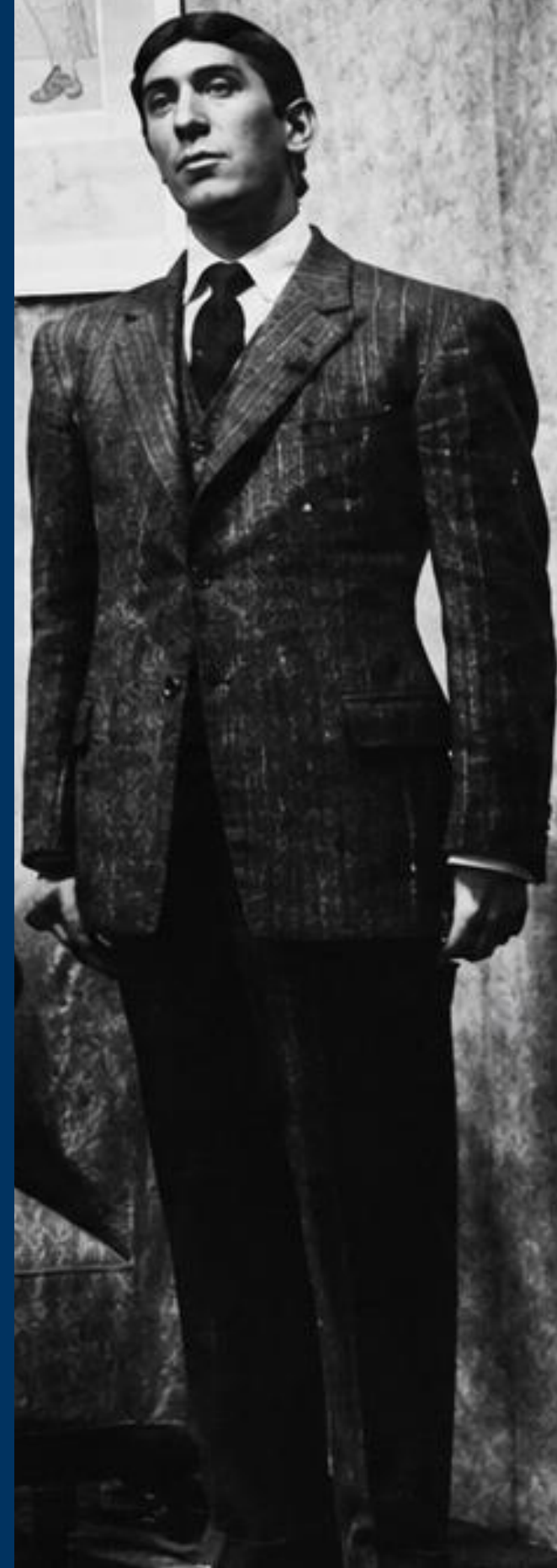


# Building Generalist Humanoid Robots

Yuke Zhu

June 3<sup>rd</sup>, 2026





# The Humanoid Robot Timeline

## “The Robot Dreams”



Rossum's Universal Robots  
(1920)

## “The Visionary Prototypes”



Honda Asimo  
(2000)

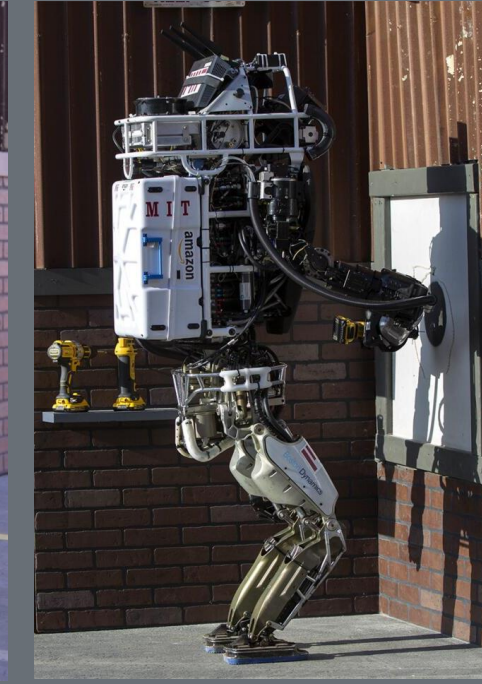


Sony QRIO  
(2003)

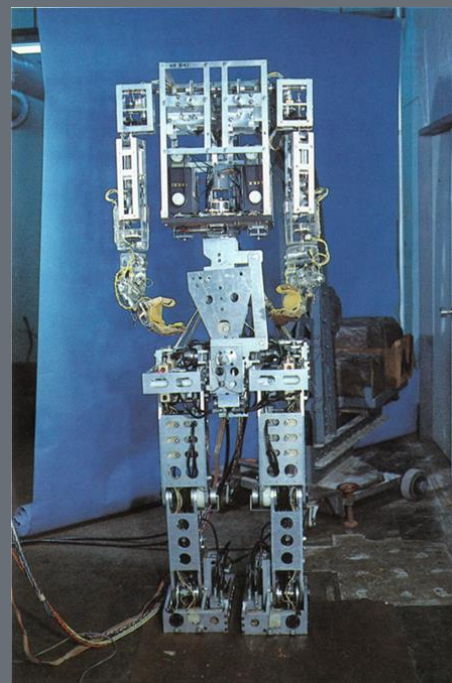
## “The Reality Check”



DARPA Robotics Challenge  
(2013-2015)



## “The Pioneer Attempts”



WABOT-1  
(1970-1973)



WABOT-2  
(1980-1984)

## “The Social Companions”

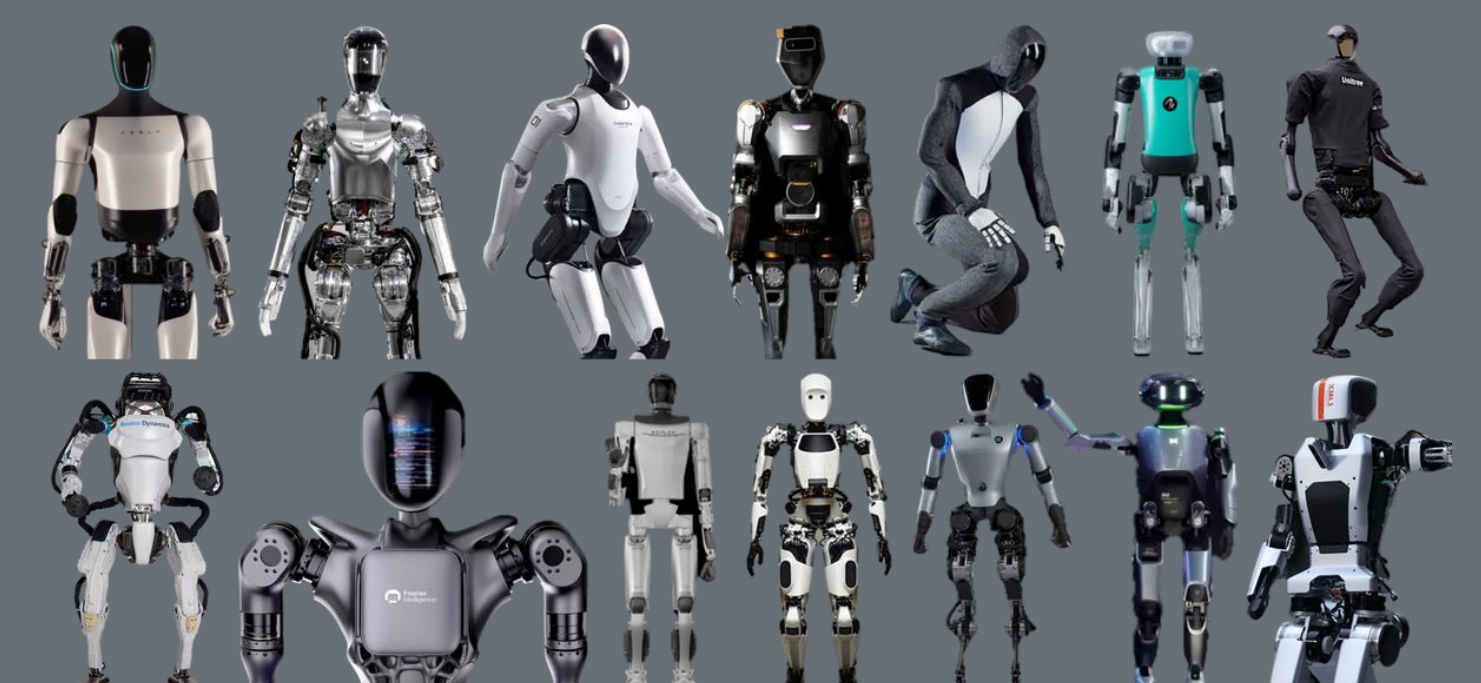


NAO  
(2008)



Pepper  
(2014)

## “The Cambrian Explosion”

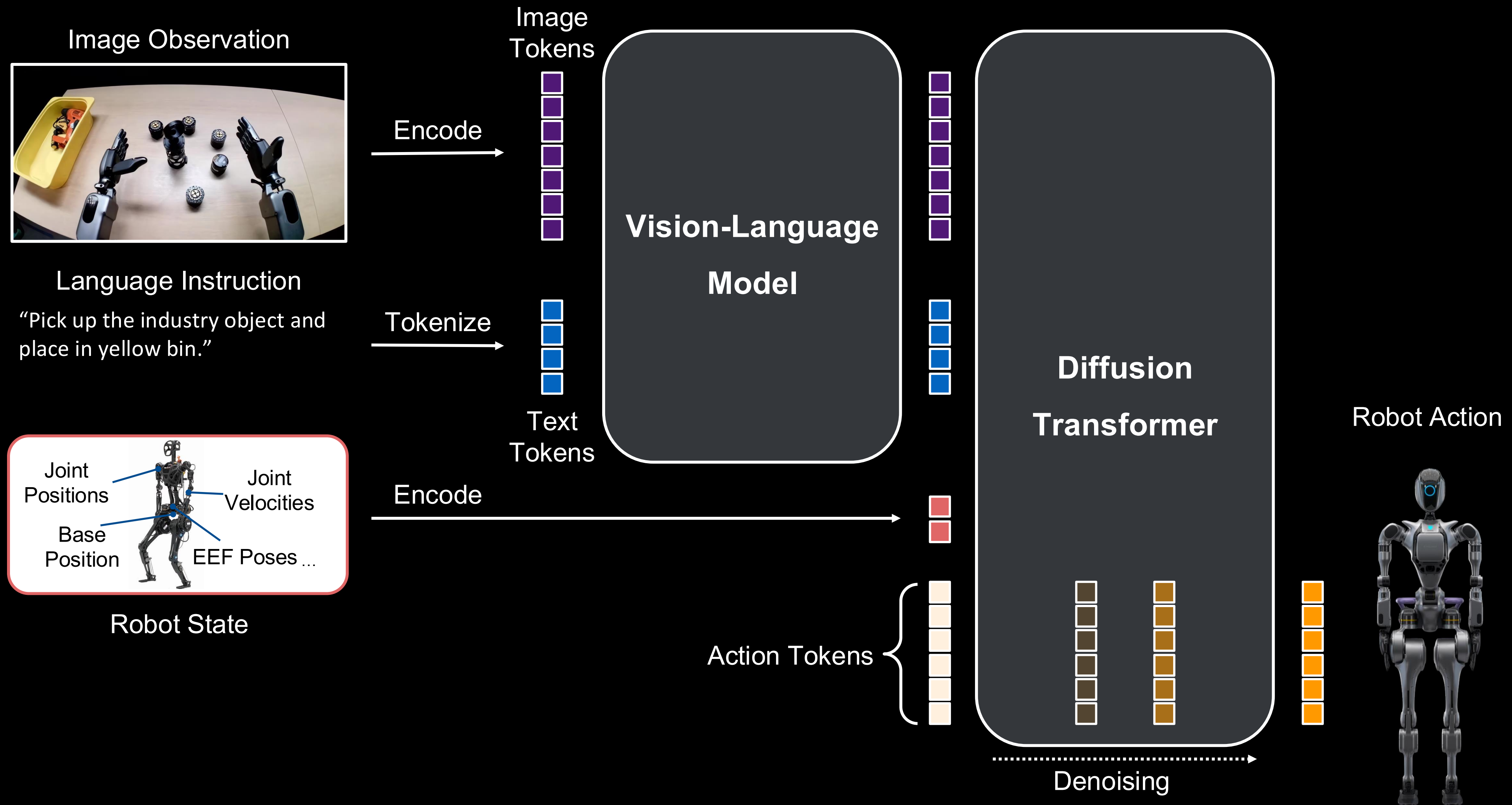


Recent Humanoid Wave  
(circa 2022-Now)

