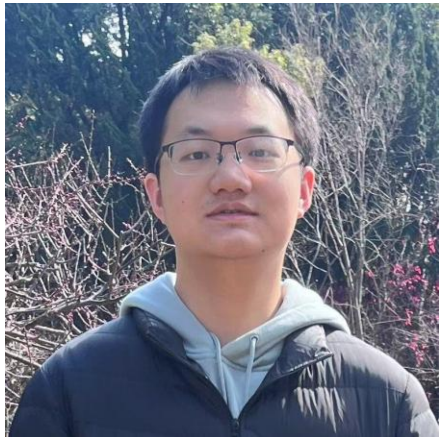
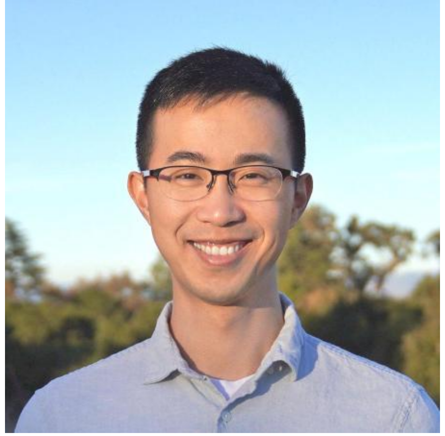
VisualMimic:

Visual Humanoid Loco-Manipulation via Motion Tracking and Generation





Shaofeng Yin* Yanjie Ze* Koven Yu C. Karen Liu† Jiajun Wu†

*Contributed Equally † Advised Equally

Stanford University

Motivation: Visual Humanoid Control

Project Goal: Develop a general* sim-to-real visual whole-body control framework for humanoid loco-manipulation.

* general: capable of performing diverse tasks; easy to train and add new tasks

Motivation: Visual Humanoid Control

Project Goal: Develop a general* sim-to-real visual whole-body control framework for humanoid loco-manipulation.

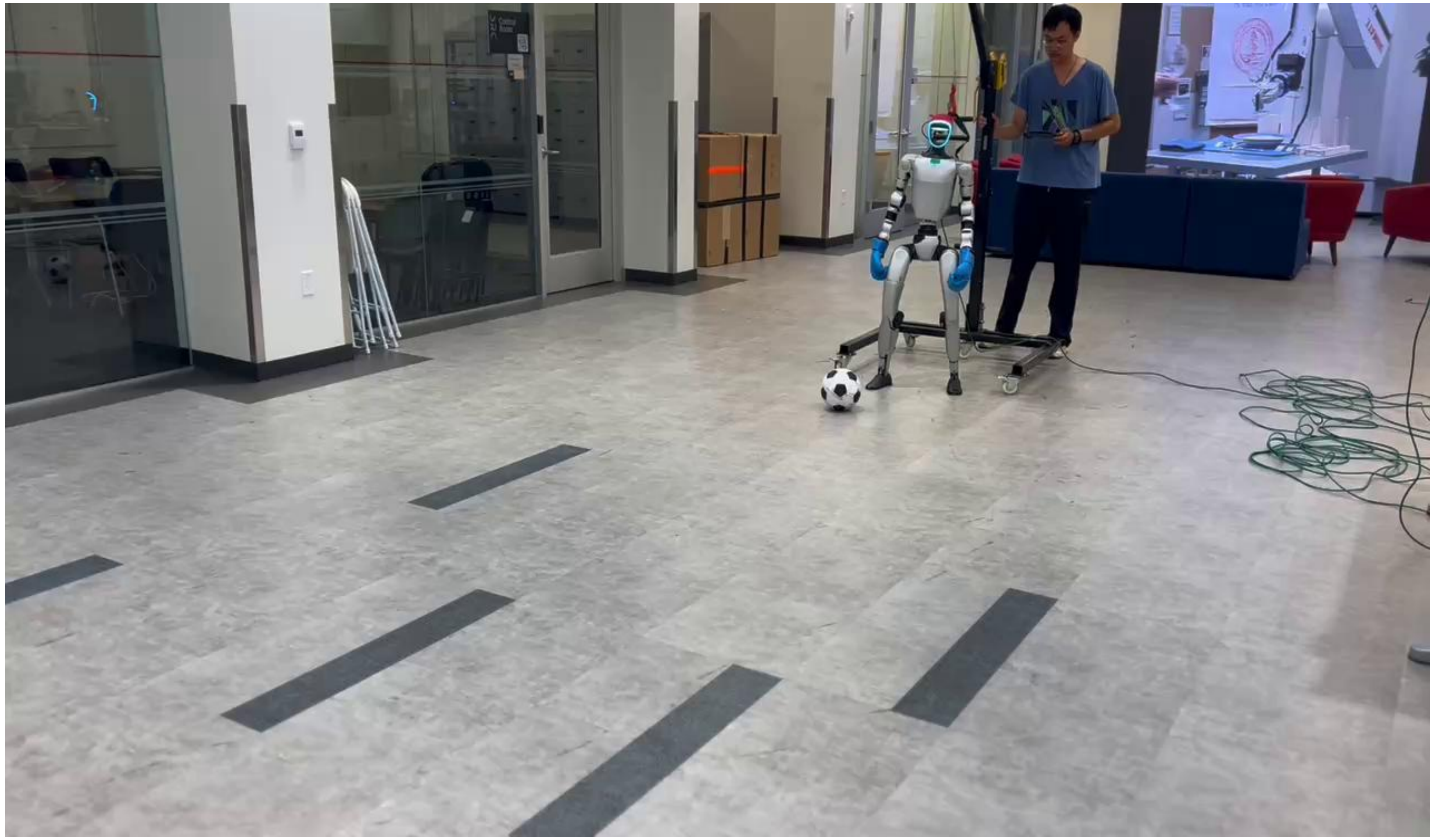
* general: capable of performing diverse tasks; easy to train and add new tasks

Motivation: Visual Humanoid Control

Project Goal: Develop a general* sim-to-real visual whole-body control framework for humanoid loco-manipulation.

* general: capable of performing diverse tasks; easy to train and add new tasks

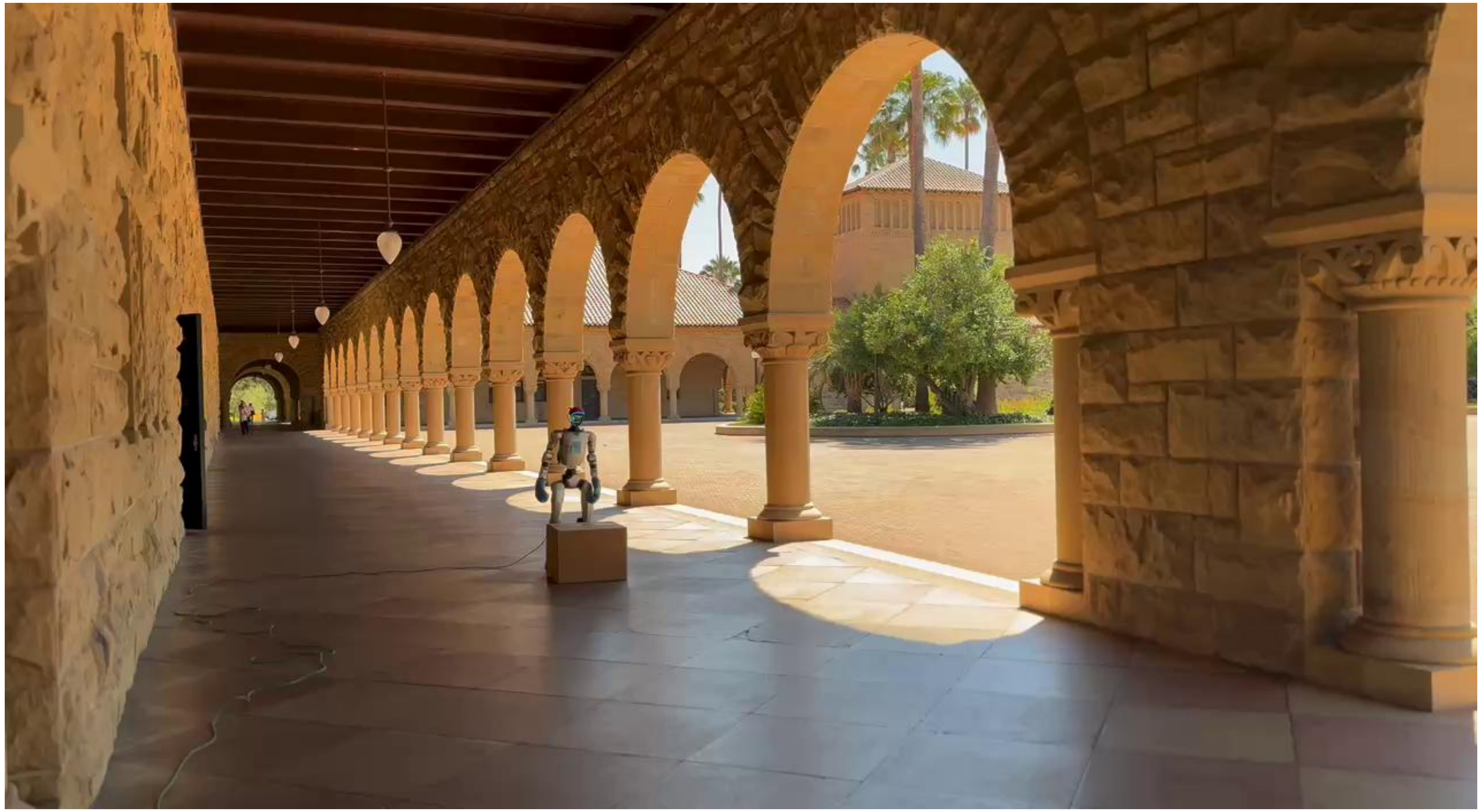
Motivation: Visual Humanoid Control



Project Goal: Develop a general* sim-to-real visual whole-body control framework for humanoid loco-manipulation.

* general: capable of performing diverse tasks; easy to train and add new tasks

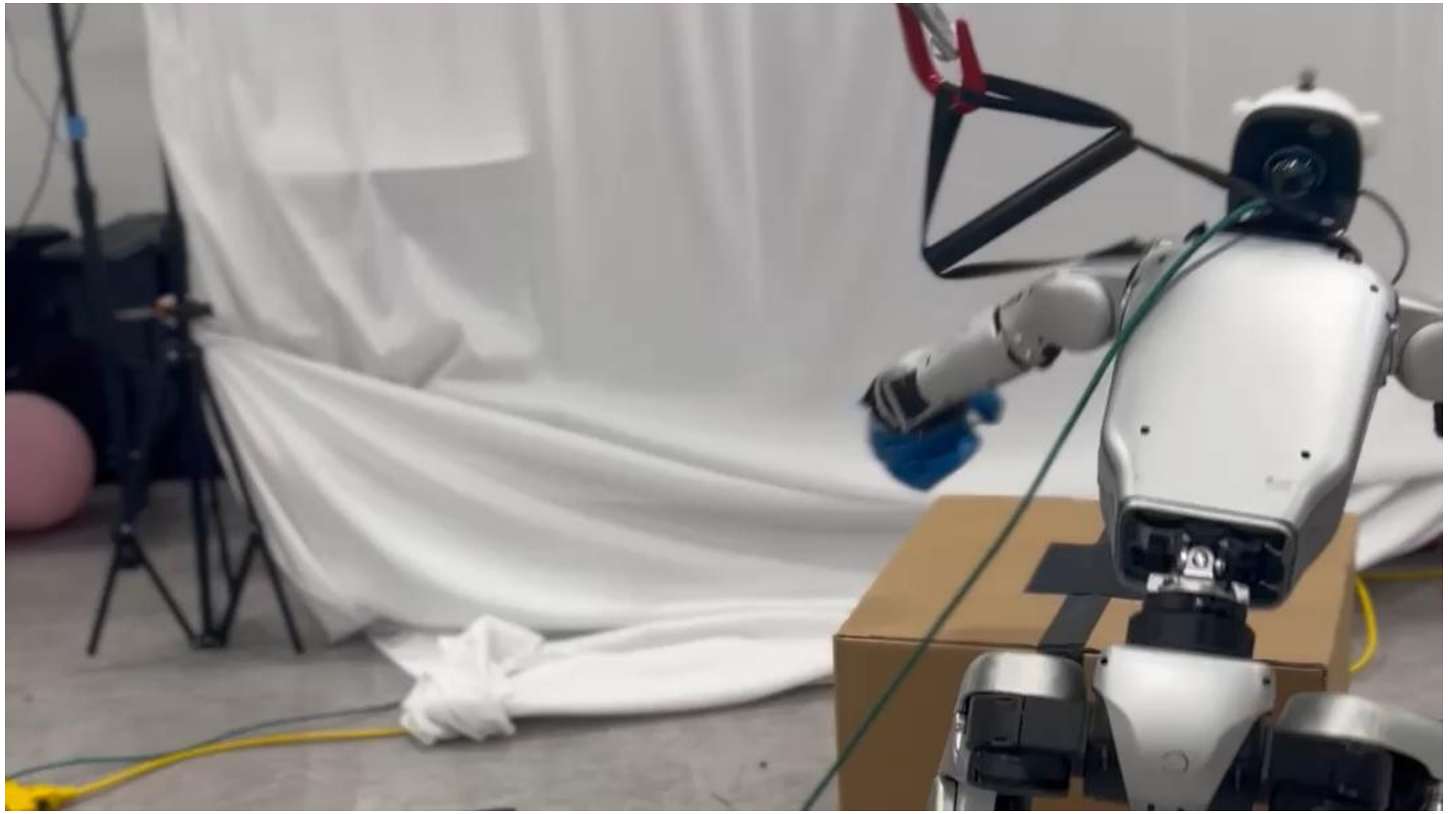
Motivation: Visual Humanoid Control



Project Goal: Develop a general* sim-to-real visual whole-body control framework for humanoid loco-manipulation.

