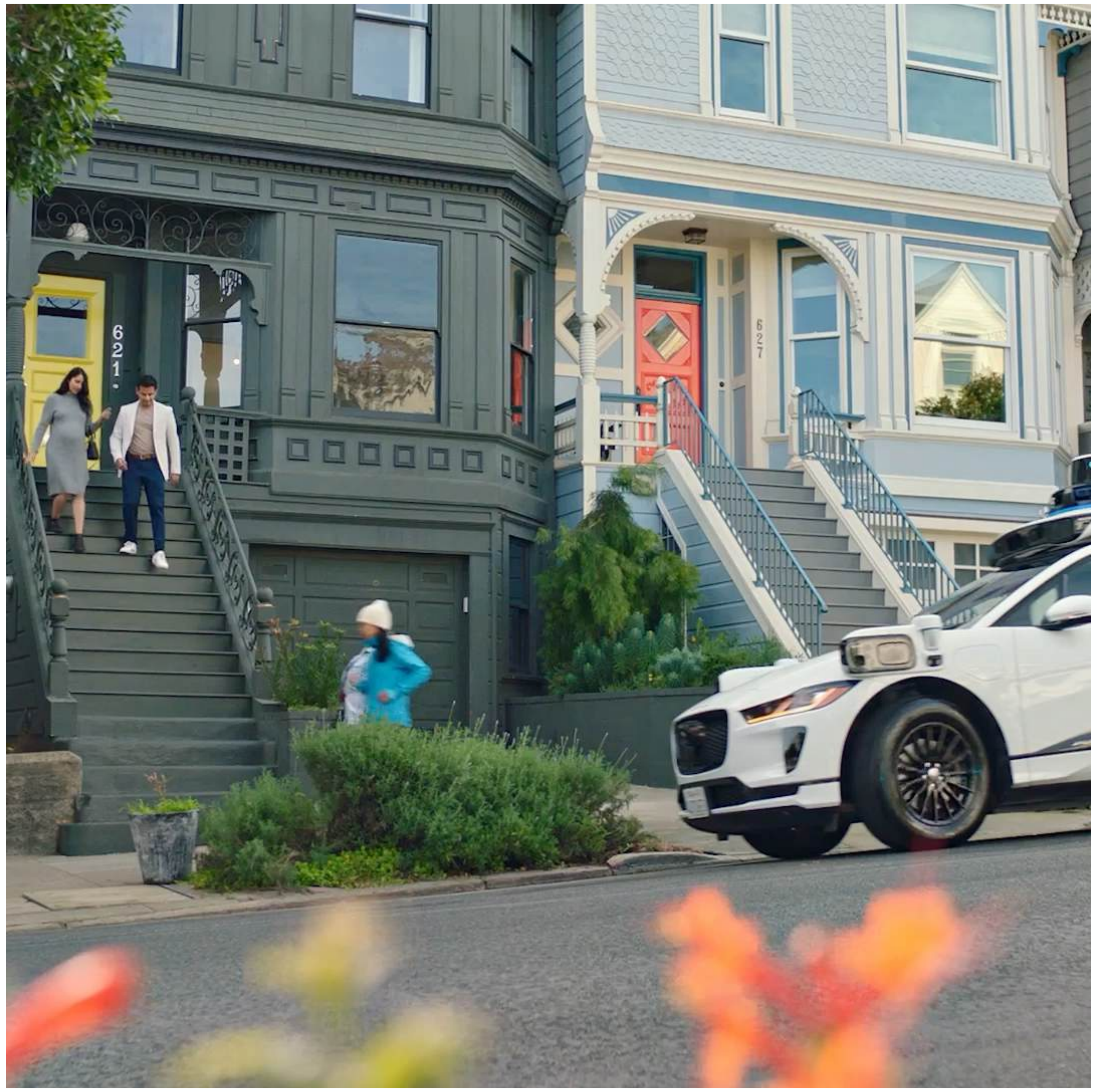
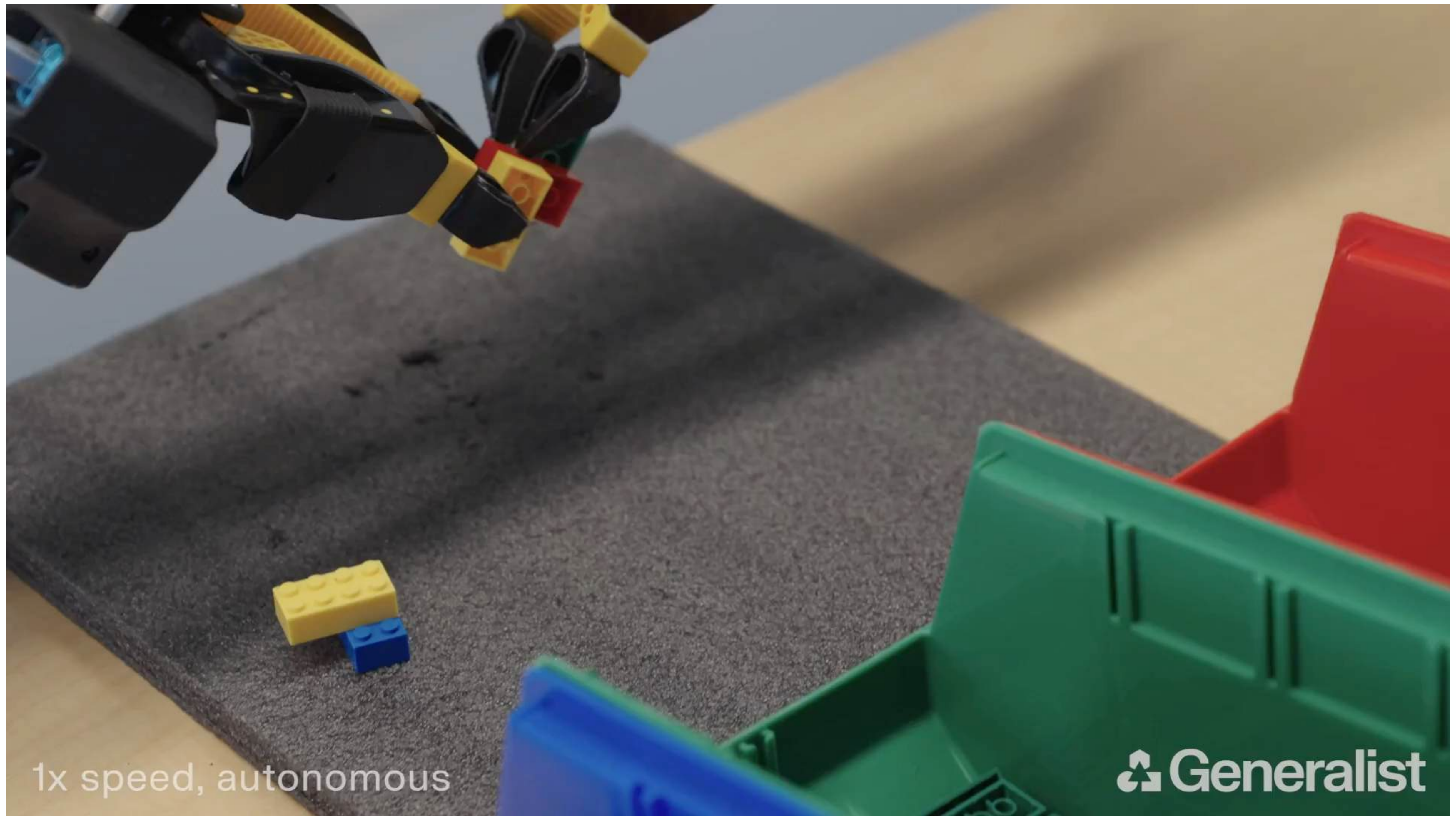
Generate Robotic Data with Spatial Intelligence

Yue Wang

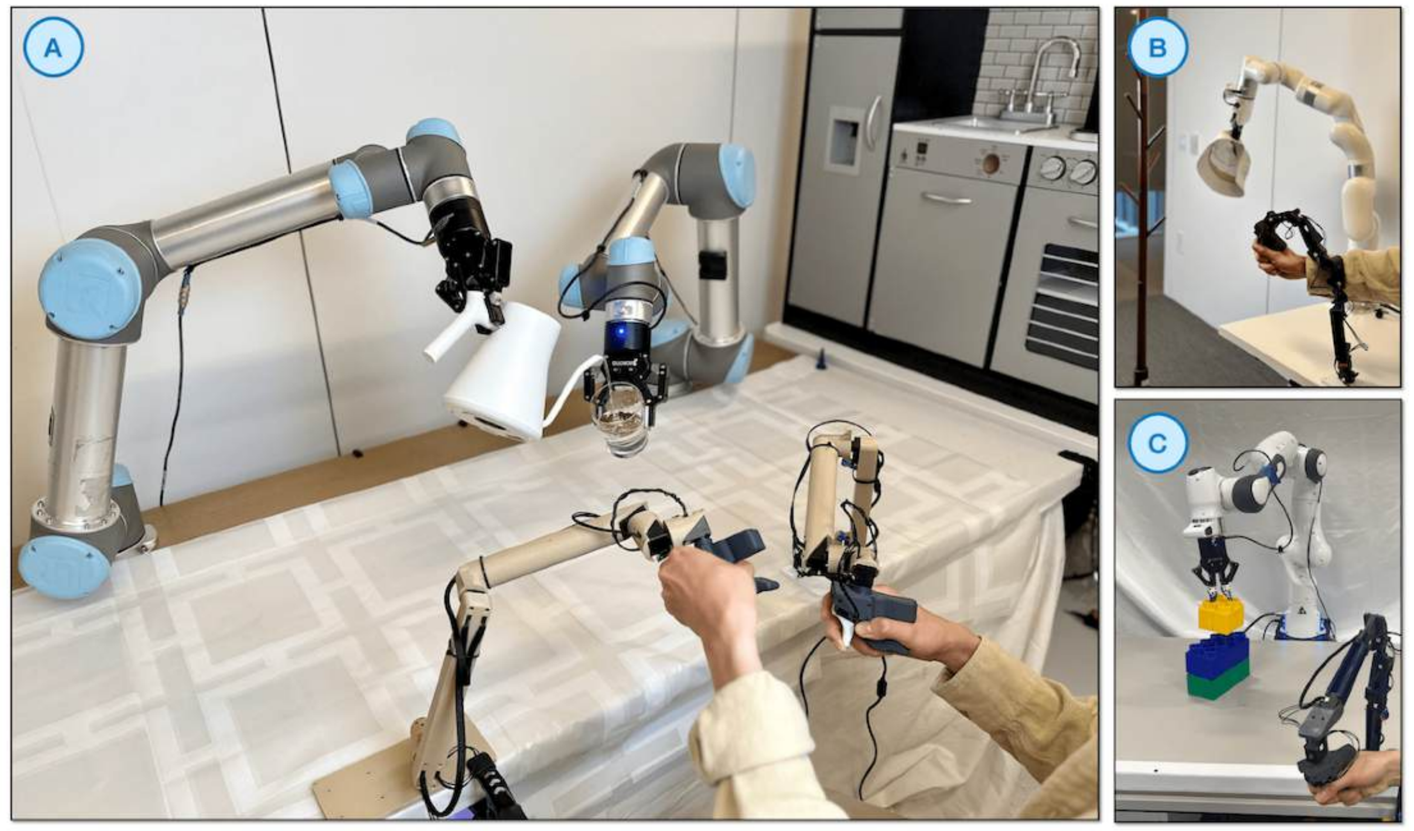
MUSI | Oct 20th, 2025

USC





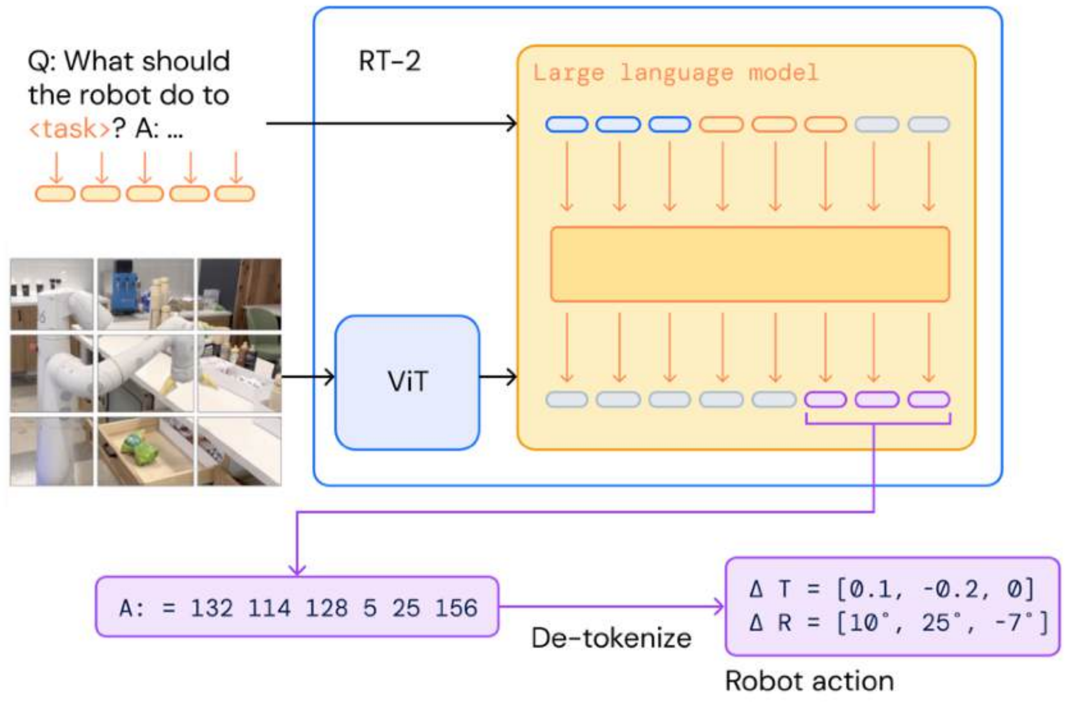
Cambrian Explosion of Robotics



Data

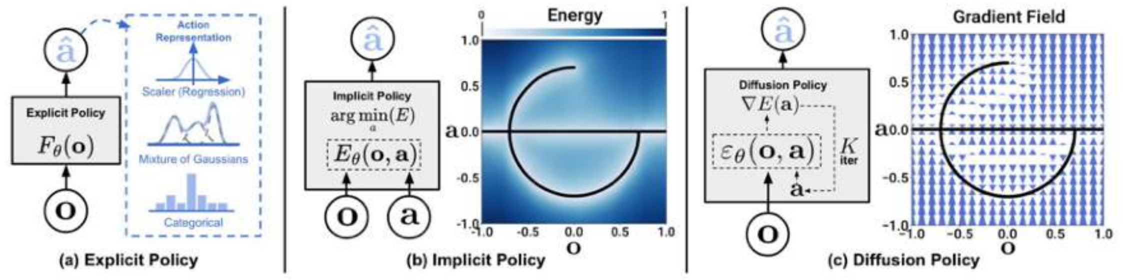
Hardware

Algorithm



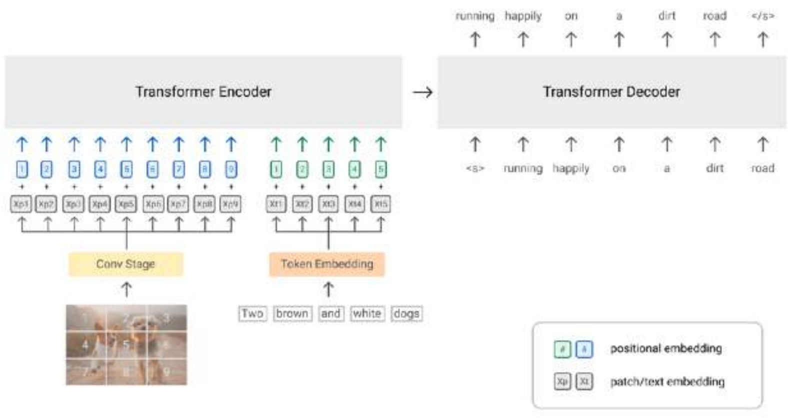
Diffusion Policy

Visuomotor Policy Learning via Action Diffusion

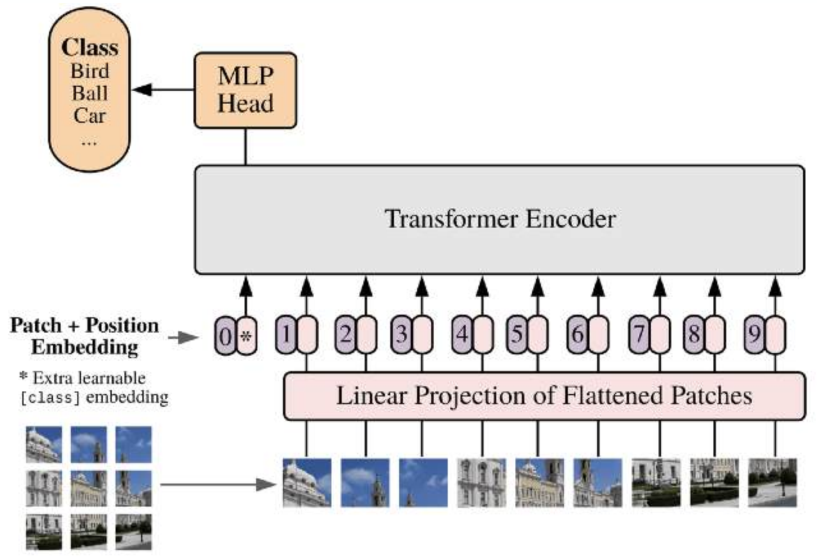


Data is the key to artificial intelligence.

