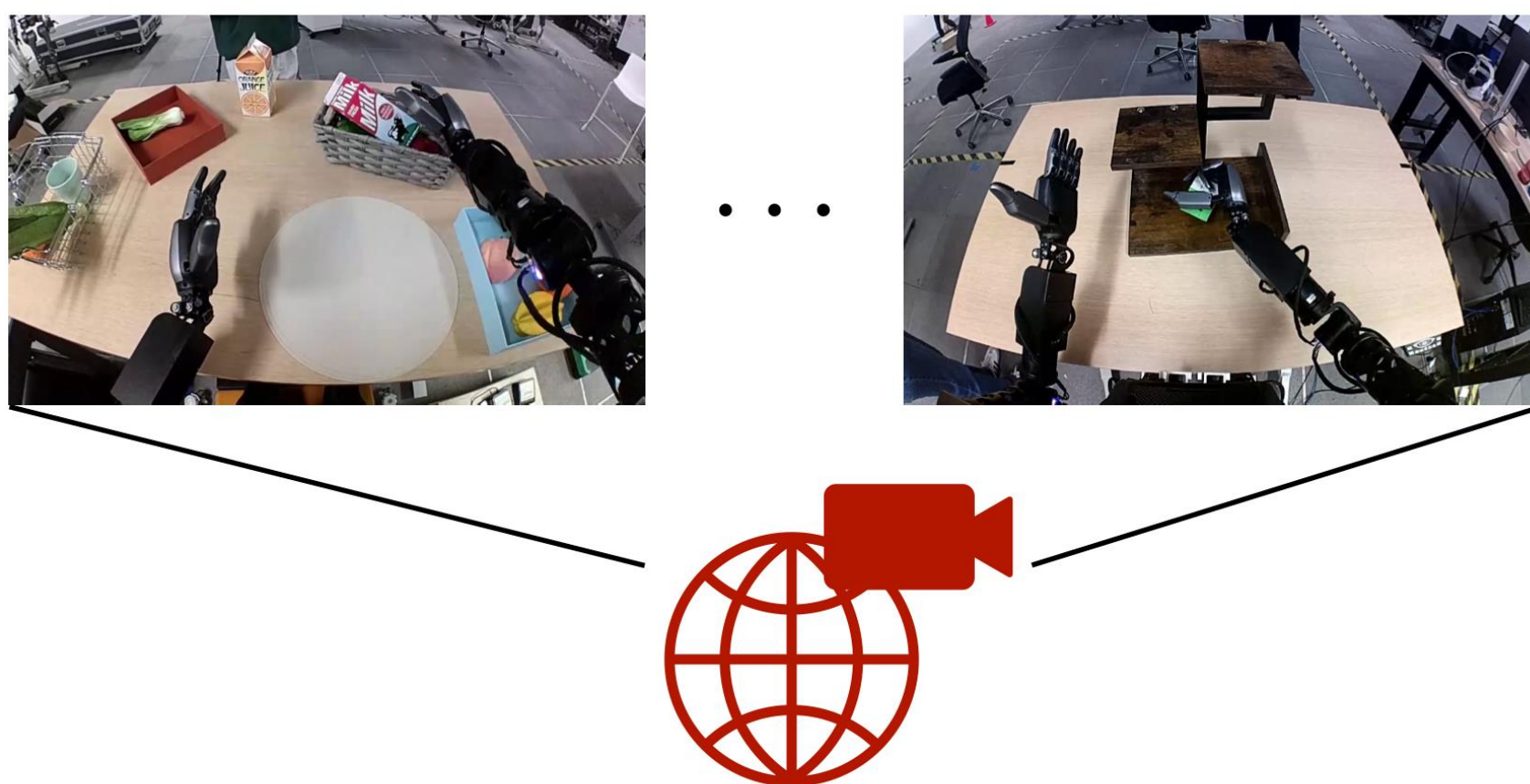


Automatically Labeled Pseudo Actions

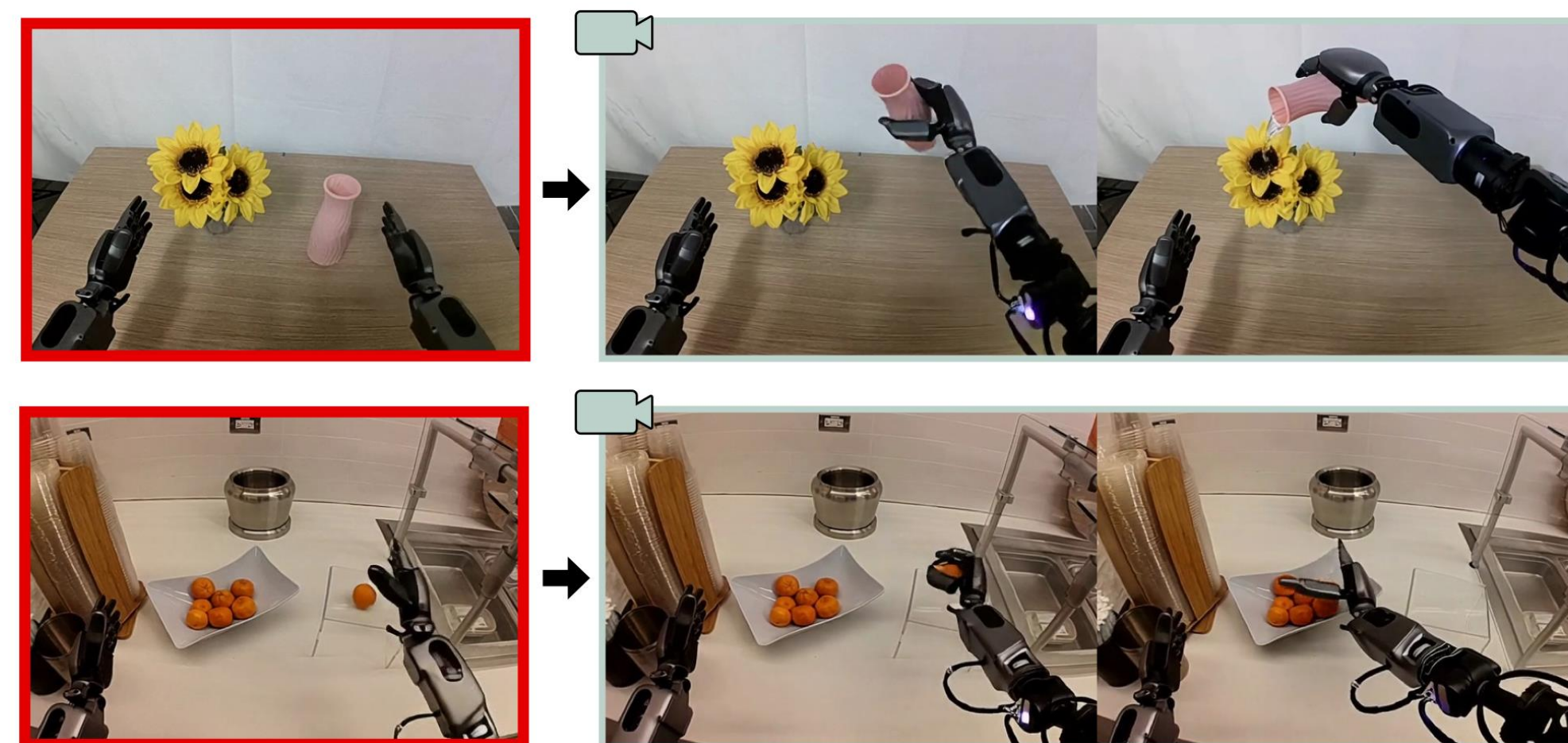
DreamGen: Data Synthesis with Video Generation Models

