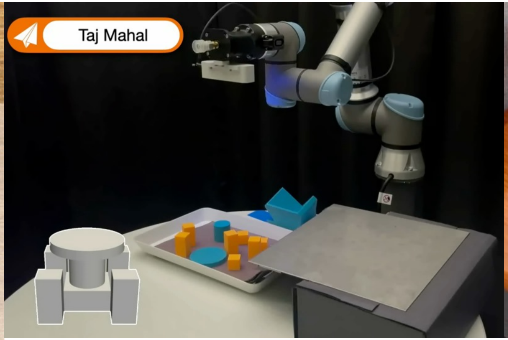
Physically Grounded Reasoning for Open-World Robot Dexterity



Kuan Fang
Department of Computer Science
Cornell University

Cornell Bowers C·IS
College of Computing
and Information Science

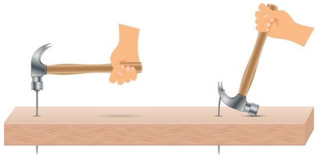

1×
Color Cubes
Clean

Toward Open-World Robot Dexterity

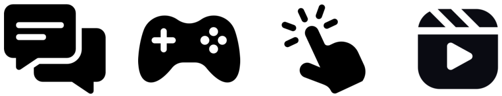
Open-World Generalization

Environment-Level

Behavior-Level

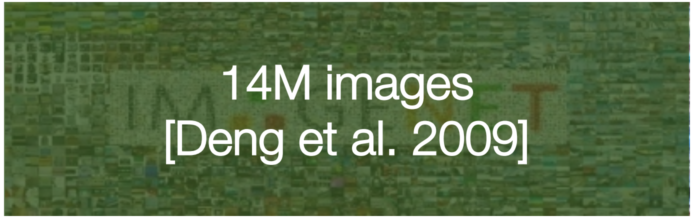


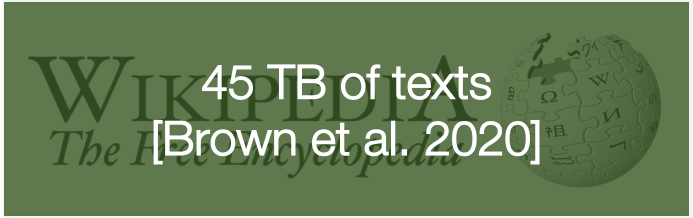
Instruction-Level



Path to foundation models: Scaling up

massive datasets





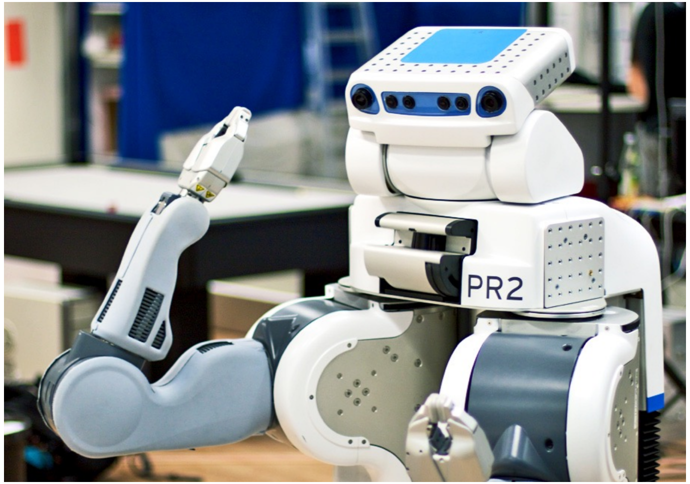
generalizable AI models

[Lin et al. 2014; Liu et al. 2015; Yu et al. 2015; Chang et al. 2015; Heilbron et al. 2015; Abu-El-Haija et al. 2016; Mo et al. 2018; He et al. 2017]

Scaling up end-to-end robot learning as the solution ?