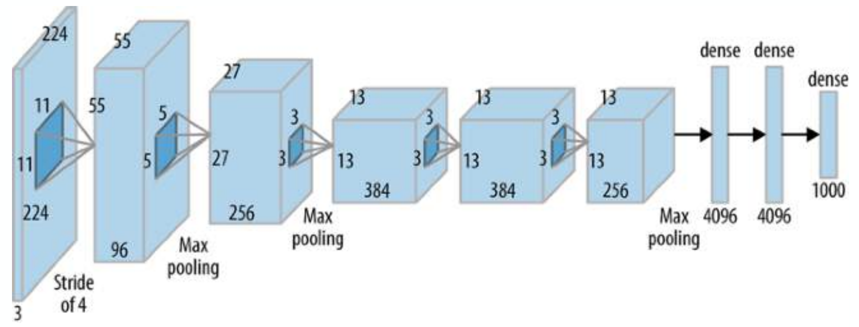
Physical AI



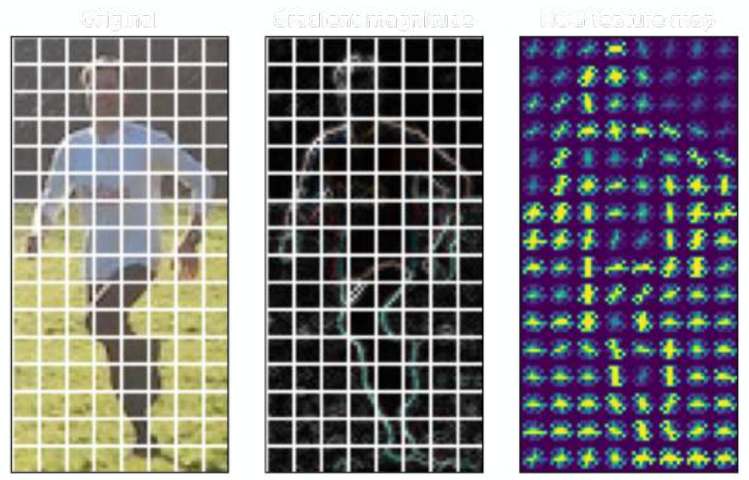
VLMs



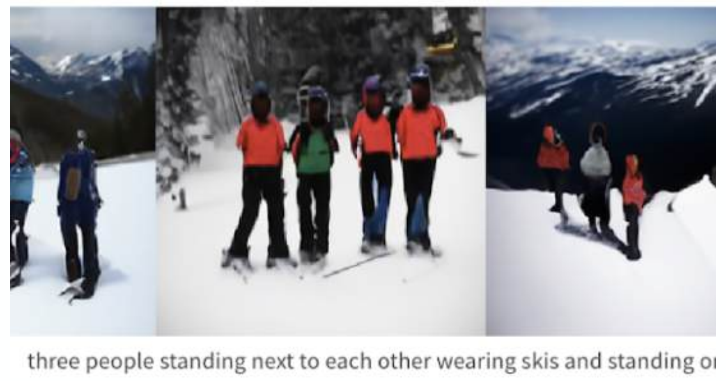
Transformers

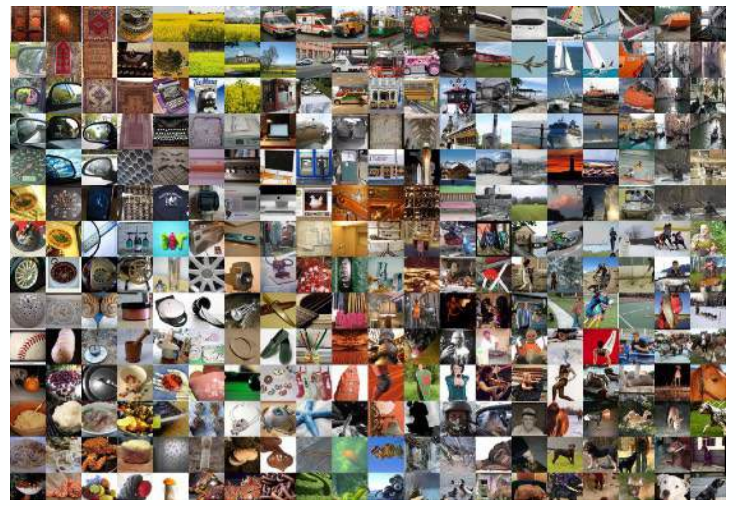


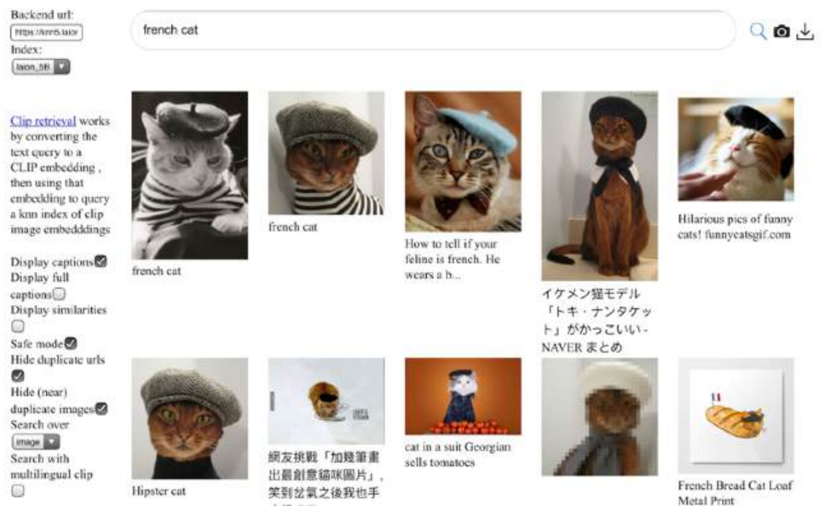
CNNs

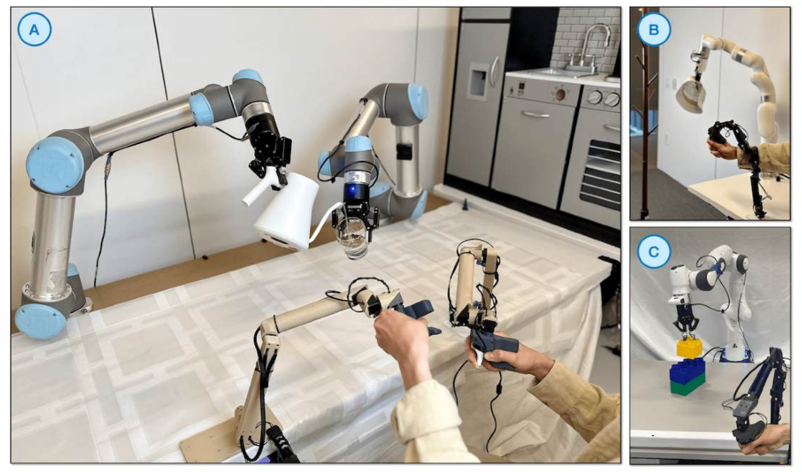


HOG+SIFT+SVM









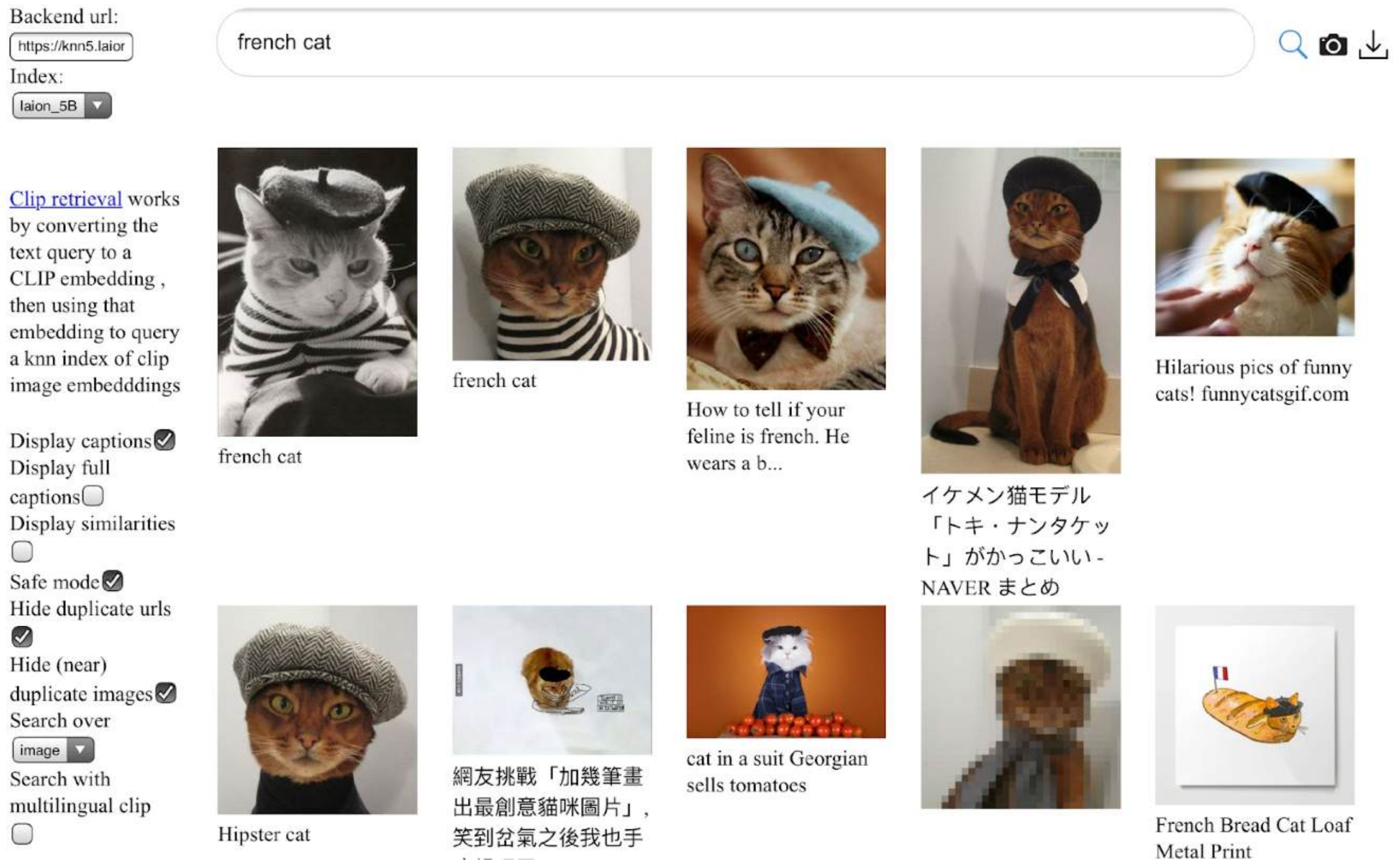
Little Data

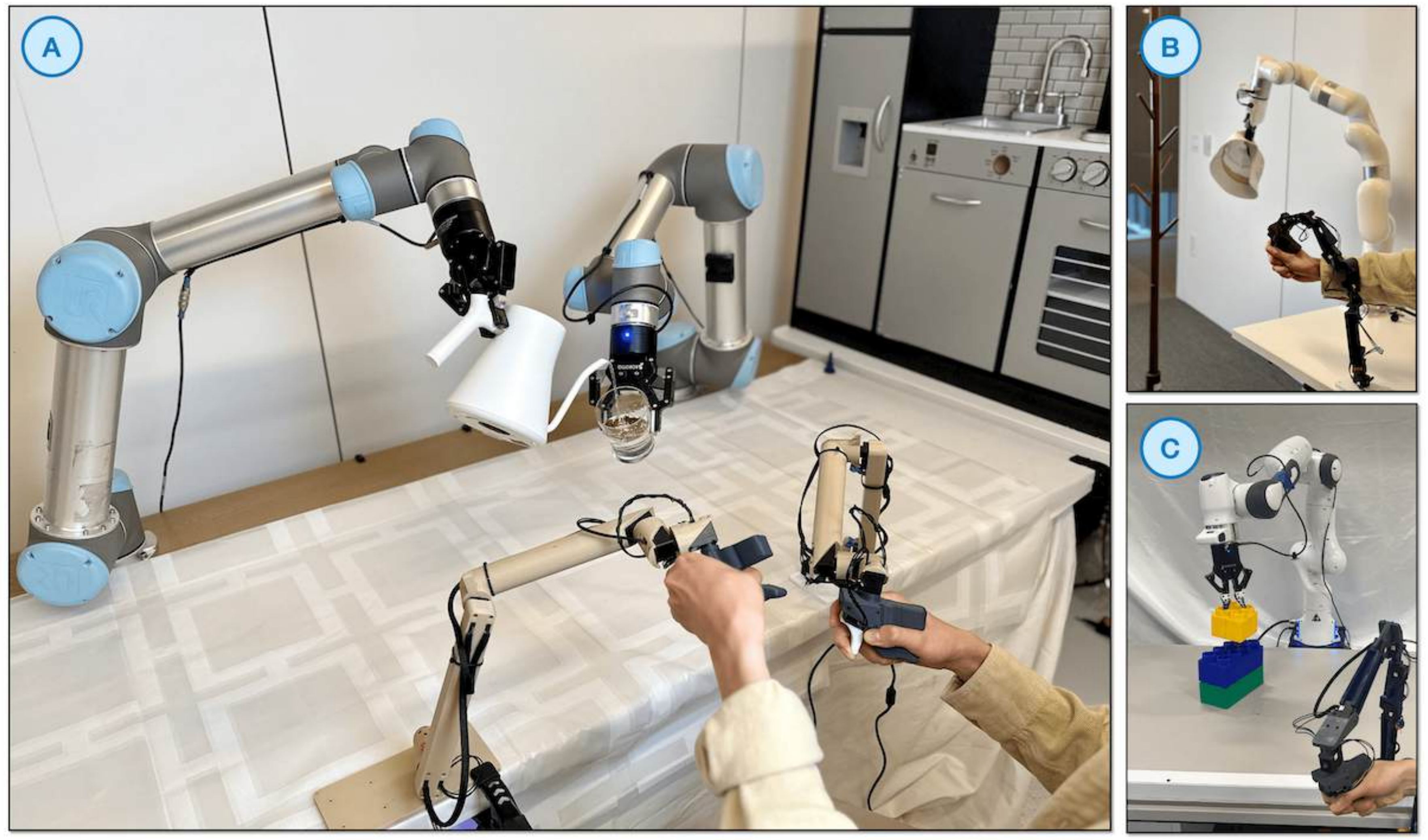
Curated

Web Scale

Multimodal

Embodied Data





< 1s

Ubiquitous

$0.01 per data point

> 60s

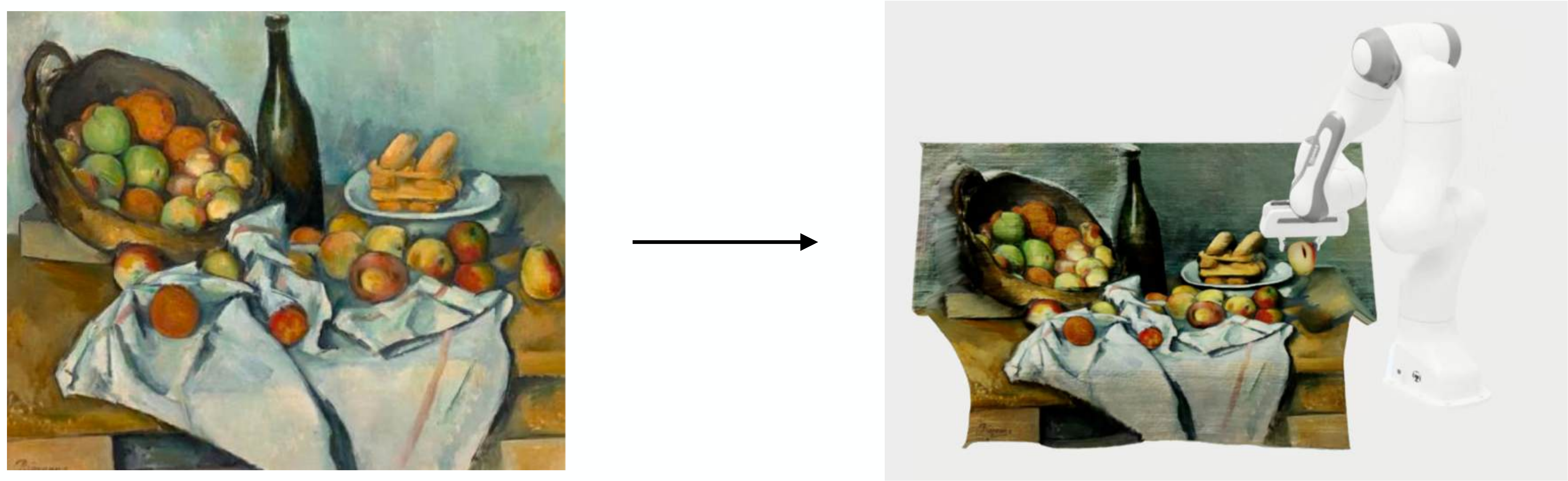
Confined to lab environments

$5 per data point

How to generate robotic data with spatial intelligence techniques?

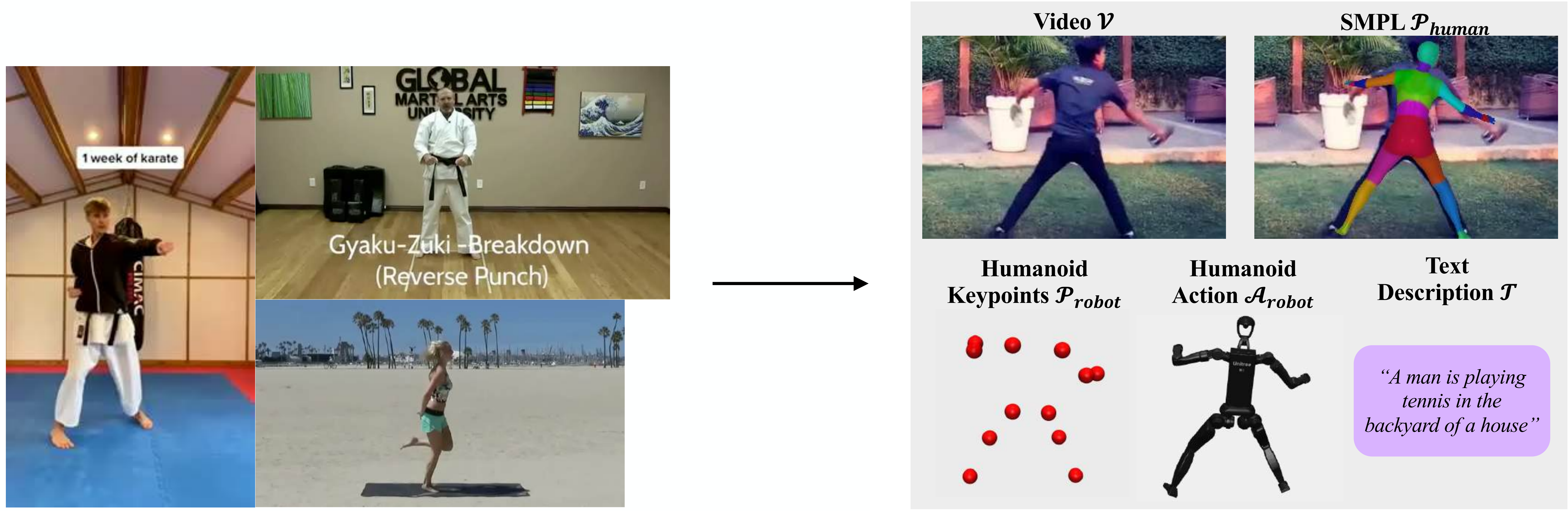
How to generate robotic data with spatial intelligence techniques?

Use Real-to-Sim Reconstruction



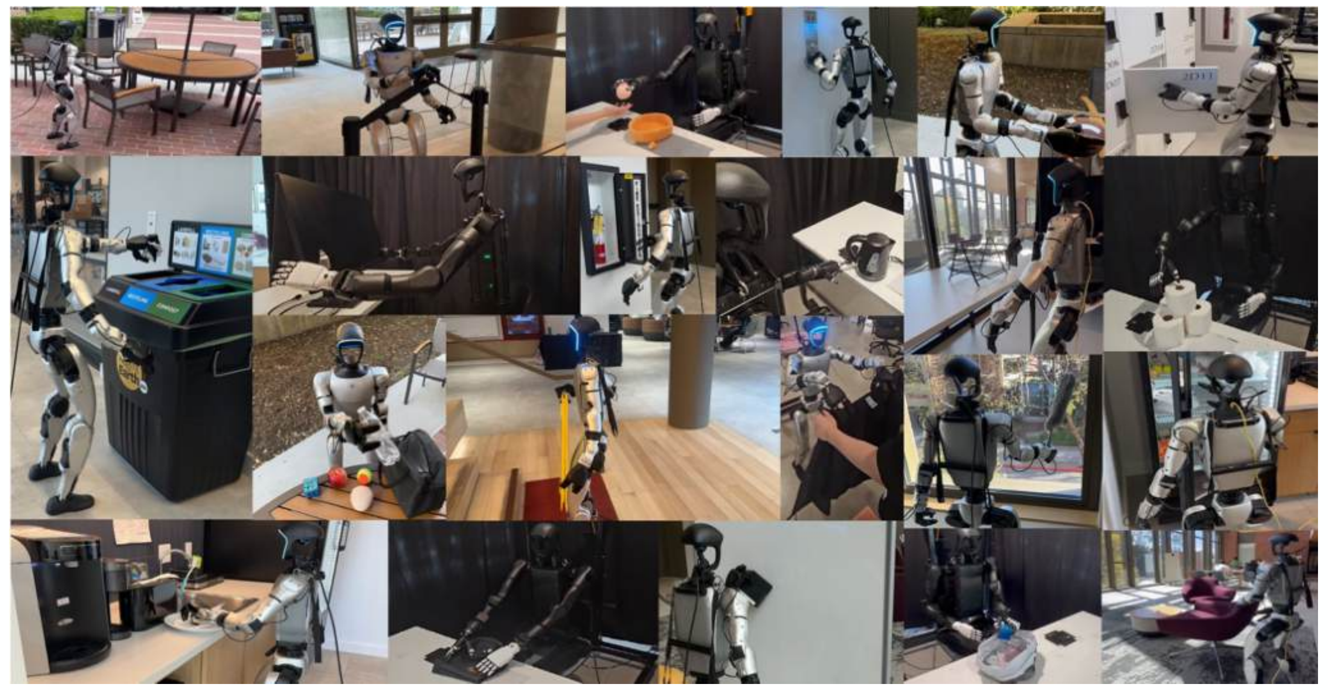
Robot Learning from Any Images. Zhao et al. CoRL 2025.