* general: capable of performing diverse tasks; easy to train and add new tasks

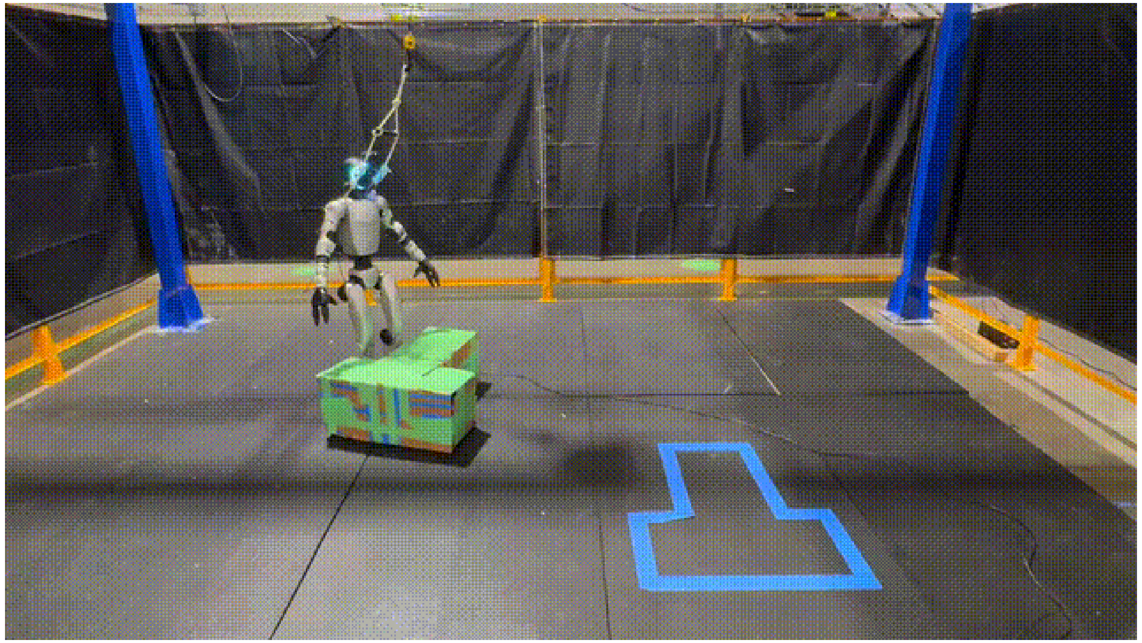
Motivation: Robust Visual Humanoid Control

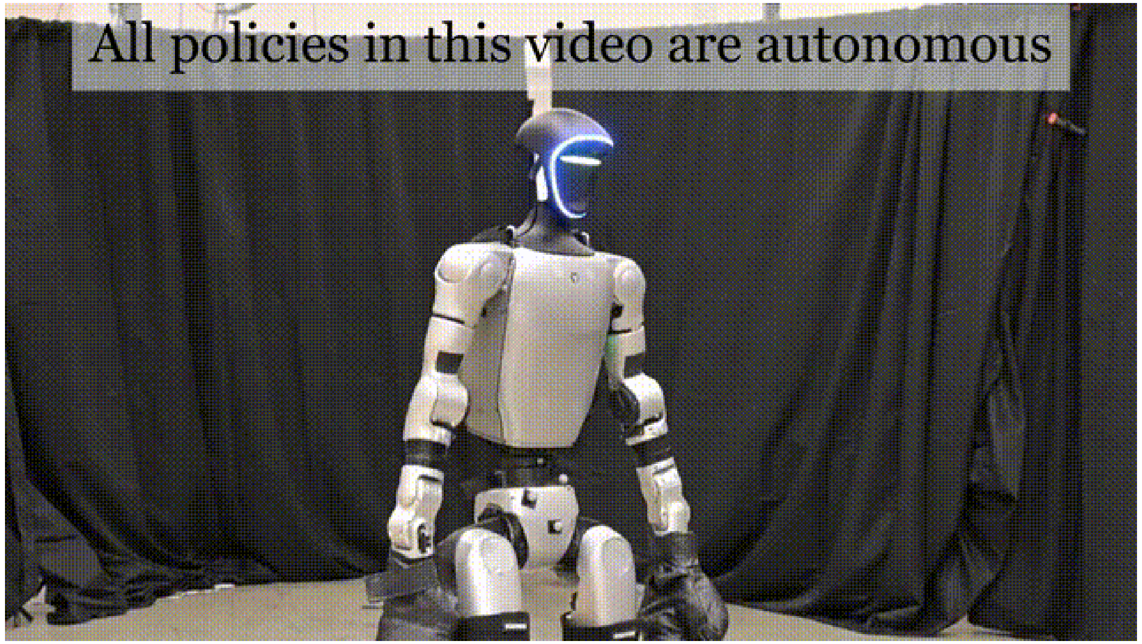


Project Goal: Develop a general* sim-to-real visual whole-body control framework for humanoid loco-manipulation.

* general: capable of performing diverse tasks; easy to train and add new tasks

Motivation: Robust Visual Humanoid Control





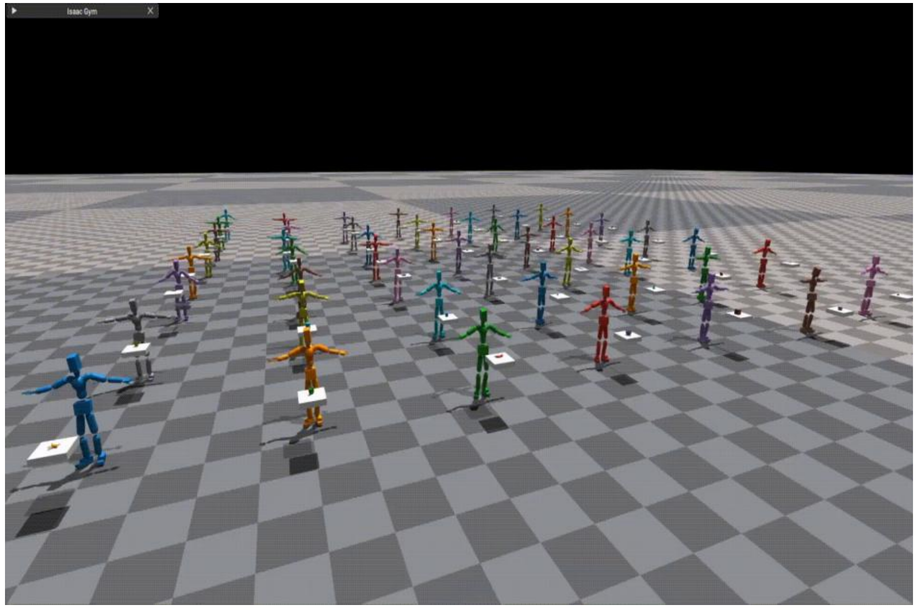
Other Approaches

Imitation Learning

"Imitation Learning"

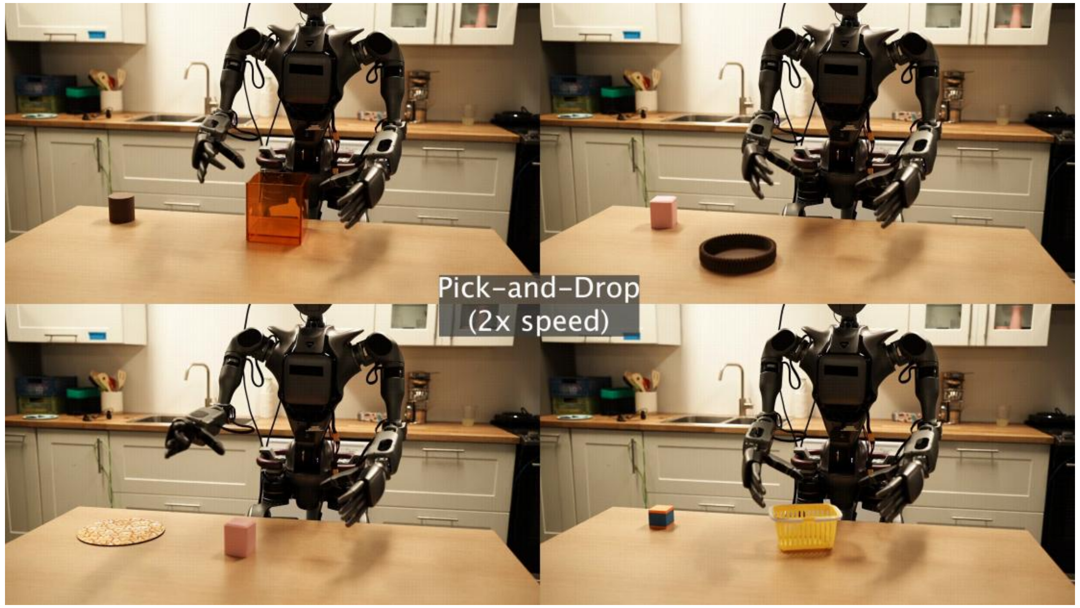
TWIST2: Scalable, Portable, and Holistic Humanoid Data Collection System, Ze et al., 2025
HDMI: Learning Interactive Humanoid Whole-Body Control from Human Videos, Weng et al., 2025