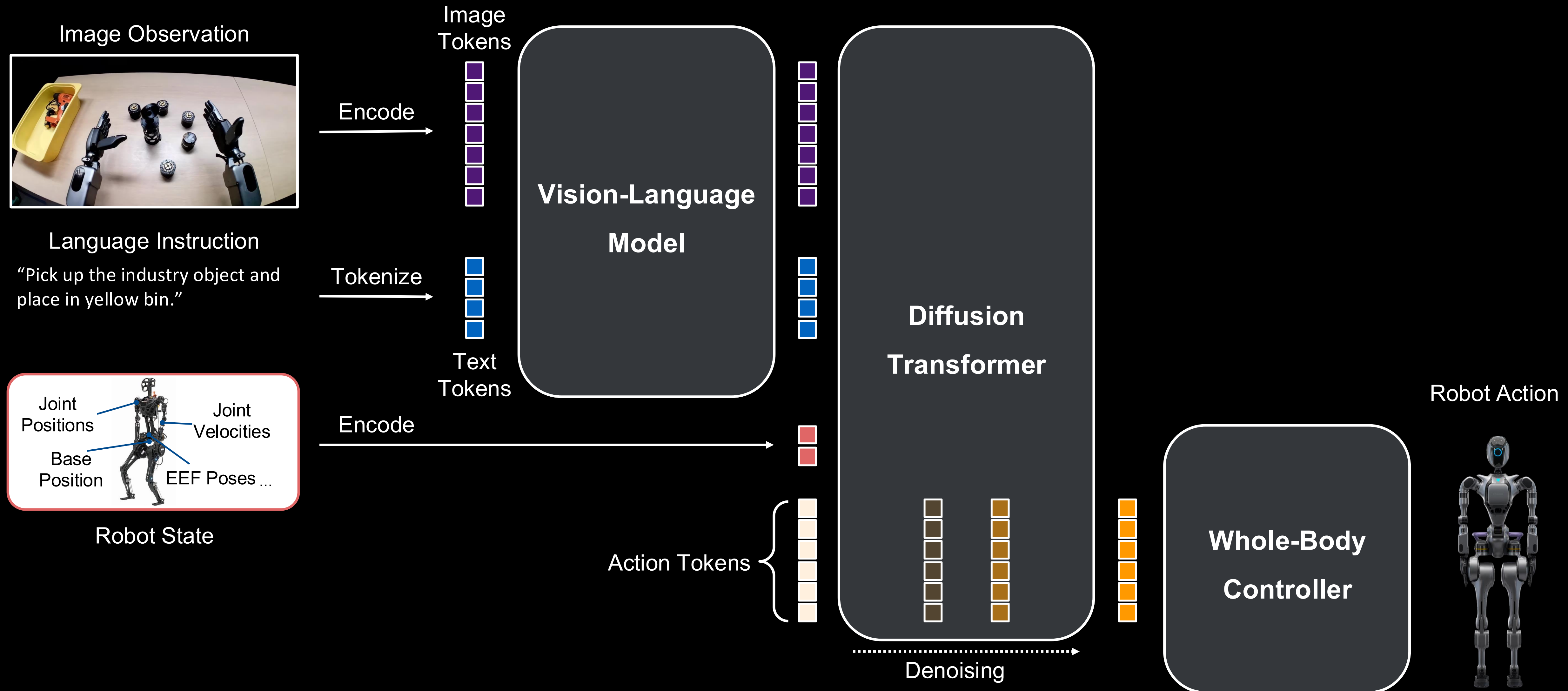
GTC 2024

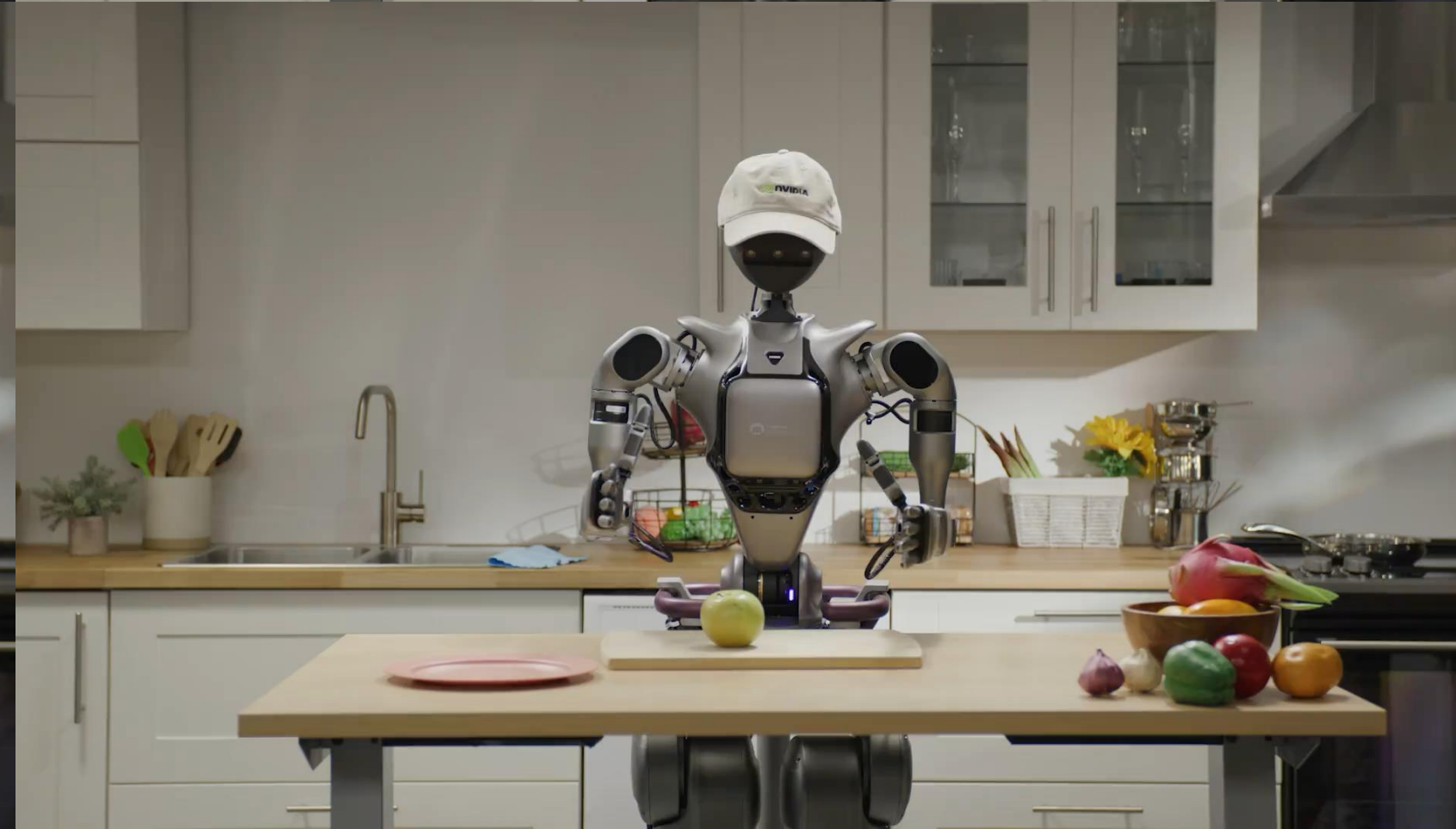
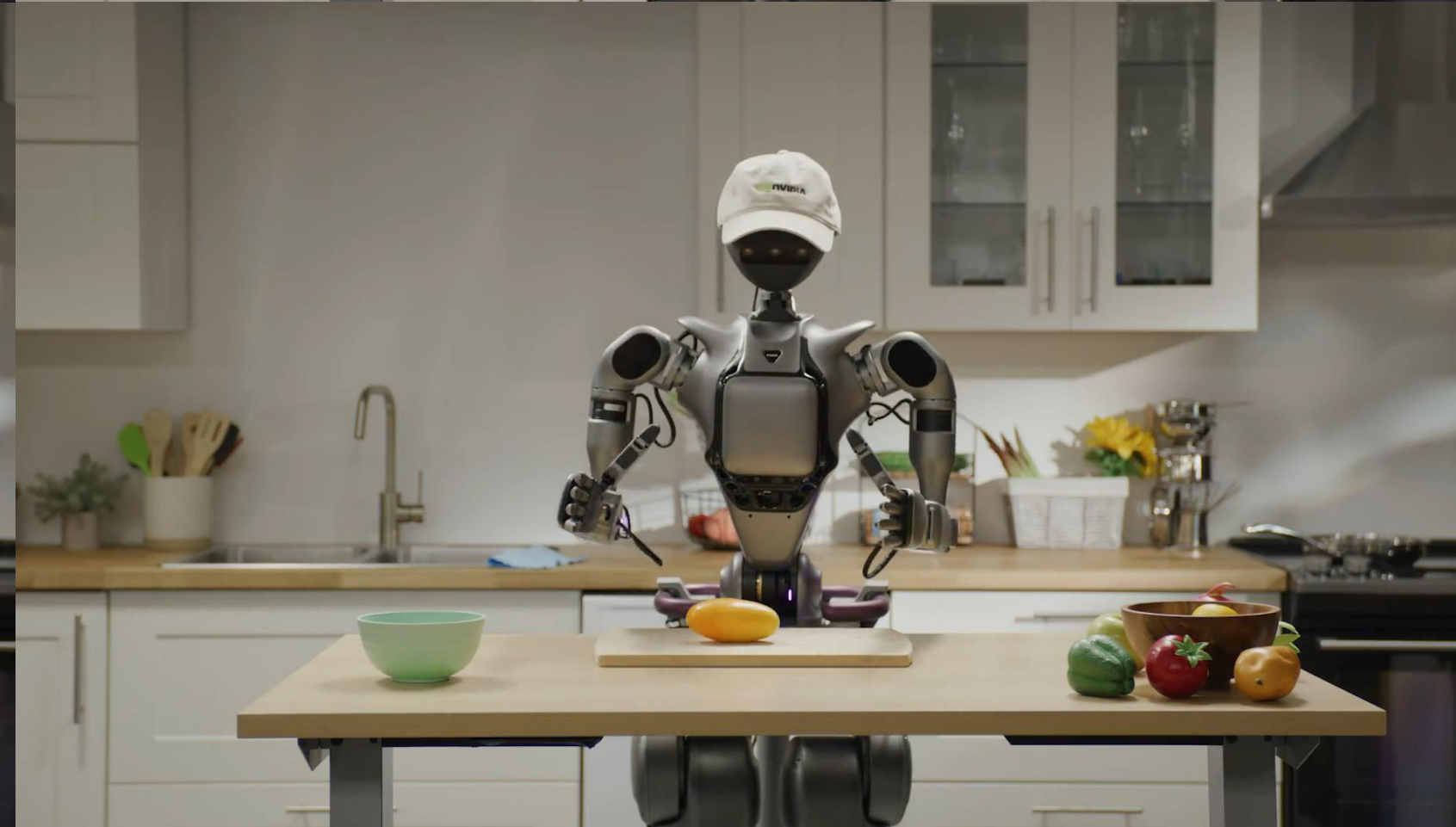
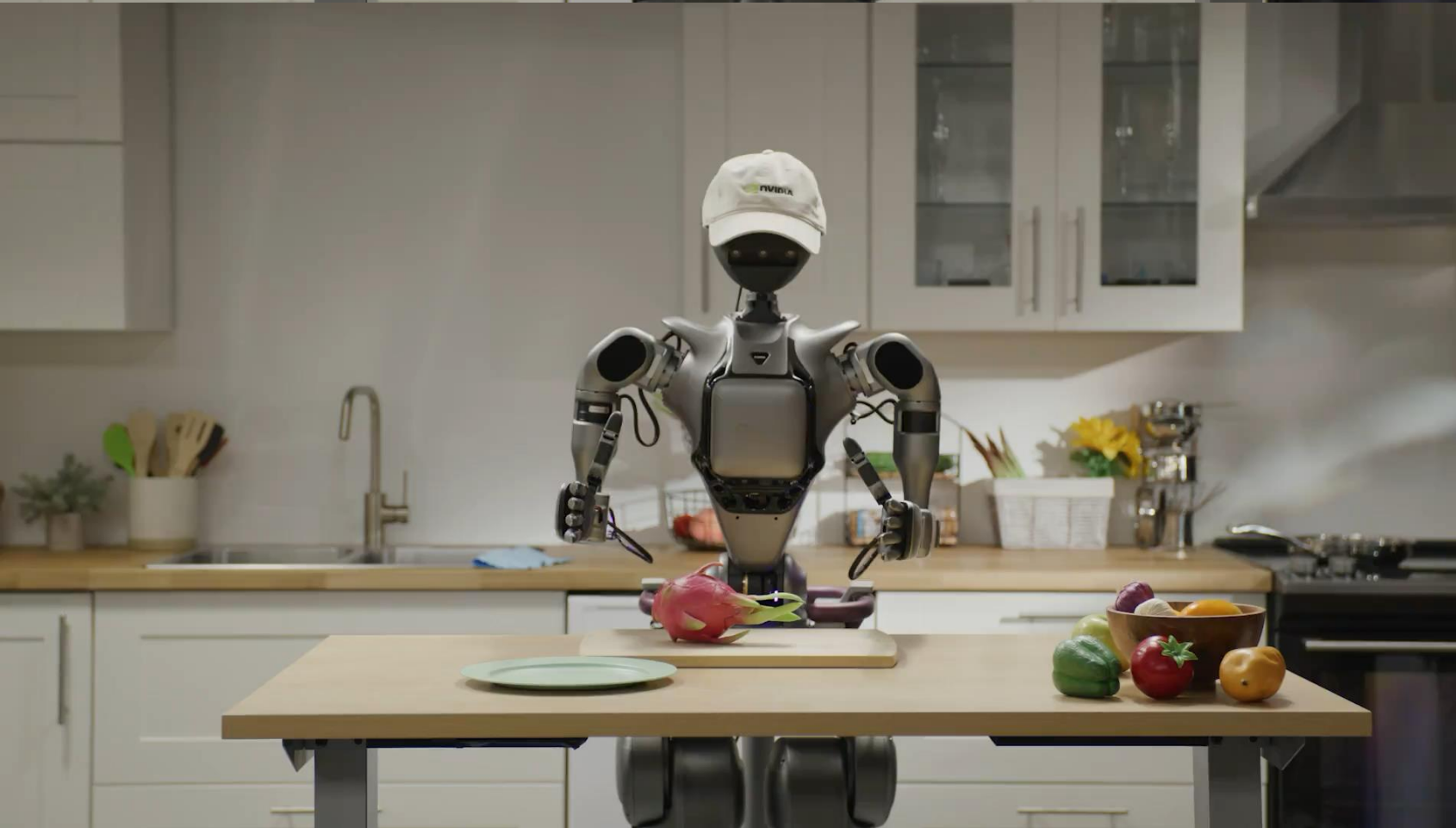