Step 1. Finetune Video World Model

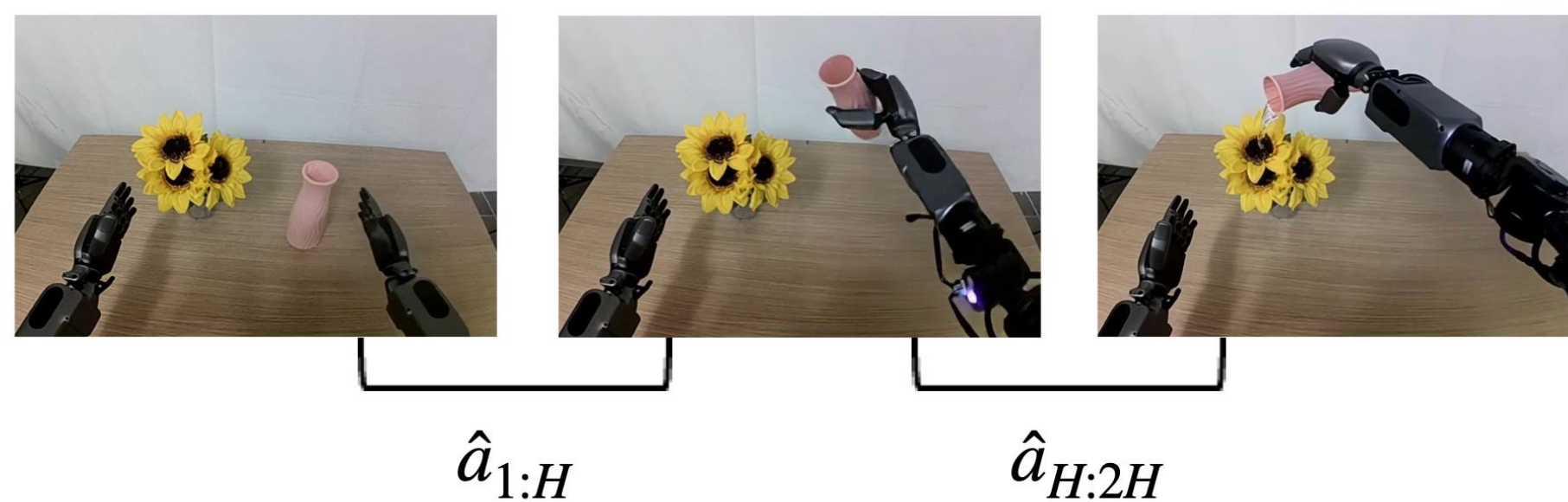
Human teleoperation data



Step 2. Rollout Video World Model

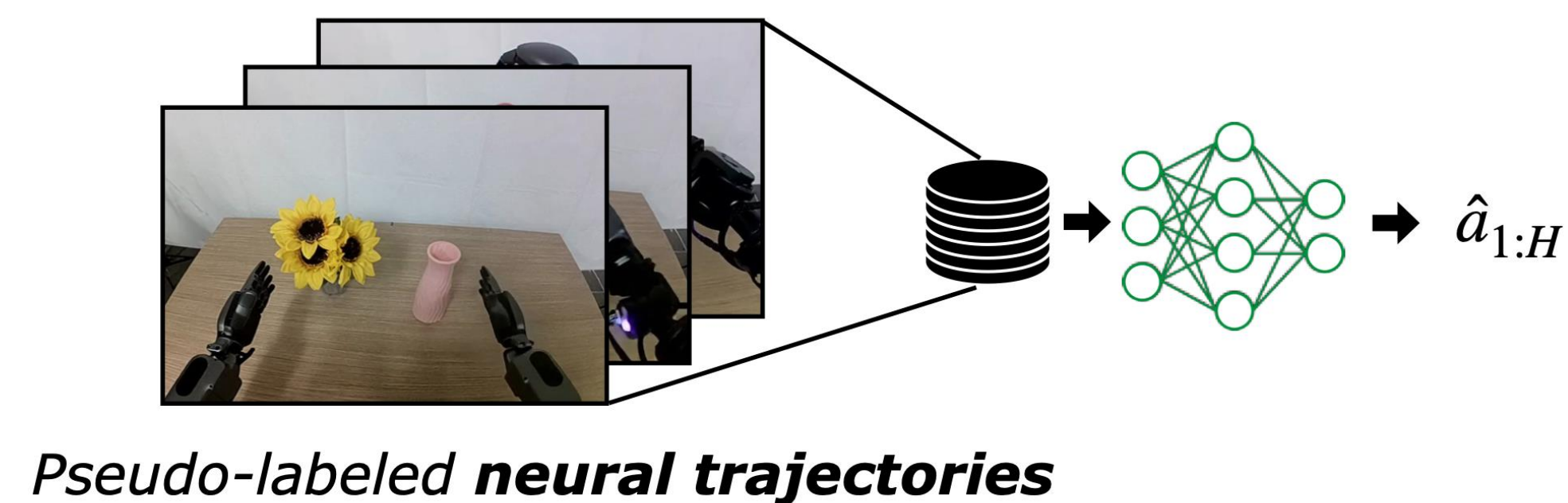


Step 3. Label Pseudo Actions



Automatically Labeled Pseudo Actions

Step 4. Visuomotor Policy Training



DreamGen: Data Synthesis with Video Generation Models

Three use cases with neural trajectories

1

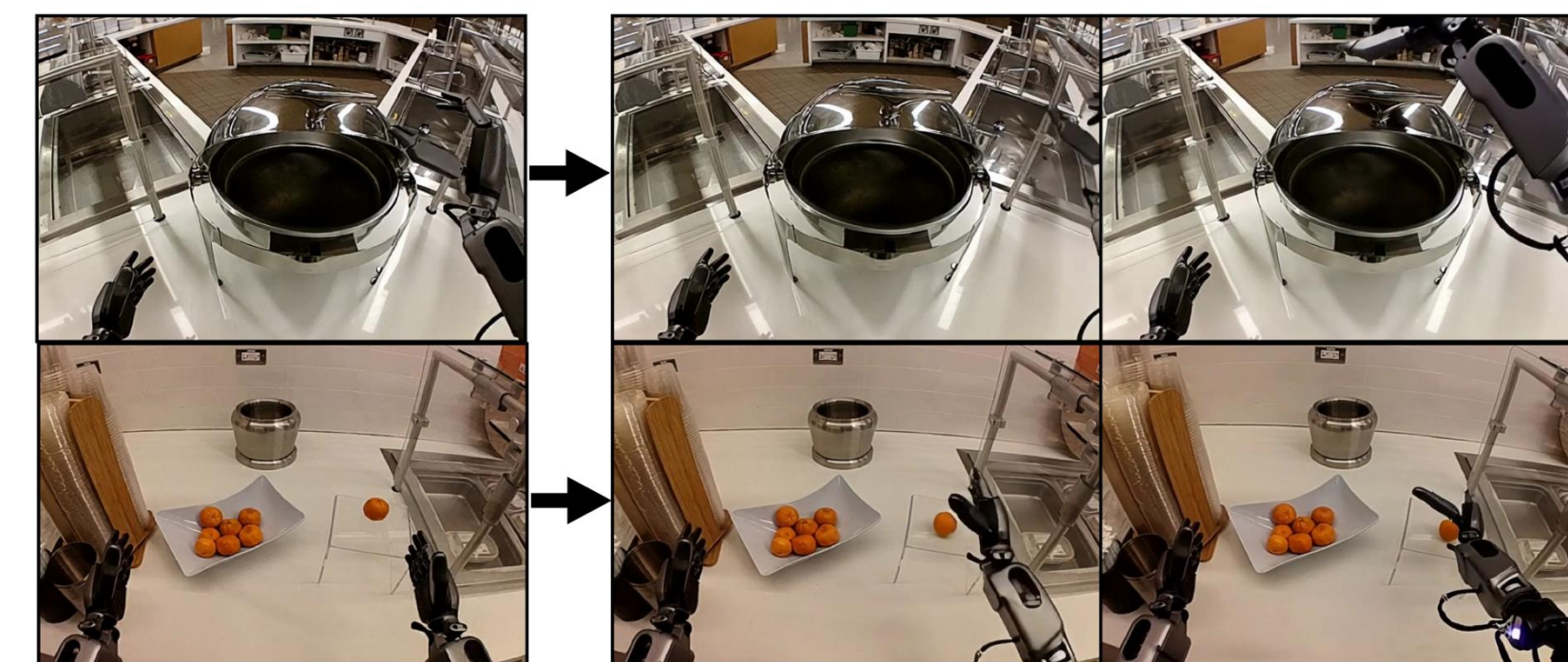
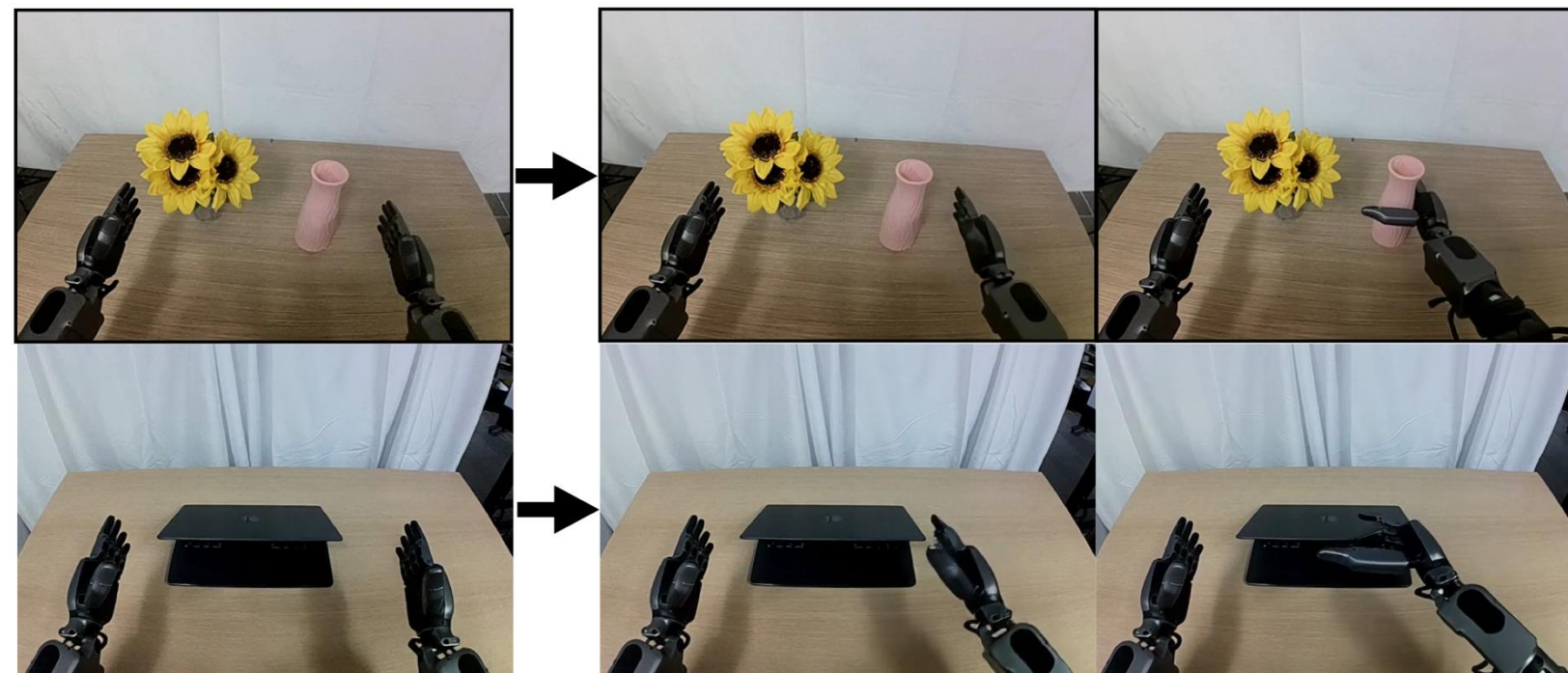
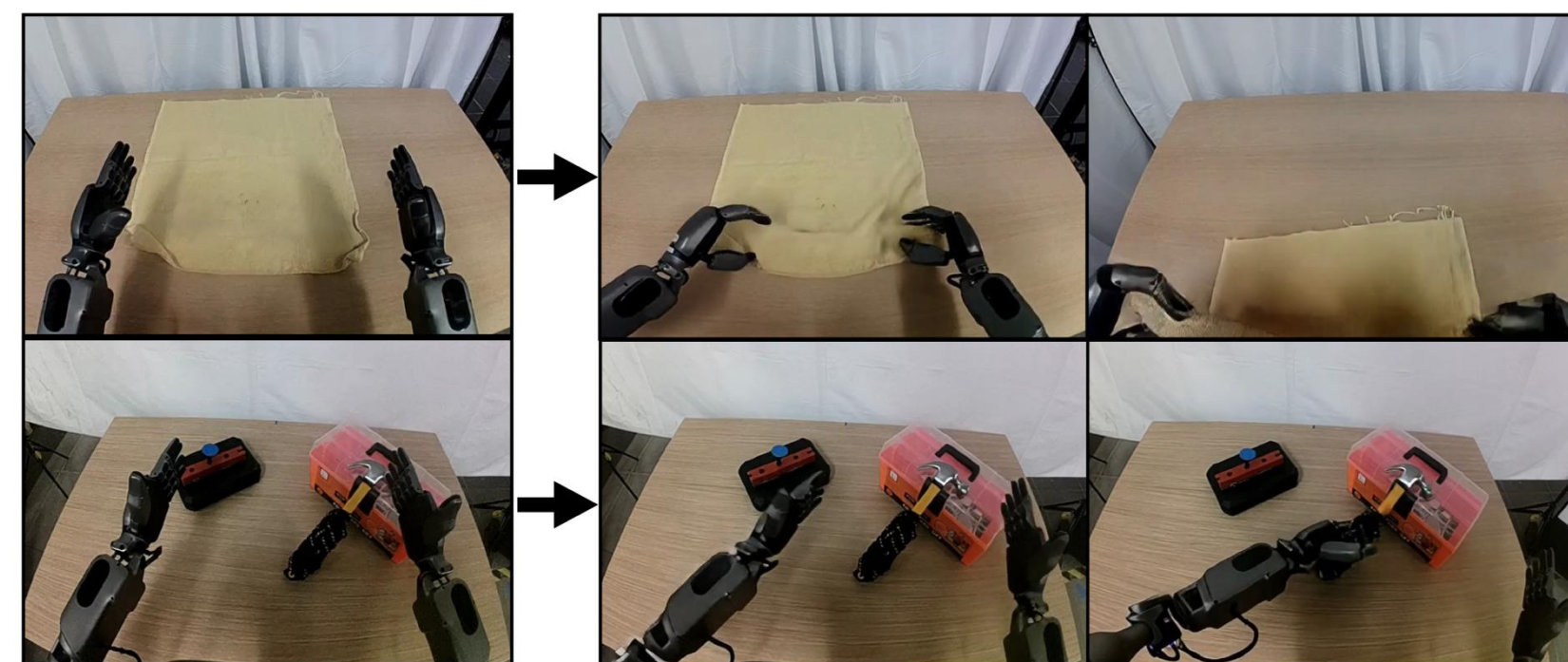
Training Data Augmentation