Two Paradims in Humanoid Policy Training



train end-to-end policy
with task-specific human motions

❌ need task-relevant motion



trained end-to-end policy
with complex RL reward design

❌ complex reward design

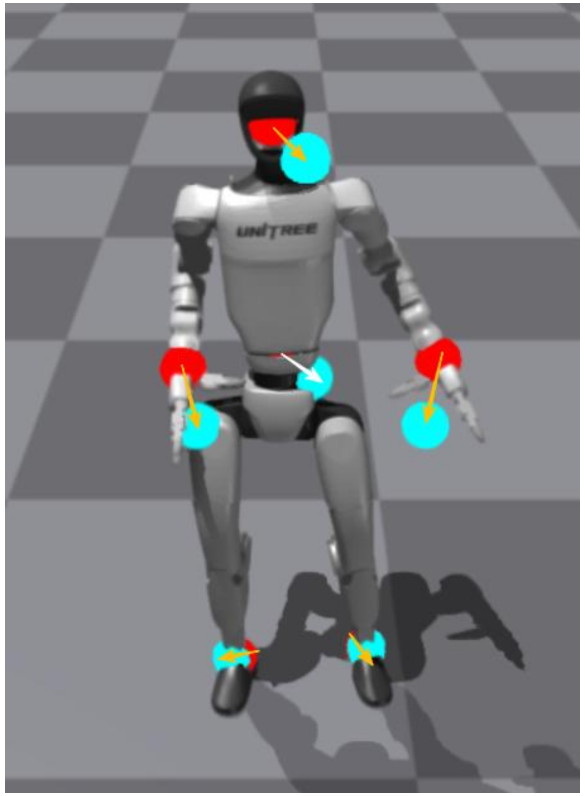
Our Project

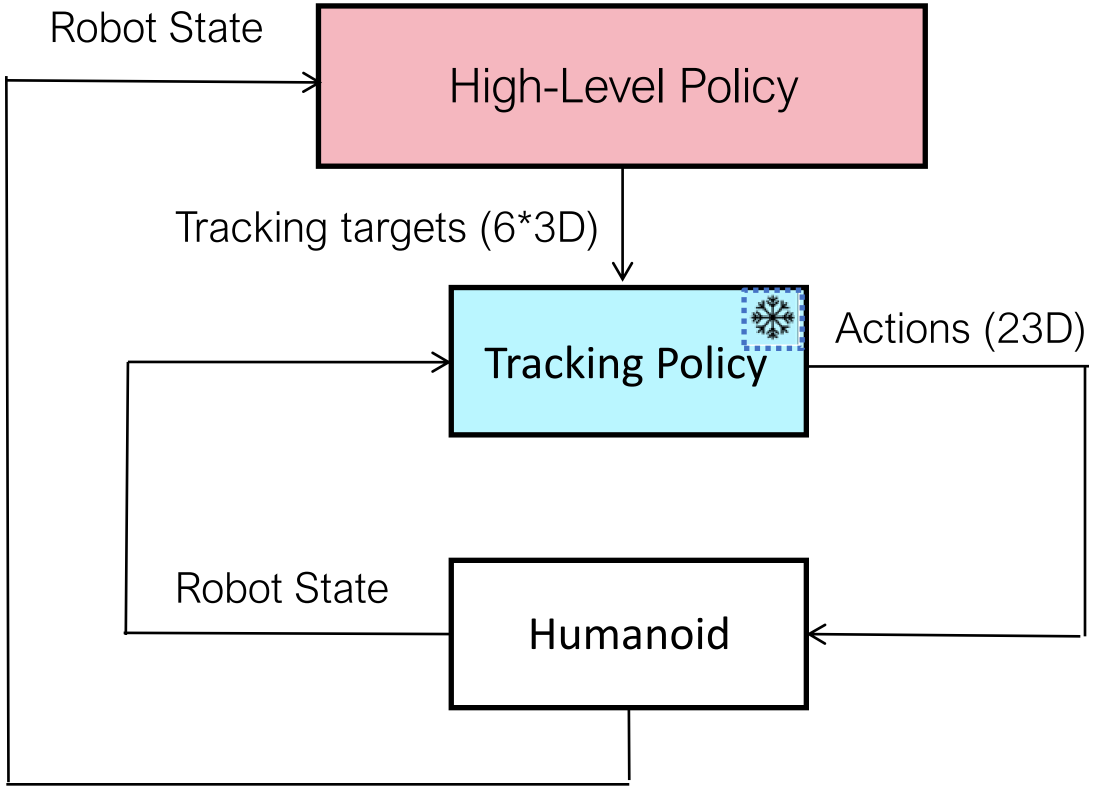
train a low-level policy with task-agnostic human motions
train high-level policies with simple RL reward

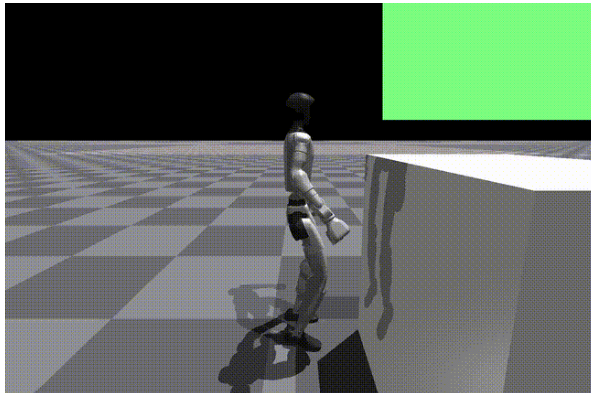
✅ no need for task-relevant motion
✅ easy reward design
✅ quickly adapt to a new task

Omnigrasp: Grasping Diverse Objects with Simulated Humanoids, Luo et al., 2024
Sim-to-Real Reinforcement Learning for Vision-Based Dexterous Manipulation on Humanoids, Lin et al., 2025

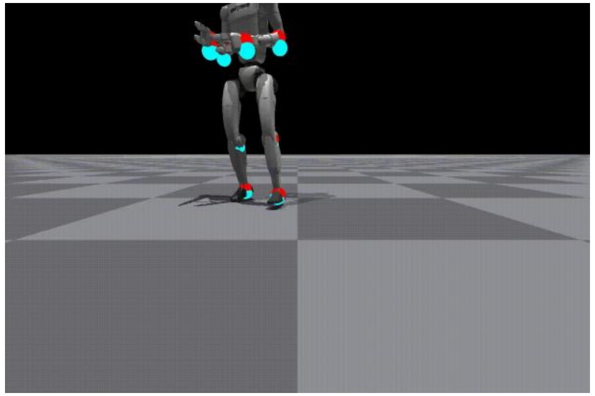
Method: Hierarchical Framework







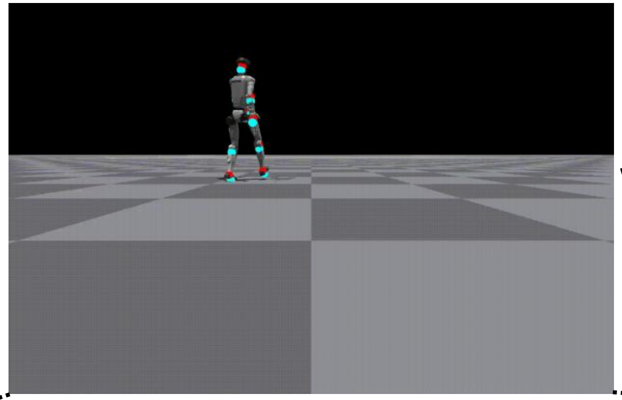
High-Level: Keypoint Generator



Low-level: General Keypoint Tracker

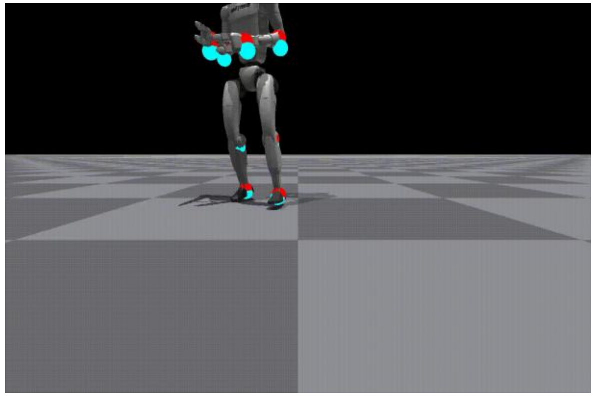
A hierarchical framework with a flexible interface that enables the high-level policy to control all body parts

Low-Level: General Keypoint Tracker (Stage 0)

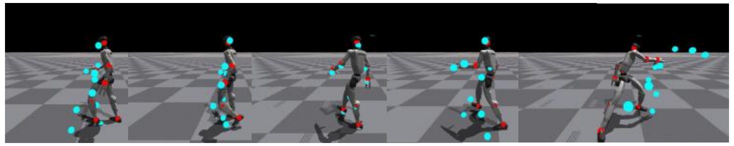


distill

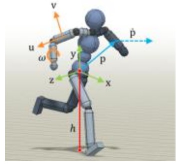
whole-body motion commands → 6 keypoints commands



(Stage 0)



20 future steps of target motion



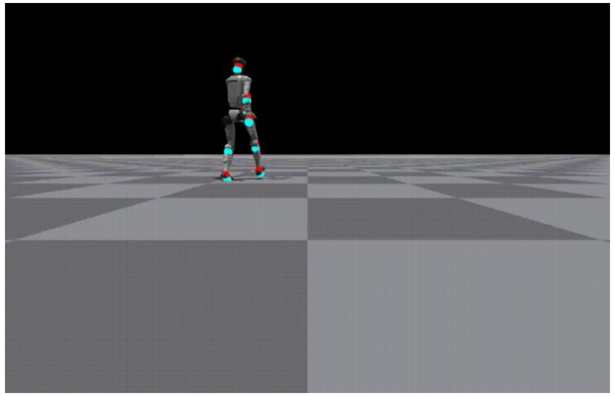
proprioceptive observation

General Keypoint Tracker

PD controller commands

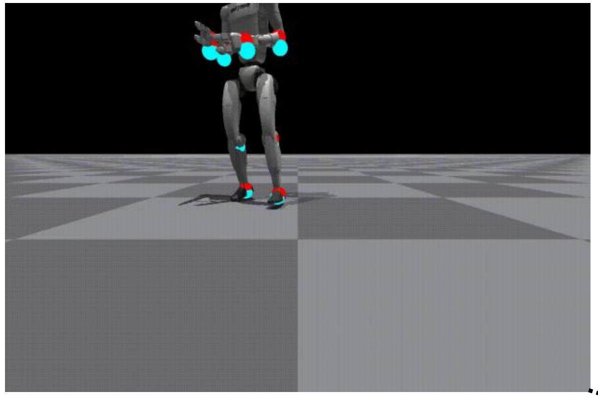
Learned through **RL** using motion tracking rewards

Low-Level: General Keypoint Tracker (Stage 1)

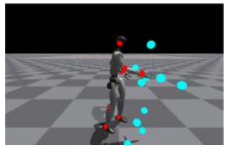


distill

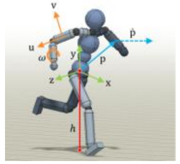
whole-body motion commands ➡ 6 keypoints commands



(Stage 1)



one future step of target motion (simplified command)

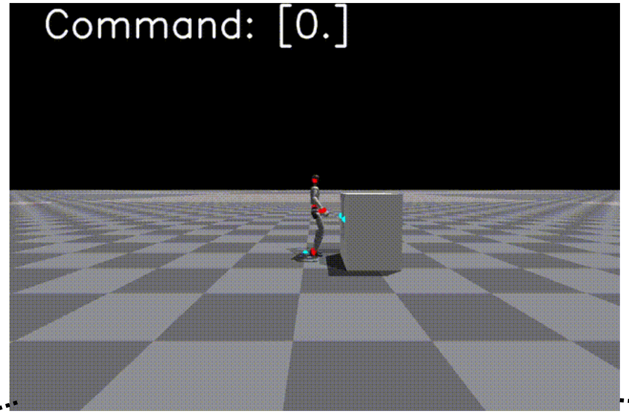


proprioceptive observation (**partial**)

→ General Keypoint Tracker → PD controller commands

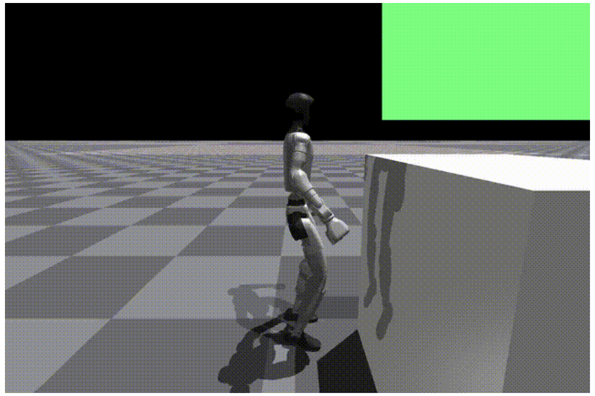
Learned through **BC** with DAgger

High-Level: Task-Specific Keypoint Generator (Stage 2)



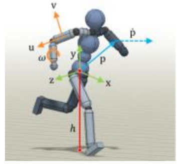
distill

state input → visual input



(Stage 2)

state-based task-relevant input
e.g. target point position
(10.0, 0.0, 0.8)

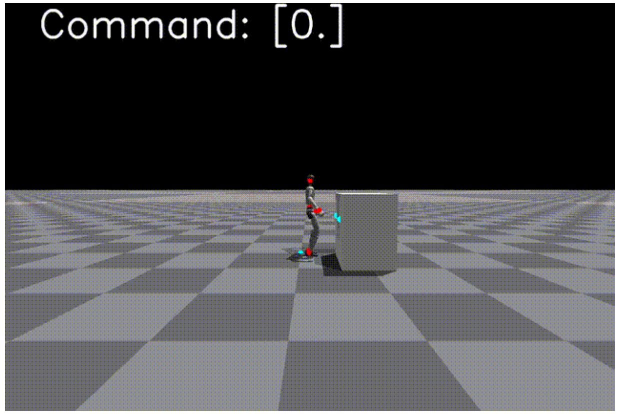


proprioceptive observation

→ Task-Specific Keypoint Generator → Tracking target for low-level

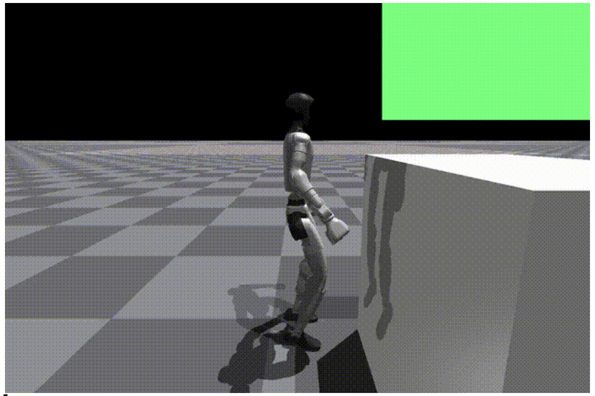
Learned through **RL** using task-specific rewards

