Building General Embodied Agents in the Physical World

Baoxiong Jia
BIGAI

Figure generated by Nano Banana Pro

About me

buzz-beater.github.io

Peking University
B.S. in CS
2014-2018

UCLA
Ph.D. in CS
2018-2022

BIGAI
Research Scientist
2022-Present

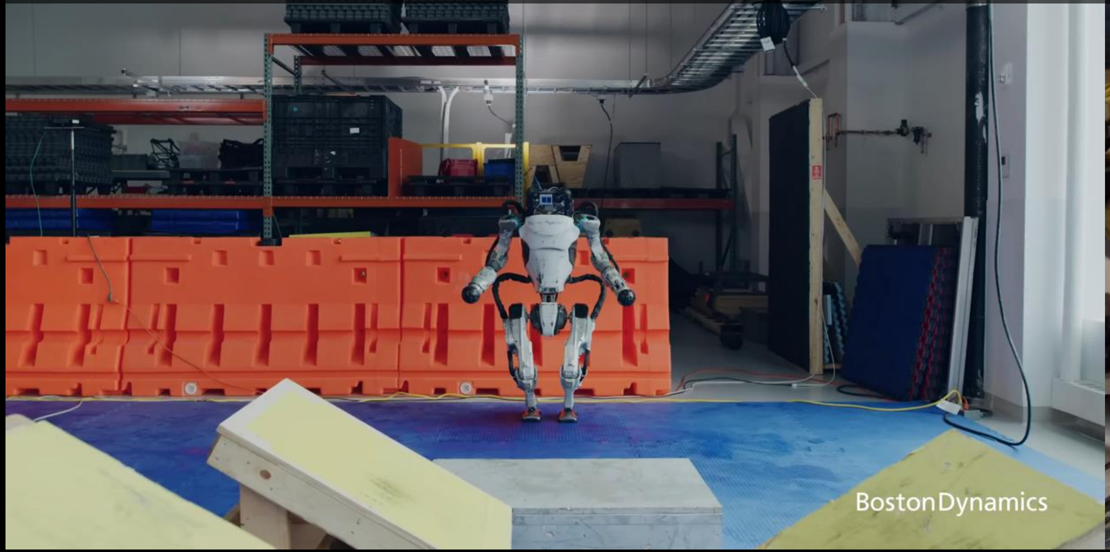

The 2025 IEEE/RSJ International Conference
on Intelligent Robots and Systems
Human-Robotics Frontier
WELCOME
@XRoboHub

Embodied AI

The embodiment hypothesis is the idea that intelligence emerges in the interaction of an agent with an environment and as a result of sensorimotor activity

Smith & Gasser, The Development of Embodied Cognition: Six Lessons from Babies, 2005

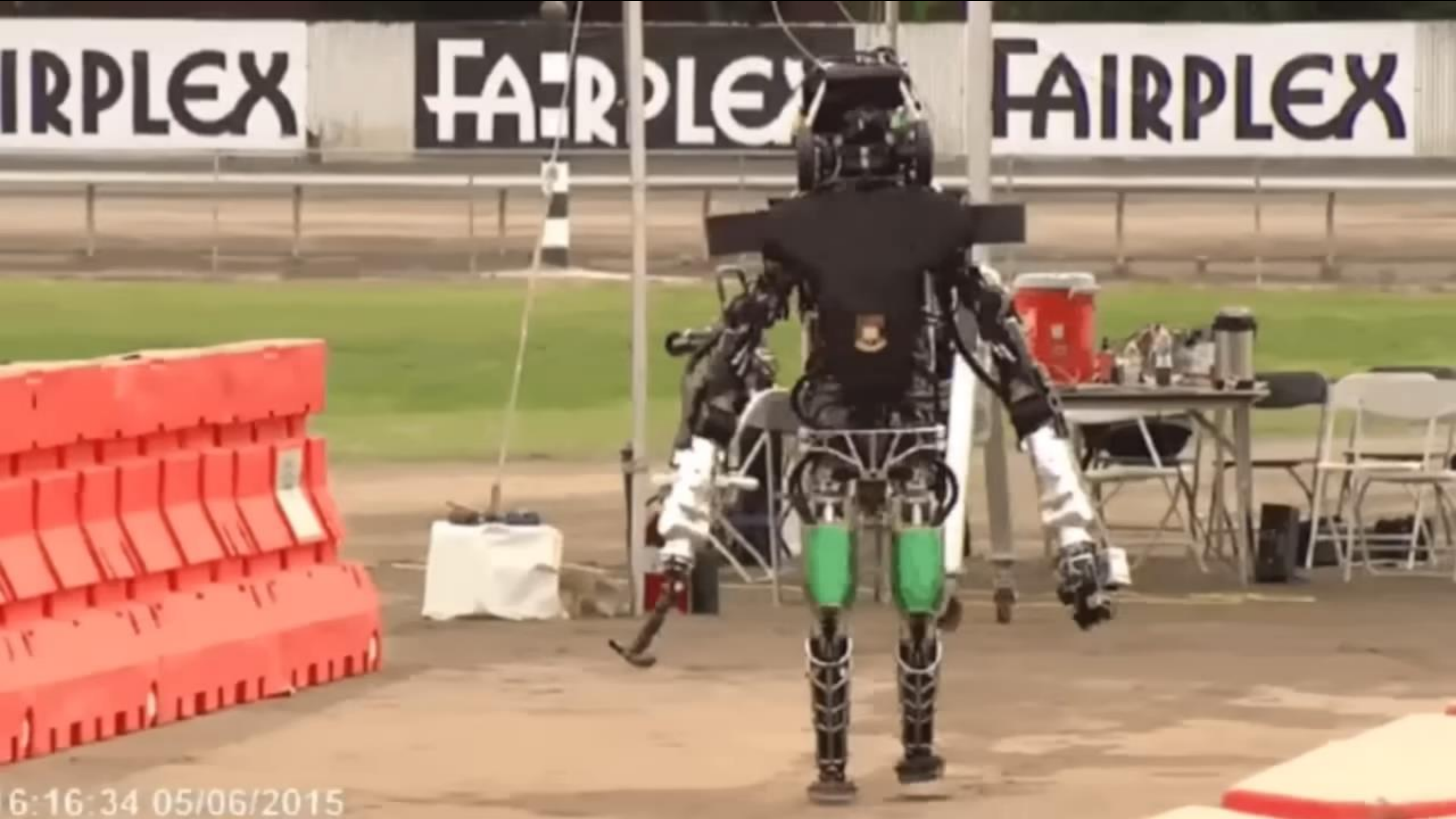
Hardware, control, and locomotion



Boston Dynamics, Atlas | Partners in Parkour, 2022
https://www.youtube.com/watch?v=tF4DML7FIWk

Interaction, reason, and plan

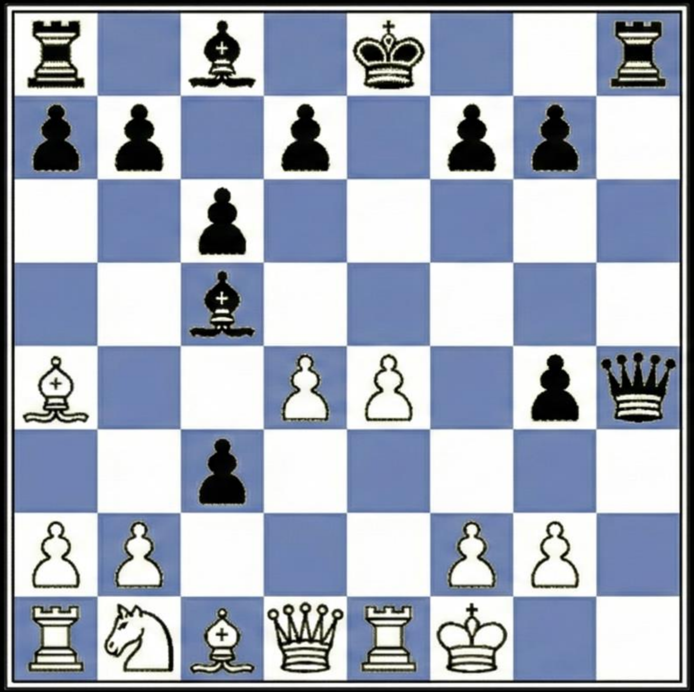
Damen et al., Scaling Egocentric Vision: The Epic-Kitchens Dataset, 2018

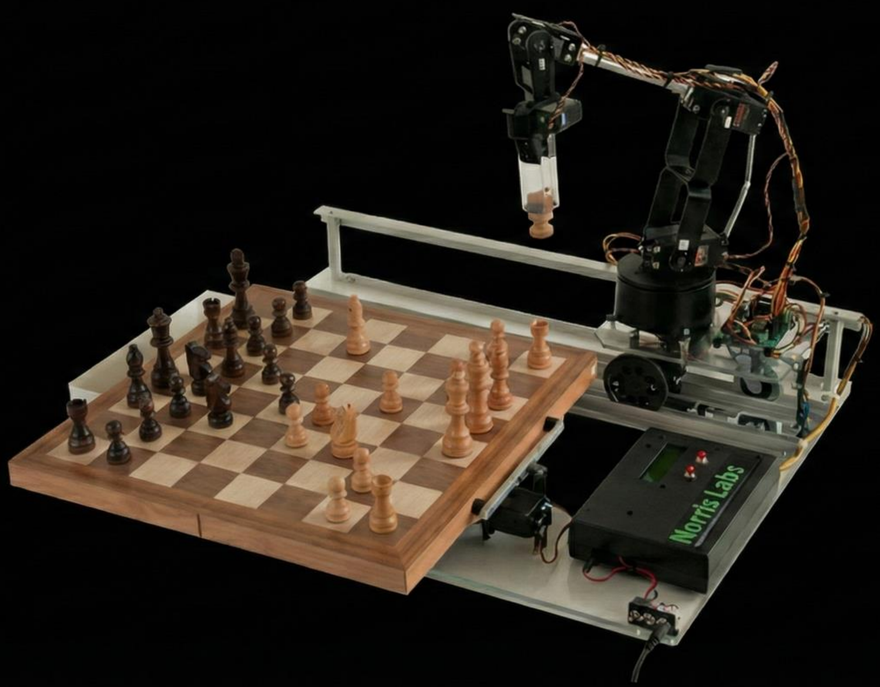

IRPLEX
FAIRPLEX
FAIRPLEX
6:16:34 05/06/2015

Moravec's Paradox

It's comparatively easy to make computers exhibit adult level performance on intelligence tests or playing checkers, and difficult or impossible to give them skills of a one-year-old when it comes to perception and mobility.

Hans Moravec, Mind Children, Harvard University Press 1988





UniTree Kungfu Kid 6.0, UniTree 2025


Walk, Run, Crawl, RL Fun, Boston Dynamics 2024


Playing Badminton against Humans, ETH 2025

UniTree B2-W Talent Awakening, UniTree 2024


GPT/Qwen
1.2B Hours
π_0
10k Hours
autonomous
GR00T
88 Hours

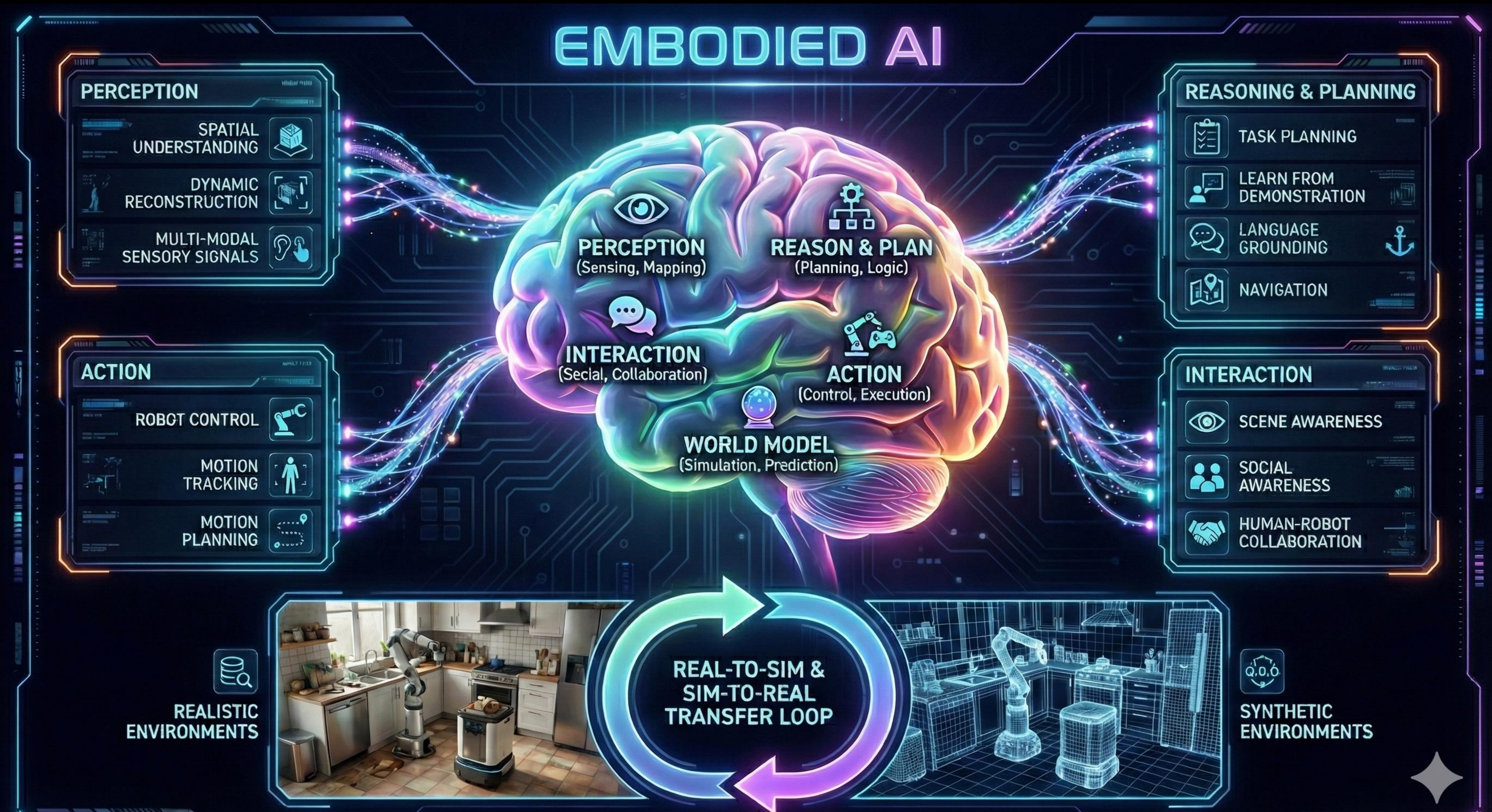


Figure generated by Nano Banana Pro

1. Building Interactable 3D Scenes for Embodied AI

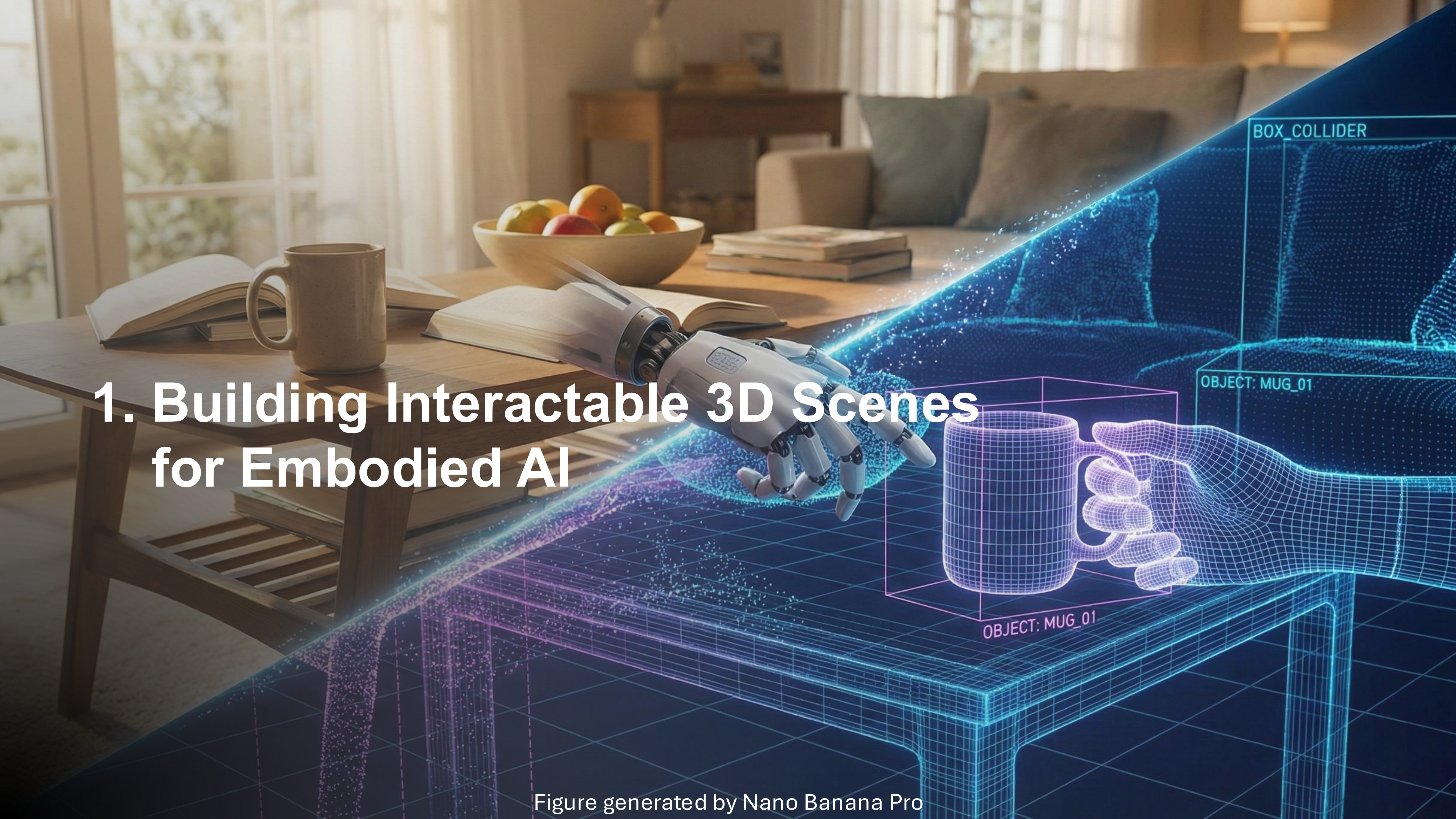
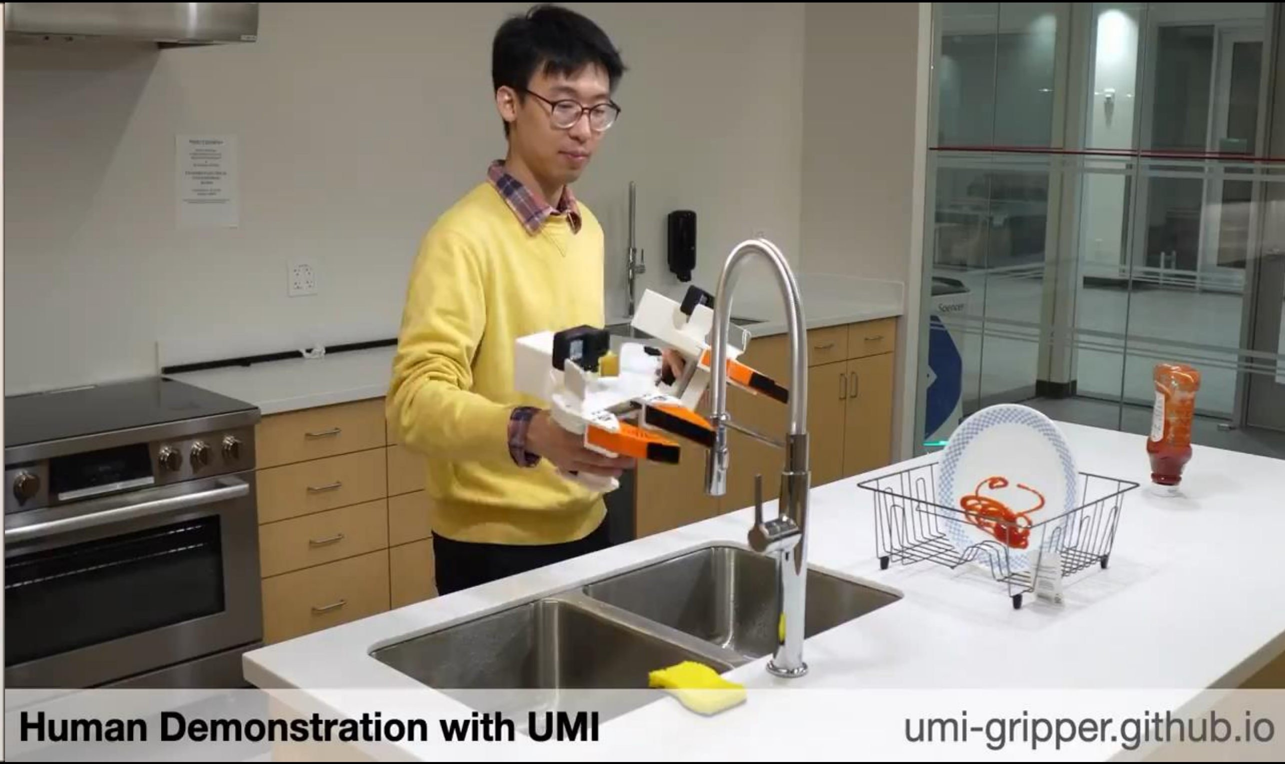


Figure generated by Nano Banana Pro

Collecting Data and Training Robot in the Real World is Expensive



Damage the environment and objects

Low-cost hardware emerging but still low efficiency

Goal of Environment Creation

- High-quality appearance understanding for grounding and reasoning
- Fine-grained geometry understanding for simulation and physics
- Solid dynamics understanding for interaction and planning

Digital Cousin Creation

MetaScenes (CVPR’25)

eractable Replicas
culated Objects
an Splatting

Dynamic Reconstruction

ArtGS (ICLR’25)

3D Scene Generation

SceneWeaver (NeurIPS’25)

Best Paper, RoboGen@IROS25

ICLR
International Conference On
Learning Representations

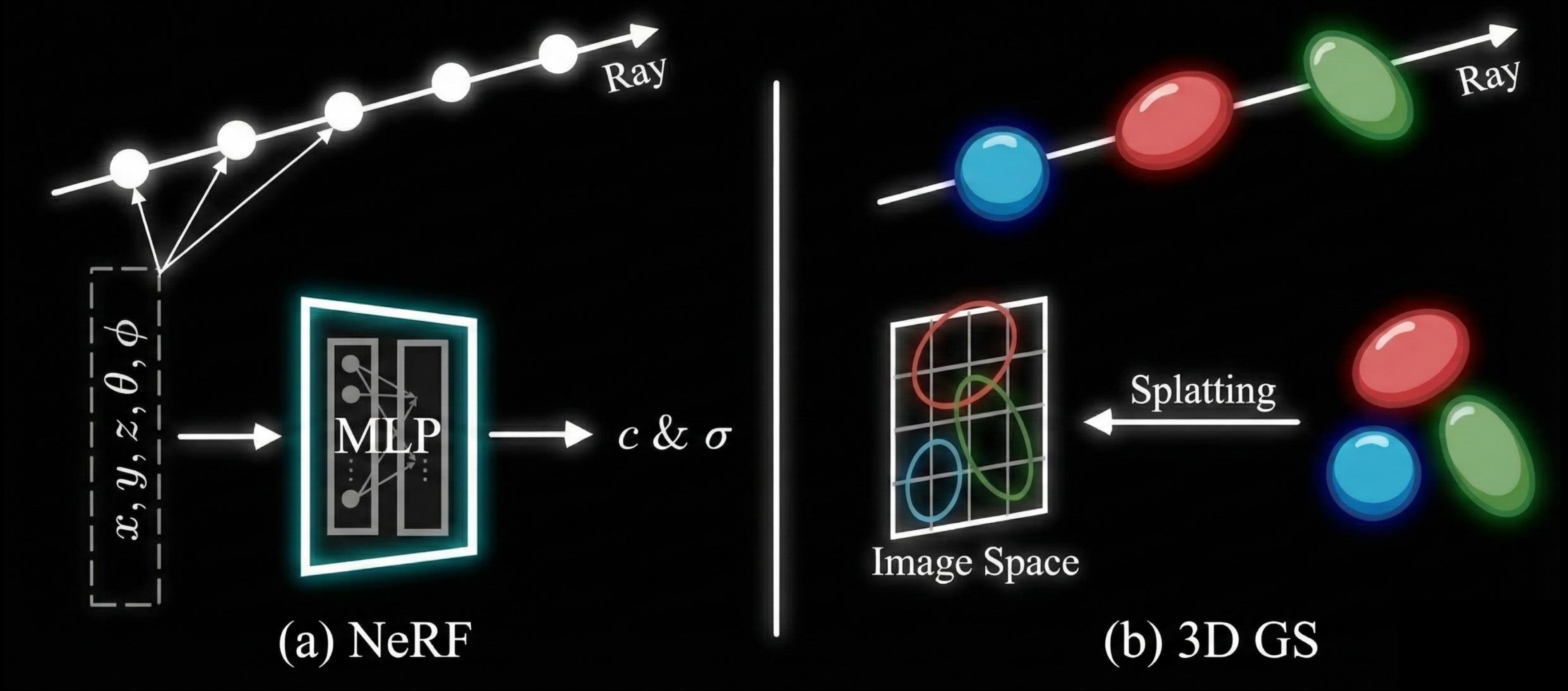
4D Gaussians for Dynamic Reconstruction

- *(ICLR'25) Building Interactable Replicas of Complex Articulated Objects via Gaussian Splatting*
- *(ArXiv'25) VideoArtGS: Building Digital Twins of Articulated Objects from Monocular Video*