2

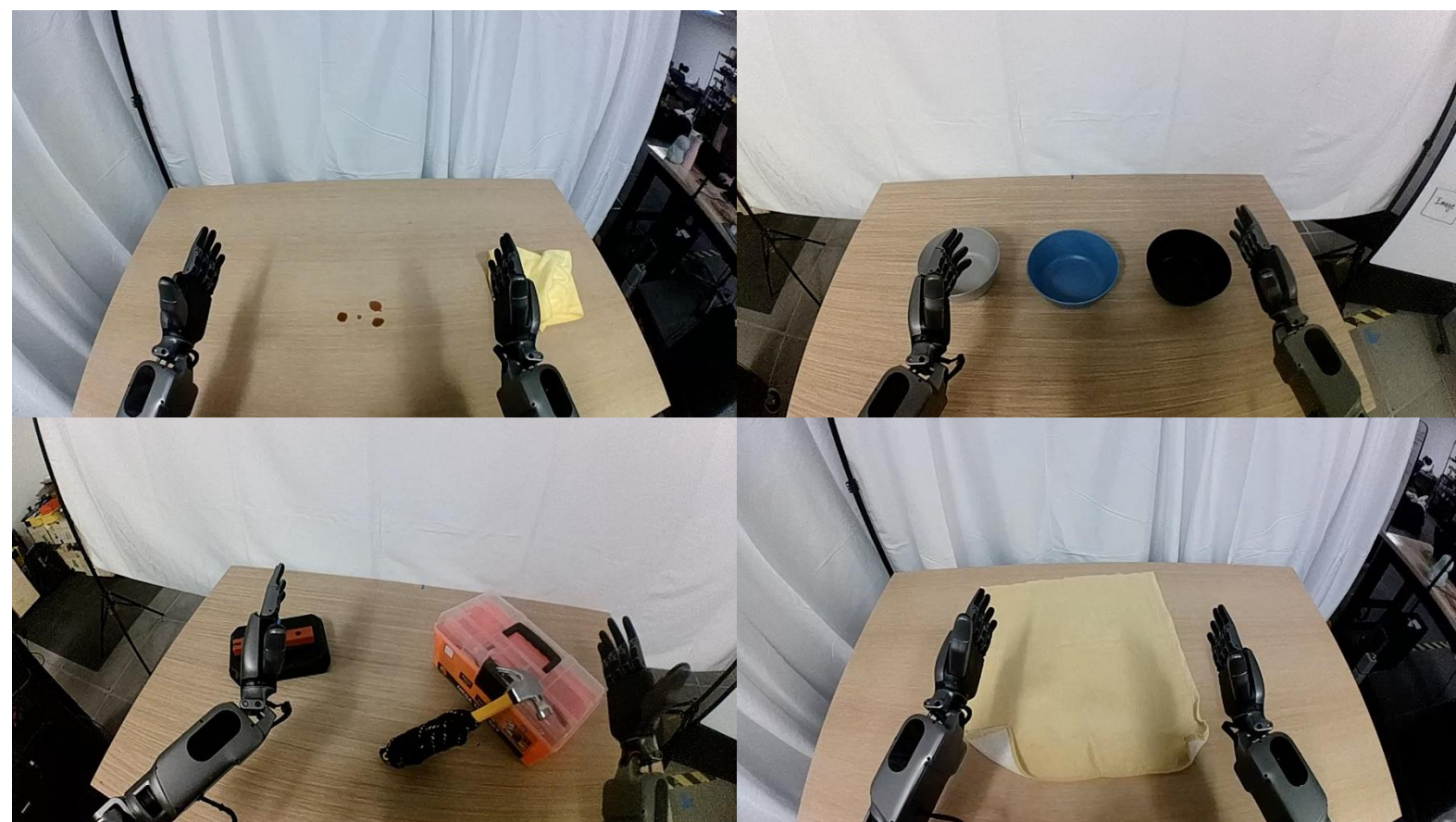
New Behavior Generalization

3

New Scene Generalization



DreamGen: Data Synthesis with Video Generation Models



10 Teleoperation Trajectories

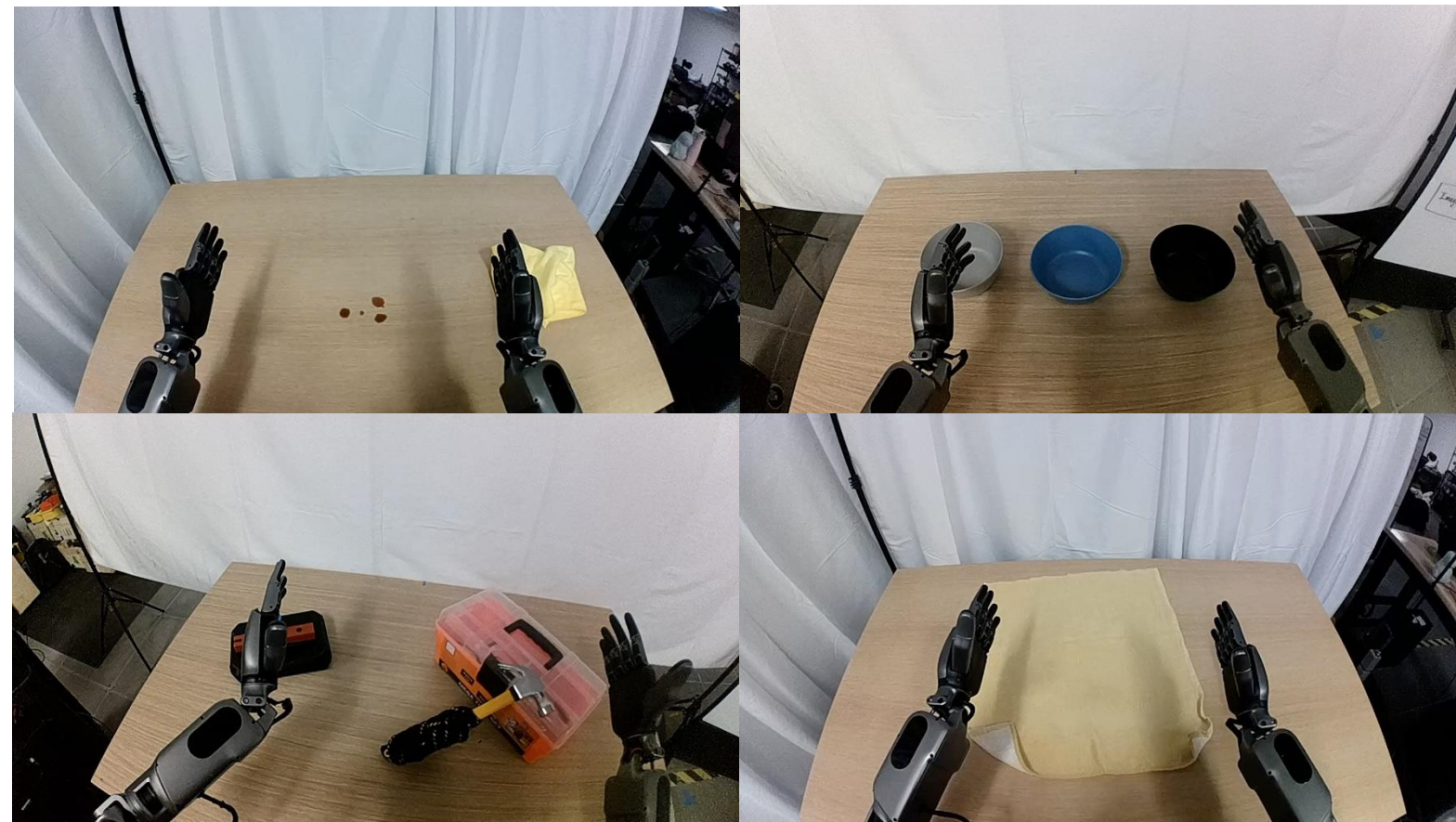


DreamGen



10k Neural Trajectories

DreamGen: Data Synthesis with Video Generation Models



10 Teleoperation Trajectories