Leverage Human Data



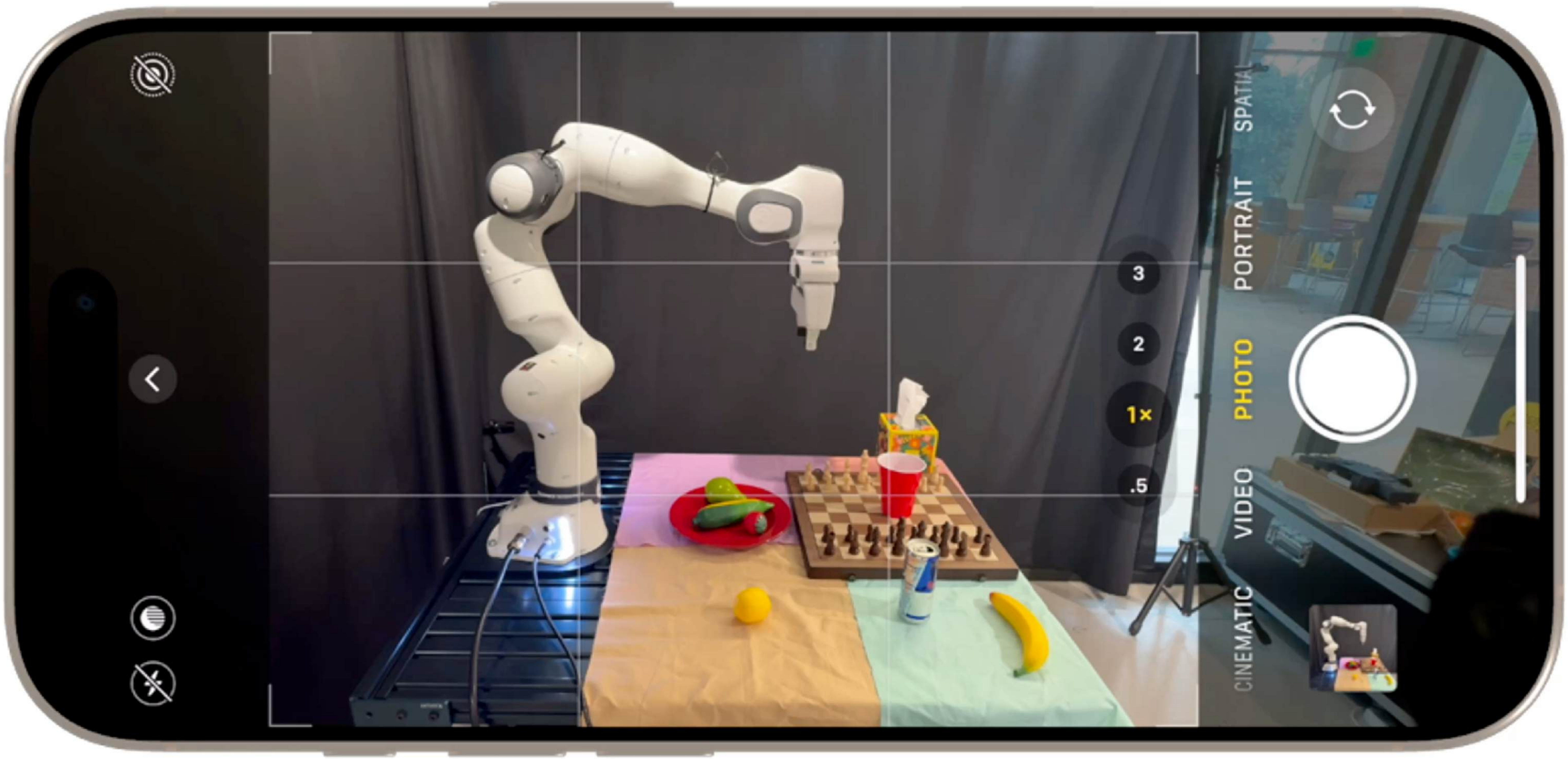
Learning from Massive Human Videos for Universal Humanoid Pose Control. Mao et al. Humanoids 2025.

Scale Teleoperation Data



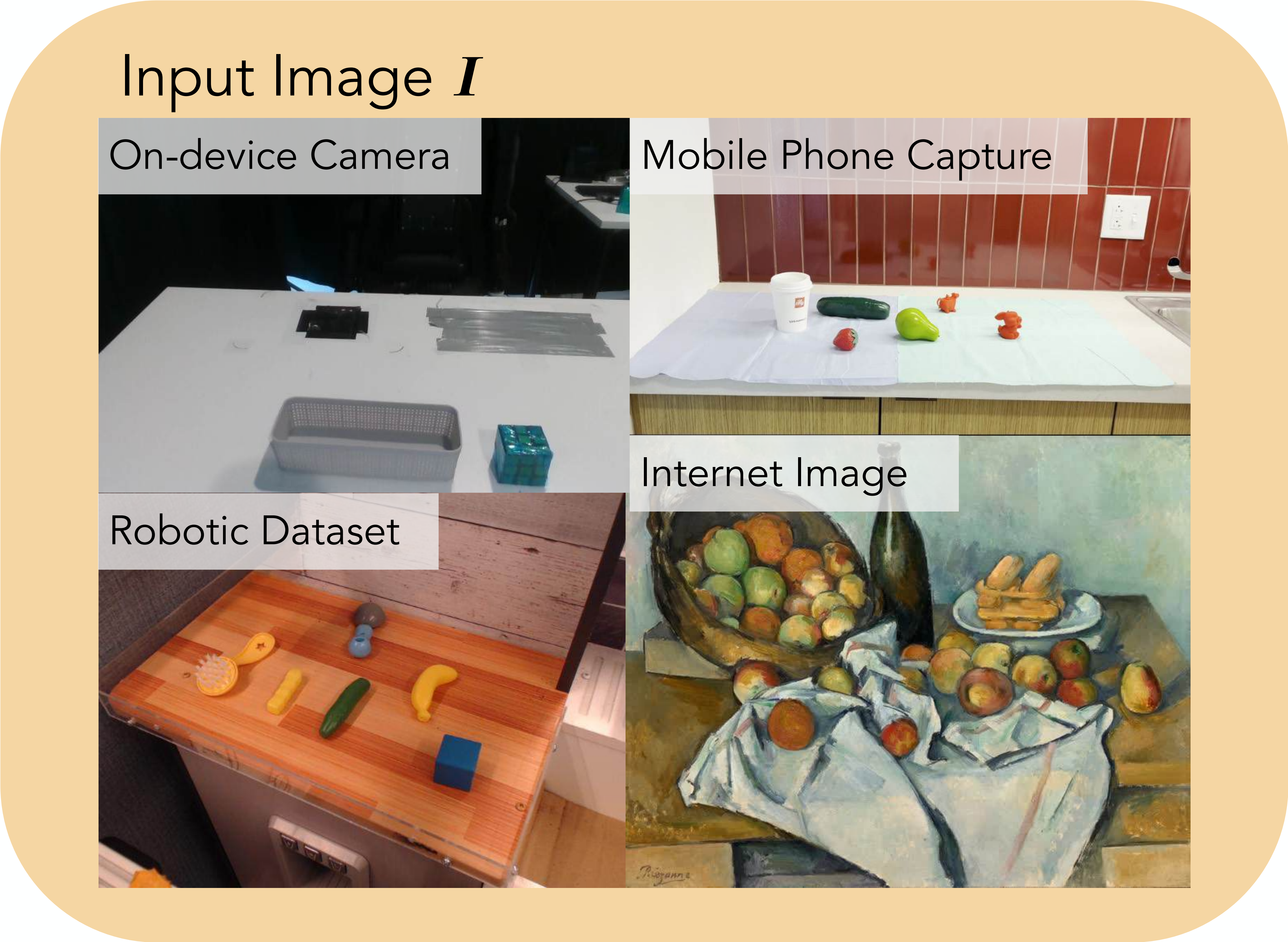
Humanoid Everyday. Jing et al. In submission.

Robot Learning from Any Images



Robot Learning from Any Images. Zhao et al. CoRL 2025.

Step-1: Recovering the Physical Scene from a Single Image



Step-1: Recovering the Physical Scene from a Single Image

Input Image $\boldsymbol{I}$

Segmentation & Inpainting

Metric Depth Prediction & Point Cloud Generation

Step-1: Recovering the Physical Scene from a Single Image

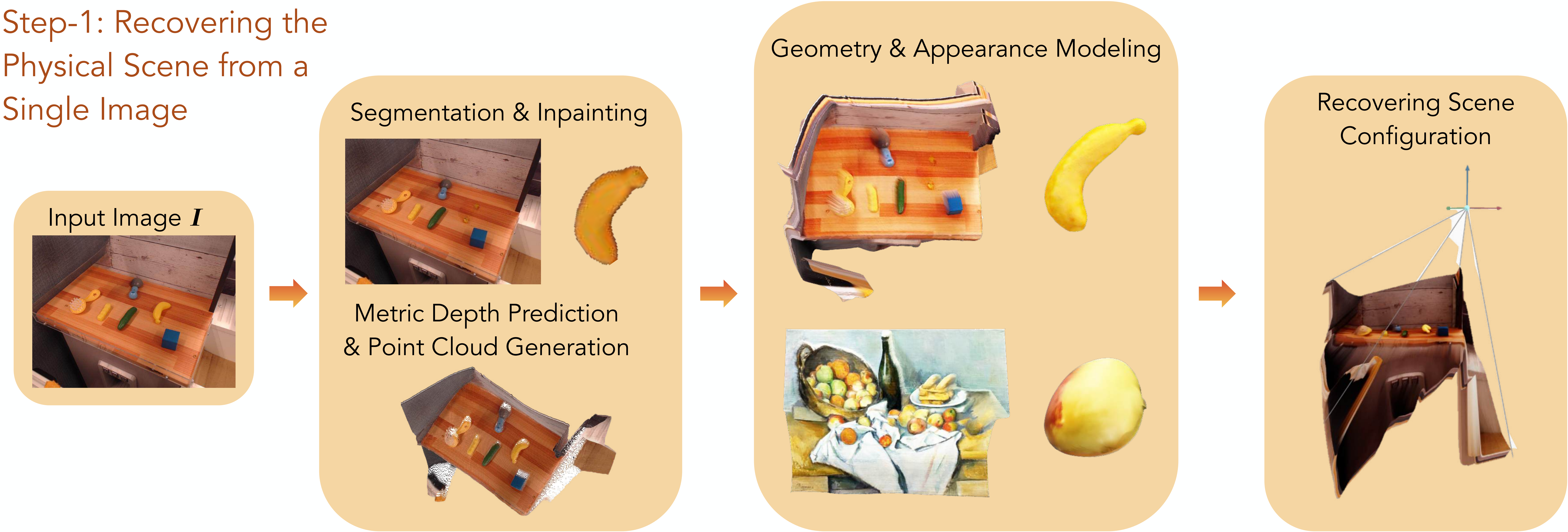
Input Image $\boldsymbol{I}$

Segmentation & Inpainting

Metric Depth Prediction & Point Cloud Generation

Geometry & Appearance Modeling

Step-1: Recovering the Physical Scene from a Single Image



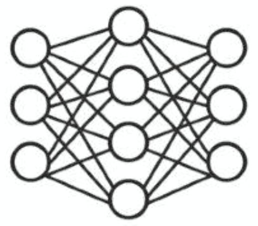

Step-1: Recovering the Physical Scene from a Single Image
Input Image I
Segmentation & Inpainting
Metric Depth Prediction & Point Cloud Generation
Geometry & Appearance Modeling
Recovering Scene Configuration
Physical Property Estimation & Robot Placement


Step-1: Recovering the Physical Scene from a Single Image
Input Image I
Segmentation & Inpainting
Metric Depth Prediction & Point Cloud Generation
Geometry & Appearance Modeling
Recovering Scene Configuration
Physical Property Estimation & Robot Placement
Step-2: Scalable Robotic Data Generation in Sim
Robotic Data Generation
clideo.com


Step-1: Recovering the Physical Scene from a Single Image
Input Image I
Segmentation & Inpainting
Metric Depth Prediction & Point Cloud Generation
Geometry & Appearance Modeling
Recovering Scene Configuration
Physical Property Estimation & Robot Placement
Step-2: Scalable Robotic Data Generation in Sim
Visual Blending
Robotic Data Generation
clideo.com


Step-1: Recovering the Physical Scene from a Single Image
Input Image I
Segmentation & Inpainting
Metric Depth Prediction & Point Cloud Generation
Geometry & Appearance Modeling
Recovering Scene Configuration
Physical Property Estimation & Robot Placement
Step-3: Robot Learning & Deployment
Real-World Deployment
Single-Image IL
VLA
Manip. Priors
Robotic Data
Robotic Images
Camera Photographs
Internet Images
Step-2: Scalable Robotic Data Generation in Sim
Visual Blending
Robotic Data Generation