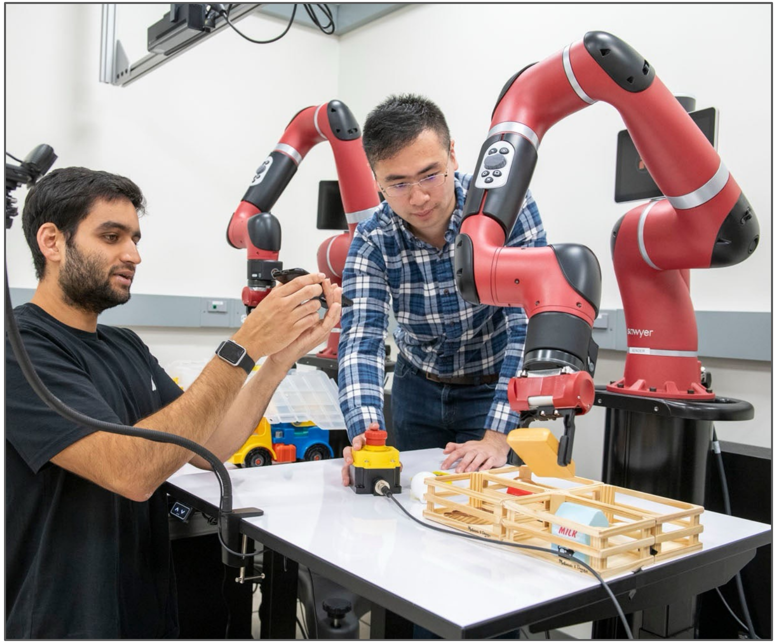
massive robot data

open-world robot dexterity

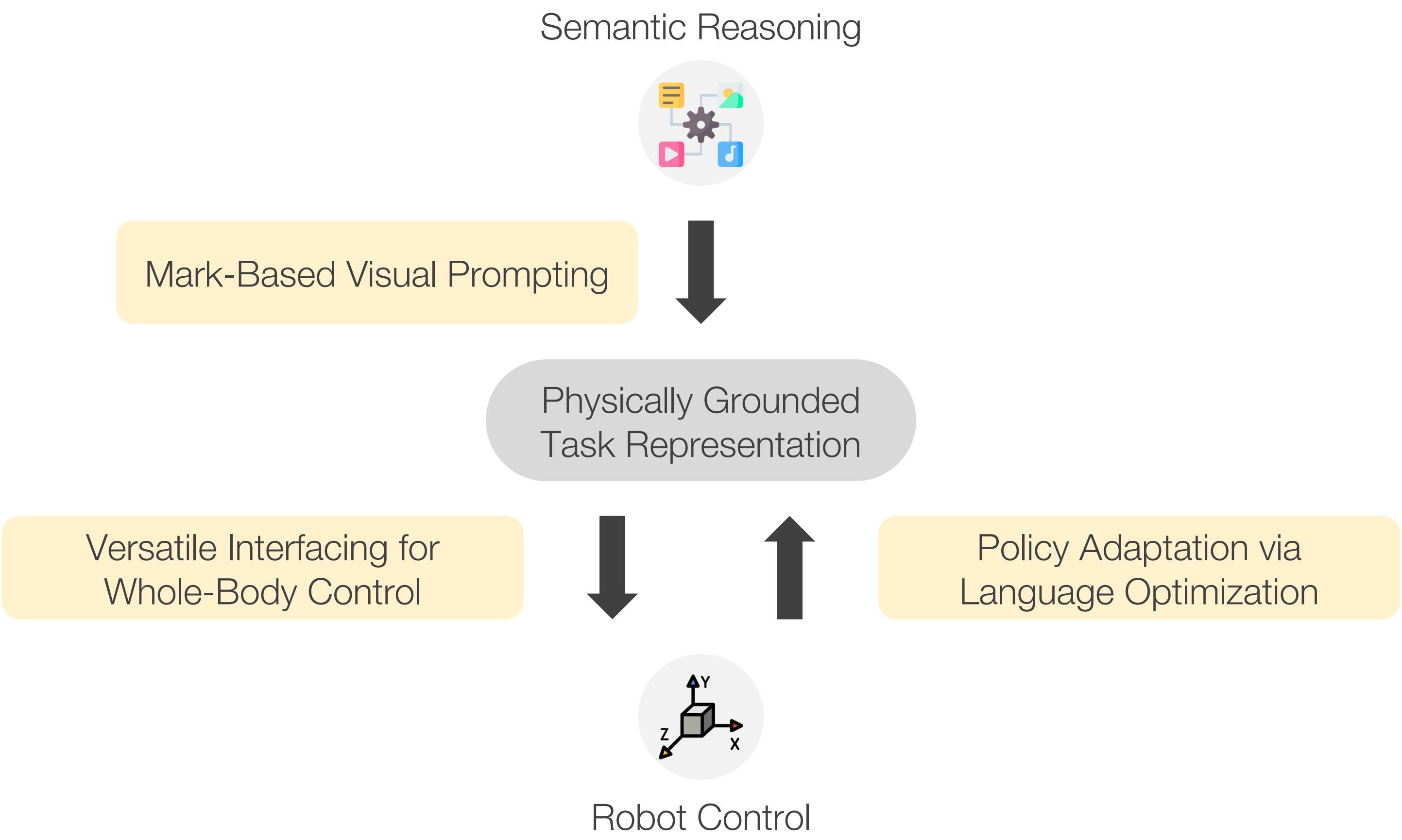


Challenges in scaling up robot learning

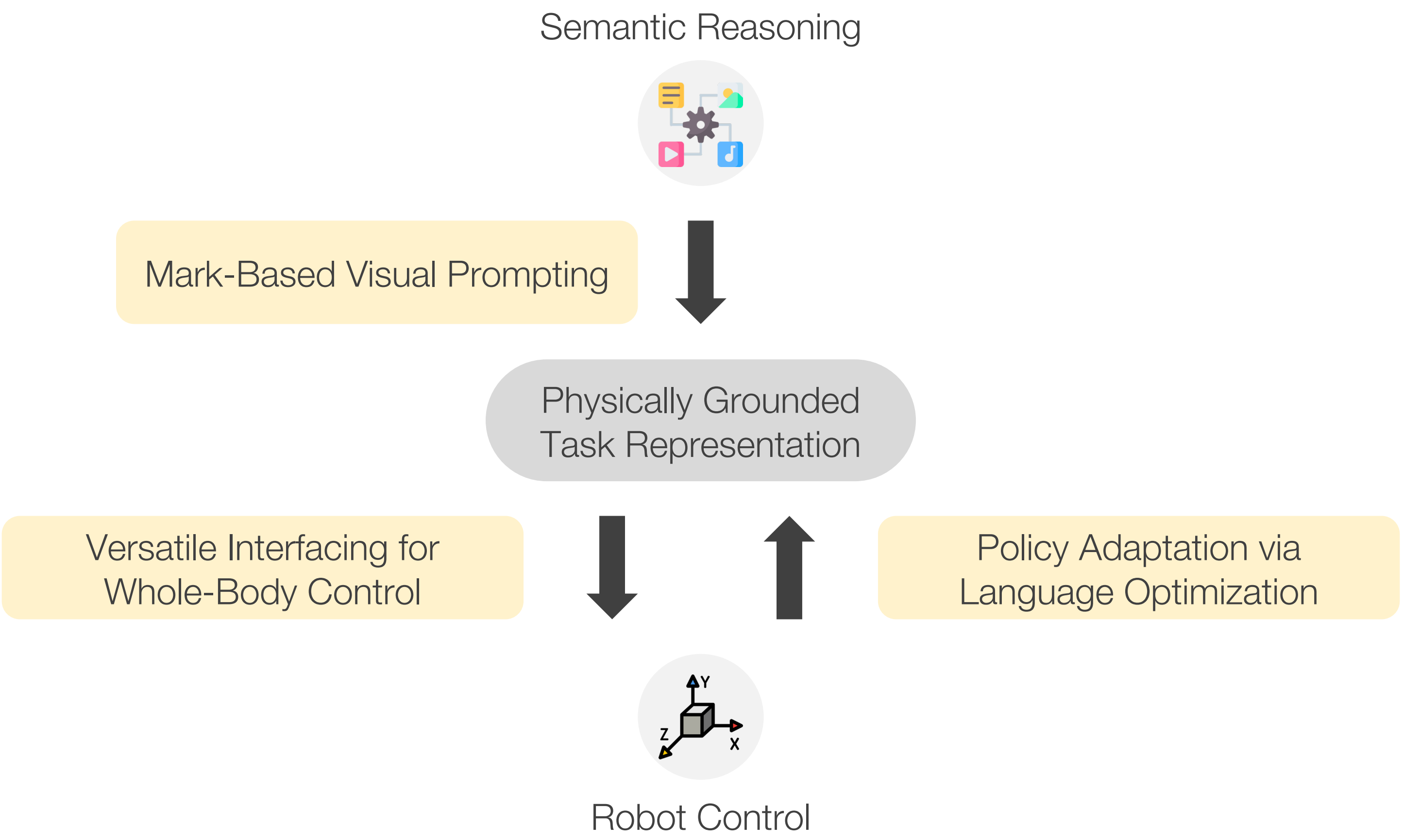
various variety of robot tasks

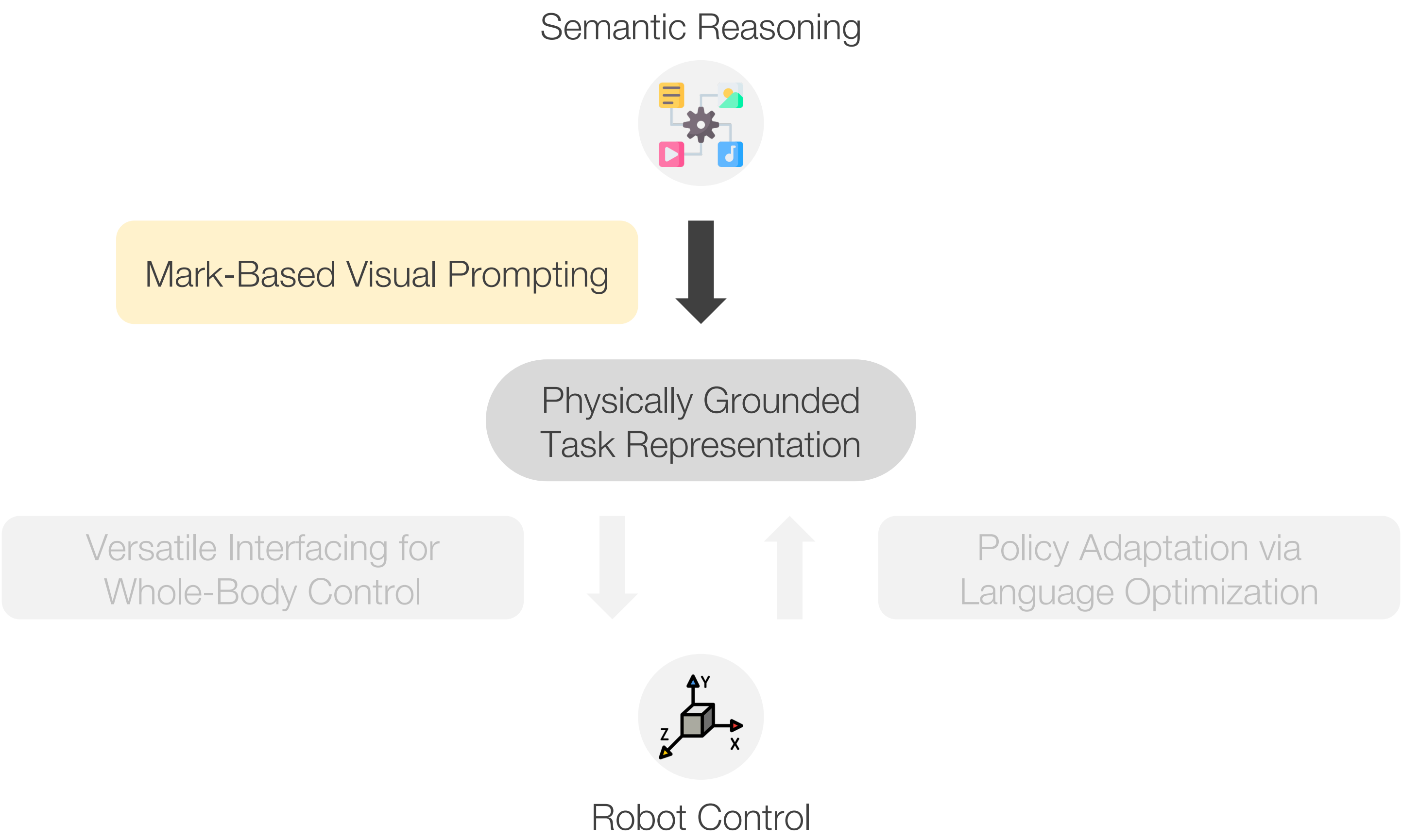


learning requires physical interactions