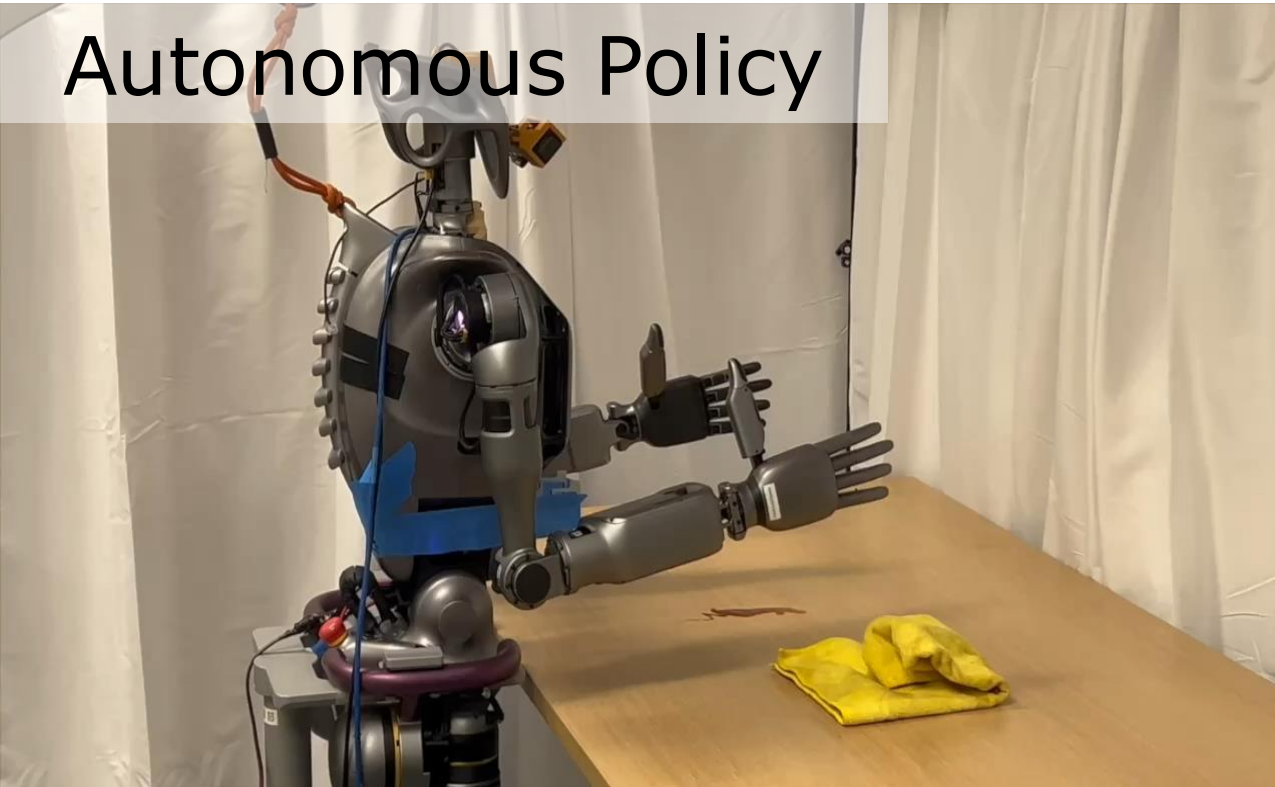
+

Co-Training

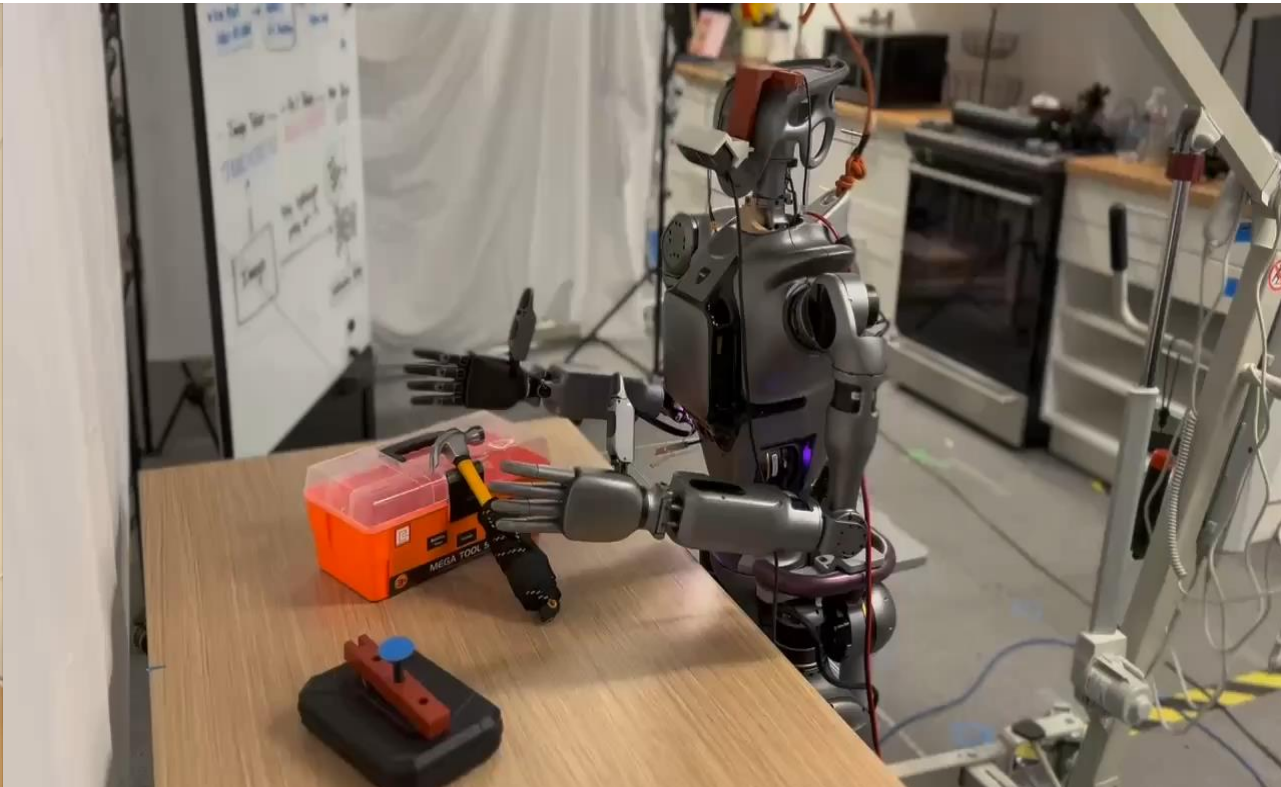


10k Neural Trajectories

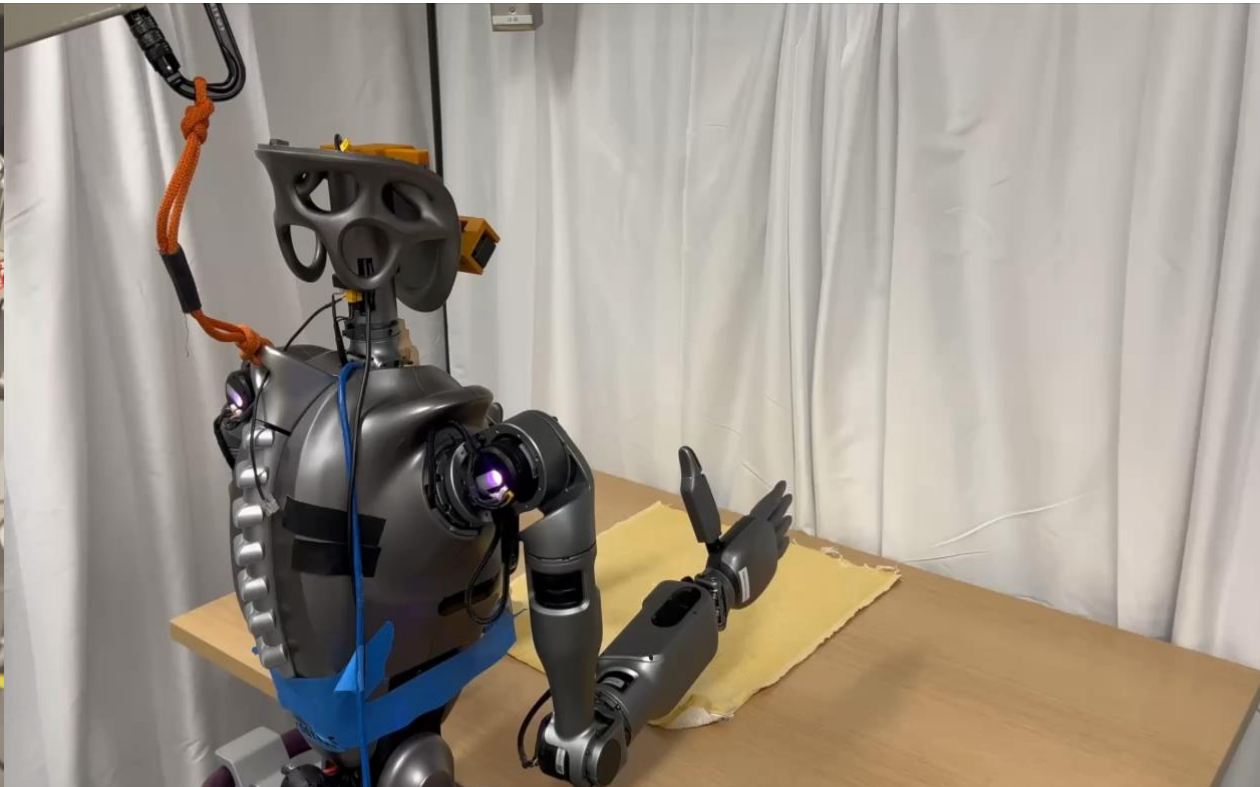
Autonomous Policy



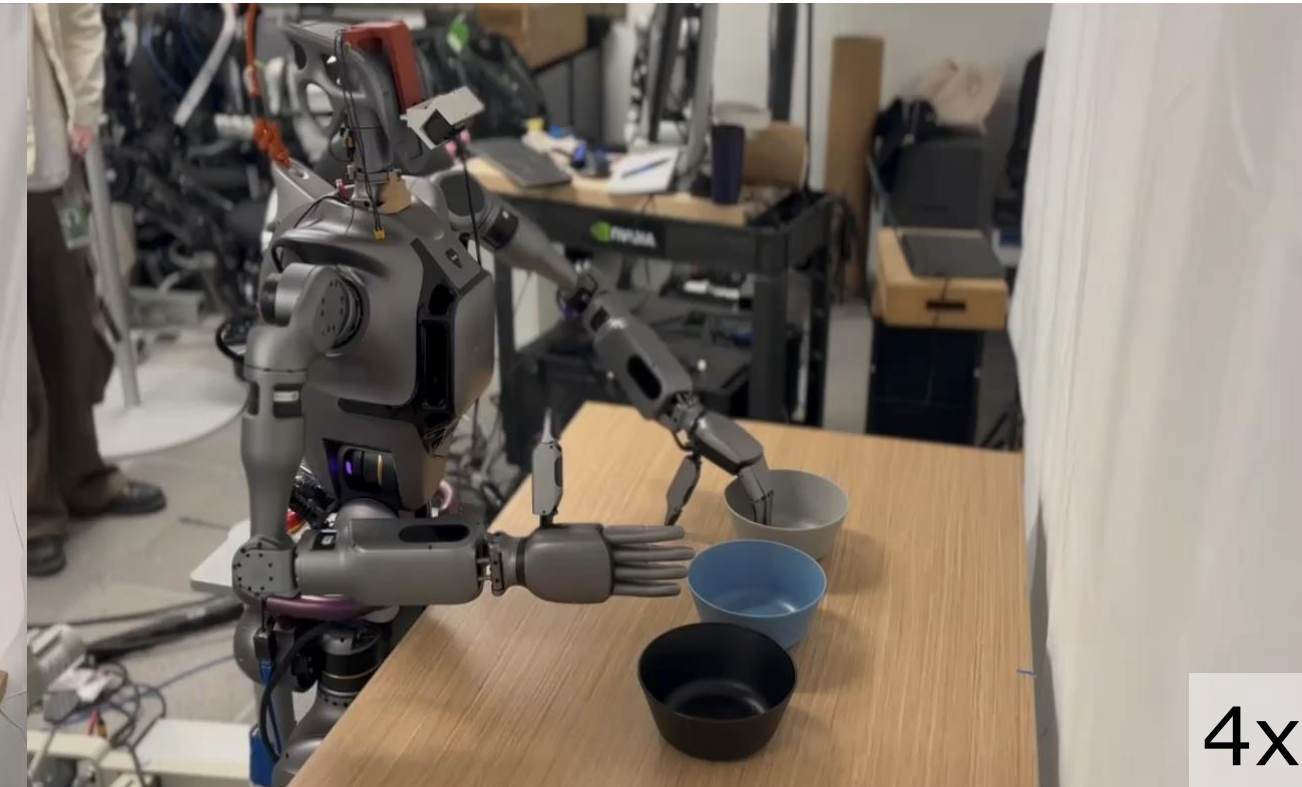
Wiping



Hammering



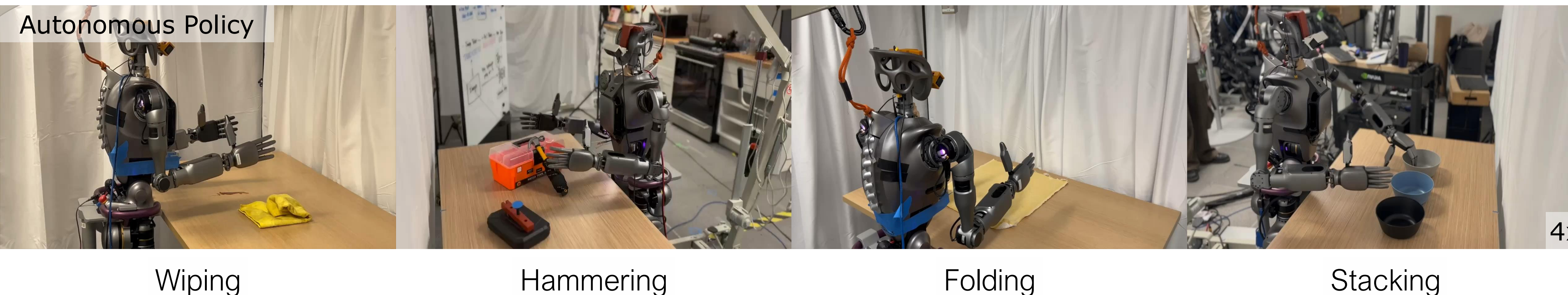
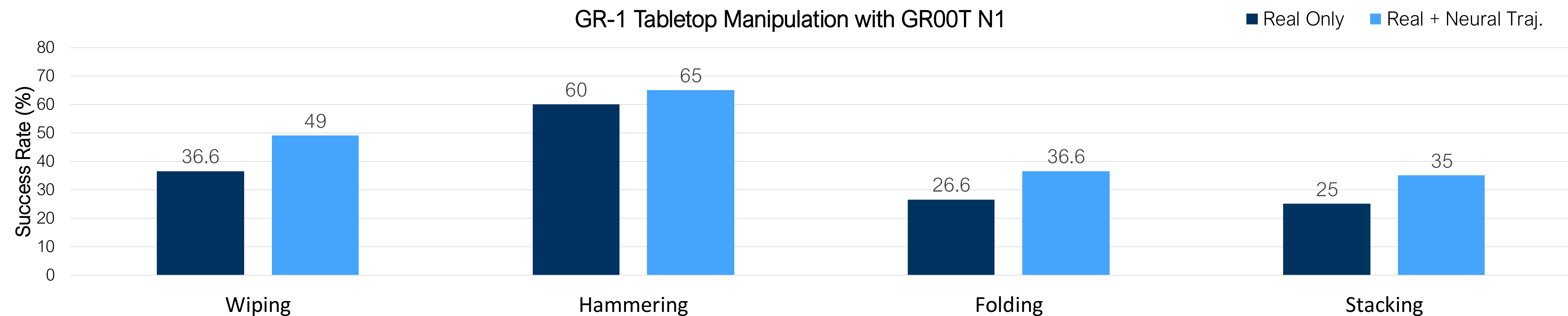
Folding