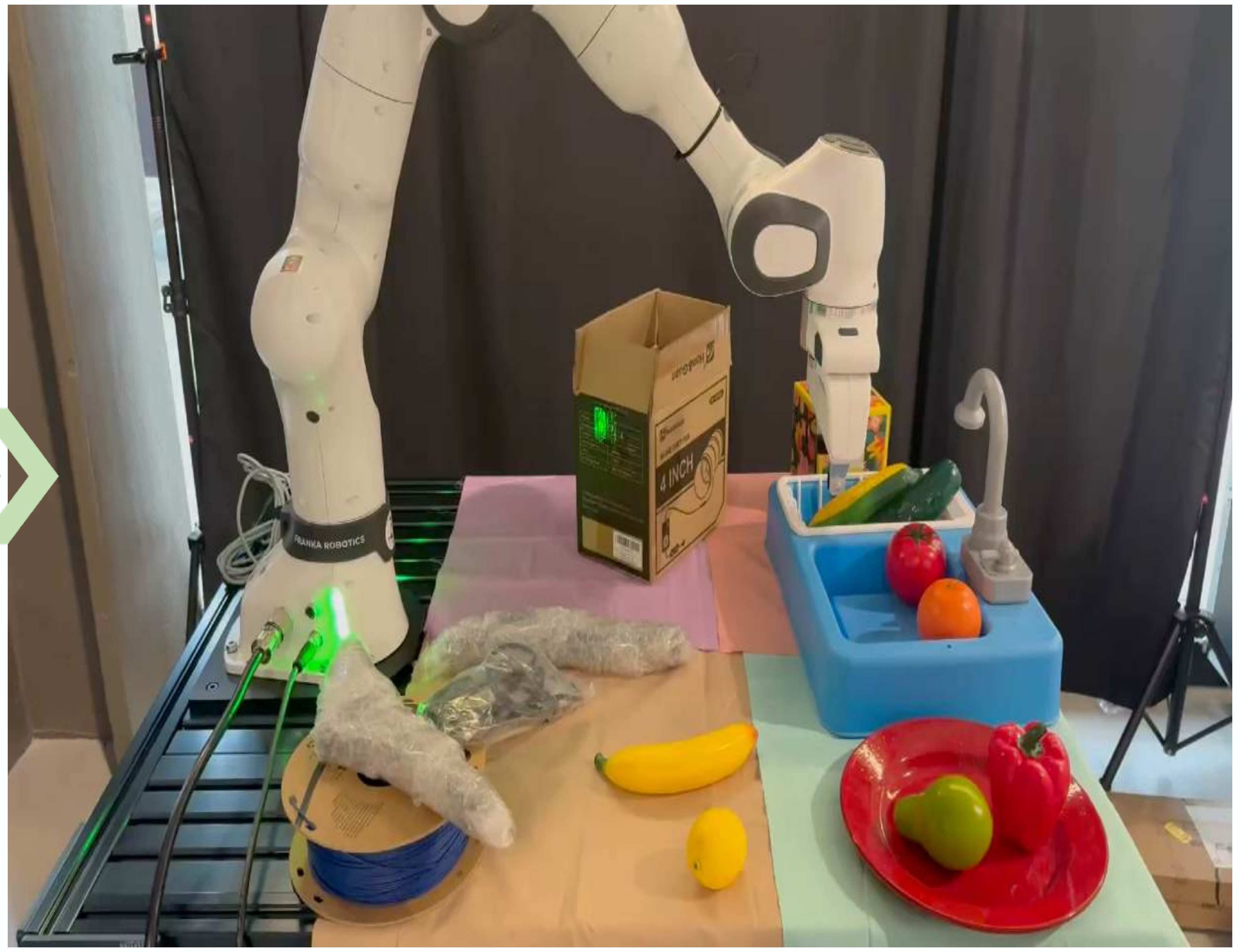
Single-
Image
Imitation



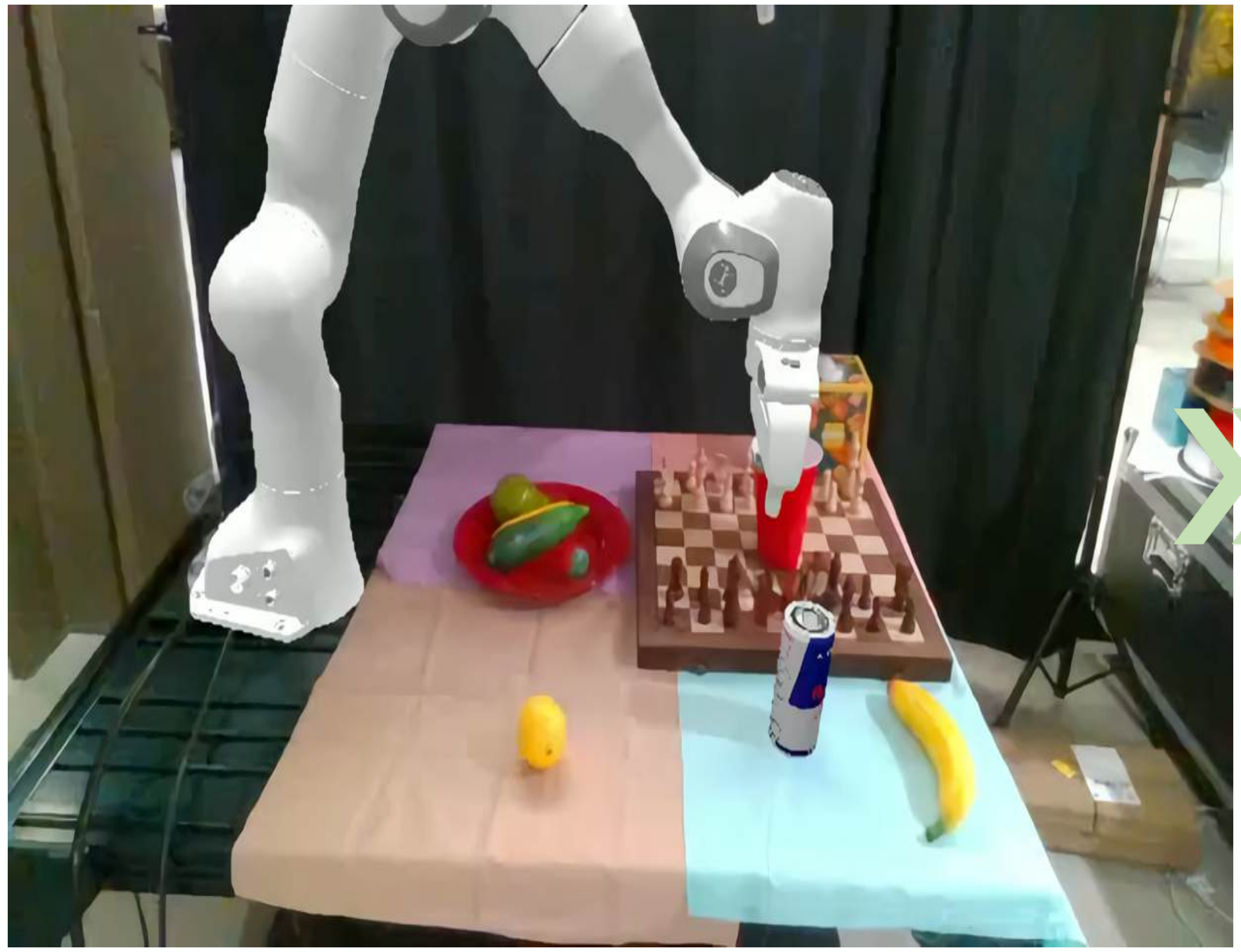
RoLA-Generated Data @ Sim

Real-world Deploy

Manipulation in
Cluttered
Scenes



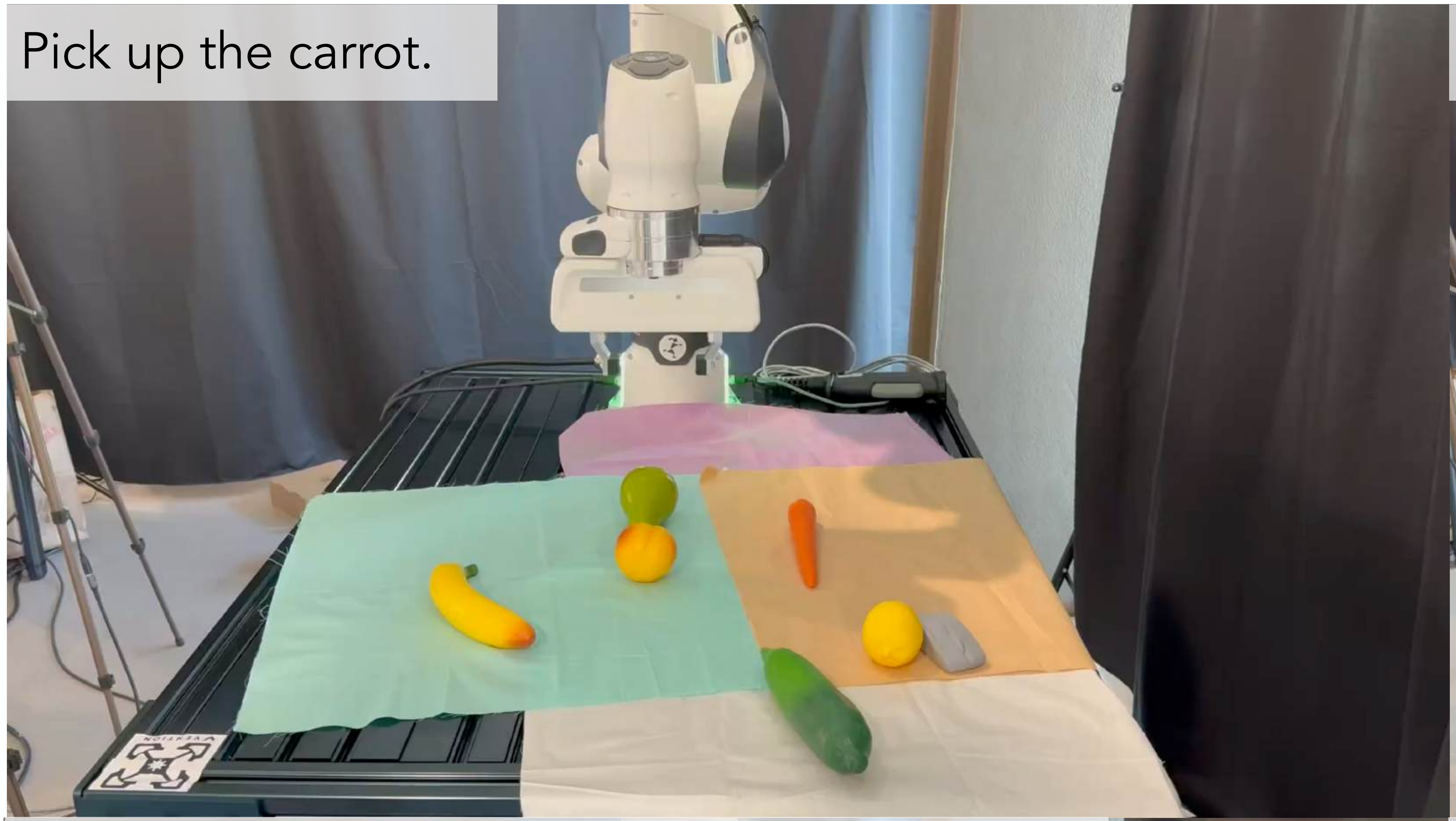
Pour Water

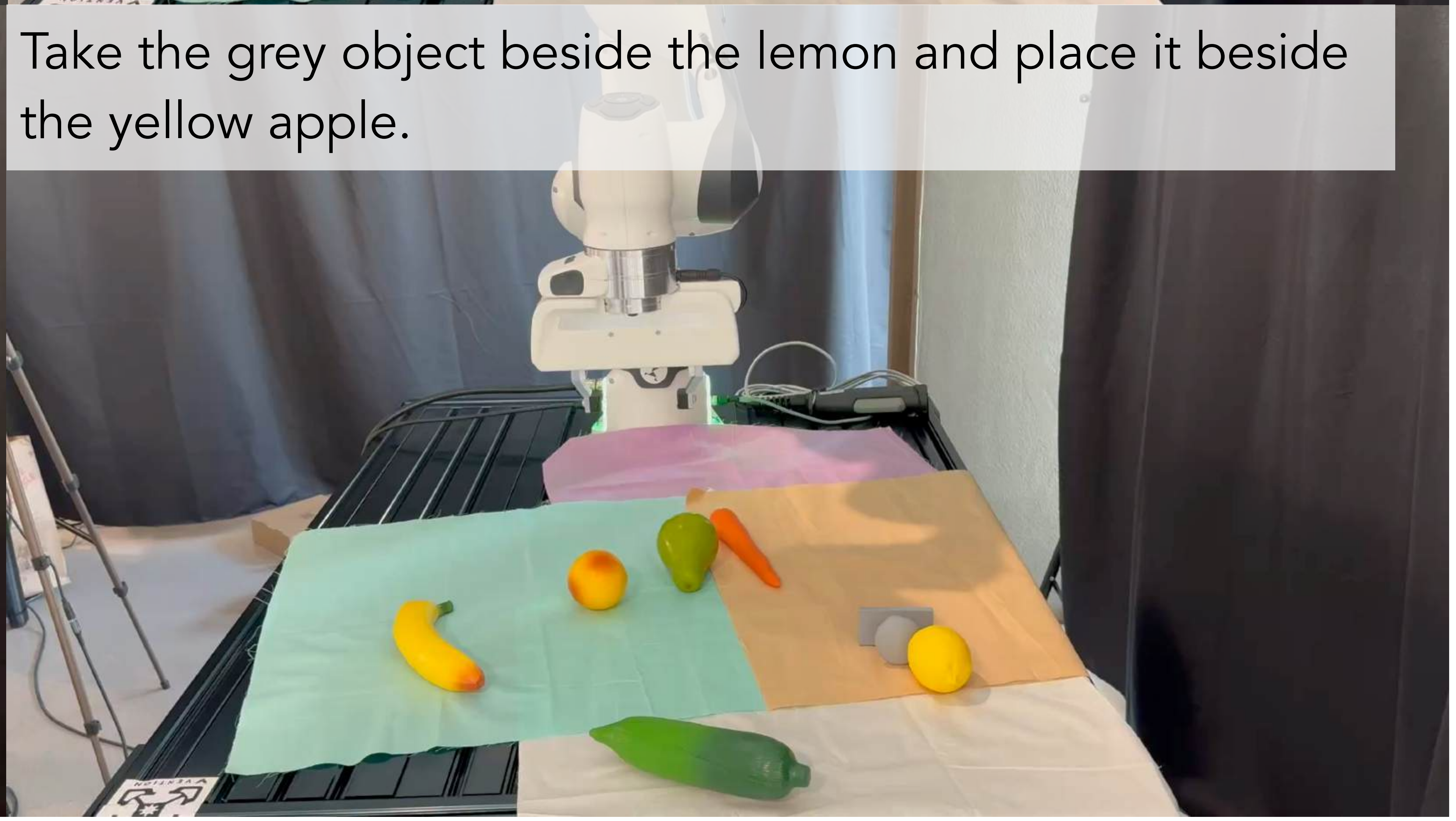


VLA

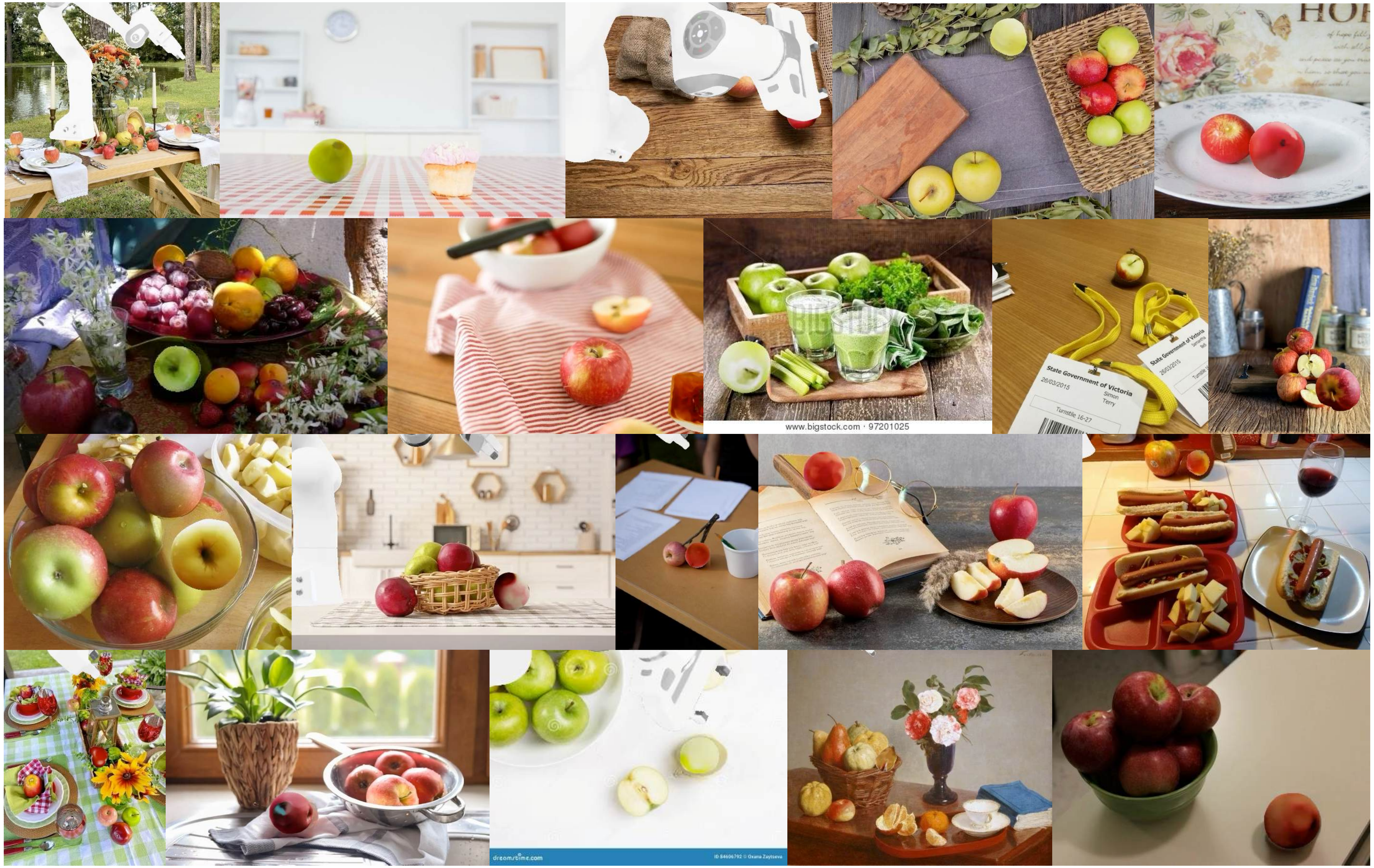
Data Collection

VLA Real-World Deployment





Manipulation Prior



Manipulation Prior

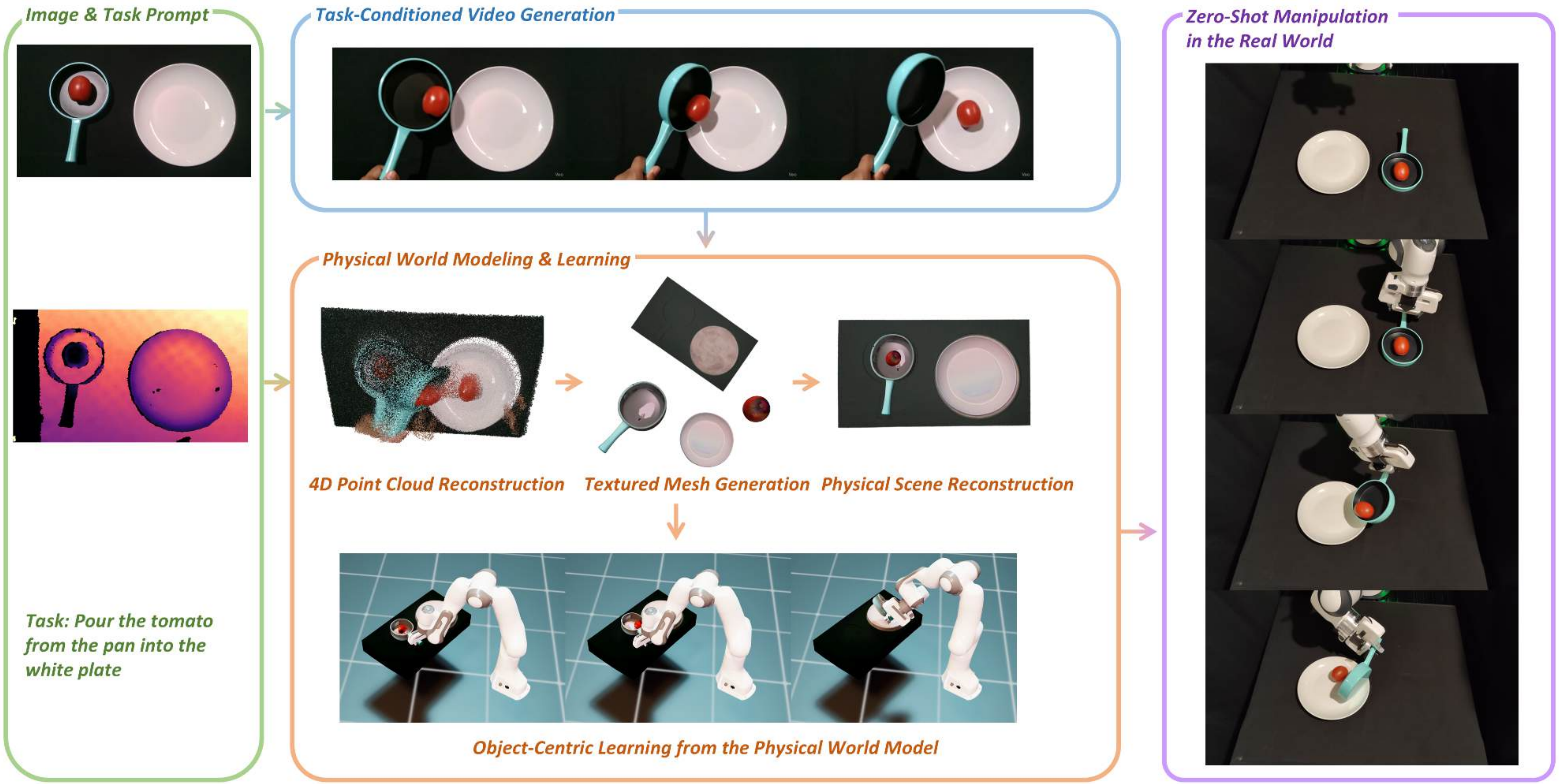
Manipulation Prior

Robot Learning from Any Images

- Data quantity and diversity are widely recognized as primary bottlenecks in scaling robot learning.
- Collecting **on-robot demonstrations** at scale demands specialized hardware and extensive labor. ☹
- Obtain robot-complete data from non-robotic images under minimal assumptions: **single image**. ☺

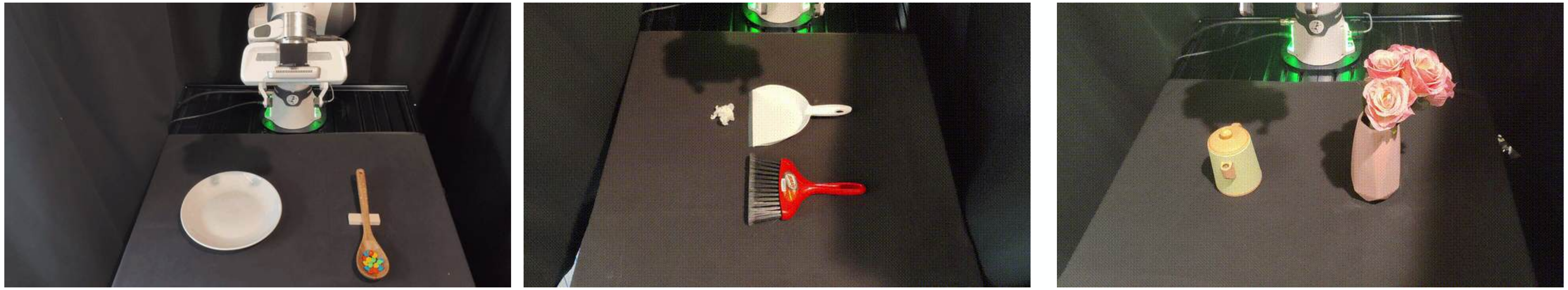
Robot Learning from Any Images. Zhao et al. CoRL 2025.

Robot Learning from A Physical World Model



Robot Learning from A Physical World Model

Video generation



Robot execution

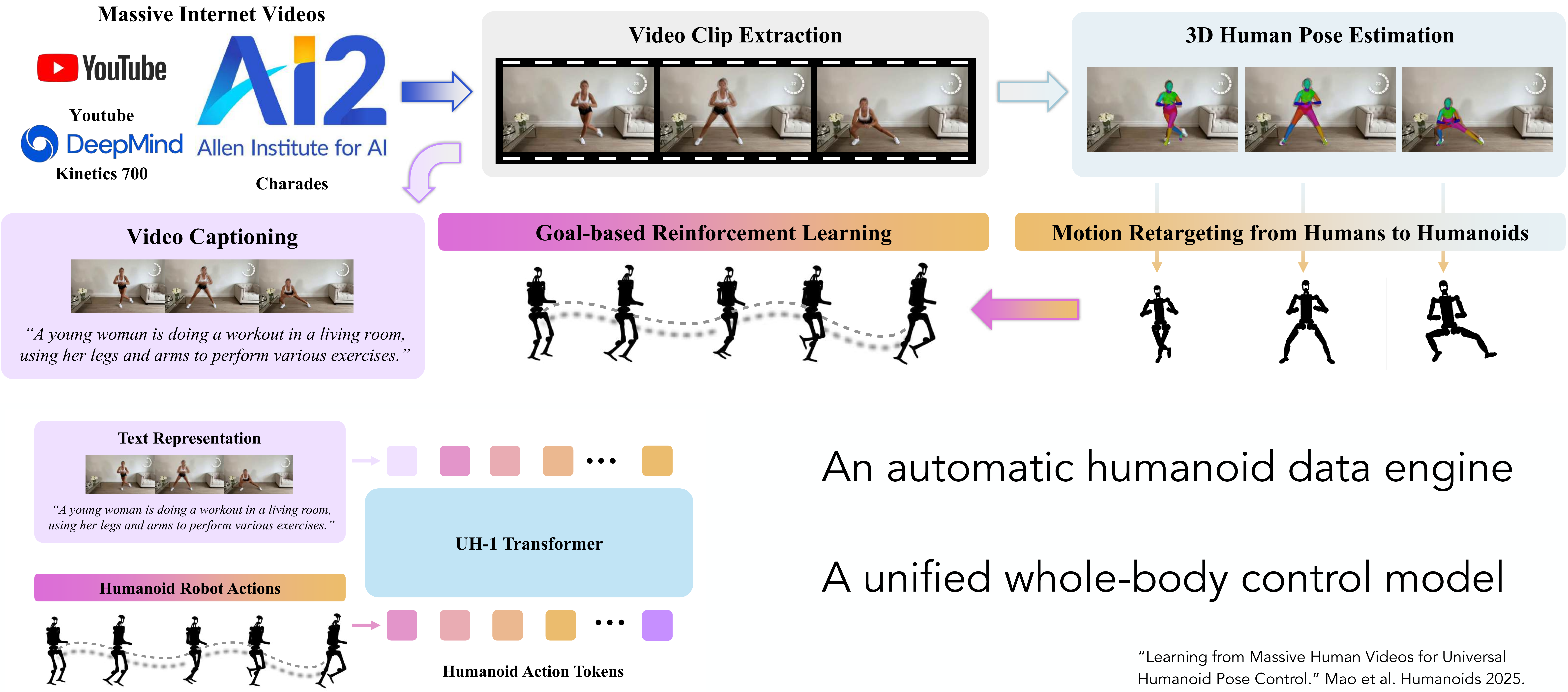
More DoFs

Not easily handled by motion model

Action retargeting is hard

How can we derive humanoid data from Internet data?

