NVIDIA

CORL 2026

GRAIL

Generating Humanoid Loco-Manipulation from World Models

Tianyi Xie[1,2], Haotian Zhang[1], Jinhyung Park[1], Zi Wang[1], Bowen Wen[1], Jiefeng Li[1], Xueting Li[1], Qingwei Ben[1], Haoyang Weng[1], Yufei Ye[1], David Minor[1], Tingwu Wang[1], Chenfanfu Jiang[2], Sanja Fidler[1], Jan Kautz[1], Linxi Fan[1], Yuke Zhu[1], Zhengyi Luo[1], Umar Iqbal[1], Ye Yuan[1]

[1]NVIDIA [2]UCLA

MOTIVATION

Robotics Has a Data Problem

Teleoperation



Motion Capture



Both need **people to record every demonstration**, so data grows only with the number of people recording it.

Teleoperation photo: The Standard (Hong Kong) · motion capture: Bones Studio SEED dataset

What if these videos could become **humanoid robot data**?

A NEW PARADIGM

Generate robot data instead of collecting it

KEY IDEA

Video World Models as Controllable Data Sources

✓ **Video world model: ask for what we need**

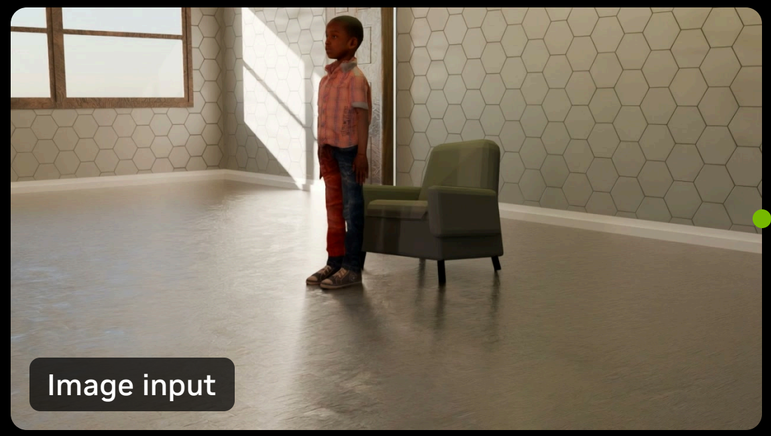
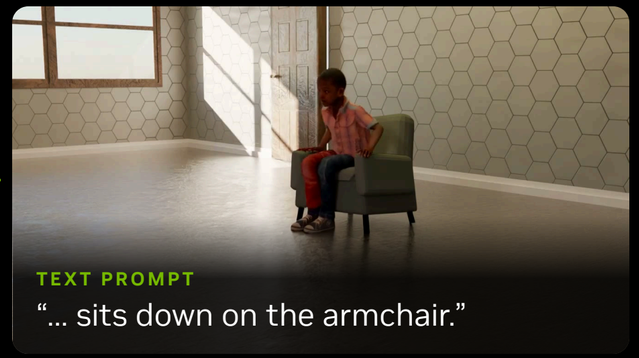


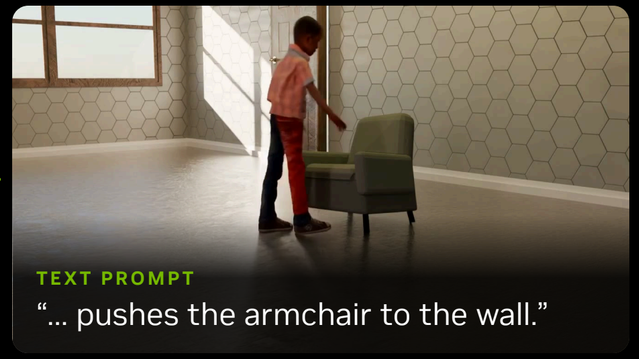
Image-to-video with MiniMax-H3

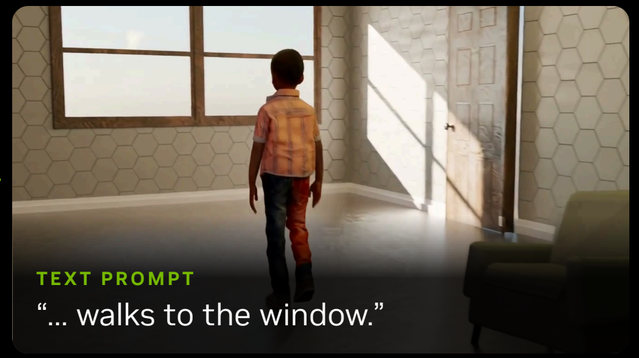
✕ **Internet videos: no control over what we get**



- **Camera cuts**
- **Unknown, moving camera**
- **Body out of frame**
- **Occlusion**
- **Meaningless motion**
- **No interaction**