# GR00T: Humanoid Robot Foundation Model



# GR00T: Humanoid Robot Foundation Model









# Data Pyramid for Generalist Humanoid Robots

## Real-Robot Data

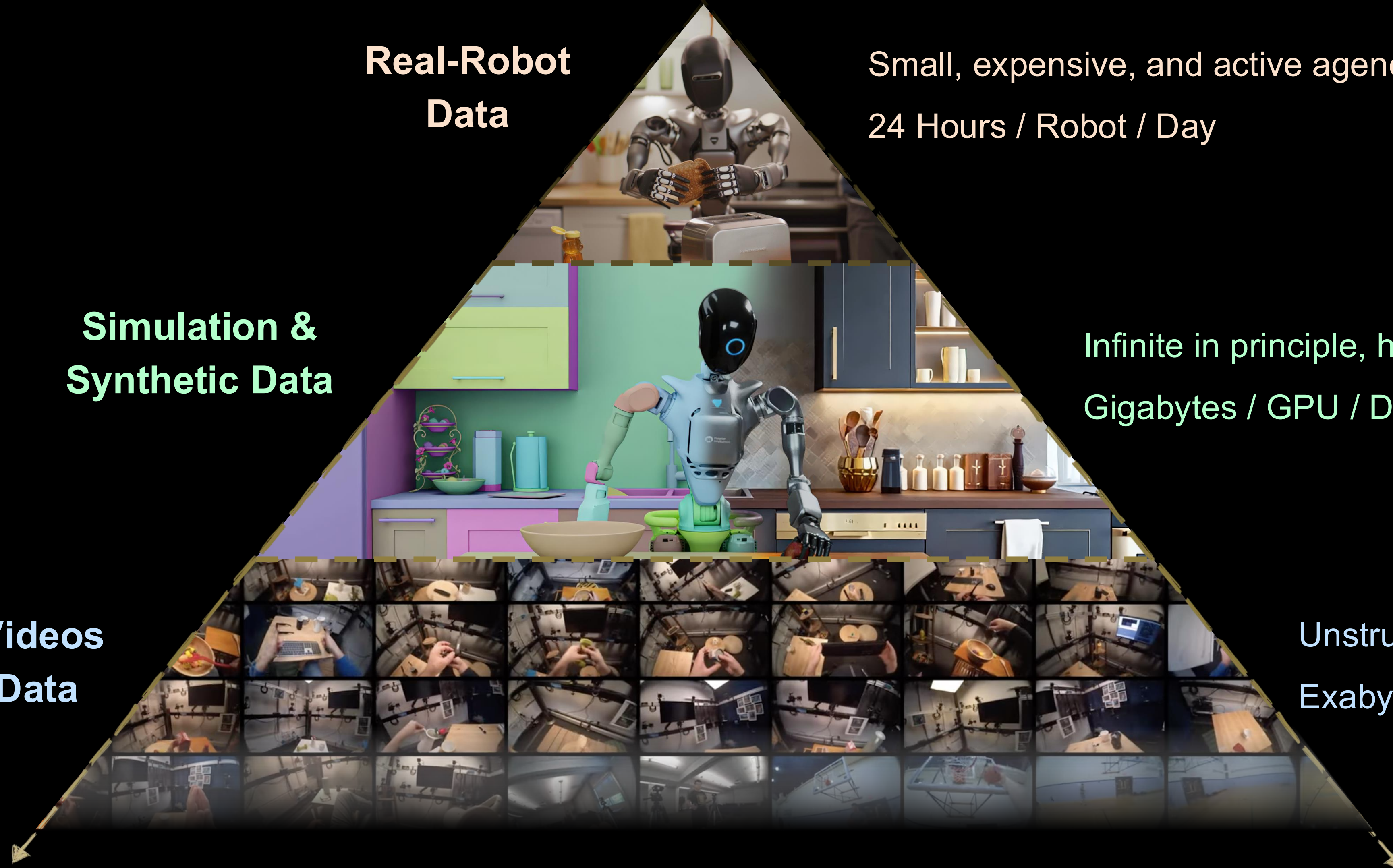
Small, expensive, and active agency  
24 Hours / Robot / Day

## Simulation & Synthetic Data

Infinite in principle, hard to scale in practice  
Gigabytes / GPU / Day

## Human Videos & Web Data

Unstructured, multimodal, passive  
Exabytes / Day



# Data Pyramid for Generalist Humanoid Robots

## Real-Robot Data

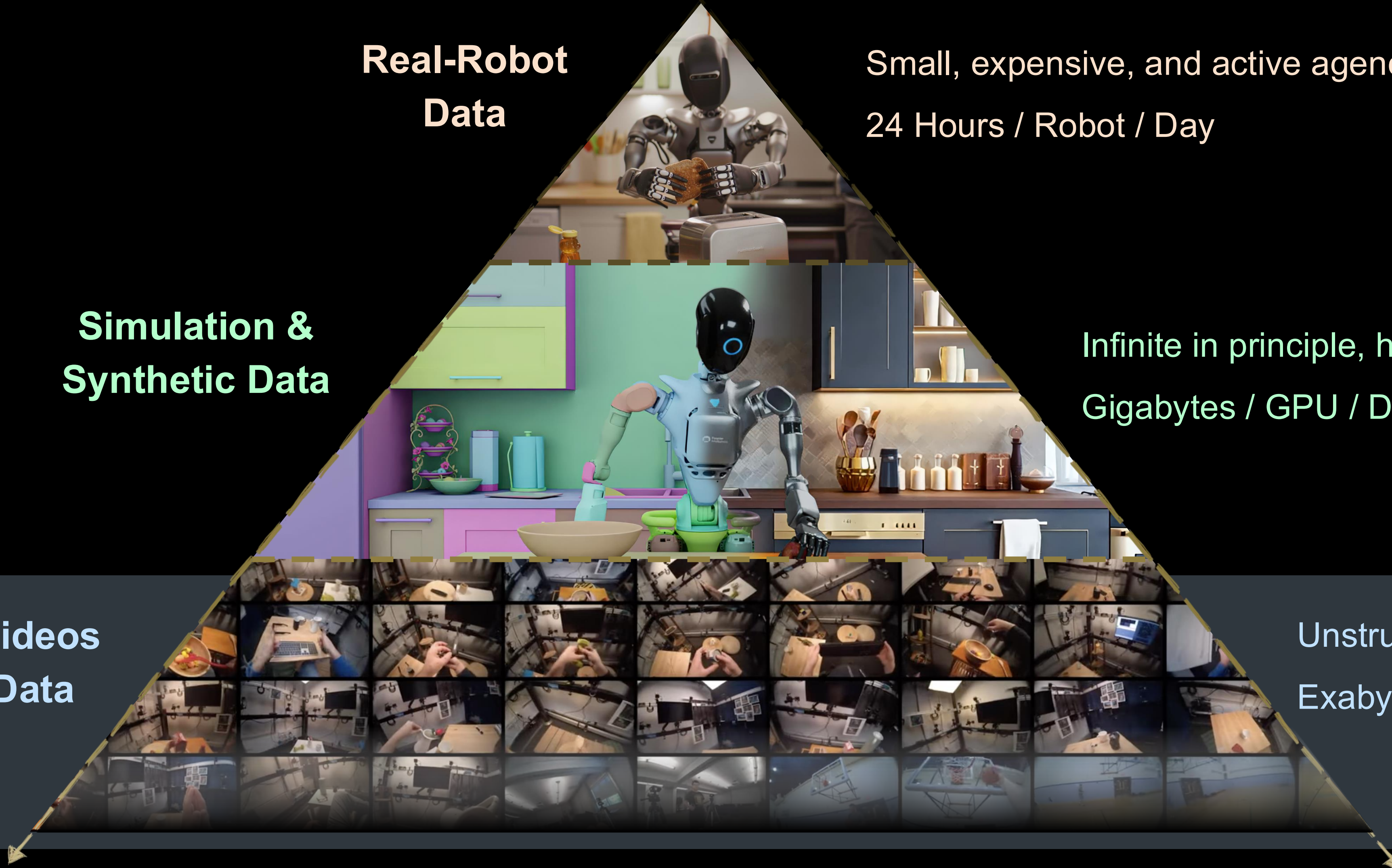
Small, expensive, and active agency  
24 Hours / Robot / Day

## Simulation & Synthetic Data

Infinite in principle, hard to scale in practice  
Gigabytes / GPU / Day

## Human Videos & Web Data

Unstructured, multimodal, passive  
Exabytes / Day



# Human Data as Scalable Training Sources

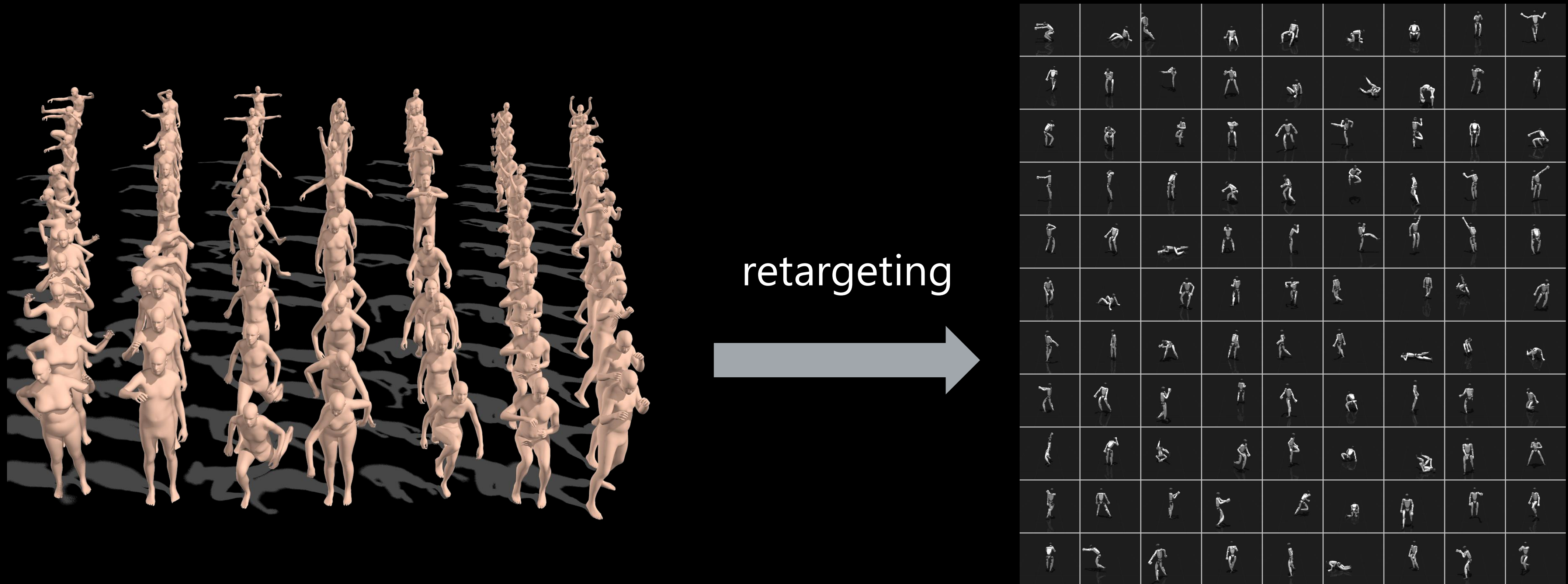


Human Motion Data



Egocentric Human Videos

# SONIC: Generalist Humanoid Whole-Body Control



**Key idea:** motion tracking as the humanoid control **super-scaler** to learn diverse behaviors without manual reward engineering.