UH-1: Learning from Massive Human Videos for Universal Humanoid Pose Control



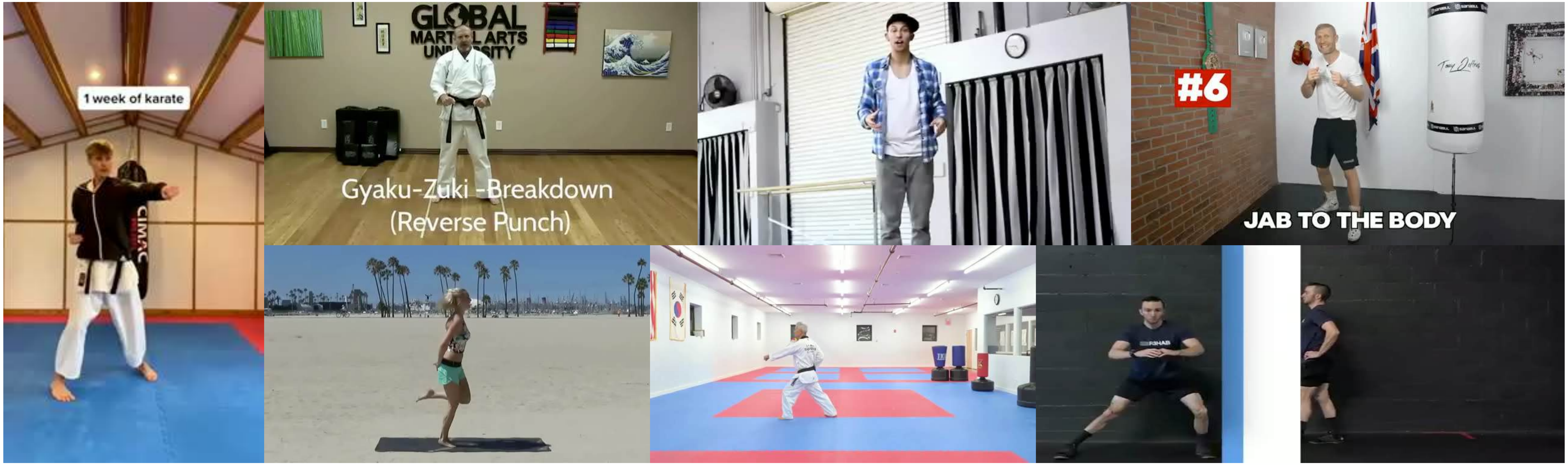
An automatic humanoid data engine

A unified whole-body control model

"Learning from Massive Human Videos for Universal Humanoid Pose Control." Mao et al. Humanoids 2025.

Data Collection

We collect 163, 800 video clips from diverse sources.

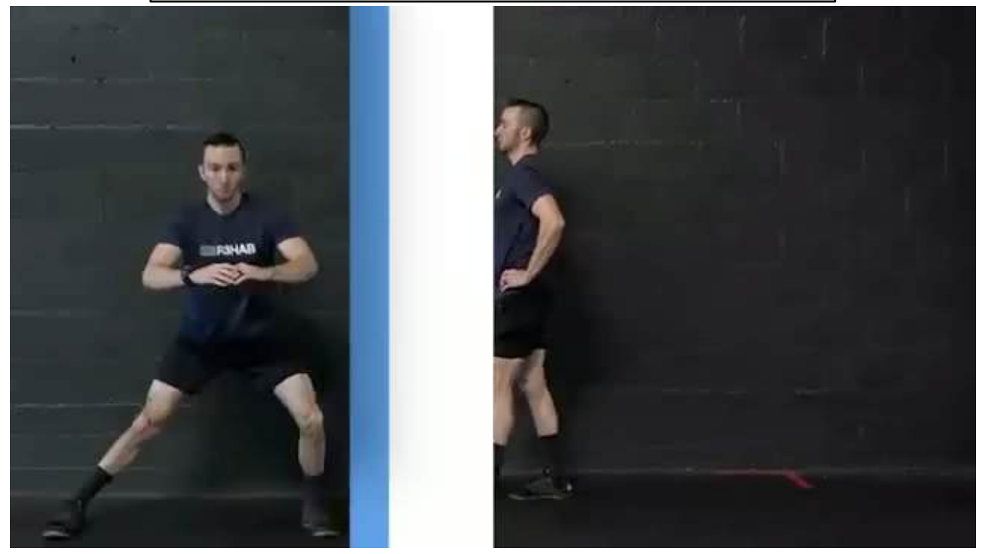


Data Collection

Videos are further annotated with captioning tools.

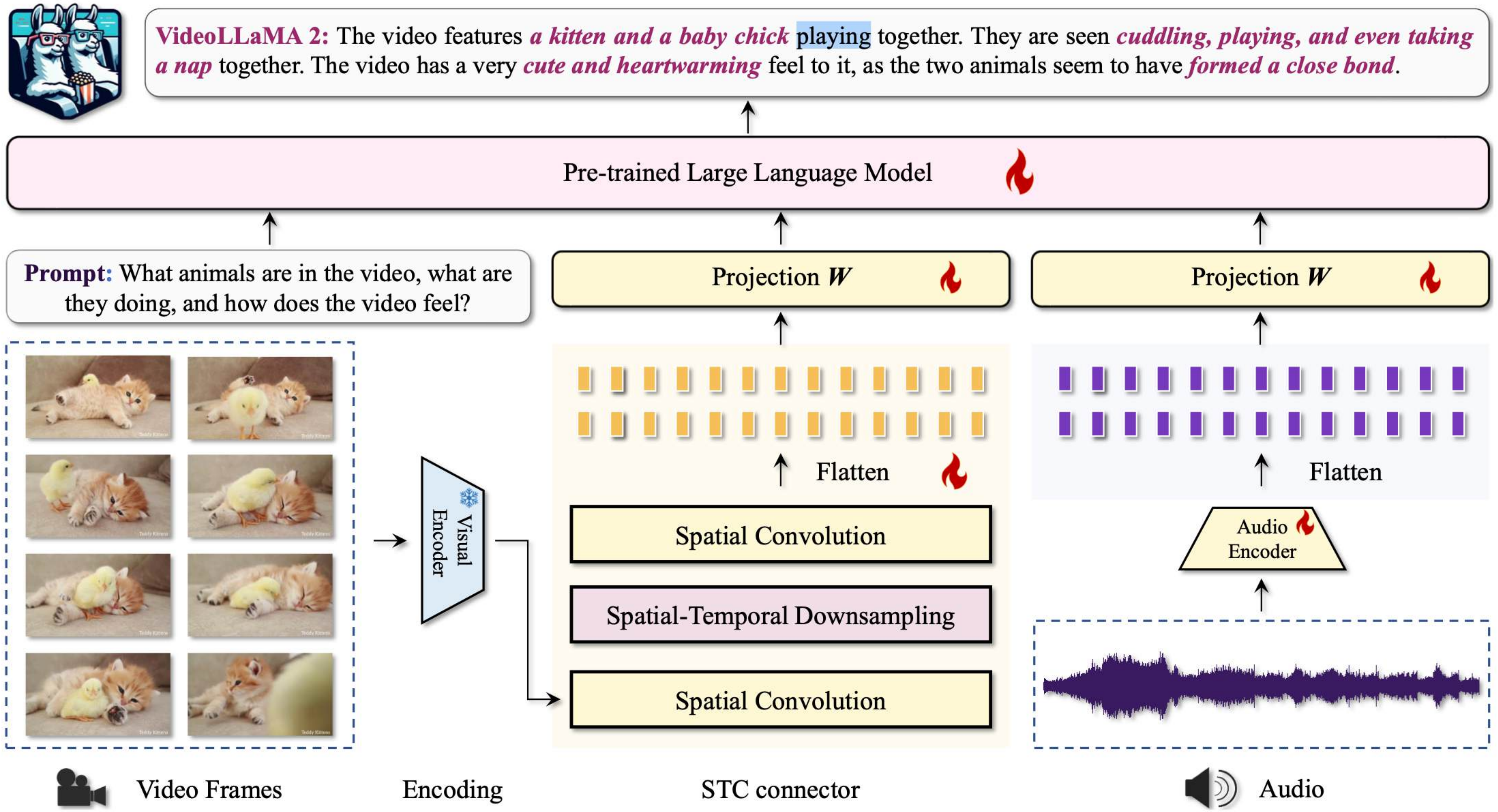
"practicing martial arts, standing."

"standing and speaking."



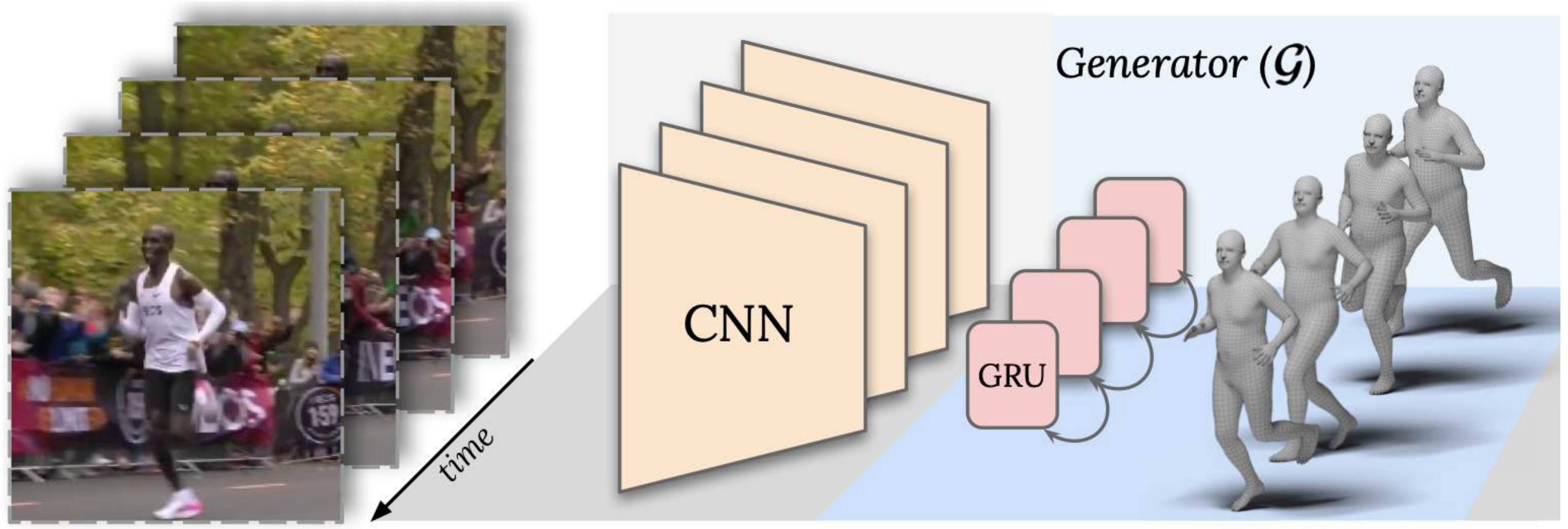
"practicing yoga."

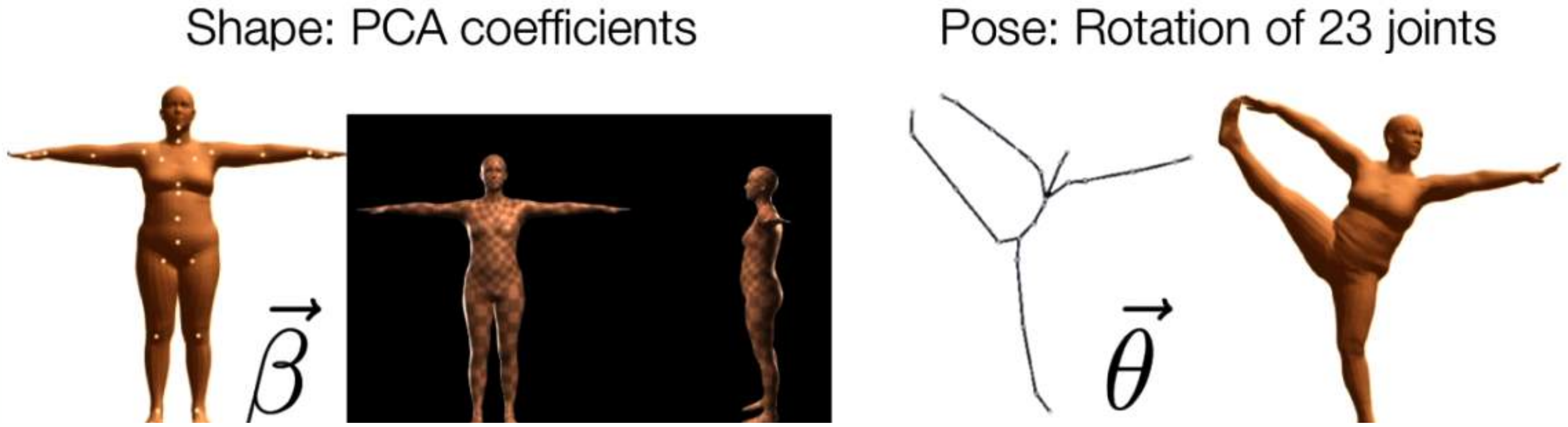
"doing squats, lunges, and jumping jacks."



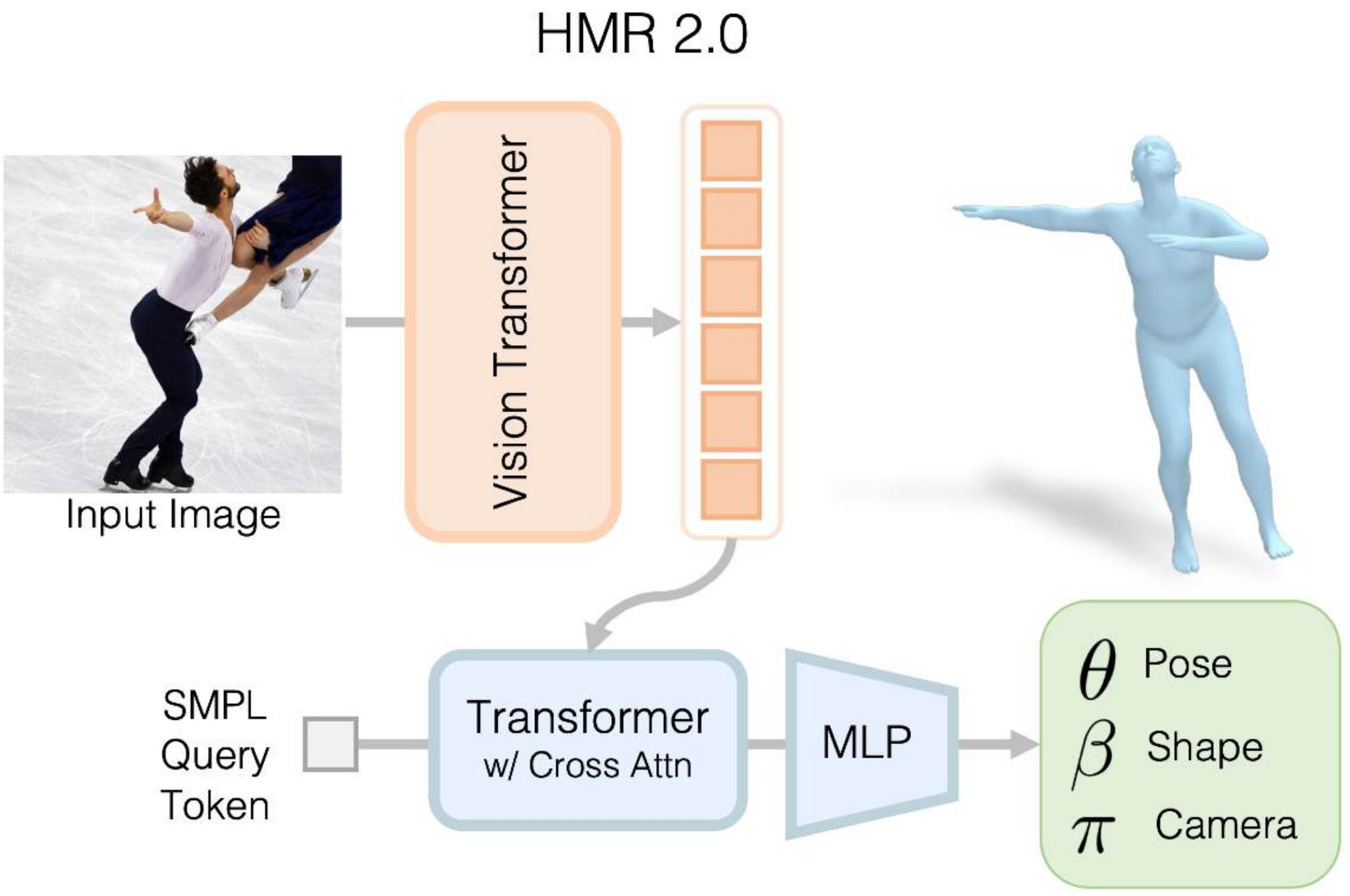
[Preprint 2024] "VideoLLaMA 2: Advancing Spatial-Temporal Modeling and Audio Understanding in Video-LLMs." Cheng et al.