Stacking

4x

DreamGen: Data Synthesis with Video Generation Models



DreamGen: Data Synthesis with Video Generation Models

Three use cases with neural trajectories

1

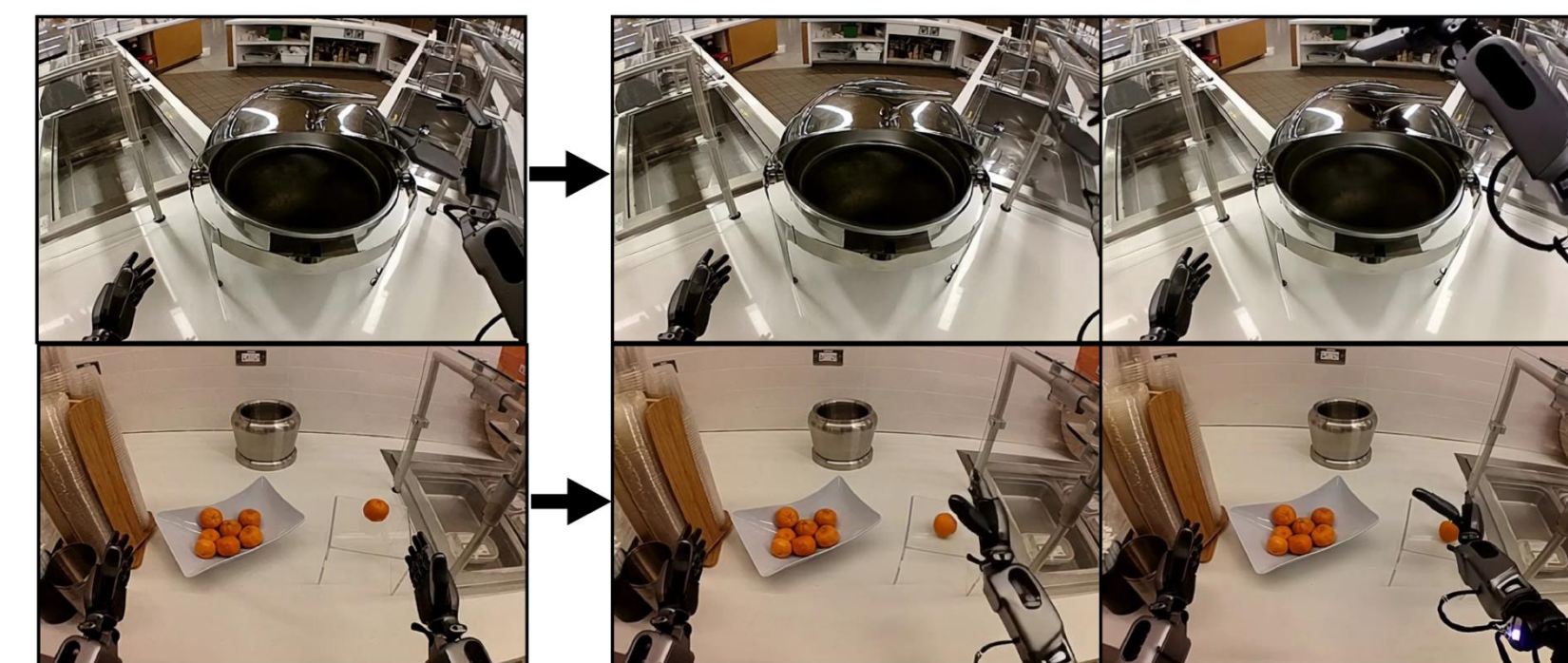
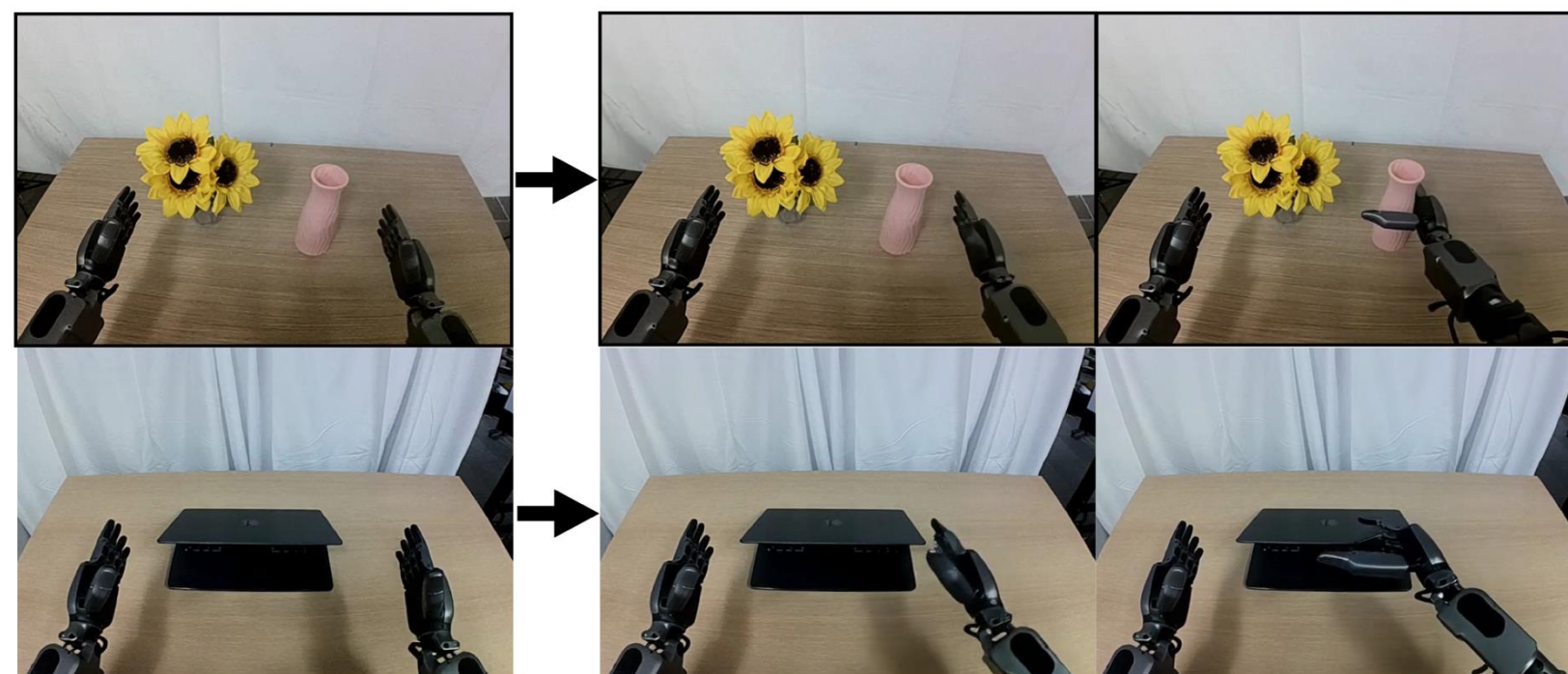
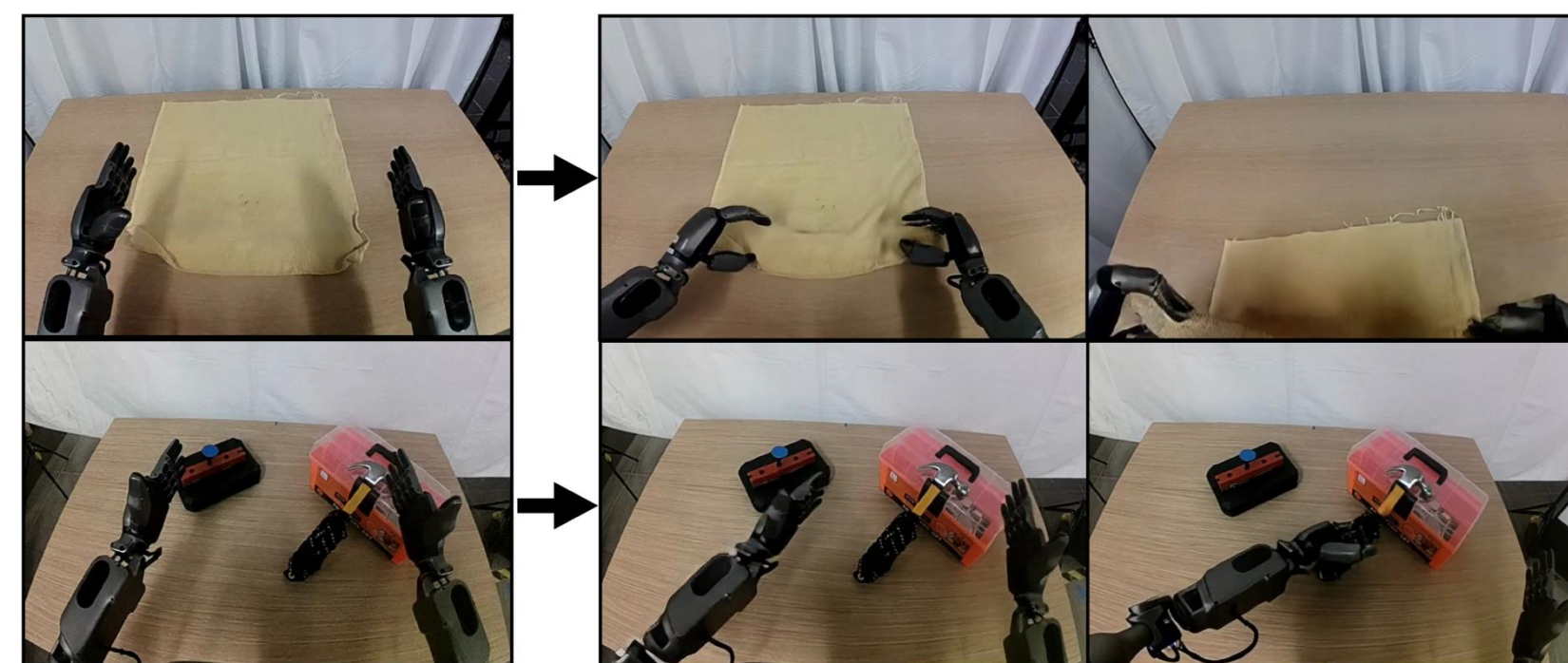
Training Data Augmentation

2

New Behavior Generalization

3

New Scene Generalization

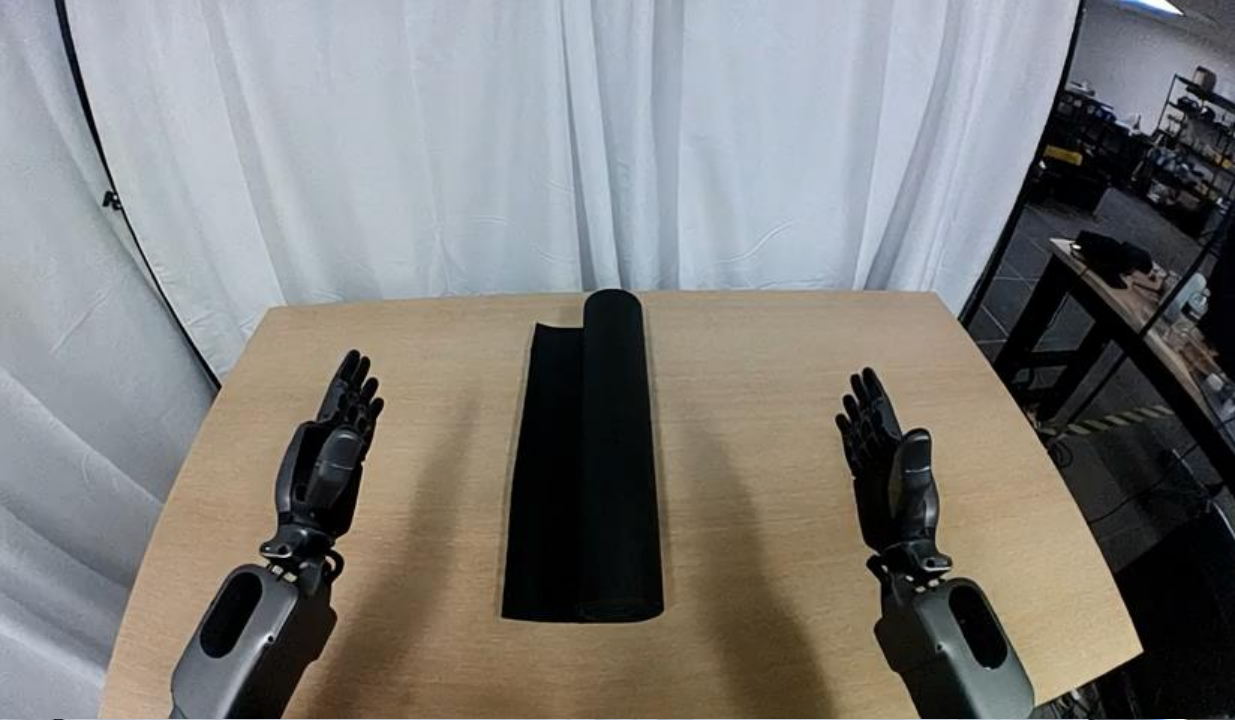




“Open the MacBook”



“Water the flowers”



“Unroll the Mat”