KEY IDEA

Video World Models as Controllable Data Sources

Easy to edit
Assets · character · scene · camera

Known 3D information
Camera · metric scale · object geometry · body shape

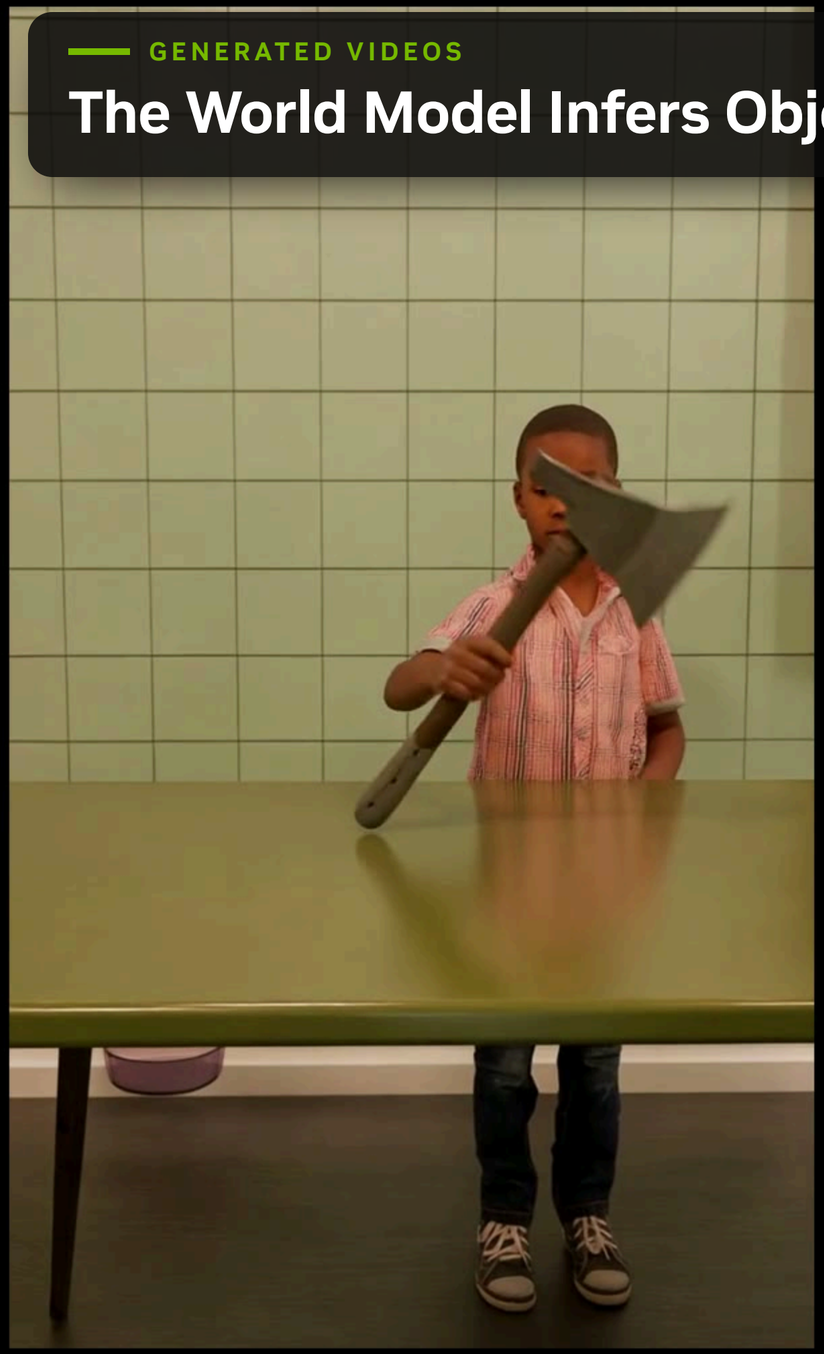
GENERATED VIDEOS

Pick-Up with Data Augmentation



GENERATED VIDEOS

The World Model Infers Object Affordance





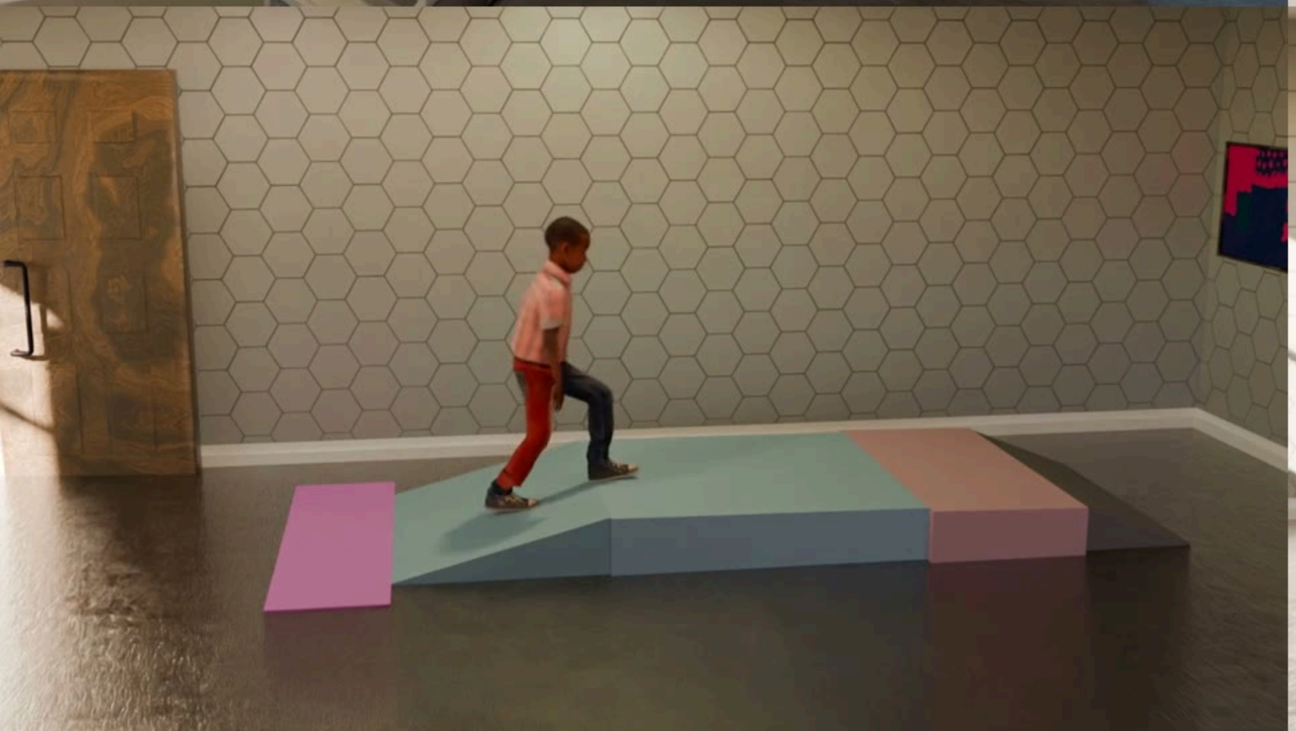


GENERATED VIDEOS