# SONIC: Generalist Humanoid Whole-Body Control

- **Model Size**
  - 1.2M → 42M parameters
  - 35x increase in capacity
- **Data Volume**
  - 100M+ frames
  - 700+ hours of motion data
- **Compute**
  - ~9,000 GPU hours
  - Enables high-fidelity tracking



<https://nvlabs.github.io/GEAR-SONIC/>



# Human Data as Scalable Training Sources



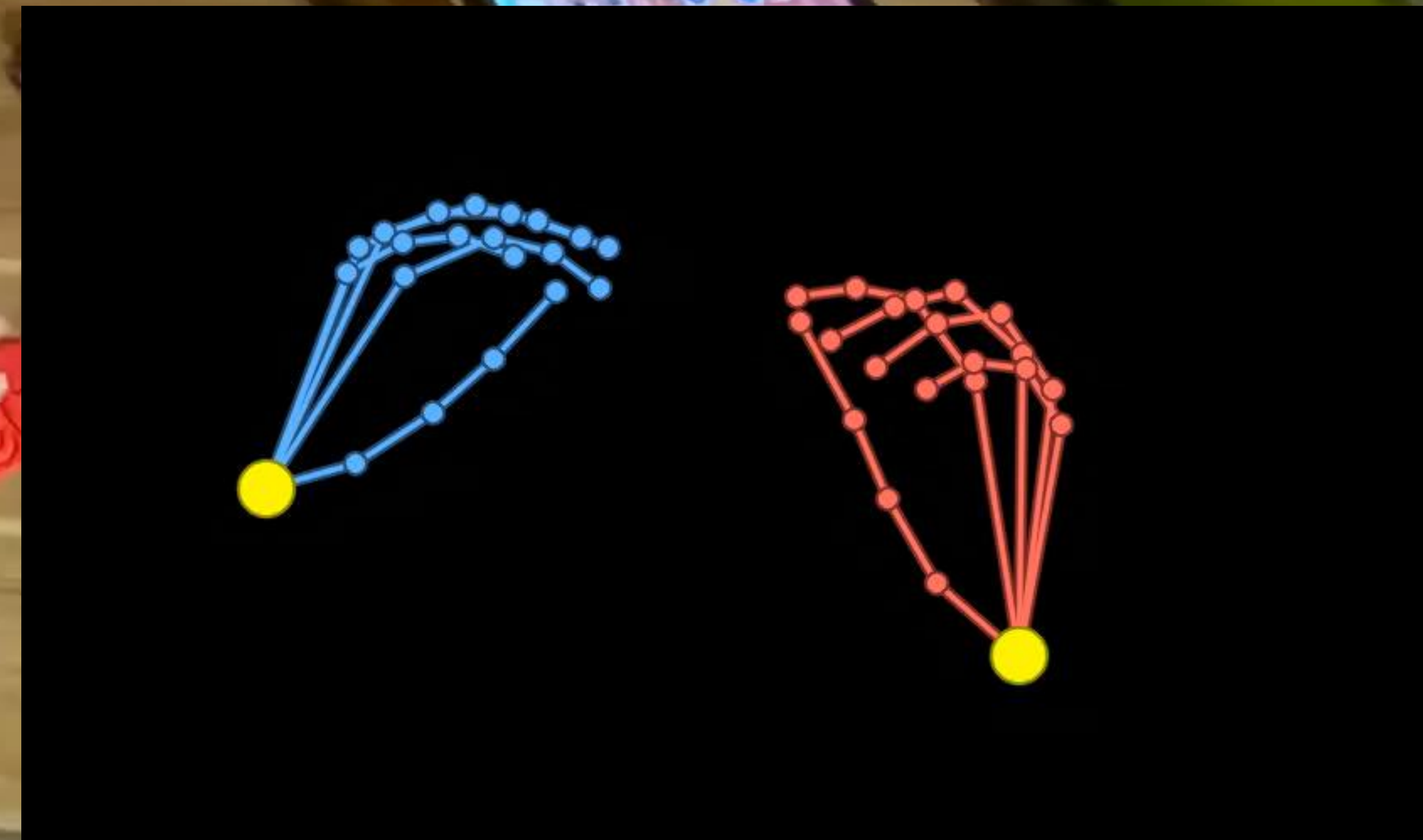
Human Motion Data



Egocentric Human Videos

# EgoScale: Learning Humanoid Dexterity from Human Videos

● ● Hand Joints  
● Wrist Poses



# EgoScale: Learning Humanoid Dexterity from Human Videos

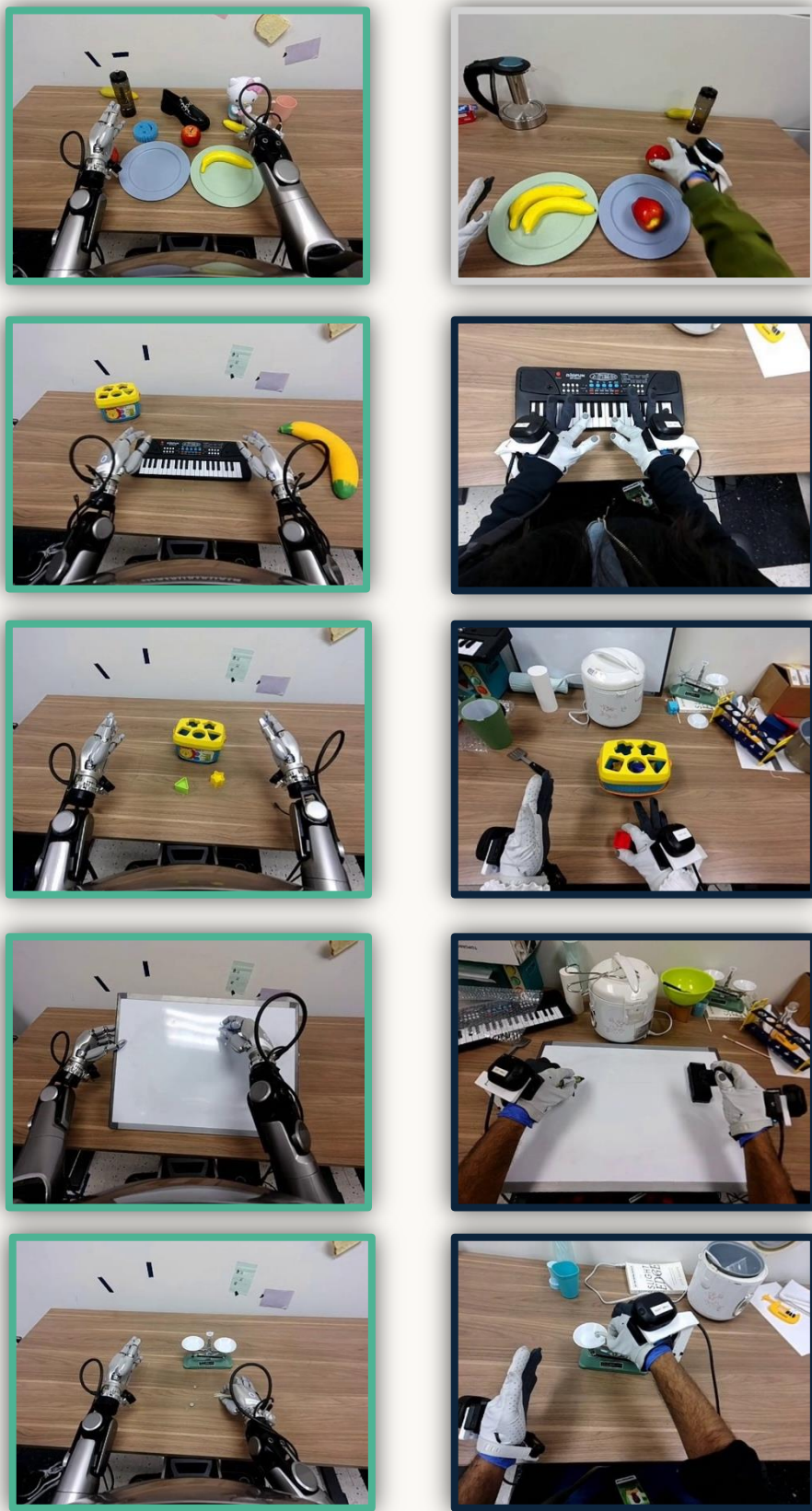