ArtGS

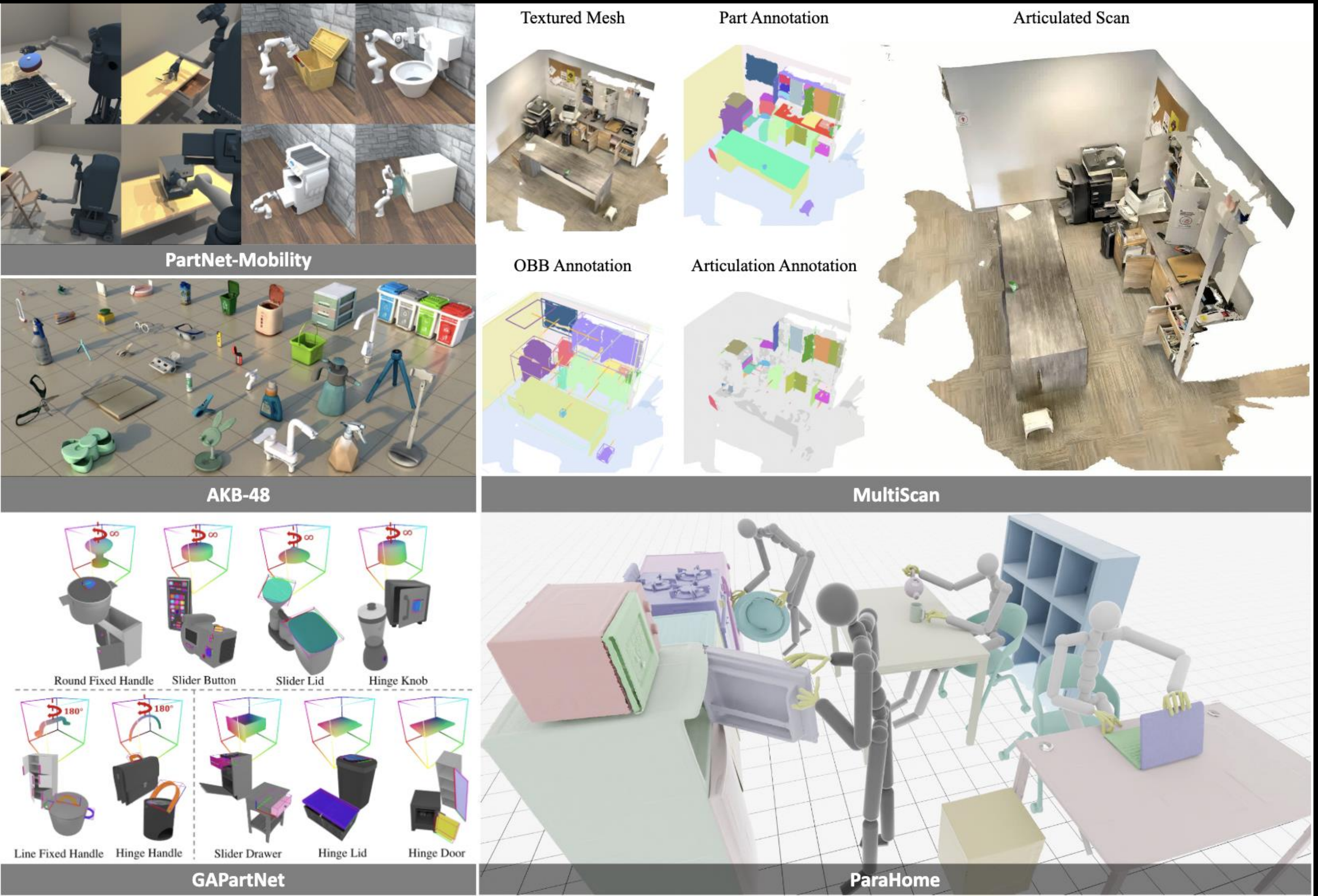
VideoArtGS

Static 3D Scene Reconstruction



Learn 3D by projecting to multiple views as supervision

Manipulation Involves Dynamic Objects



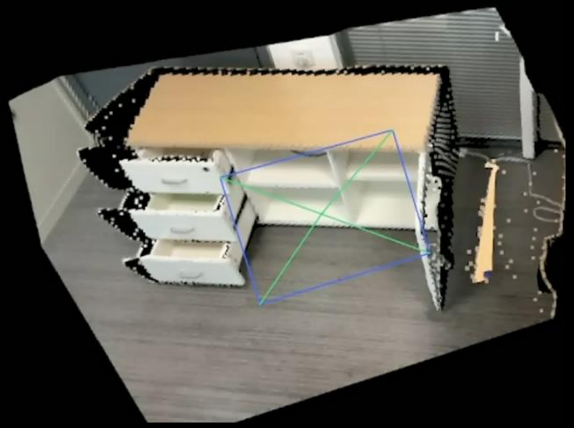
In reality, we deal with dynamic, articulated objects whose geometry and shape change during interaction, making them difficult to reconstruct

Image Supervision is Ambiguous for Articulation Learning

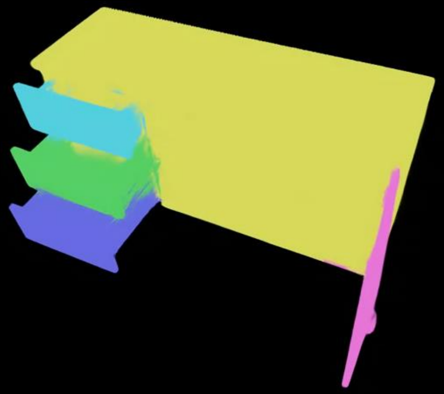
Key Challenge: The observed pixel motion results from four entangled factors:

Camera trajectory

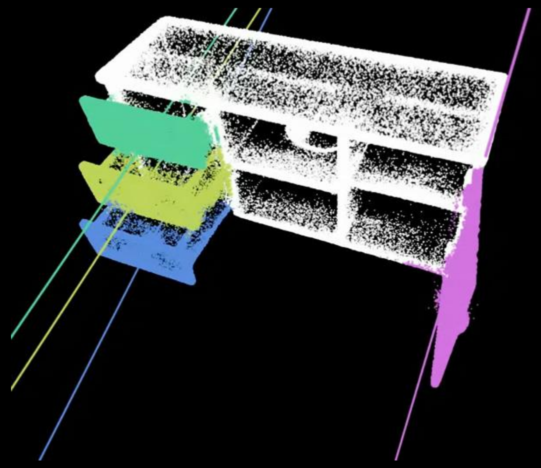
Object Geometry



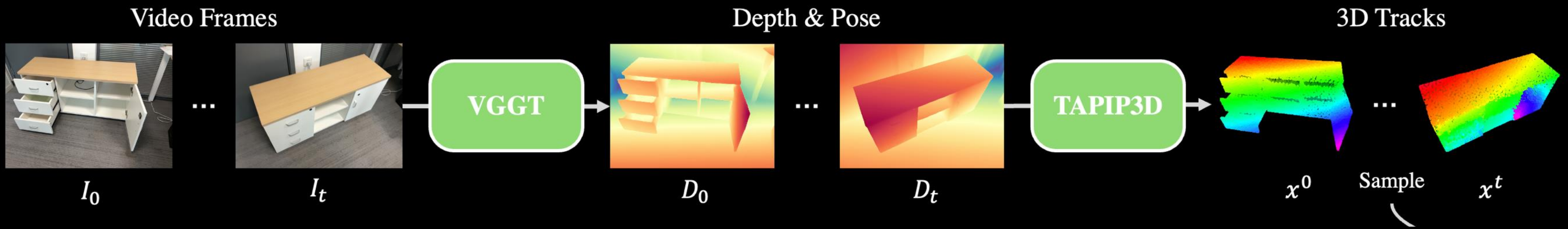
Part Segmentation



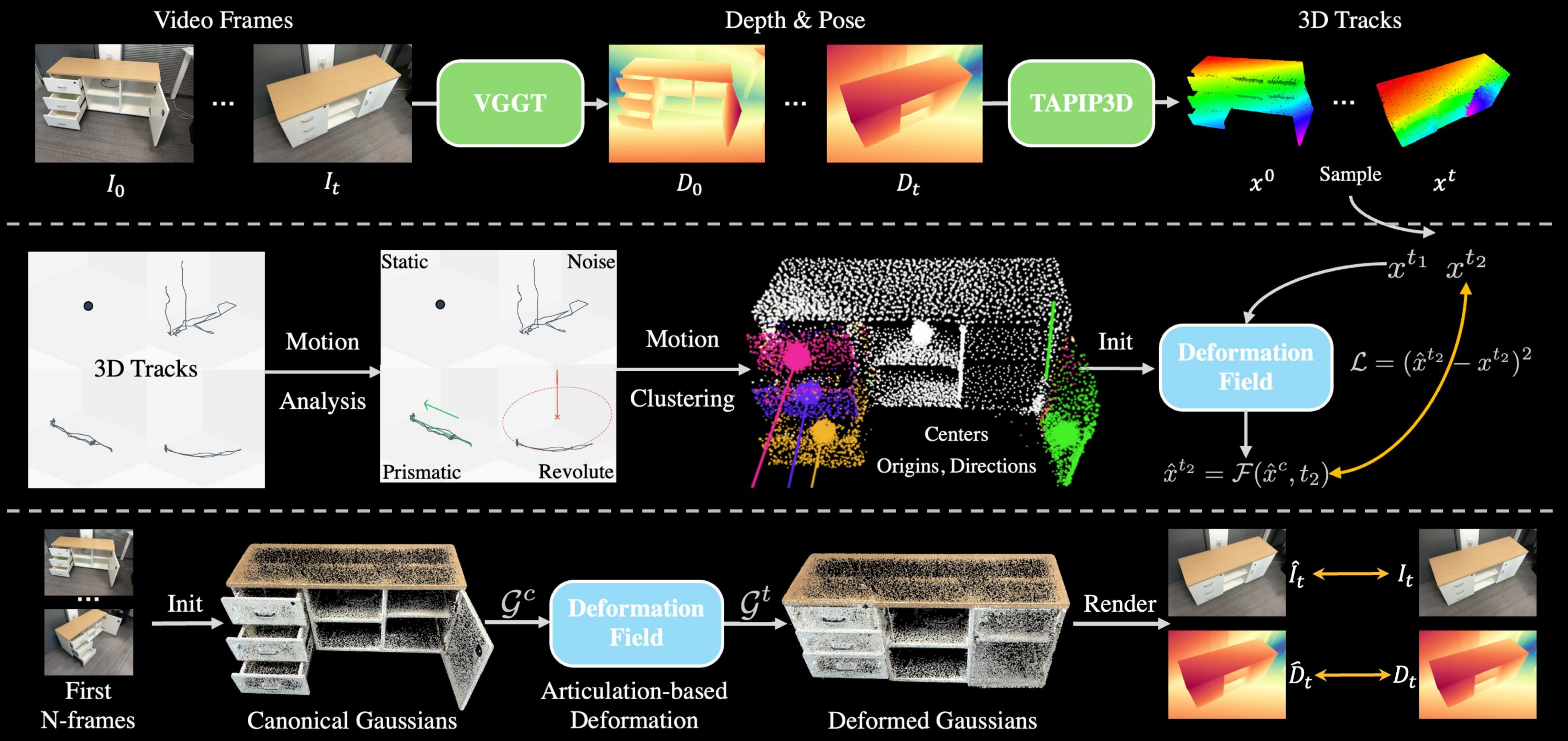
Articulation Dynamics



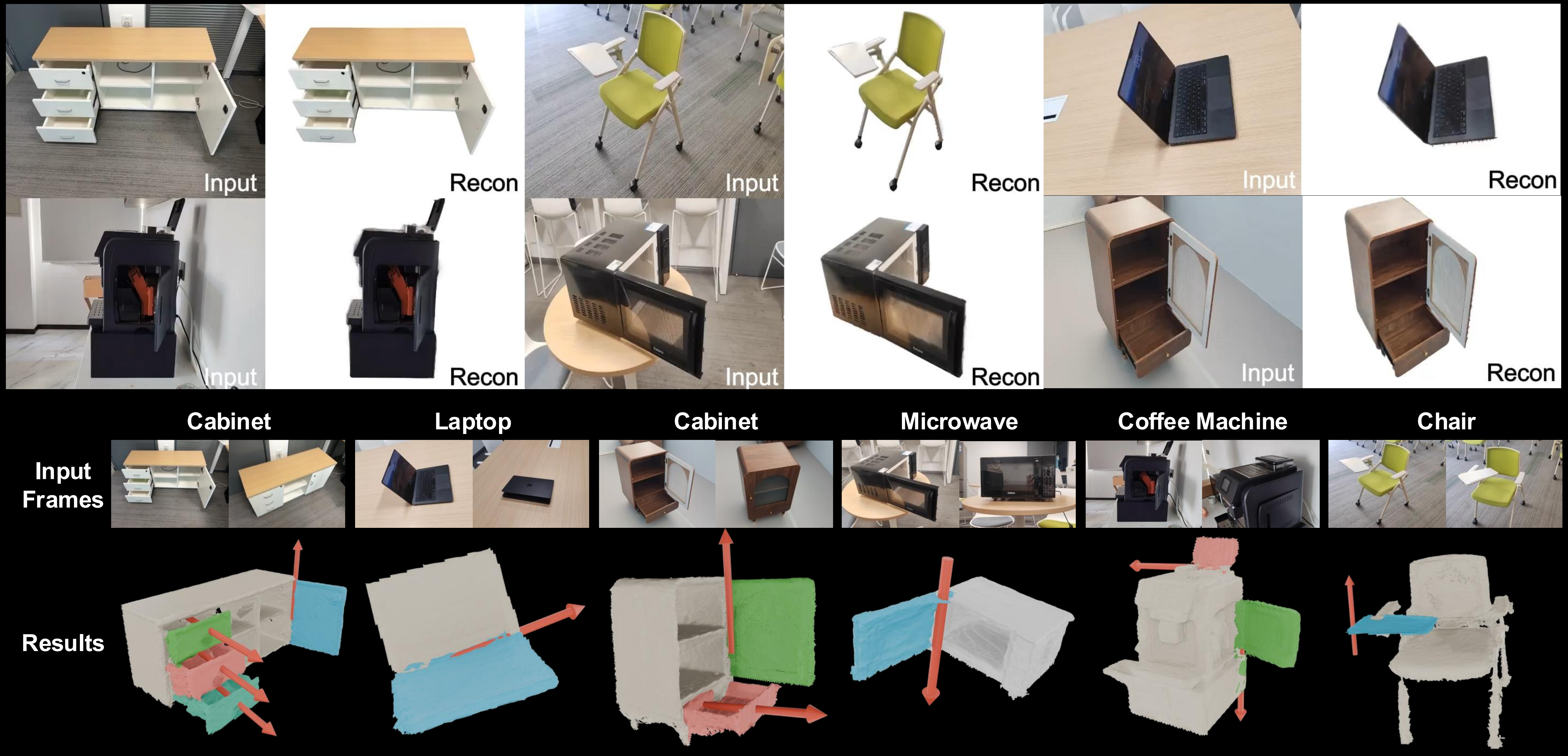
Camera, Depth, Tracks Estimation



Geometry Reconstruction & Articulation Learning



Real-world Experiments




Data Capture
Input
Recon
Gaussian

Limitations & Takeaways

- Difficult to scale-up
  - Feed-forward reconstruction
  - Active camera trajectory selection

- Limited quality for simulation
  - Physical priors (e.g. plane) during reconstruction
  - Existing assets and generative models as guidance for reconstruction
  - System vs. Model?

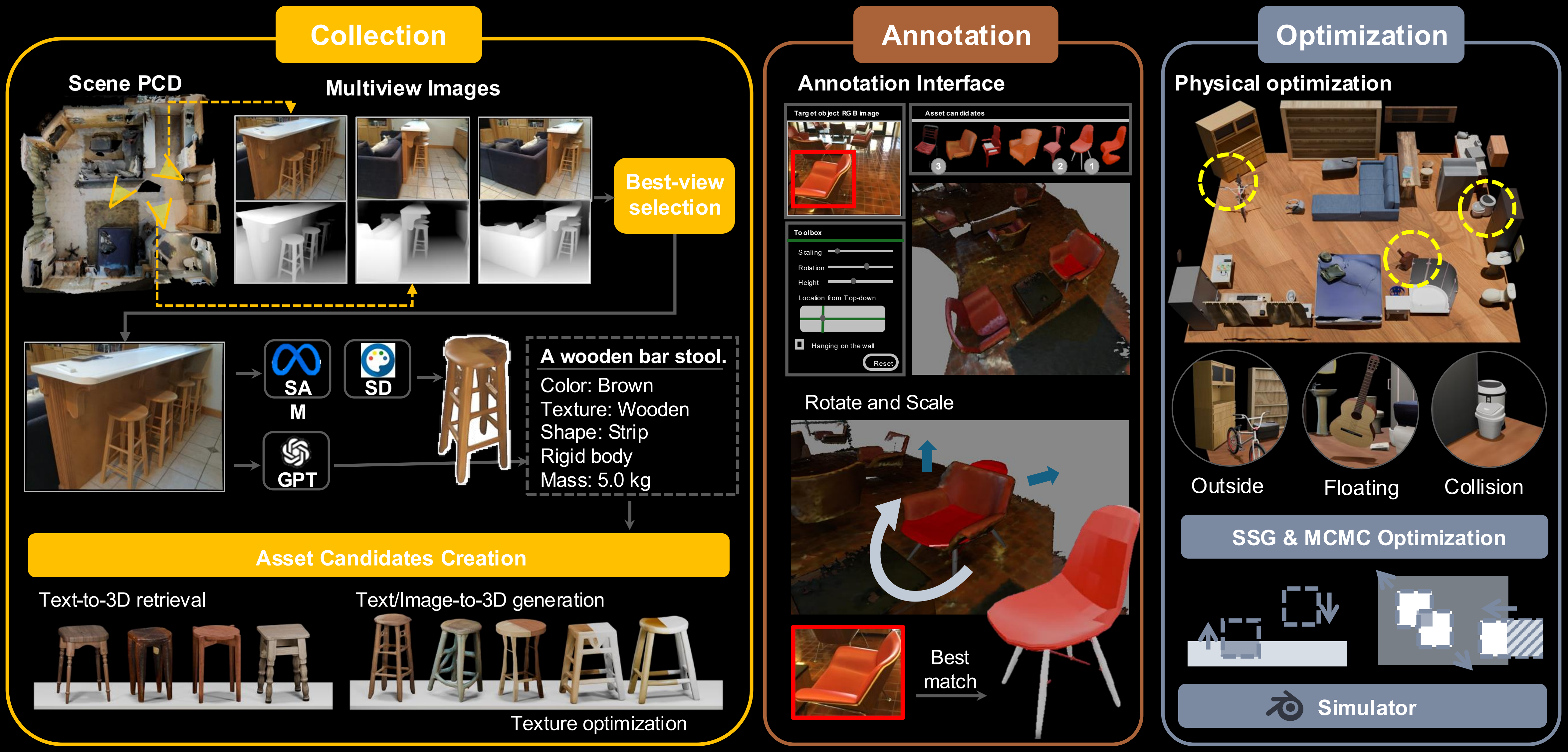
ArtGS: Building Interactable Replicas of Complex Articulated Objects via Gaussian Splatting

Digital Cousin Creation with 3D AIGC

- *(CVPR'25) MetaScenes: Towards Automated Replica Creation for Real-World 3D Scans*

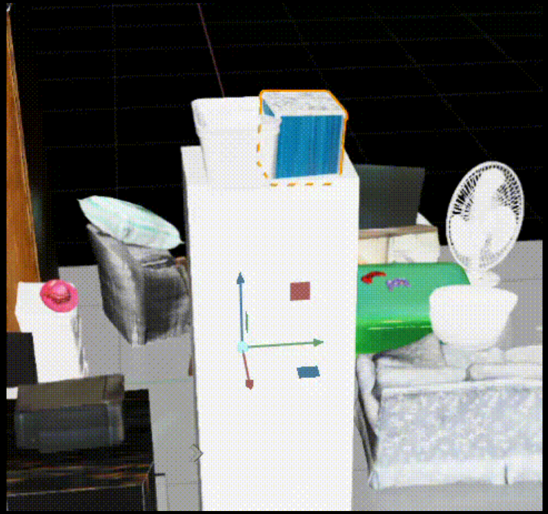
MetaScenes

Leveraging AIGC & Online Assets for Digital Twin Creation

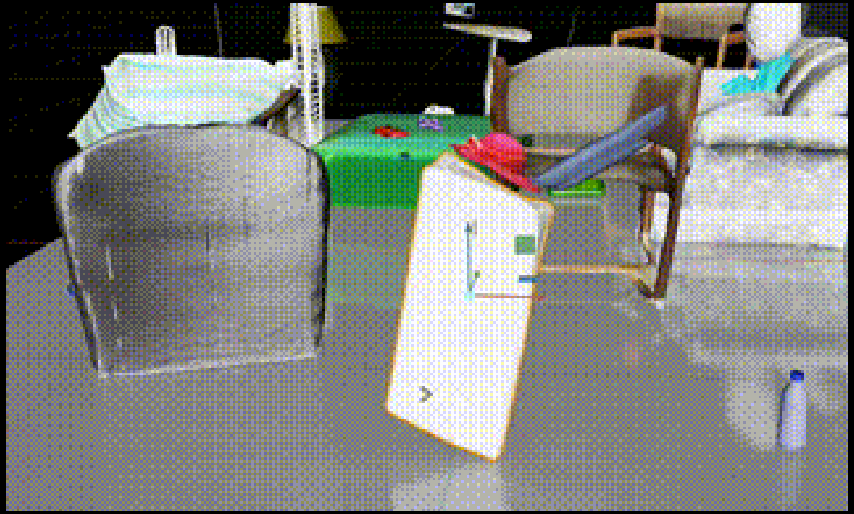


Limitations & Takeaways

- Unstable generation quality
  - Physical-based post-optimization
  - Better 3D generative models (e.g. SAM3D)



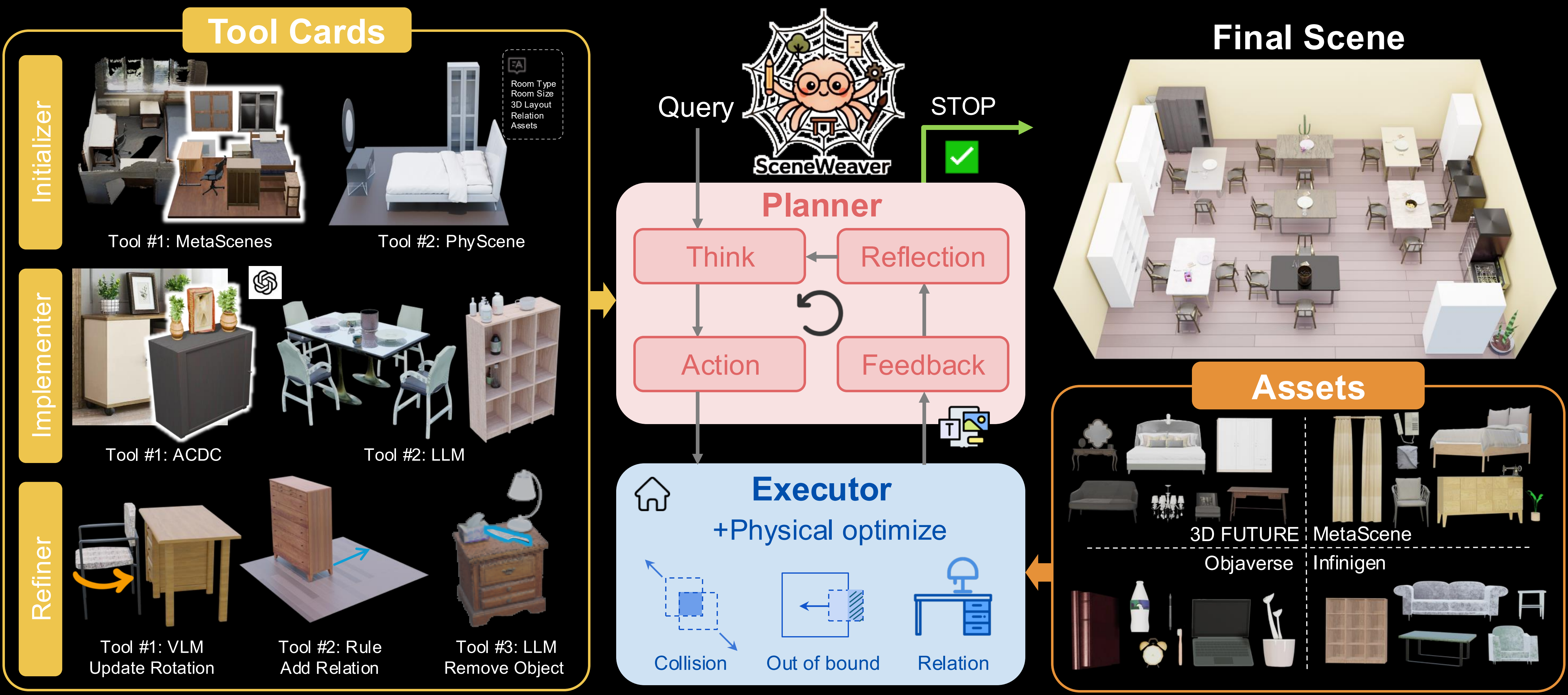
- Sequential error in the pipeline
  - Better orchestration of tools
  - Iterative refinement of the generated results?