Human Motion Representation



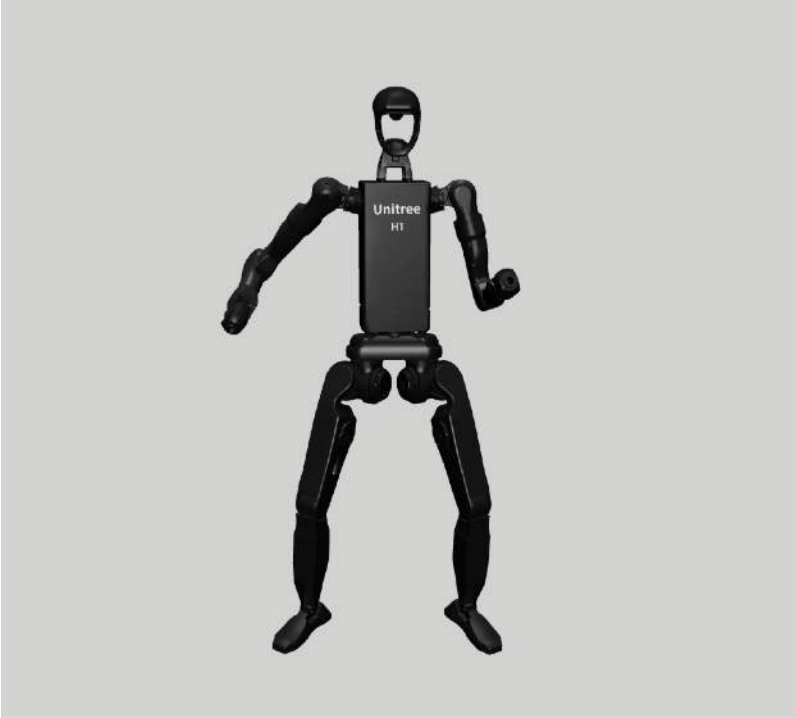


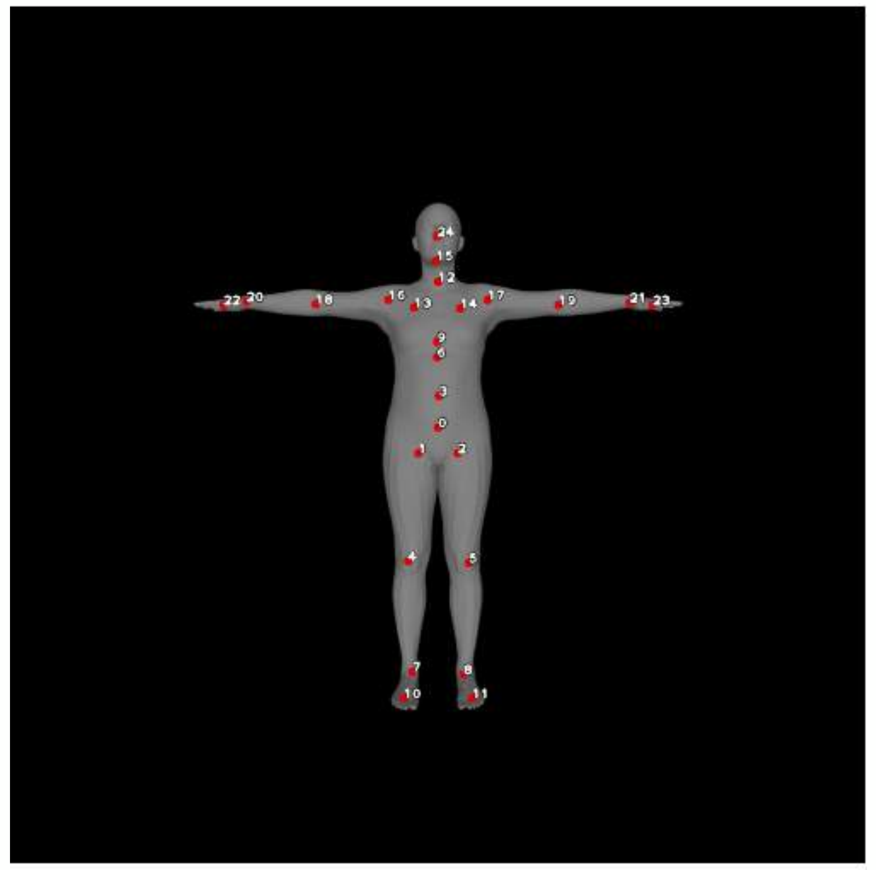
SMPL Model

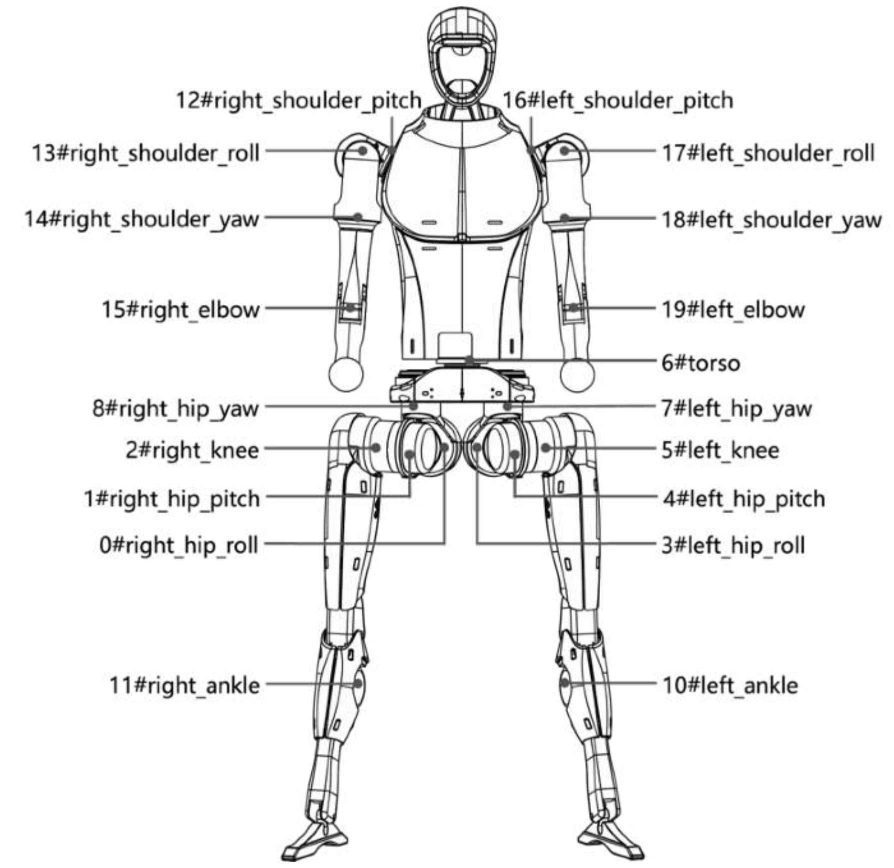


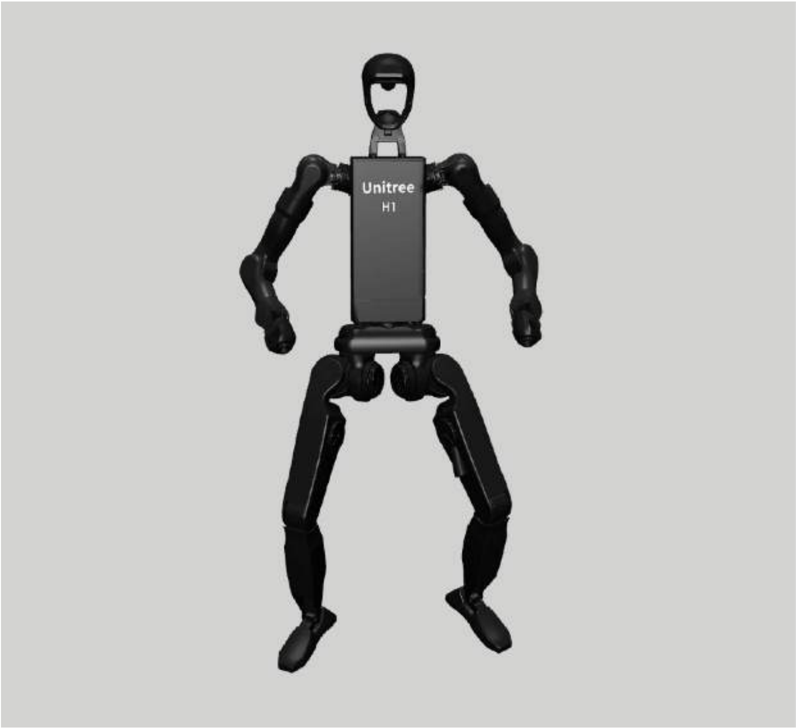
[Kocabas et al., CVPR 2020]
[Goel et al., CVPR 2024]

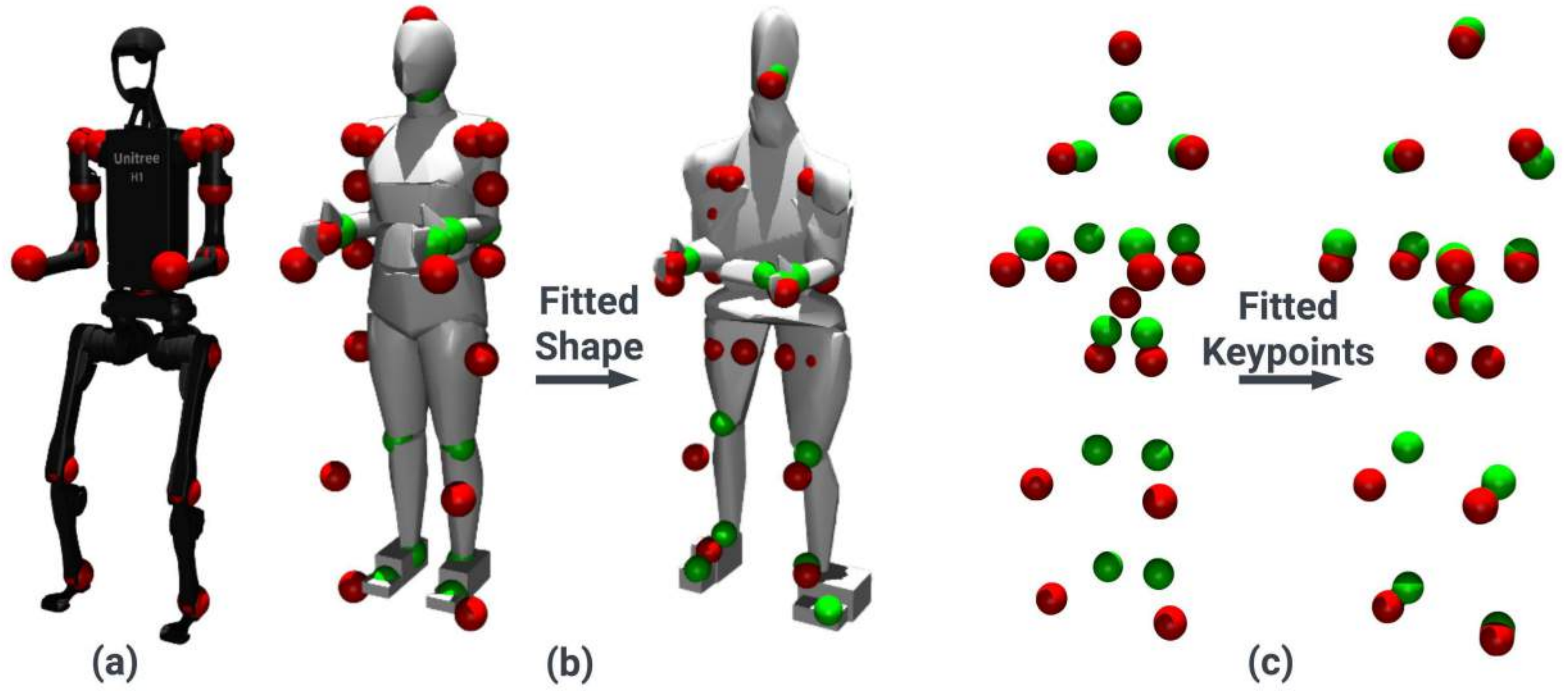
Human-to-Humanoid Motion Retargeting





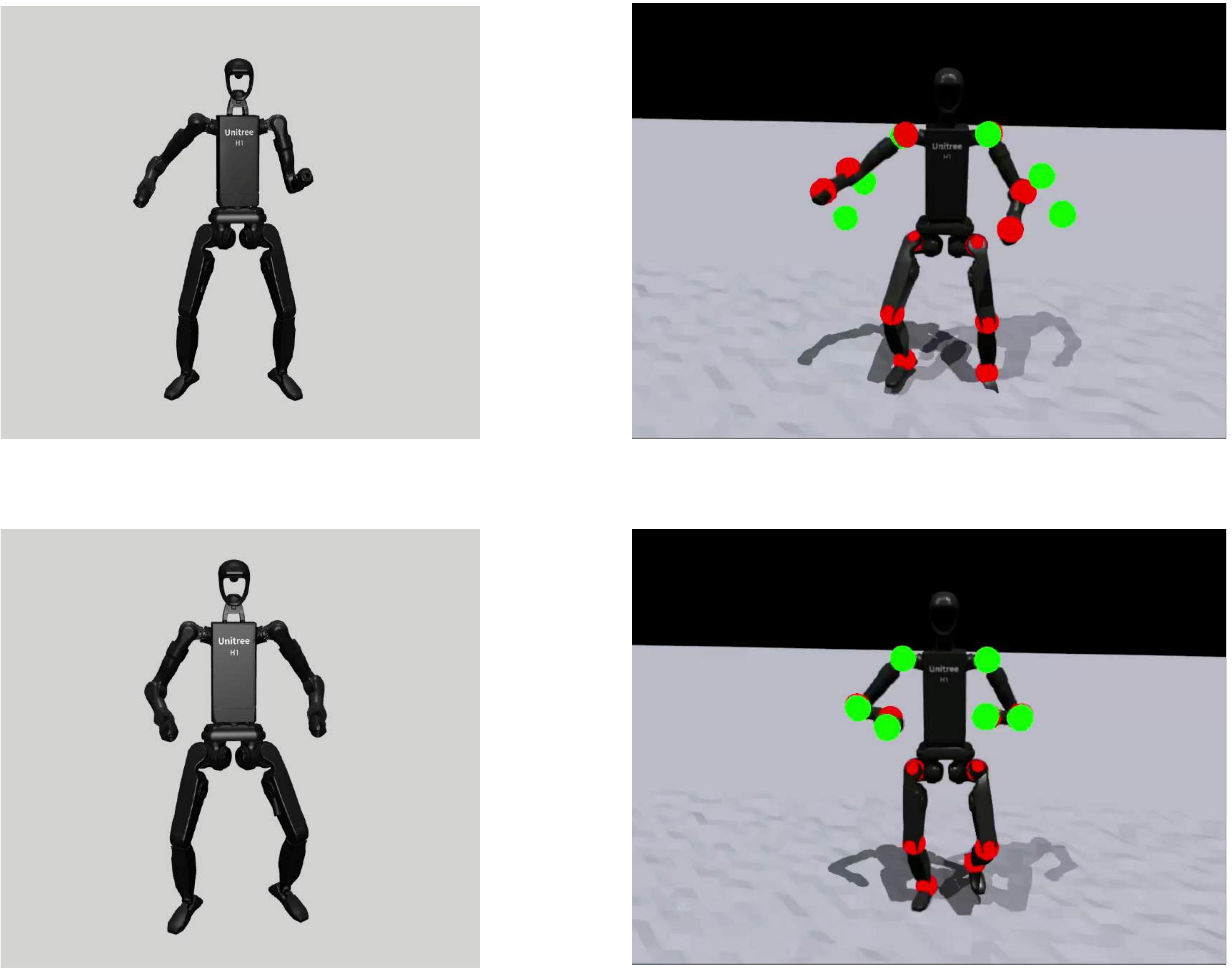


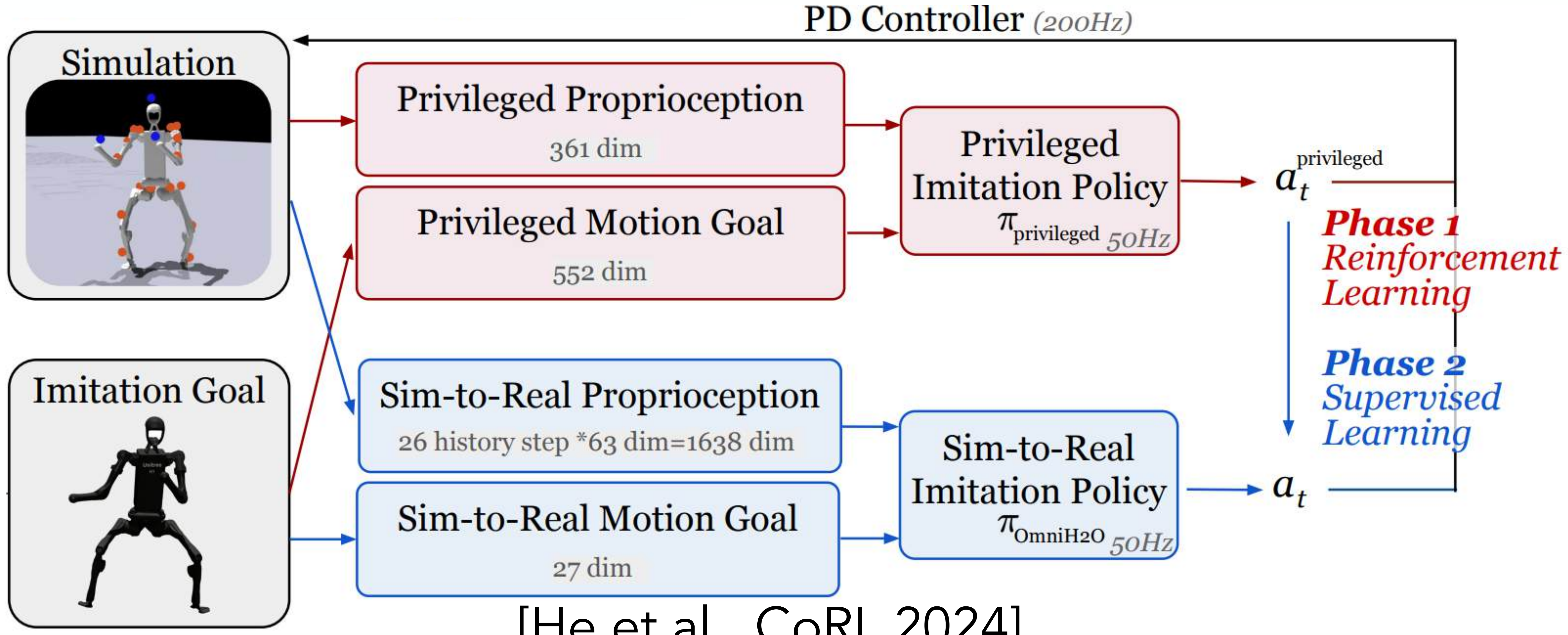




$$
\begin{aligned}
&\min_{\beta} ||\mathcal{P}^T_{joints} - \mathcal{P}^T_{robot}||_2, \\
&\text{s.t.} \quad \mathcal{P}^T_{joints} = F_{fk}(\mathcal{P}_{human}(\beta, \theta^T, t_{root})),
\end{aligned}
$$

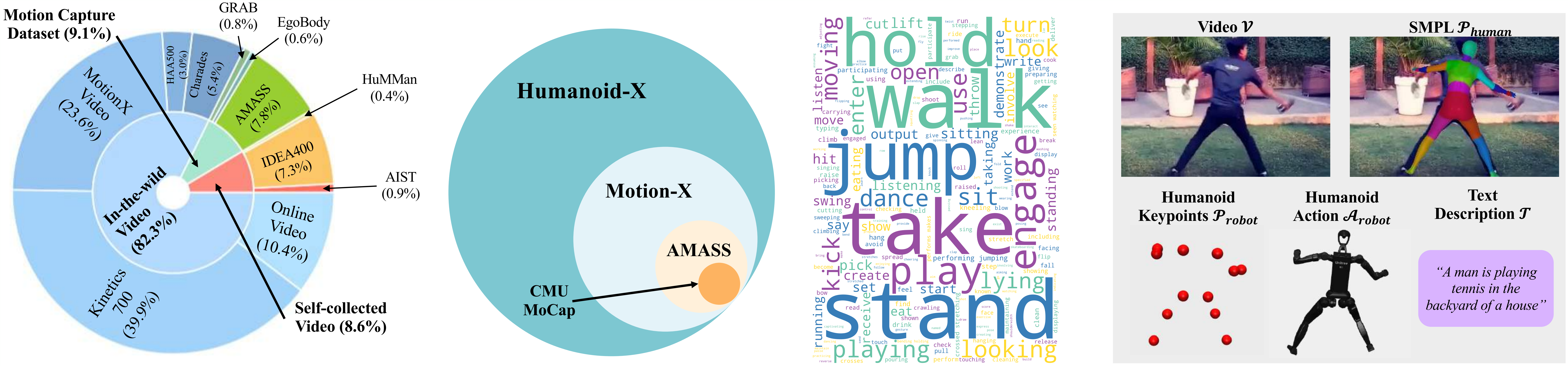
Sim-to-Real Adaptation





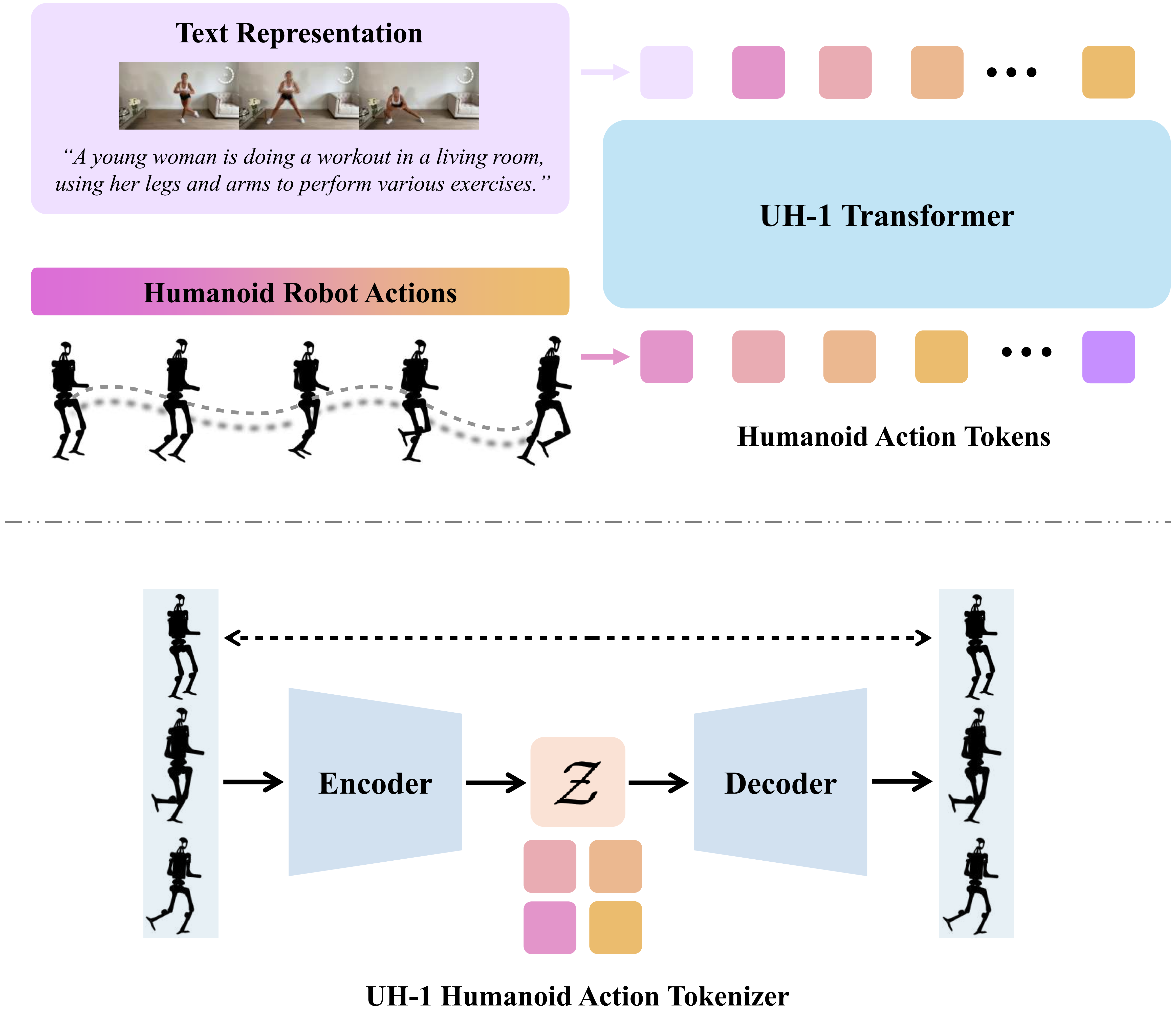
[He et al., CoRL 2024]

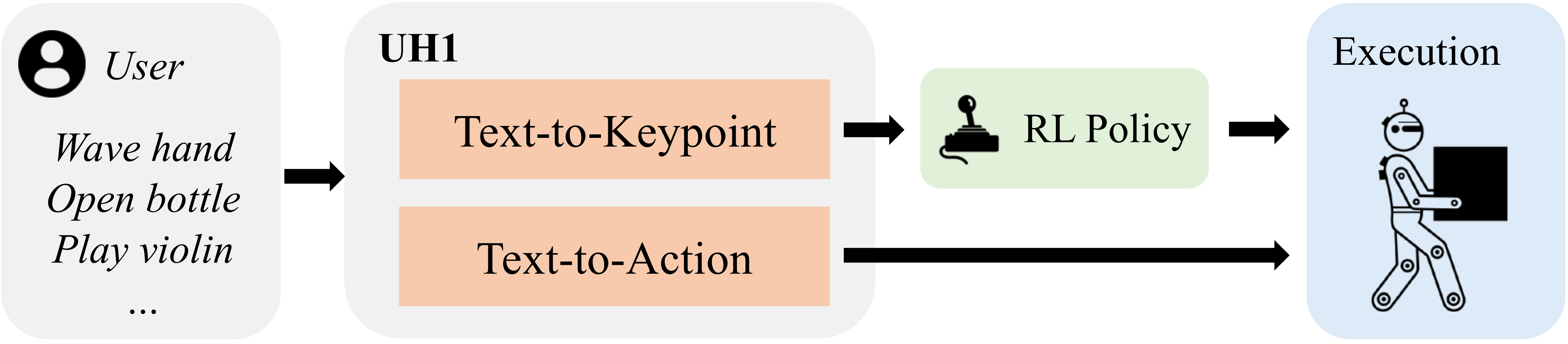
Dataset



(a) Data Distribution (b) Data Scale (c) Vocabulary Diversity (d) Motion Sample

Universal Humanoid (UH-1) Architecture





One Model, Two Control Modes

Research Questions

- **Universal Pose Control with UH-1**: Does UH-1 model enable universal humanoid robot pose control based on text commands?