High-Level: Task-Specific Keypoint Generator (Stage 3)



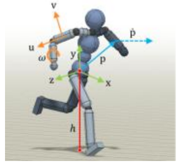
distill

state input → visual input



(Stage 3)

Depth image



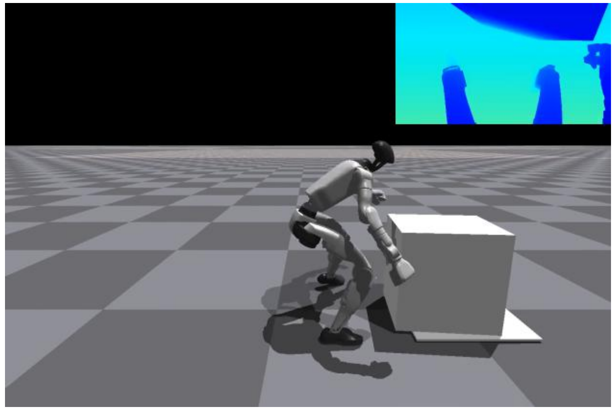
proprioceptive observation (partial)

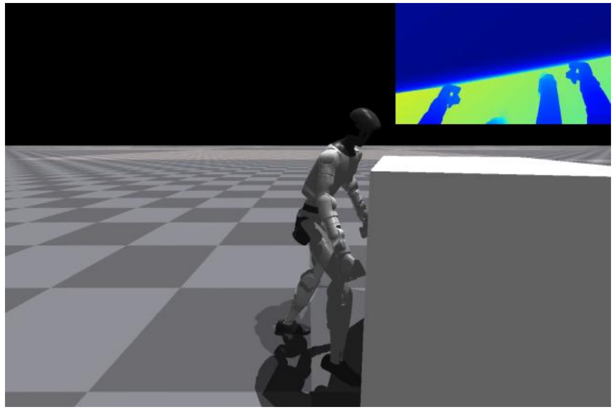
→ Task-Specific Keypoint Generator → Tracking target for low-level

Learn through **BC** with DAgger

High-Level: Reward Design

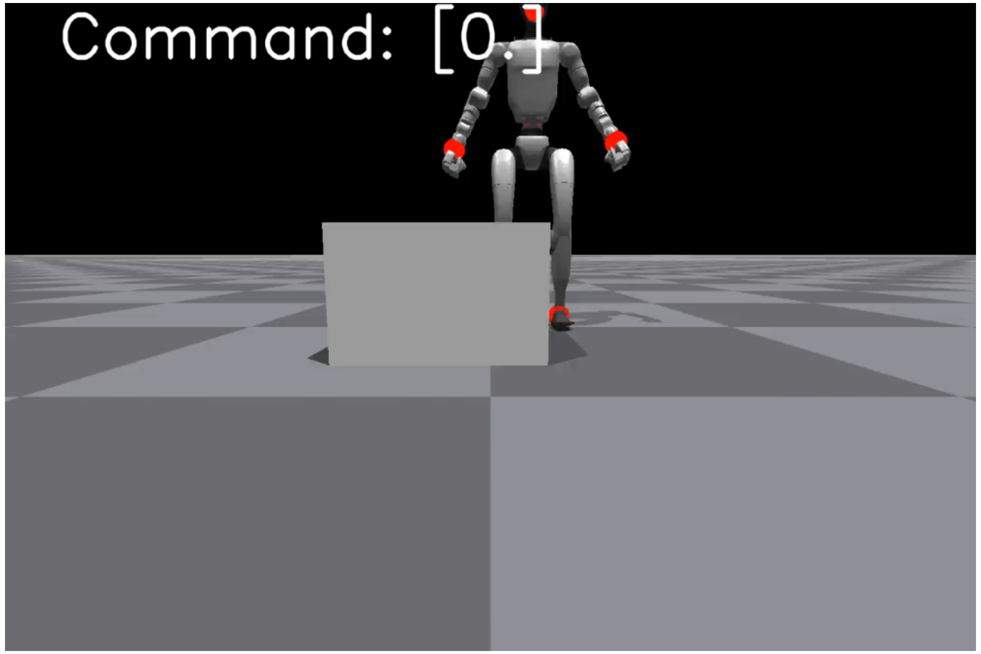
- Approach
  - $R_{\text{approach}}(t) = e^{-0.1d(t)}$
  - $R_{\text{approach}}(t) = \dfrac{2\,e^{-0.1d_1(t)}e^{-0.1d_2(t)}}{e^{-0.1d_1(t)} + e^{-0.1d_2(t)}}$
- Forward
  - $R_{\text{forward}}(t) = \tanh\Big(10[x_{\text{obj}}(t) - \max_{t'<t} x_{\text{obj}}(t')]_+\Big)$
- Force
  - $R_{\text{force}}(t) = e^{-0.1[F_{\text{des}} - F_{\text{obj}}(t)]_+}$



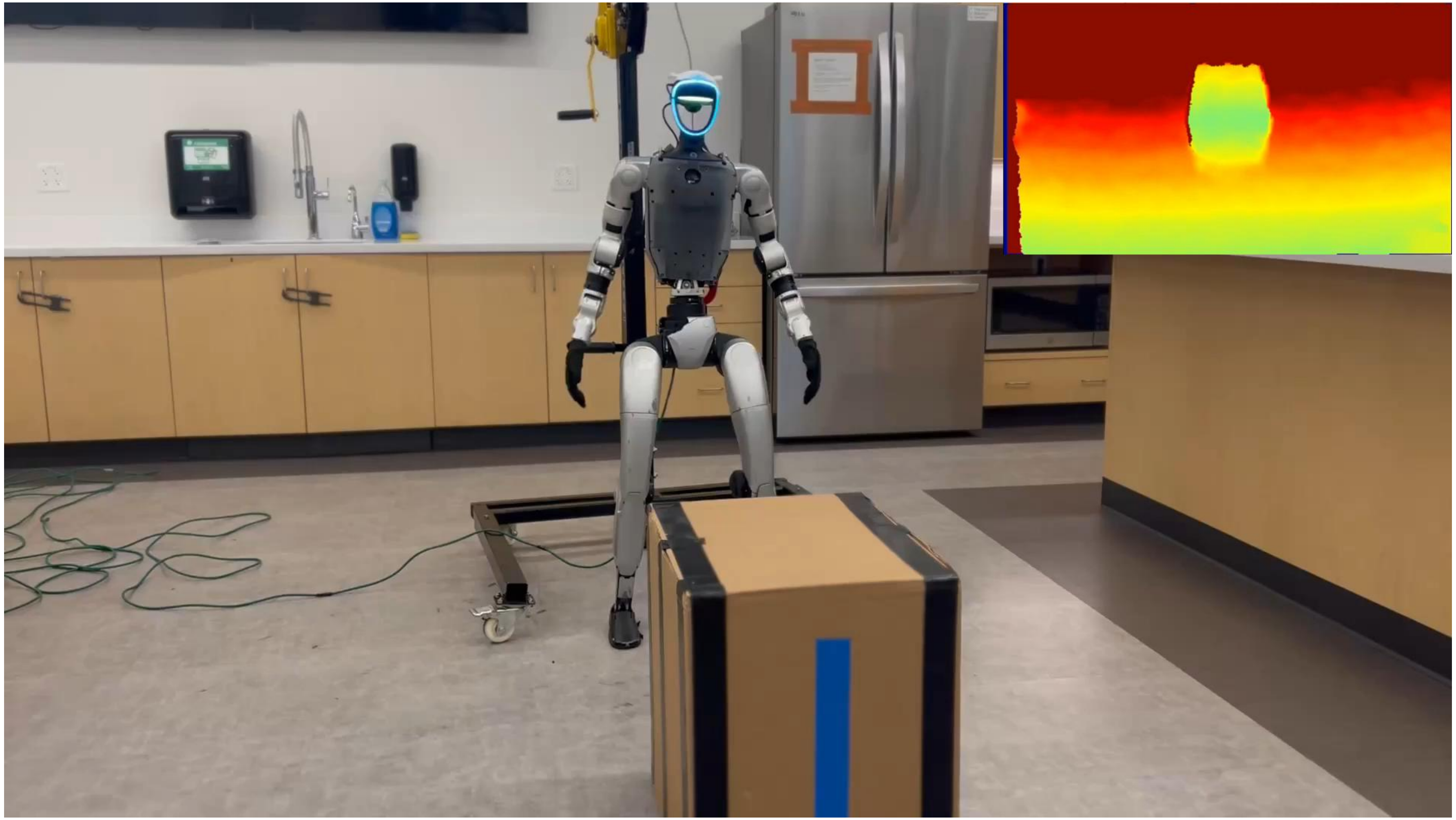


High-Level: Reward Design

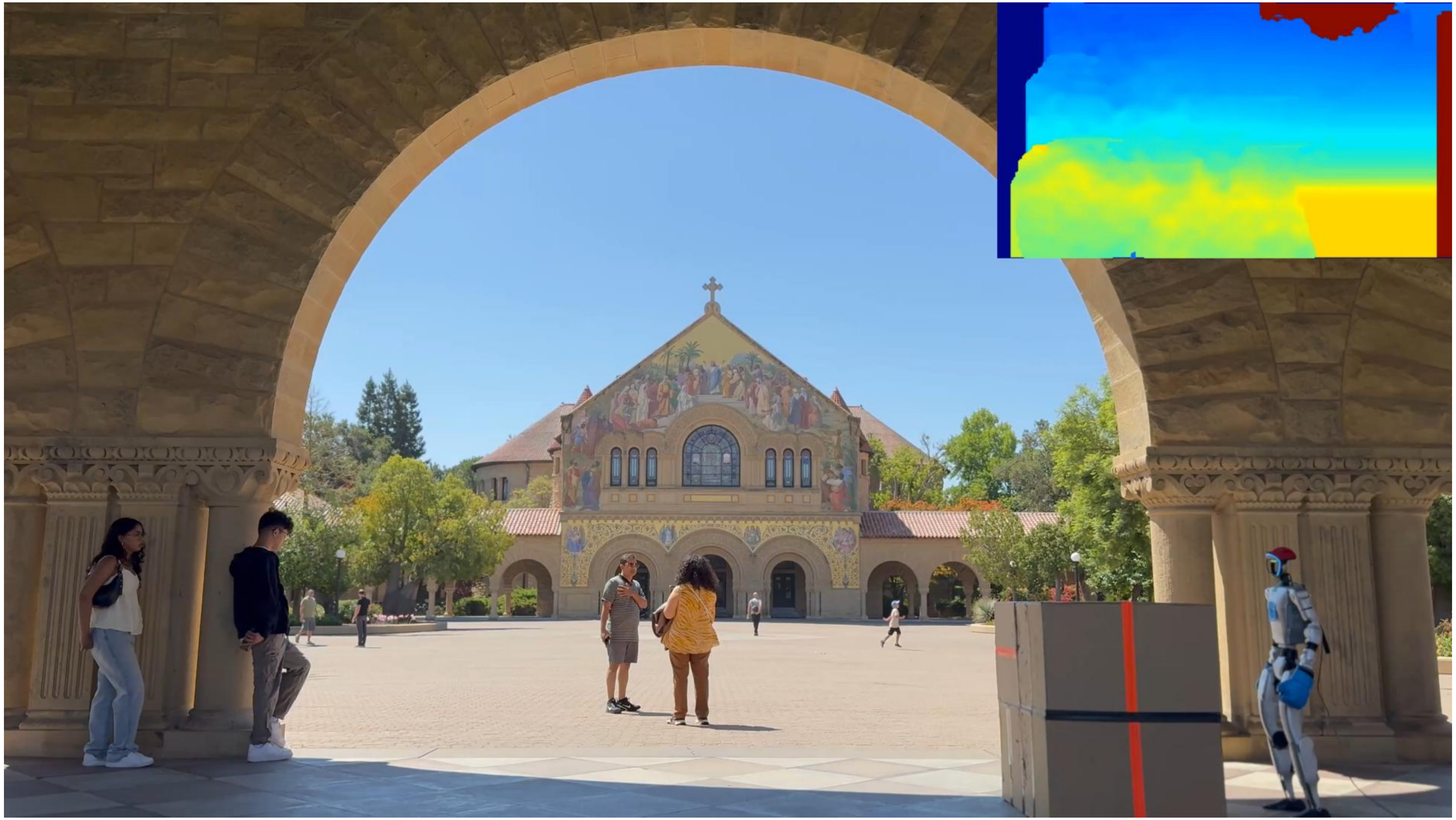
- Look at object
  - $R_{\text{look}}(t) = -\left(\arccos(\hat{\mathbf{f}}_{\text{body}} \cdot \hat{\mathbf{d}}_{\text{obj}})\right)^2$
- Drift
  - $R_{\text{drift}}(t) = -\tanh\Big(10\big[\,|y_{\text{obj}}(t)| - \max_{t'<t} |y_{\text{obj}}(t')|\,\big]_+\Big).$