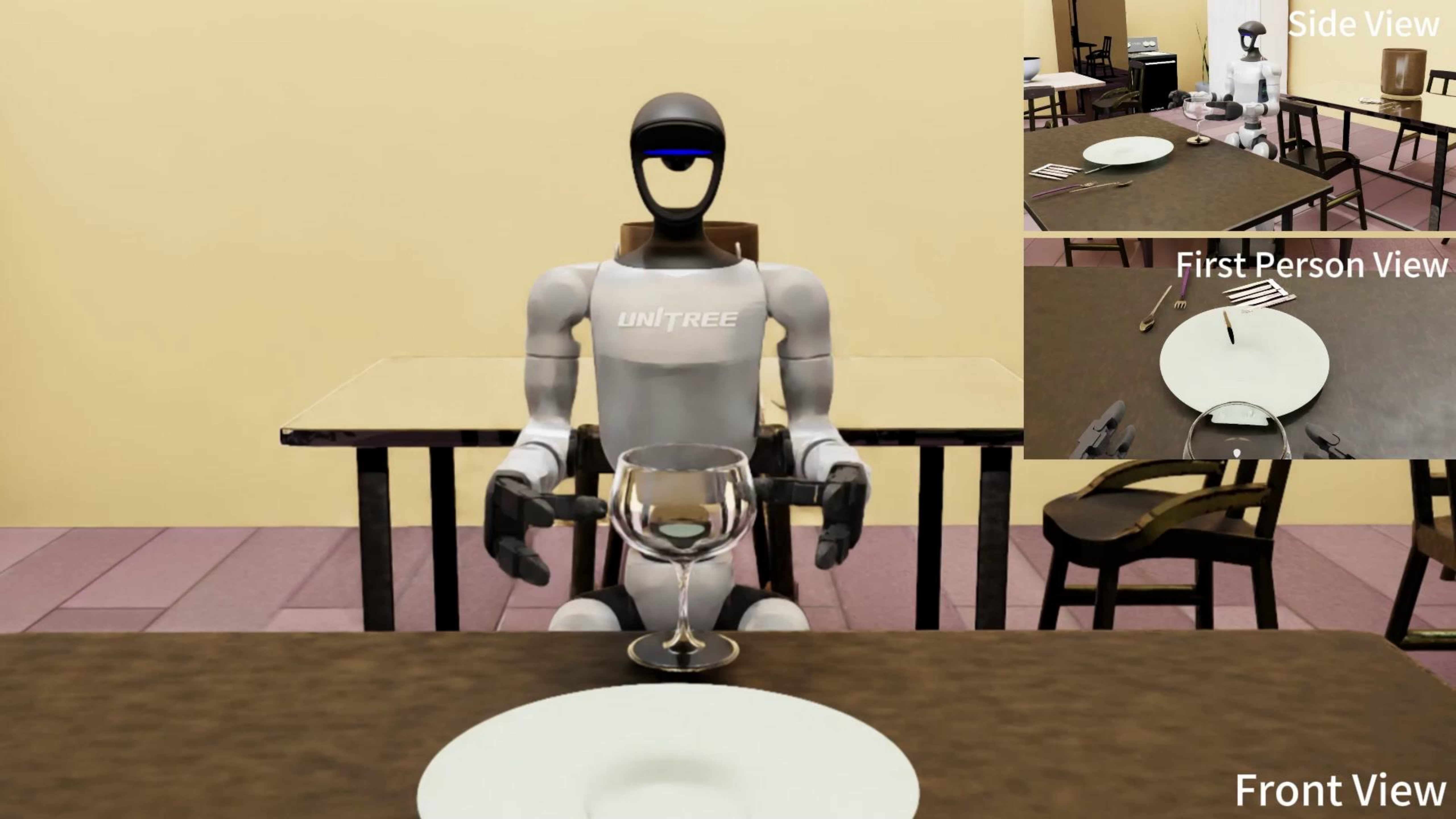
Agentic Tool-Use for 3D Scene Generation

- *(NeurIPS'25) SceneWeaver: All-in-One 3D synthesi with an Extensible and Self-Reflective Agent (Best Paper, RoboGen@IROS'25)*

SceneWeaver

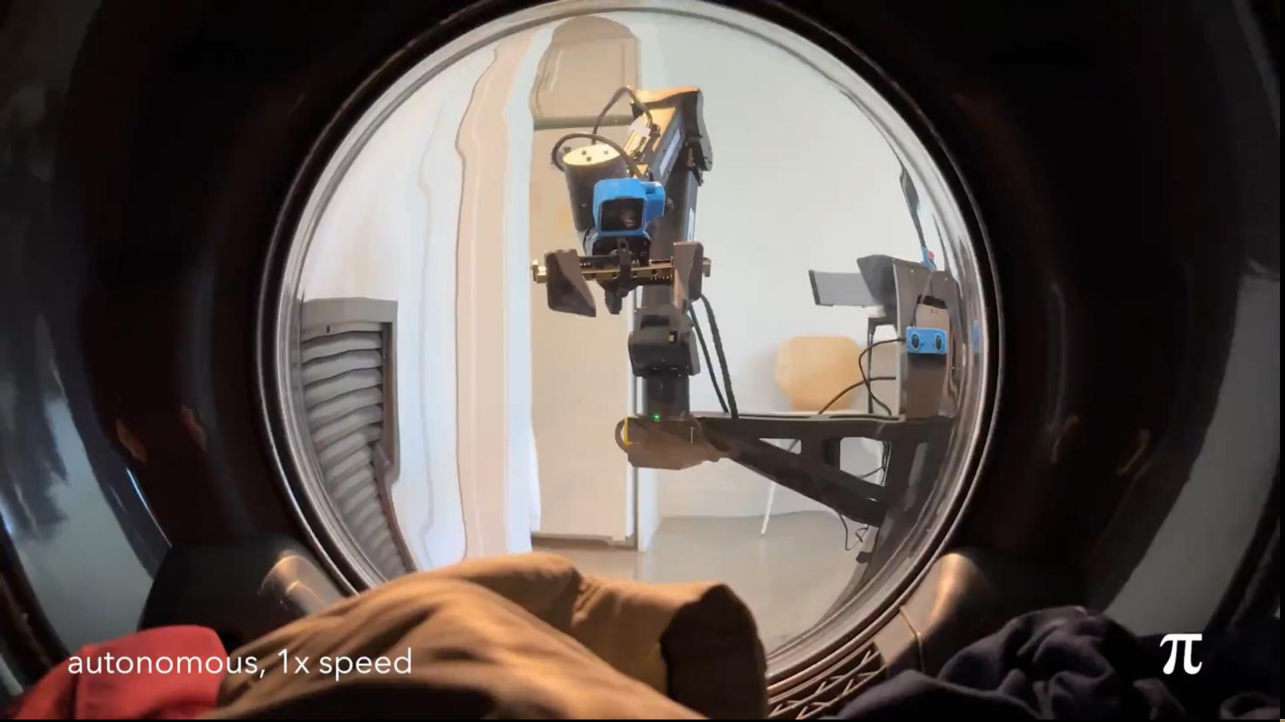


Tool Cards
Initializer
Room Type
Room Size
3D Layout
Relation
Assets
Tool #1: MetaScenes
Tool #2: PhyScene
Implementer
Tool #1: ACDC
Tool #2: LLM
Refiner
Tool #1: VLM
Update Rotation
Tool #2: Rule
Add Relation
Tool #3: LLM
Remove Object
Query
SceneWeaver
STOP
Planner
Think
Reflection
Action
Feedback
Executor
+Physical optimize
Collision
Out of bound
Relation
Final Scene
Assets
3D FUTURE
MetaScene
Objaverse
Infinigen


Side View
First Person View
UNITREE
Front View

2. Bridging Reasoning and Action for General Embodied AI Agents

Figure generated by Nano Banana Pro

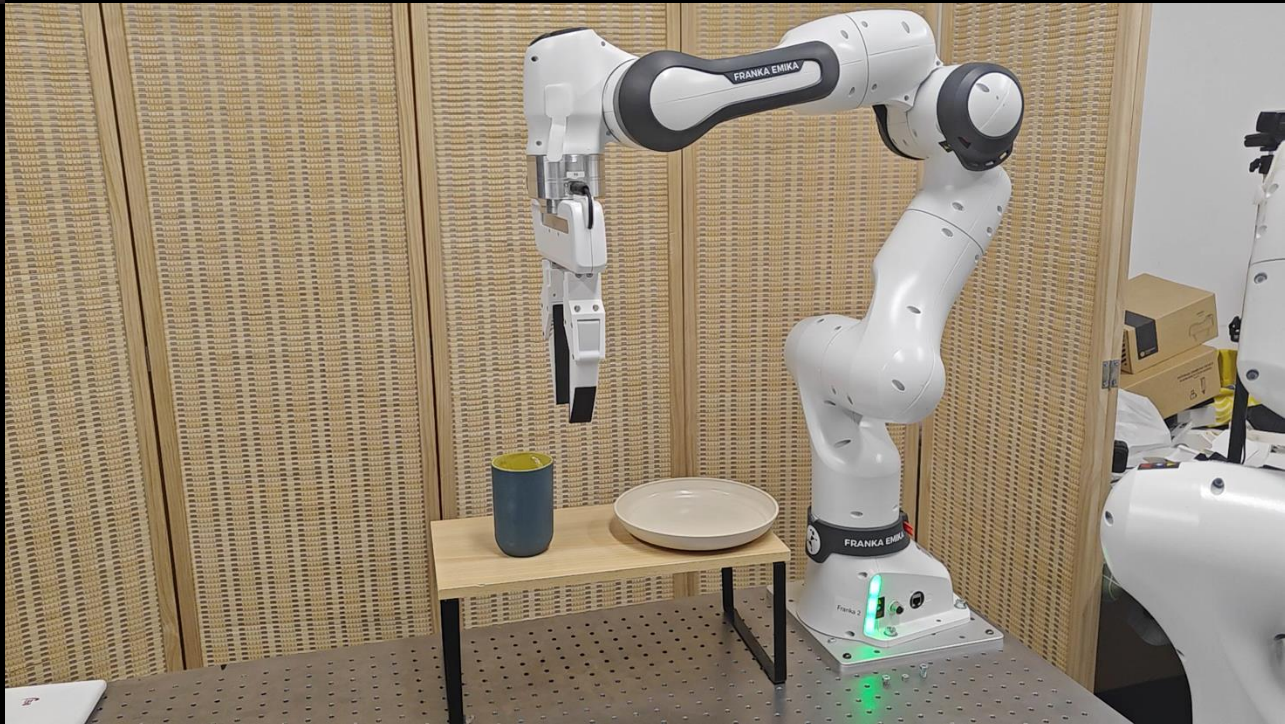


Training Generalist Policies

- Leveraging large-scale pre-trained VLMs
- Pre-trained with large-scale data
- Still limited generalizability on tasks and embodiments

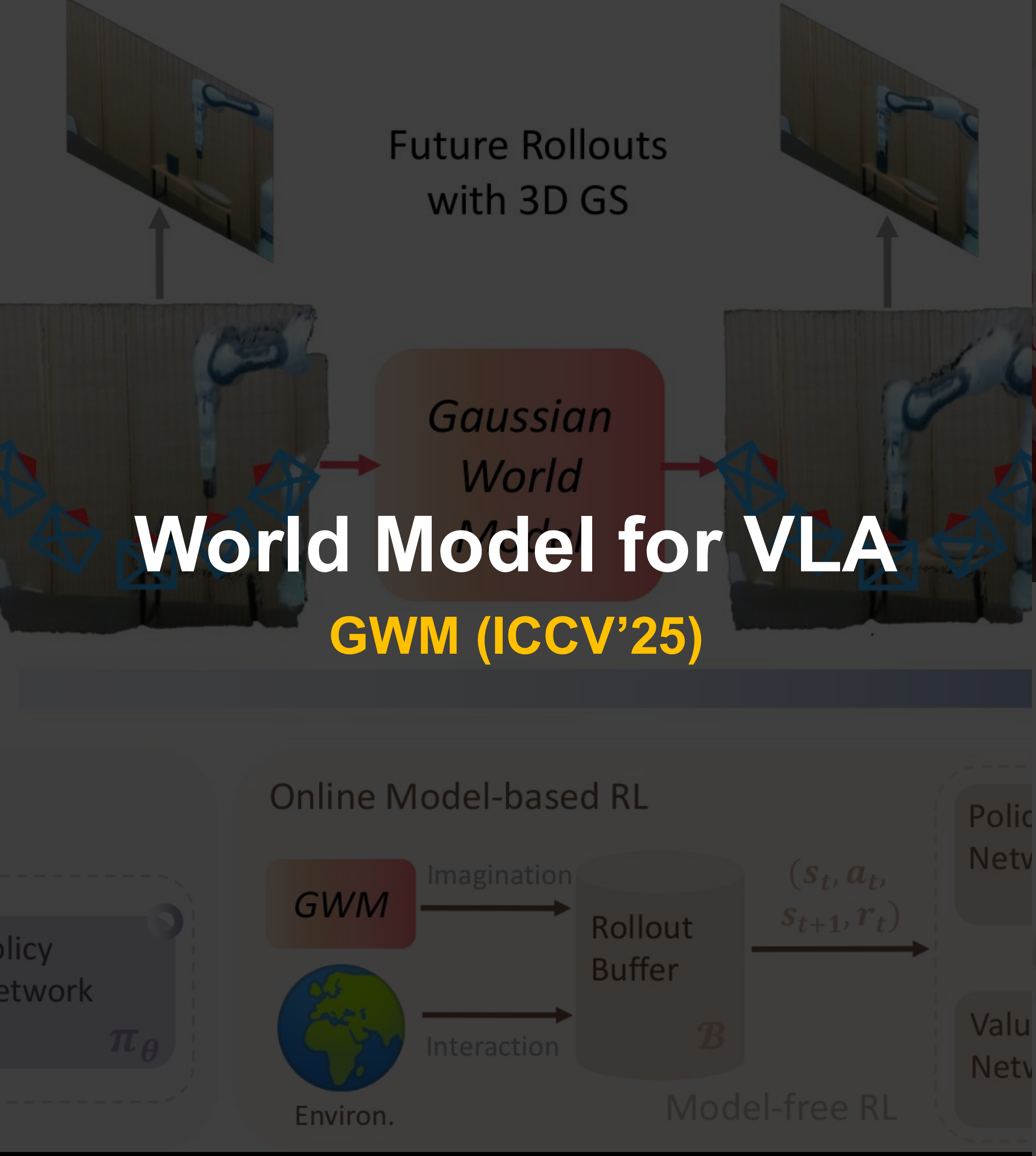
Adapting to Your Specific Scene

- Rolling out “almost” successful trajectories but hard to improve
- Can only afford few-shot demonstrations
- Sensitive to capturing modalities and viewpoints



Goal of Reasoning and Planning

- Capable of finishing diverse tasks according to instructions
- Enable spatial understanding for existing VLMs for better backbones
- Efficient representation for effective learning from high-cost data

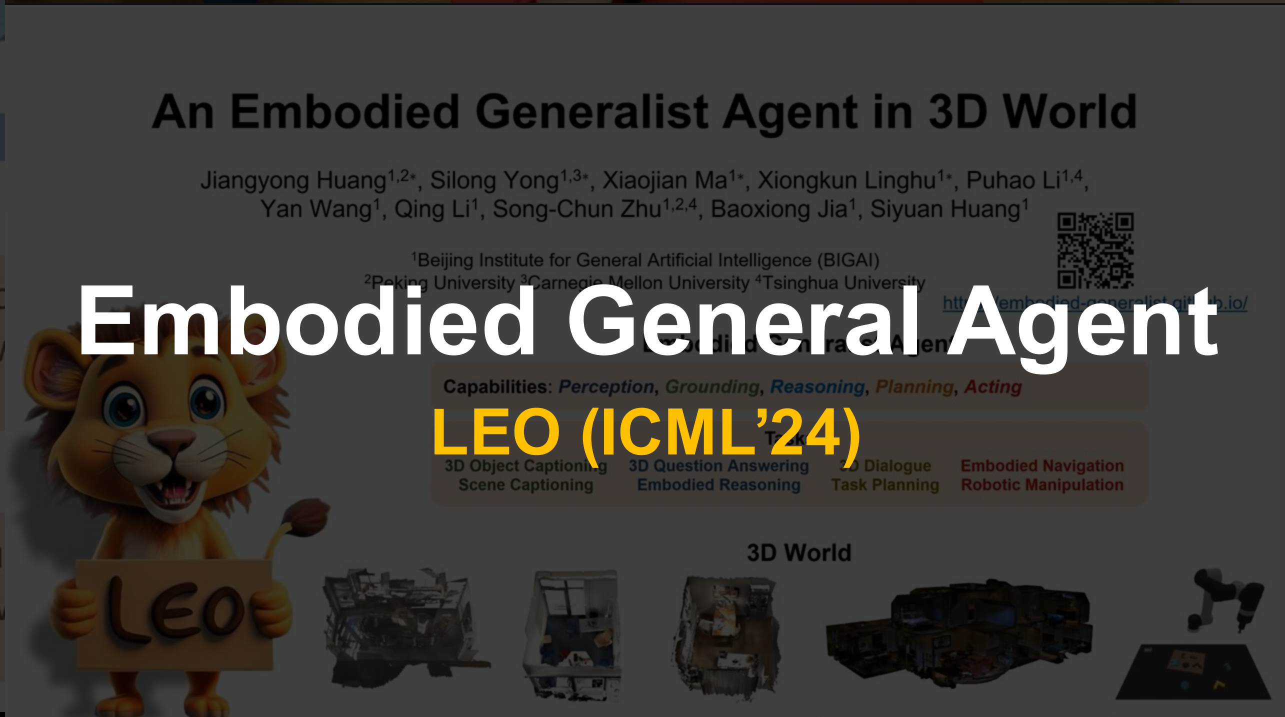


World Model for VLA

GWM (ICCV'25)

Open-Vocabulary Planning

COME-Robot (ICRA'25)



Embodied General Agent

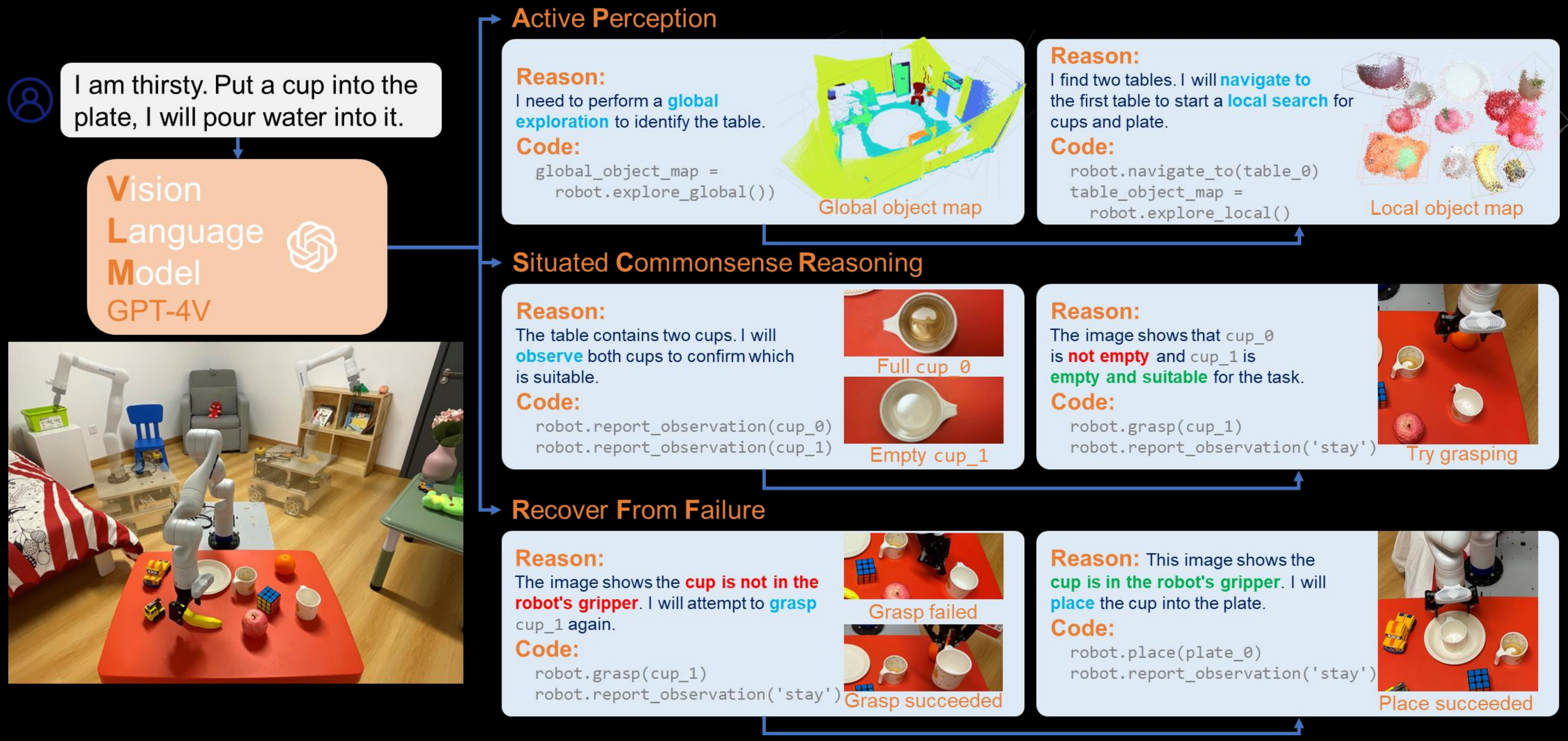
LEO (ICML'24)

Open-Vocabulary Mobile Manipulation

- *(ICRA'25) Closed-Loop Open-Vocabulary Mobile Manipulation with GPT-4V*

Come-Robot

LLM-based Closed-Loop Open-Vocabulary Manipulation



Zhi et al., Closed-Loop Open-Vocabulary Mobile Manipulation with GPT-4V, ICRA 2025


I am hungry. Could you give me some food? And pass me a cup of juice.
15x
Active
Perception

Problem?