Scene-Aware Interactions









GENERATED VIDEOS

Interacting with People

Toward human–robot interaction data



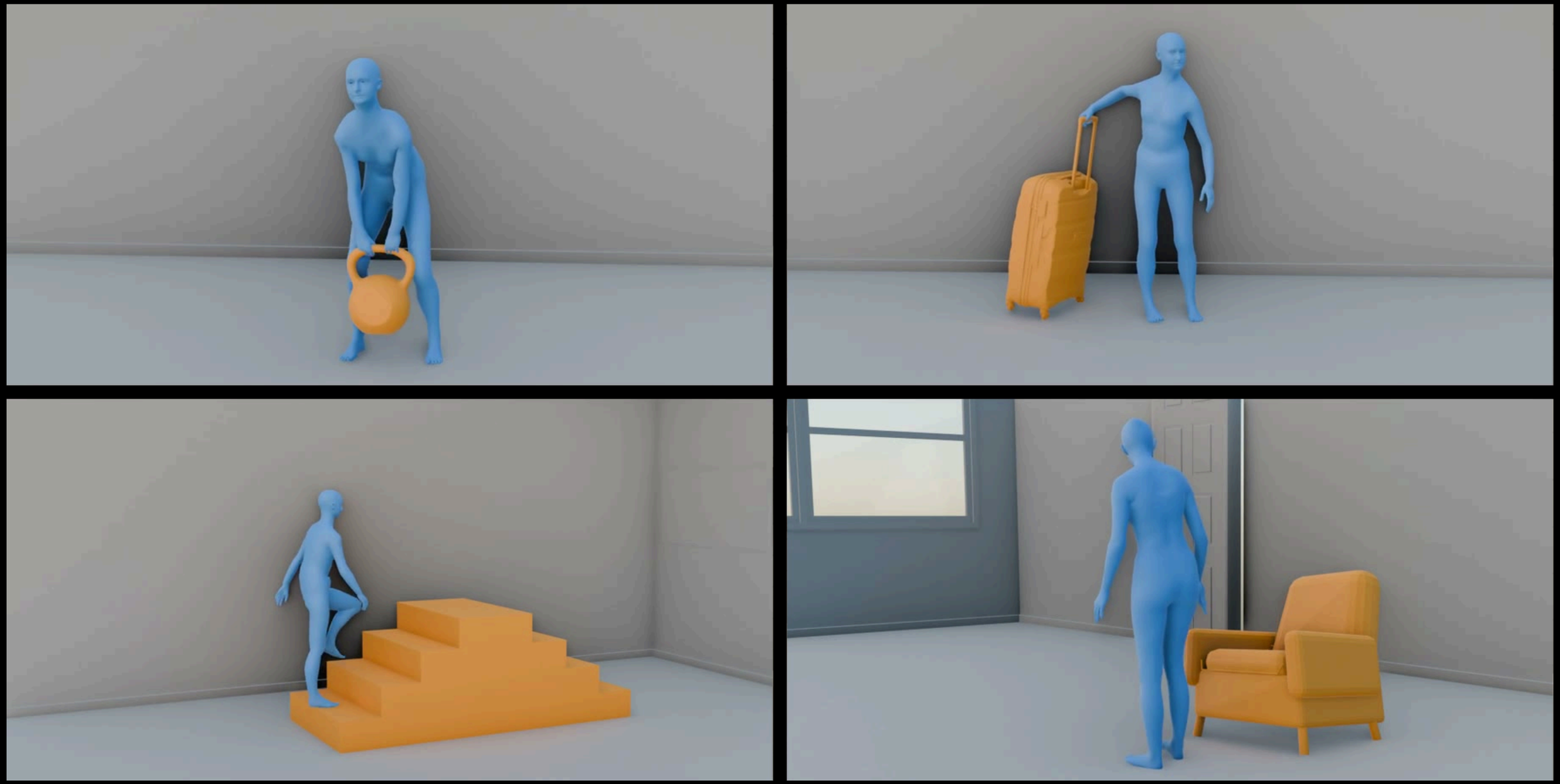


RECONSTRUCT

Metric 4D Reconstruction

Human GEM-X **Object** FoundationPose **Known scene**

+ joint optimization of human, object and scene

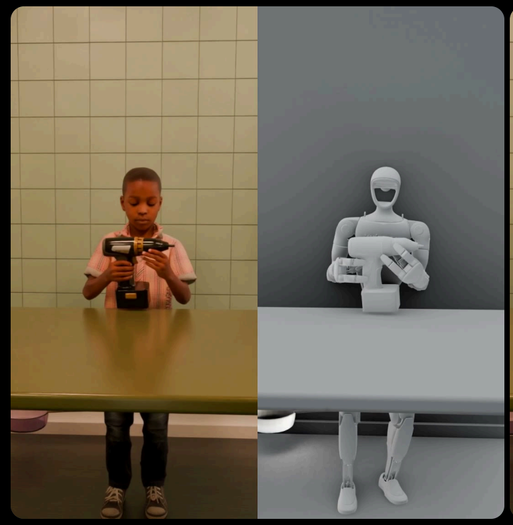



RETARGET
Reconstruction → Unitree G1
Jason · 4D reconstruction
Retargeted Unitree G1

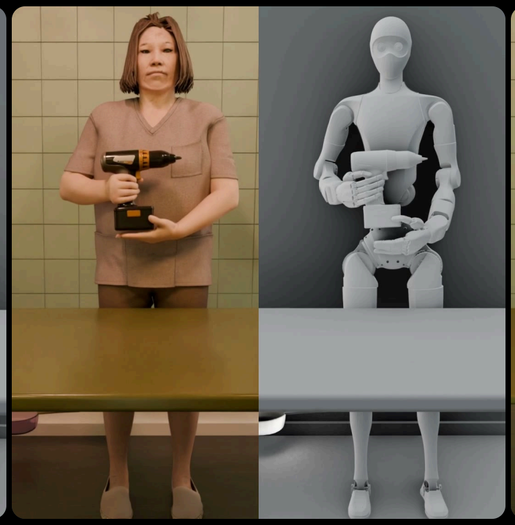
RETARGET

Multiple Embodiments

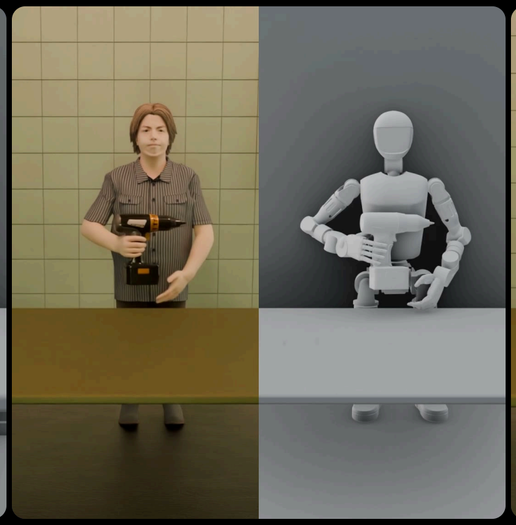
No spatial retargeting required



Unitree G1 · 1.32 m



Unitree H2 · 1.83 m



Booster T1 · 1.18 m



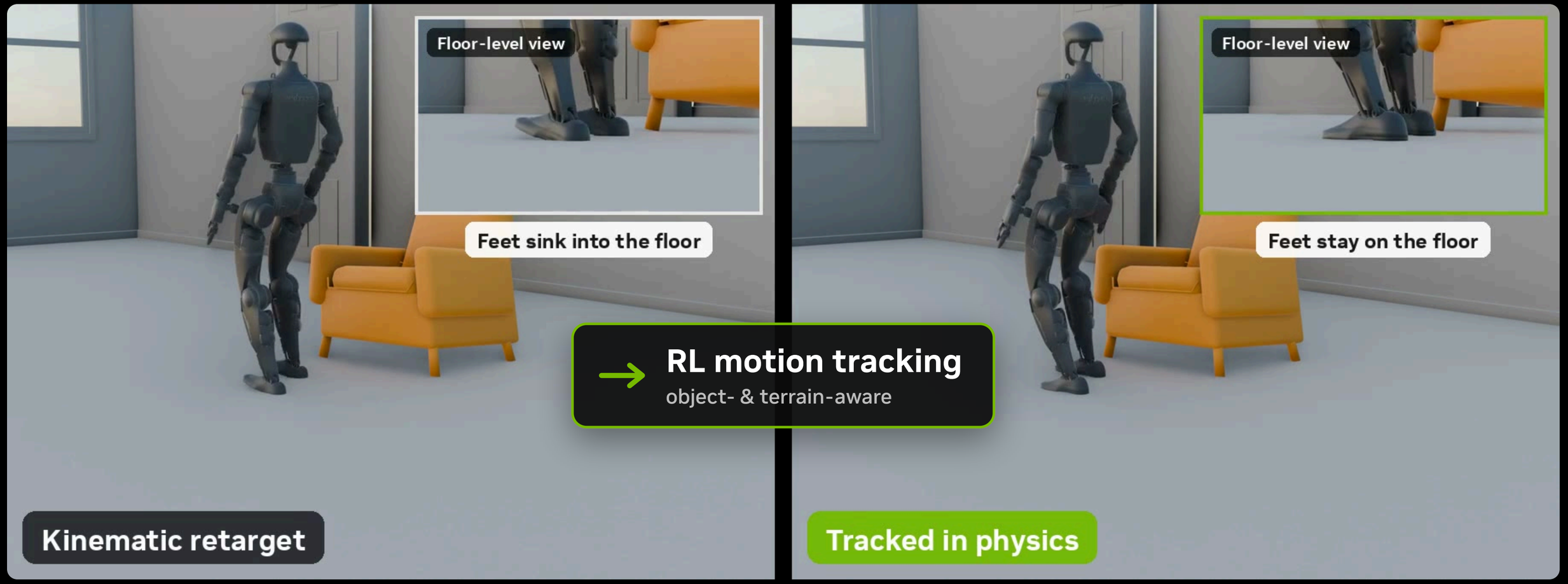
AgiBot X2 Ultra · 1.32 m



AgiBot A3 · 1.72 m

Each robot picks up the drill differently: we generate videos with the **target embodiment in mind** from the start.

Grounding Kinematic Motions in Physics



Object- and terrain-aware RL motion tracking grounds every generated motion in physics.

DATASET

A Dataset Generated by GRAIL

~250K downloads on Hugging Face

huggingface.co/datasets/nvidia/PhysicalAI-Robotics-Locomanipulation-GRAIL



35,000+

physically validated loco-manipulation motions,
and generating more

DEPLOY

Training Visual Policies in Simulation

Domain randomization · distilled from the trackers

Rendered egocentric RGB · head camera

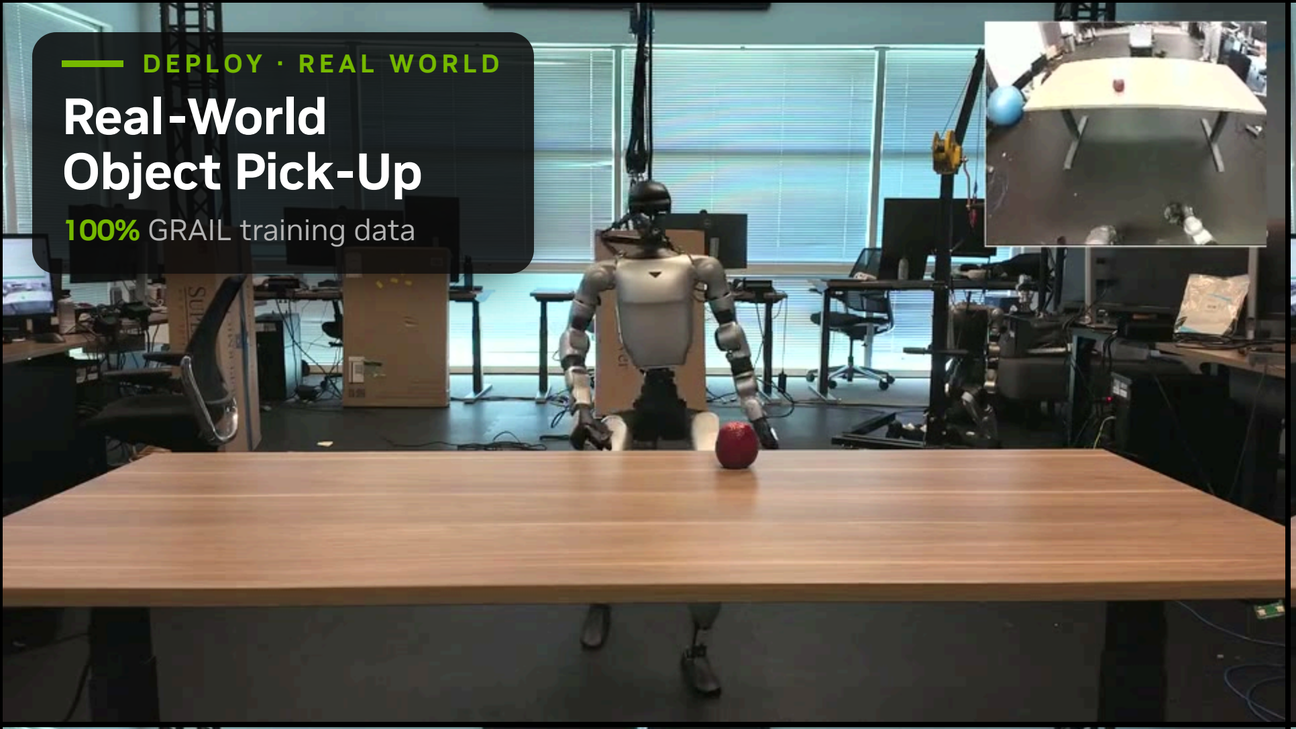


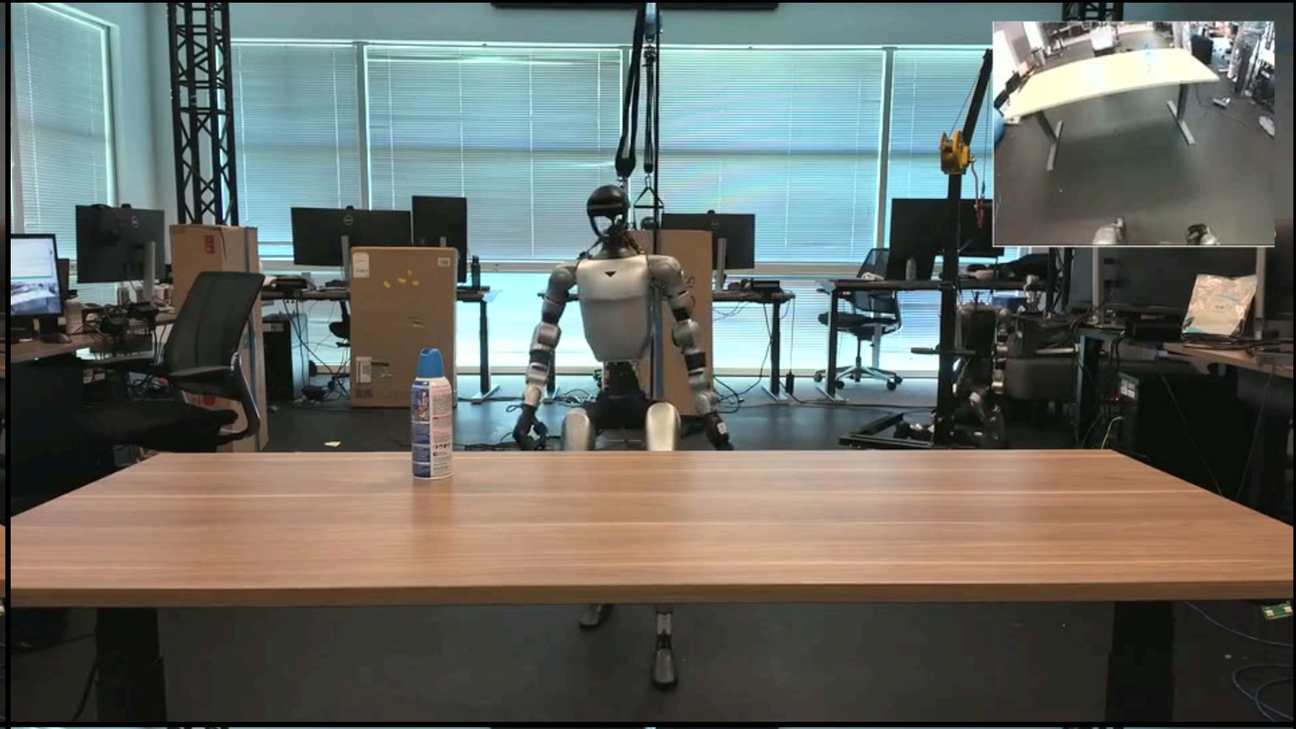
Parallel simulated environments

DEPLOY · REAL WORLD

Real-World Object Pick-Up

100% GRAIL training data



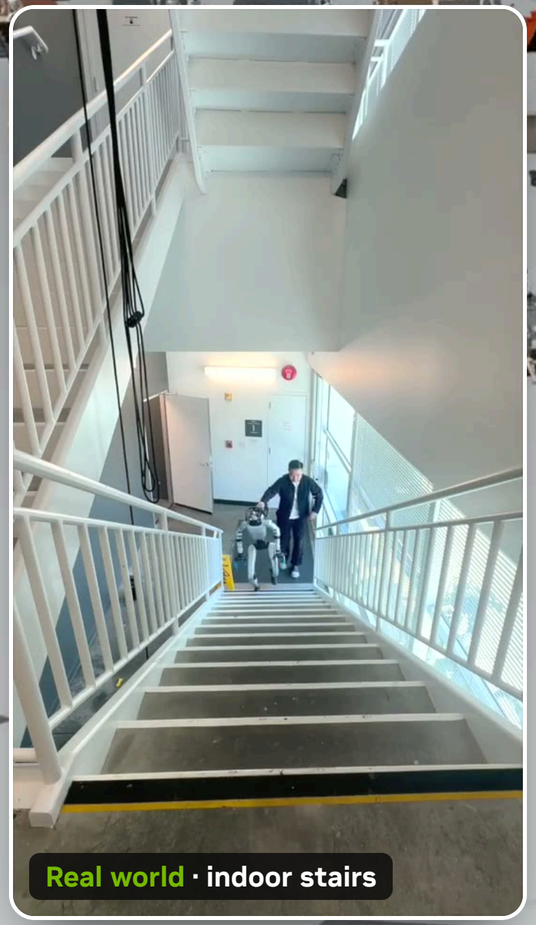






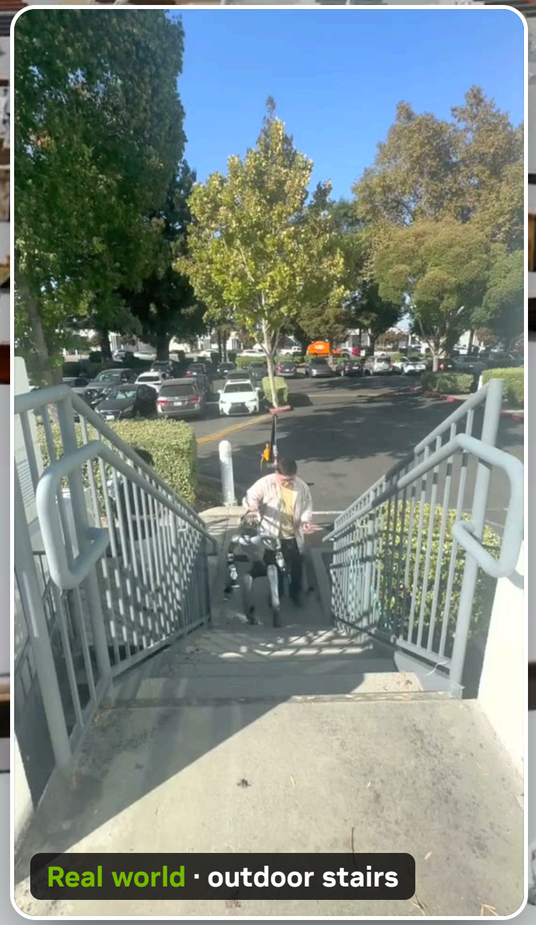
80% success on unseen objects

ZERO-SHOT SIM-TO-REAL



Real world · indoor stairs

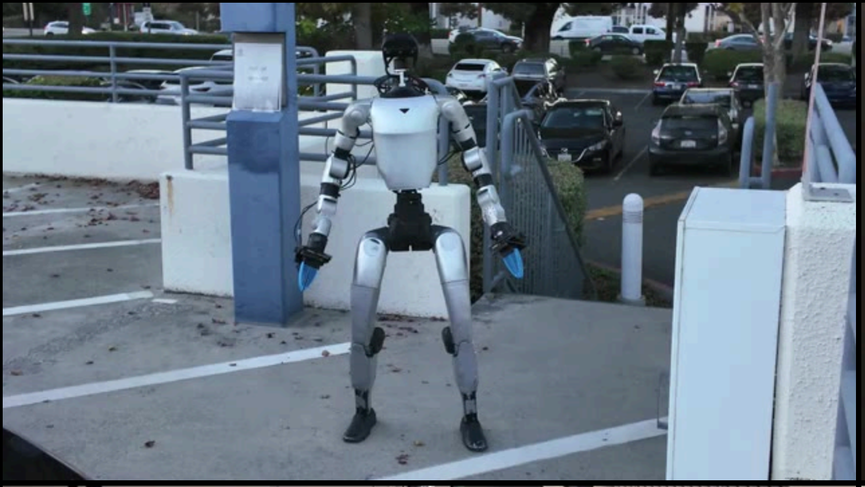
Simulation · GRAIL stair motions

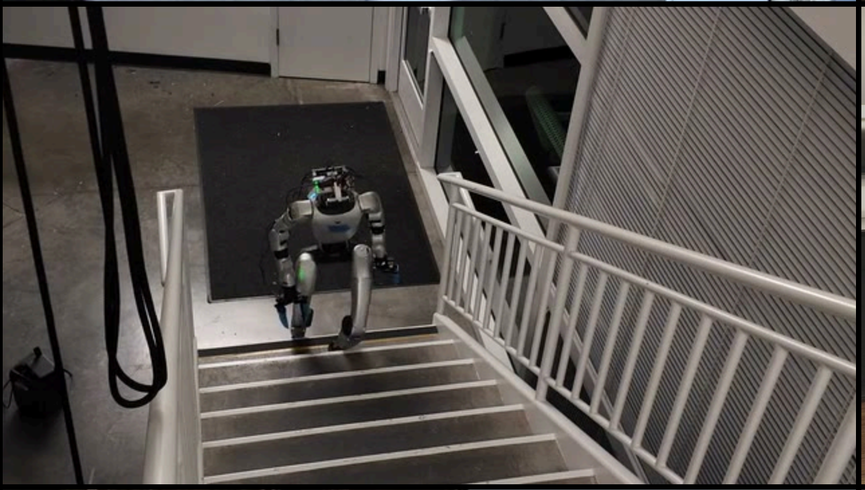


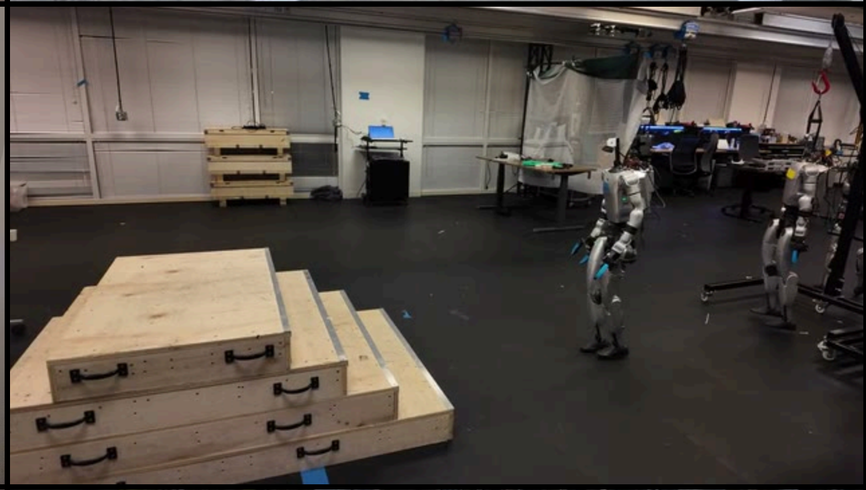
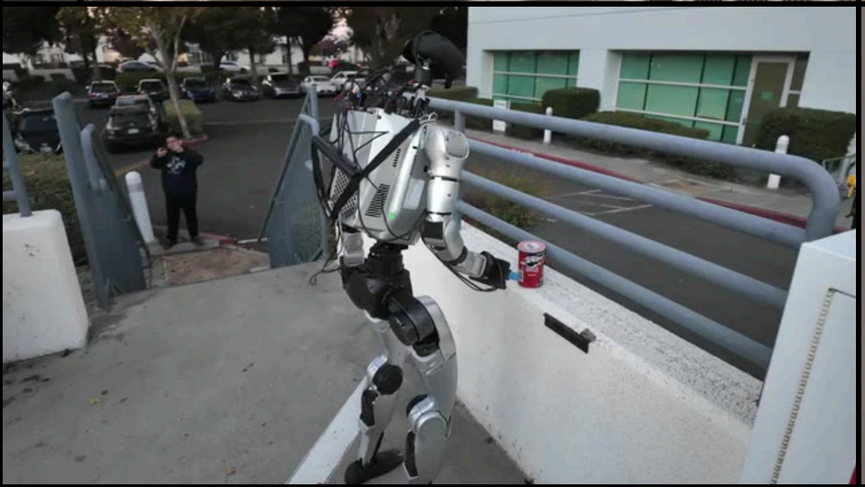
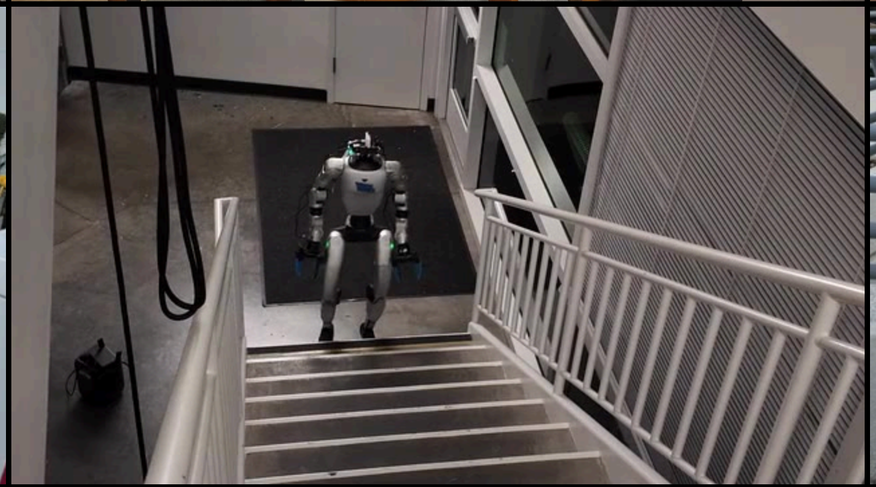
Real world · outdoor stairs

DEPLOY

Sim-to-Real Stair Climbing




SUMMARY

GRAIL: a new paradigm for scaling robot learning

We envision a future where **data is no longer the bottleneck**.



Thank you.

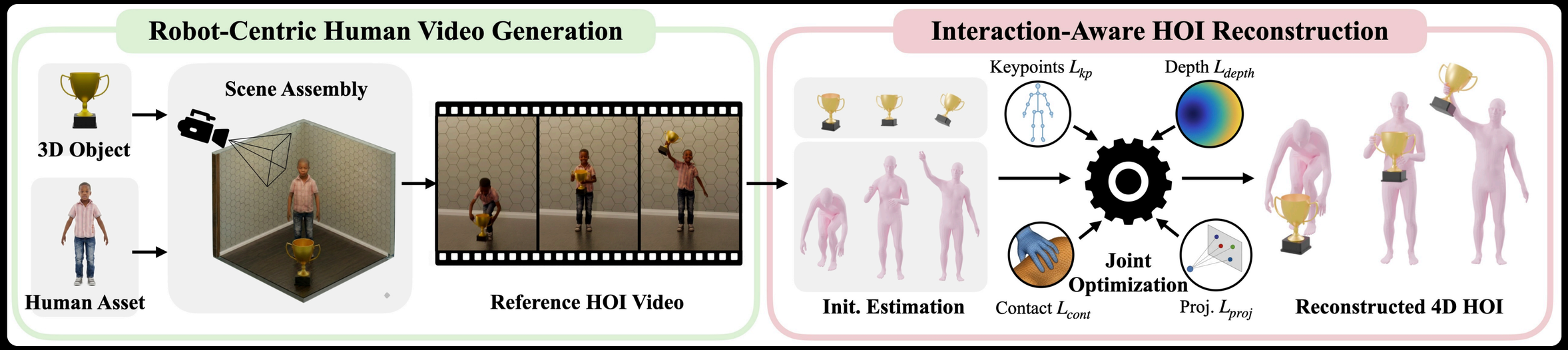
Our code and dataset are released. Feel free to check them out!

Code

Dataset

BACKUP

Generation and 4D HOI Reconstruction



Robot-centric human video generation

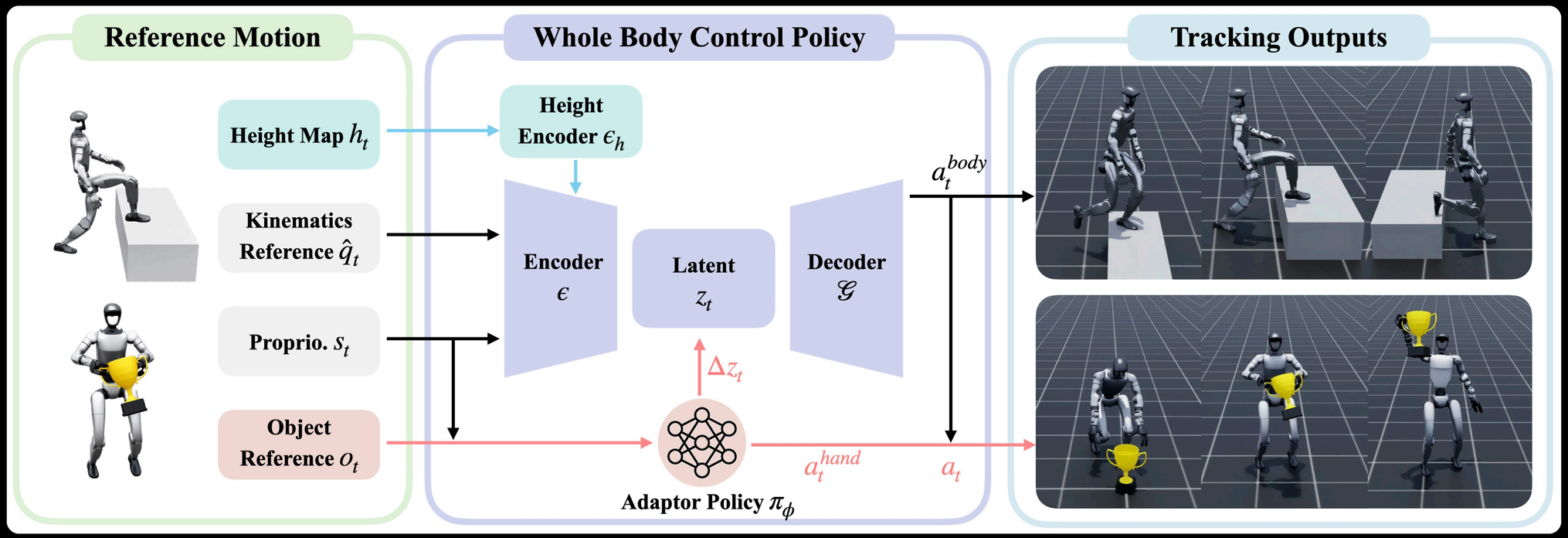
Render frame 0 of our scene; the video world model animates it.

Interaction-aware 4D HOI reconstruction

Refine keypoints, depth, and contact against the known scene: metric 4D HOI.

BACKUP

One Task-General Whole-Body Tracker



Whole-body controller

Pretrained encoder → latent → decoder.

Object-aware adaptor

Controller frozen; latent residual + hand open/close.

Scene-aware tracker

Height map input; fine-tuned end to end.