## Pre-training

20,854 Hours Human Videos



## Mid-training

Human-robot Aligned Play data  
50 Hours Human + 4 Hours Robot

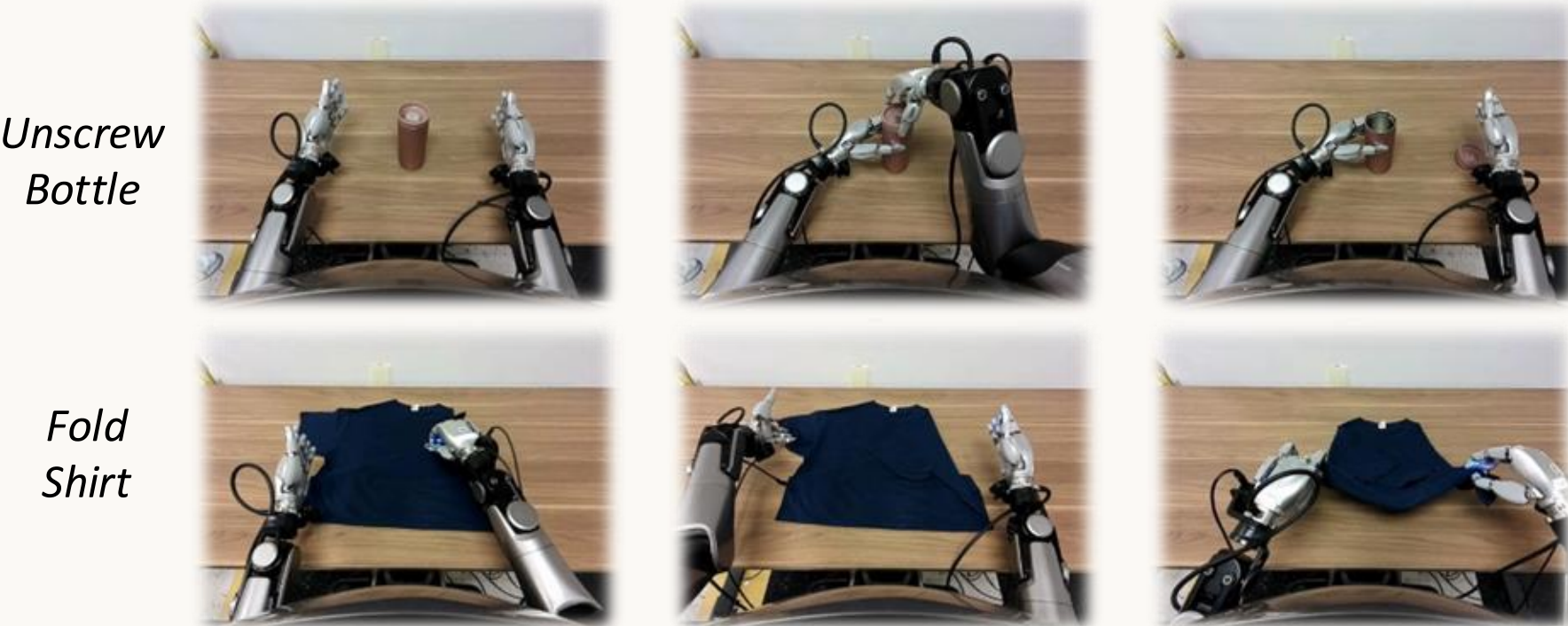


## Post-training

Efficient Learning on Highly Dexterous Tasks

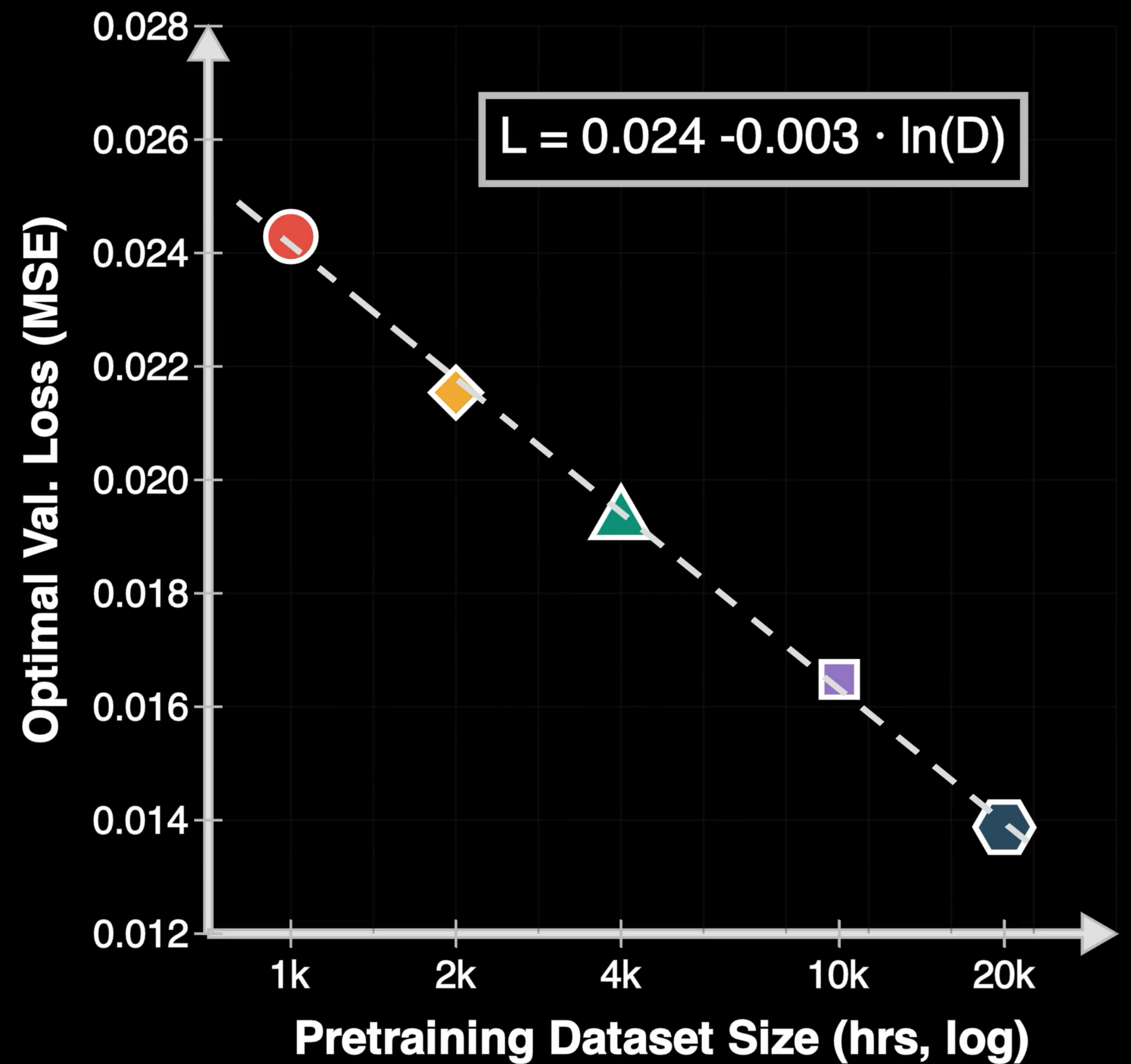
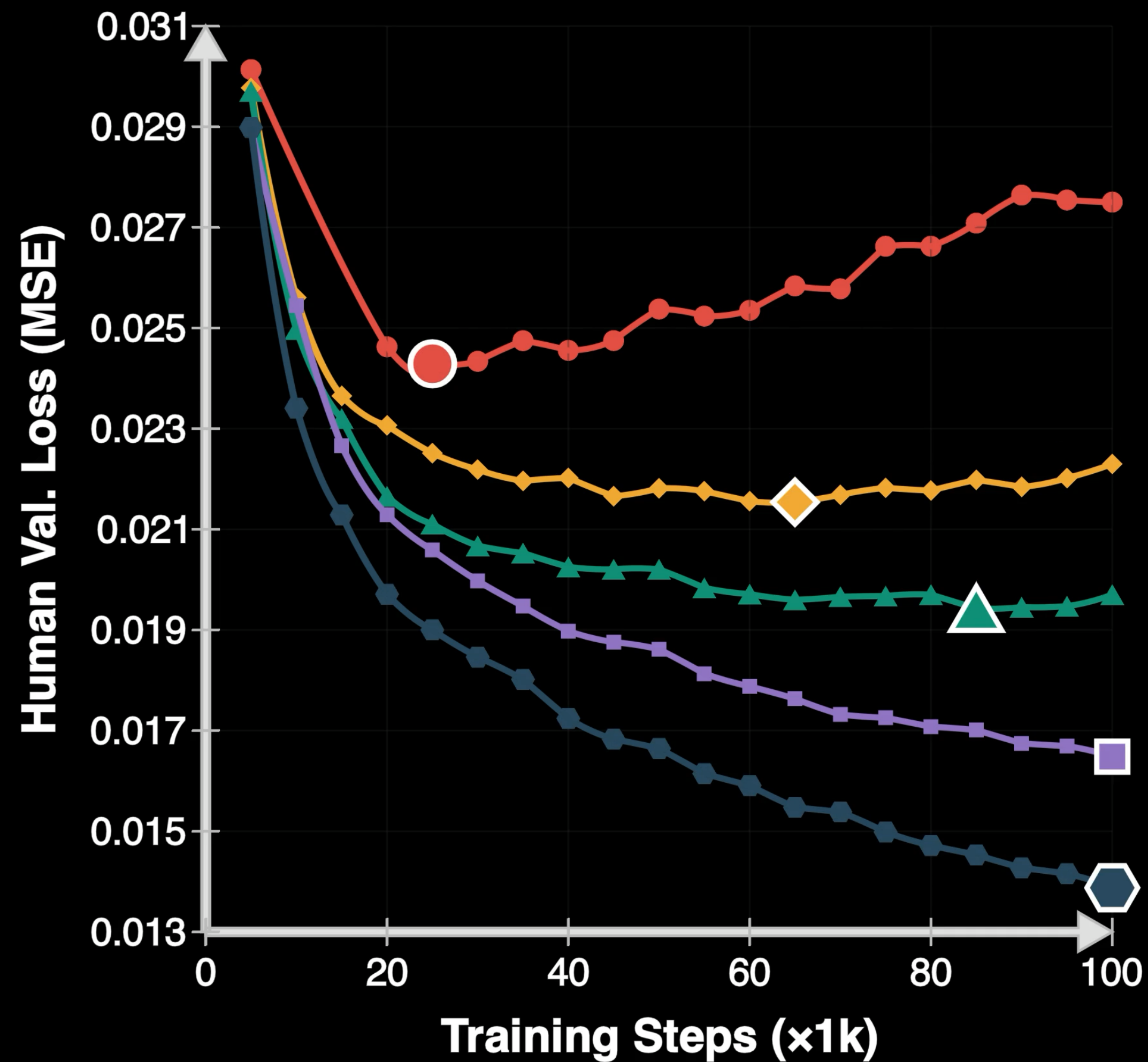


## One-Shot on Unseen Dexterous Tasks

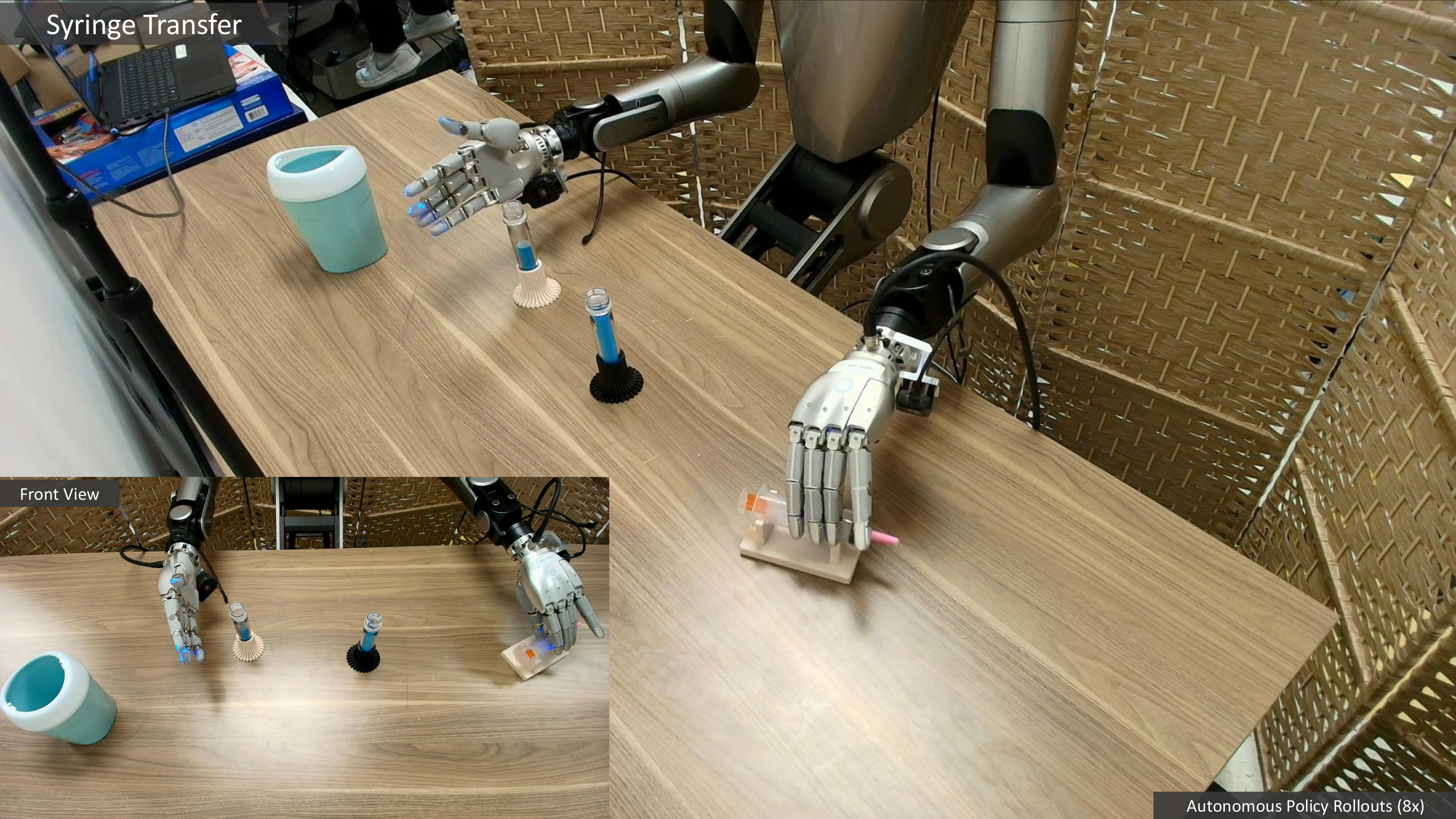


# EgoScale: Learning Humanoid Dexterity from Human Videos

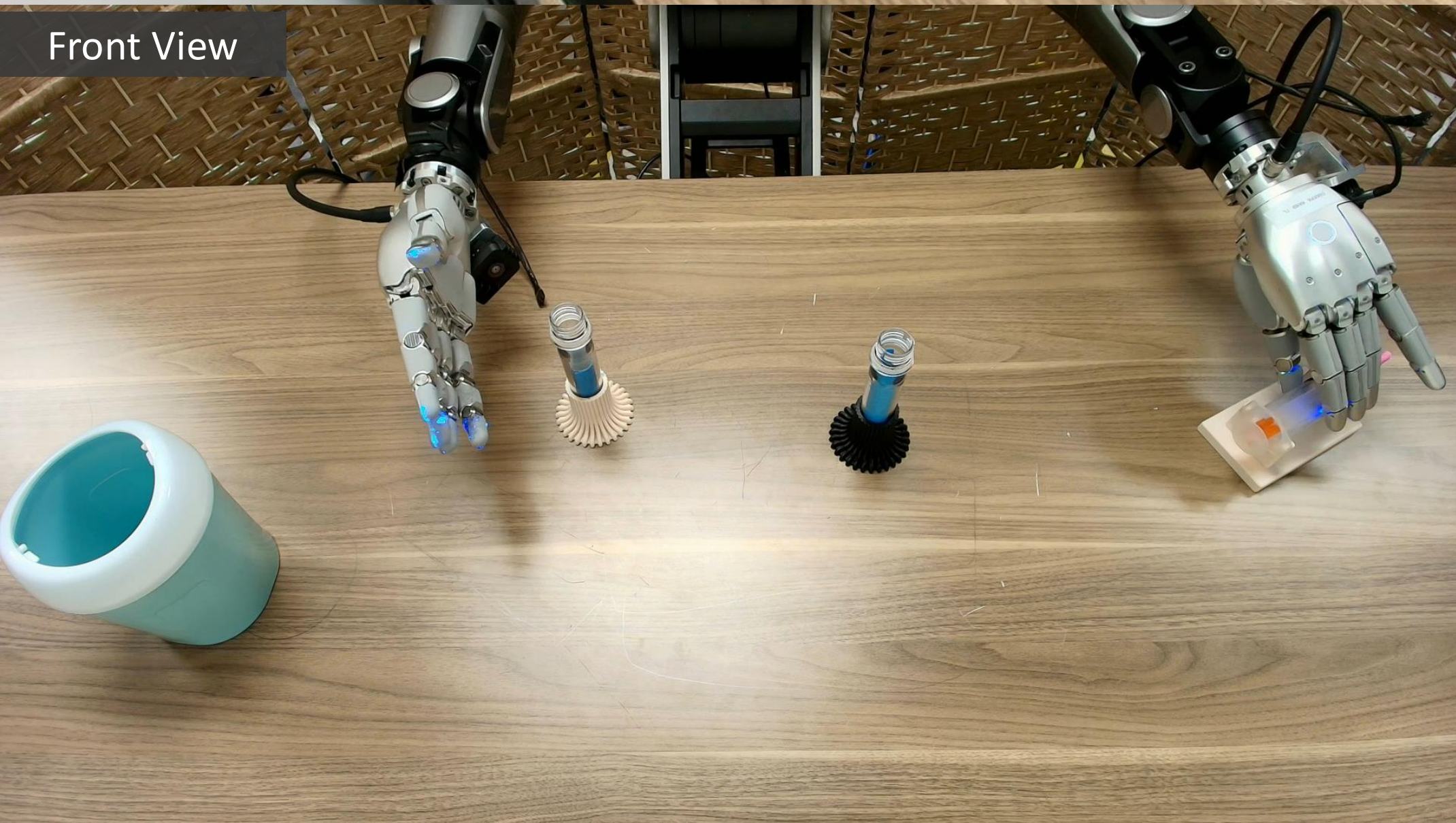
● 1k hrs    ◆ 2k hrs    ▲ 4k hrs    ■ 10k hrs    ◆ 20k hrs