No learning, just inference

ICML
International Conference
On Machine Learning

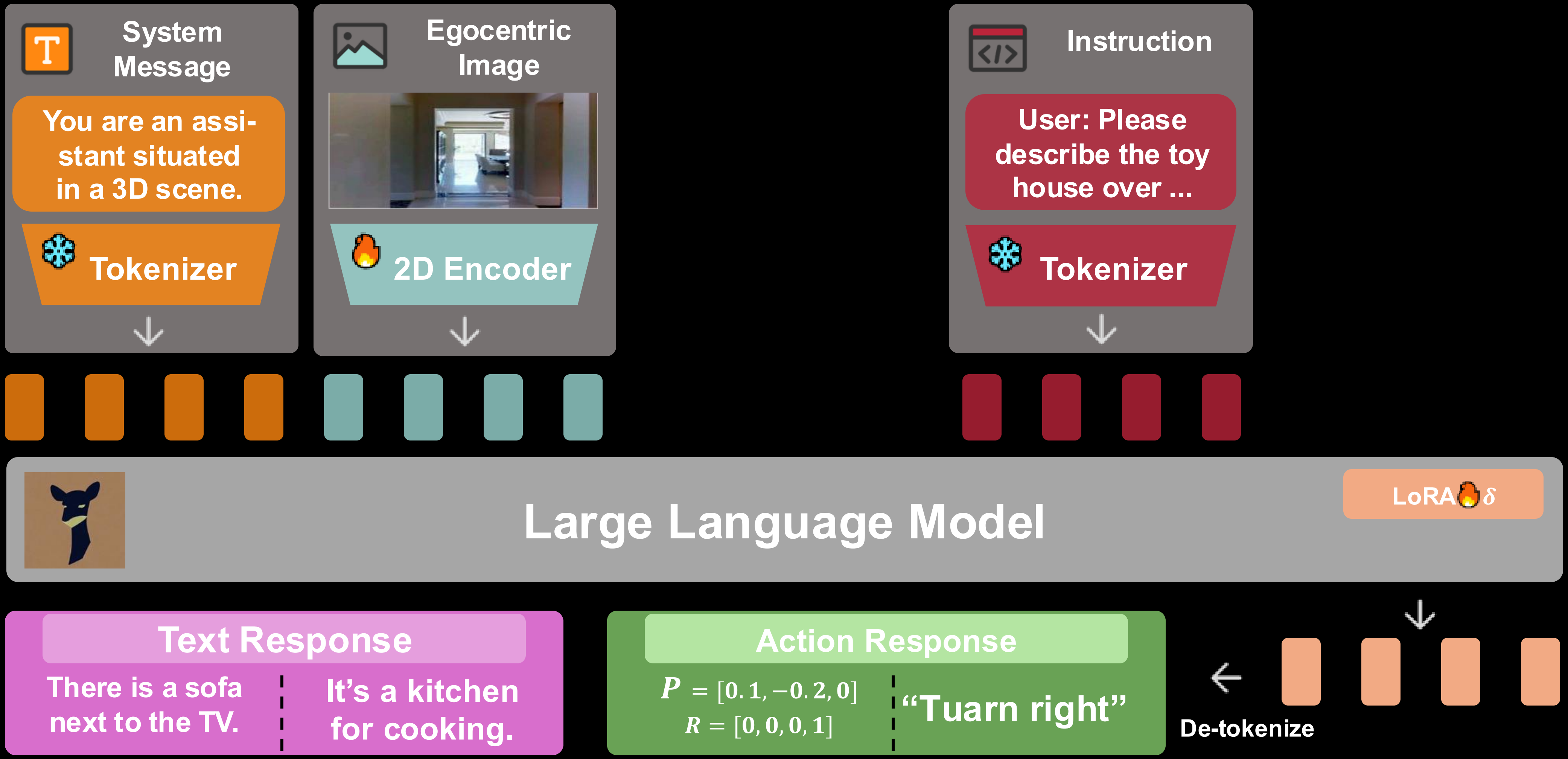
Improving Spatial Understanding for VLAs

- *(ICML'24) LEO: An Embodied Generalist Agent in 3D World*
- *(ArXiv'25) LEO-VL: Efficient Scene Representation for Scalable 3D Vision-Language Learning*

LEO

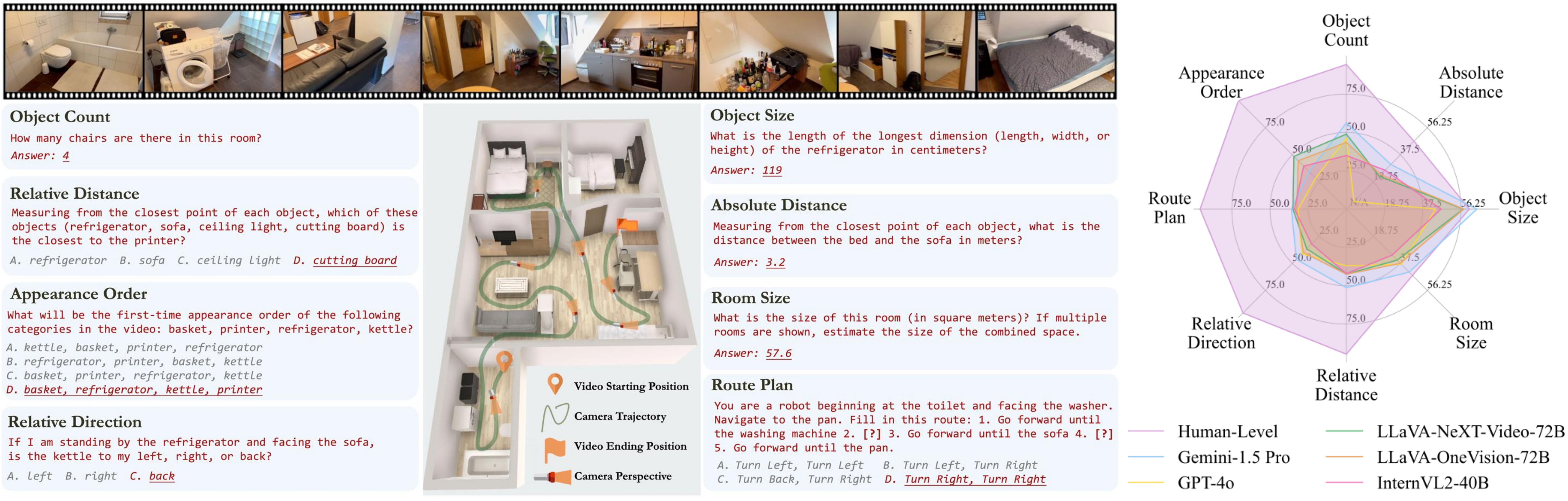
LEO-VL

Vision-Language-Action Models



Can 2D MLLMs Understand 3D Scenes?

VSI-Bench (CVPR 2025)

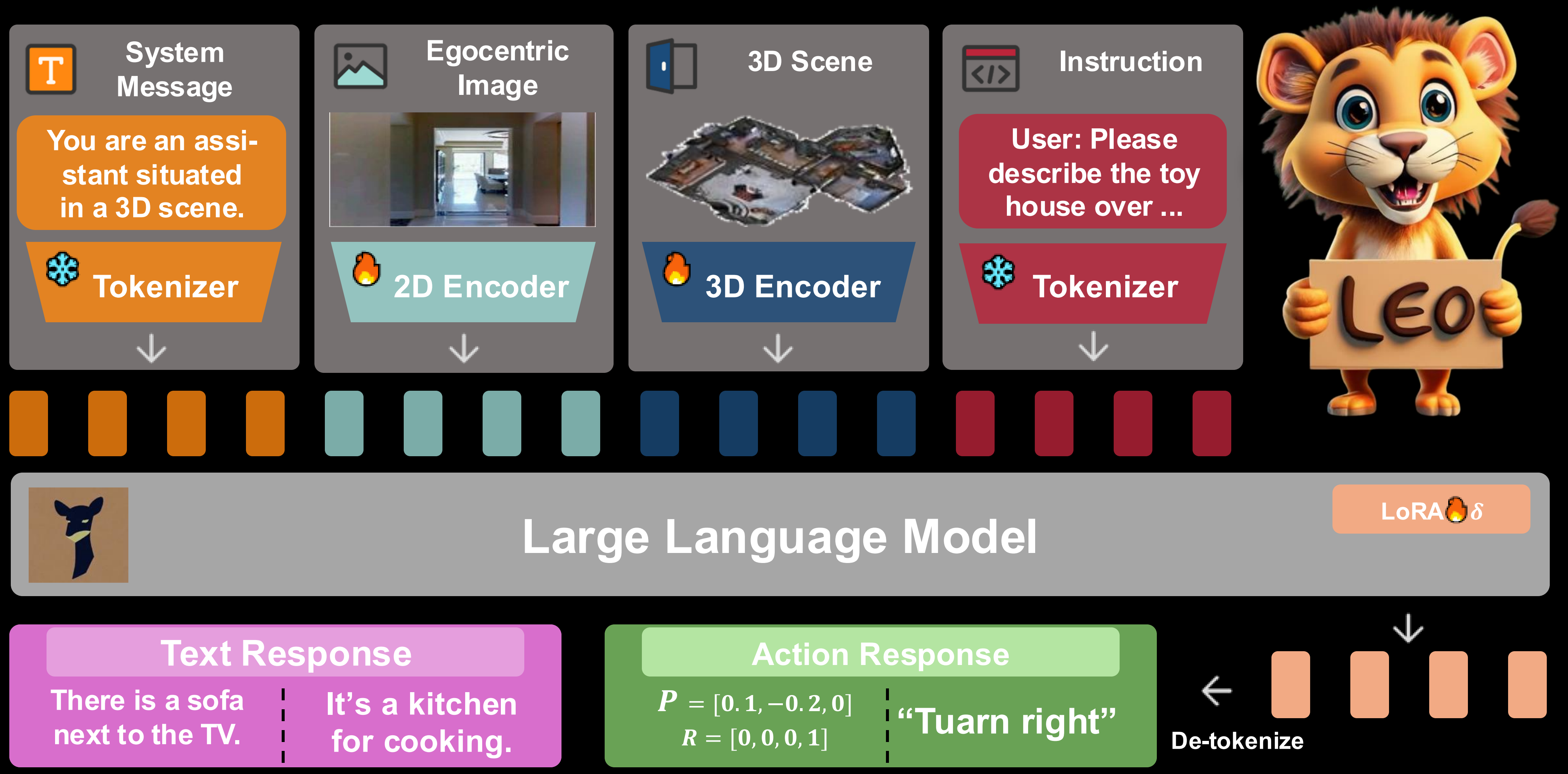


Despite powerful understanding of images and videos, current 2D MLLMs significantly lag far behind humans in 3D scene understanding, especially spatial reasoning.

Real Deployment?

Depth is a problem, but normally not equipped and used by VLAs

Vision-Language-Action Models



An Embodied Generalist Agent in 3D World

Jiangyong Huang[1,2]*, Silong Yong[1,3]*, Xiaojian Ma[1]*, Xiongkun Linghu[1]*, Puhao Li[1,4],
Yan Wang[1], Qing Li[1], Song-Chun Zhu[1,2,4], Baoxiong Jia[1], Siyuan Huang[1]

[1]Beijing Institute for General Artificial Intelligence (BIGAI)
[2]Peking University [3]Carnegie Mellon University [4]Tsinghua University

https://embodied-generalist.github.io/

Embodied Generalist Agent

Capabilities: ***Perception***, ***Grounding***, ***Reasoning***, ***Planning***, ***Acting***

Tasks

3D Object Captioning
Scene Captioning

3D Question Answering
Embodied Reasoning

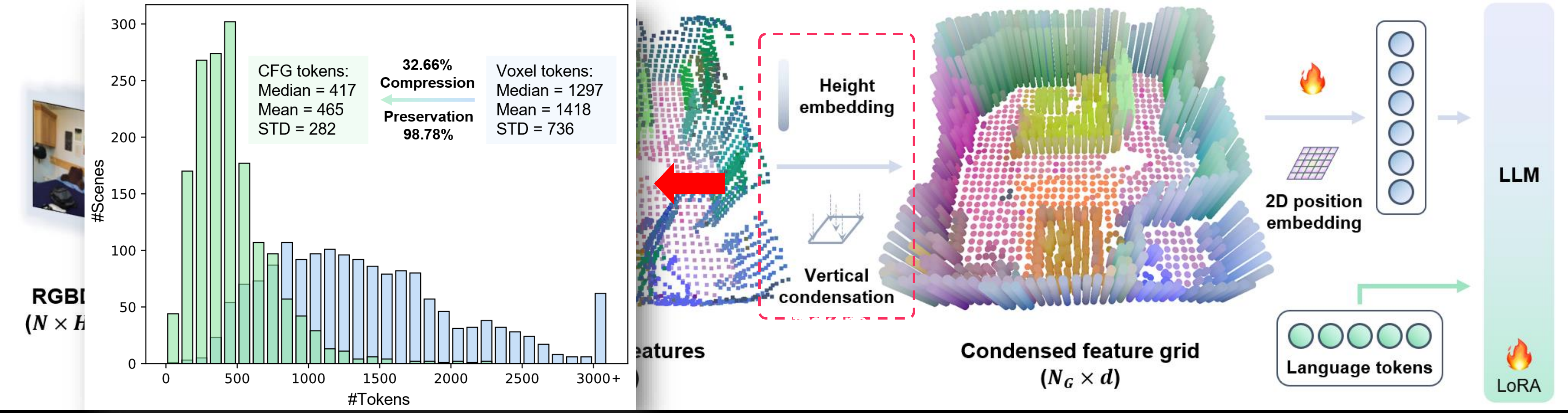
3D Dialogue
Task Planning

Embodied Navigation
Robotic Manipulation

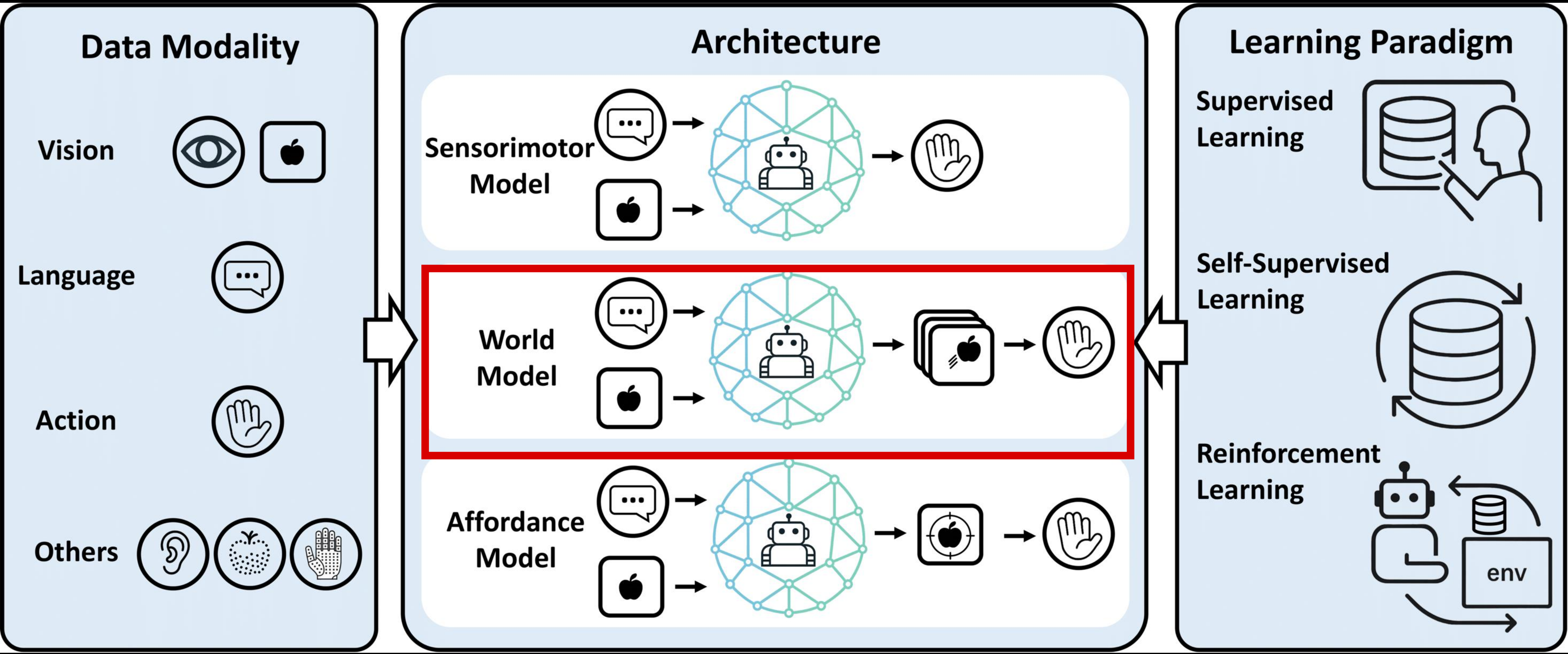
3D World

Efficient Representation Bridging 2D-3D Perception

	Pros	Cons
3D perception	Explicit 3D structure	Complex pre-processing pipelines, learning difficulty
2D perception	Strong capability	Significant computation overhead (thousands of tokens)

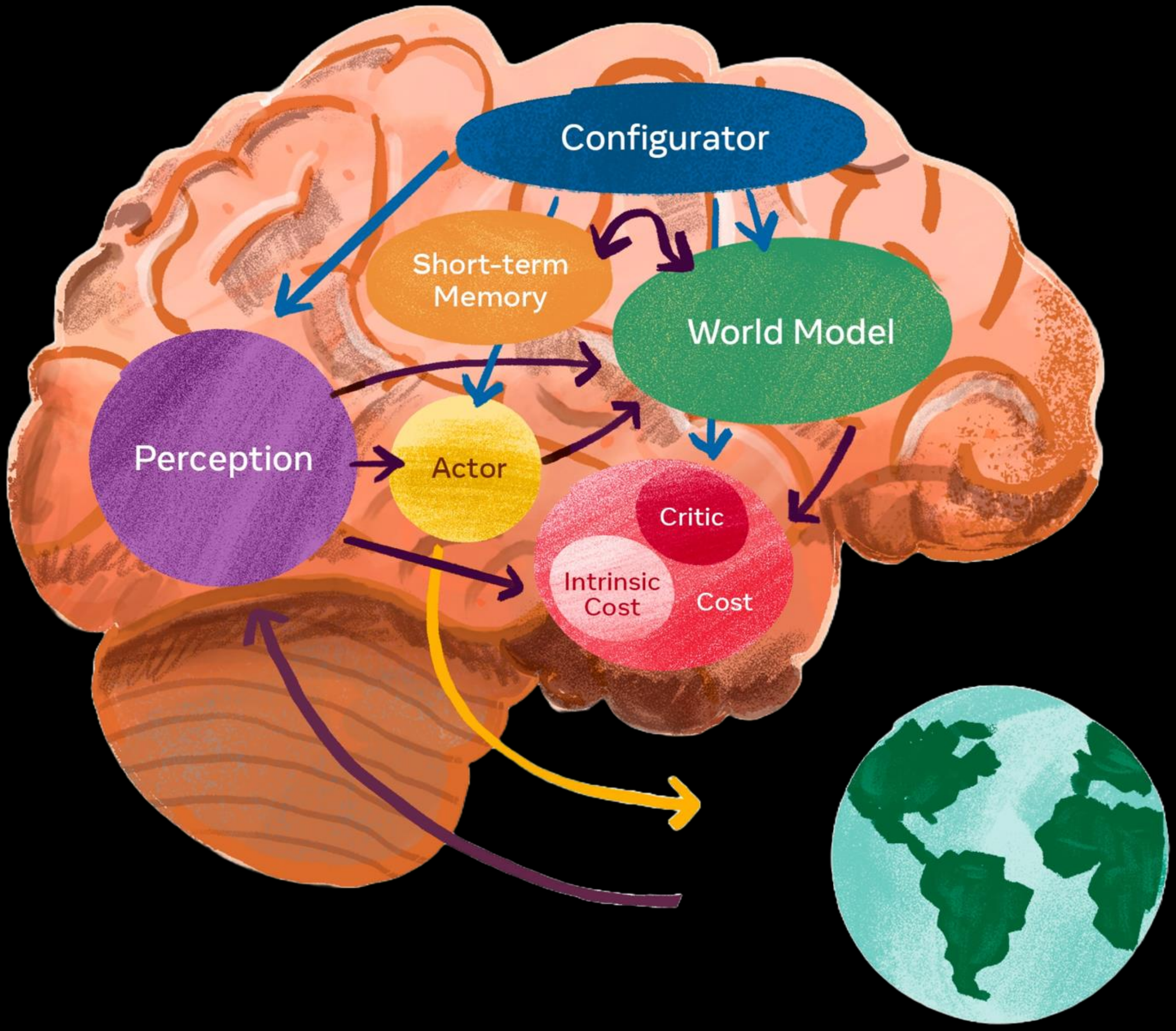


A Closer Look at VLA Model Design



Vision-Language-Action Models for Robotics (IEEE Access 2025)

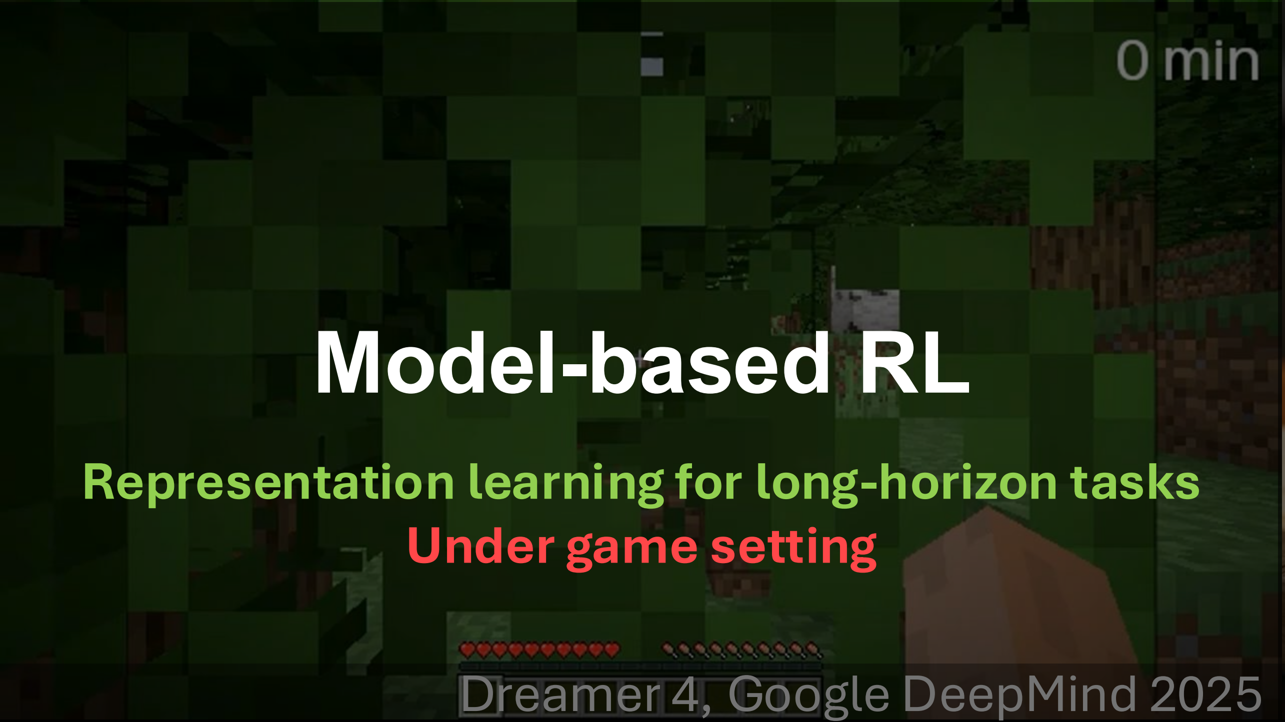
World Models



"If the organism carries a small-scale model of external reality and of its own possible actions within its head, it is able to try out various alternatives, conclude which is the best of them, react to future situations before they arise, utilize the knowledge of past events in dealing with the present and future."

— Kenneth Craik (1943)

Figure credit: JEPA Blog, Meta 2022



Model-based RL

Representation learning for long-horizon tasks

Under game setting

Dreamer 4, Google DeepMind 2025

Video Generation

Flexible conditional generation

Weak physical consistency / modeling of action

Veo 3.1, Google Deepmind 2025

Dream to Generate

Latent Action Learning

Aligning video generation with latent actions

Limited by the view-point

DreamGen, NVIDIA GEAR 2025

Spatial Representations

World modeling with 3D Gaussians

Interactiveness for robot manipulation?