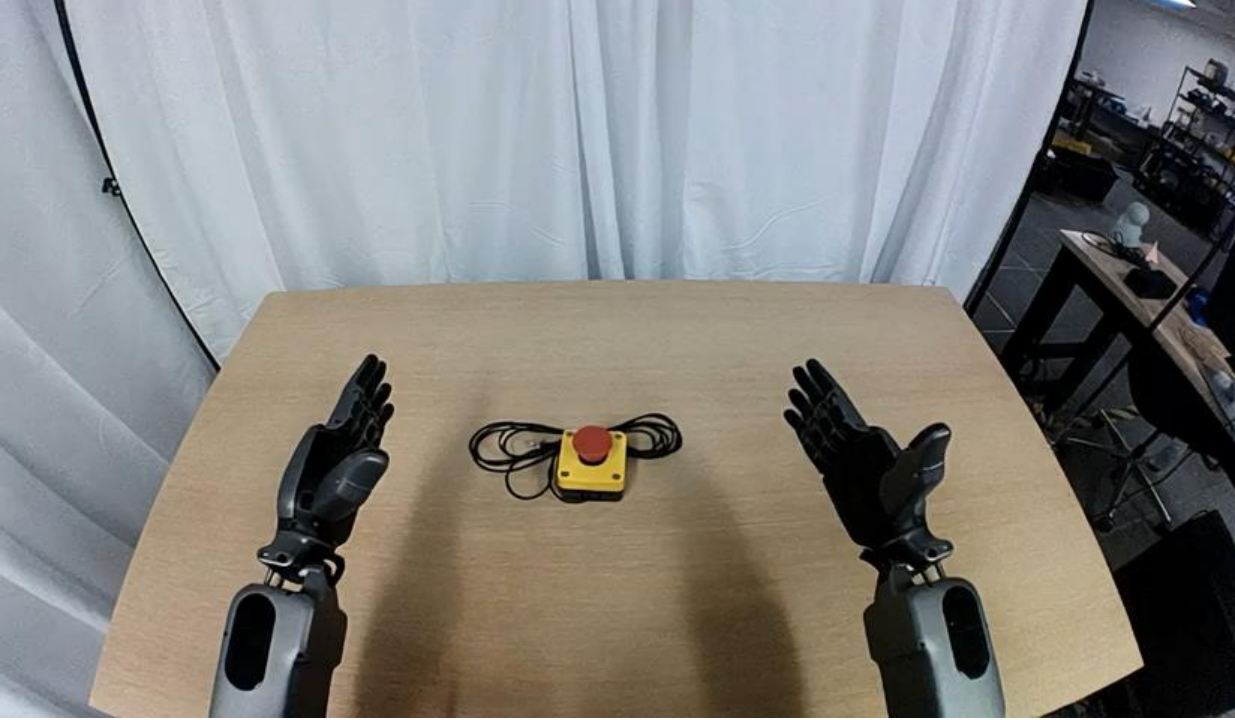
“Close the microwave”



“Type on the keyboard”



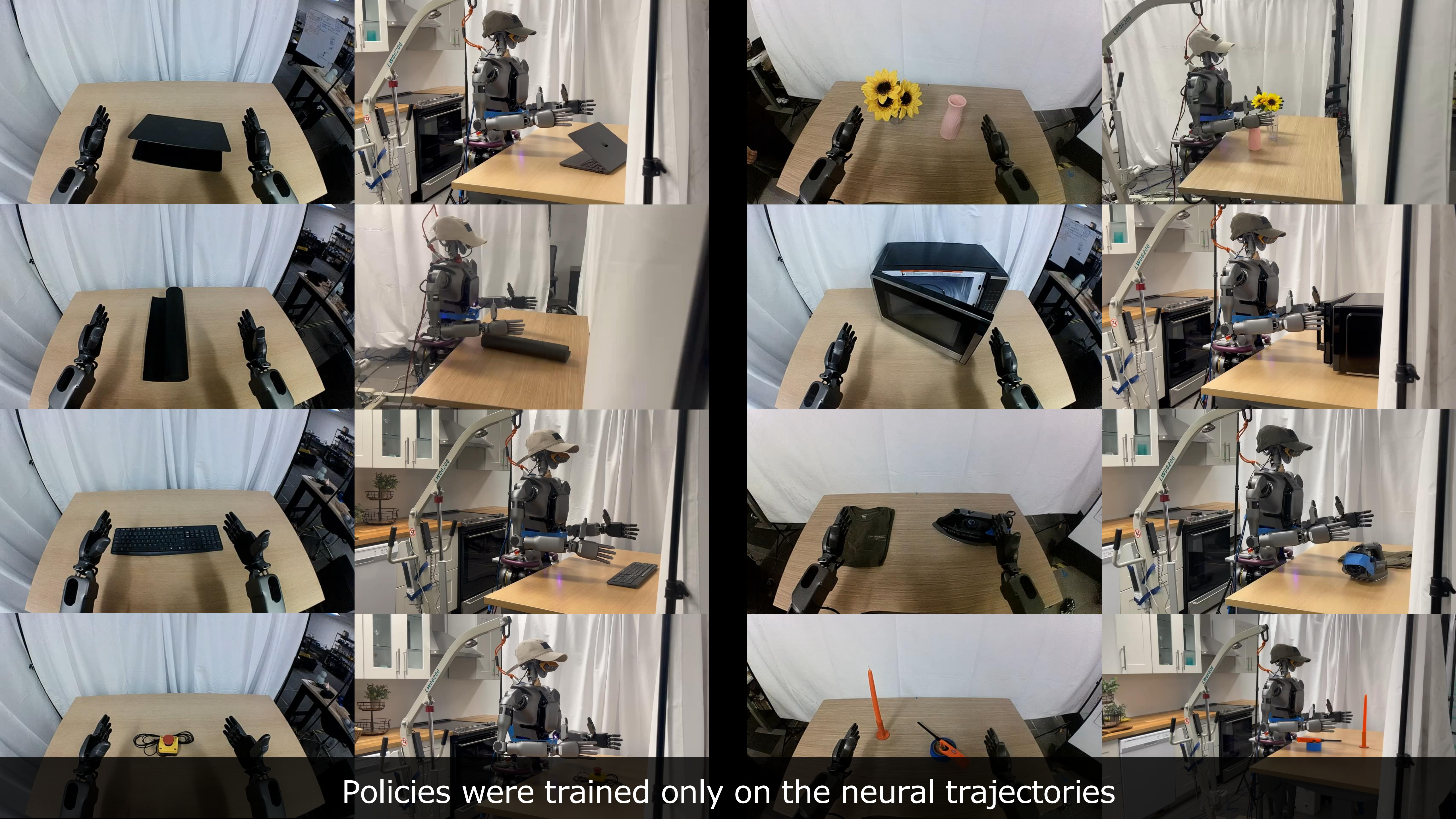
“Iron the shirt”



“Press the red button”



“Light the candle”



Policies were trained only on the neural trajectories

DreamGen: Data Synthesis with Video Generation Models

Three use cases with neural trajectories

1

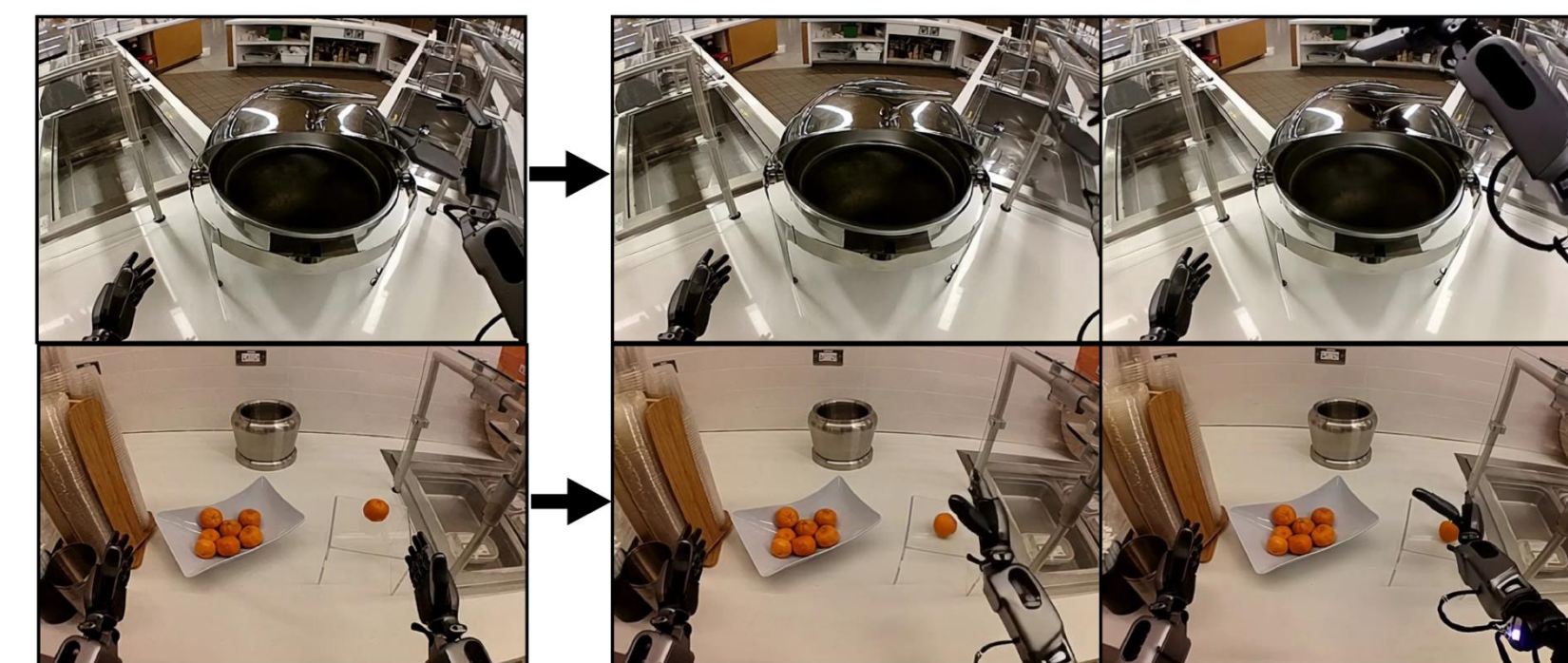
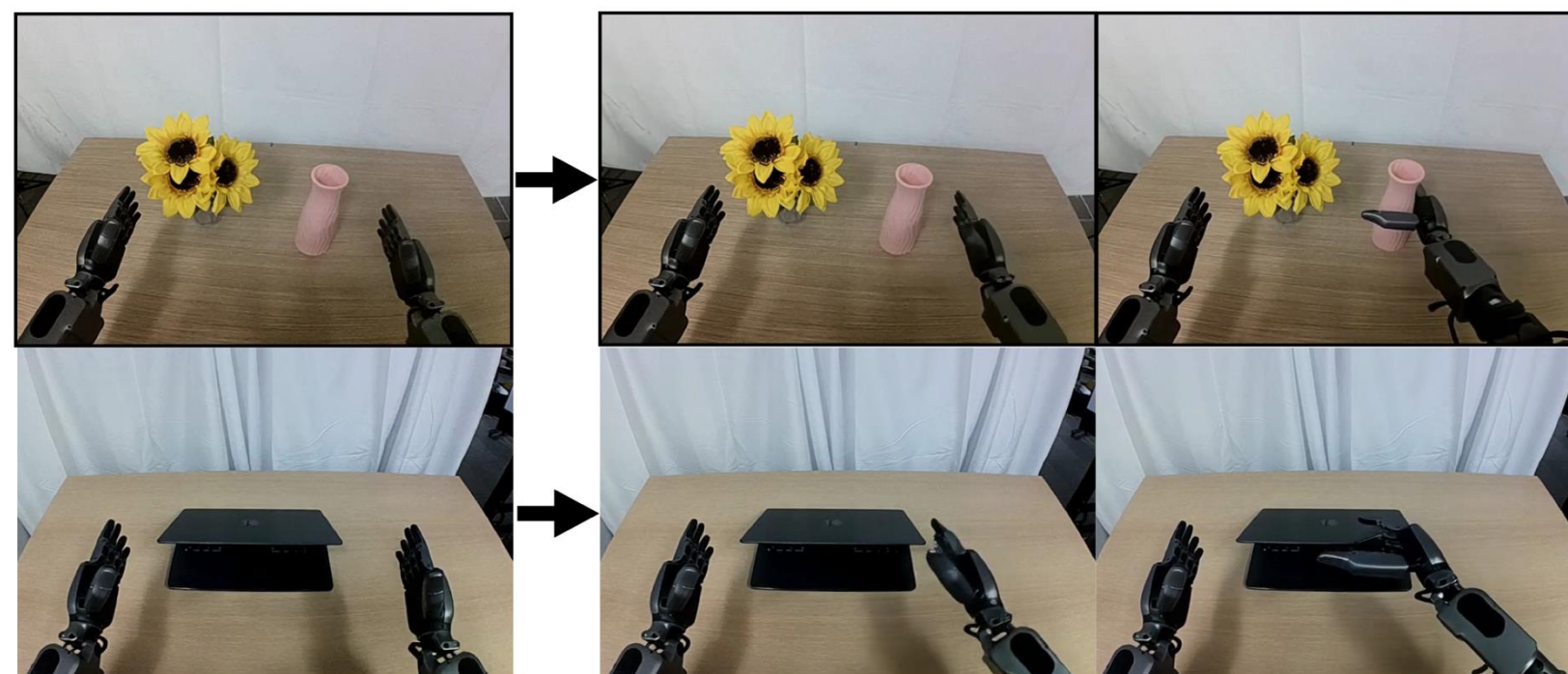
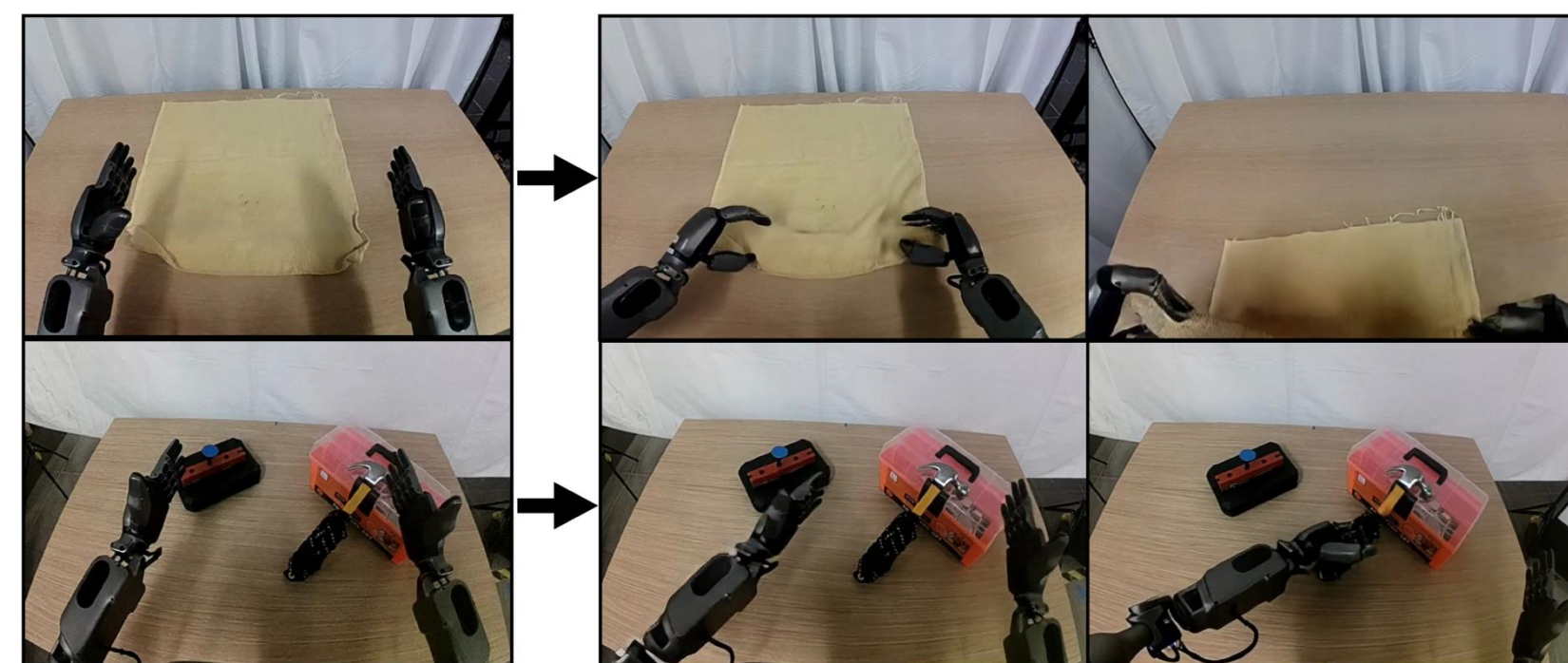
Training Data Augmentation