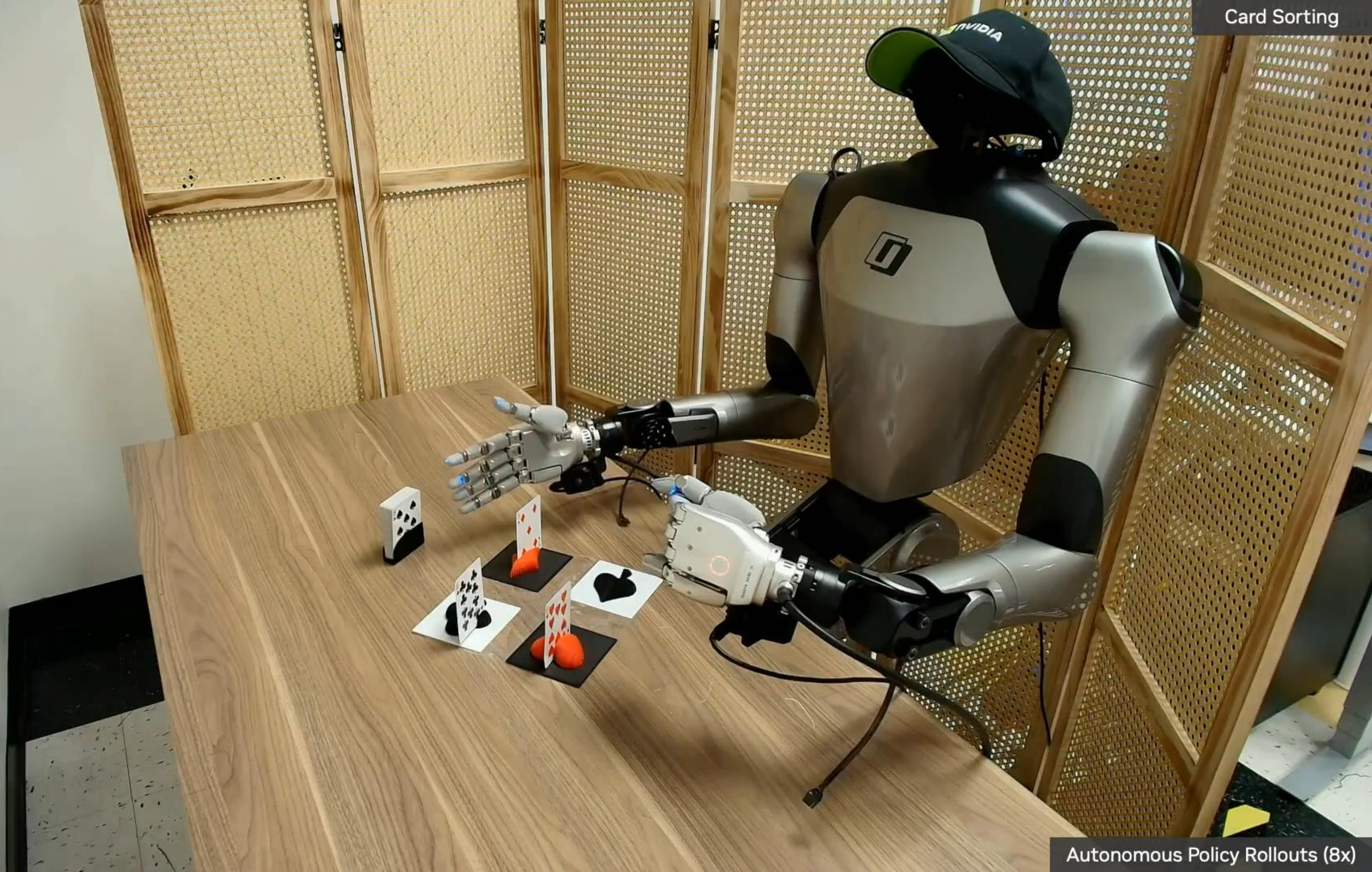


# Syringe Transfer

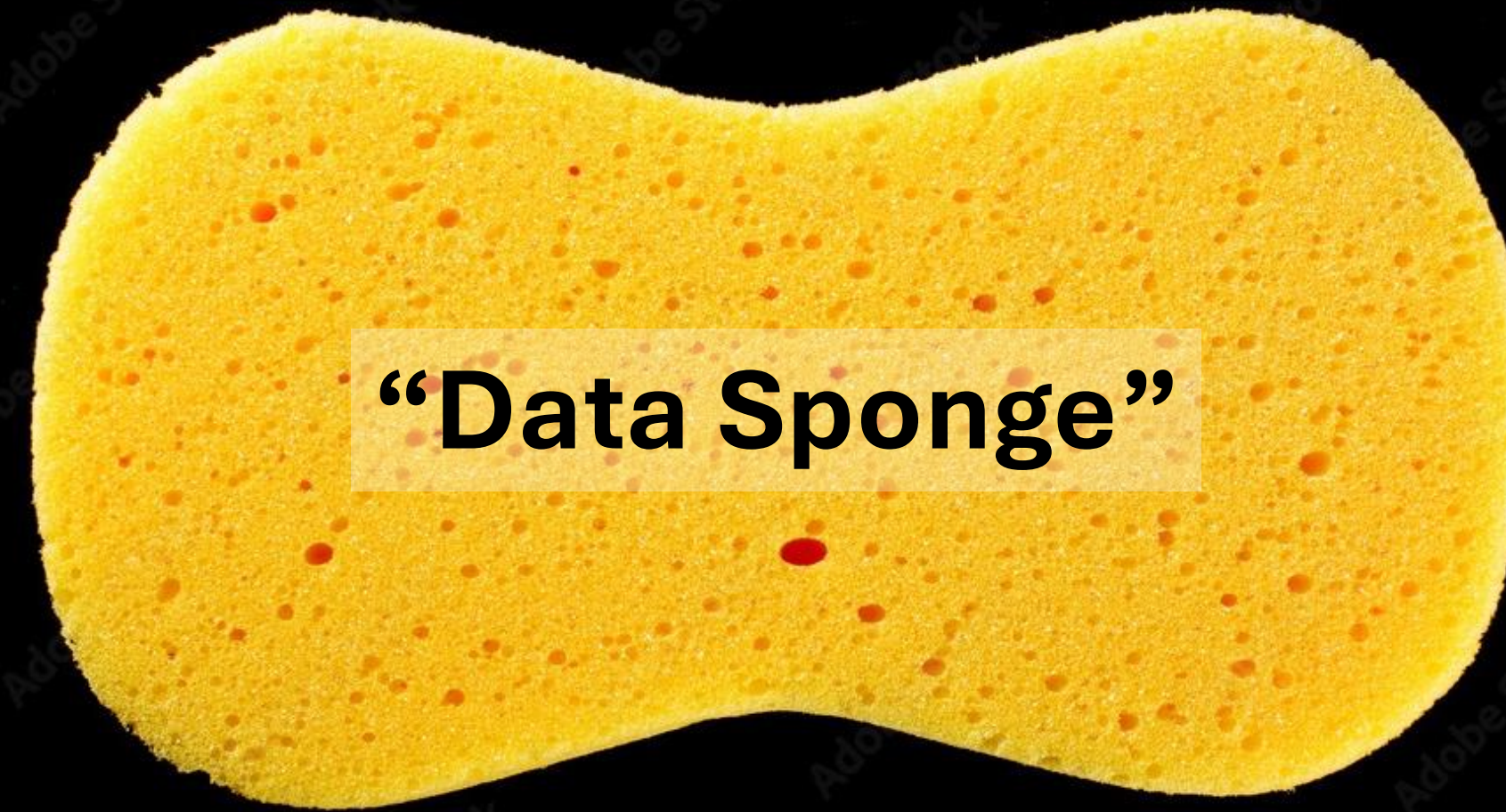


## Front View





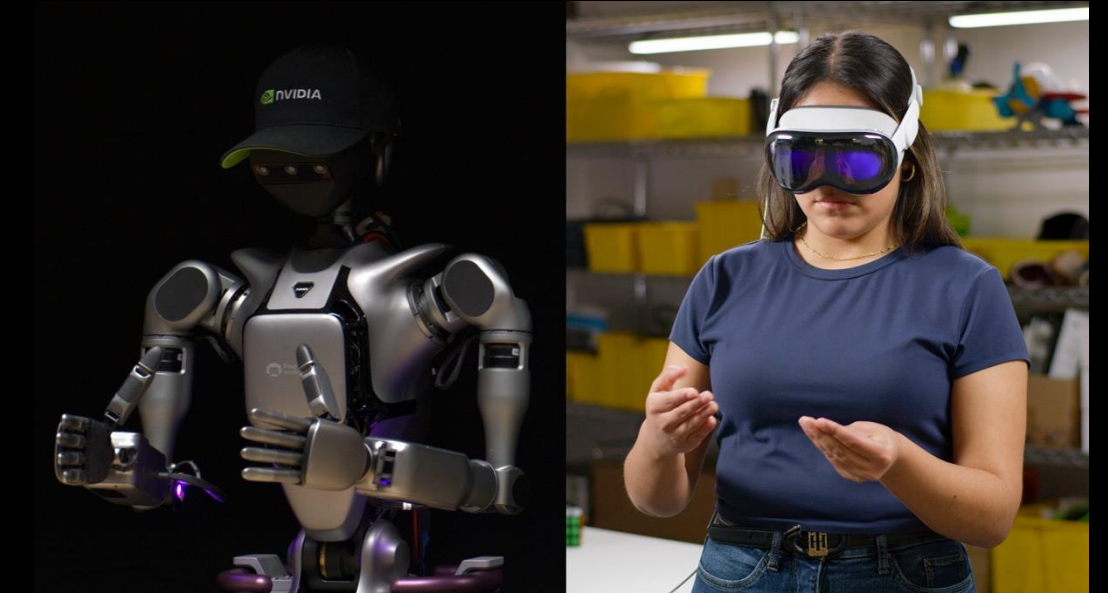
# World Models for Robotics



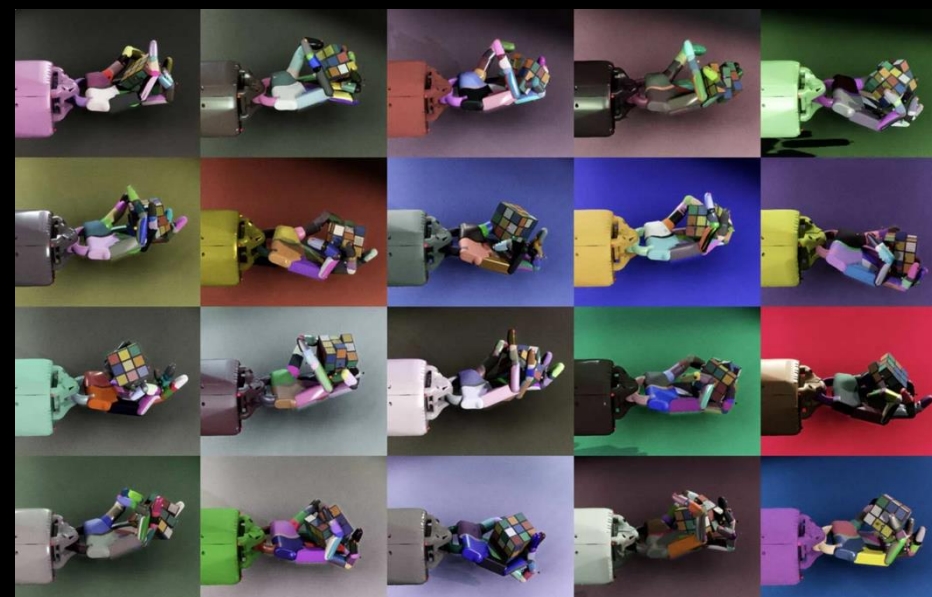
# World Models for Robotics



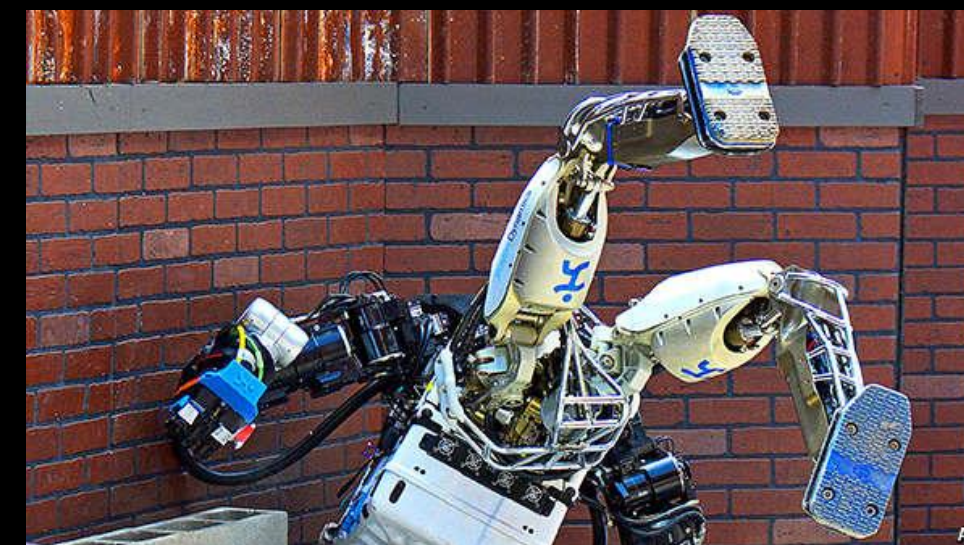
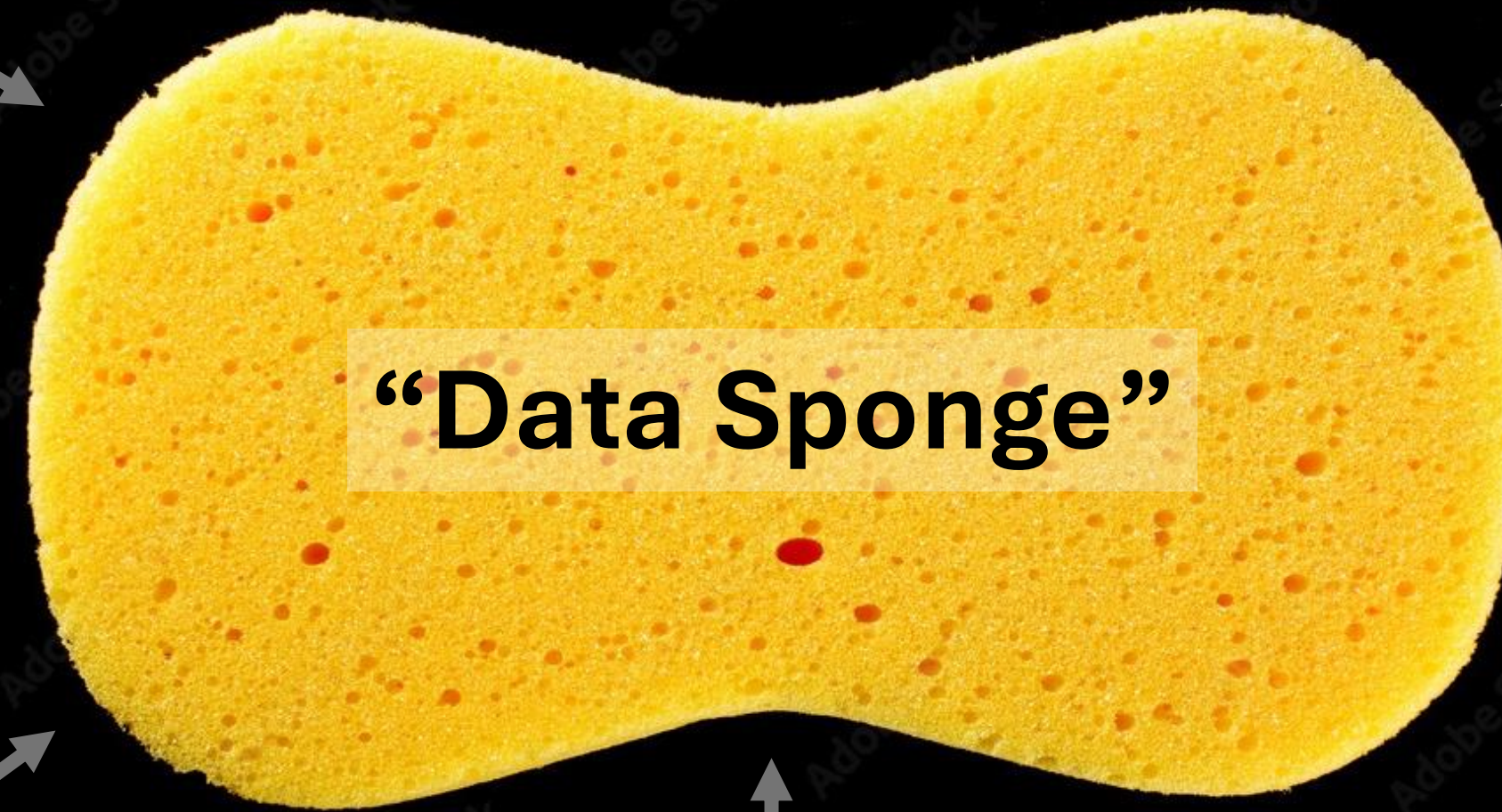