Marble, WorldLabs 2025

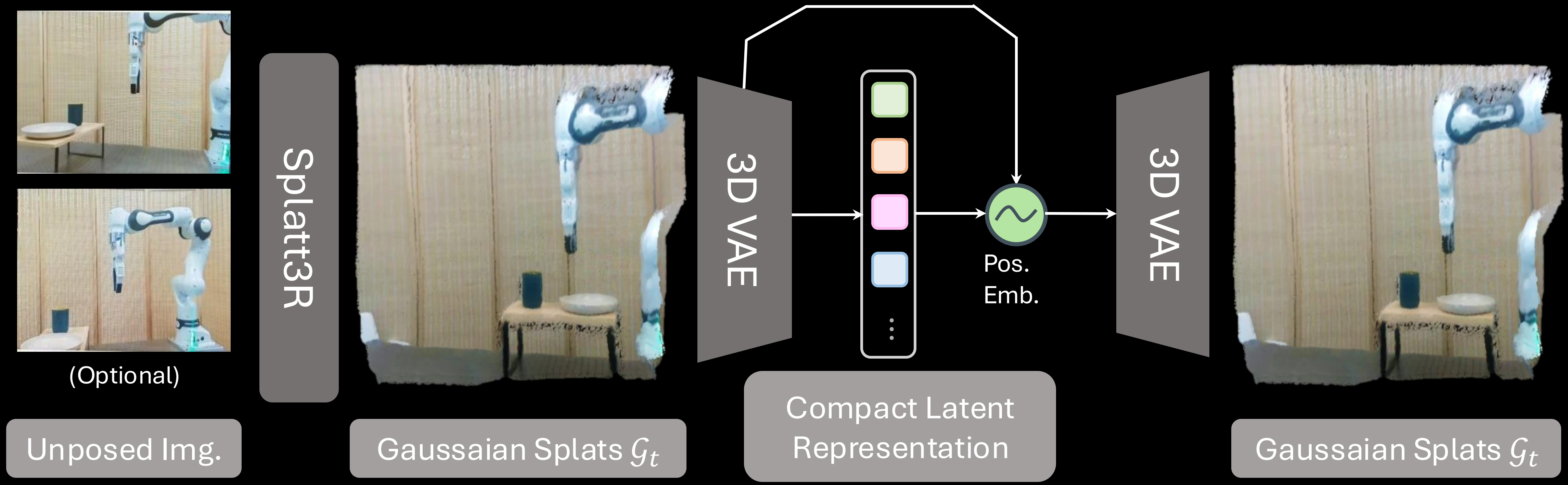
ICCV OCT 19-23, 2025 HONOLULU HAWAII

Scalable World Modeling with 3D Gaussians

- *(ICCV'25) GWM: Towards Scalable Gaussian World Modeling for Robotic Manipulation*

GWM

Encoding 3D Gaussians into Latent Space

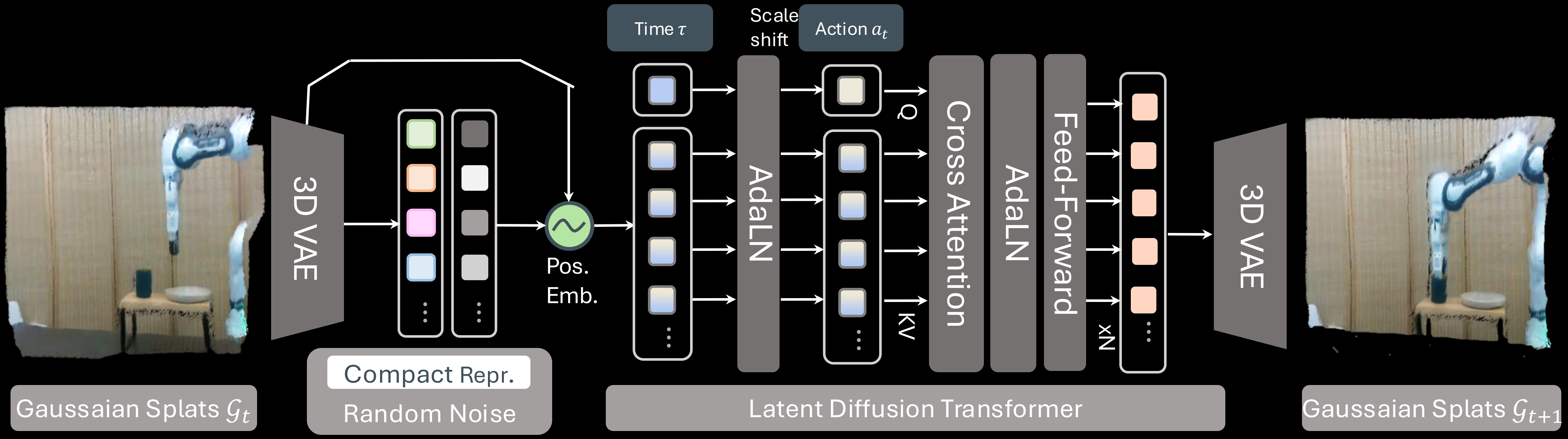


Feed-Forward 3D Gaussian Reconstruction

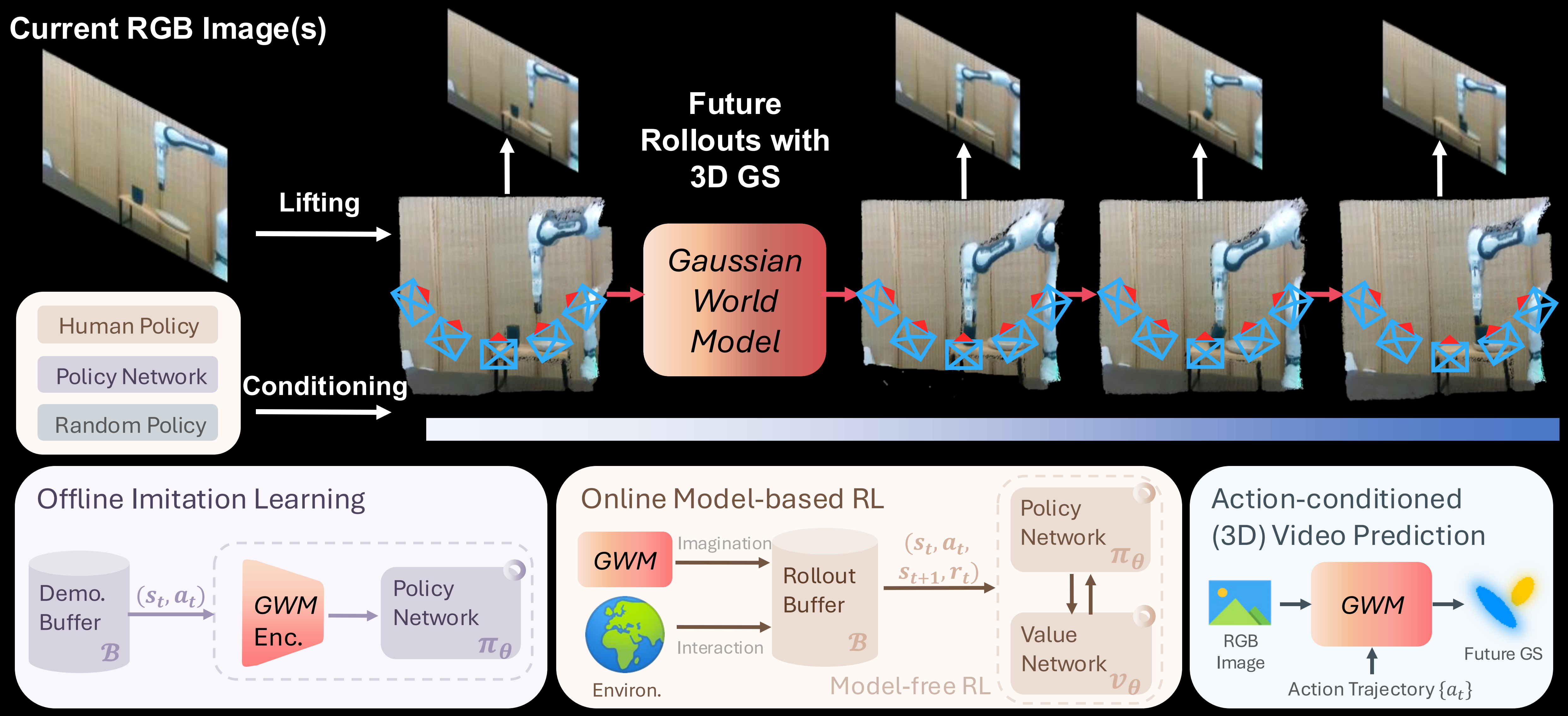
FPS-based Subsampling Query-based Encoding

Rendering / Geometry Supervision

GWM: Gaussian World Model

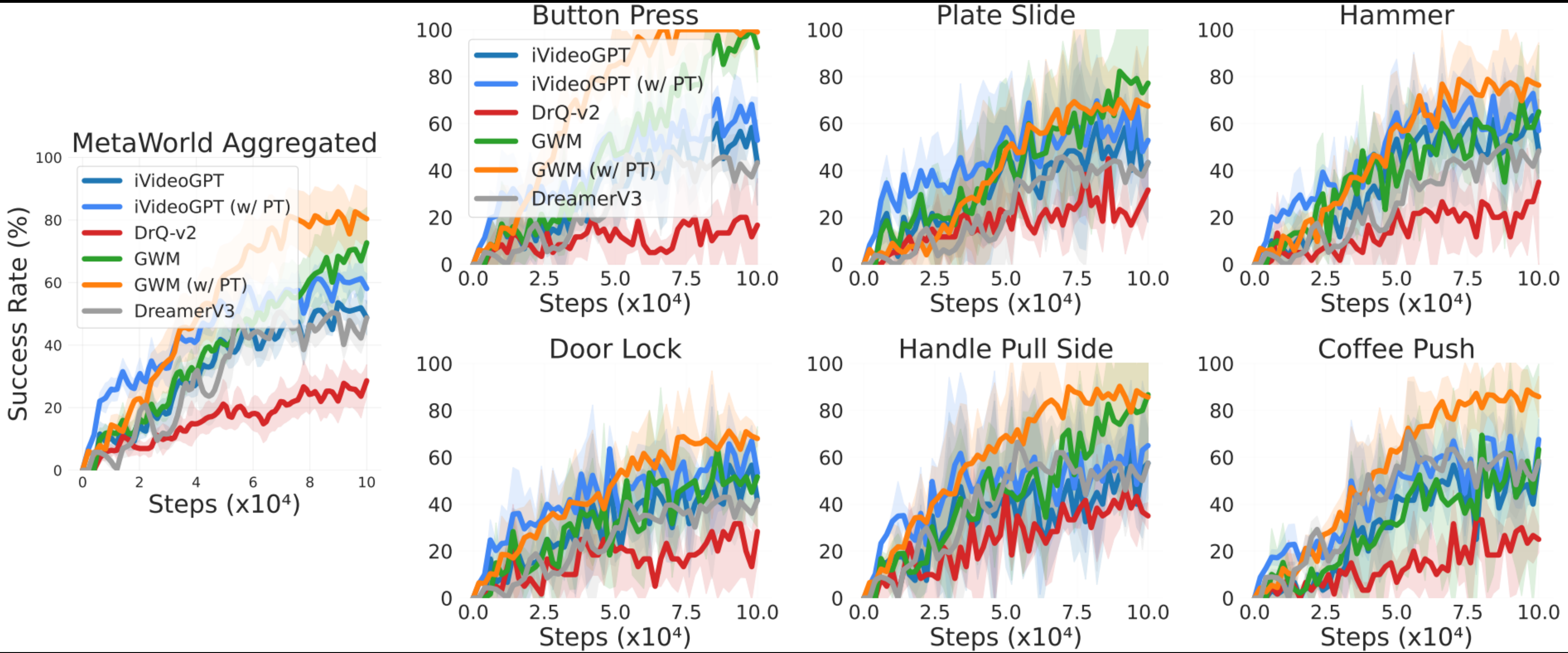


DiT-based Dynamics Learning and Prediction



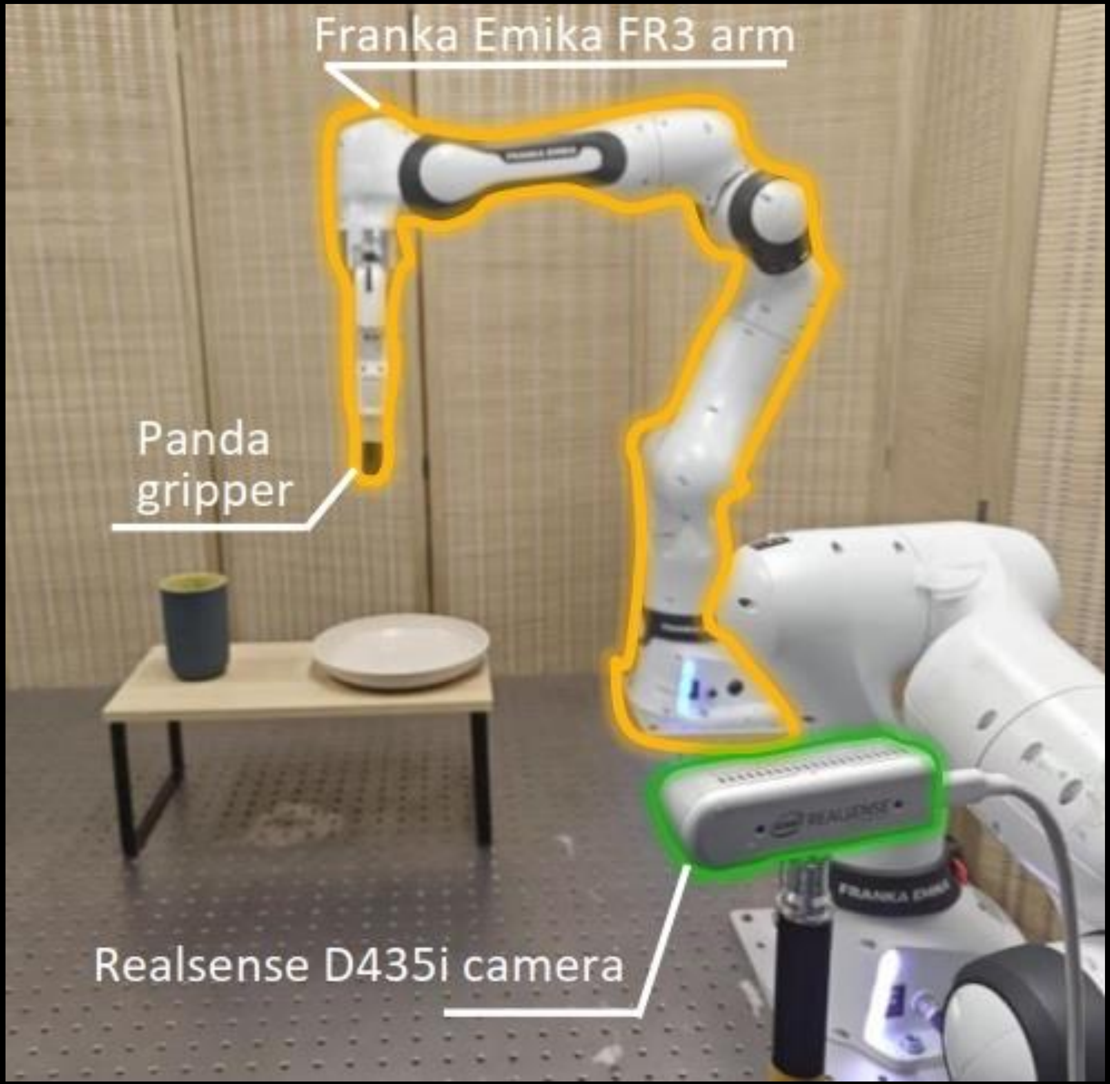
Current RGB Image(s)
Lifting
Future Rollouts with 3D GS
Gaussian World Model
Human Policy
Policy Network
Random Policy
Conditioning
Offline Imitation Learning
Demo. Buffer
$\mathcal{B}$
(s_t, a_t)
GWM Enc.
Policy Network
π_θ
Online Model-based RL
GWM
Imagination
Environ.
Interaction
Rollout Buffer
$\mathcal{B}$
(s_t, a_t, s_{t+1}, r_t)
Policy Network
π_θ
Value Network
v_θ
Model-free RL
Action-conditioned (3D) Video Prediction
RGB Image
GWM
Future GS
Action Trajectory $\{a_t\}$

GWM for Online Model-based RL



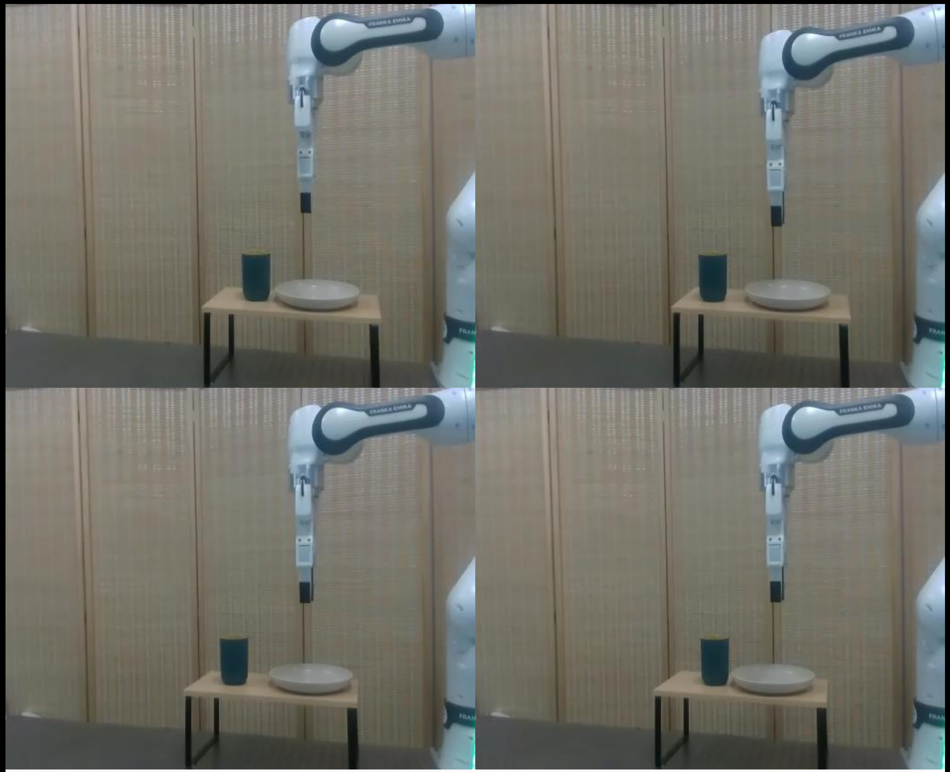
Additional reward learning on top of GWM for online RL

GWM for Real-World Robot Manipulation



Diffusion Policy

DP w/ GWM



Comparsion

FRANKA-PNP	Diffusion Policy	GWM (Ours)
Cup distractor	6/10	**7/10**
Plate distractor	1/5	**3/5**
Table distractor	0/5	**3/5**
Total	7/20	**13/20**

3. Evolving Force-Aware Control Skills for Human-Robot Interaction

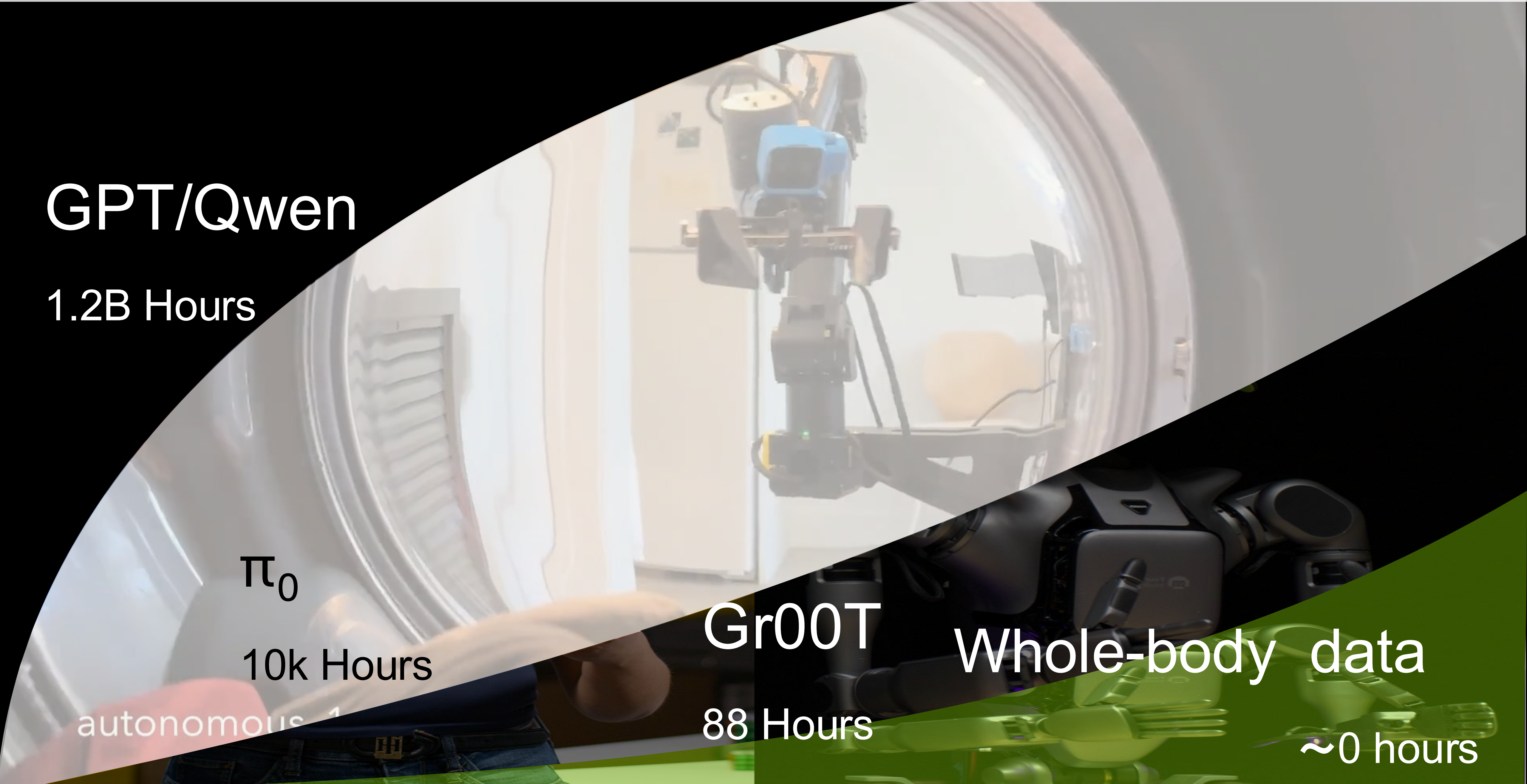
Figure generated by Nano Banana Pro

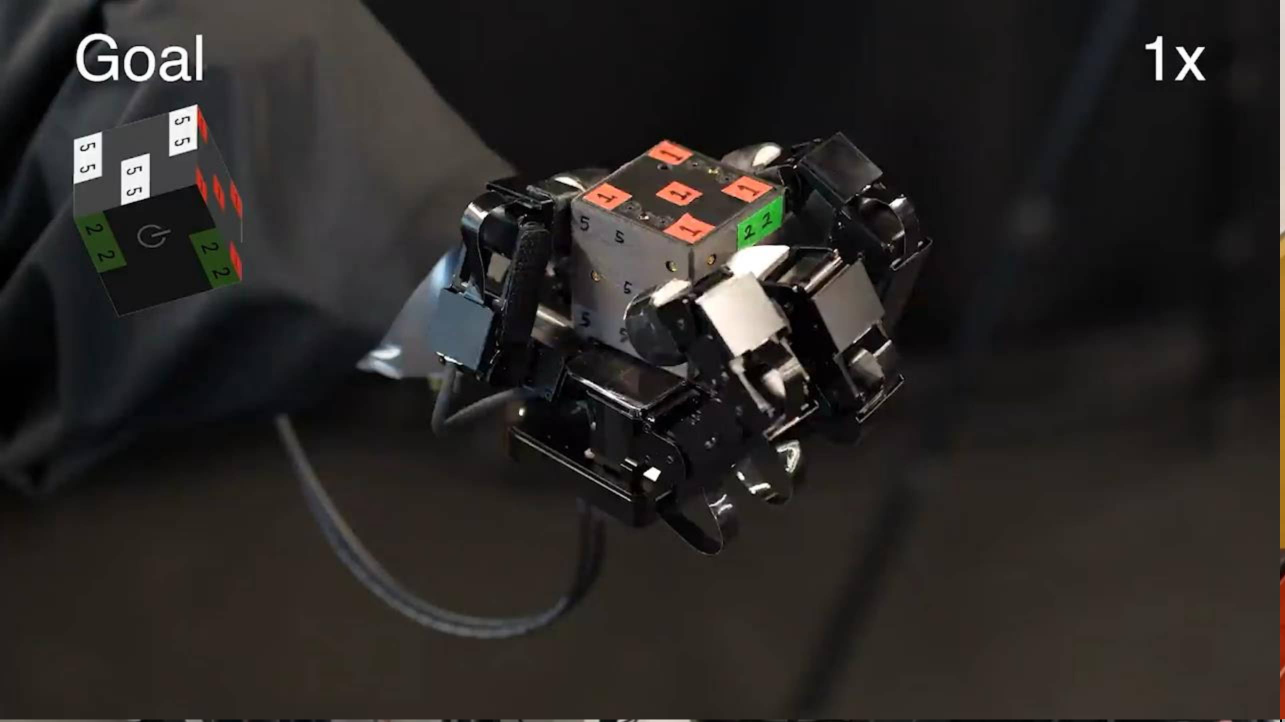
Can Humanoids Interact at This Level?

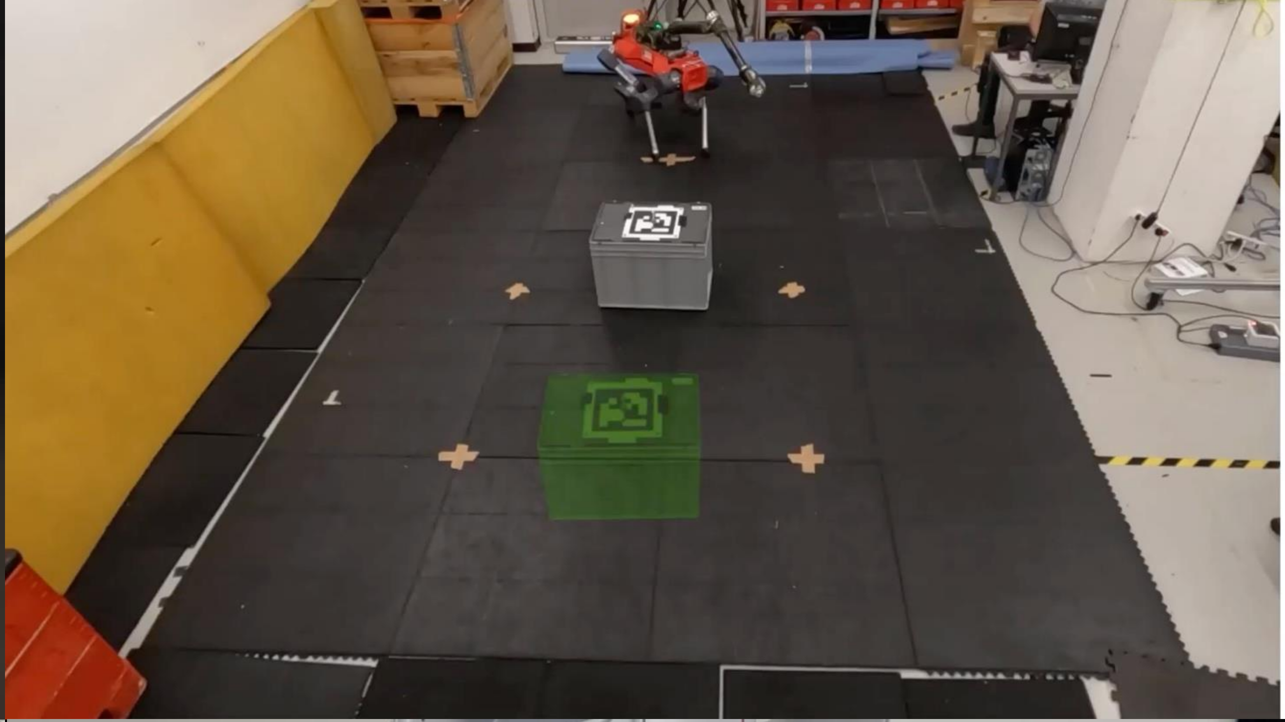
Humans effortlessly **squat to retrieve objects** from the ground and then **walk to another distant place**.