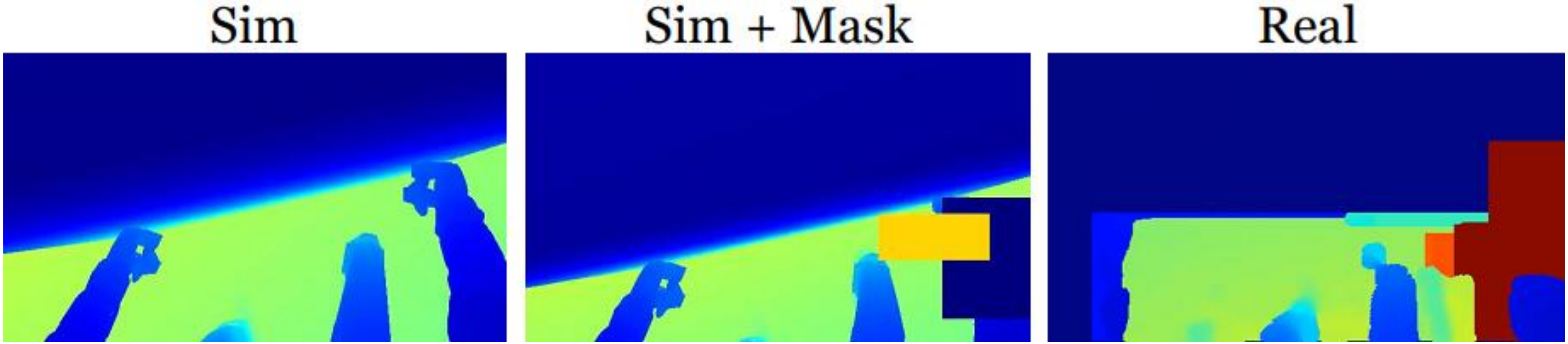
Real World: Visual Mask



Real World: Visual Mask

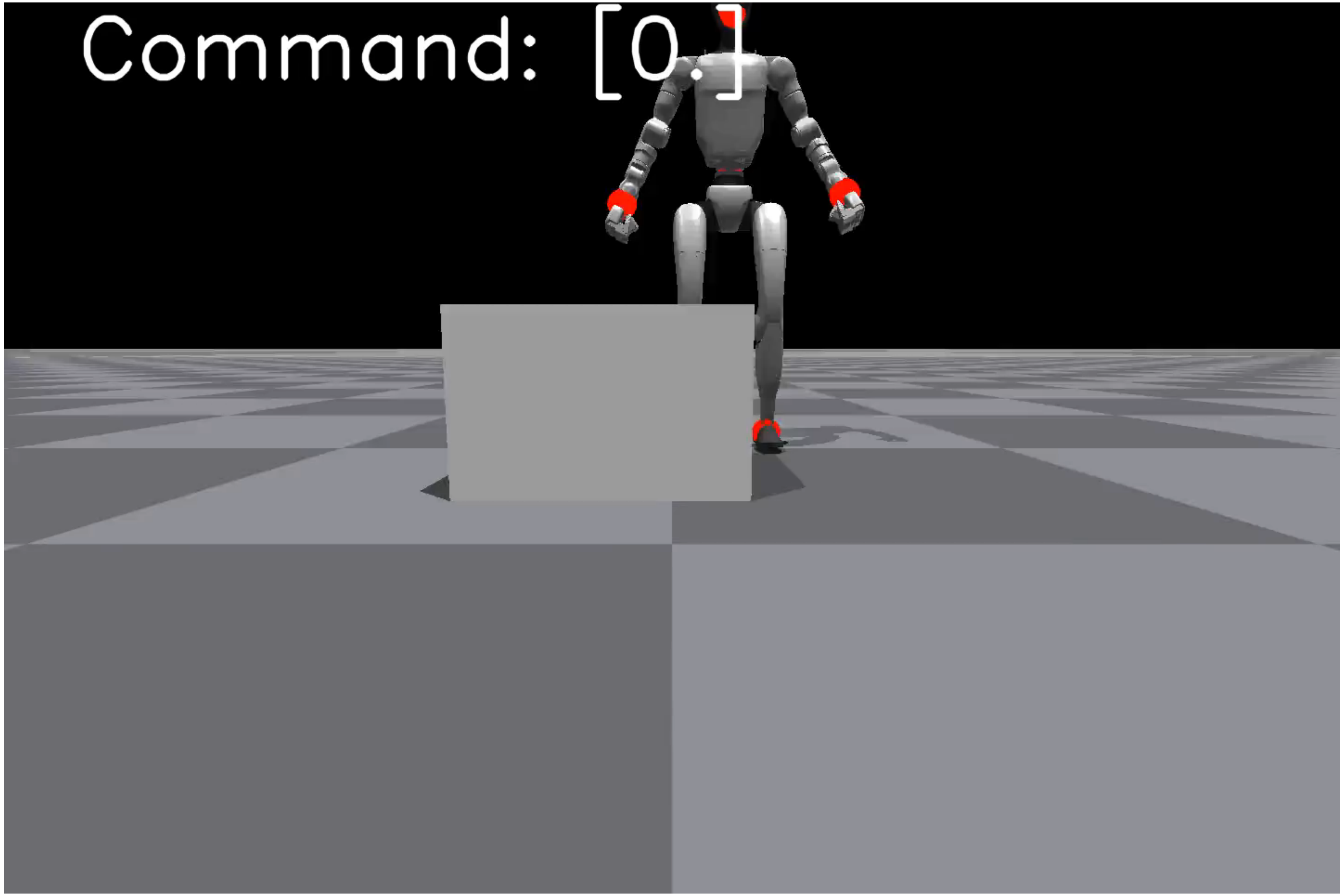


Real World: Visual Mask

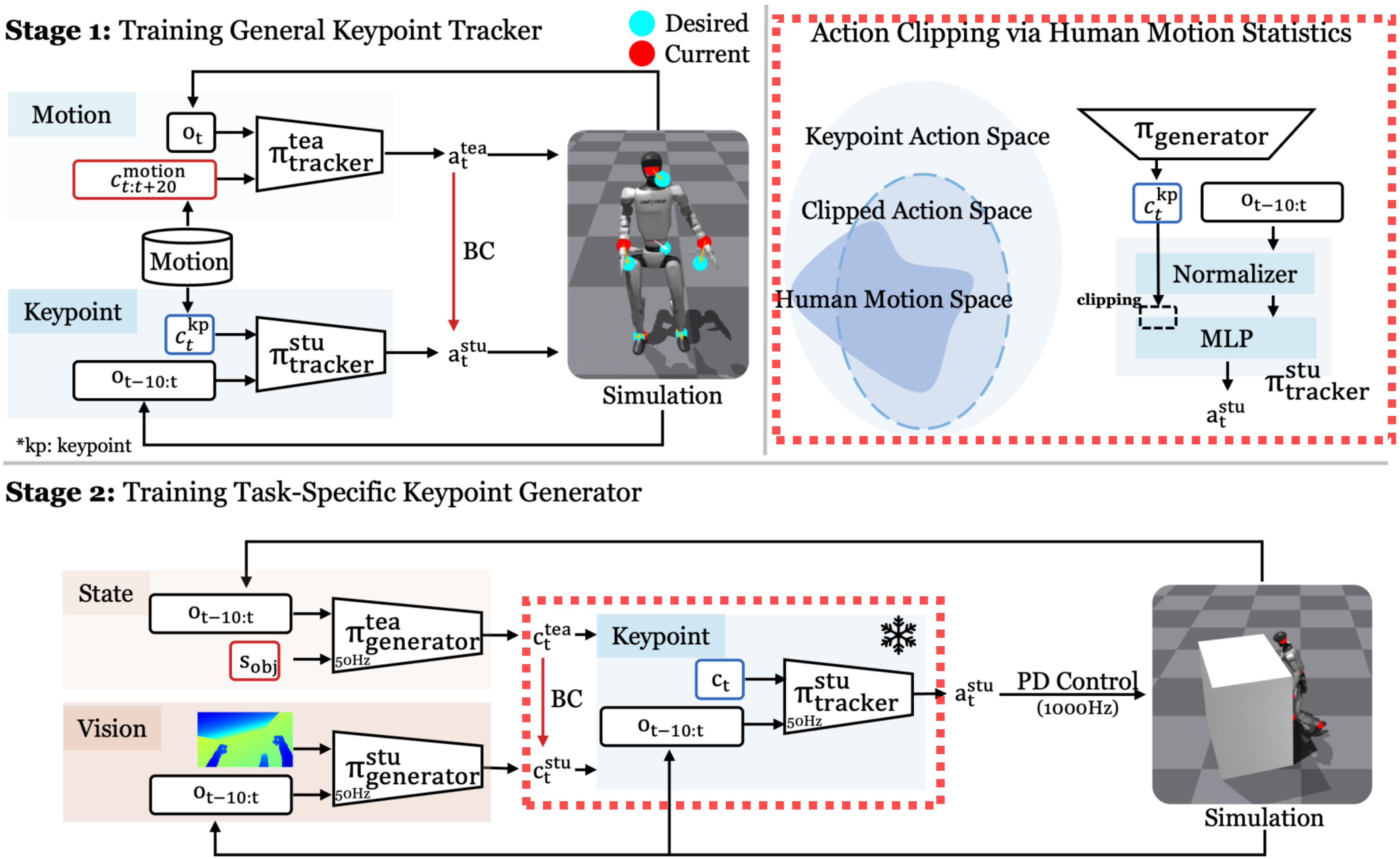


- Six independently sampled rectangular masks
- Each mask covers **up to 25%** of the image area
- Each mask has a **10% probability** of being applied

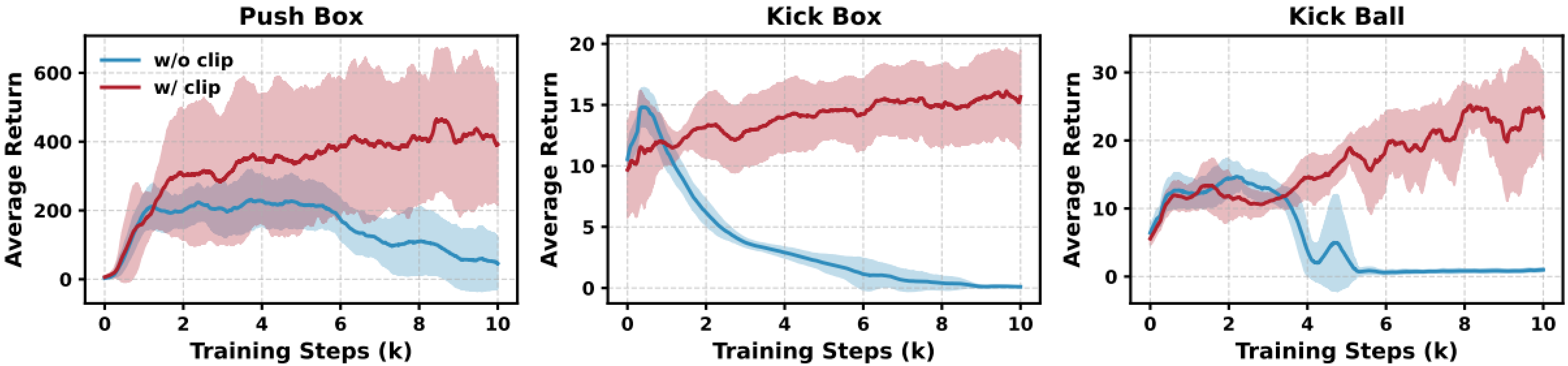
Real World: Binary Command



Important Tricks 1 (Action Clip)



Important Tricks 1 (Action Clip)