2

New Behavior Generalization

3

New Scene Generalization





“Weight the orange”



“Put the pear in the basket”



“Put the tangerine
in the bowl”



“Throw the cup away”



“Lift the bucket”



“Open the lid”



“Water the flowers”



“Close the lid”



Policies were trained only on the neural trajectories

Co-Training: Learning across the Data Pyramid

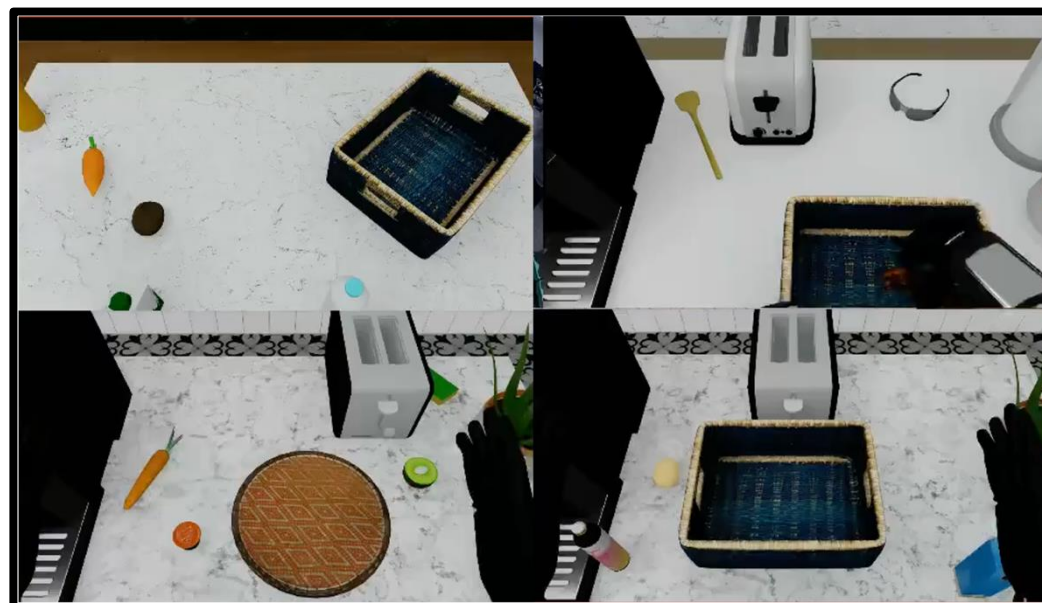
**Real-Robot
Data (1x)**

Human Teleoperation



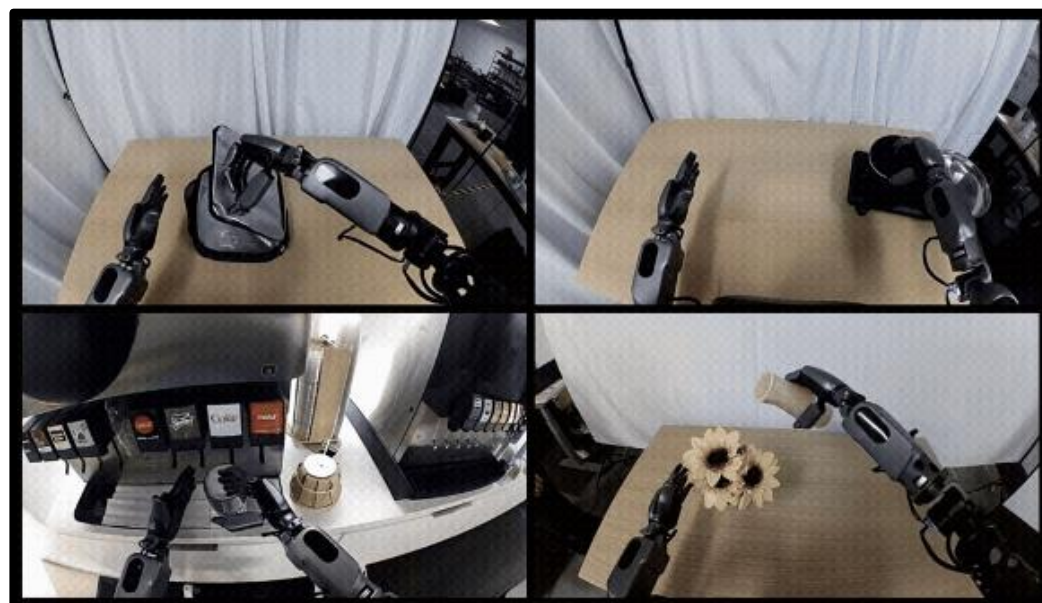
**Simulated
Trajectories
(100x)**

DexMimicGen



**Neural
Trajectories
(100x)**

DreamGen



From **Robot Hardware**

- real sensor observations and real motor actions
- costly and time-consuming to collect

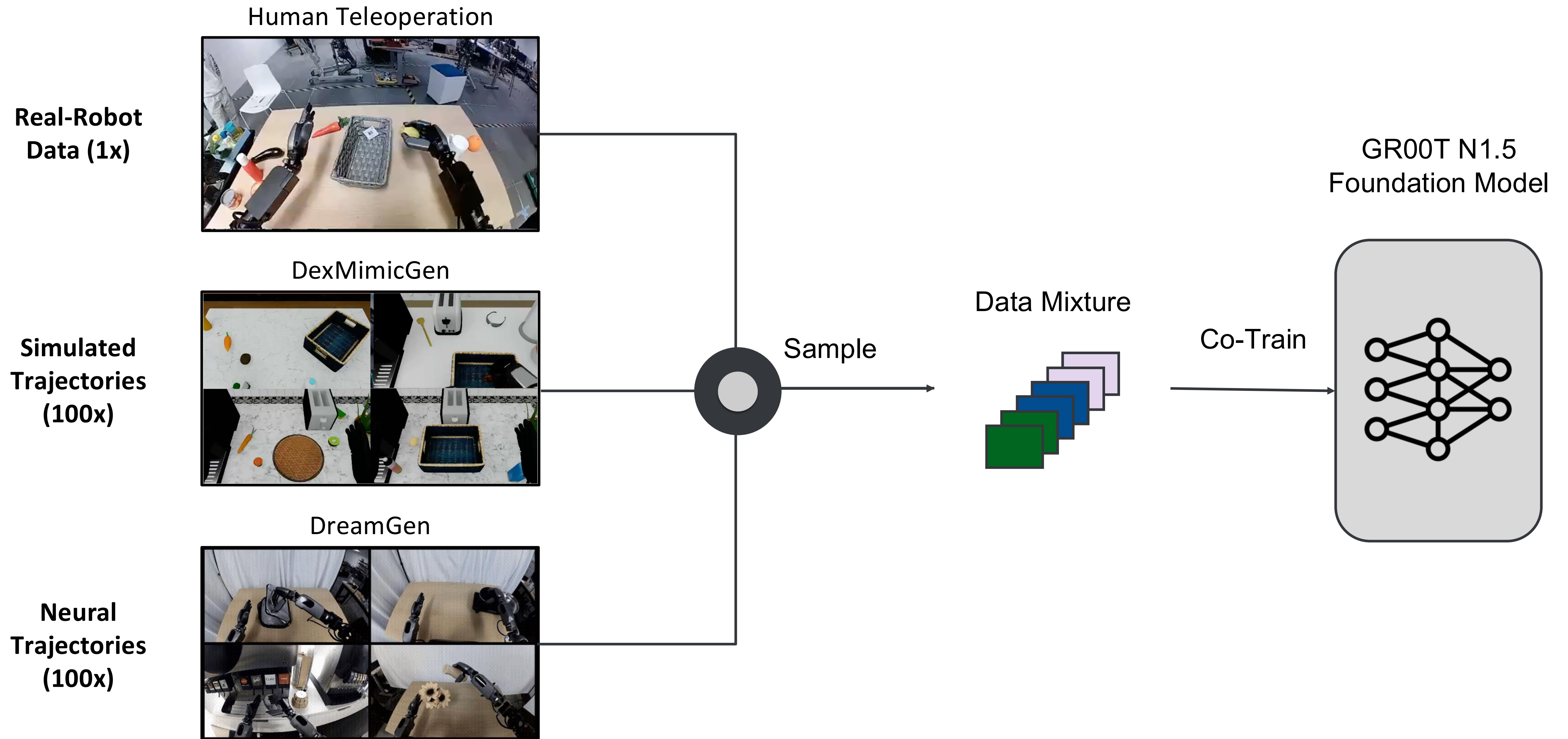
From **Physics Engines**

- rendering gaps and high-quality actions
- physically grounded and annotation-rich