Enable holistic and long-horizon humanoid–scene interaction

Whole-body demonstrations is significantly limited



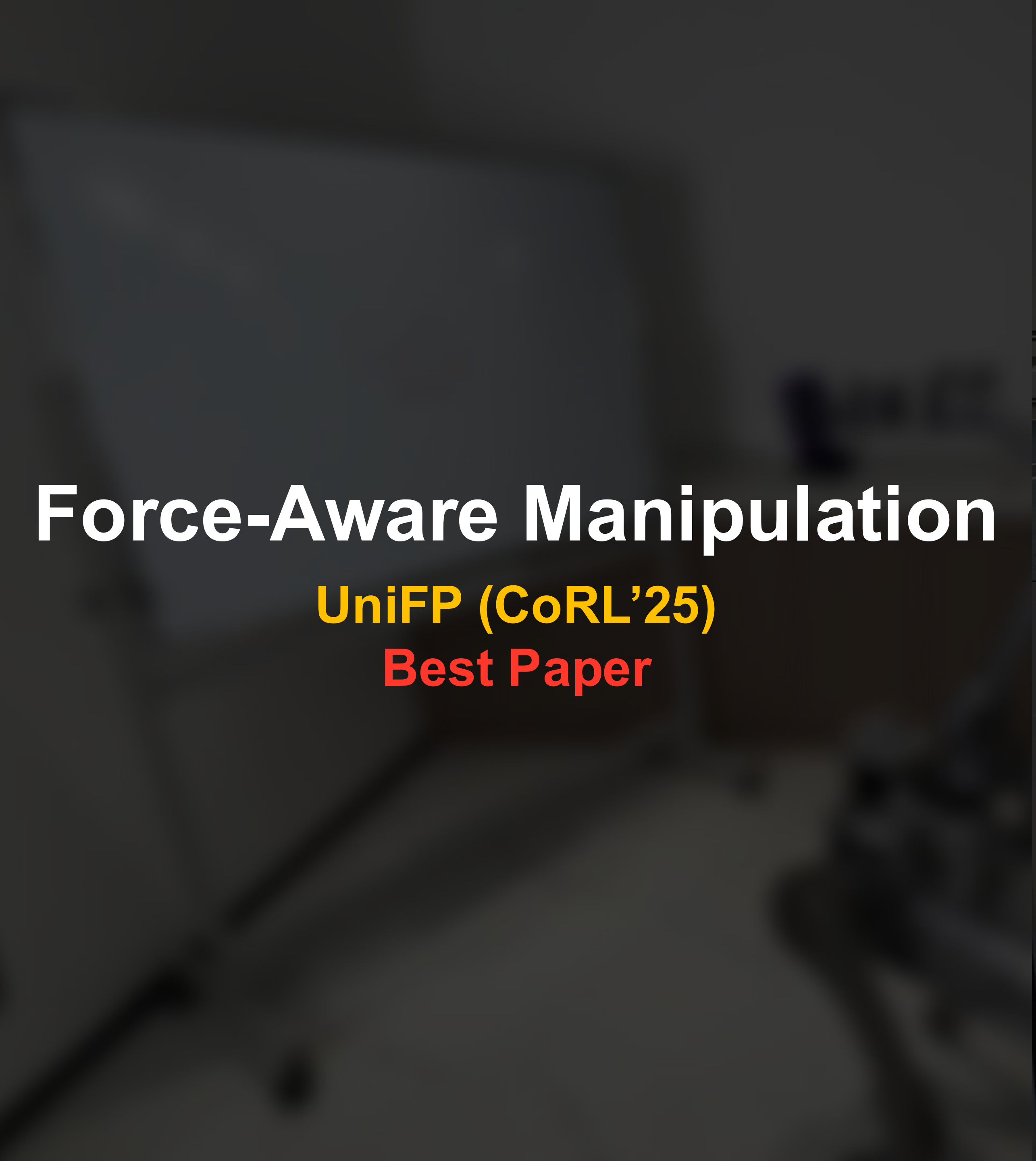




Contact-Rich Workshop @ ICRA 2025

Goal of Action and Control

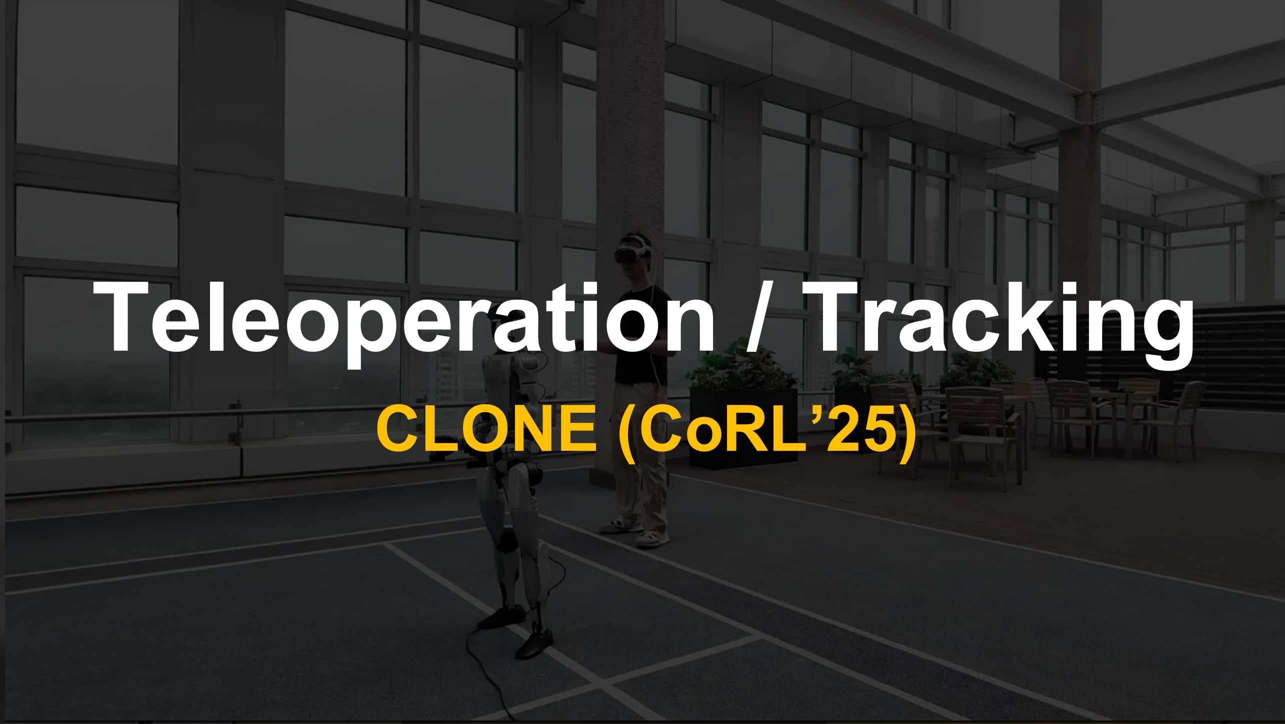
- Enable agile and stable whole-body control for humanoid robots
- Mitigate the missing force modality for contact-rich manipulation tasks
- Safe and helpful human-robot collaboration patterns



Force-Aware Manipulation

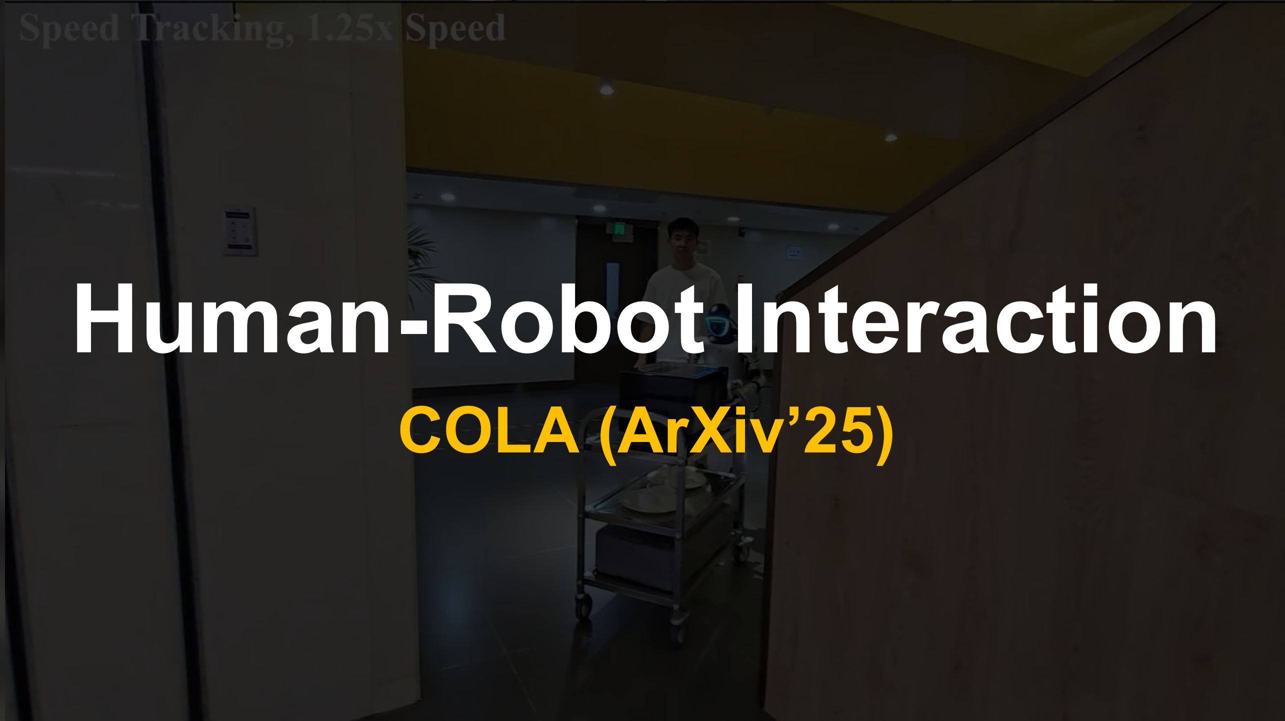
UniFP (CoRL’25)

Best Paper



Teleoperation / Tracking

CLONE (CoRL’25)



Human-Robot Interaction

COLA (ArXiv’25)

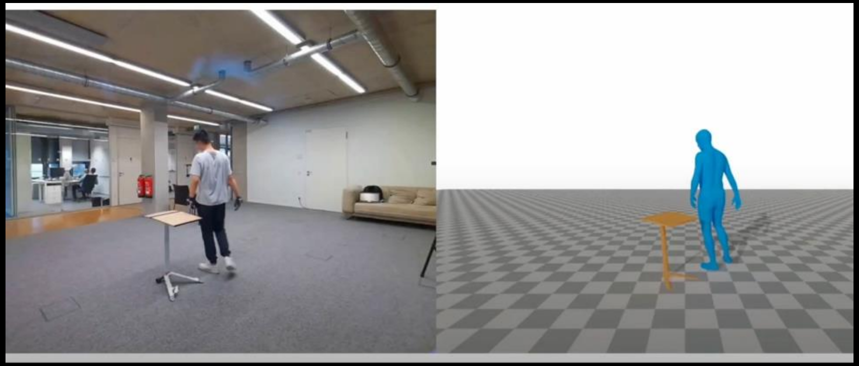
Agile Humanoid Whole-Body Teleoperation

- *(CoRL'25) CLONE*: Closed-Loop Whole-Body Humanoid Teleoperation for Long-Horizon Tasks

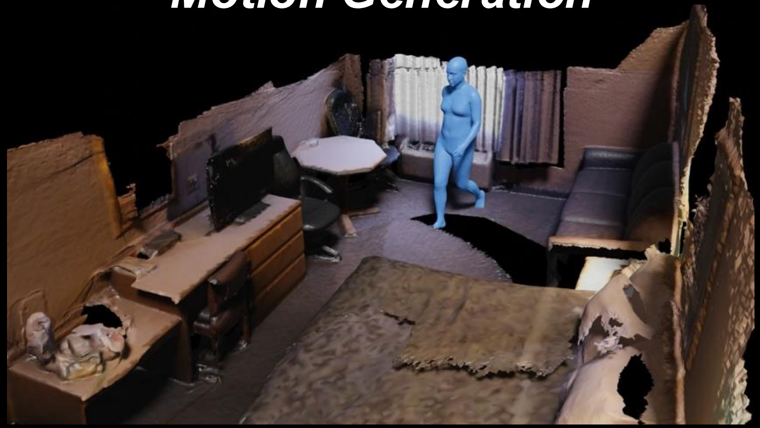
CLONE

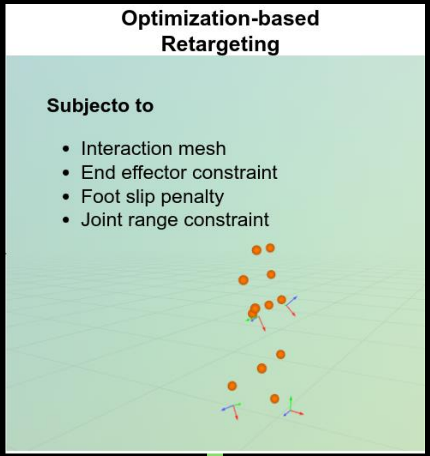
Driving a Physical Humanoid with Human Motion

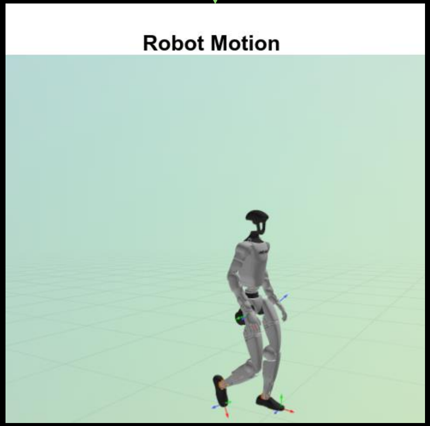
MoCap



Motion Generation







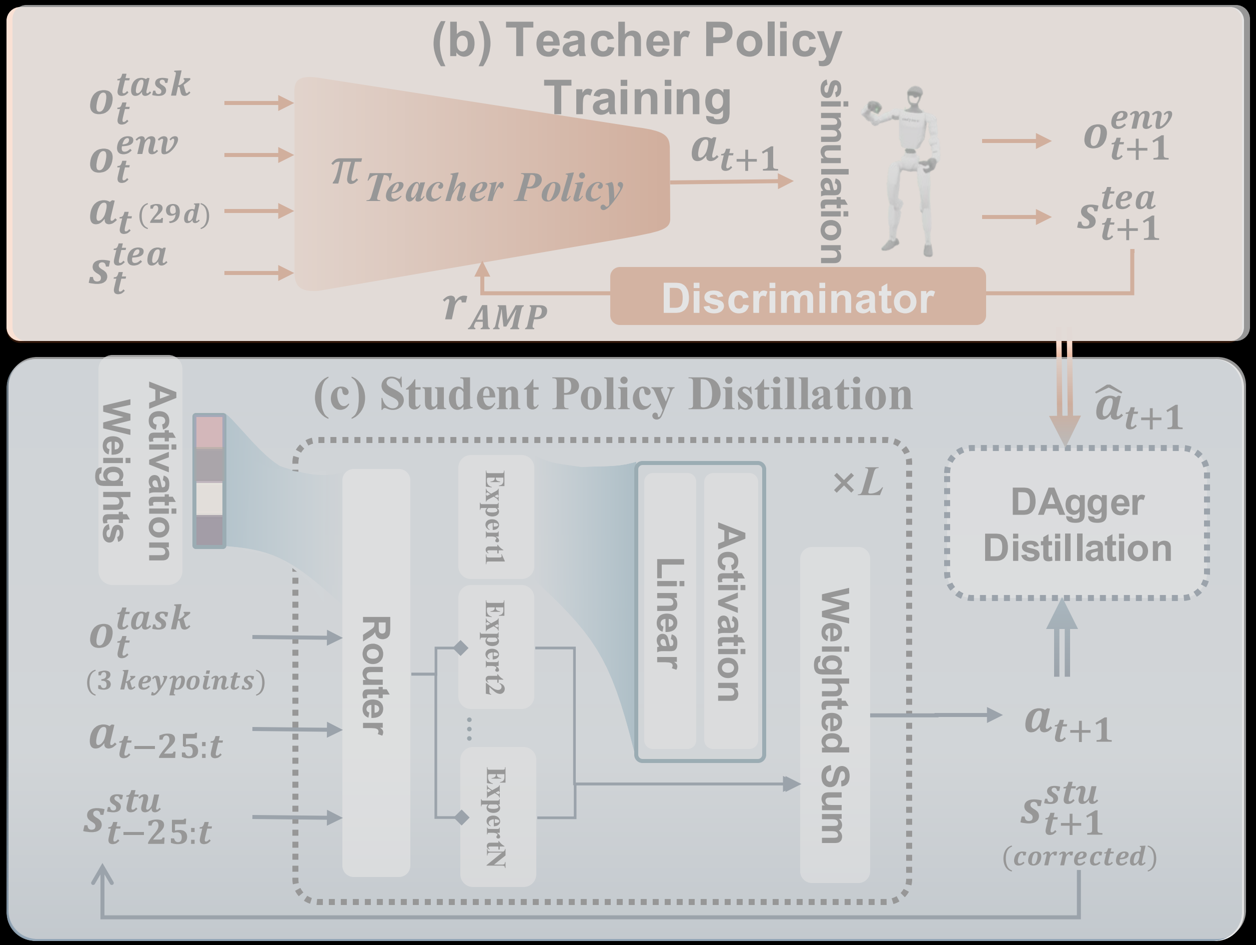
LAFAN

Humanoid Controller

MLP
MoE
Transformer
PID
MPC
…

Human Motion → ***Motion Retargeting*** → ***Controller*** → ***Robot Execution***

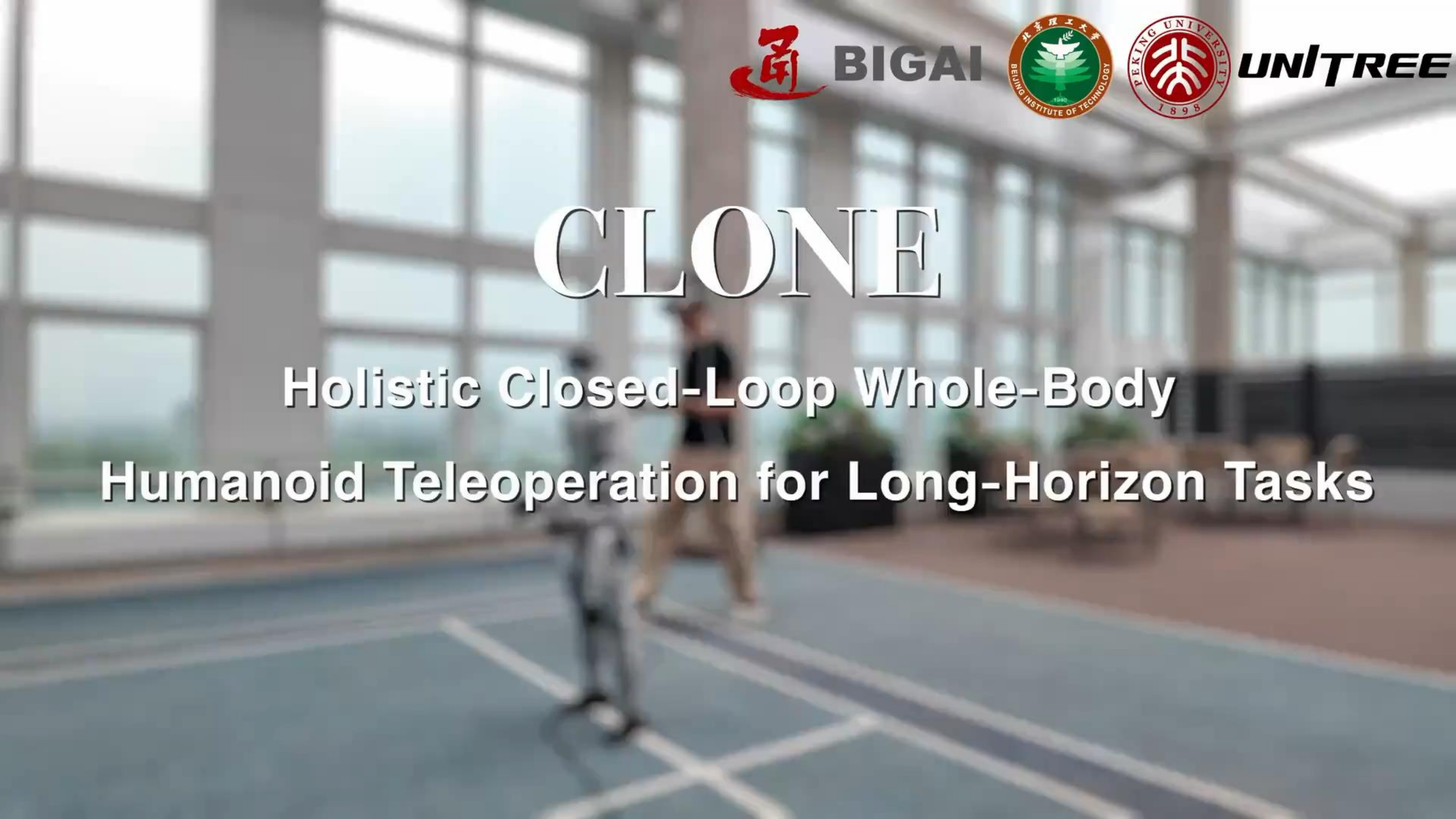
CLONE: Humanoid Whole-Body Teleoperation



Learning a teacher policy with privileged information for human motion tracking

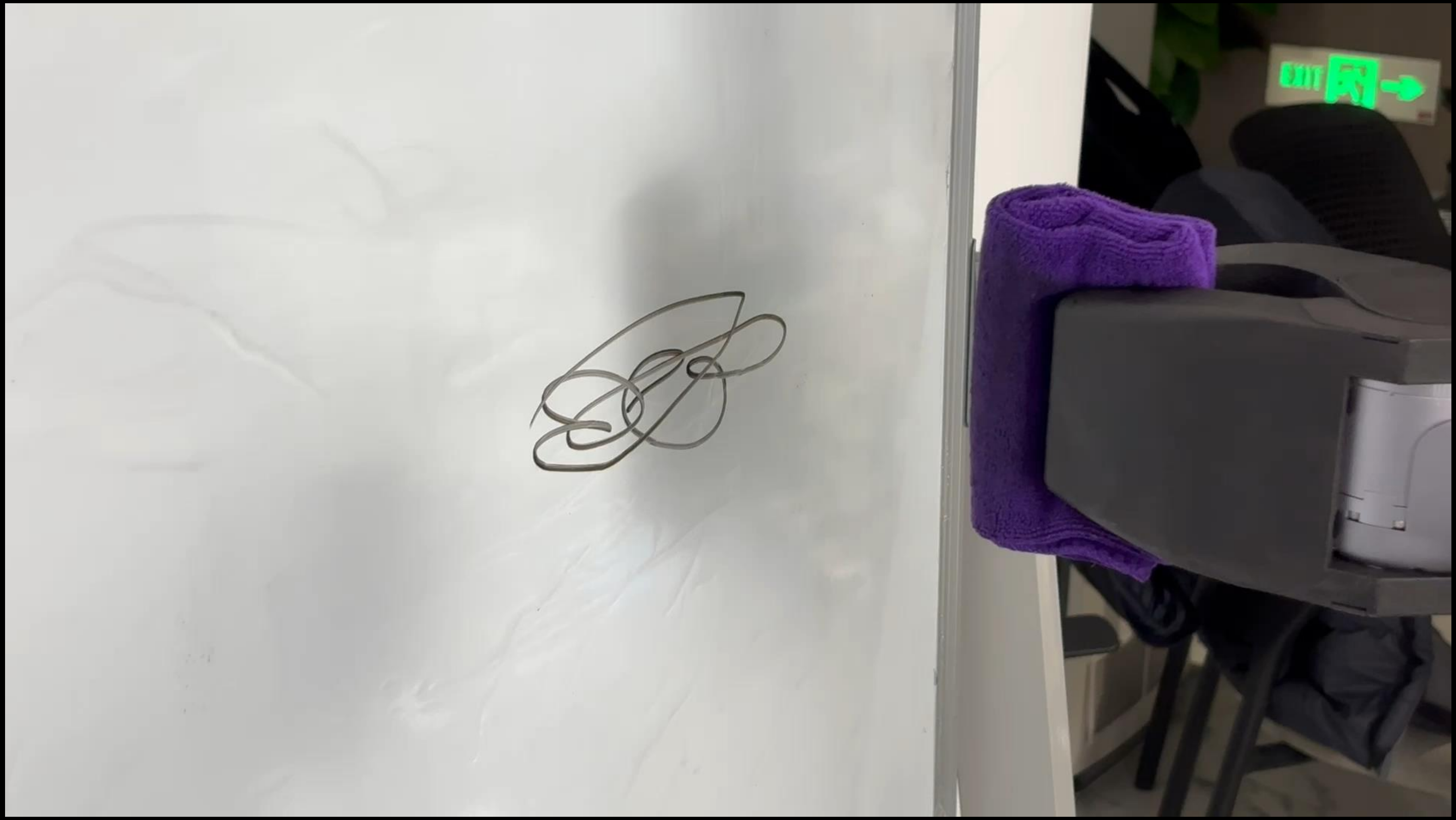
Distilling a MoE-based student policy with Behavior Cloning (Dagger)

Li et al., CLONE:Closed-Loop Whole-Body Humanoid Teleoperation for Long-Horizon Task, CoRL 2025


BIGAI
UNITREE
CLONE
Holistic Closed-Loop Whole-Body
Humanoid Teleoperation for Long-Horizon Tasks

OK, can we let the robot help us wipe the whiteboard first after meeting?