Internet Videos



Real-Robot Trajectories



Synthetic Data

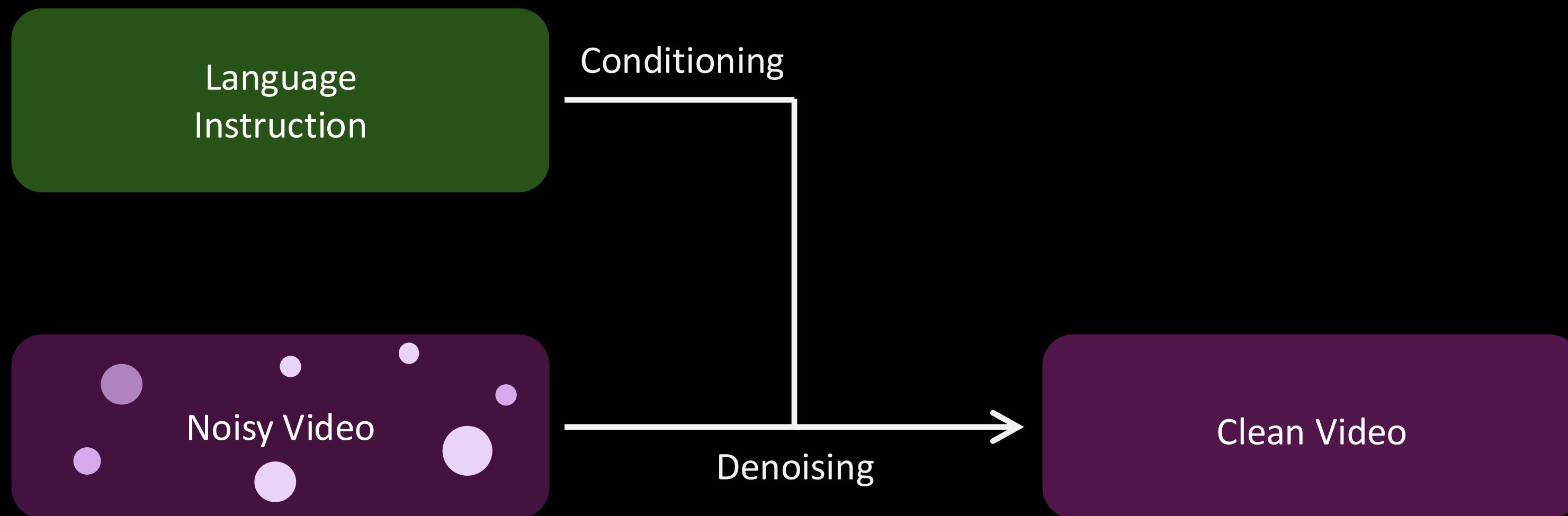


Failure Experiences



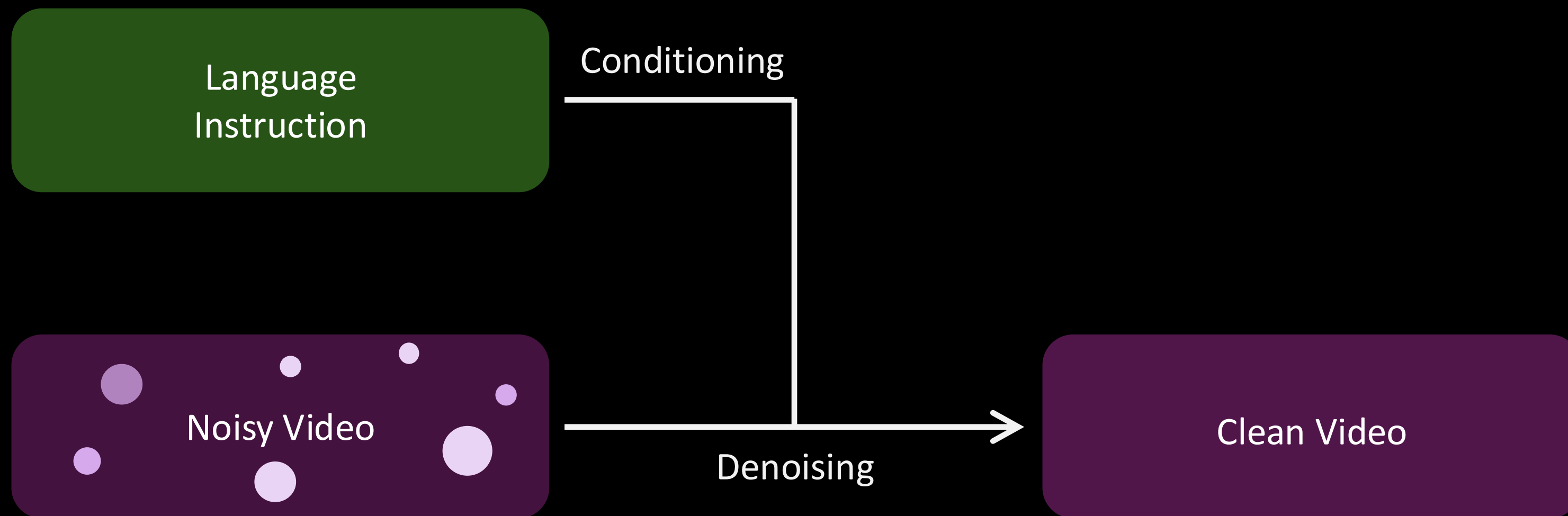
Multimodal Human Data

# DreamZero: World Action Models are Zero-shot Policies



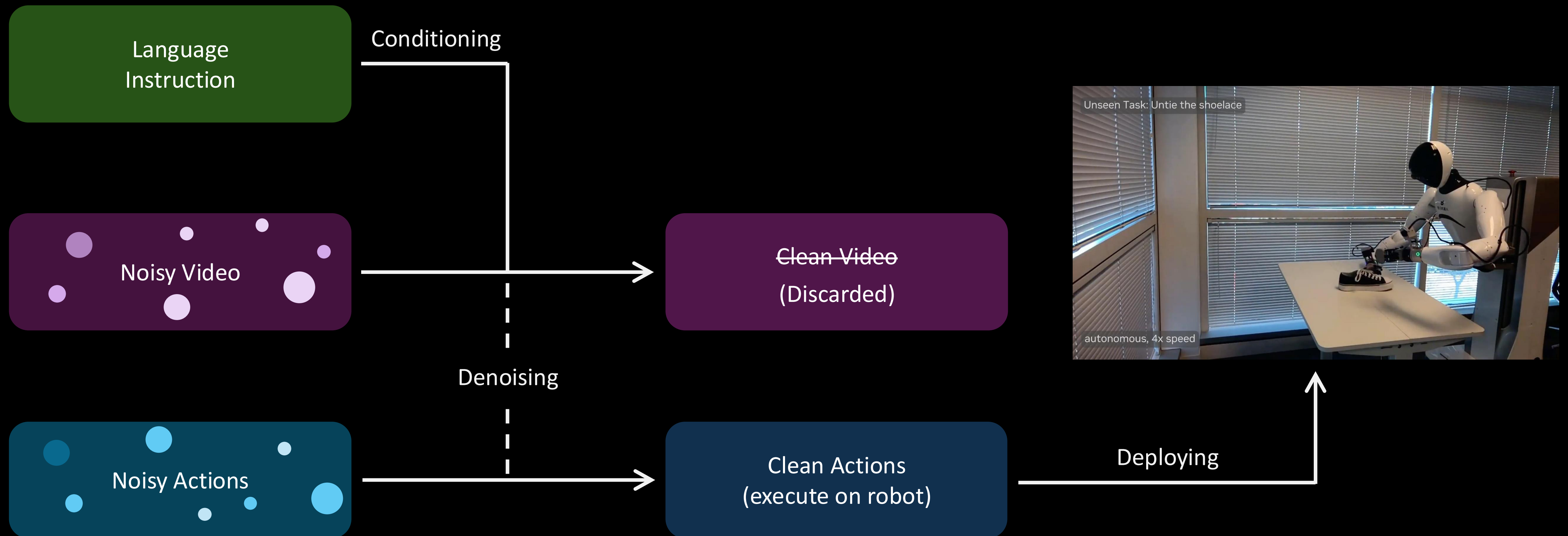
[Video credit: OpenAI Sora]