- **Scalability and Generalization with Humanoid-X**: Does the large-scale Humanoid-X dataset facilitate scalable training and improve the generalization ability of UH-1?

- **Real-World Deployment of UH-1**: Can UH-1 model be deployed on real humanoid robots to enable reliable robotic control in real-world environments?

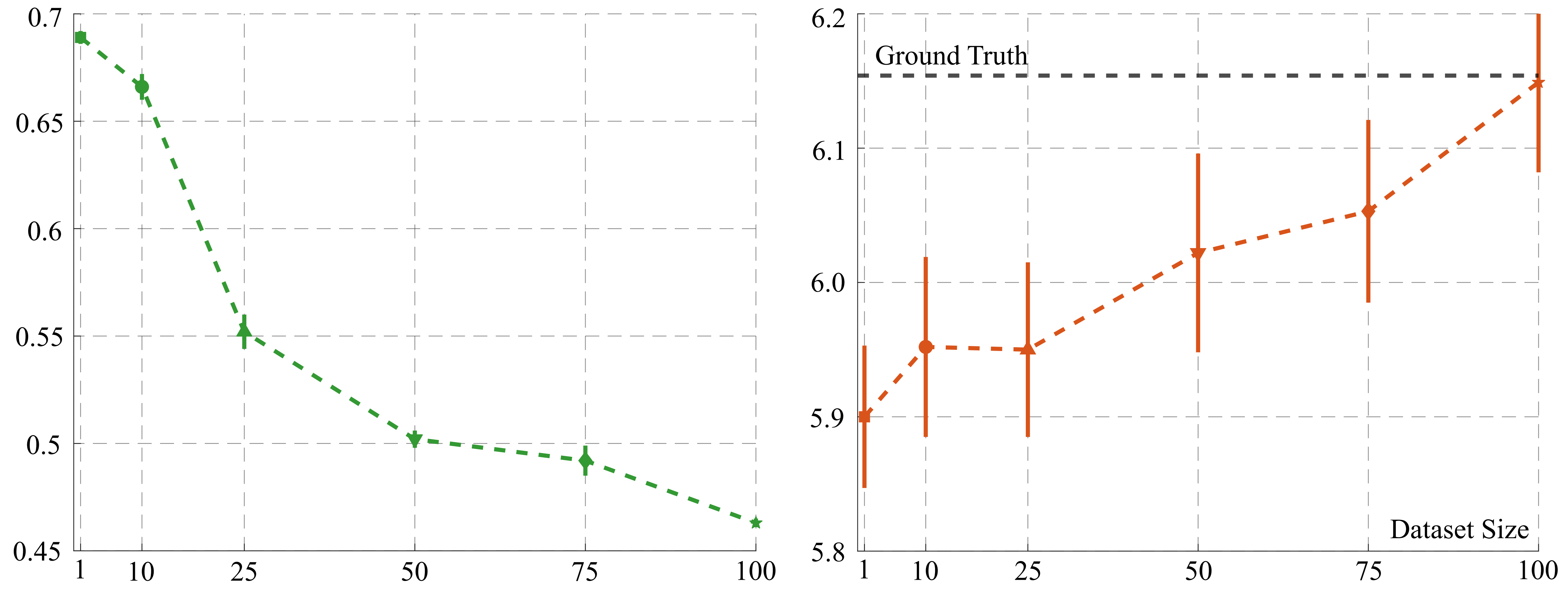
Universal Pose Control with UH-1

- Baseline models: Motion Diffusion Model (MDM) and Text-to-Motion GPT (T2M-GPT)

Methods	FID ↓	MM Dist ↓	Diversity ↑	R Precision ↑
Oracle	$0.005^{\pm.001}$	$3.140^{\pm.010}$	$9.846^{\pm.062}$	$0.780^{\pm.003}$
MDM [57]	$0.582^{\pm.051}$	$5.921^{\pm.034}$	$10.122^{\pm.078}$	$0.617^{\pm.007}$
T2M-GPT [71]	$0.667^{\pm.109}$	$3.401^{\pm.017}$	$\mathbf{10.328}^{\pm\mathbf{.099}}$	$0.734^{\pm.004}$
UH-1 (ours)	$\mathbf{0.445}^{\pm\mathbf{.078}}$	$\mathbf{3.249}^{\pm\mathbf{.016}}$	$10.157^{\pm.106}$	$\mathbf{0.761}^{\pm\mathbf{.003}}$

Scalable Learning with Humanoid-X

- Increasing data size leads to consistent performance improvement.
- Pre-training on Humanoid-X helps generalization.

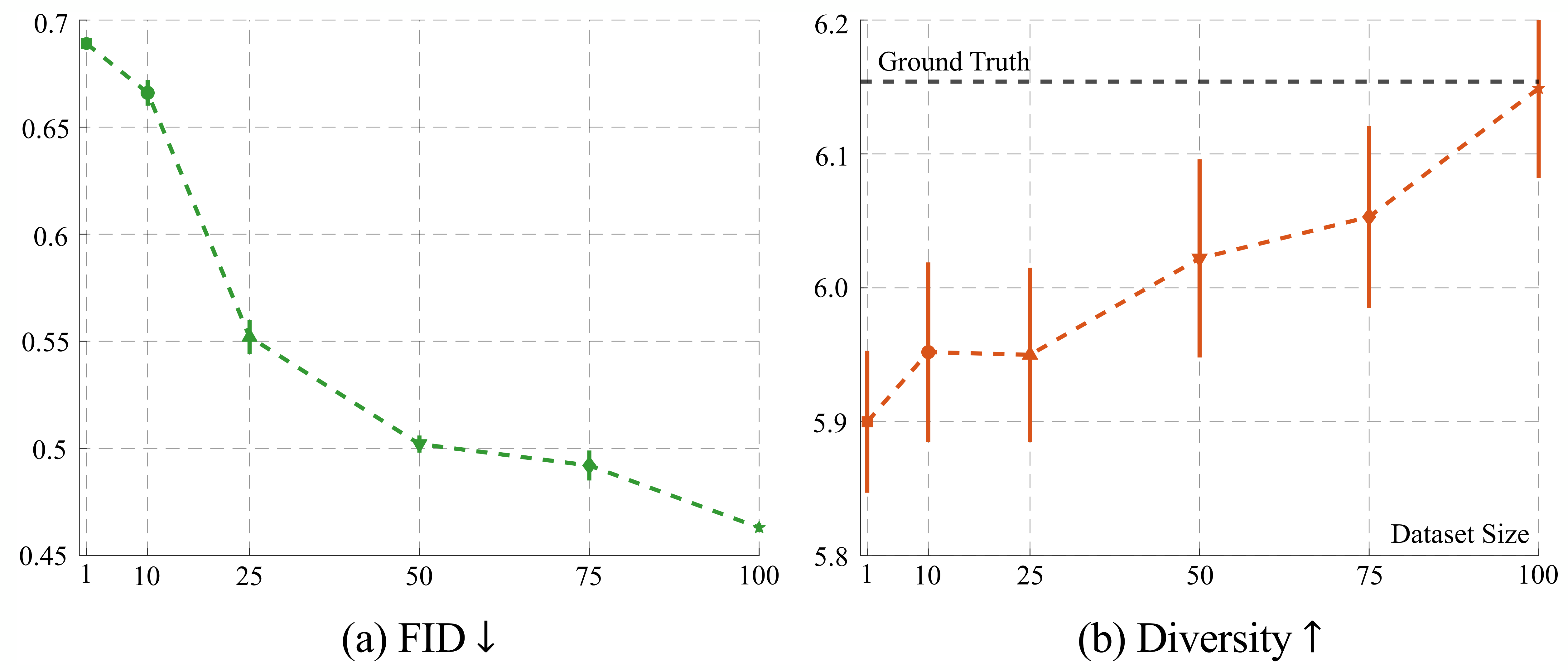


(a) FID ↓ (b) Diversity ↑

Dataset	FID ↓	MM Dist ↓	Diversity ↑	R Precision ↑
Oracle	$0.005^{\pm.001}$	$3.140^{\pm.010}$	$9.846^{\pm.062}$	$0.780^{\pm.003}$
HumanoidML3D	$0.445^{\pm.078}$	$3.249^{\pm.016}$	$10.157^{\pm.106}$	$0.760^{\pm.003}$
Humanoid-X	$\mathbf{0.379}^{\pm.046}$	$\mathbf{3.232}^{\pm.008}$	$\mathbf{10.221}^{\pm.100}$	$\mathbf{0.761}^{\pm.003}$

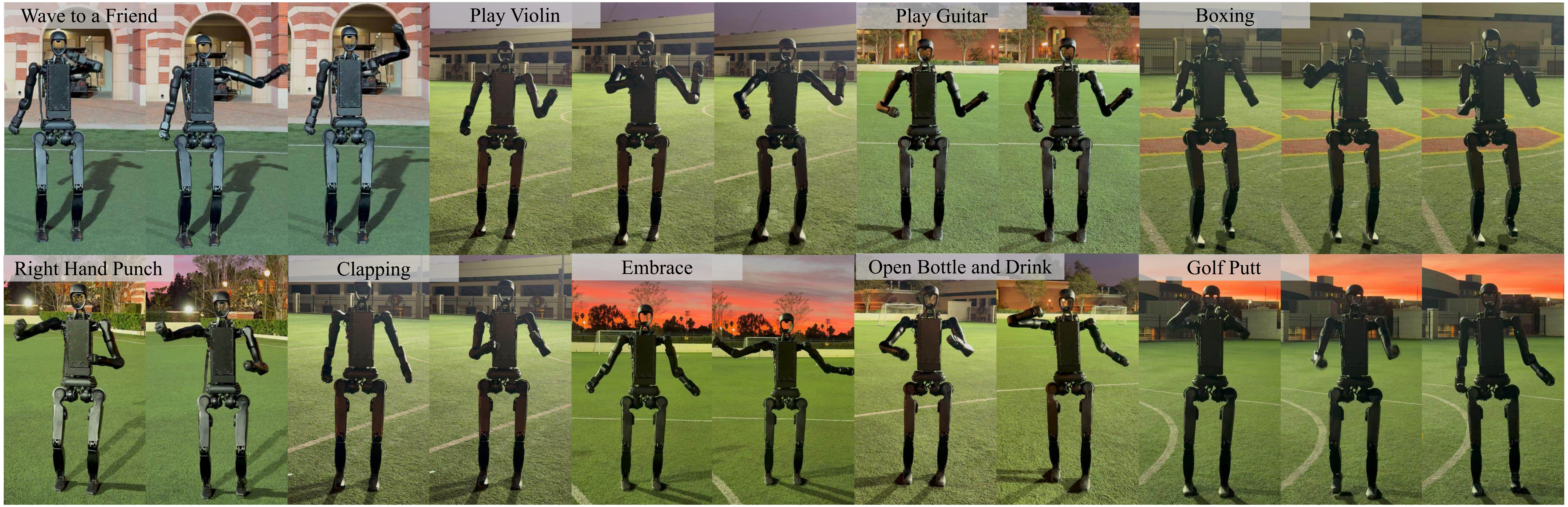
Scalable Learning with Humanoid-X

- Increasing data size leads to consistent performance improvement.
- Pre-training on Humanoid-X helps generalization.



Dataset	FID ↓	MM Dist ↓	Diversity ↑	R Precision ↑
Oracle	$0.005^{\pm.001}$	$3.140^{\pm.010}$	$9.846^{\pm.062}$	$0.780^{\pm.003}$
HumanoidML3D	$0.445^{\pm.078}$	$3.249^{\pm.016}$	$10.157^{\pm.106}$	$0.760^{\pm.003}$
Humanoid-X	$\mathbf{0.379}^{\pm.046}$	$\mathbf{3.232}^{\pm.008}$	$\mathbf{10.221}^{\pm.100}$	$\mathbf{0.761}^{\pm.003}$

Real-World Deployment of UH-1



Humanoid Everyday

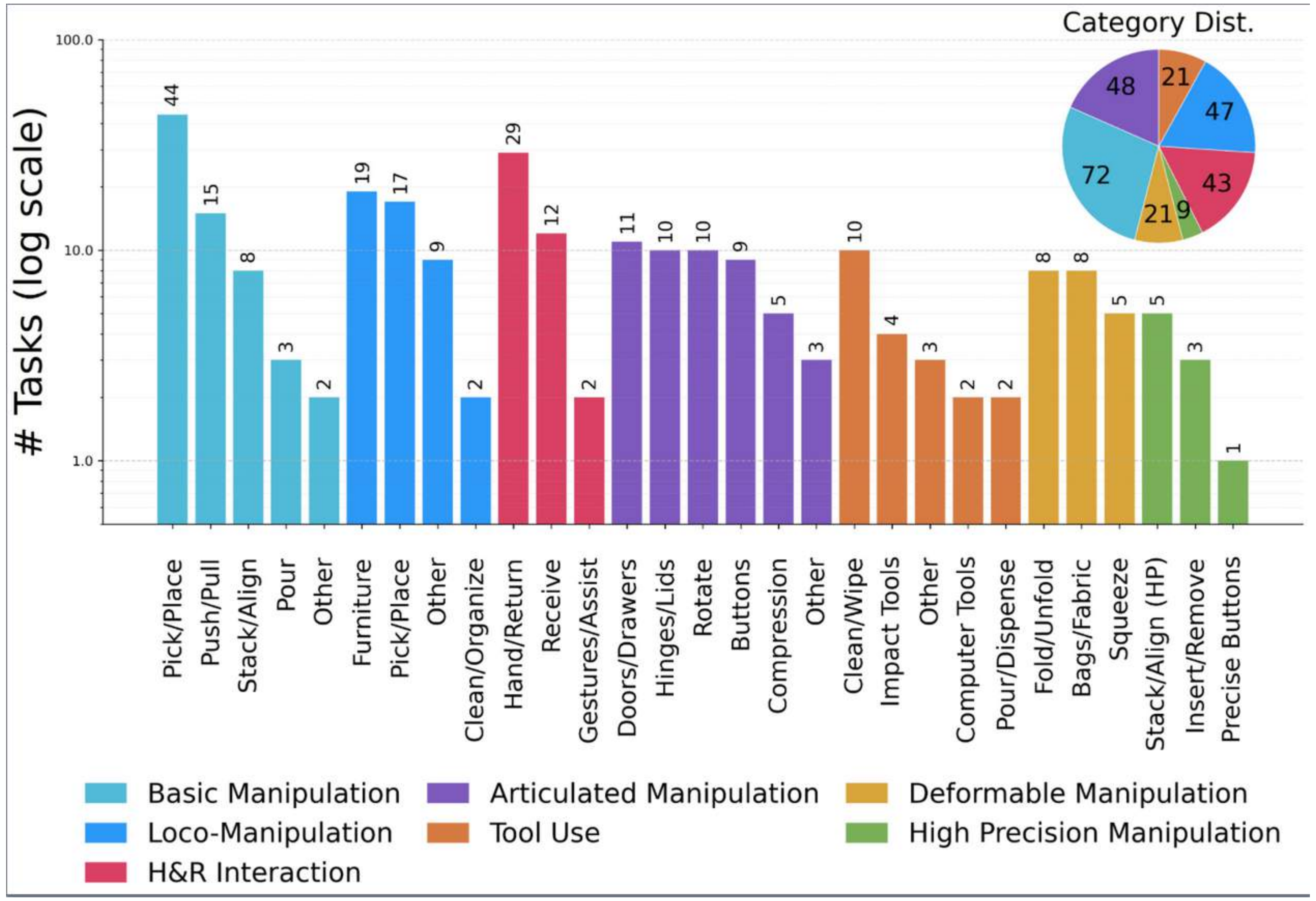
A high-frequency humanoid dataset spanning diverse everyday tasks.