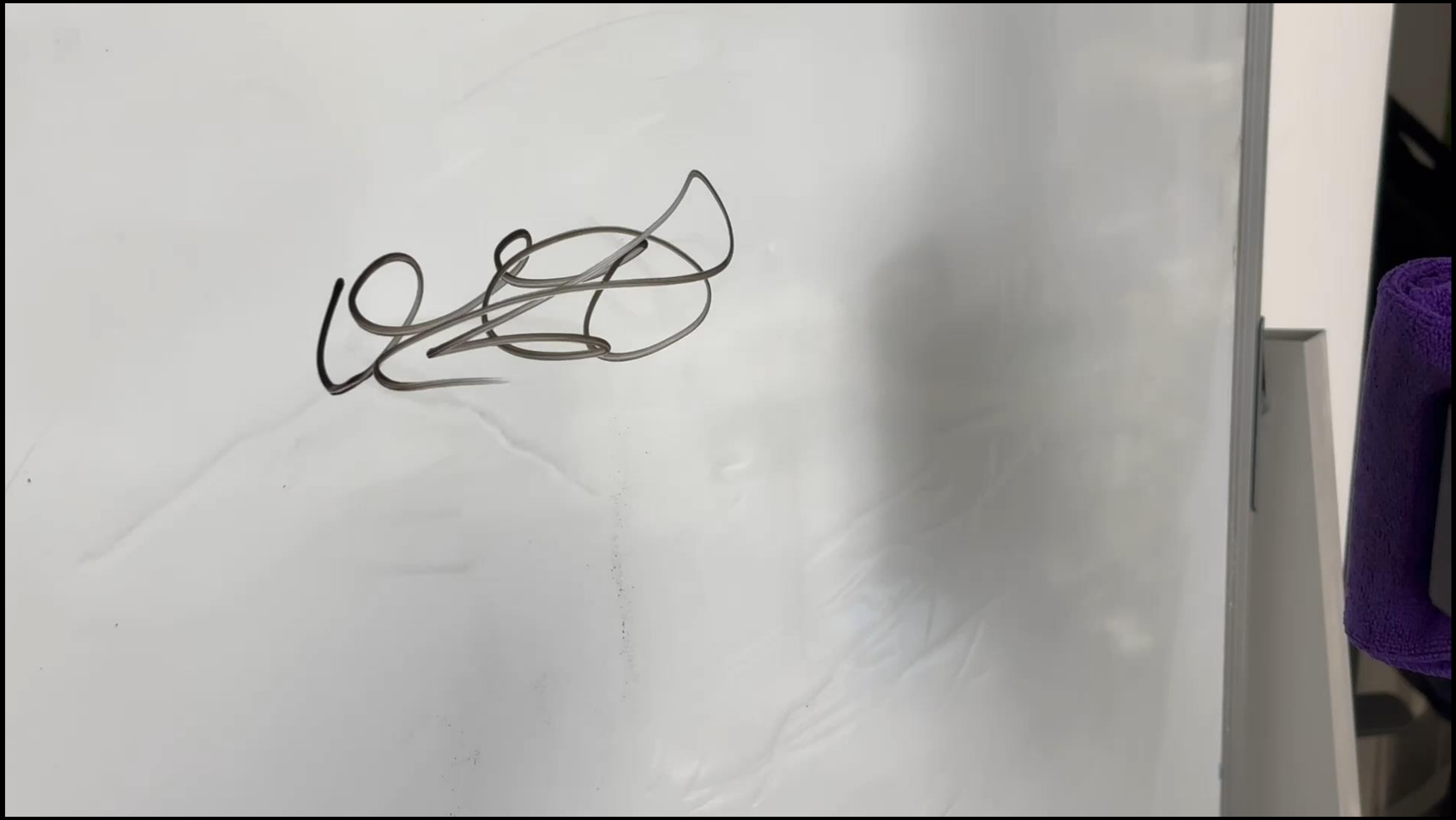
Let me collect the data and imitation learning will solve the rest ☺



Data collection is a disaster

OK, can we let the robot help us wipe the whiteboard first after meeting?

Let me collect the data and imitation learning will solve the rest ☺



Of course, the learned policy failed no matter how much data used ☹

Unified Force-Position Control Policy

- *(CoRL'25 Best Paper) Learning a Unified Policy for Position and Force Control in Legged Loco-Manipulation*

UniFP

Revisiting the control formulation

mass-spring-damper system

$$\boldsymbol{F} = K\left(\mathrm{x} - \boxed{\mathrm{x}^{\mathrm{cmd}}}\right) + \cancel{D\left(\dot{\mathrm{x}} - \boxed{\dot{\mathrm{x}}^{\mathrm{cmd}}}\right)} + \cancel{M\left(\ddot{\mathrm{x}} - \boxed{\ddot{\mathrm{x}}^{\mathrm{cmd}}}\right)}$$

$$\mathrm{x} = \mathrm{x}^{\mathrm{cmd}} + \frac{F}{K}$$

And if the end effector moves really slowly...

Revisiting the control formulation

mass-spring-damper system

$$\boldsymbol{F} = K\left(\mathrm{x} - \mathrm{x}^{\mathrm{cmd}}\right) + \cancel{D\left(\dot{\mathrm{x}} - \dot{\mathrm{x}}^{\mathrm{cmd}}\right)} + \cancel{M\left(\ddot{\mathrm{x}} - \ddot{\mathrm{x}}^{\mathrm{cmd}}\right)}$$

$$\mathrm{x} = \mathrm{x}^{\mathrm{cmd}} + \frac{F}{K}$$

Force can be estimated via position offsets!

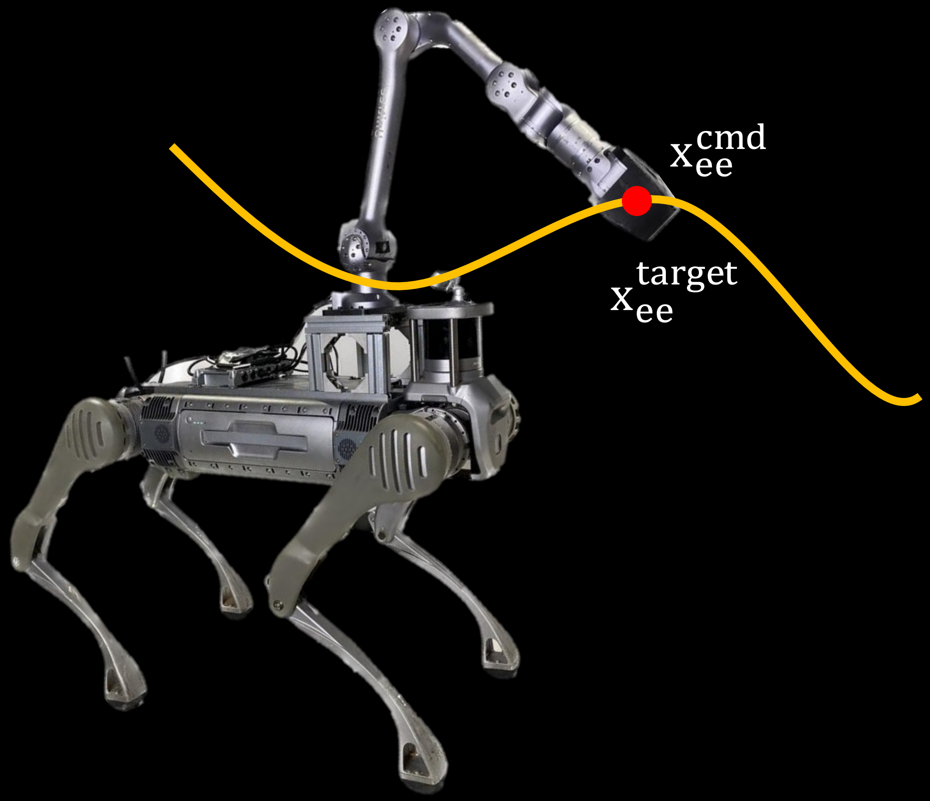
Tracking the force-adjusted position enables joint force-position control.

Formulating forces with positions

$$F = K\left(x - x^{\mathrm{cmd}}\right)$$

Position control

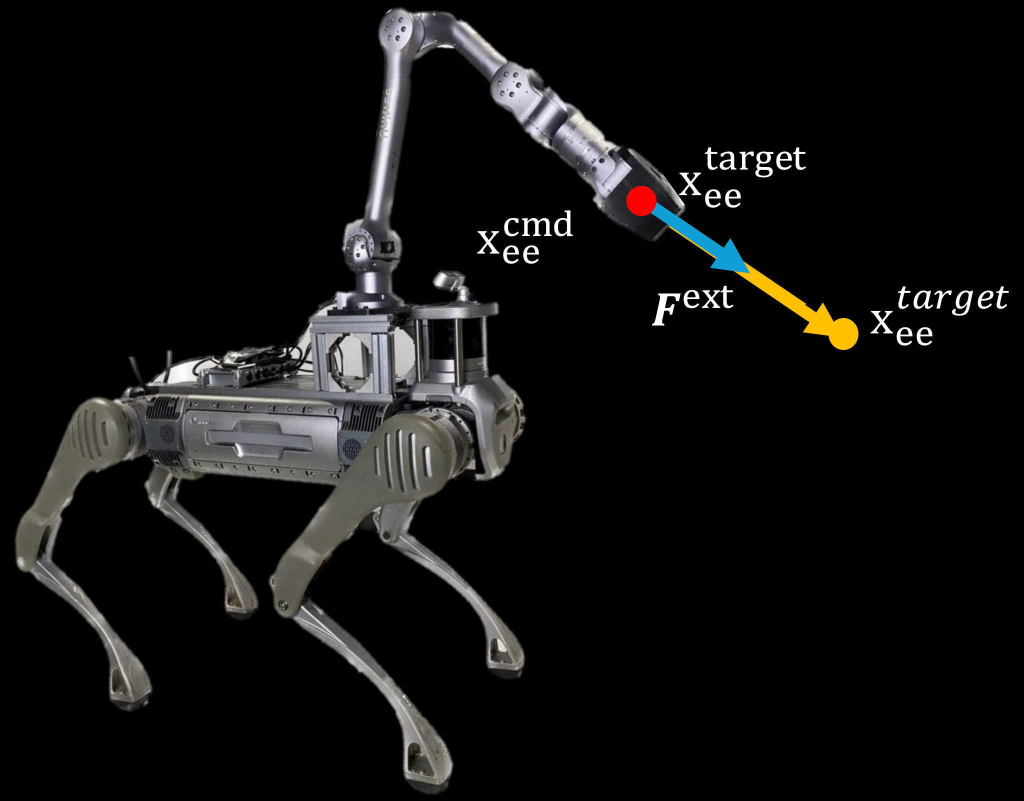
$$x^{\mathrm{target}} = x^{\mathrm{cmd}}$$



Formulating forces with positions

$$\boldsymbol{F} = K\left(\mathrm{x} - \mathrm{x}^{\mathrm{cmd}}\right)$$

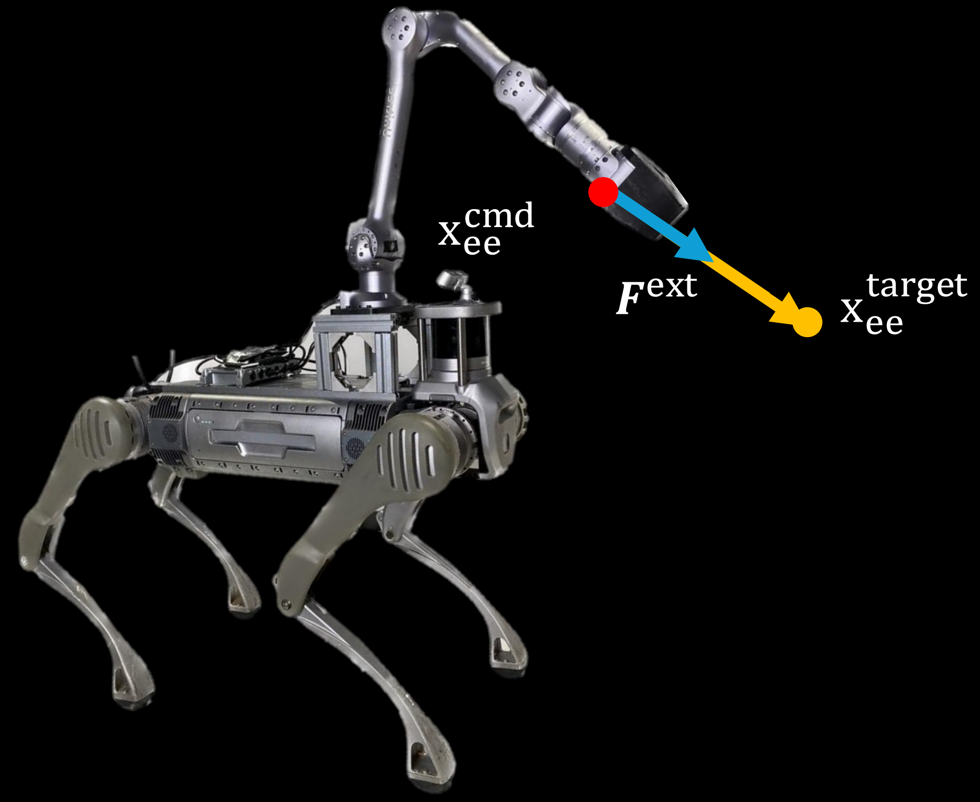
When with external force

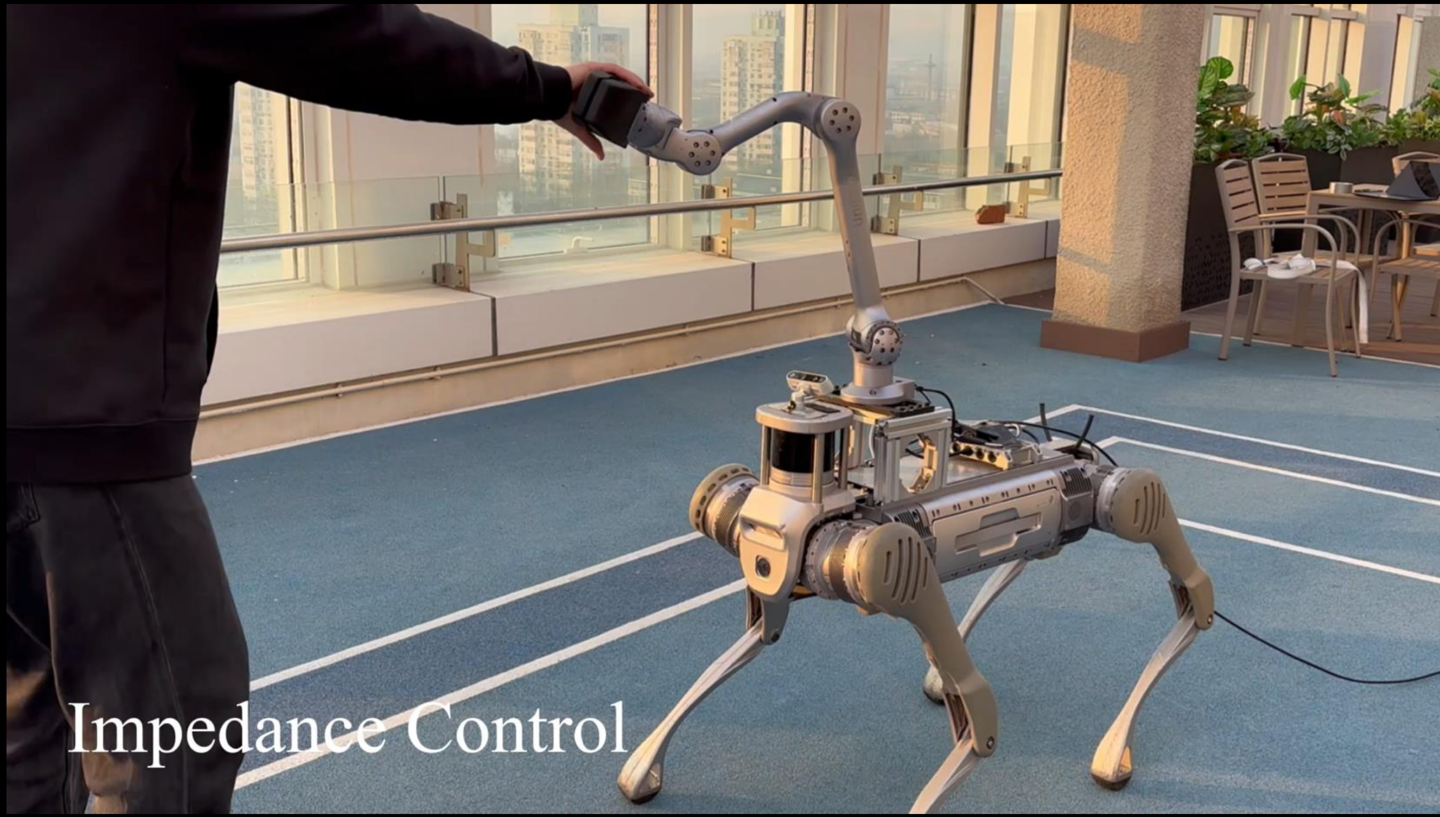


Formulating forces with positions

$$\boldsymbol{F} = K\left(\mathrm{x} - \mathrm{x}^{\mathrm{cmd}}\right)$$

Impedance control $$\mathrm{x}^{\mathrm{target}} = \mathrm{x}^{\mathrm{cmd}} + \frac{\boldsymbol{F}^{\mathrm{ext}}}{K}$$





Revisiting the control formulation

mass-spring-damper system

$$\boldsymbol{F} = \cancel{K\left(\mathrm{x} - \mathrm{x}^{\mathrm{cmd}}\right)} + D\left(\dot{\mathrm{x}} - \dot{\mathrm{x}}^{\mathrm{cmd}}\right) + \cancel{M\left(\ddot{\mathrm{x}} - \ddot{\mathrm{x}}^{\mathrm{cmd}}\right)}$$

And if we care about the locomotion

Revisiting the control formulation

mass-spring-damper system

$$\boldsymbol{F} = \cancel{K\left(\mathrm{x} - \mathrm{x}^{\mathrm{cmd}}\right)} + D\left(\dot{\mathrm{x}} - \dot{\mathrm{x}}^{\mathrm{cmd}}\right) + \cancel{M\left(\ddot{\mathrm{x}} - \ddot{\mathrm{x}}^{\mathrm{cmd}}\right)}$$

$$\dot{\mathrm{x}} = \dot{\mathrm{x}}^{\mathrm{cmd}} + \frac{F}{D}$$

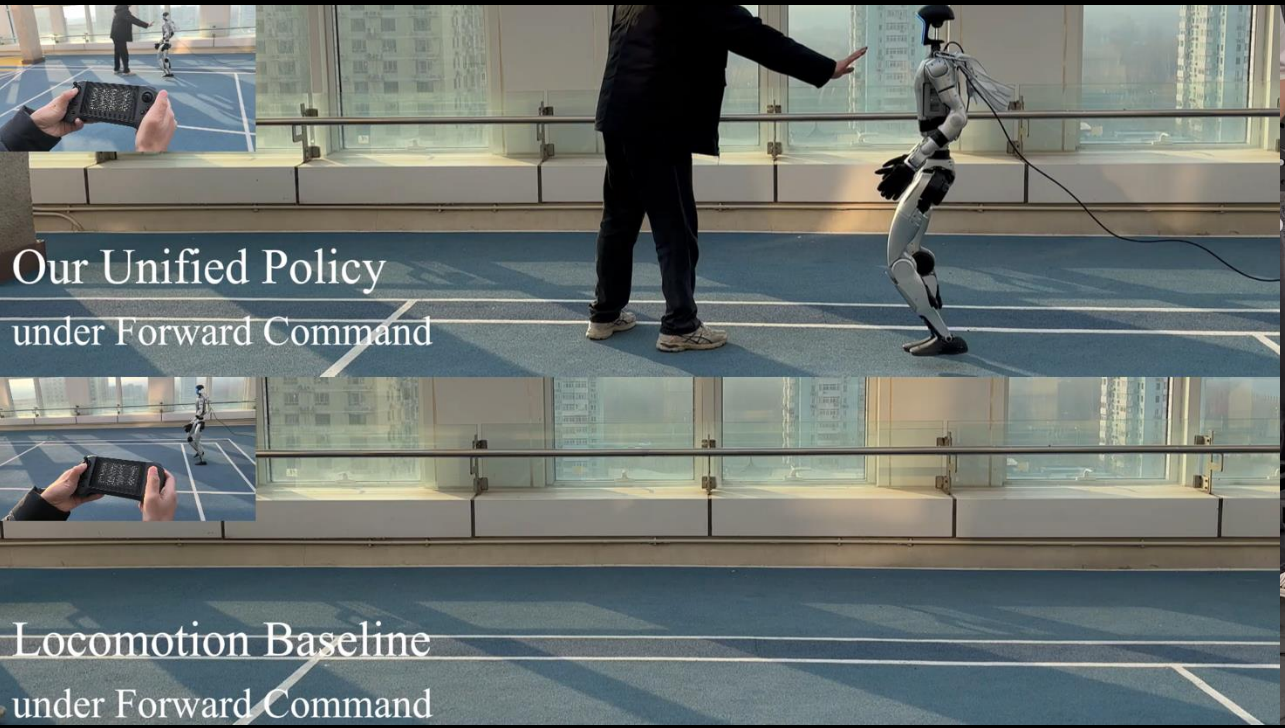
Force-adjusted velocity enables compliant locomotion

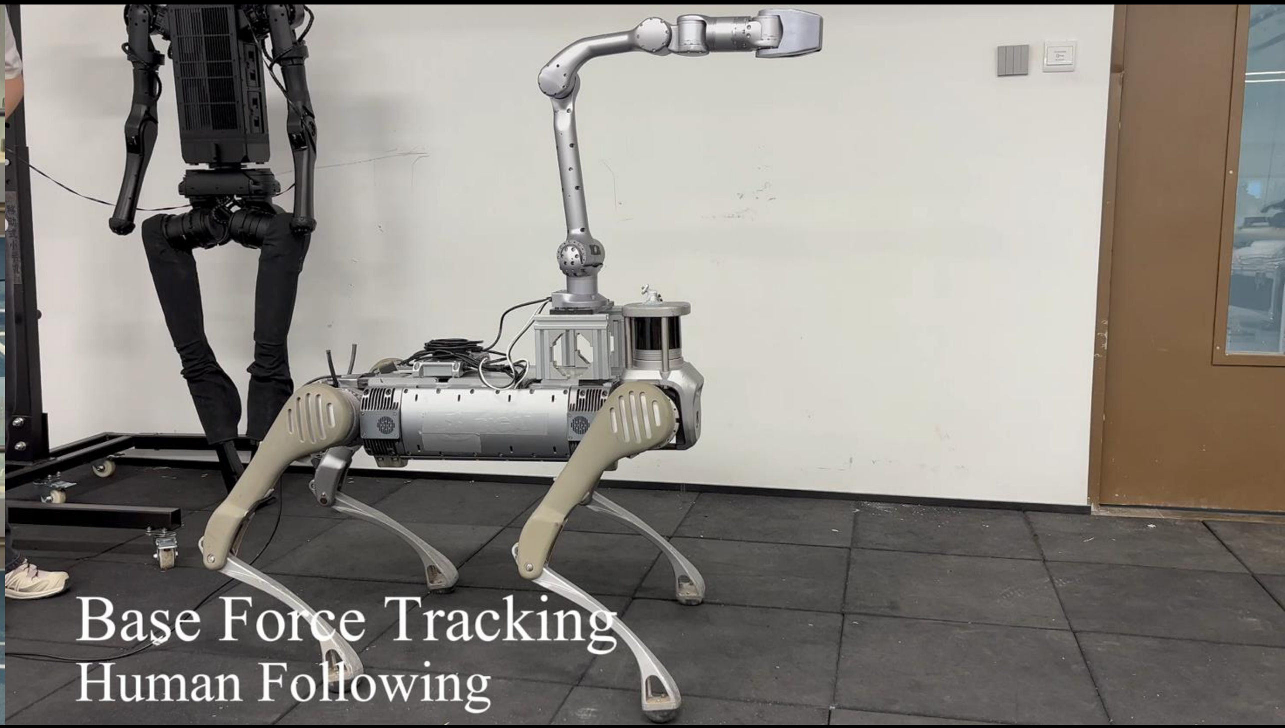
Formulating forces with velocities

$$\boldsymbol{F} = D\left(\dot{\mathrm{x}} - \dot{\mathrm{x}}^{\mathrm{cmd}}\right)$$