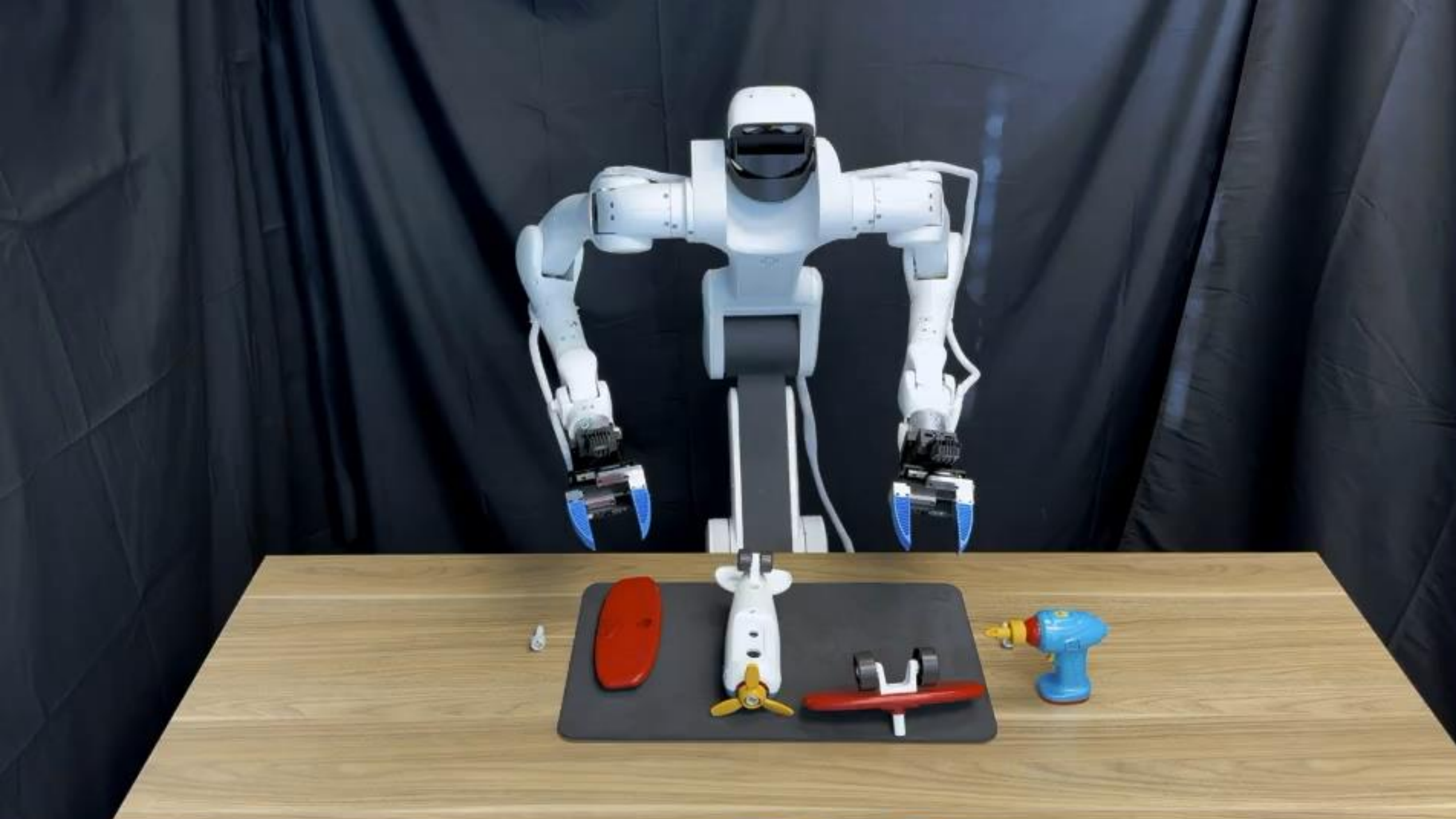
# DreamZero: World Action Models are Zero-shot Policies

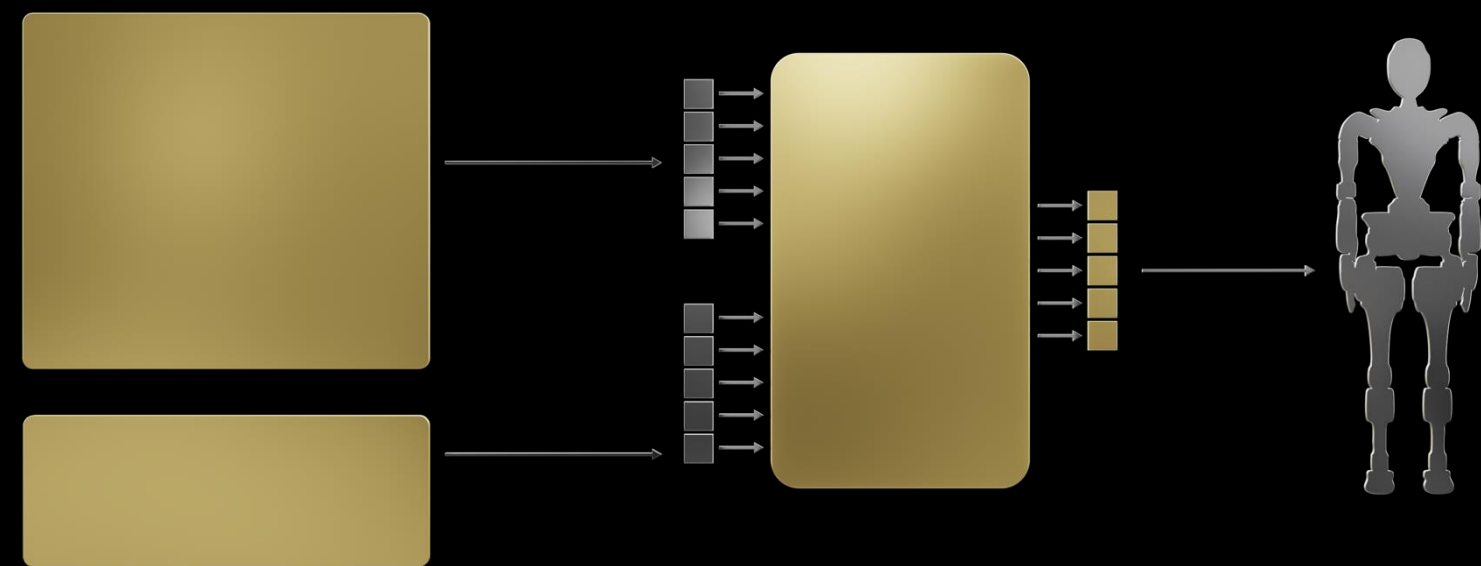


“Pour nuts into the container.”

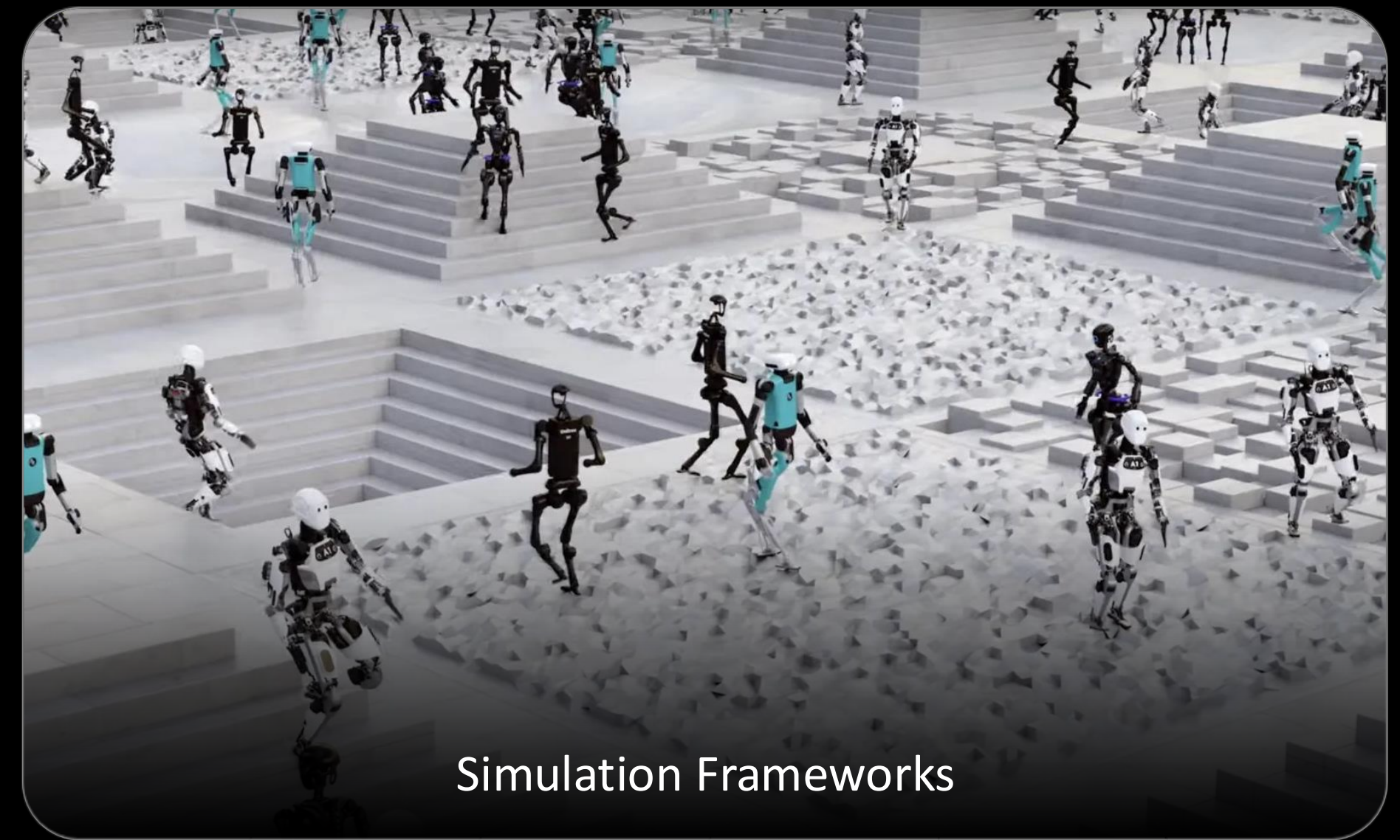
# DreamZero: World Action Models are Zero-shot Policies



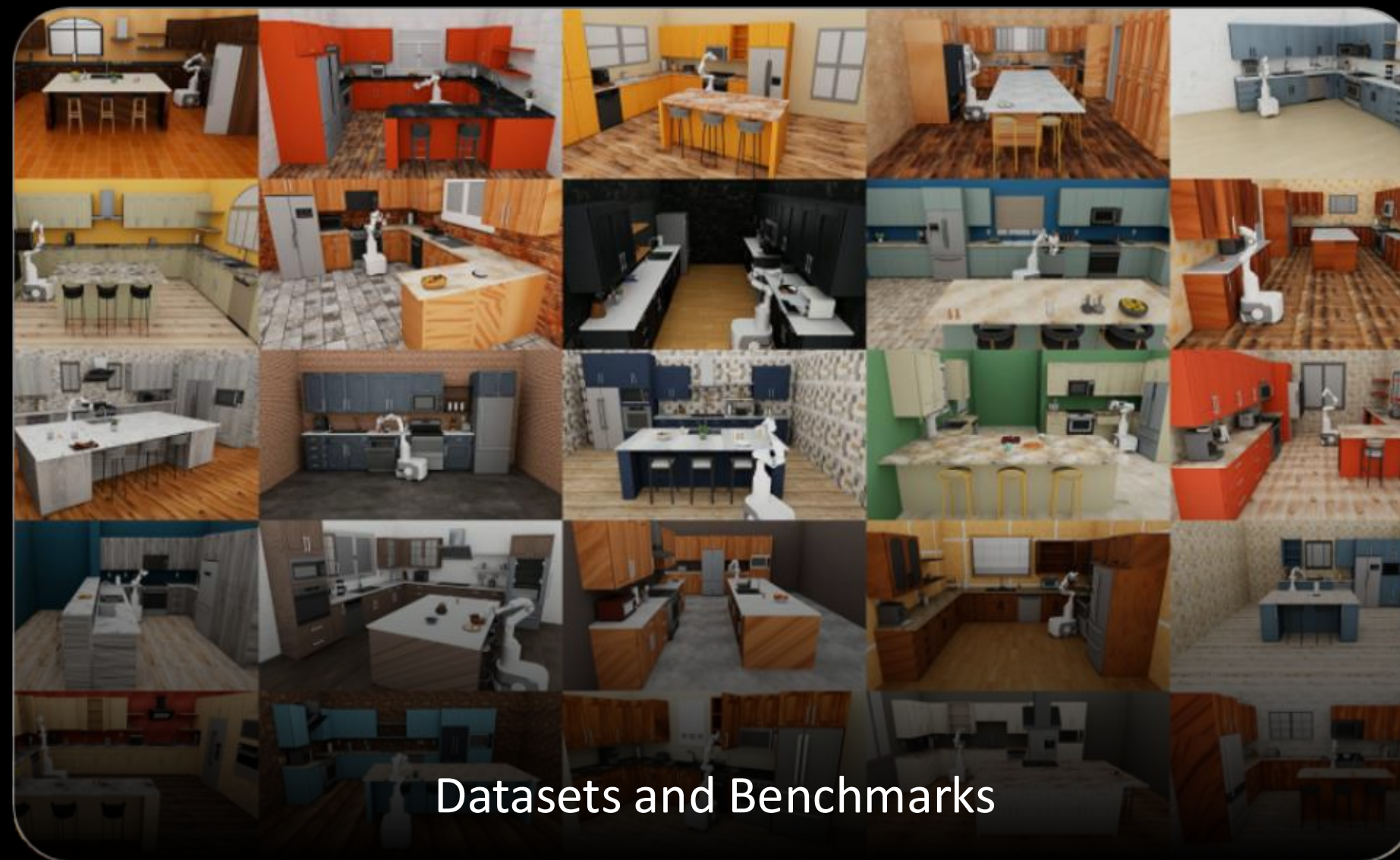




Open Robot Foundation Models



Simulation Frameworks



Datasets and Benchmarks



Reference Hardware Designs

UT Austin Robot Perception and Learning Group

 [rpl.cs.utexas.edu](https://rpl.cs.utexas.edu)

NVIDIA Generalist Embodied Agent Research Lab

 [research.nvidia.com/labs/gear](https://research.nvidia.com/labs/gear)