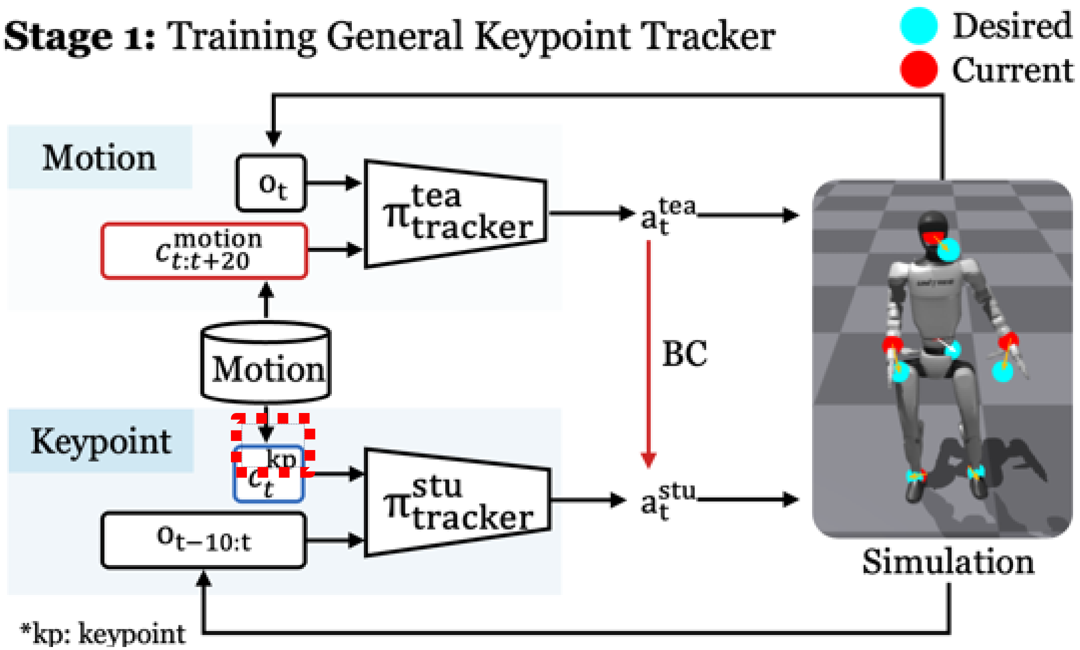
Method	**Push Box**			**Kick Box**			**Kick Ball**		
	Distance [m] ↑	Forward [m] ↑	Drift [m] ↓	Distance [m] ↑	Forward [m] ↑	Drift [m] ↓	Distance [m] ↑	Forward [m] ↑	Drift [m] ↓
State-based									
Ours	**152 ± 36**	**151 ± 29**	13 ± 4	**78 ± 3**	**78 ± 3**	**0 ± 0**	**189 ± 3**	**189 ± 3**	4 ± 1
w/o clip	68 ± 118	67 ± 94	11 ± 16	3 ± 5	3 ± 4	**0 ± 0**	12 ± 15	12 ± 12	1 ± 1
Vision-based									
Ours	**37 ± 28**	**19 ± 15**	21 ± 12	**55 ± 5**	**30 ± 3**	33 ± 3	**135 ± 6**	**121 ± 8**	47 ± 12
w/o clip	10 ± 18	9 ± 12	4 ± 5	6 ± 7	5 ± 3	3 ± 3	1 ± 1	0 ± 1	**0 ± 0**

Important Tricks 2 (Noise Augmentation)

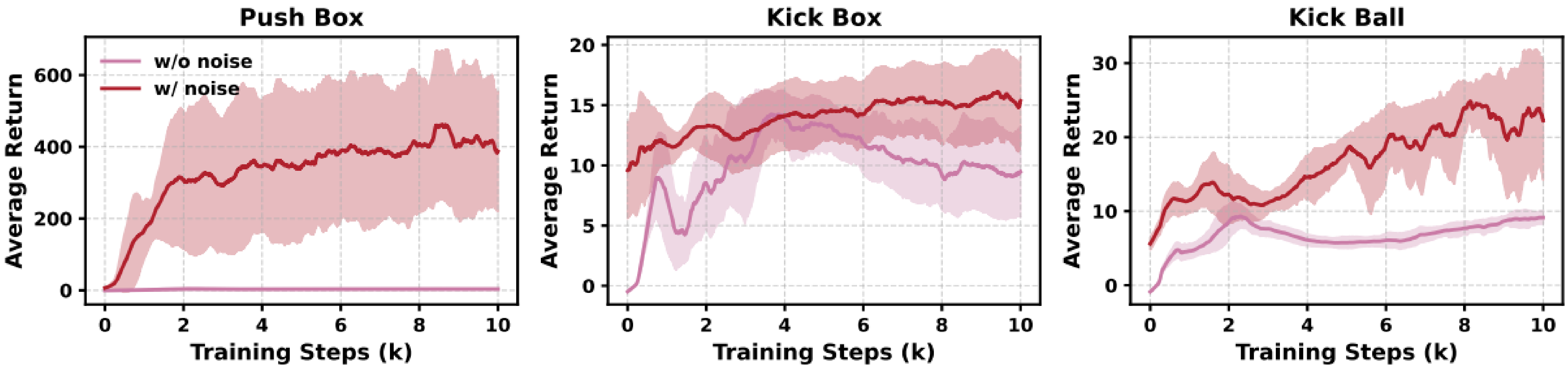
Implementation Details

- Add noise for student input (Stage 1)
  - relative noise
  - $X_{\text{noised}} = X \cdot \lambda_i, \quad \forall i \in \{1, \ldots, n\}, \quad \lambda_i \in [0.5, 1.5]$

- Why adding noise
  - Kind of data augumentation

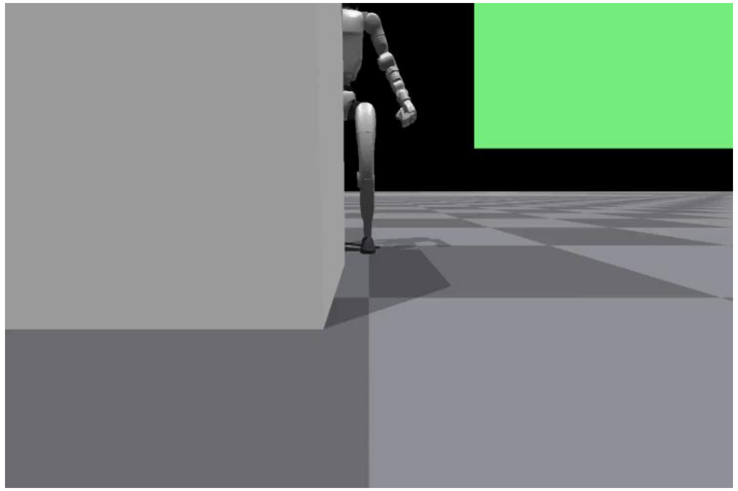


Important Tricks 2 (Noise Augmentation)

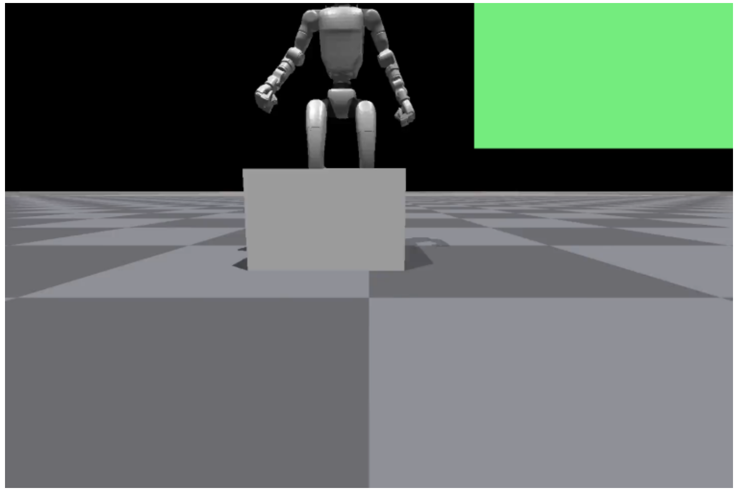


Method	Push Box			Kick Box			Kick Ball		
	Distance [m] ↑	Forward [m] ↑	Drift [m] ↓	Distance [m] ↑	Forward [m] ↑	Drift [m] ↓	Distance [m] ↑	Forward [m] ↑	Drift [m] ↓
State-based									
Ours	**152 ± 36**	**151 ± 29**	13 ± 4	**78 ± 3**	**78 ± 3**	**0 ± 0**	**189 ± 3**	**189 ± 3**	4 ± 1
w/o noise	2 ± 1	2 ± 1	**0 ± 0**	30 ± 24	30 ± 20	1 ± 0	136 ± 8	136 ± 7	4 ± 0
Vision-based									
Ours	**37 ± 28**	**19 ± 15**	21 ± 12	**55 ± 5**	**30 ± 3**	33 ± 3	**135 ± 6**	**121 ± 8**	47 ± 12
w/o noise	2 ± 1	2 ± 1	**0 ± 0**	25 ± 7	11 ± 4	15 ± 3	86 ± 7	77 ± 7	30 ± 8