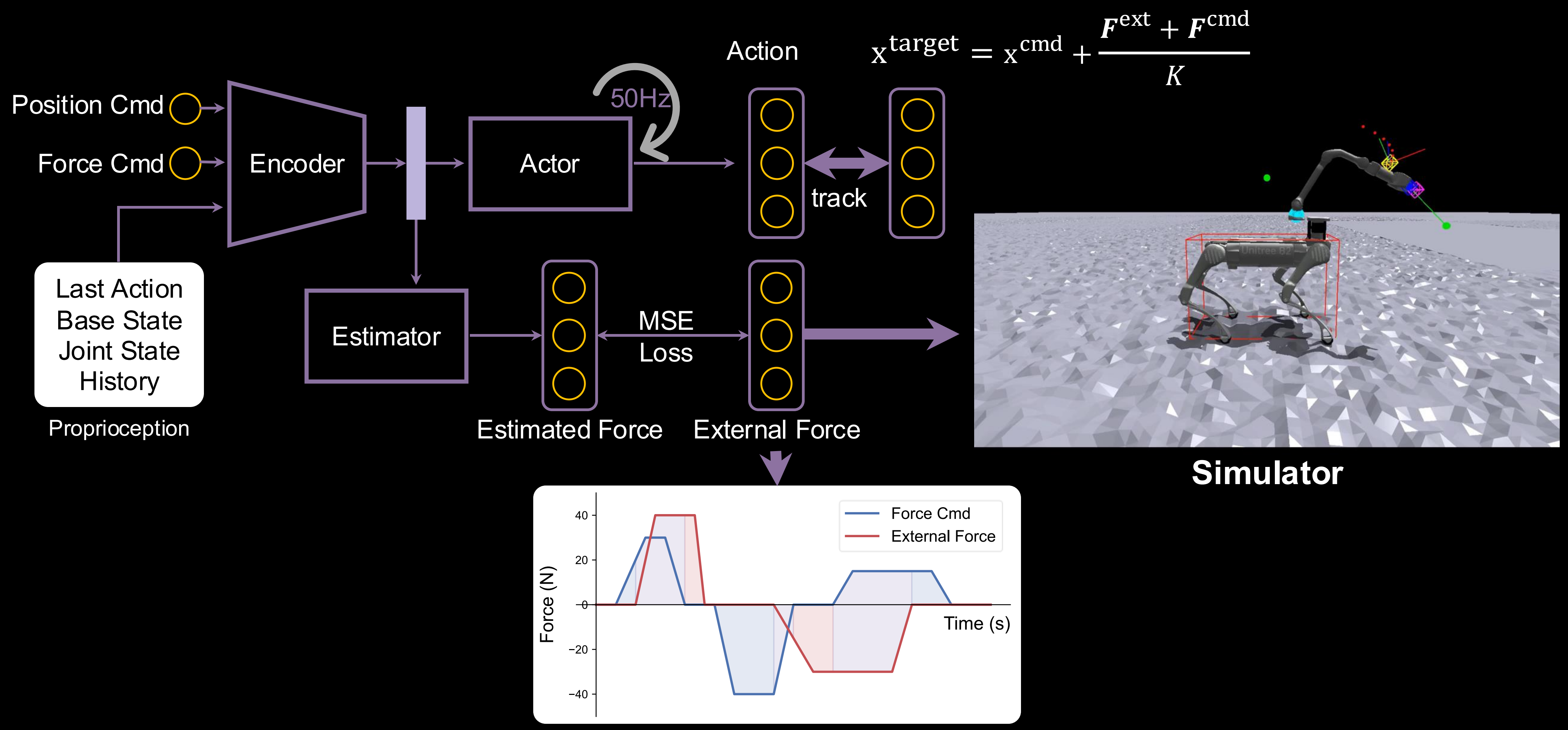
Compliant locomotion

$$\dot{\mathrm{x}}^{\mathrm{target}} = \dot{\mathrm{x}}^{\mathrm{cmd}} + \frac{F^{\mathrm{ext}}}{D}$$

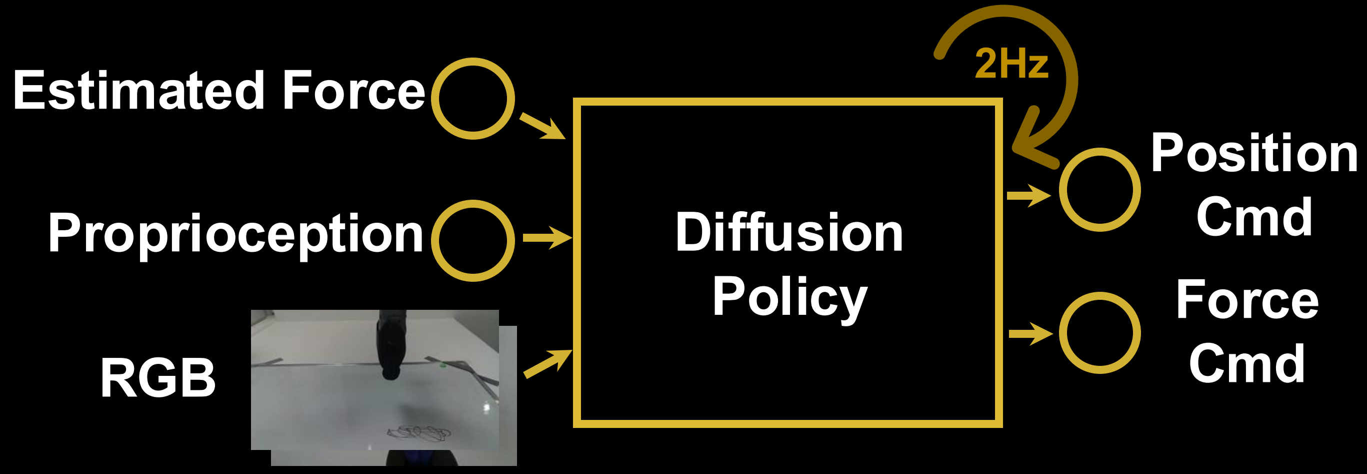


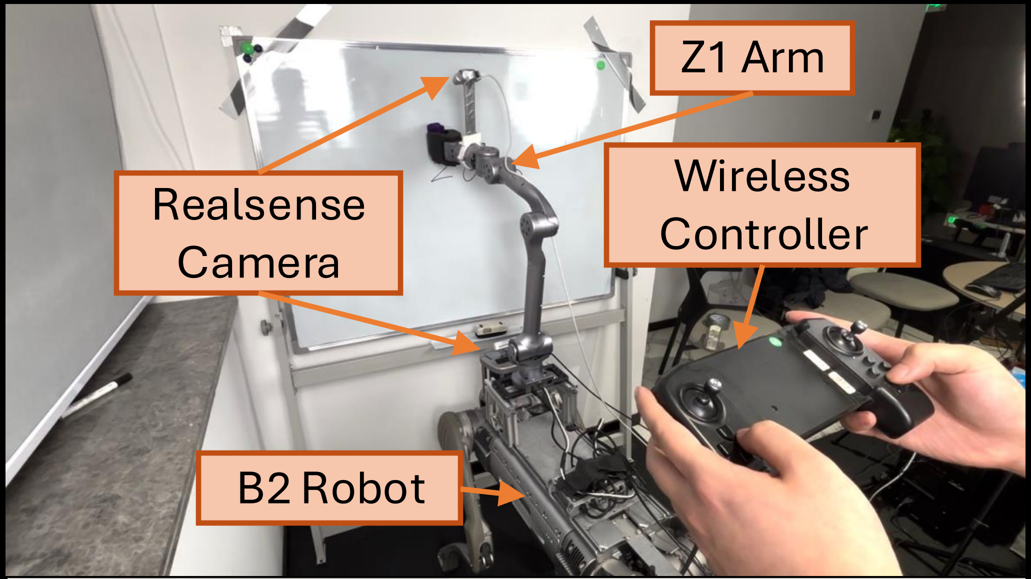


UniFP via RL with force-position sampling in simulator



UniFP for force-aware real-world imitation learning





- Data collection with estimated forces
- Imitation learning with position and force command targets
- Inference with UniFP

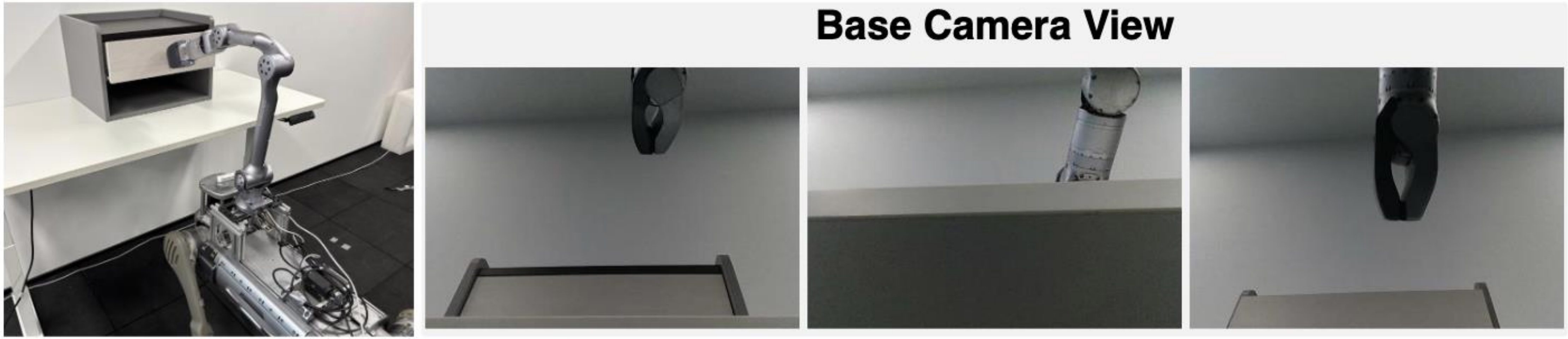
UniFP for force-aware real-world imitation learning

Tested on 4 tasks with each task taking 50 demonstrations

UniFP for force-aware real-world imitation learning

Table A.3: Imitation learning results (**50 trials per task**)

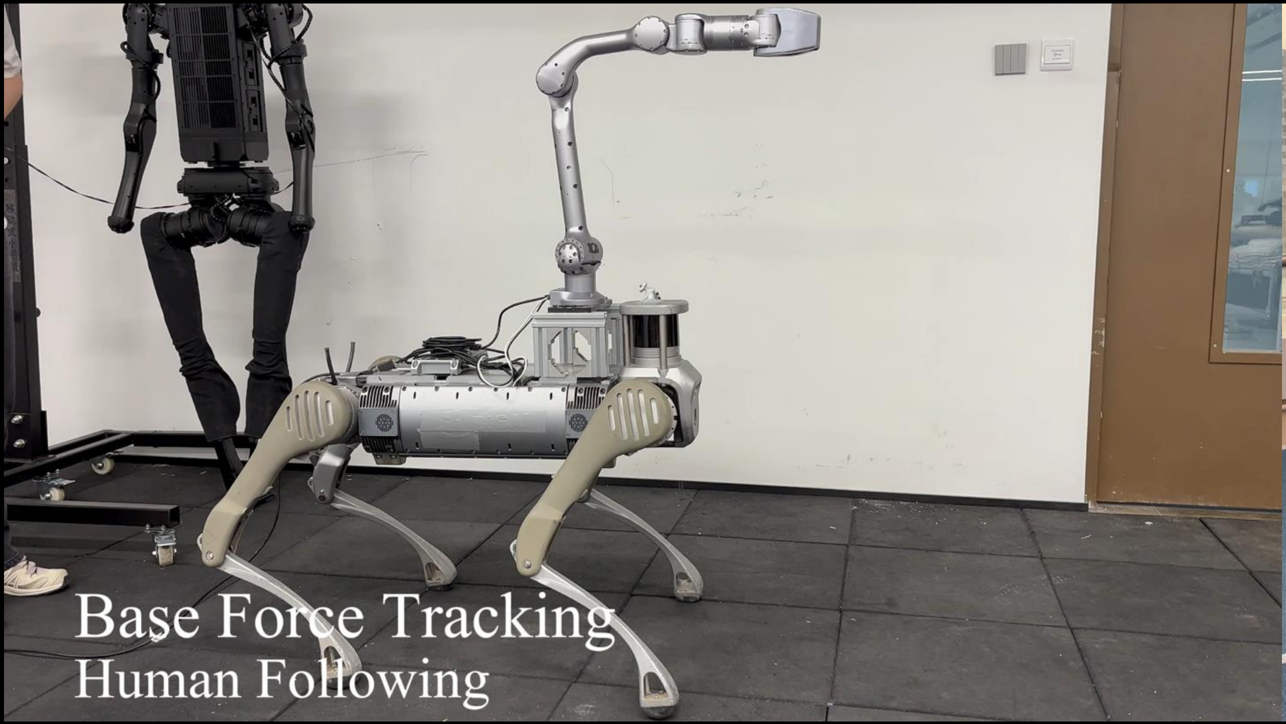
Task	wipe-blackboard	open-cabinet	close-cabinet	open-drawer-occlusion
w/o Force	0.22	0.36	0.30	0.30
w/ Force	0.58	0.70	0.72	0.76



Achieves ~39.5% higher success rate than the vanilla DP policy


Robot Exercising in Gym
Kettlebell Pulling
Tug-of-war
Throw the trash

So how is this important…



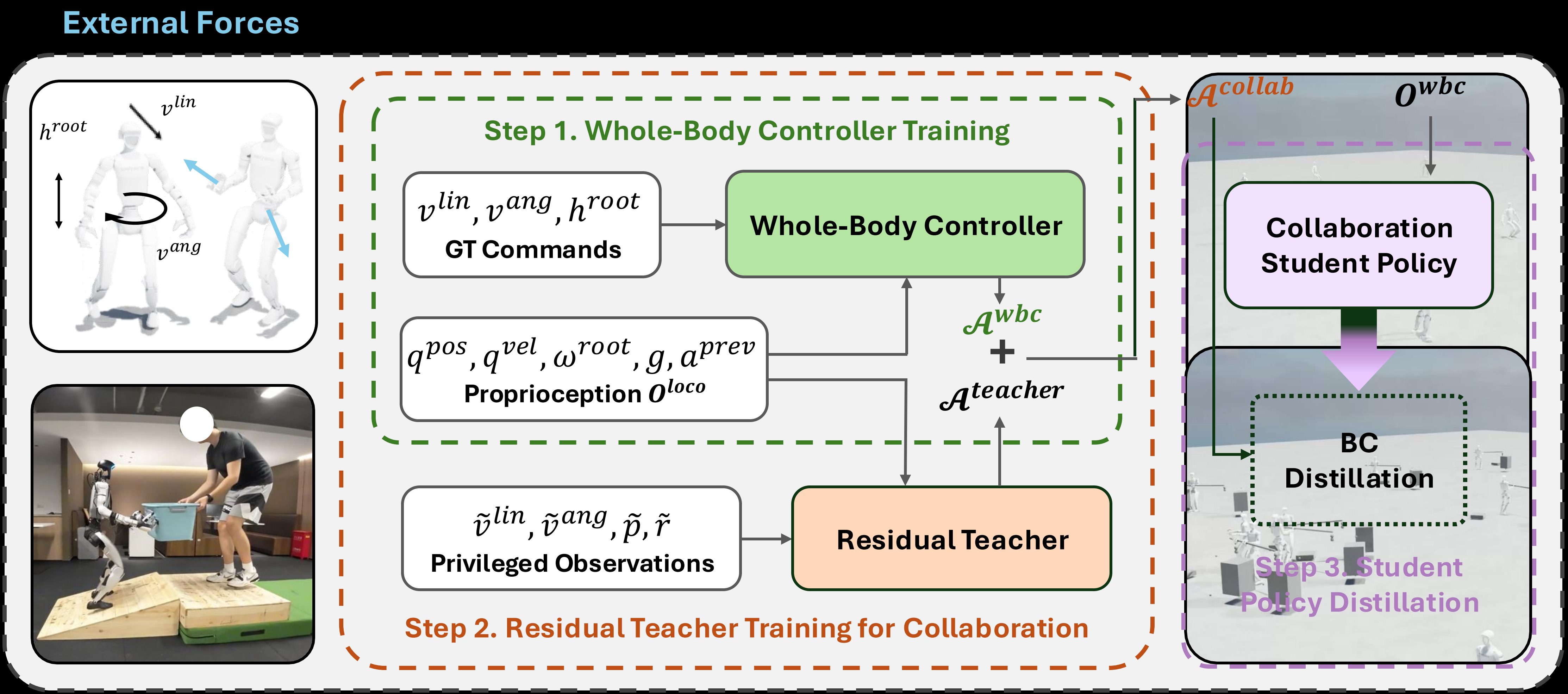
Movement Tracking

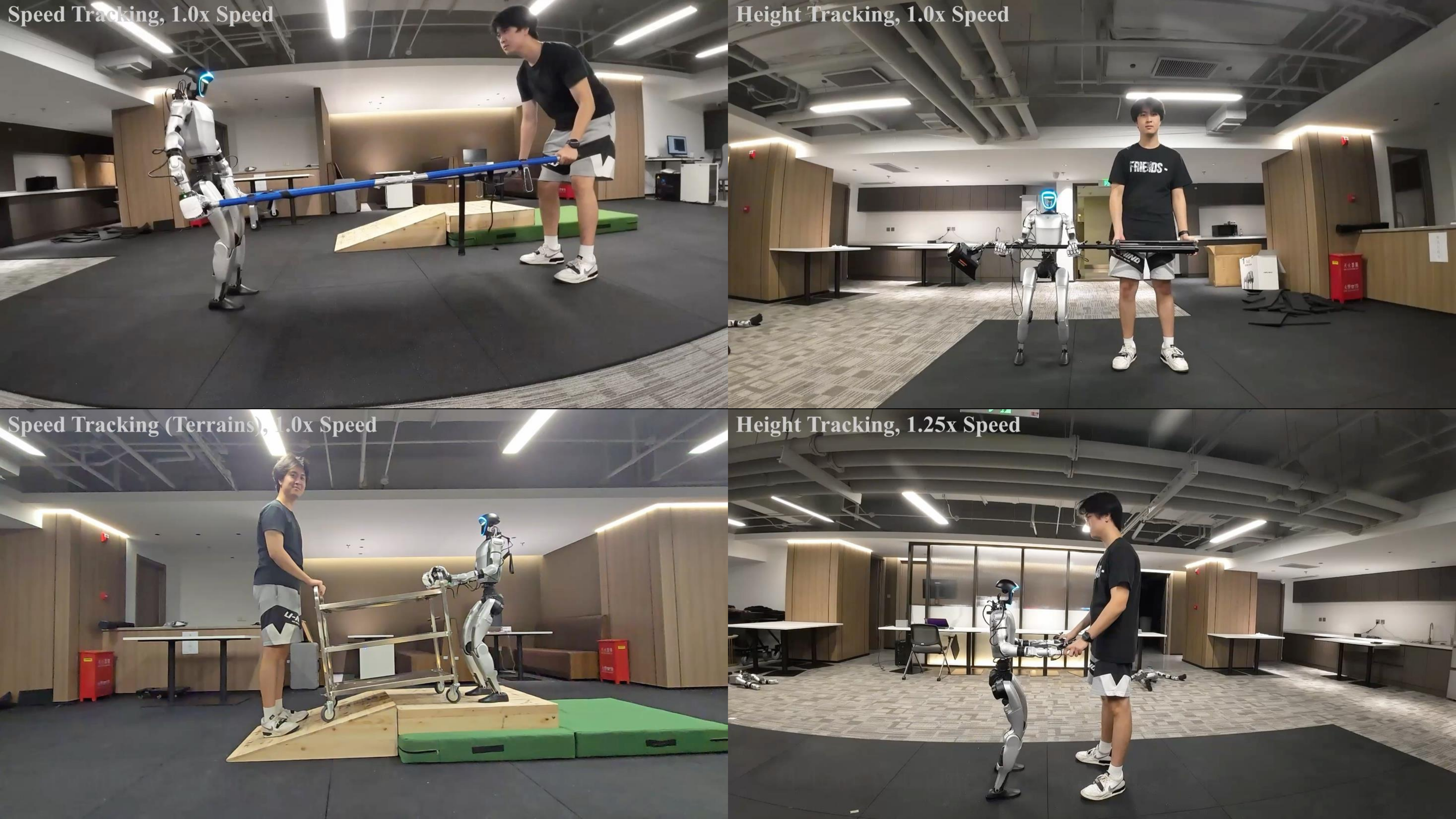
Compliant Holding

The composition of these behaviors works for human-robot collaboration


EXIT

COLA for collaborative object carrying




Speed Tracking, 1.0x Speed
Height Tracking, 1.0x Speed
Speed Tracking (Terrains), 1.0x Speed
Height Tracking, 1.25x Speed


Long Distance Testing (102.4m Total), 1.0x Speed


Speed Tracking, 1.0x Speed
'Shopping Assistant' with Payload (15kg)

Summary & Takeaways

- ➢ Data scaling and unification as the main obstacle of EAI
    - ❖ Different embodiments, low-cost hardwares
    - ❖ Real2Sim2Real, synthetic augmentation, world models etc.

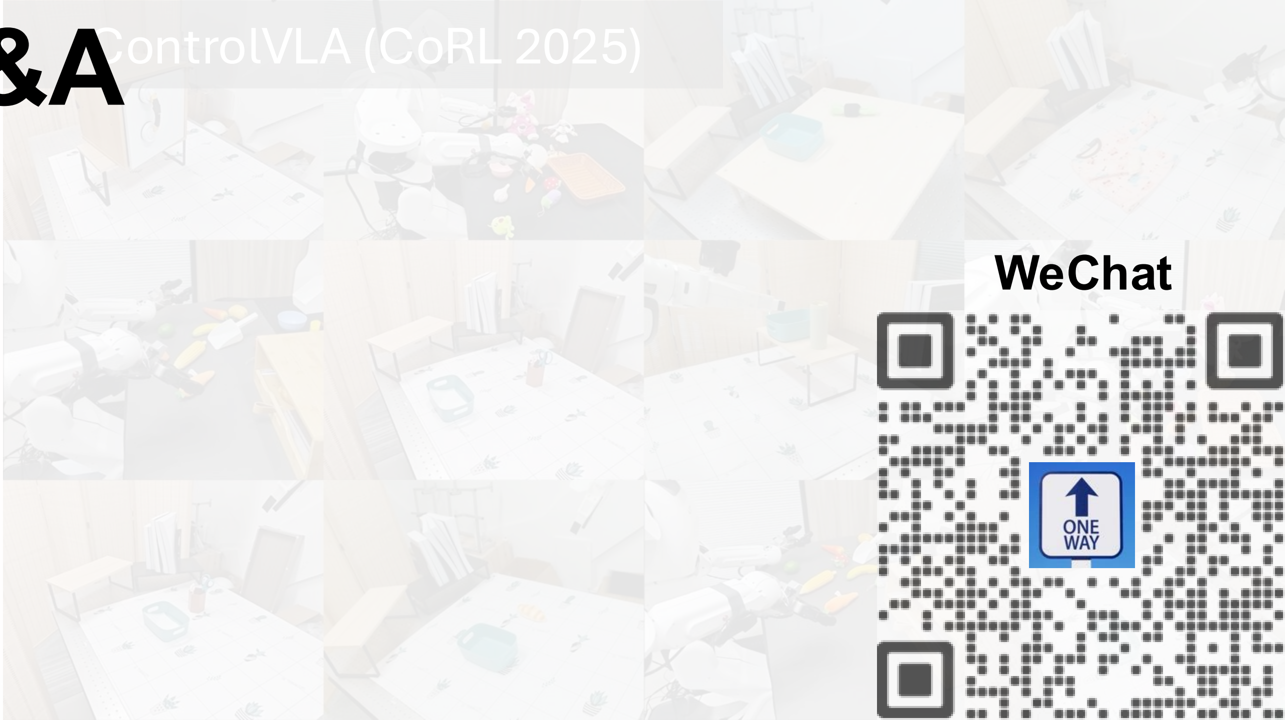
Summary & Takeaways

- Data scaling and unification as the main obstacle of EAI
  - Different embodiments, low-cost hardwares
  - Real2Sim2Real, synthetic augmentation, world models etc.
- General reasoning and acting capabilities for robot tasks
  - Aligning MLLMs for planning and interaction, efficient representations
  - Injecting spatial understanding capabilities for VLA models

Summary & Takeaways

- ➢ Data scaling and unification as the main obstacle of EAI
  - ❖ Different embodiments, low-cost hardwares
  - ❖ Real2Sim2Real, synthetic augmentation, world models etc.
- ➢ General reasoning and acting capabilities for robot tasks
  - ❖ Aligning MLLMs for planning and interaction, efficient representations
  - ❖ Injecting spatial understanding capabilities for VLA models
- ➢ Agile and safe robot control for human-robot interaction
  - ❖ Recover the missing force modality for compliance policies
  - ❖ Safe control behaviors over VLA for human-robot interaction

Thank you
Q&A



Webpage

WeChat