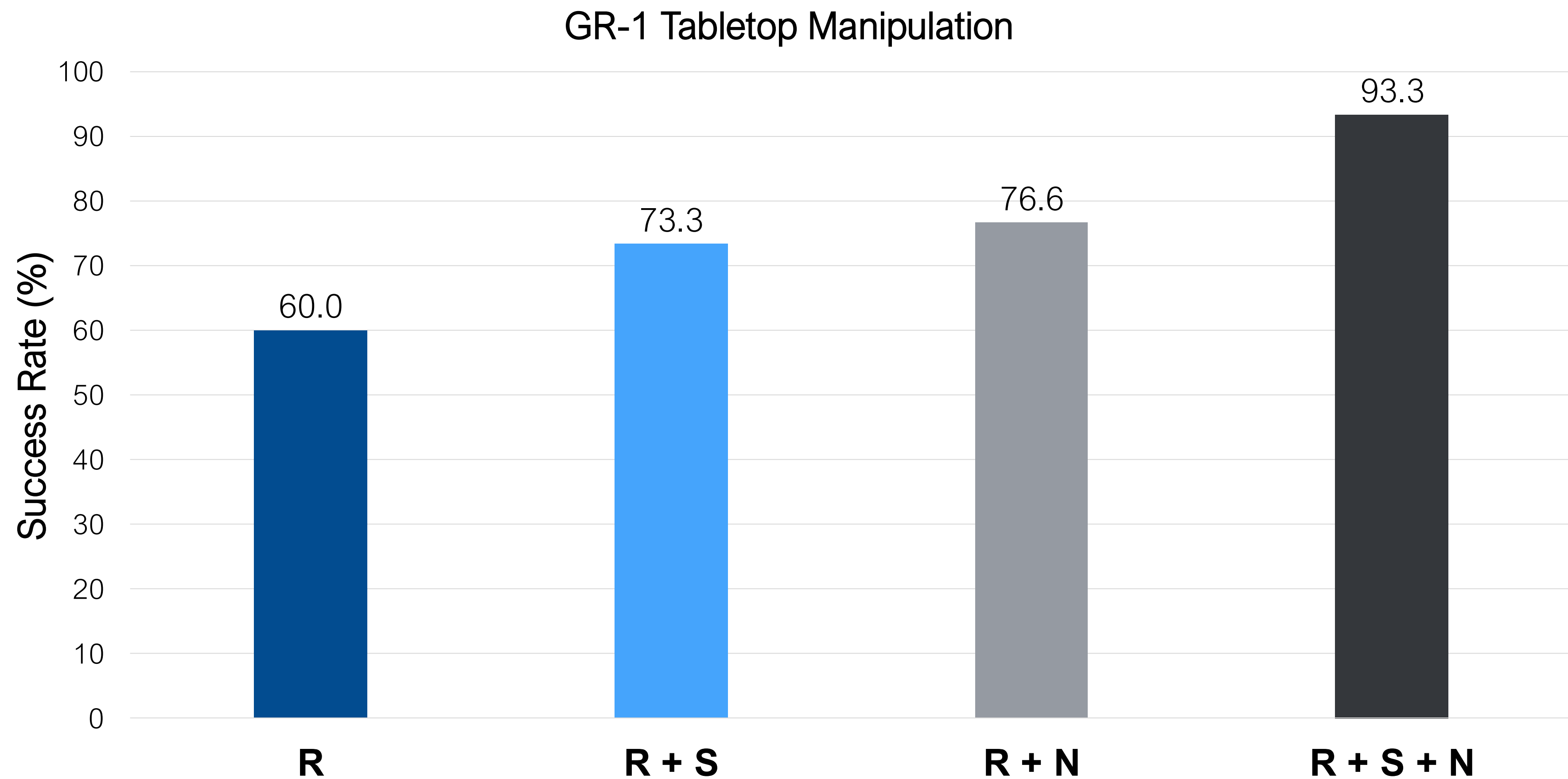
From **Video Generation Models**

- realistic observations and predicted (noisy) actions
- open-ended and simulating complex physics effects

Co-Training: Learning across the Data Pyramid



Co-Training: Learning across the Data Pyramid



R: Real-Robot Data

S: Simulated Trajectories

N: Neural Trajectories



Some Closing Thoughts

Humanoid loco-manipulation is the killer application.

- Colossal challenges in mechanics, compute & power, sensing, control, and cognition...
- Fewer humanoid dancing/Kung Fu demos. Let's put the robots to work!

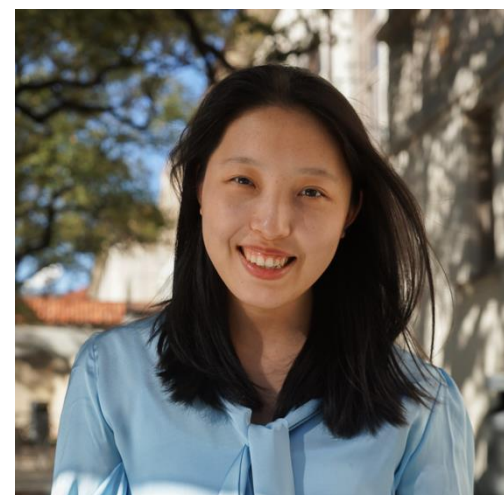
“People who are really serious about robot learning should make their own robot hardware.” (tweet from @yukez)

- Today's humanoid hardware makes the AI problems artificially harder.
- Dexterous hands, short of tactile sensing, remain the Achilles' heel of humanoid robots.

A long yet bright path toward making humanoids widely adopted products

- We are still in the infancy of humanoid commercialization, yet progress is accelerating.
- Long-term societal need and huge total addressable market (TAM); the question is not *if*, but *when*.

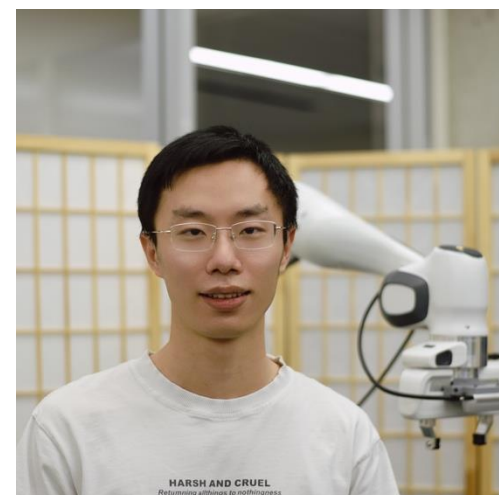
Acknowledgement



Huihan Liu



Tairan He



Zhenyu Jiang



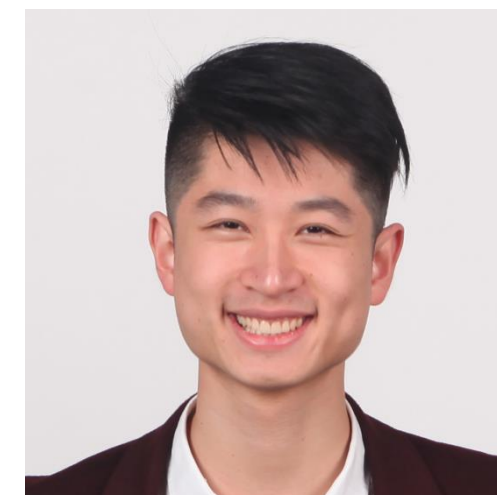
Seonghyeon Ye



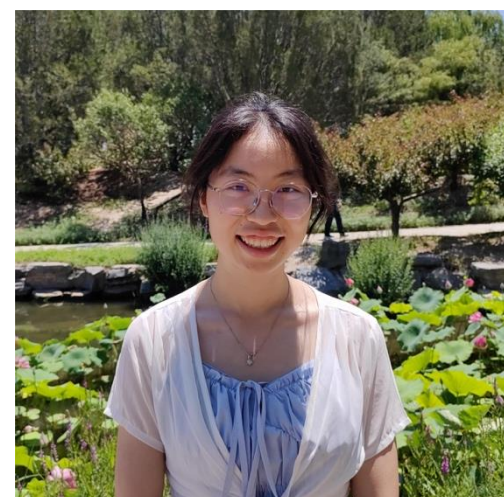
Mingyo Seo



Soroush Nasiriany



Kevin Lin



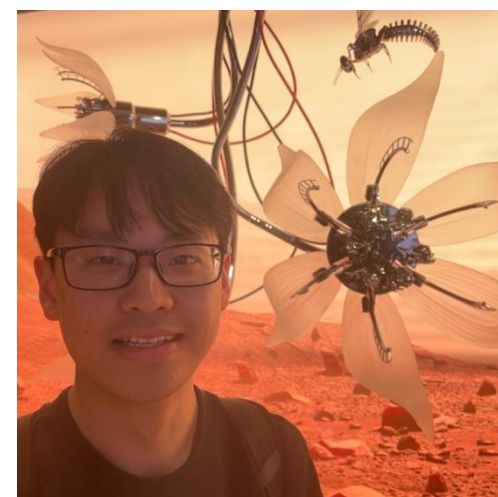
Jinhan Li



Johan Bjorck



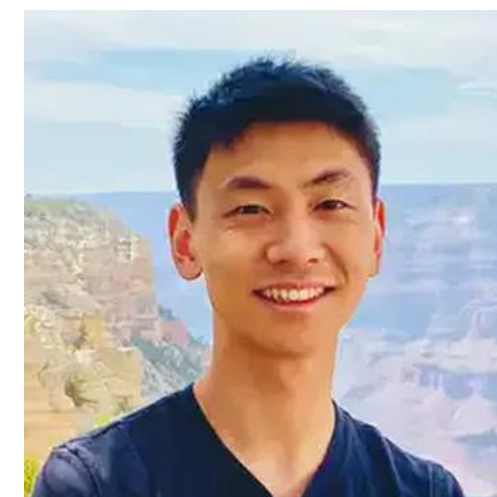
Zhengyi (Zen) Luo



Lei Yu



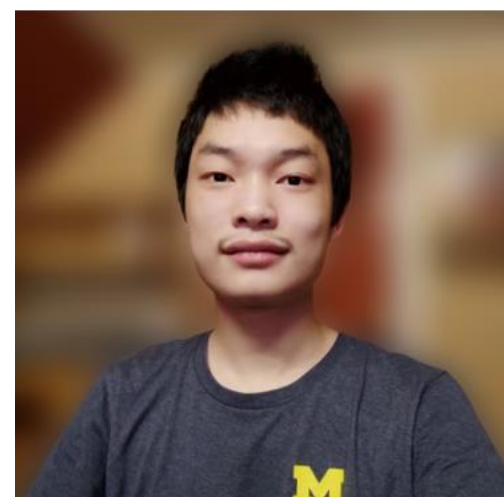
Jim Fan



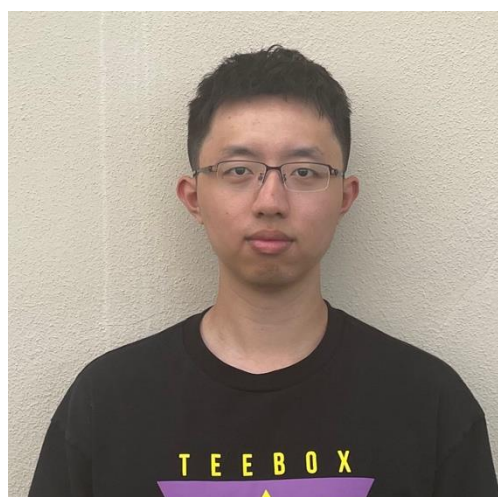
Lawrence Chen



Ajay Mandlekar



Yuqi (Leo) Xie



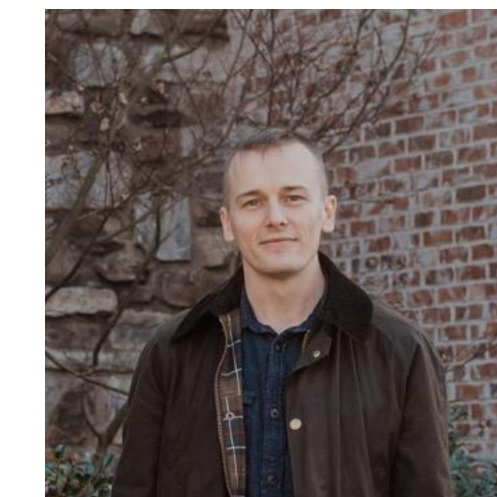
Jiannan Xiang



Abhiram Maddukuri



Zongyu Lin



Scott Reed



Joel Jang



Zhenjia Xu

Funding
Sources:

