Foundation Models for Embodied Intelligence

Hang Zhao

Assistant Professor

Tsinghua University
Shanghai Qi Zhi Institute

Dec. 2025

MARS Lab

MARS Lab Overview

MARS Lab

Hang Zhao

Assistant Professor
Principal Investigator

We build robots at Galaxea

Multimedia Computing

We train AI models that understand and generate multi-modal data such as images, audios, videos and text.

Autonomous Driving

We develop the next-generation autonomous driving software stack: from visual scene understanding to neural planning.

Robotics

We make robots that learn from multiple sensory inputs to interact with the environment.

Industry Sponsors & Collaborators

Li Auto
GALAXEA
Mercedes-Benz
BOSCH

Foundation Models for Embodied Intelligence


Driving


Manipulation


Navigation


Locomotion

Vision-Language-Action Models?

Applying LLMs on robotics…

We still don't have native robotic foundation models

Foundation Models for Embodied Intelligence

Driving

Manipulation

Navigation

Locomotion

HRI
Task Planning

↑

Language Foundation Models
Reasoning Intelligence

Navigation
Motion Planning

↑

Vision Foundation Models
Spatial Intelligence

Locomotion
Whole Body Control

↑

Control Foundation Models
Athletic Intelligence

Foundation Models for Embodied Intelligence


Driving


Manipulation


Navigation


Locomotion

DriveVLM, CoRL 2024
DriveAgent-R1, arXiv 2025
Galaxea G0, arXiv 2025

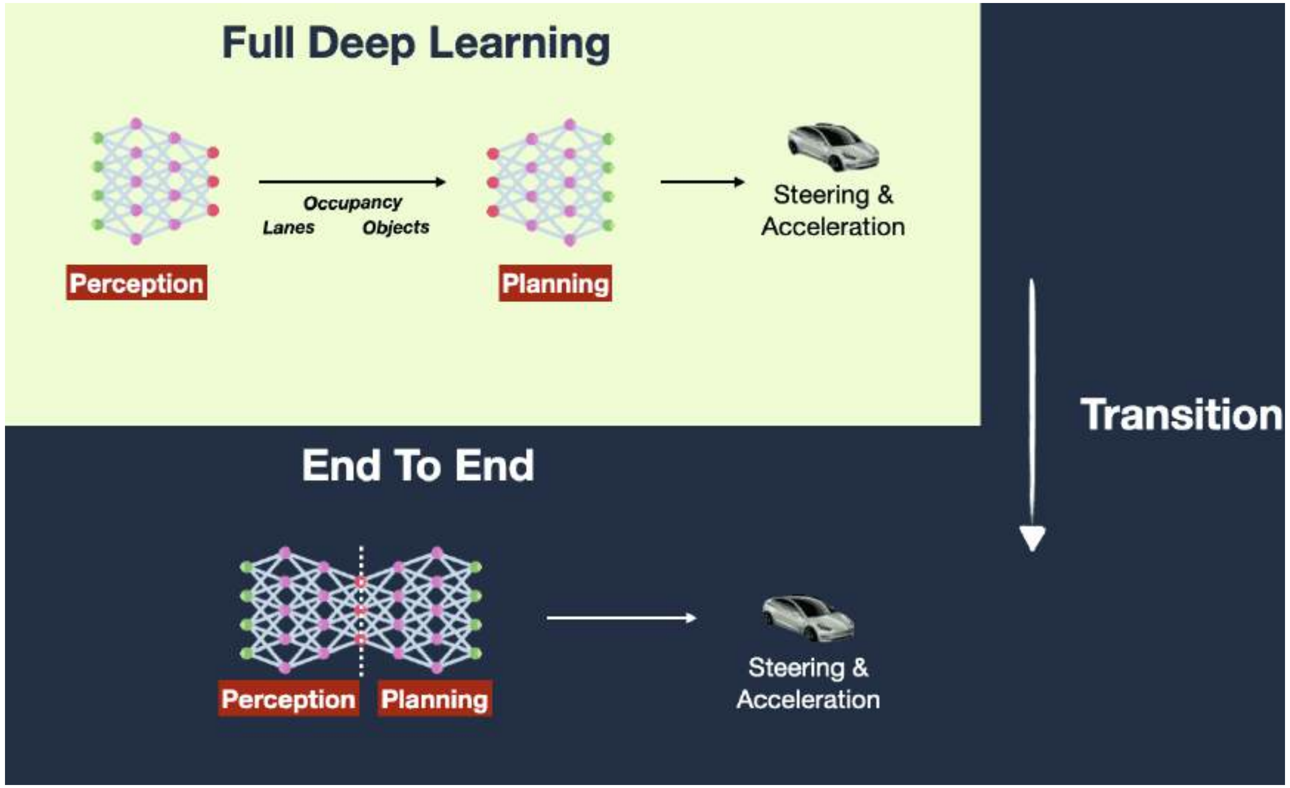
Language Foundation Models
Reasoning Intelligence

Vision Foundation Models
Spatial Intelligence

Control Foundation Models
Athletic Intelligence

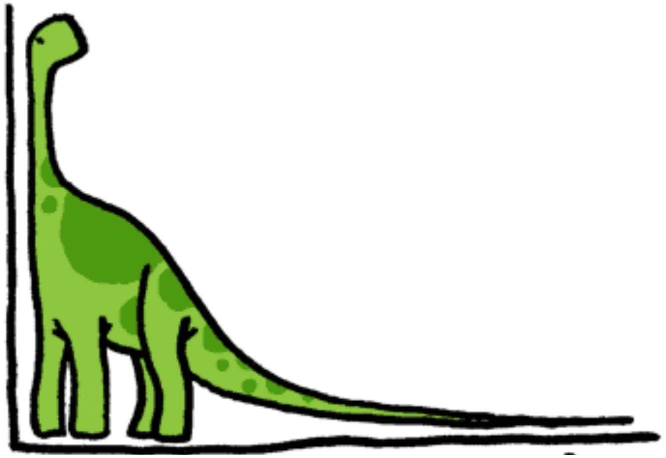
End2end models are dominating Self-driving

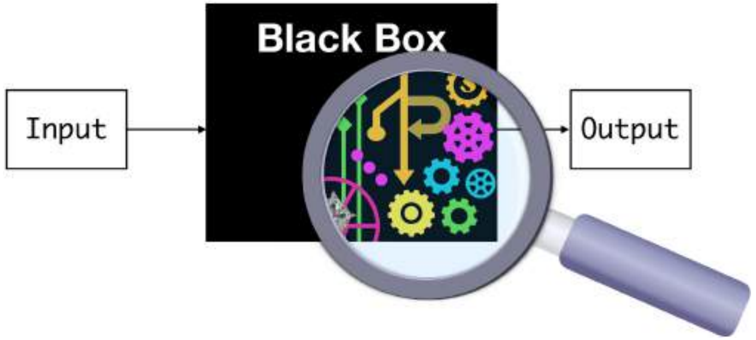
- Tesla has moved from modular deep learning models to end2end models since FSD V12



End2end models do not lead us to L4 Self-driving

- L4 means no drivers, AI need to deal with all the long-tail cases
- Long-tail dilemma: there is not enough long-tail data to train AI models
- Black-box models lack interpretability and controllability





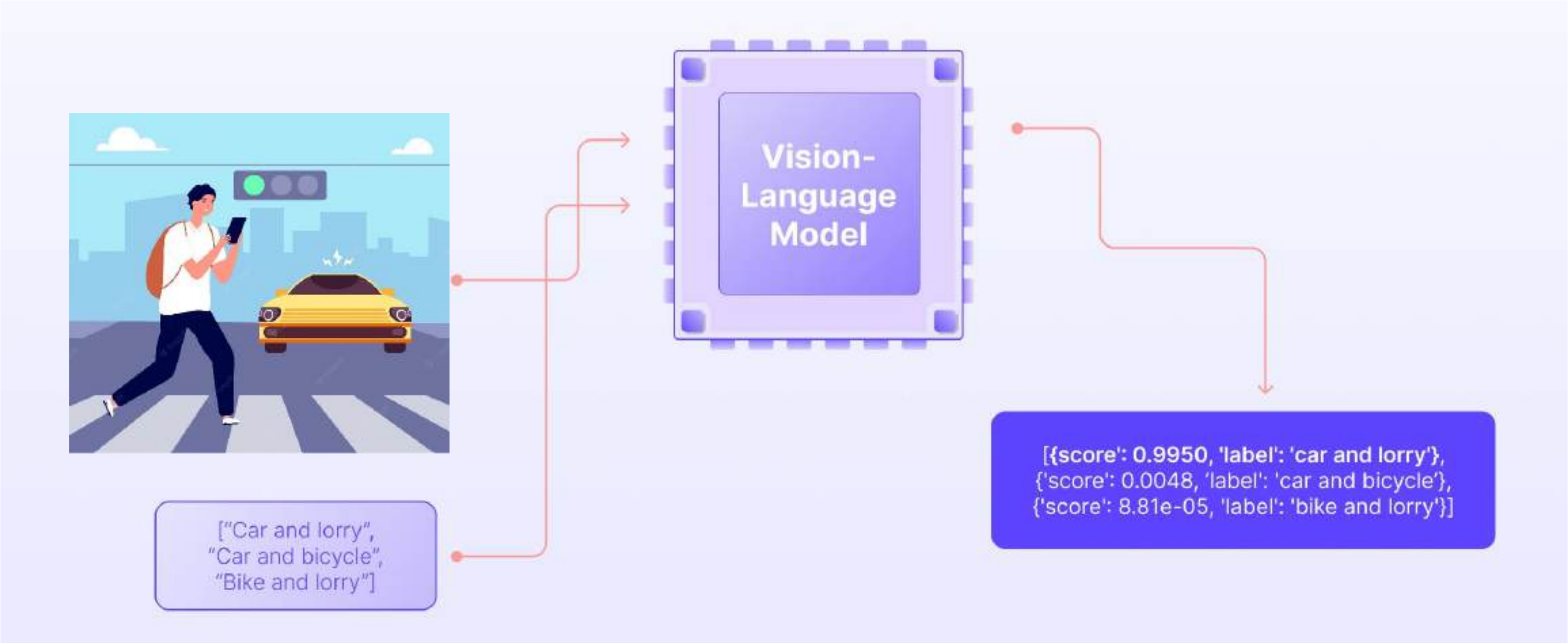
“Should AGI happen before L4 Self-driving?”

Scene Understanding has been bottlenecking self-driving

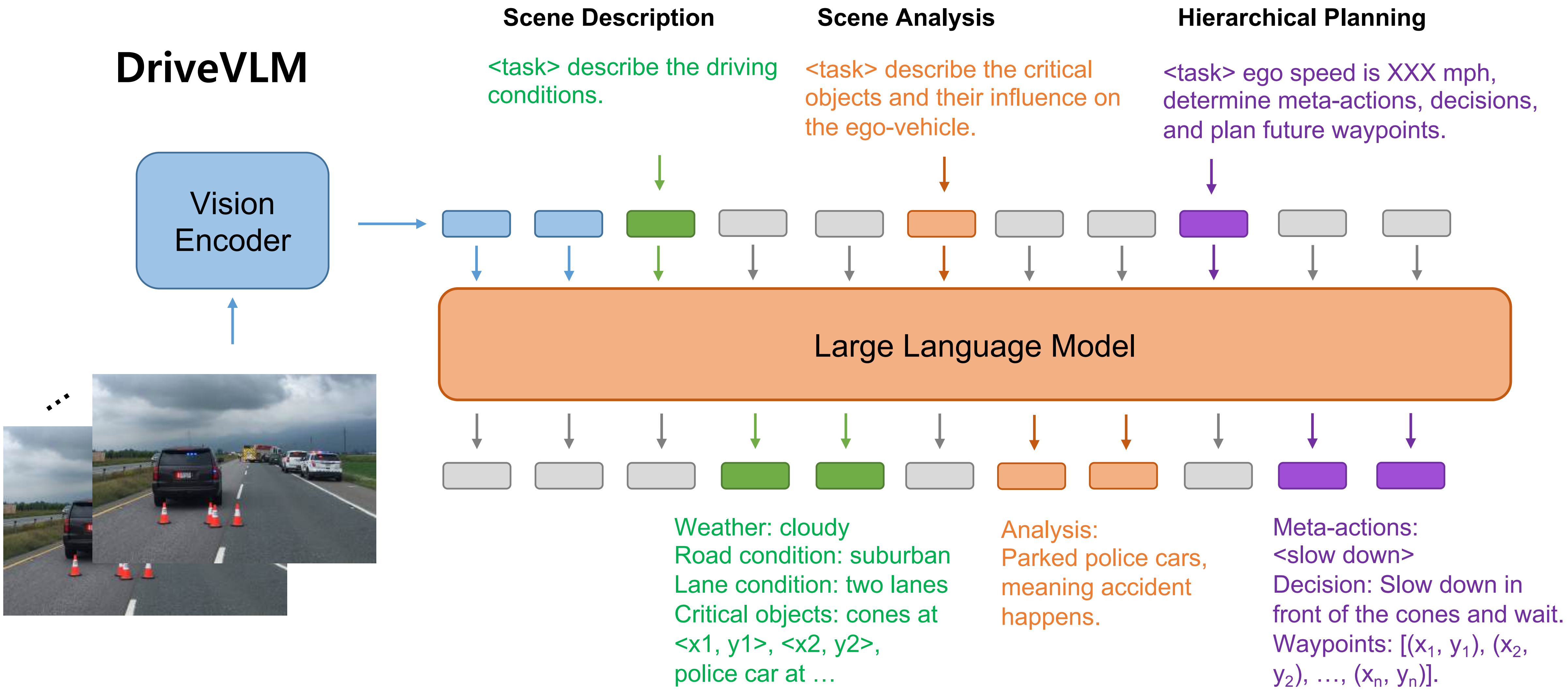
Credit to Waymo 2016

How do VLMs help to solve hard cases in AD?

- **The Vision Encoder** offers **dense observations**, with **rich semantic attributes.**
- **The LLM** connects the visual elements together, performs holistic **scene understanding, reasoning,** and **decision making.**

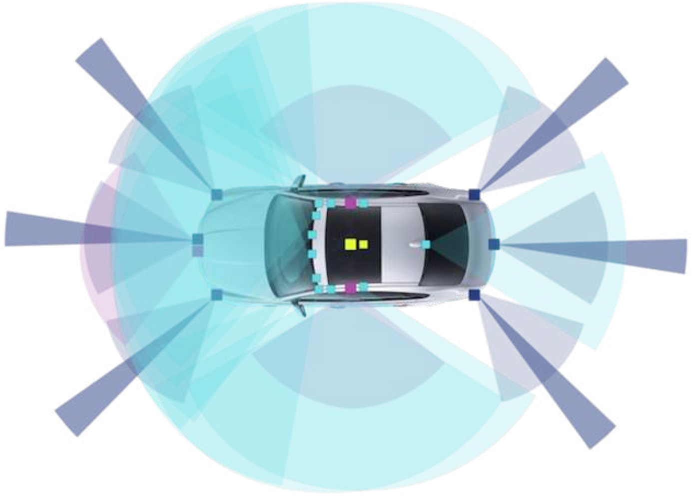


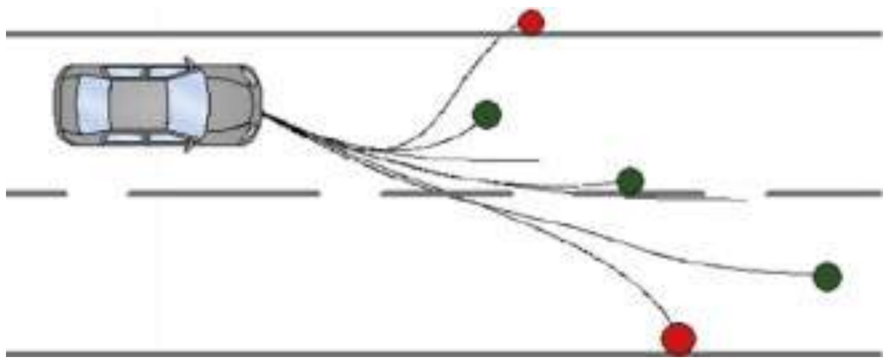
Our first attempt on end2end VLM in April 2023



DriveVLM - Limitations

- VLMs have poor 3D spatial grounding ability.
  - Perception: false positives/negatives, inaccurate distance
  - Planning: jagged trajectories, prone to collision
- Computing bottleneck: hard to deploy on the car for real-time inference.



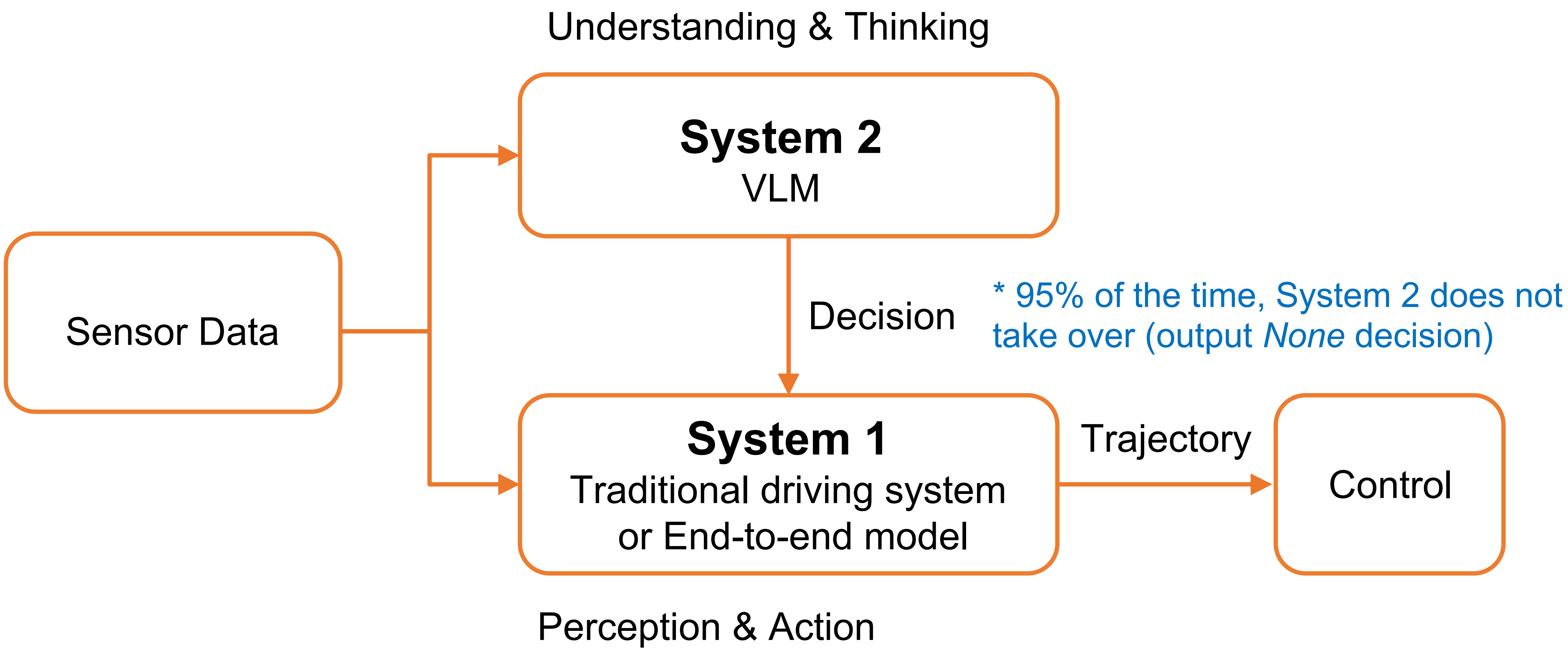


Dual System for Self-driving

System 1	System 2
Fast System	Slow System
Intuitive Maneuver	**Rational Decision**
95%	5%
• **In simple scenarios** • Based on memory and experience • Fast decision	• **In hard and long-tail scenarios** • Logical thinking • Slow and focused

DriveVLM, Xiaoyu Tian, Junru Gu, et al. CoRL 2024

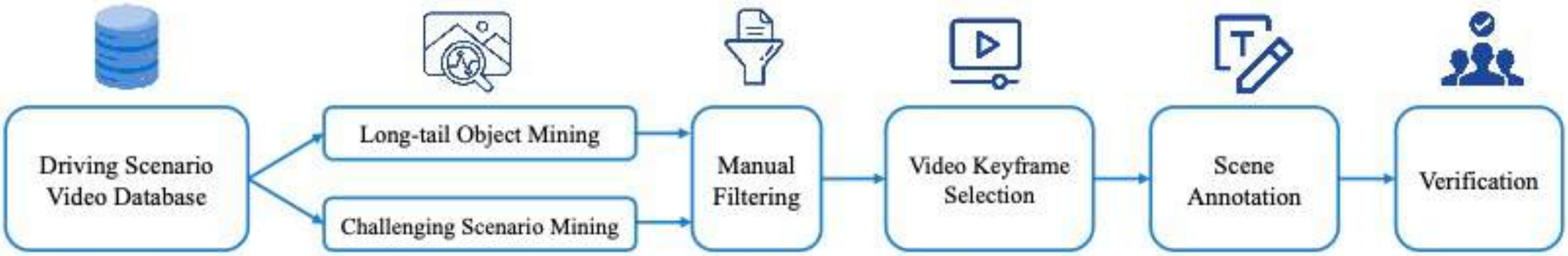
DriveVLM-Dual



DriveVLM, Xiaoyu Tian, Junru Gu, et al. CoRL 2024

Dataset Construction Pipeline

- Constructed a dataset with complex and long-tail scenarios by mining from ***~billions of miles*** of real-world driving data.



SUP-AD Dataset
Complex and Long-tail Driving Scenarios

SUP-AD Dataset Annotations

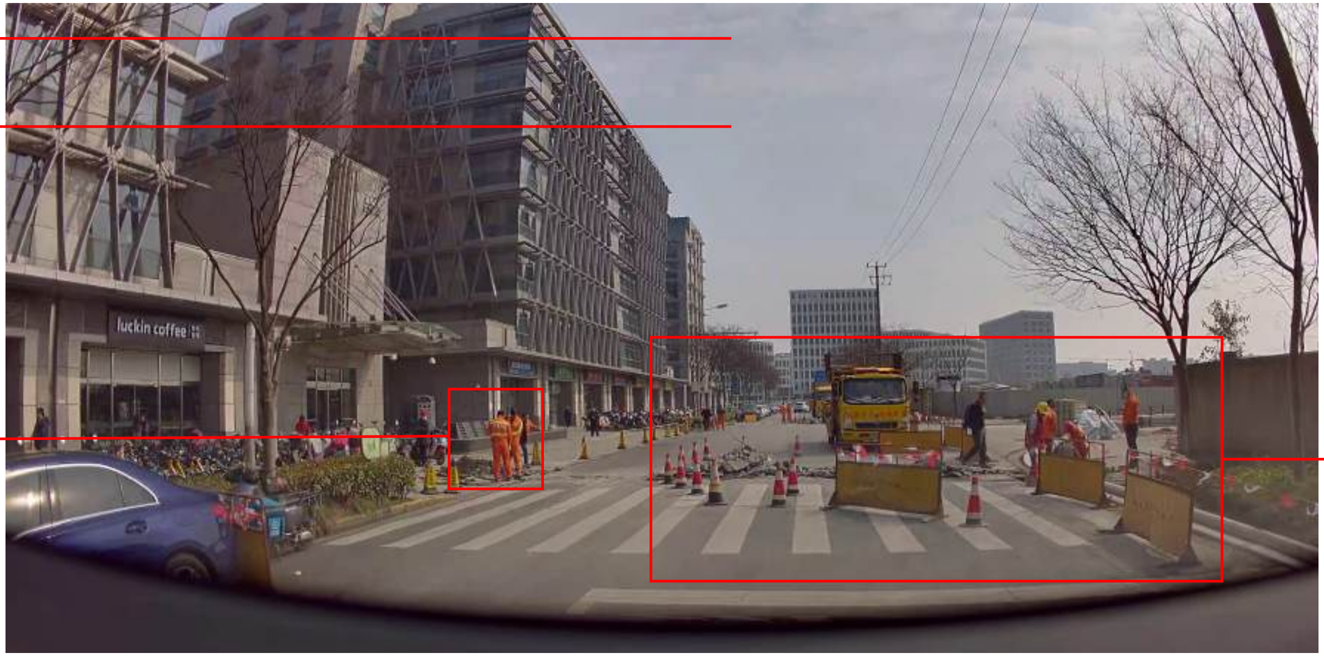
Scene Summary: The ego vehicle is moving at a constant speed along the current lane, with ongoing road construction work ahead; there are three construction workers working on the left side of the lane at the roadside.

Weather: Sunny

Time: Daytime

Critical Object:

Class:	Three Construction Workers
Characteristics:	Construction work on the side of the lane to the left of the host vehicle
Influence:	Affects the normal speed of the host vehicle



Road Environment: Construction

Critical Object:

Class:	Construction Zone
Characteristics:	Road repair in front of the host vehicle lane
Influence:	Affects the host vehicle to drive straight normally

Meta Action: ["Slow down”, "Change lane to the left”, "Go straight slowly"]

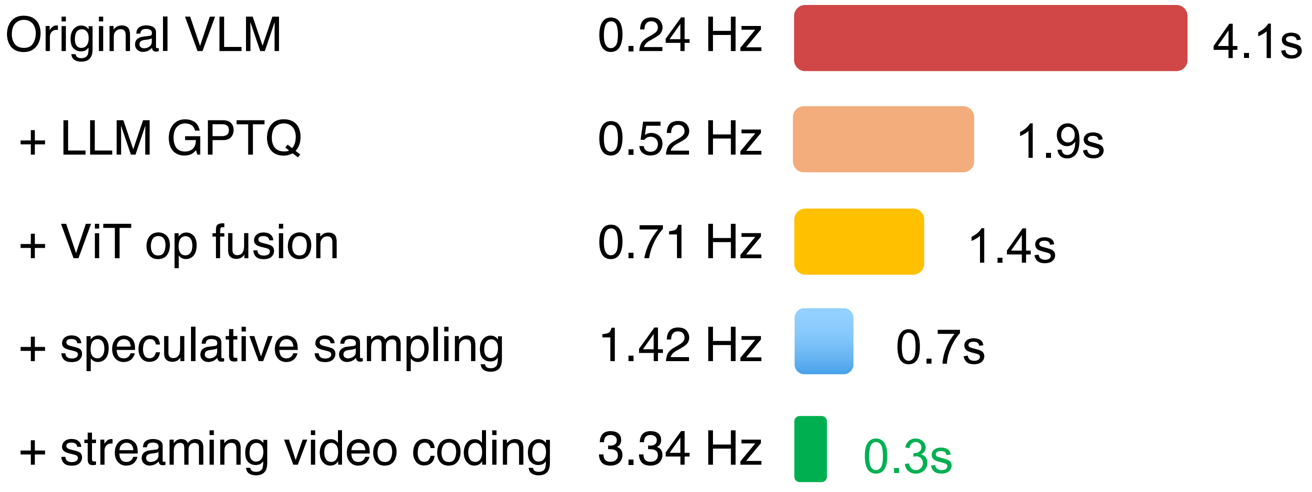
Decision Description: Decelerate and change lanes to the left, keeping a safe distance from the construction workers on the left front side.

SUP-AD Dataset
Complex and Long-tail Driving Scenarios

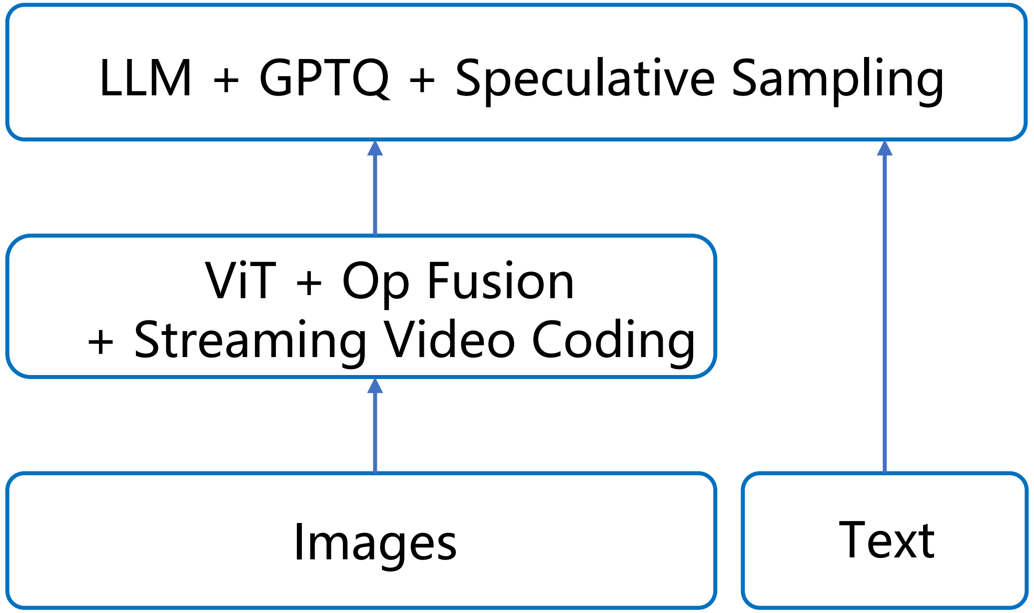
SUP-AD Dataset
Complex and Long-tail Driving Scenarios

DriveVLM-Dual Onboard Deployment

2.2 Billion Parameters



Inference speed comparison



Model architecture optimization

Credit to Li Auto

DriveVLM

The Convergence of Autonomous Driving and Large Vision-Language Models

TSINGHUA MARS LAB LI AUTO

Dealing with super complex scenarios?

Front View

Speed:
3.07 km/h

Navigation：
Turn right at the current intersection

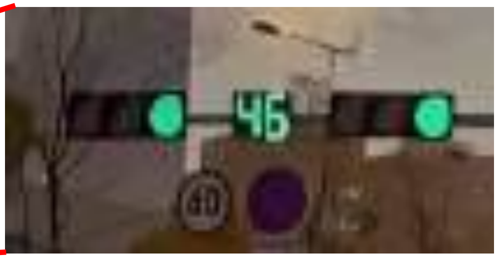
Visual Reasoning

Front View

Speed:
3.07 km/h

Navigation：
Turn right at the current intersection

Description The scene shows an intersection with multiple vehicles and pedestrians. There are motorcyclists and cars nearby, requiring careful navigation. I need check the traffic light status.



{"view_index": "front", "bbox": [80, 140, 120, 160], "description": "the traffic light at the intersection"}

The traffic light is green and there are 46 seconds left, which means I need to wait cautiously for pedestrians to cross. I will check front-right next.

{"frame_index": "current", "view_index": "front_right"}

The front-right view captures motorcyclists and pedestrians crossing the intersection, indicating a busy area.

Reasoning Given that the current speed is low and road condition is busy, stopping is necessary before proceeding with the turn. After ensuring that there are no pedestrians in front of the vehicle, accelerating with a right turn aligns with the navigation command.

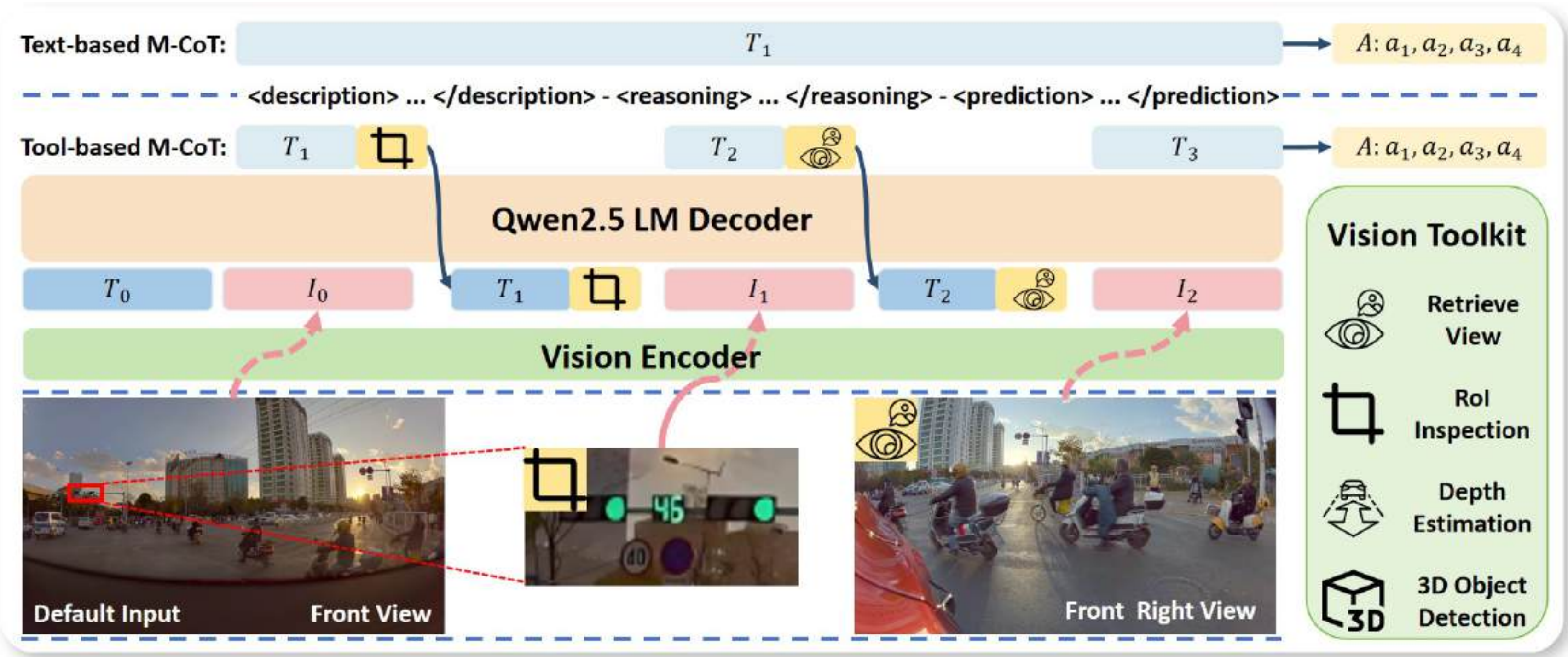
Prediction The initial actions involve stopping and assessing the environment. After ensuring safety, accelerating with a right turn.

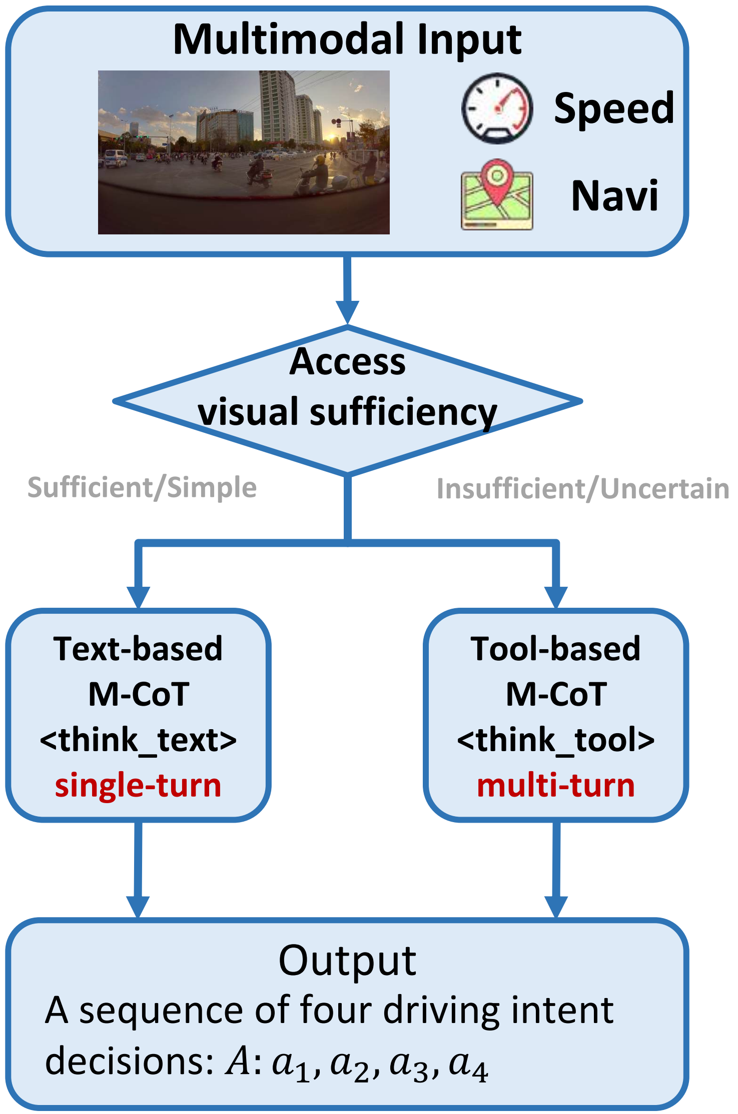
["Stop, Straight", "Stop, Straight", "Stop, Right Turn", "Accelerate, Right Turn"]

DriveAgent-R1

Advancing VLM-based Autonomous Driving with Active Perception and Hybrid Thinking, arXiv:2507.20879

Step 1: Assess visual information sufficiency
Step 2: Select thinking mode
Sufficient: <think_text> (single-turn)
Insufficient: <think_tools> (multi-turn)





Vision Toolkit

Retrieve View

- **On-demand request** for images from 6 cameras
- **Historical Memory Pool**
  - ✓ 5-second cache of all views.
  - ✓ Enables "look back" capability.
  - ✓ Avoids processing entire video clips.

RoI Inspection

- **Provides a "zoom-in" capability**
  - ✓ Utilizing the native grounding capability of VLMs.
- **Inspects specific high-res Regions of Interest**
  - ✓ Confirms critical details (e.g., traffic light status, sign text).

Depth Estimation

- Uses a SOTA monocular algorithm
- **Provides a high-fidelity depth map**
  - ✓ Helps understand object distances & spatial layout.

3D Detection

- **Open-vocabulary** 3D object detection tool
- **Exceptional flexibility**
  - ✓ Dynamically specifies novel objects to detect based on context.

Results

Front View

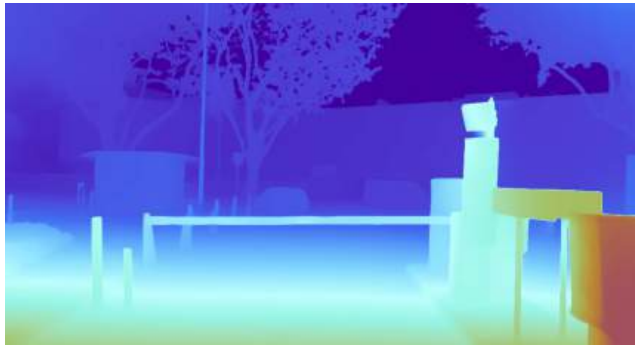
Speed:
5.20 km/h

Navigation：
Continue straightafter passing the barrier gate

Reasoning under $\mathcal{M}_{tool}$

Description The vehicle is approaching a barrier gate at the intersection. There are traffic cones on the left side, indicating a controlled entry point.

Reasoning Given the low speed and the presence of traffic cones, it's necessary to decelerate initially to ensure safety as the vehicle approaches the barrier gate. Once safely past the gate, maintaining a steady speed while continuing straight aligns with the navigation command.



{"view_index": "front"}

The depth map shows that the barrier gate and traffic cones are relatively close, with need to decelerate initially. The road ahead appears clear after the gate.

Prediction The vehicle should decelerate initially, once past the gate, maintain a steady speed and continue straight.

["Decelerate, Straight", "Decelerate, Straight", "Keep Speed, Straight", "Accelerate, Straight"]

Results

Front View

Speed:
12.8 km/h

Navigation：
go straight through the intersection

Reasoning under M_{tool}

Description The vehicle ahead is partially on the road. A motorcycle is also present on the right side of the road. The current speed is relatively low at 12.78 km/h.

{"view_index": "front", "bbox": [205, 180, 330, 225], "description": "the vehicle ahead"}

The RoI Inspection image confirms the presence of a vehicle and a motorcycle on the road, indicating a minor collision.

Reasoning Given the minor collision and the low speed, it's necessary to decelerate initially to ensure safety. Due to the ongoing situation, stopping is prudent until further assessment is possible.

Prediction Decelerating twice ensures safety, followed by maintaining a steady speed. Stopping ensures adherence to the traffic accident scenario.

["Decelerate, Straight", "Decelerate, Straight", "Keep Speed, Straight", "Stop, Straight"]

Manipulation is highly similar to Driving

✓Learning paradigm
- Imitation learning from human demonstrations
- Steering wheels *vs.* Teleoperation device

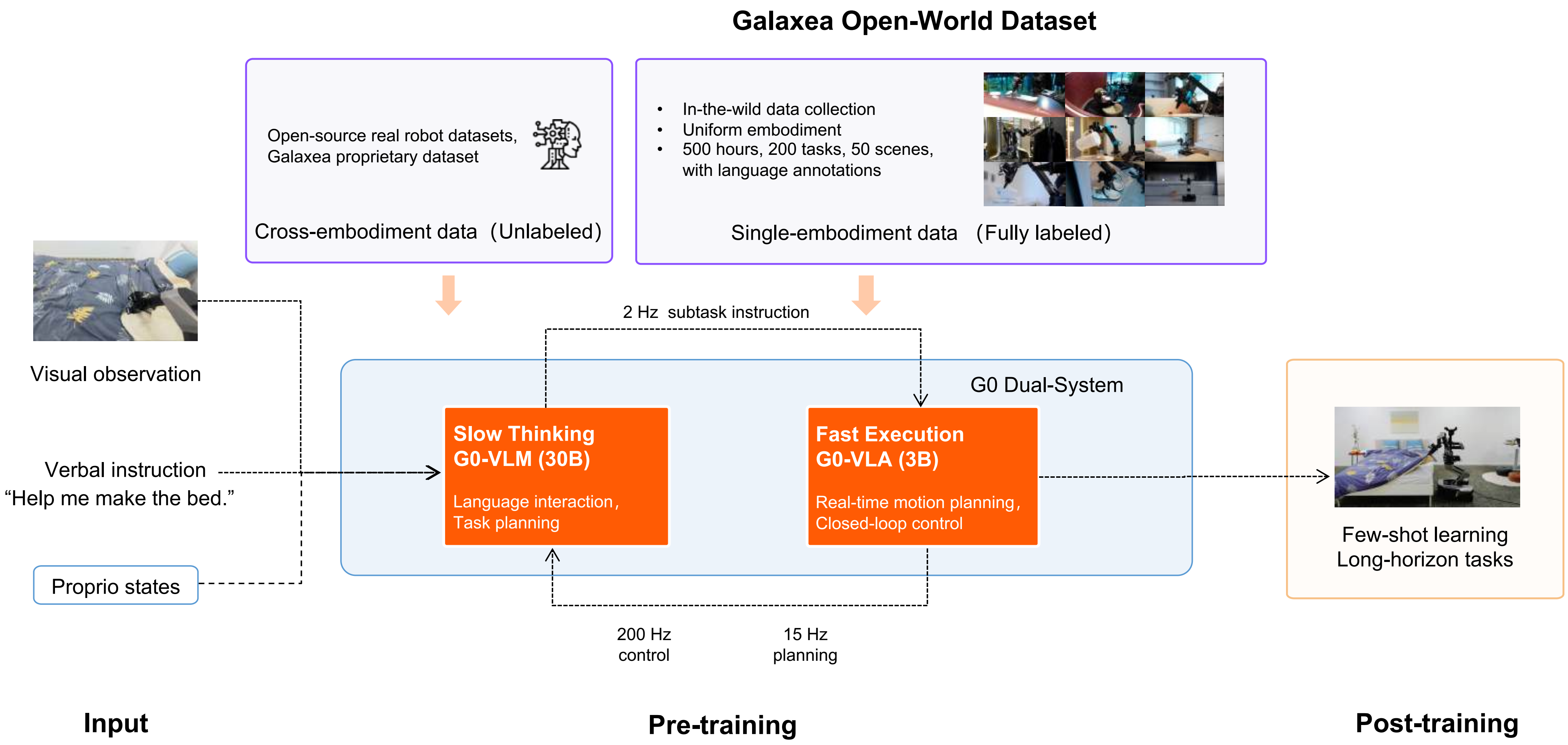
✓Input: camera images

✓Output: trajectories / joint angles of the arms and grippers

Galaxea R1 Lite
Co-design with Physical Intelligence

π-0.5， Physical Intelligence

Galaxea G0: Dual-System VLA



Galaxea Open-World Dataset

Galaxea Open-World Dataset

Galaxea Open-World Dataset Open-Source Review

Download count across two platforms exceeds **120k**

Dominate the rankings on Huggingface and Modelscope

Download count on Hugging Face 75.1k

Ranking of total download counts for robot real-world datasets	Global NO. 1
Comprehensive trend ranking of robot datasets	Global NO. 1
Ranking of total download counts for robot datasets	Global NO. 2

Download count on ModelScope 47.4k

Comprehensive popularity ranking of datasets

Global NO. 1

Foundation Models for Embodied Intelligence

Driving

Manipulation

Navigation

Locomotion

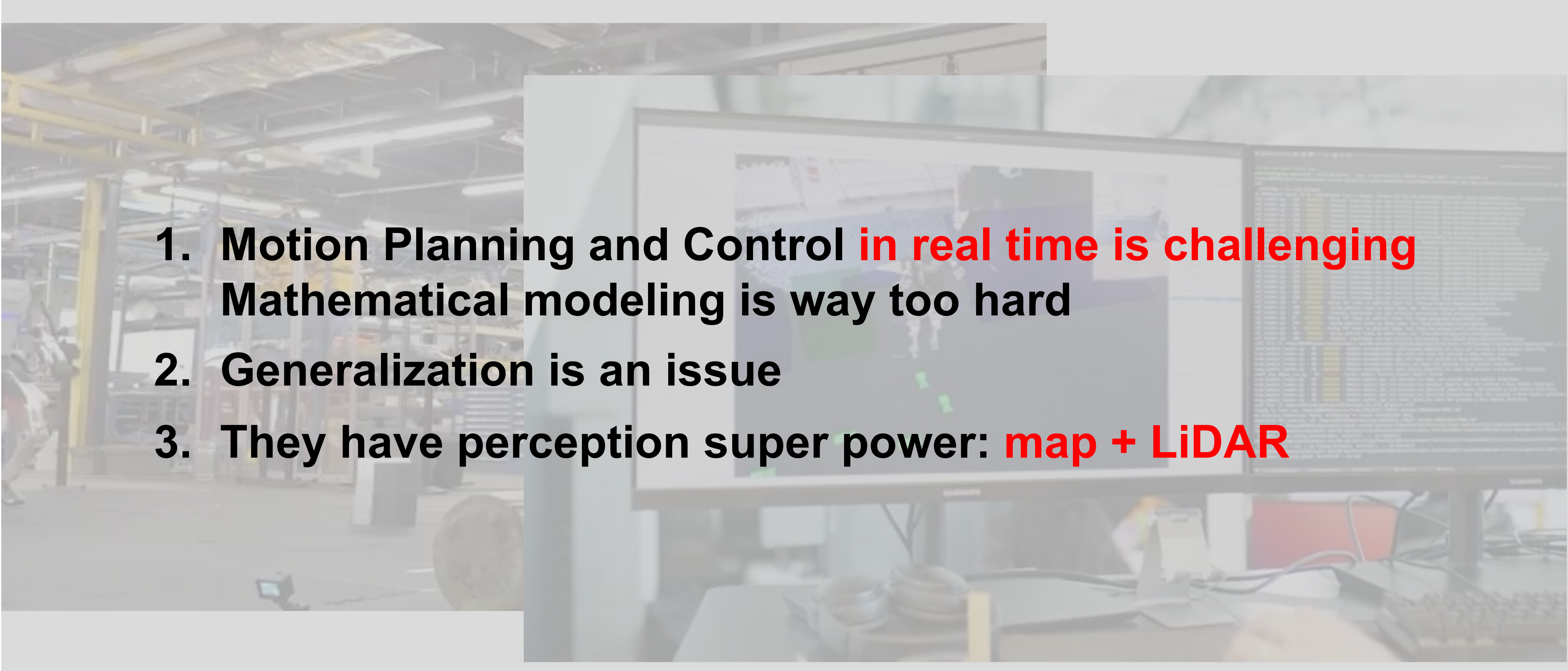
Robot Parkour Learning, CoRL 2023
Humanoid Parkour Learning, CoRL 2024
Embrace Collisions, CoRL 2025
VR-Robo, RAL 2025

Language Foundation Models
Reasoning Intelligence

Vision Foundation Models
Spatial Intelligence

Control Foundation Models
Athletic Intelligence

Perceptive Locomotion



1. Motion Planning and Control **in real time is challenging**
   Mathematical modeling is way too hard
2. Generalization is an issue
3. They have perception super power: **map + LiDAR**

Perceptive RL locomotion

It is important to show, how perceptive RL makes traditionally hard control problem easy

Why perceptive RL locomotion?

Robot Parkour Learning

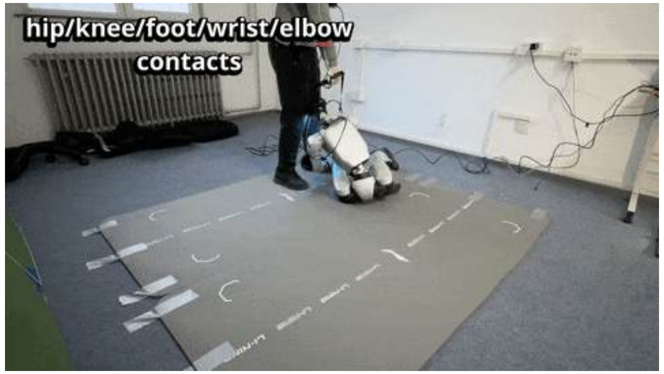
Zhuang et al., CoRL 2023 Best System Paper Finalist

Humanoid Parkour Learning

Zhuang et al, CoRL 2024

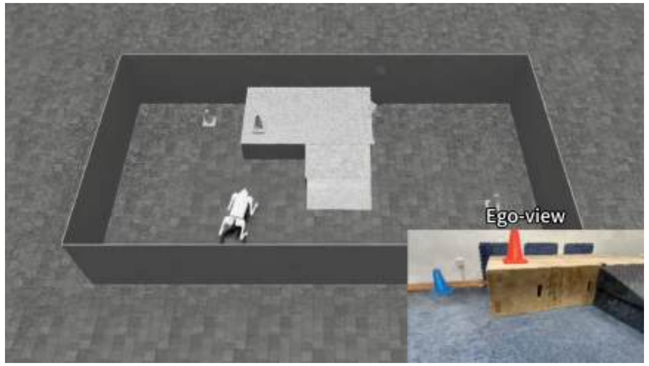
Embrace Collisions

Zhuang and Zhao, CoRL 2025

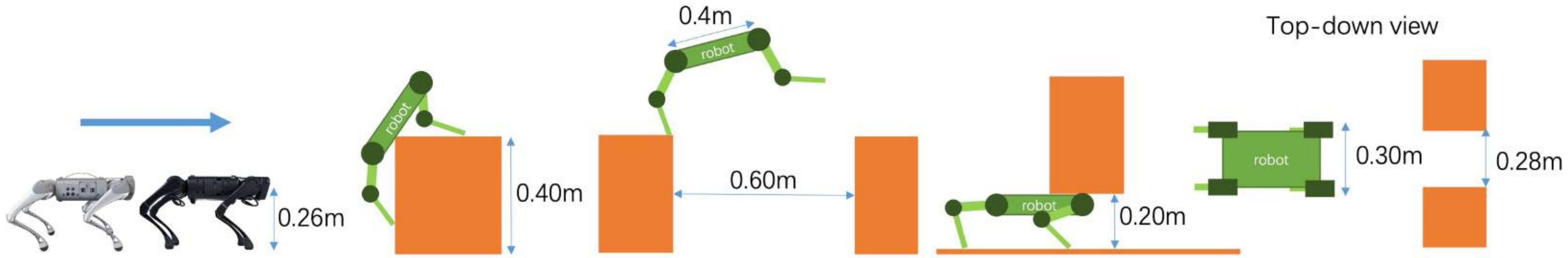


VR-Robo

Zhu et al., RAL 2025



Parkour is a great task to work with



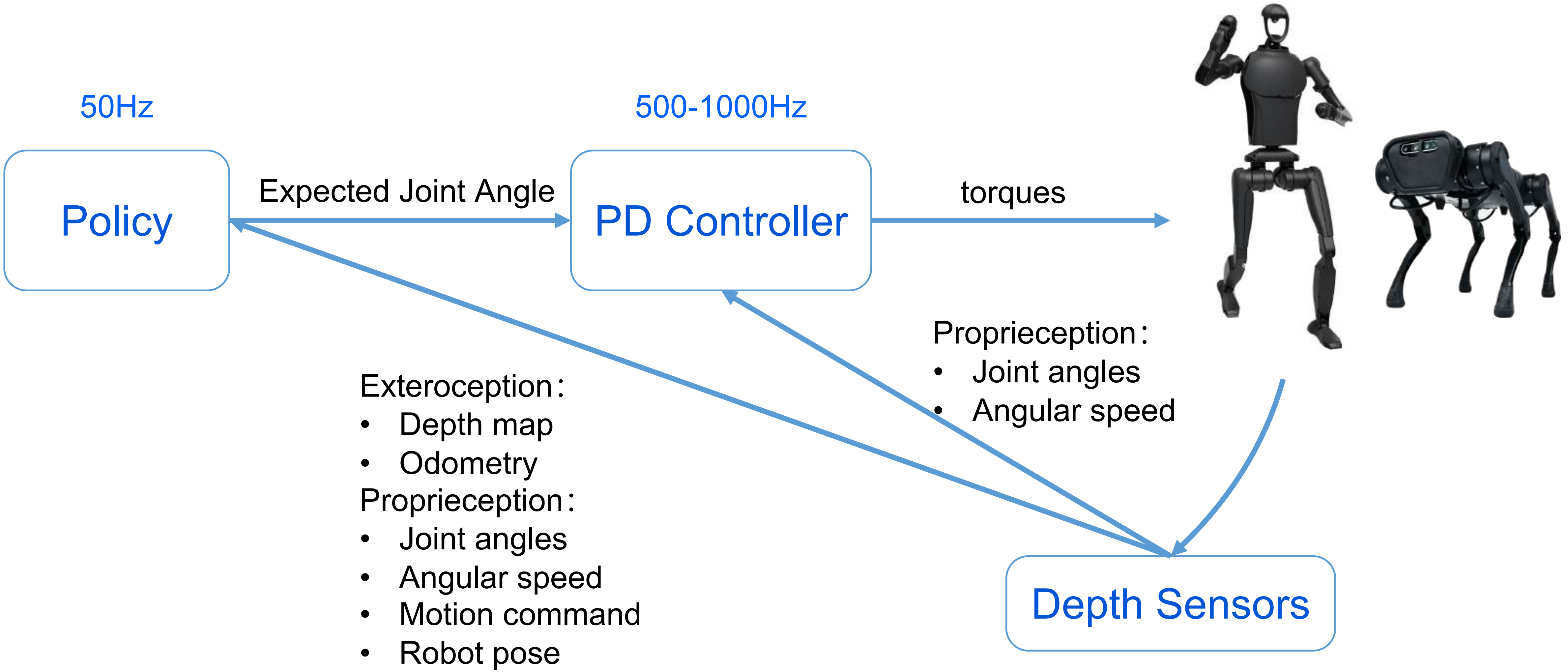
Robot Parkour Learning

Zhuang et al., CoRL 2023 Best System Paper Finalist

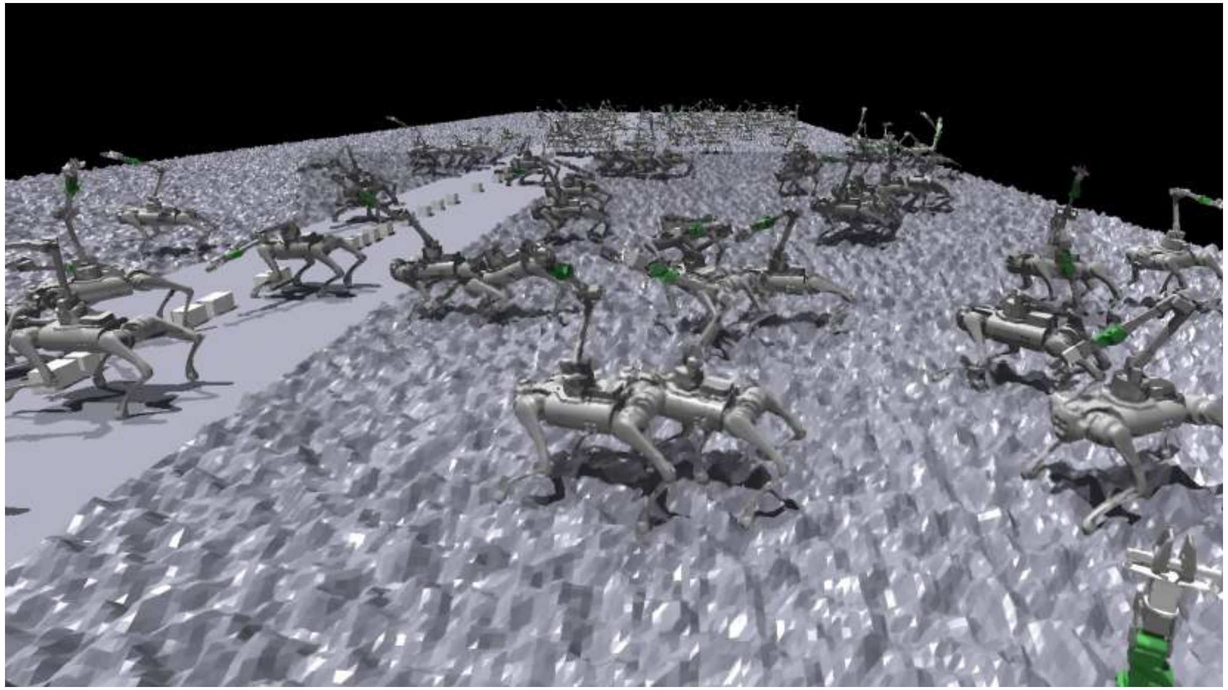
Humanoid Parkour Learning

Zhuang et al, CoRL 2024

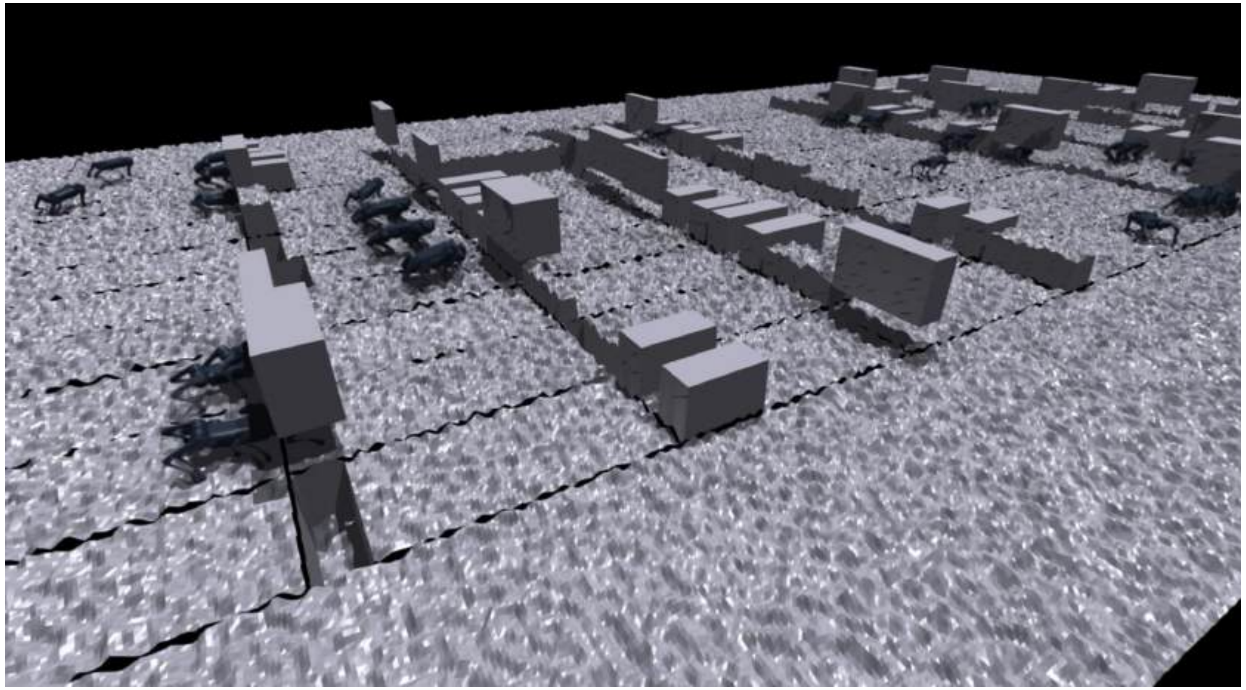
The overall feedback system



First, fractal noise in the simulator helps with robust walking

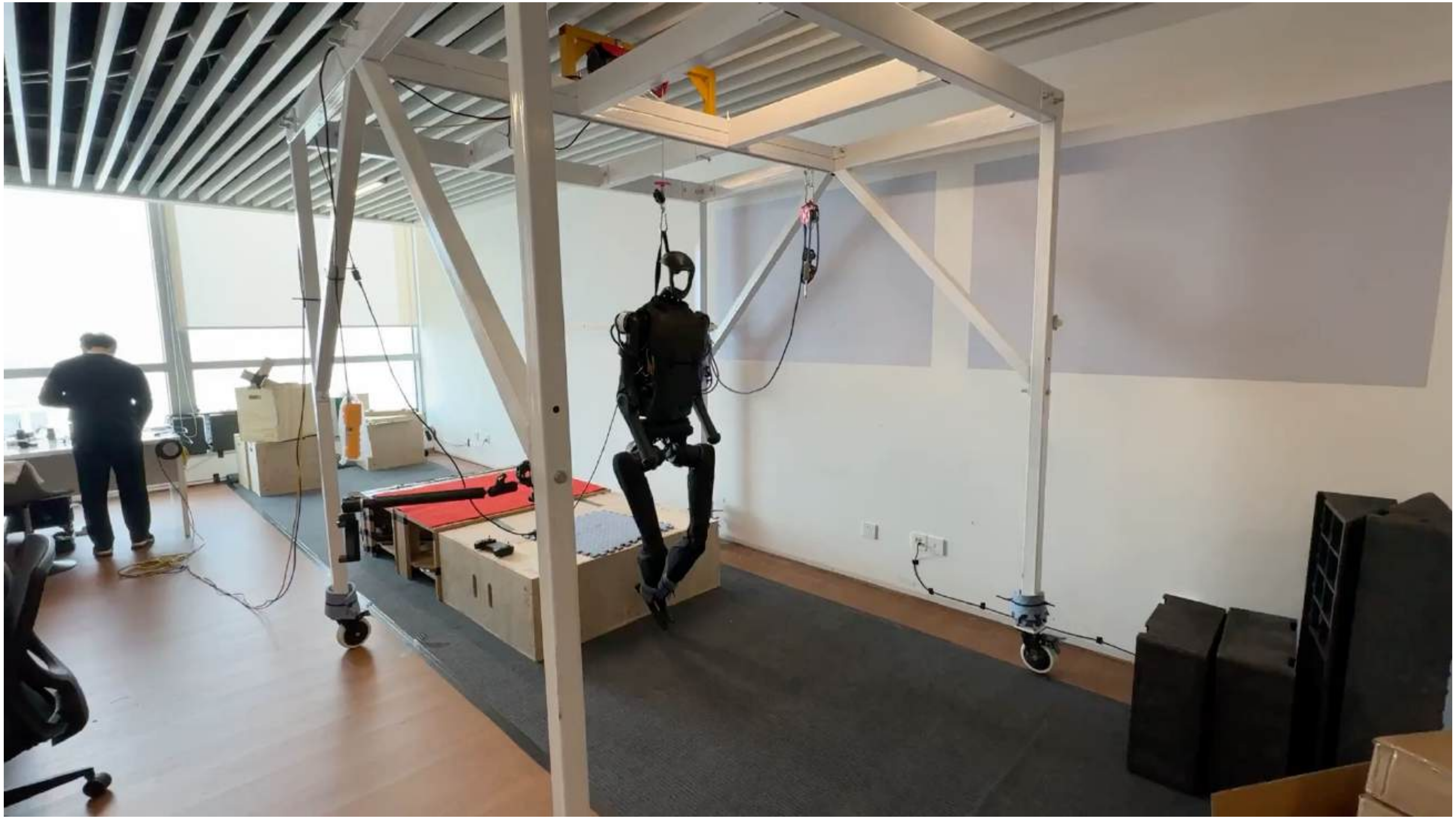


Z. Fu, X. Cheng, and D. Pathak, 'Deep Whole-Body Control: Learning a Unified Policy for Manipulation and Locomotion', in Conference on Robot Learning (CoRL), 2022.



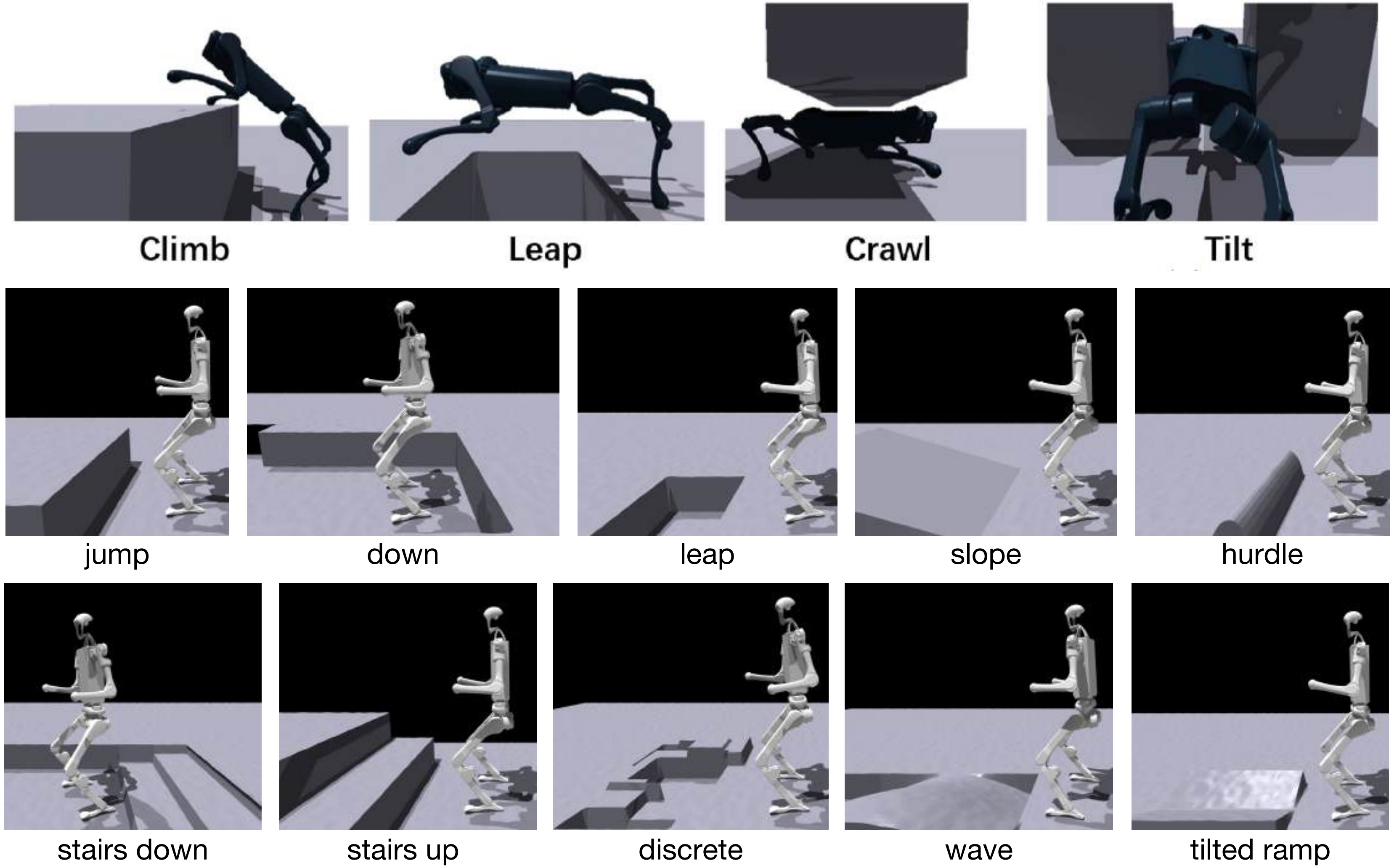
Z. Zhuang et al., 'Robot Parkour Learning', in Conference on Robot Learning (CoRL), 2023.

Our first indoor walking test



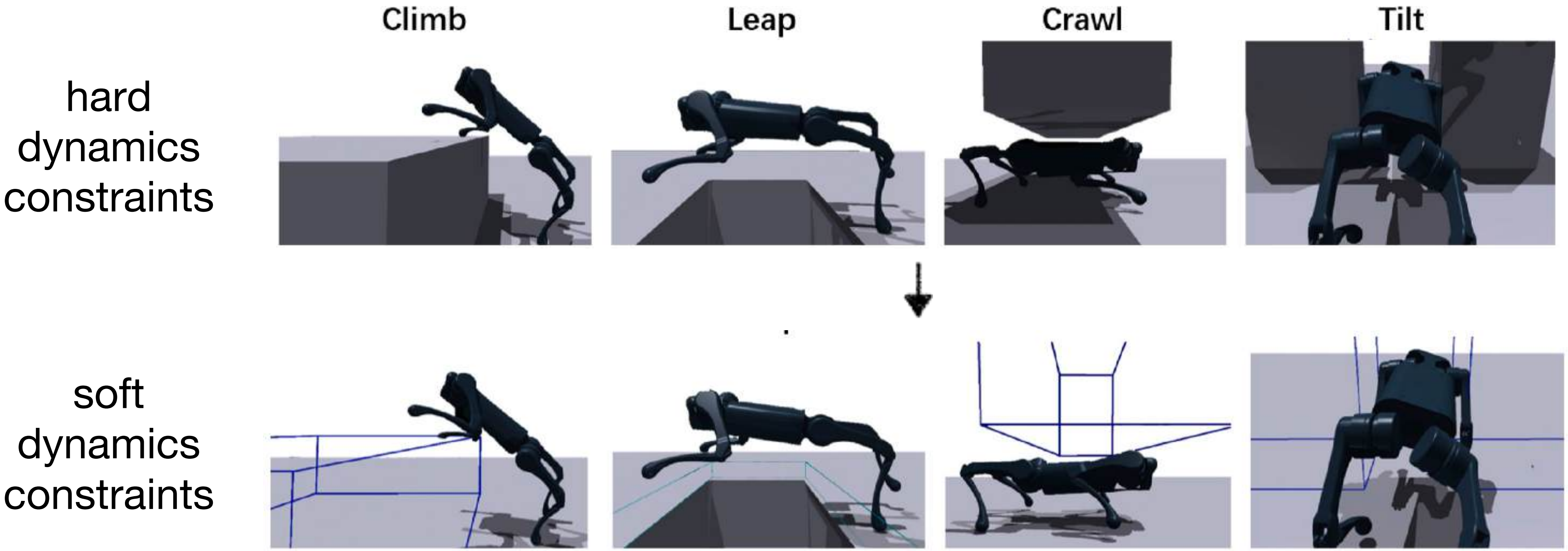
Outdoor running

Next, adding challenging terrains into the simulator



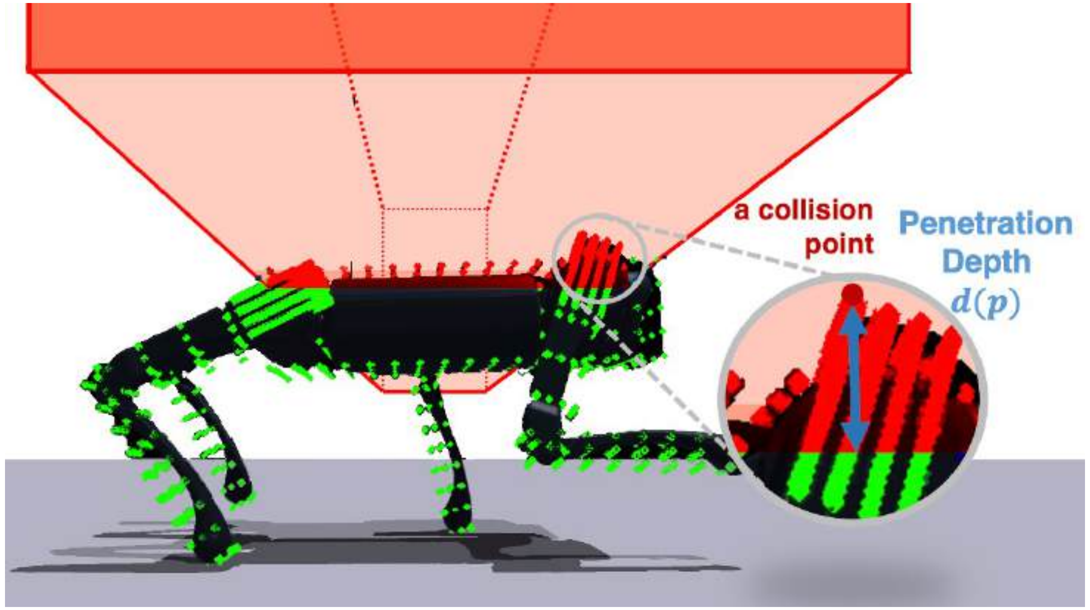
Additionally, curriculum learning helps

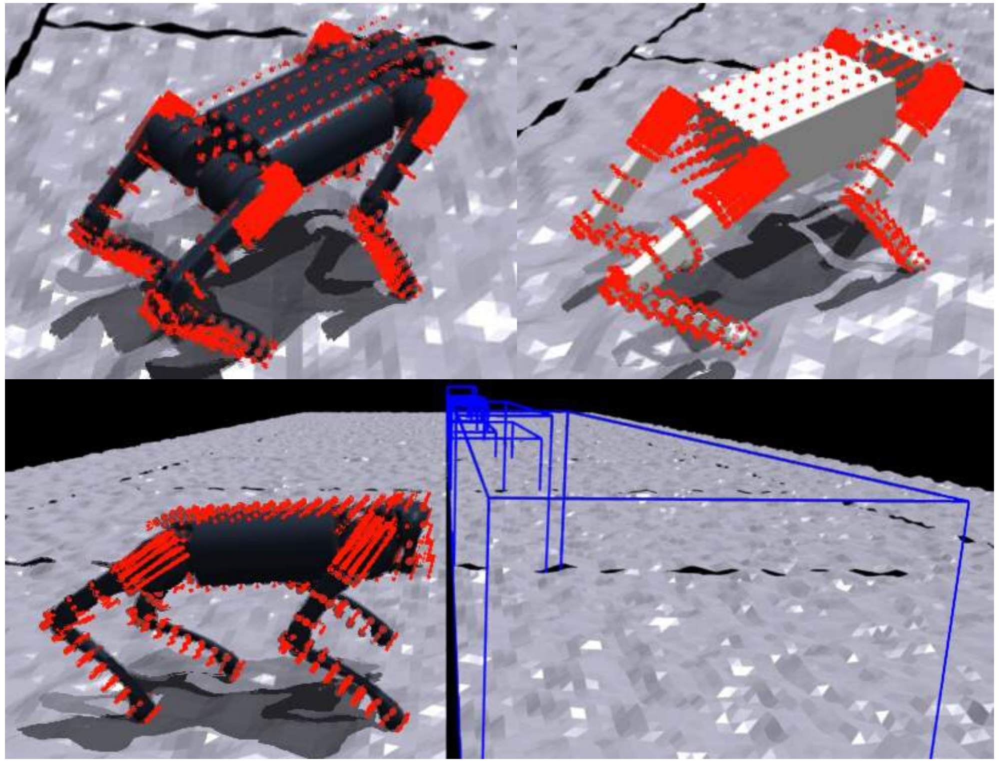
- curriculum learning and soft-dynamics constraints



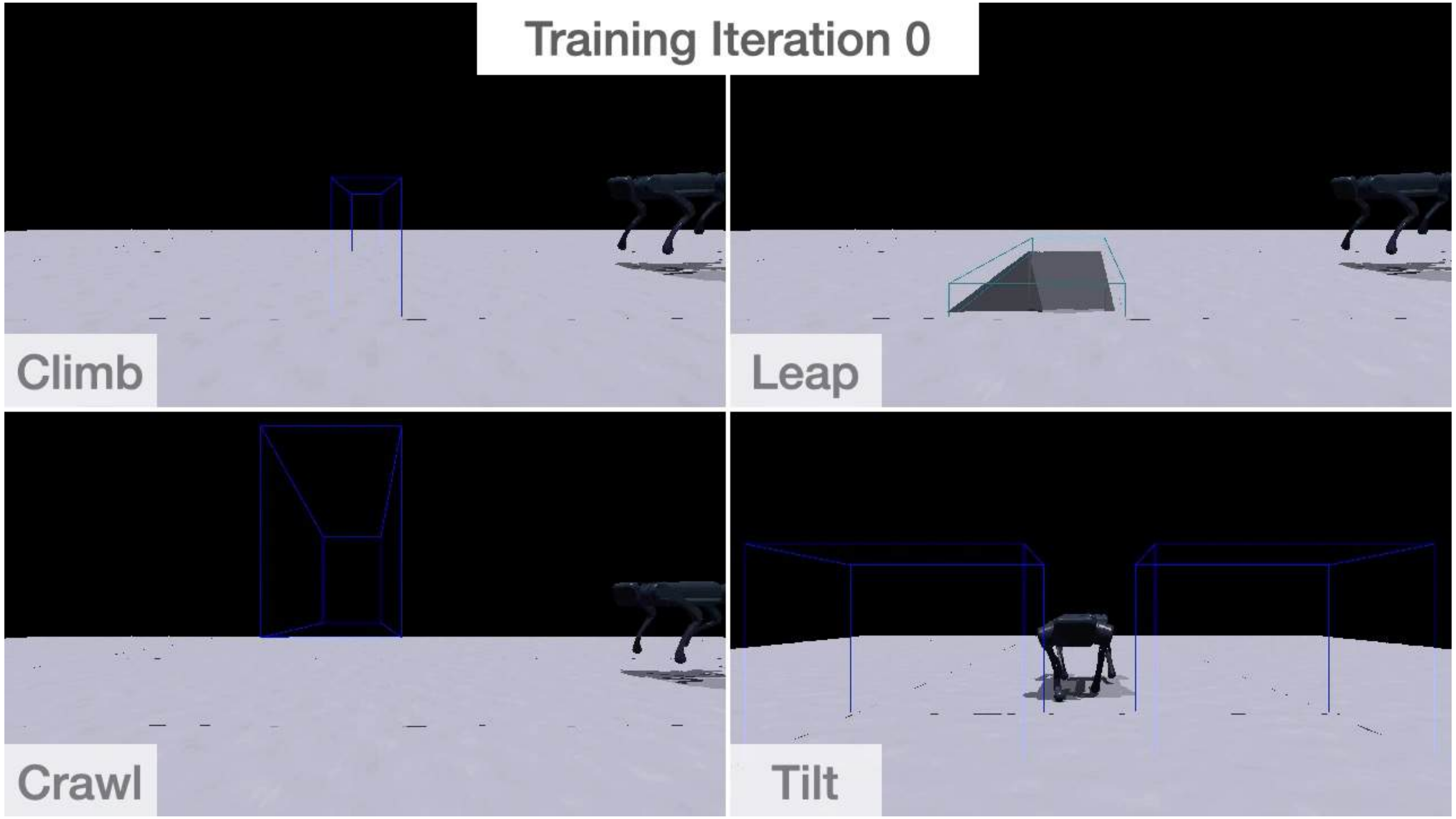
Soft Dynamics Constraints

- “softly” penalize penetration
- reward = - (# of penetration points + penetration depth)
- robots gradually learn to obey “physics”

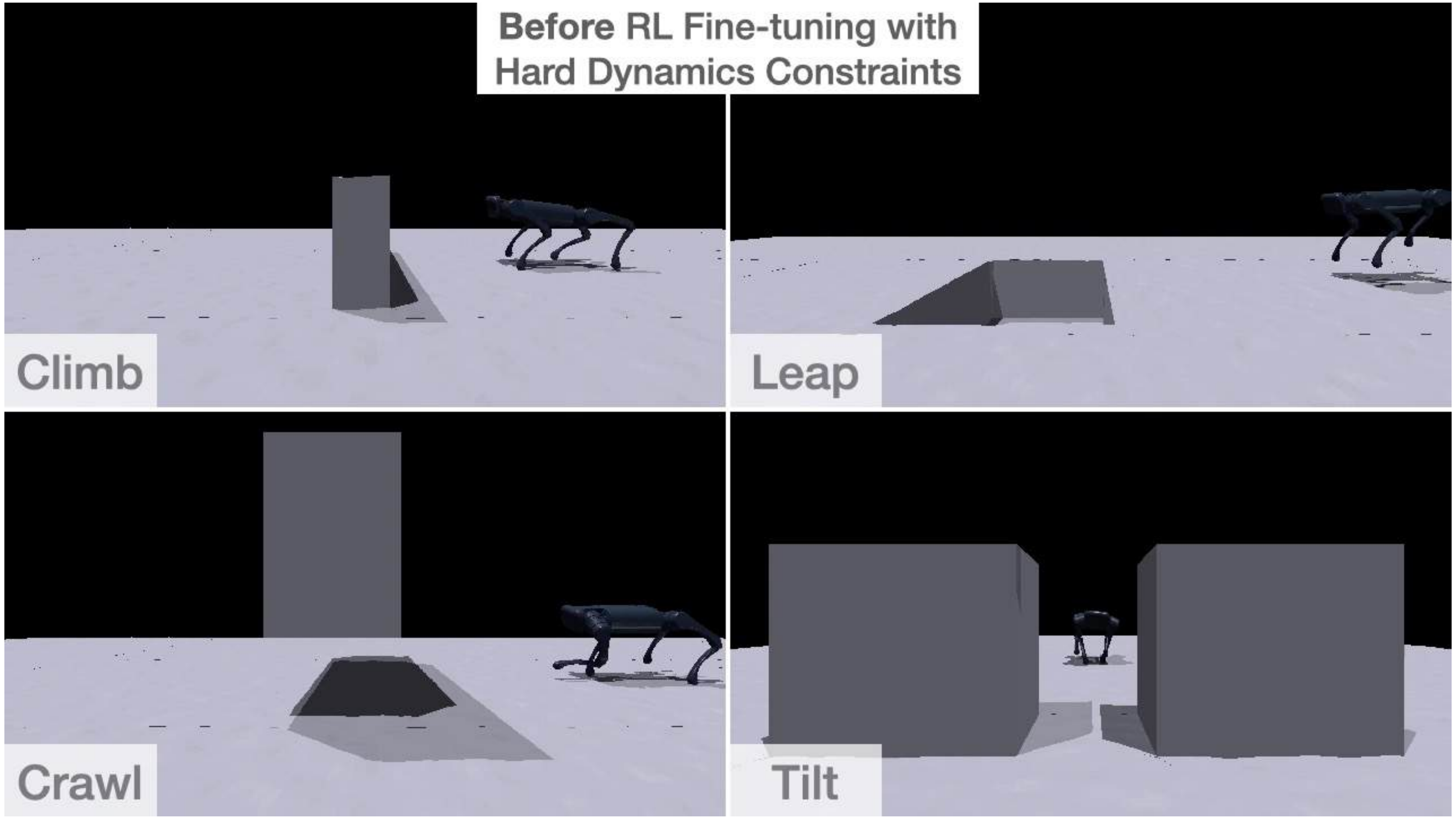




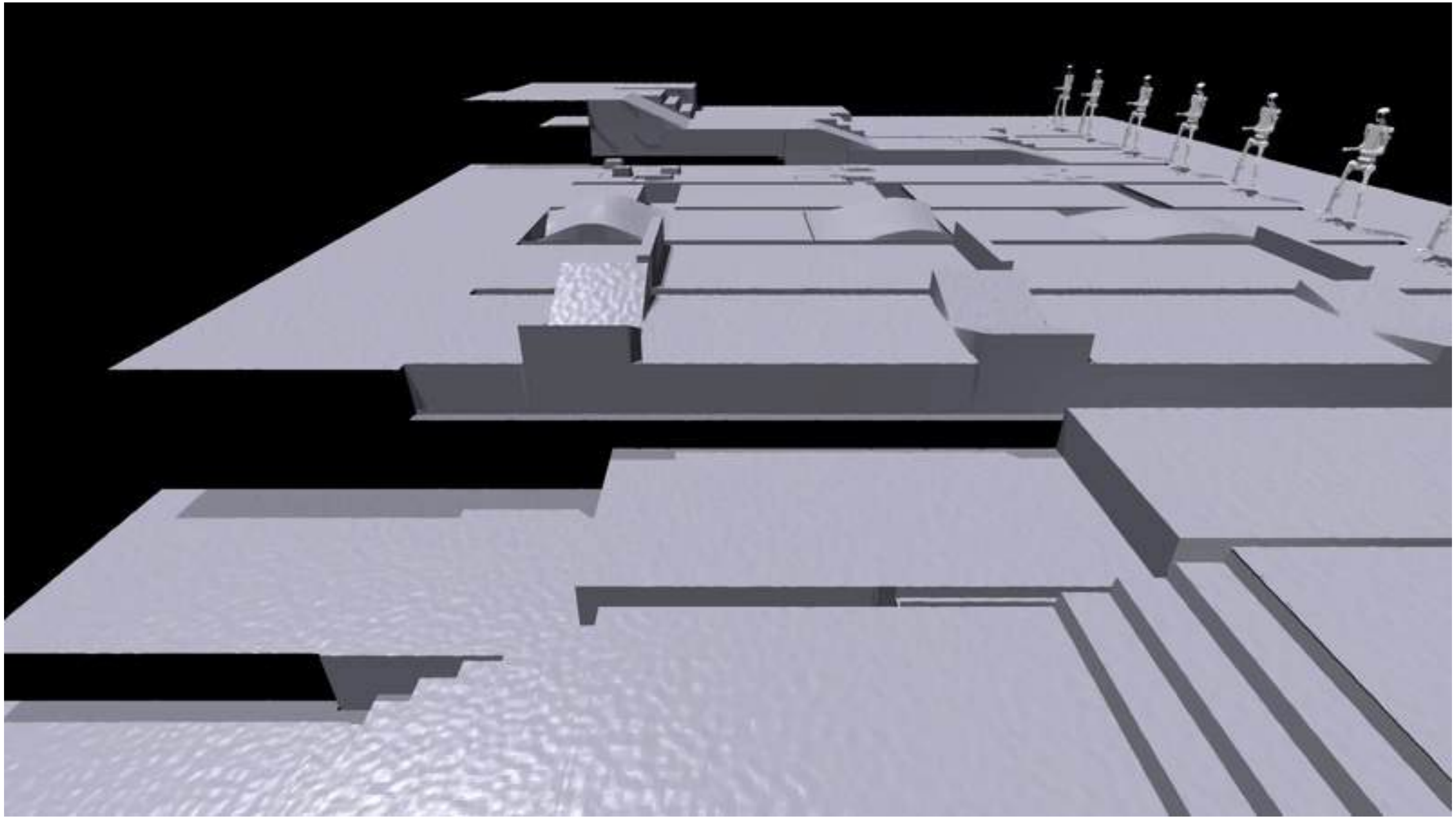
Learning with Soft Dynamics Constraints



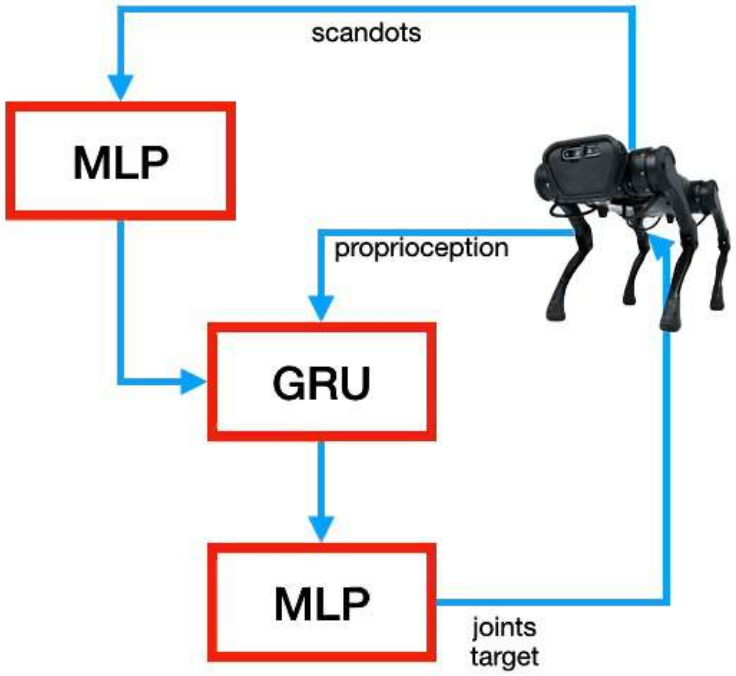
Finetune by imposing hard dynamics constraints



One policy to master them all



Finally, incorporate perception by distillation

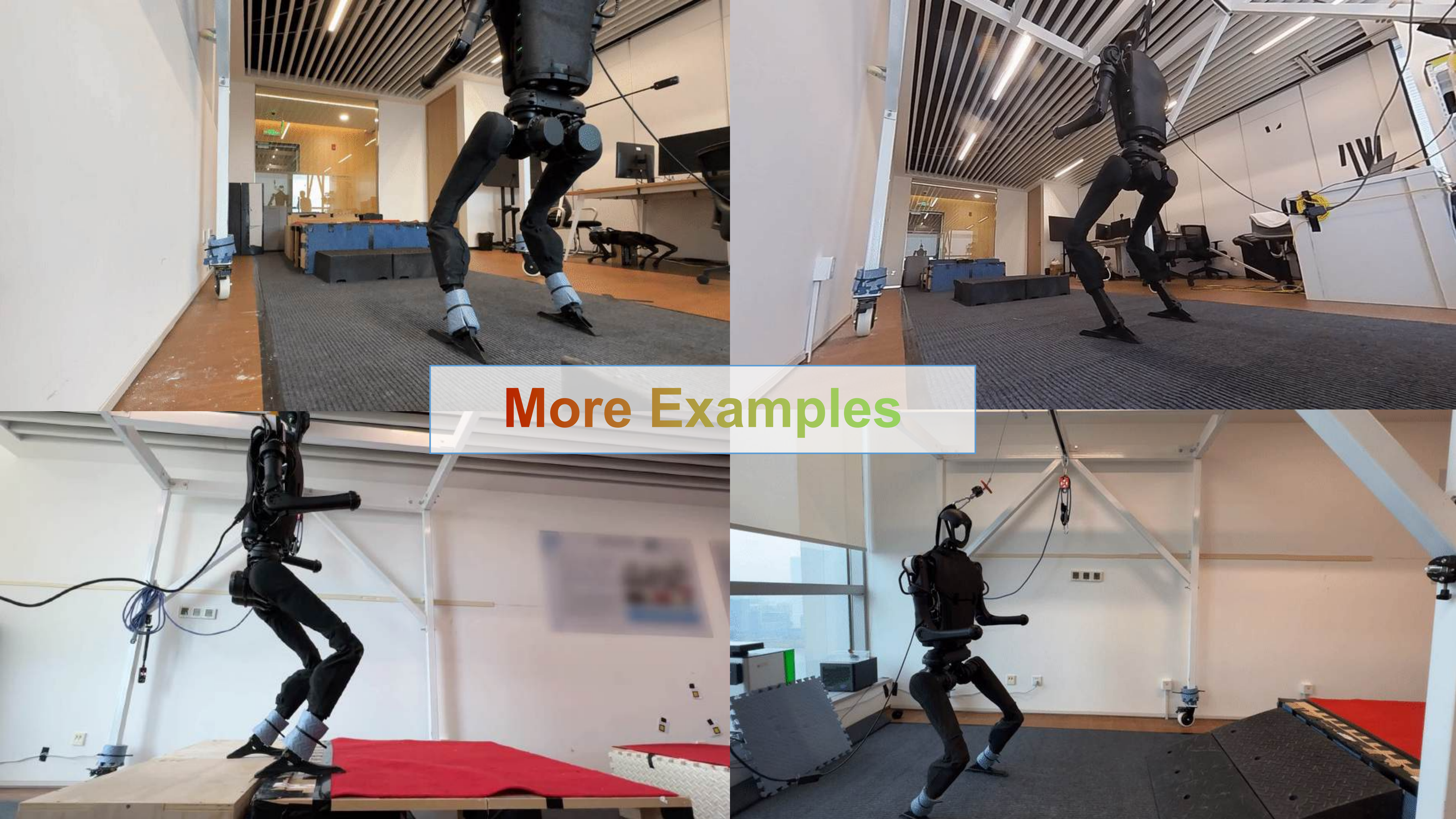


Quadruped Parkour

Humanoid Parkour

Humanoid Parkour

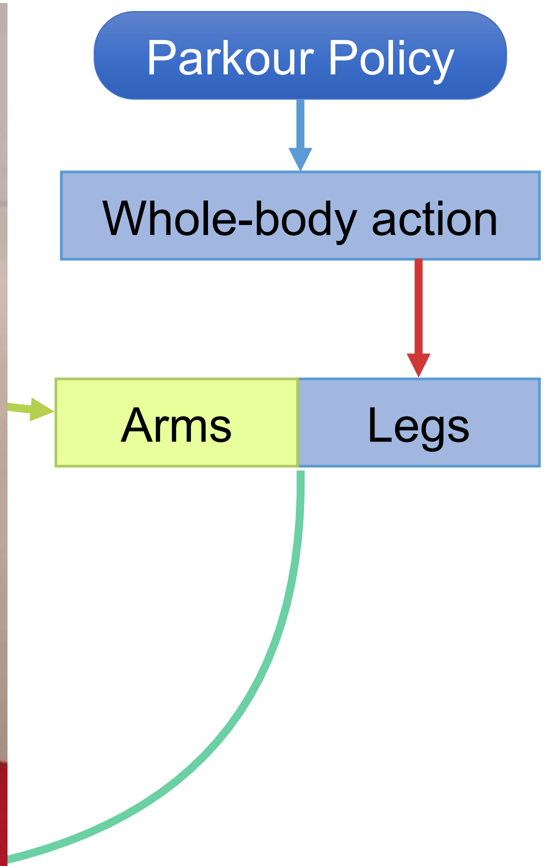
More Examples



Emerging behavior of re-trying

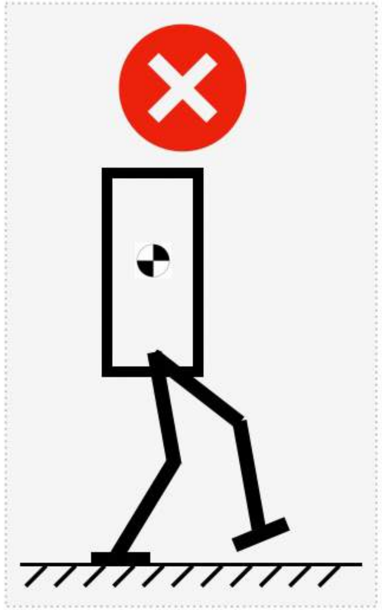
Emerging upper limb motions

Humanoid teleoperation is a free lunch

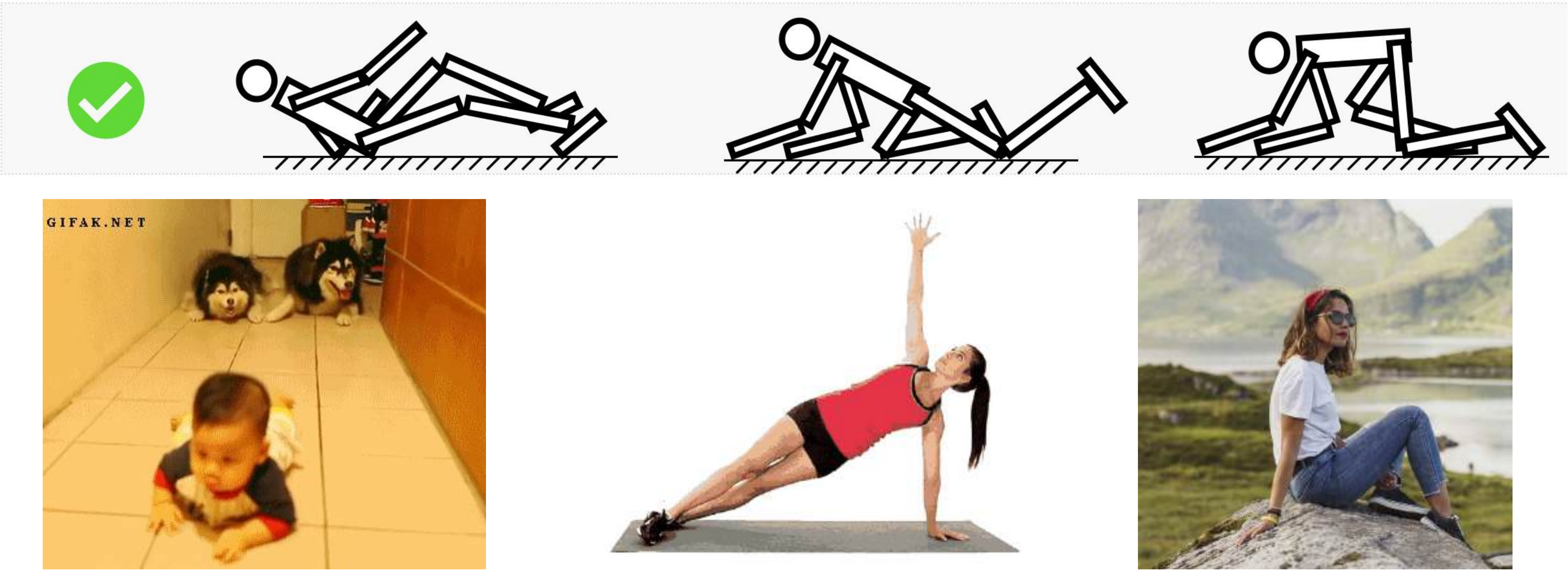


Bipedal *vs.* Humanoid

- Bipedal robots : two feet interact with the ground
- Humanoid robots : use all body parts to interact with the world



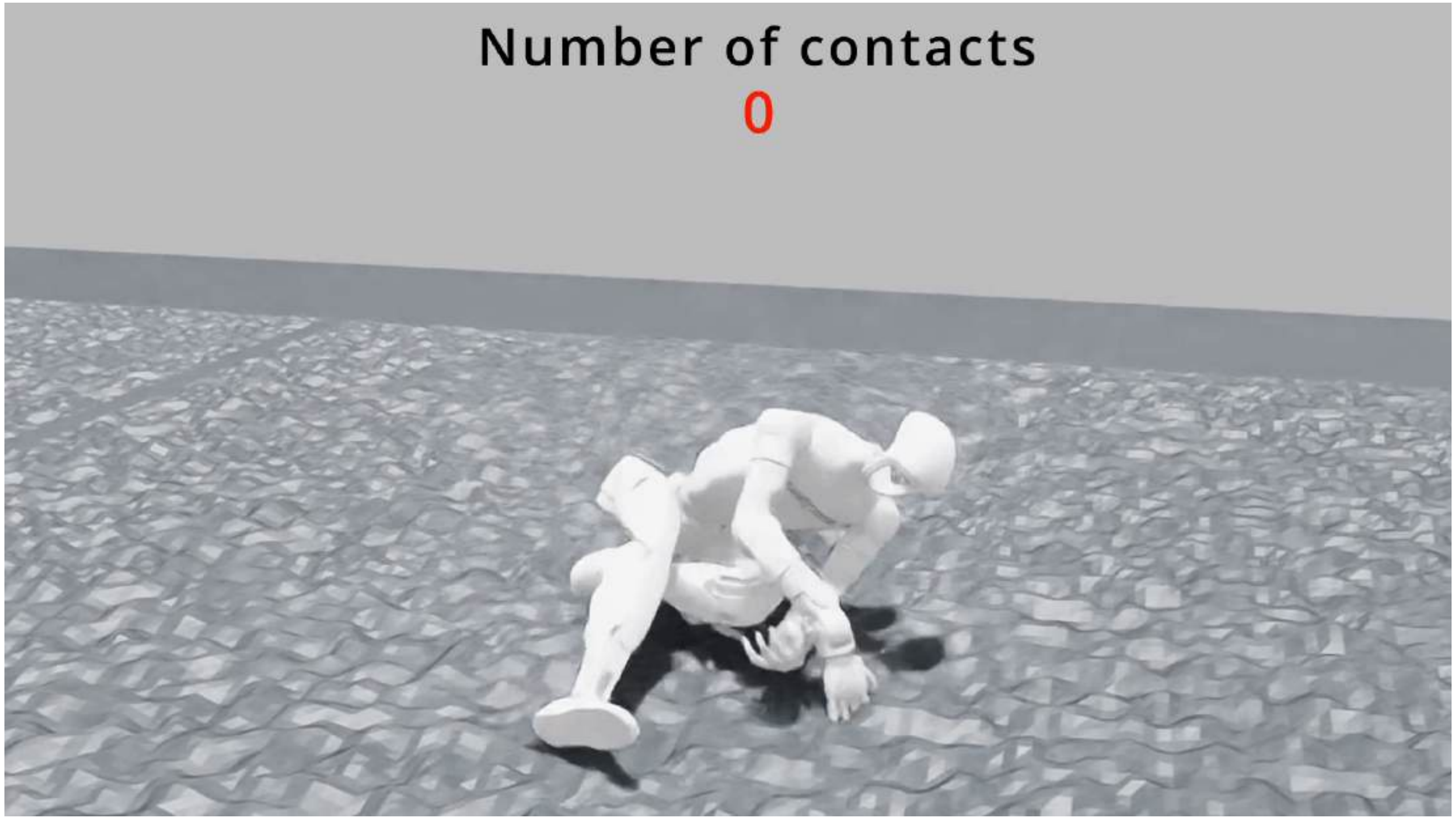
bipedal



whole body

Embrace Collisions

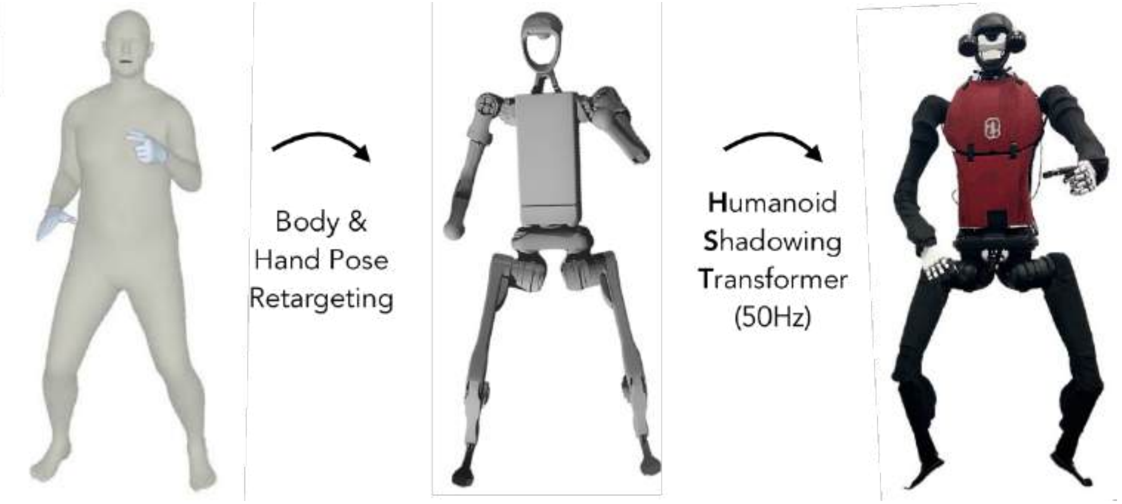
- Humanoid motions with chaotic contracts is extremely difficult to model



Motion instruction following

Predefined motions *vs.* Motion instruction following

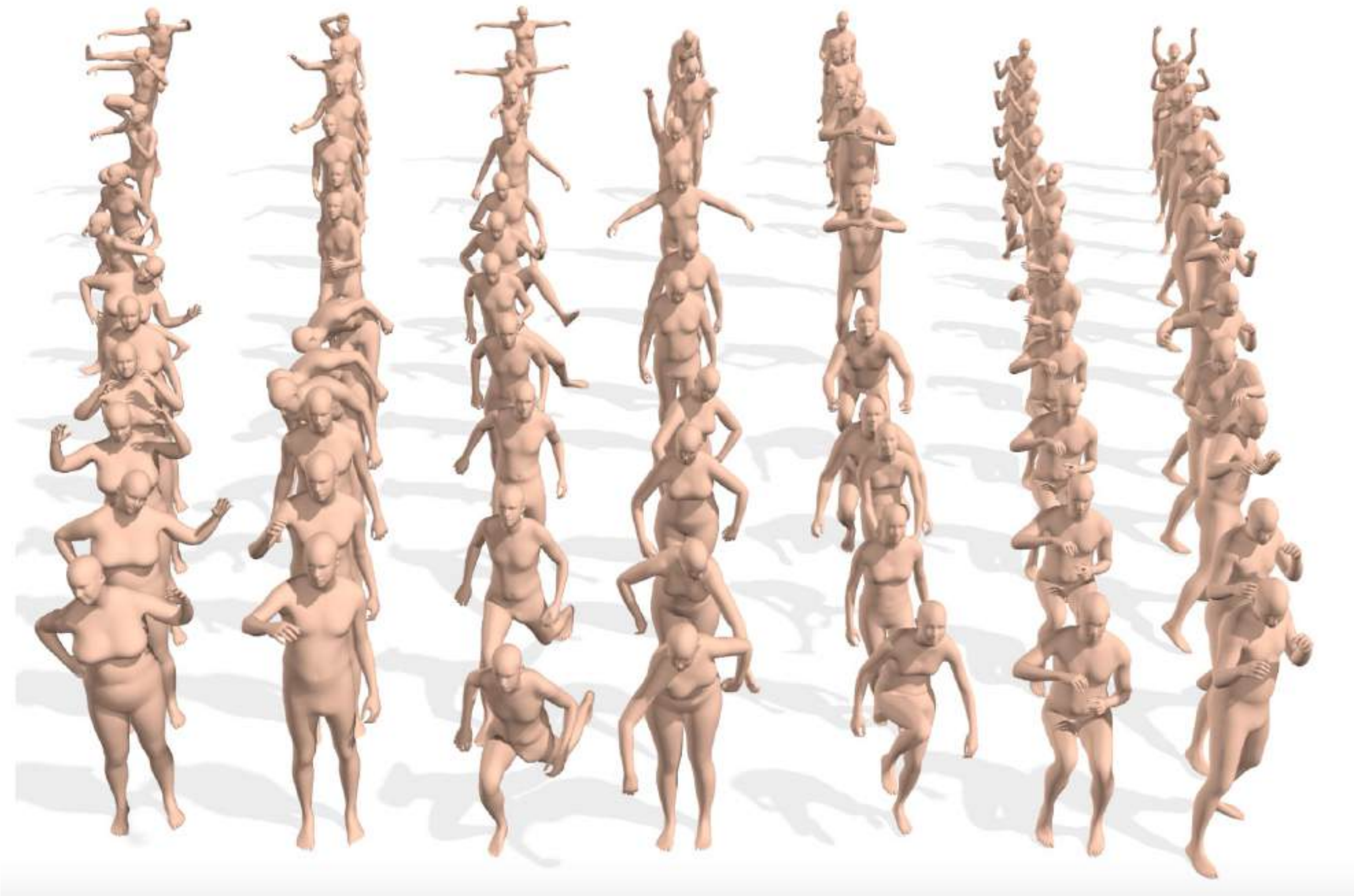
Pre-defined motions



Pic credit to HumanPlus

Motion instruction following for humanoids

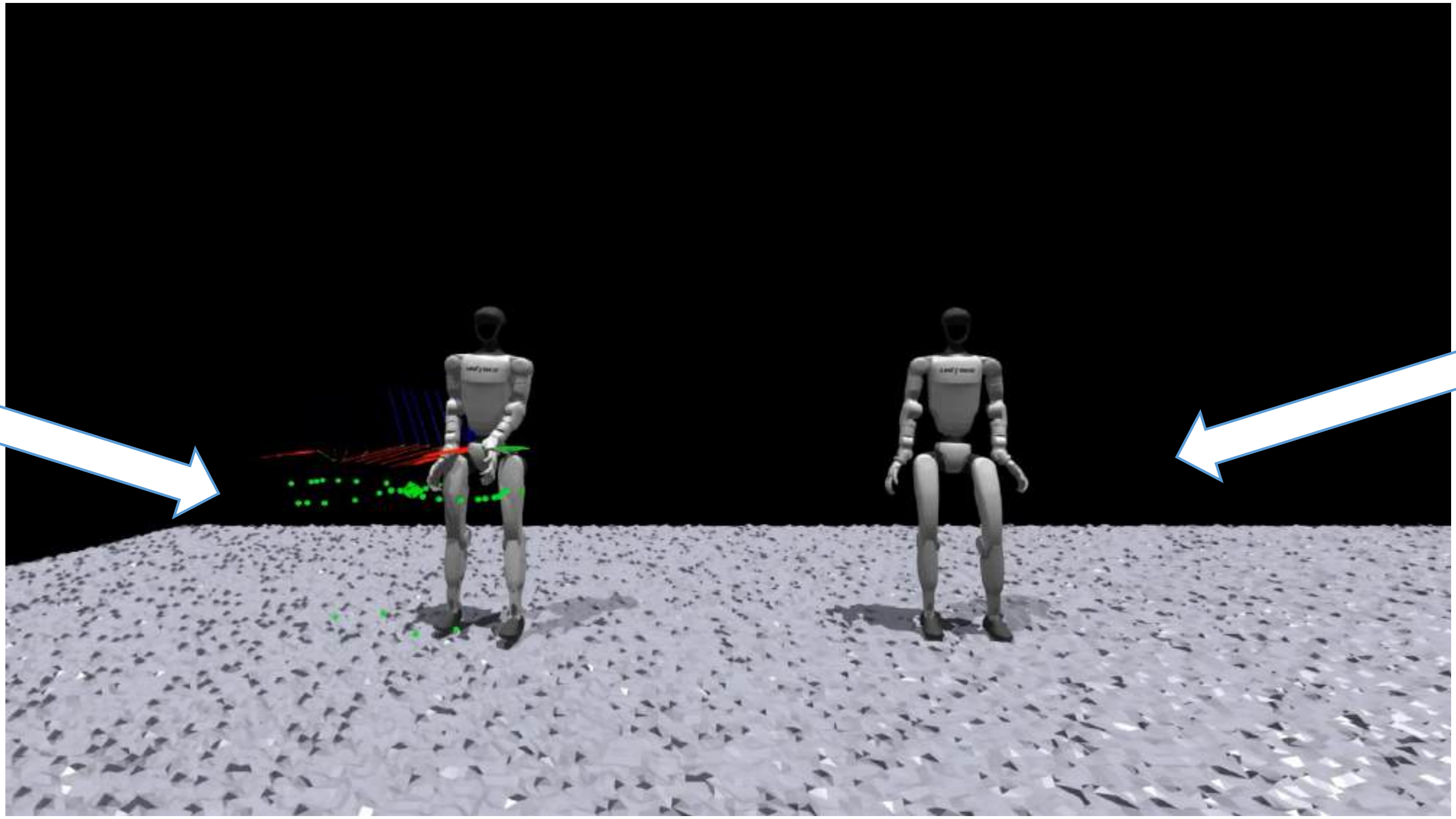
Motion dataset curation



AMASS: Archive of Motion Capture As Surface Shapes

The dataset is to blame

Our policy



Motion command

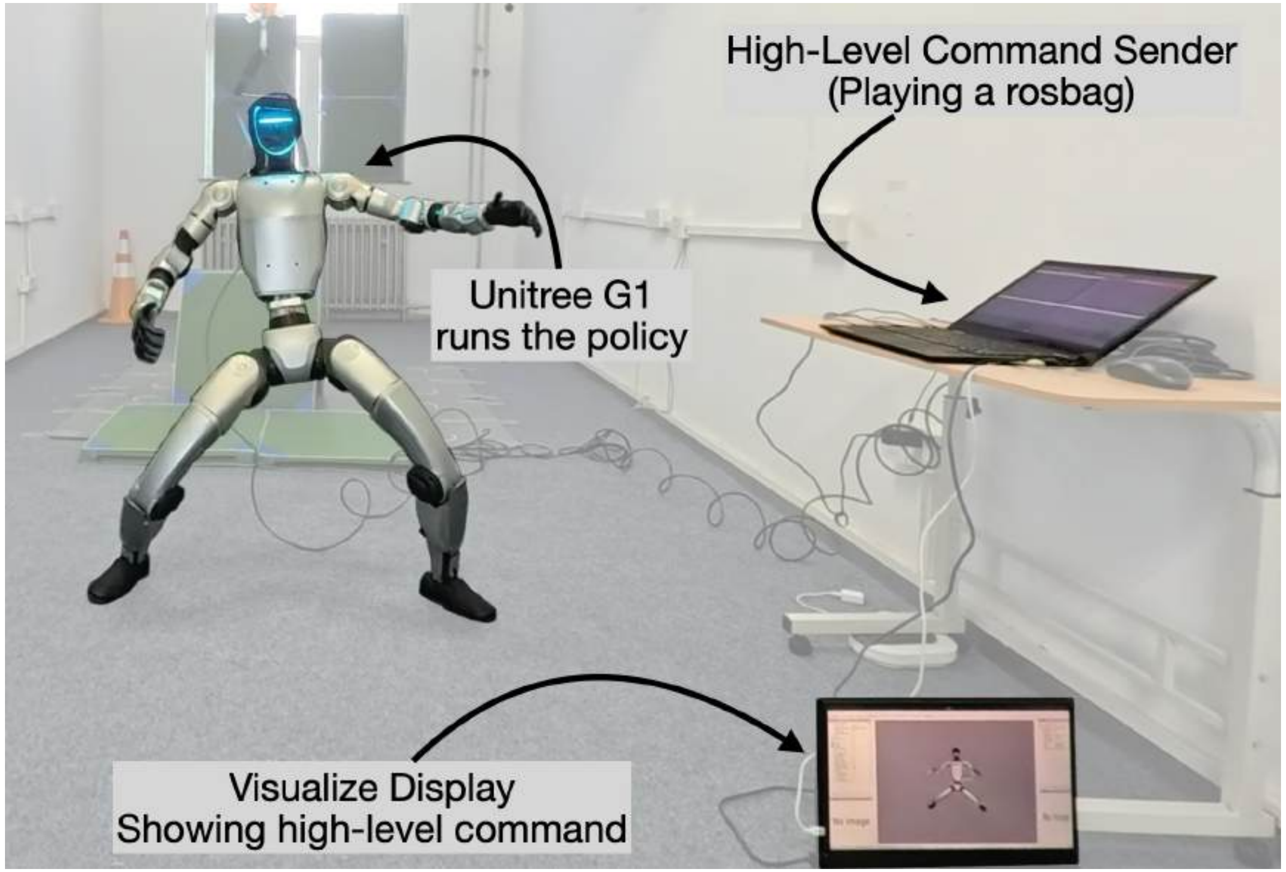
Motion dataset curation

Time to build our own dataset!

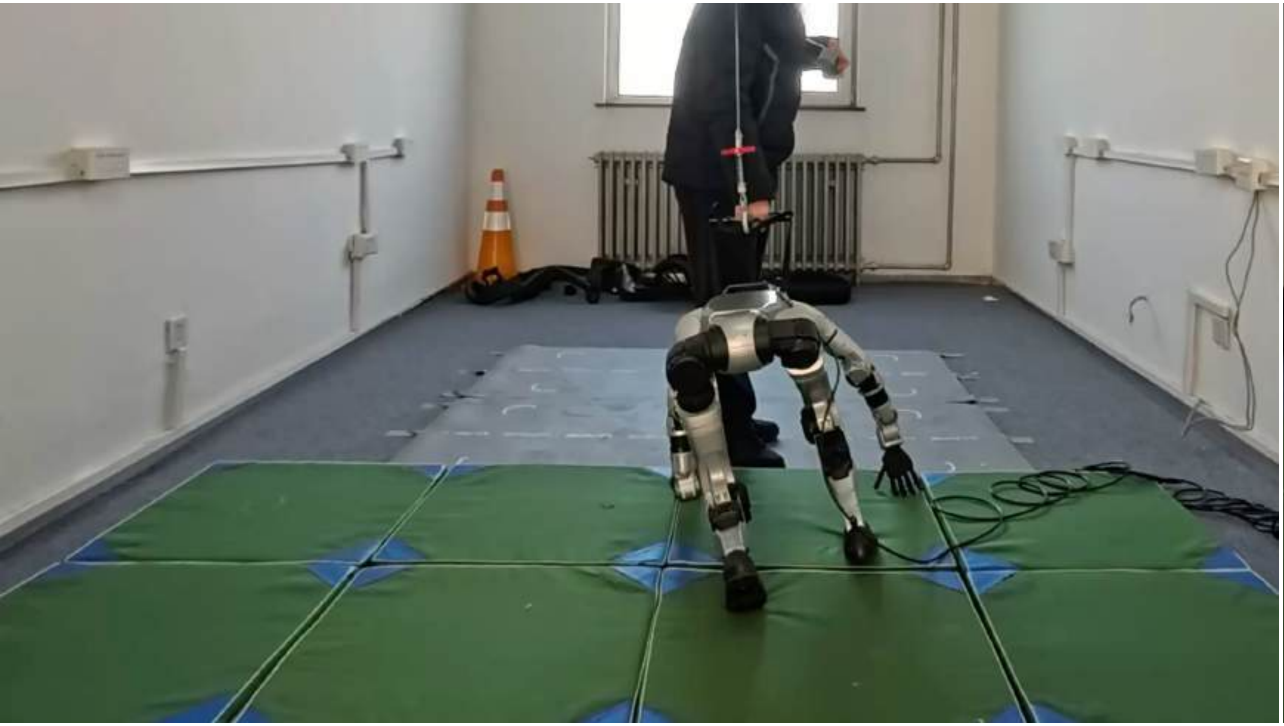
- ✓ AMASS/CMU/140
- ✓ 
- ✓ 
- ✓ 

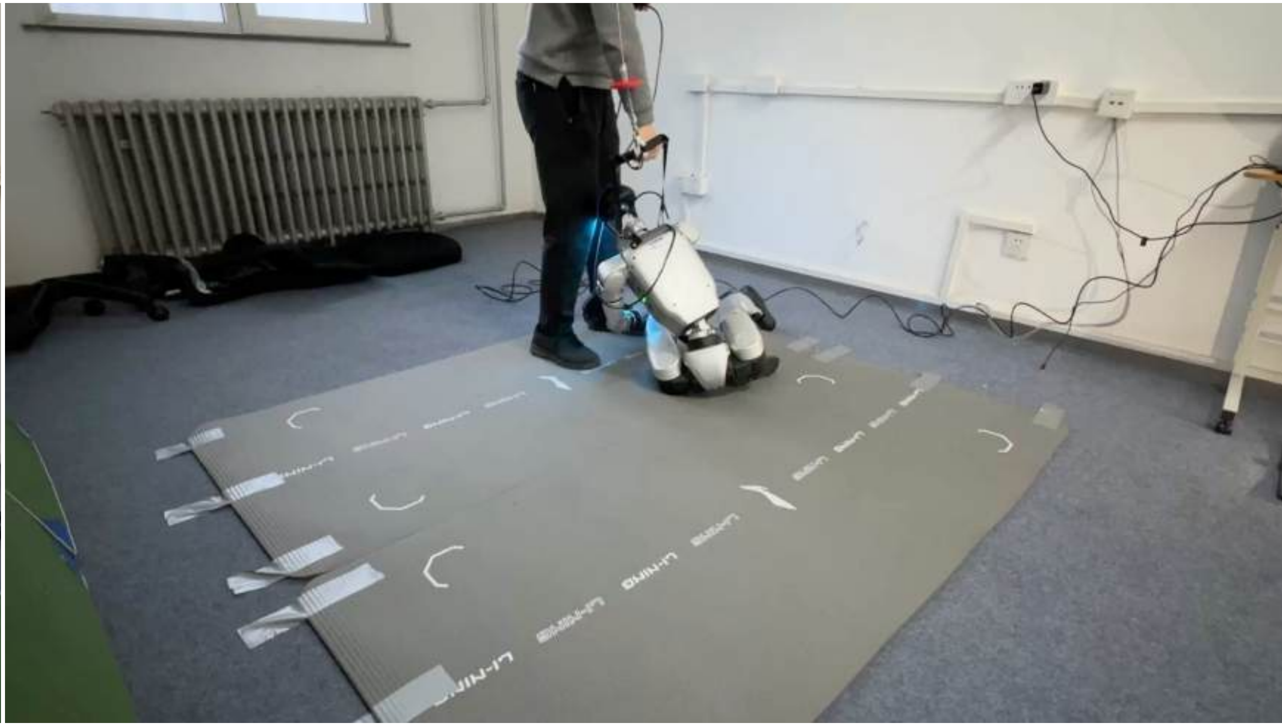
Onboard Deployment

- Laptop: sending high-level commands in at low frequency (with latency)
- G1 robot: running whole-body control policy on Jetson Orin

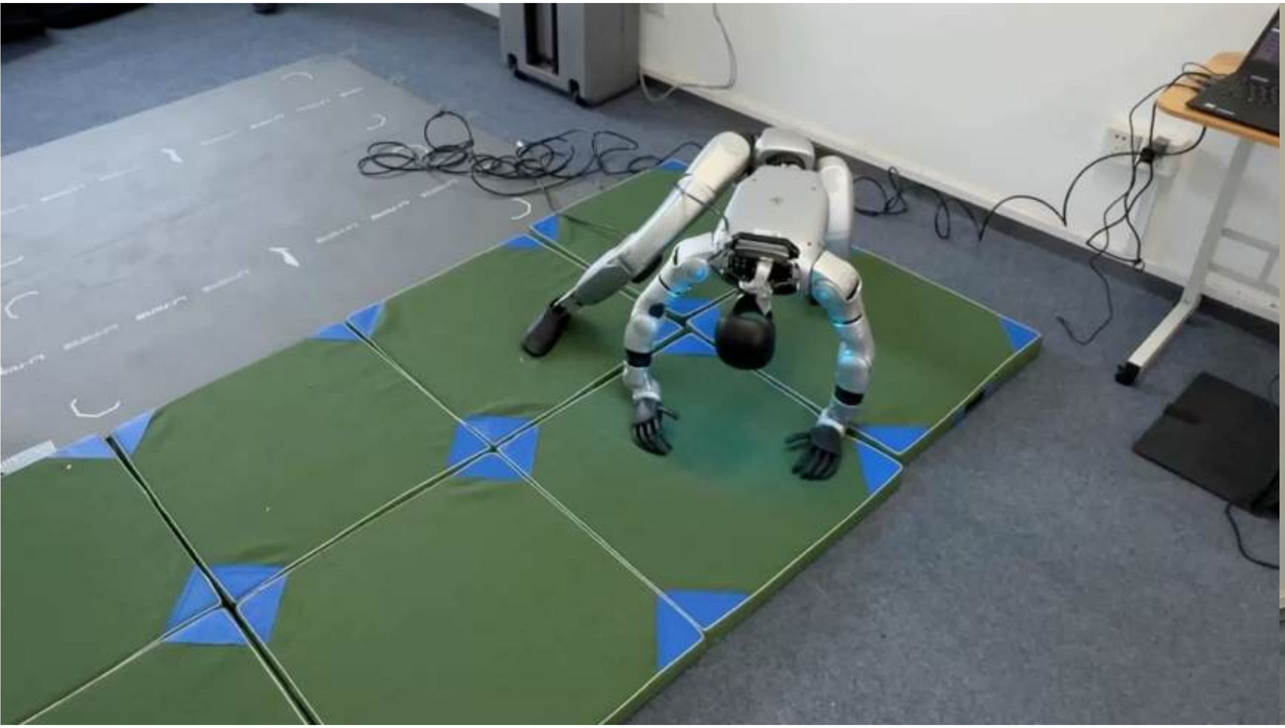


Ground Whole-body Motions





Ground Whole-body Motions: Jiu-jitsu



Getting Up

Remove the Perception Super Power → towards RGB Perception

RGB Only *Policy Training and Deployment*

VR-Robo

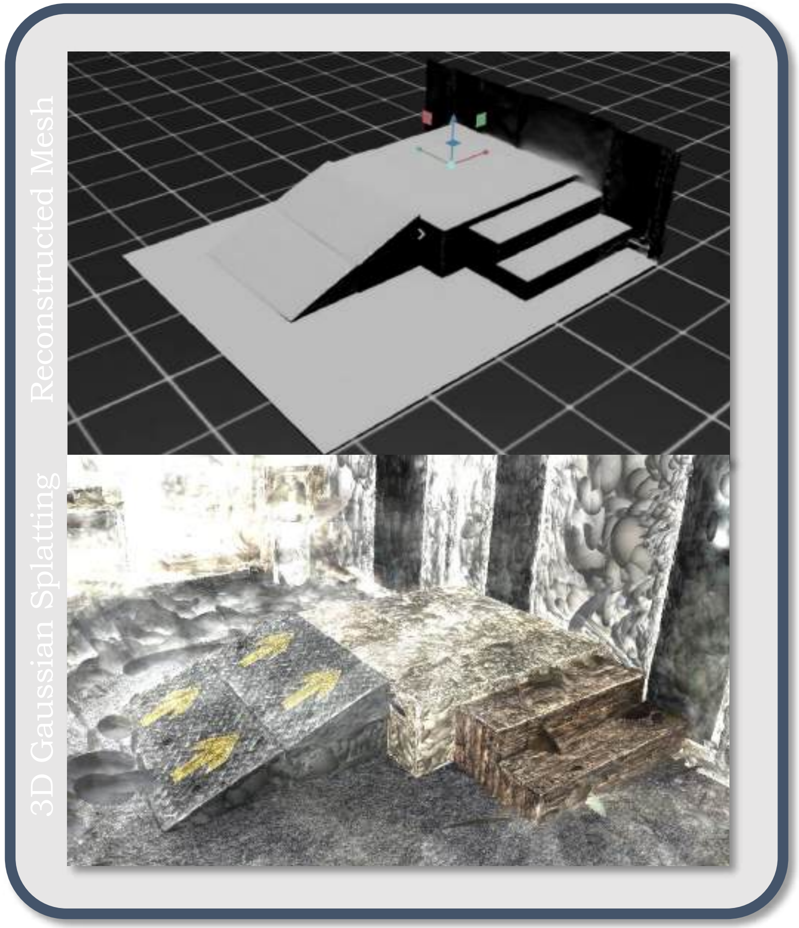
A ***Real-to-Sim-to-Real*** System for Robot Navigation and Locomotion

Shaoting Zhu, Linzhan Mou, Derun Li, Baijun Ye, Runhan Huang, Hang Zhao

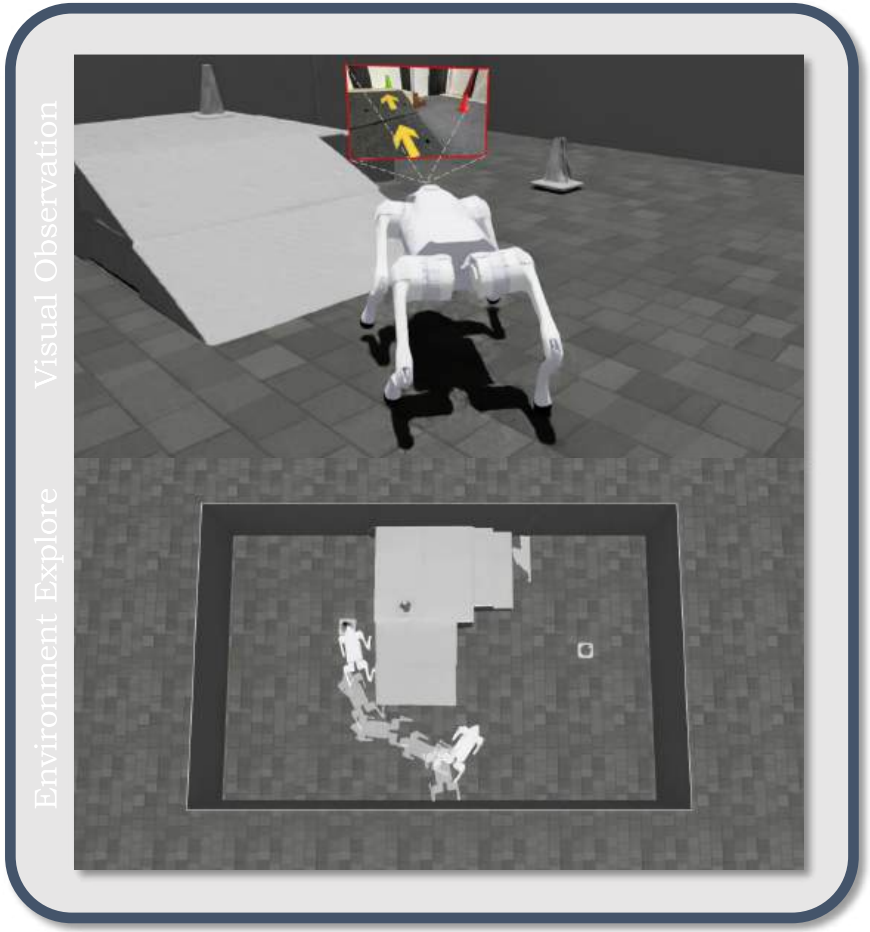
Real2Sim2Real

Real2Sim ⟶  ⟶ *Sim2Real*

Geometry-consistent Reconstruction

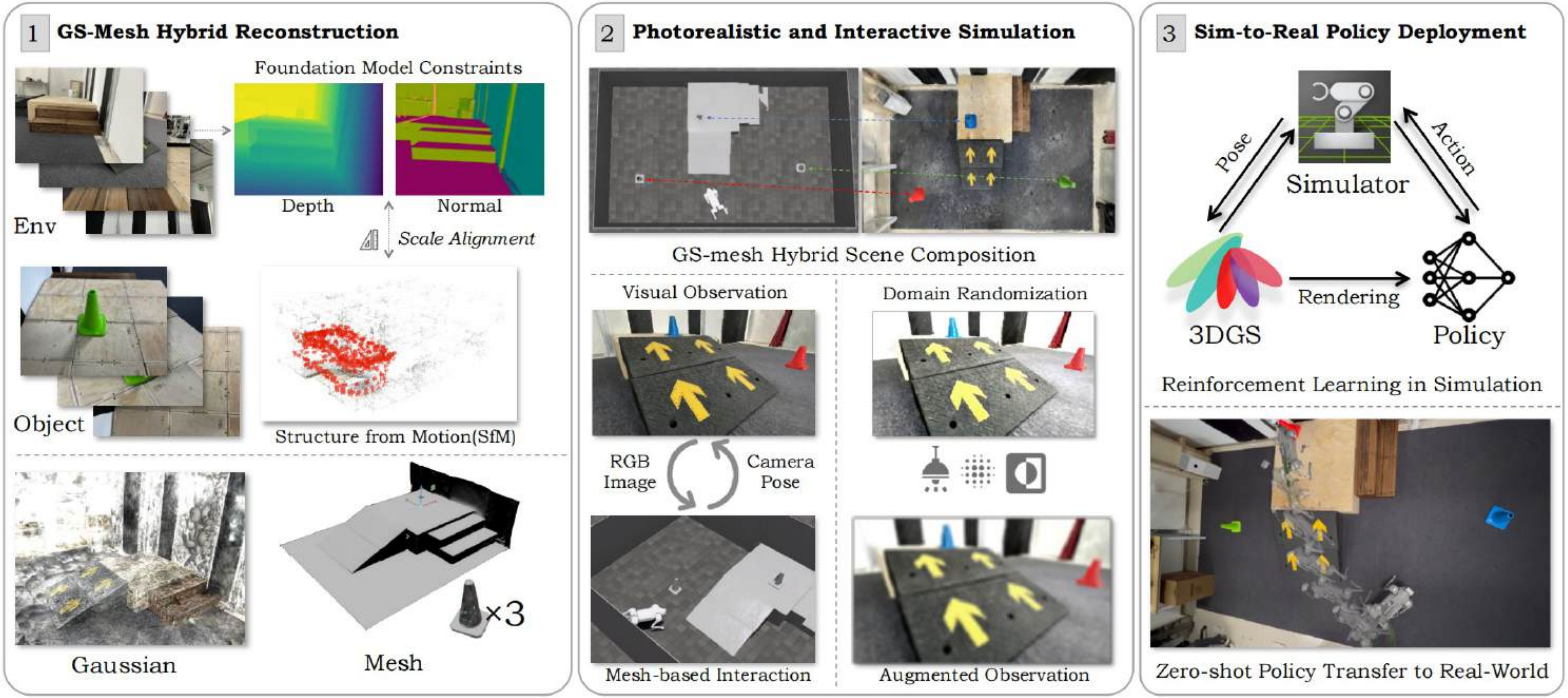


RL Policy Training in Simulation



Zero-shot Real-world Deployment

RL policy training in GS-Mesh Hybrid simulation



Demos

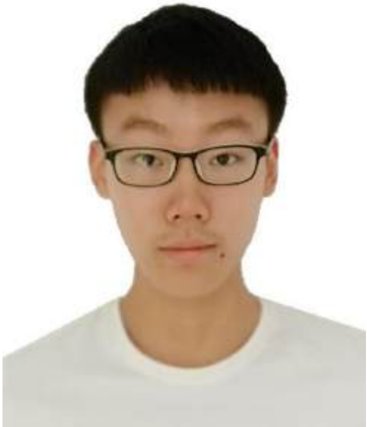

VR-Robo
A Real-to-Sim-to-Real Framework for Visual Robot Navigation and Locomotion
Shaoting Zhu, Linzhan Mou, Derun Li, Baijun Ye, Runhan Huang, Hang Zhao
清华大学 交叉信息研究院
GALAXEA
上海期智研究院
SHANGHAI QI ZHI INSTITUTE
上海交通大学

Takeaways

- LLMs/VLMs can effectively help self-driving cars and robots to understand and plan tasks.
- Perceptive locomotion with RL is a promising approach to bridge high-level planning and low-level control.

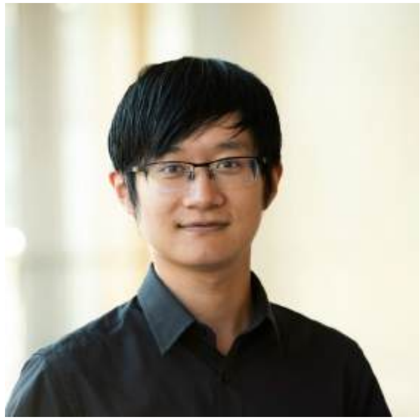
Contributors

- I sincerely thank our lab members for turning cool research ideas into reality.



Junru Gu

Yicheng Liu



Ziwen Zhuang

Shaoting Zhu

MARS Lab

THE END THANKS