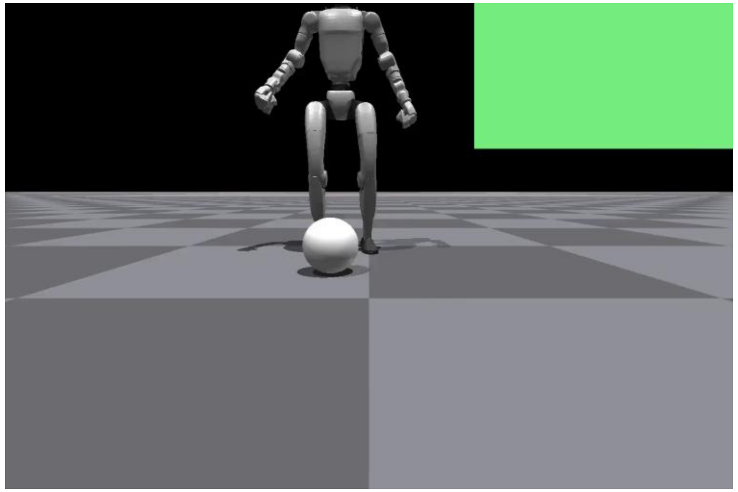
Results



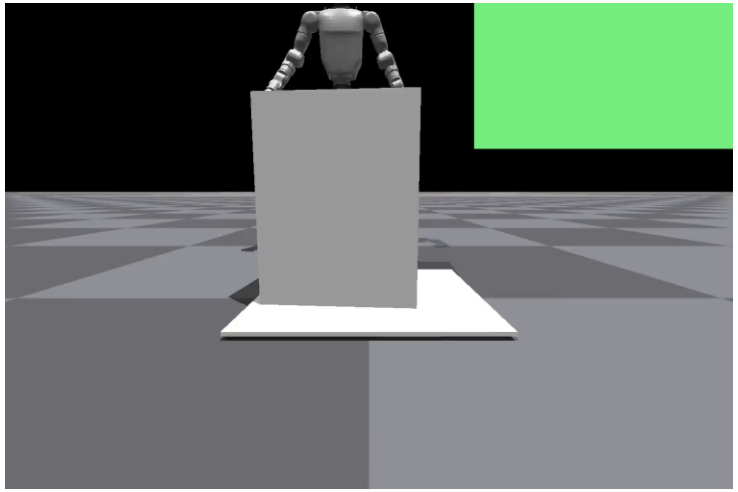
Push Box



Kick Box

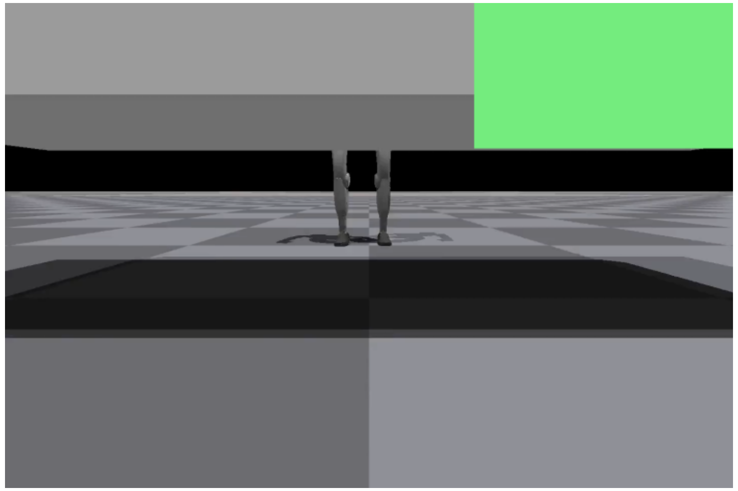


Kick Ball

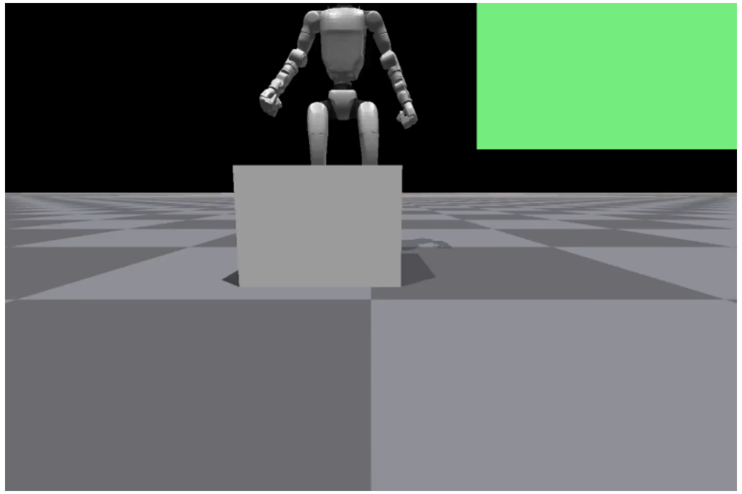


Lift Box

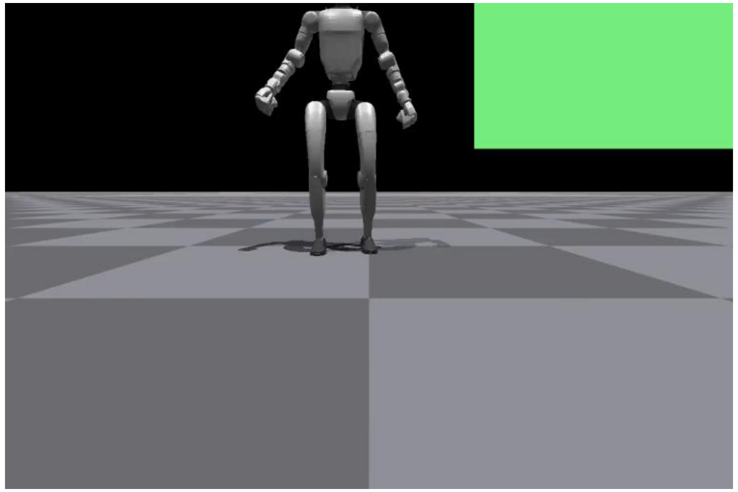
Results



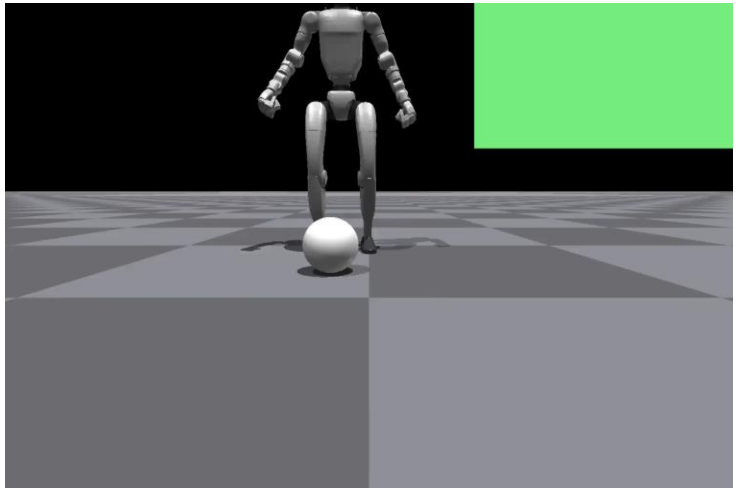
Push Cube



Large Kick



Reach Box



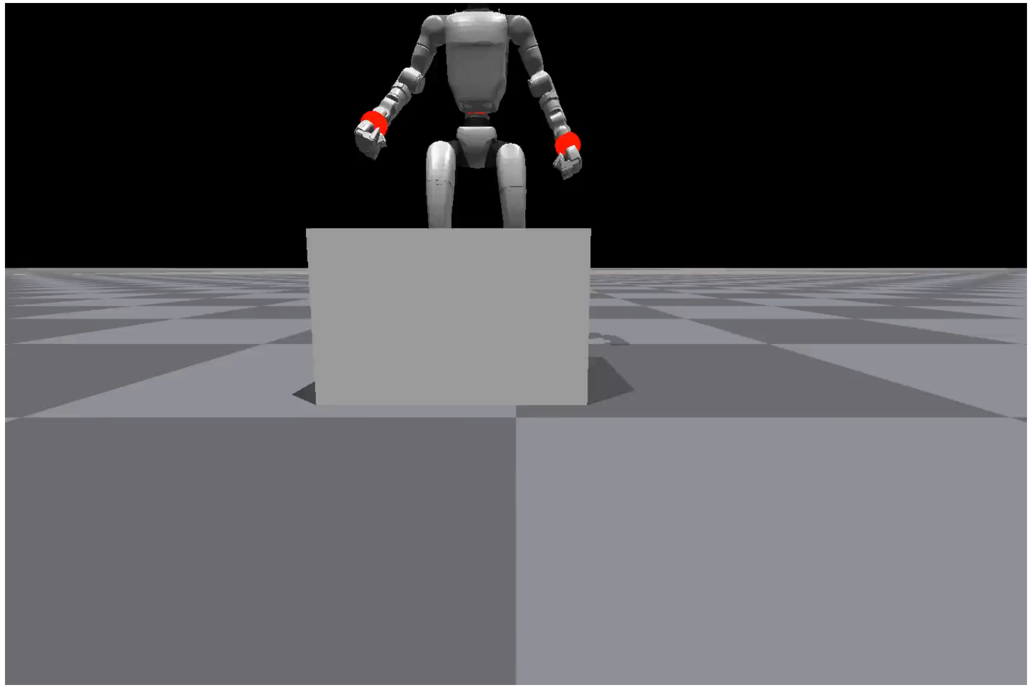
Balance Ball

Results (Simulation Evaluation)

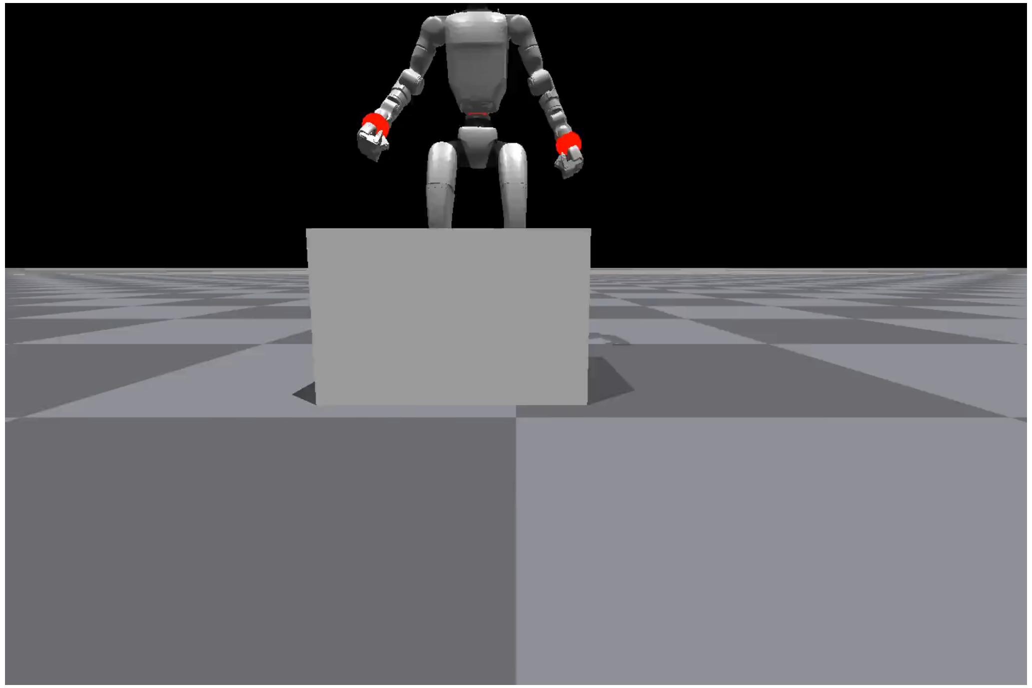
Method	Push Box			Kick Box		
	Distance [m] ↑	Forward [m] ↑	Drift [m] ↓	Distance [m] ↑	Forward [m] ↑	Drift [m] ↓
teacher	152 ± 36	151 ± 29	13 ± 4	78 ± 3	78 ± 3	0 ± 0
stu w/ vision	**37 ± 28**	**19 ± 15**	21 ± 12	**55 ± 5**	**30 ± 3**	33 ± 3
stu w/o vision	2 ± 0	2 ± 0	**1 ± 0**	0 ± 0	0 ± 0	**0 ± 0**
Method	**Large Kick**			**Kick Ball**		
	Distance [m] ↑	Forward [m] ↑	Drift [m] ↓	Distance [m] ↑	Forward [m] ↑	Drift [m] ↓
teacher	8 ± 1	7 ± 1	2 ± 0	189 ± 3	189 ± 3	4 ± 1
stu w/ vision	**6 ± 0**	**6 ± 0**	2 ± 0	**135 ± 6**	**121 ± 8**	47 ± 12
stu w/o vision	4 ± 0	4 ± 0	**1 ± 0**	1 ± 0	1 ± 0	**0 ± 0**

Method	Lift Box			Reach Box		
	Height [m] ↑	Box Fall Rate [%] ↓	Alive [s] ↑	Velocity [m/s] ↑	Collision Rate [%] ↓	Alive [s] ↑
teacher	1 ± 0	34 ± 25	38 ± 13	4 ± 0	0 ± 0	60 ± 0
stu w/ vision	**1 ± 0**	30 ± 23	**30 ± 7**	**4 ± 0**	**0 ± 0**	**42 ± 6**
stu w/o vision	0 ± 0	**15 ± 21**	6 ± 4	**4 ± 0**	**0 ± 0**	18 ± 6
Method	**Balance Ball**			**Push Cube (Tabletop)**		
	Force [N] ↑	Foot Fall Rate [%] ↓	Alive [s] ↑	Error [cm] ↓	Finish Time [s] ↓	Alive [s] ↑
teacher	21 ± 2	0 ± 0	34 ± 8	9 ± 3	4 ± 1	58 ± 1
stu w/ vision	**24 ± 1**	**0 ± 0**	**45 ± 7**	**21 ± 2**	**20 ± 8**	**57 ± 0**
stu w/o vision	6 ± 0	**0 ± 0**	5 ± 1	57 ± 22	43 ± 8	51 ± 10

Analysis Results (Training Pipeline)