260 Tasks · 7 categories · 30 Hz · 9 Modalities · Open Source

Dataset: Diverse collection of humanoid tasks

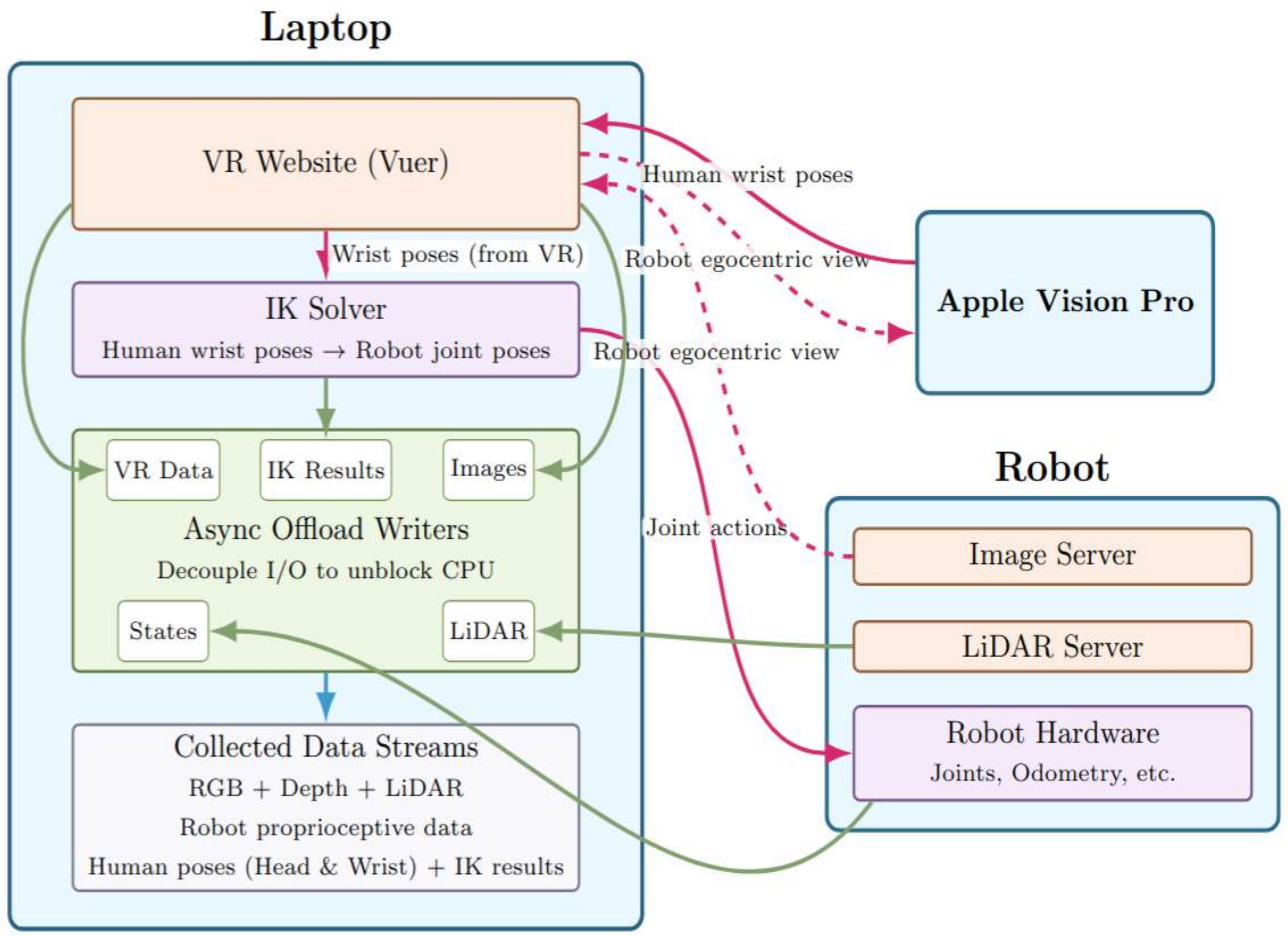
- Covers 10.3K trajectories, 3M+ frames, and 260 tasks using Unitree G1 and H1

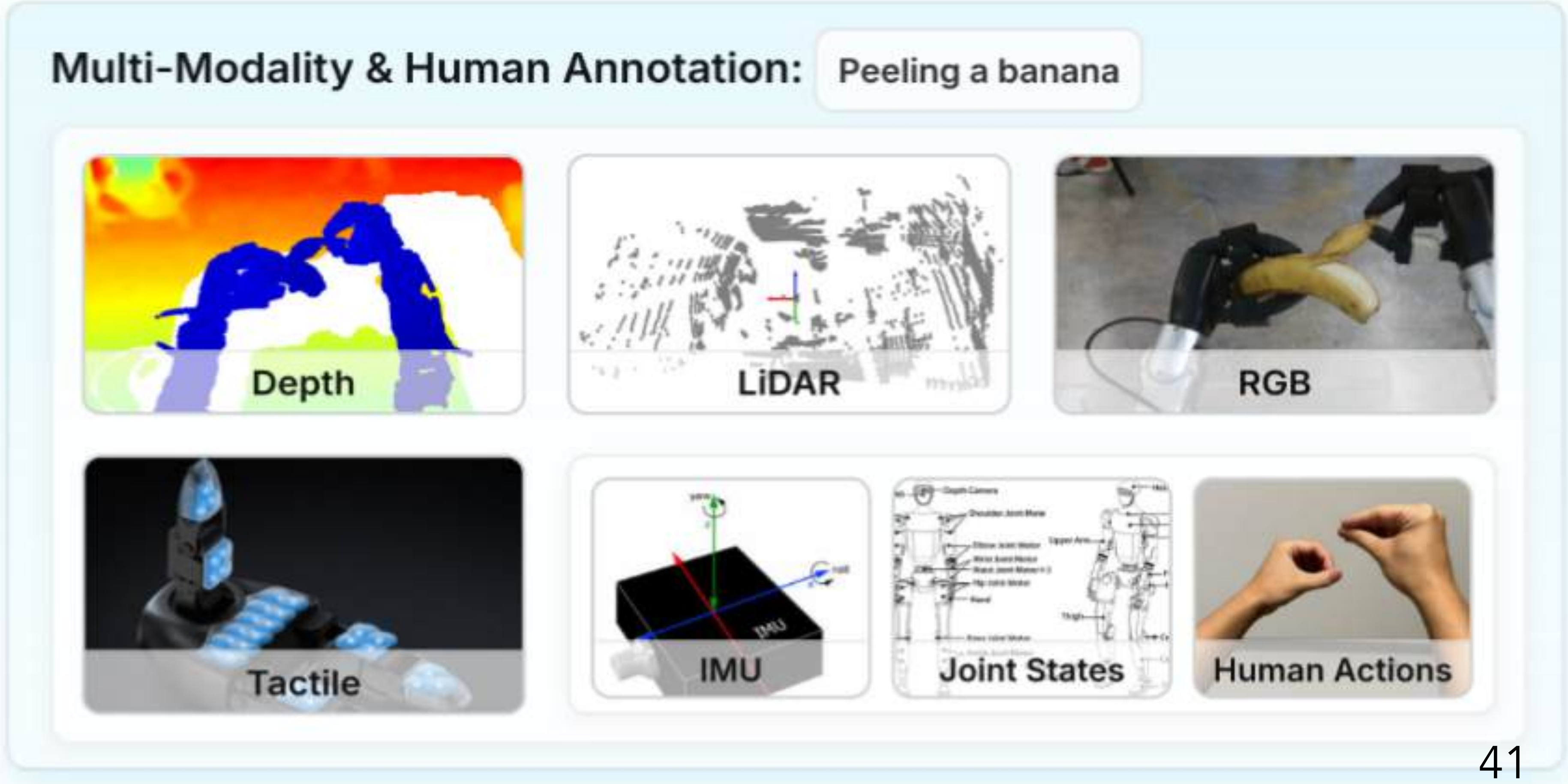
- Includes bipedal loco-manipulation and human-robot Interaction that are rare in other datasets



Dataset: efficient data collection pipeline

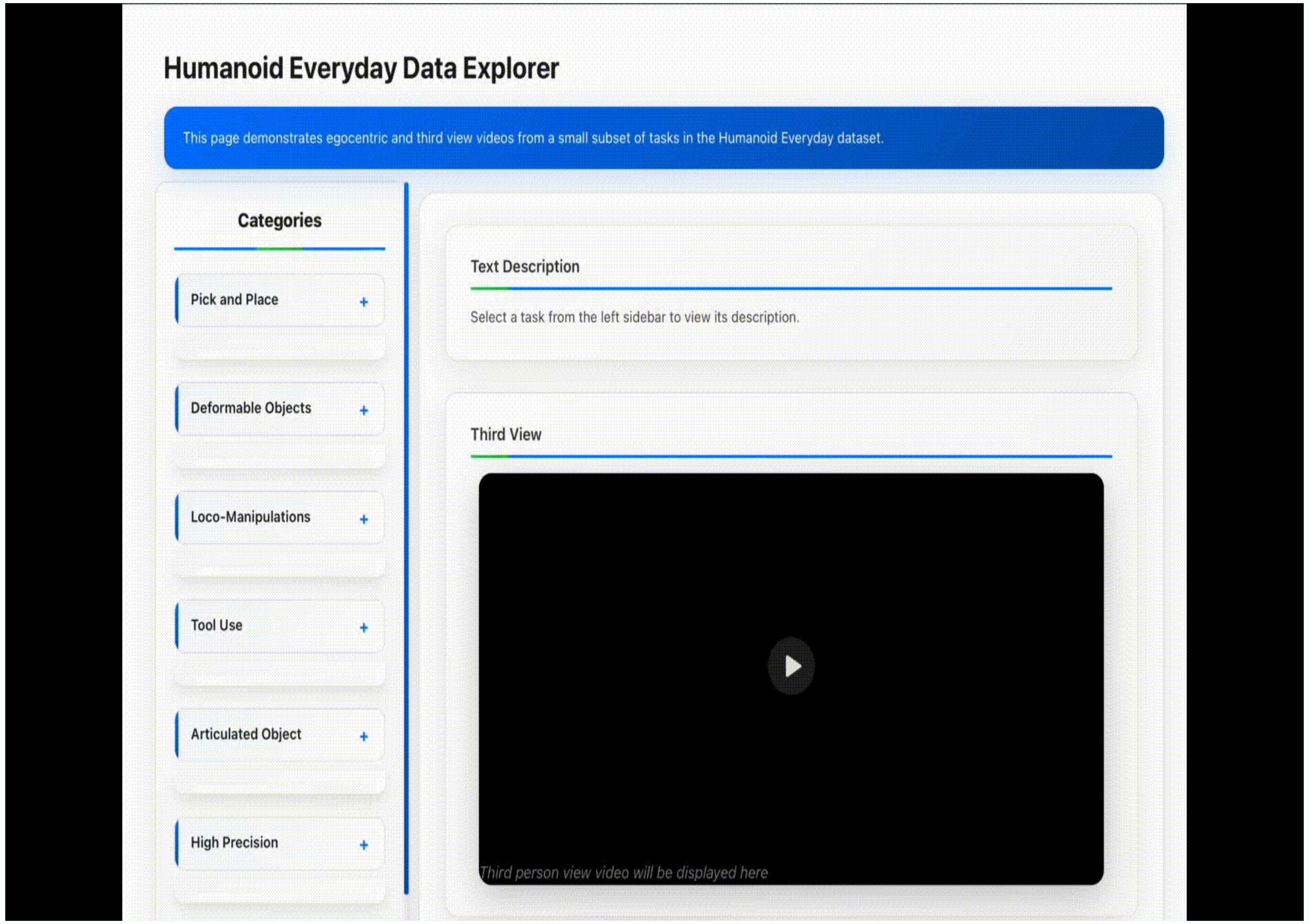
- Collection Pipeline:
  - Offloaded I/O keeps control loop fast and responsive.
- Improved Performance:
  - Reduced control delay from 500ms to 20ms
  - Halved data collection time
- 30hz multi-modality streams collected:
  - RGB+Depth+LiDAR
  - Proprioceptives: Joint States, Tactile, Odometry, IMU
  - Human Actions+Task Descriptions





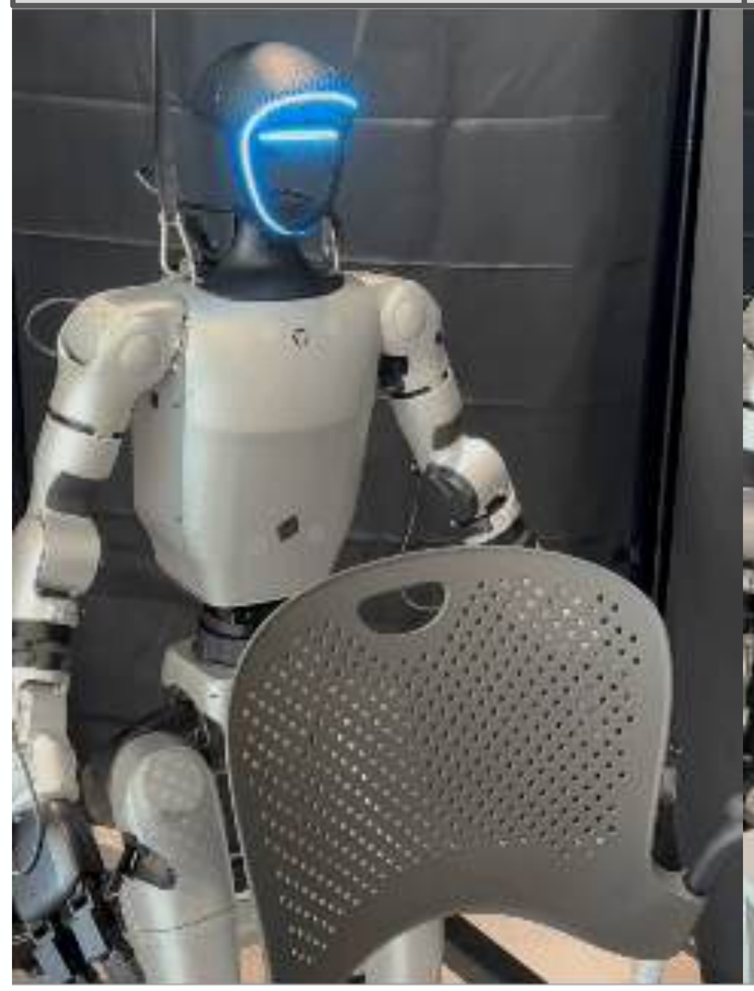
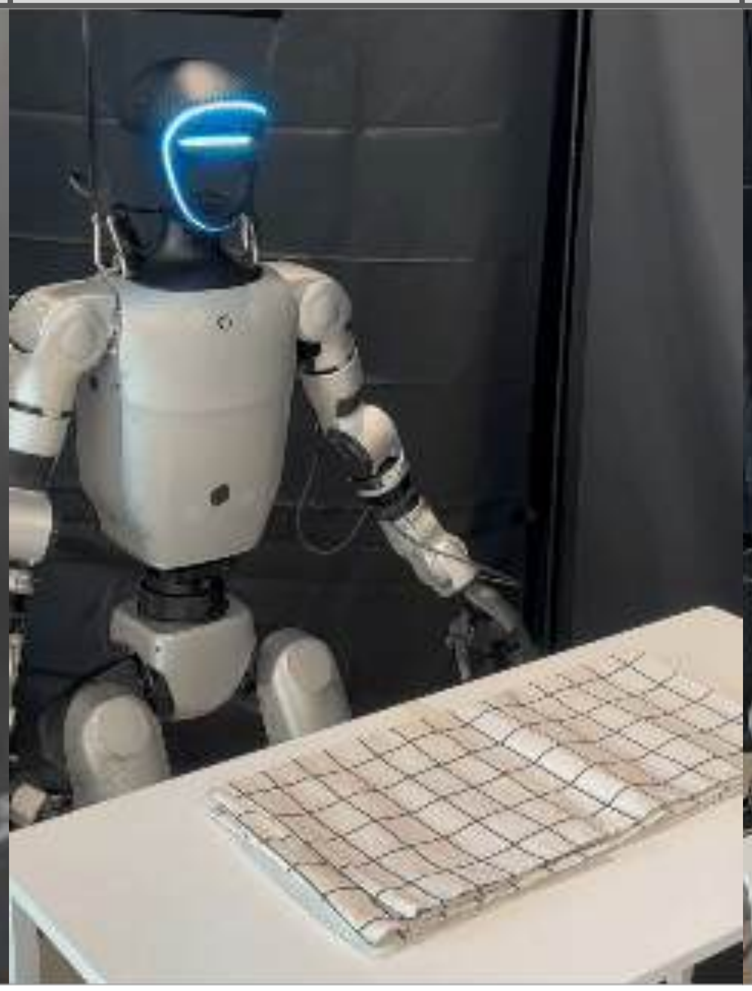
Dataset: Data viewer

- Data viewer contains 50 sample tasks from all of our categories
- Structure
    - Text Description
    - Third View Video
    - Egocentric Video
    - Point Cloud/Depth Visualization



Policy Inference: Imitation Learning + VLA

- We run inference using different imitation learning policies and VLA models on different manipulation tasks.

ACT	DP3	OpenVLA	OpenVLA	$\pi_{0.5}$	GR00T
					

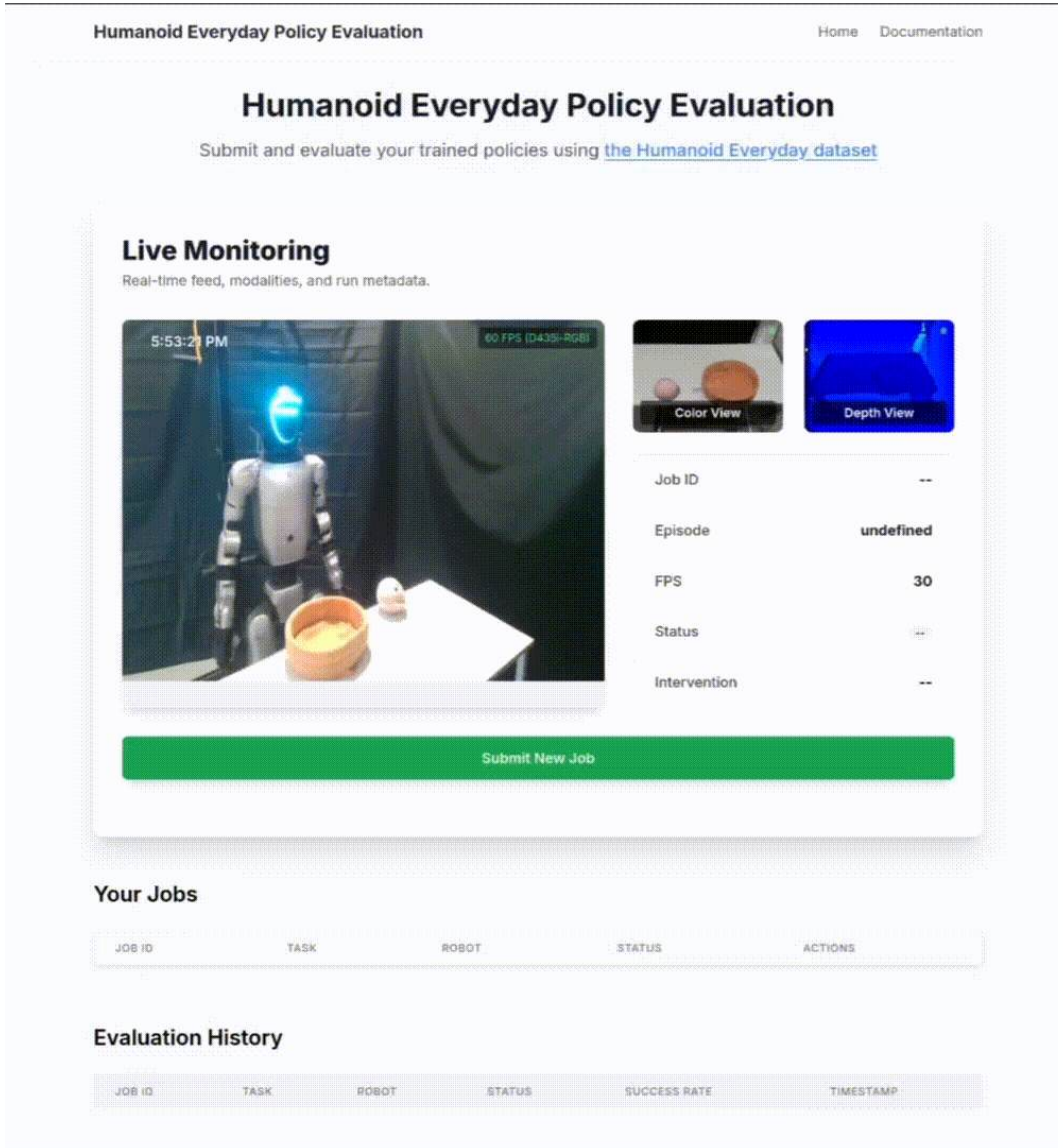
Results

Task Category	Task	DP	DP3	ACT	OpenVLA	π_0-FAST	$\pi_{0.5}$	GR00T N1.5
Articulate	Rotate chair	100%	90%	100%	70%	100%	100%	100%
Tool Use	Use eraser to wipe the desk	0%	70%	0%	30%	40%	40%	0%
Basic	Put dumpling toy into plate	30%	20%	70%	30%	60%	30%	80%
Deformable	Fold towel on the desk	0%	20%	0%	40%	20%	40%	50%
HRI	Hand over dumpling toy	40%	40%	70%	60%	30%	40%	100%
Loco-Manip.	Walk to grab door handle	30%	0%	0%	30%	10%	0%	30%
High Precision	Insert rose into vase	0%	0%	0%	10%	0%	0%	0%
Average		29%	34%	34%	39%	37%	36%	51%

- VLA models with pretrained priors outperform imitation learning policies.
- GR00T N1.5 achieves the best overall performance.
- All policies perform poorly on high-difficulty manipulation tasks.

Evaluation: Cloud-based Evaluation Platform

- Website for evaluating policies trained on the *Humanoid Everyday* dataset
- Streams real robot data and records success rates
- Supports remote inference (user policy server)



Humanoid Everyday

A high-frequency humanoid dataset spanning diverse everyday tasks.

- 260 Tasks
- 7 categories
- 30 Hz
- 9 Modalities
- Open Source

Acknowledgement

Robot Learning from Any Images: Siheng Zhao, Jiageng Mao

Universal Humanoid (UH1): Jiageng Mao, Siheng Zhao

Humanoid Everyday: Hongyi Jing, Zhenyu Zhao, William Liu

NSF TOYOTA RESEARCH INSTITUTE NVIDIA Google DOLBY CapitalOne Qualcomm