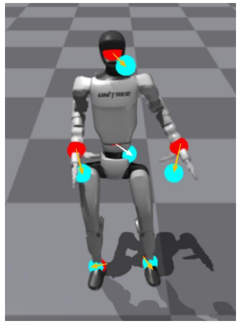


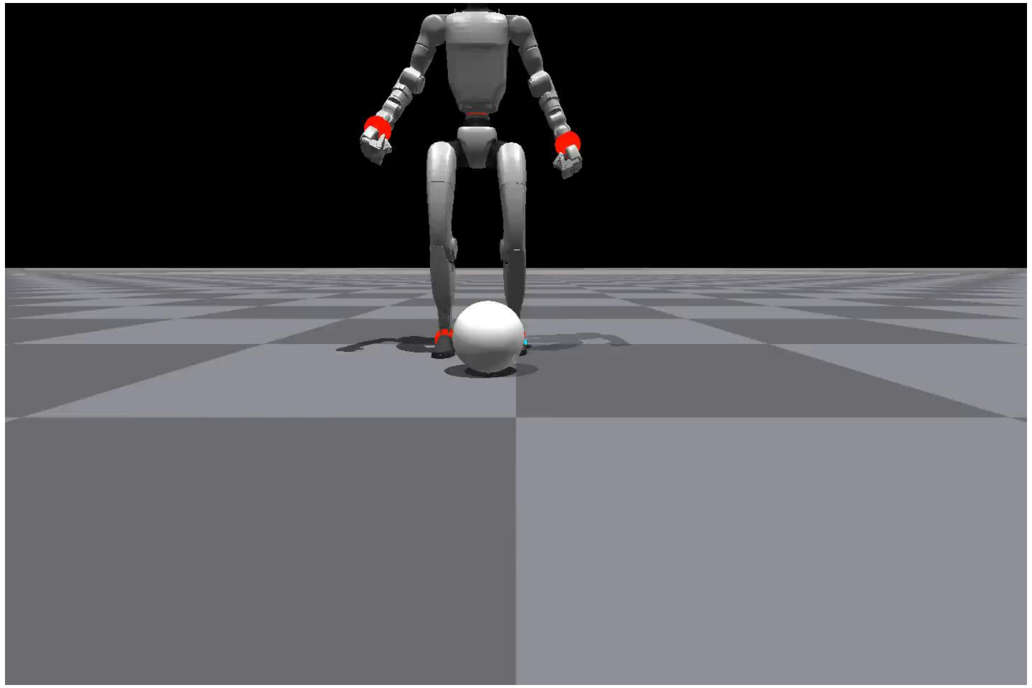
VisualMimic



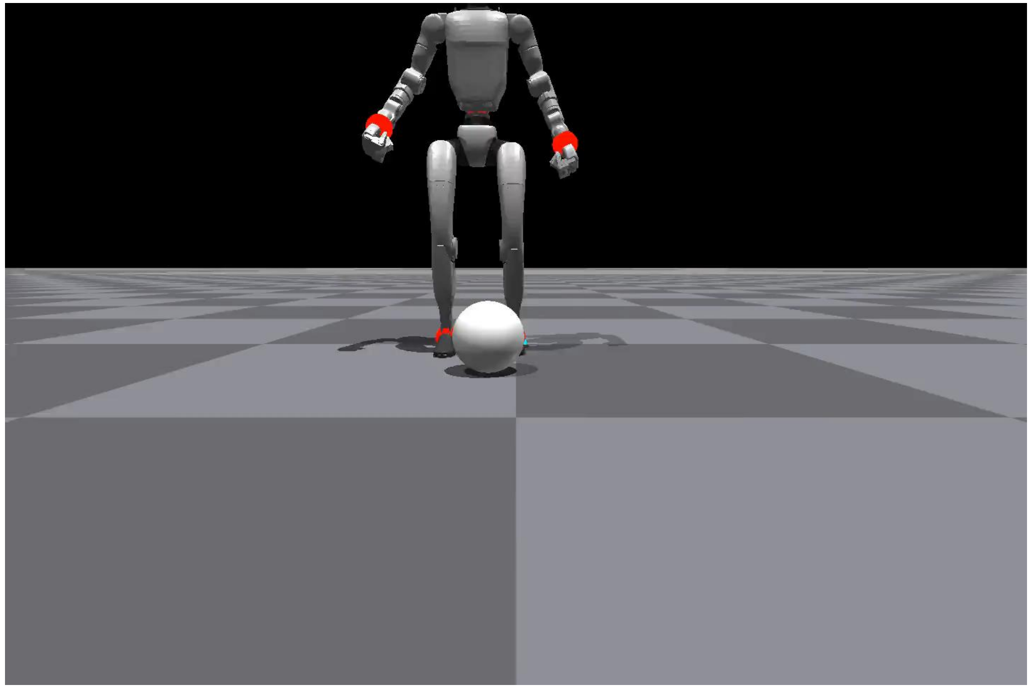
VisualMimic (w/o Stage 0&1 Distillation)

Analysis Results (Interface Design)



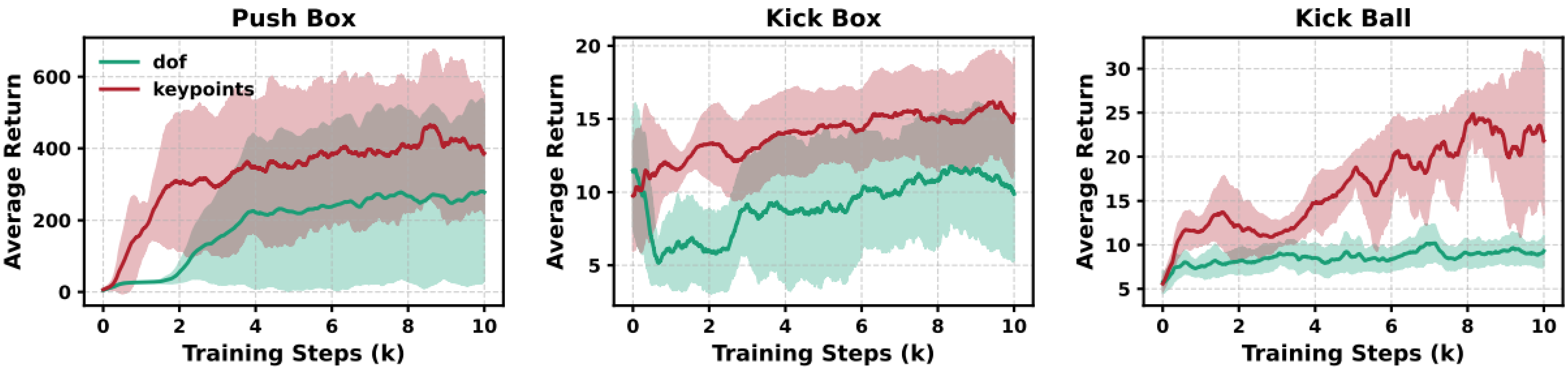


VisualMimic



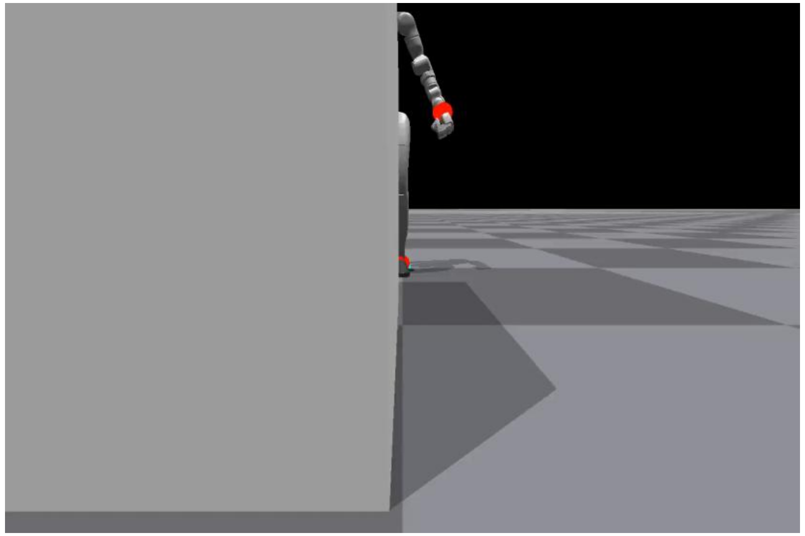
VisualMimic (w/ 3 point interface)

Analysis Results (Interface Design)

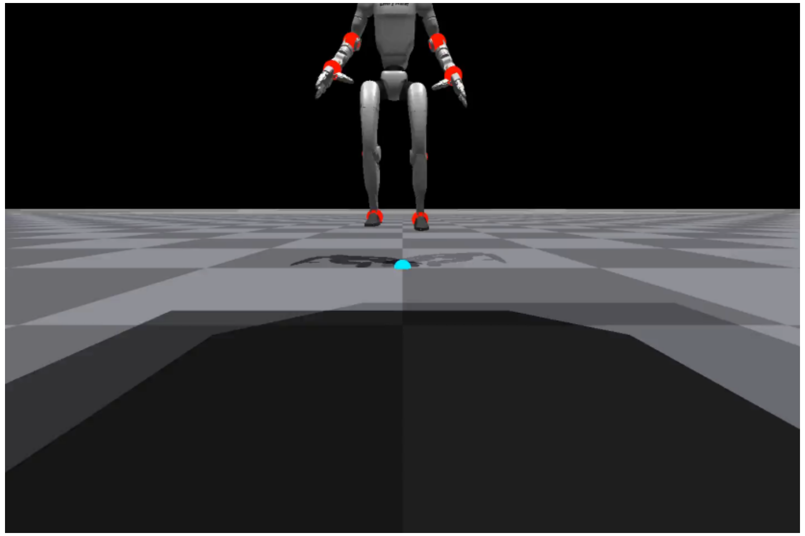


Method	Push Box			Kick Box			Kick Ball		
	Distance [m] ↑	Forward [m] ↑	Drift [m] ↓	Distance [m] ↑	Forward [m] ↑	Drift [m] ↓	Distance [m] ↑	Forward [m] ↑	Drift [m] ↓
State-based									
Ours	**152 ± 36**	**151 ± 29**	13 ± 4	**78 ± 3**	**78 ± 3**	**0 ± 0**	**189 ± 3**	**189 ± 3**	4 ± 1
DoF as Interface	10 ± 9	8 ± 6	5 ± 3	40 ± 34	40 ± 28	**0 ± 0**	0 ± 0	0 ± 0	**0 ± 0**
Vision-based									
Ours	**37 ± 28**	**19 ± 15**	21 ± 12	**55 ± 5**	**30 ± 3**	33 ± 3	**135 ± 6**	**121 ± 8**	47 ± 12
DoF as Interface	10 ± 2	6 ± 1	6 ± 1	5 ± 4	1 ± 0	4 ± 3	0 ± 0	0 ± 0	**0 ± 0**

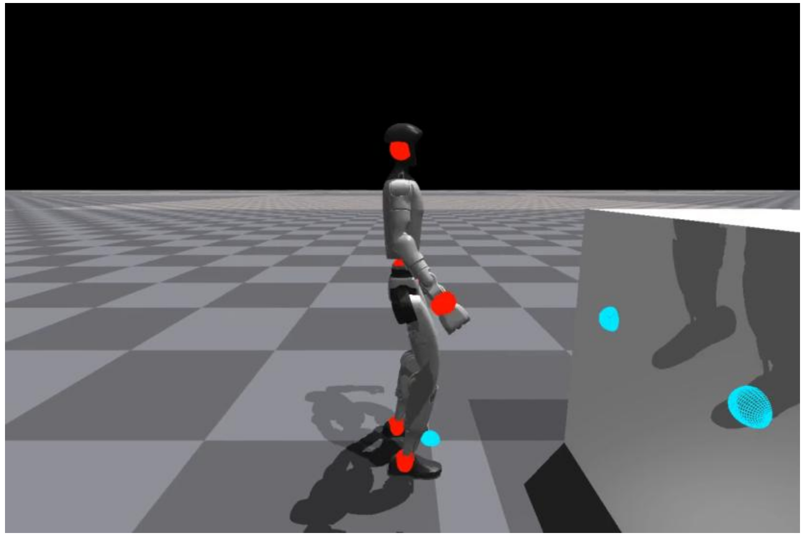
Analysis Results (Whole-body Dexterity)



Foot and Hand



Shoulder



Two Hands

Analysis Results

Method	Push Box			Kick Box			Kick Ball		
	Distance [m] ↑	Forward [m] ↑	Drift [m] ↓	Distance [m] ↑	Forward [m] ↑	Drift [m] ↓	Distance [m] ↑	Forward [m] ↑	Drift [m] ↓
State-based									
Ours	**152 ± 36**	**151 ± 29**	13 ± 4	**78 ± 3**	**78 ± 3**	**0 ± 0**	**189 ± 3**	**189 ± 3**	4 ± 1
w/o noise	2 ± 1	2 ± 1	**0 ± 0**	30 ± 24	30 ± 20	1 ± 0	136 ± 8	136 ± 7	4 ± 0
w/o clip	68 ± 118	67 ± 94	11 ± 16	3 ± 5	3 ± 4	**0 ± 0**	12 ± 15	12 ± 12	1 ± 1
DoF as Interface	10 ± 9	8 ± 6	5 ± 3	40 ± 34	40 ± 28	**0 ± 0**	0 ± 0	0 ± 0	**0 ± 0**
Local-Frame Tracker	38 ± 27	30 ± 23	16 ± 15	45 ± 7	45 ± 5	1 ± 0	109 ± 23	109 ± 19	7 ± 1
Vision-based									
Ours	**37 ± 28**	**19 ± 15**	21 ± 12	**55 ± 5**	**30 ± 3**	33 ± 3	**135 ± 6**	**121 ± 8**	47 ± 12
w/o noise	2 ± 1	2 ± 1	**0 ± 0**	25 ± 7	11 ± 4	15 ± 3	86 ± 7	77 ± 7	30 ± 8
w/o clip	10 ± 18	9 ± 12	4 ± 5	6 ± 7	5 ± 3	3 ± 3	1 ± 1	0 ± 1	**0 ± 0**
DoF as Interface	10 ± 2	6 ± 1	6 ± 1	5 ± 4	1 ± 0	4 ± 3	0 ± 0	0 ± 0	**0 ± 0**
Local-Frame Tracker	14 ± 11	7 ± 5	8 ± 6	27 ± 15	16 ± 9	15 ± 6	38 ± 17	34 ± 13	12 ± 4
Visual RL	25 ± 16	11 ± 6	16 ± 9	0 ± 0	0 ± 0	**0 ± 0**	0 ± 0	0 ± 0	**0 ± 0**
Blind									
Ours w/o vision	2 ± 0	2 ± 0	1 ± 0	0 ± 0	0 ± 0	0 ± 0	1 ± 0	1 ± 0	0 ± 0

Takeaway & Limitation

- RL is a beast
  - Powerful
  - Without proper DR or termination, it might bite
- Human motion matters
- Surprising facts
  - The visual sim2real gap isn’t as large as expected.
  - Distilled visual policies still lag far behind their state-based counterparts.

https://visualmimic.github.io

Code, Video

Thank You!

Contact: yinshaofeng04@gmail.com

Website: https://operator22th.github.io/

I am applying for PhD programs in Fall 2026.

I’d love to connect and discuss research opportunities.