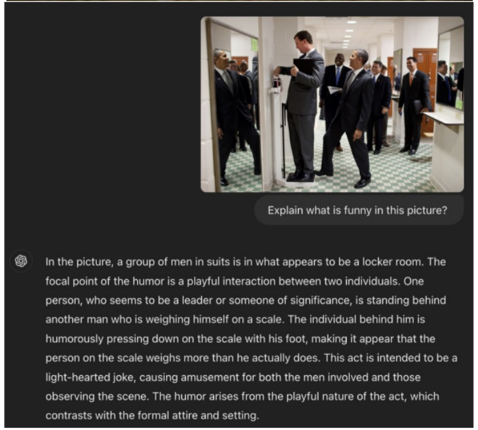
Semantic Reasoning
Mark-Based Visual Prompting
Physically Grounded Task Representation
Versatile Interfacing for Whole-Body Control
Policy Adaptation via Language Optimization
Y
Z
X
Robot Control


Semantic Reasoning
Mark-Based Visual Prompting
Physically Grounded
Task Representation
Versatile Interfacing for
Whole-Body Control
Policy Adaptation via
Language Optimization
Y
Z
X
Robot Control


Semantic Reasoning
Mark-Based Visual Prompting
Physically Grounded Task Representation
Versatile Interfacing for Whole-Body Control
Policy Adaptation via Language Optimization
Y
Z
X
Robot Control

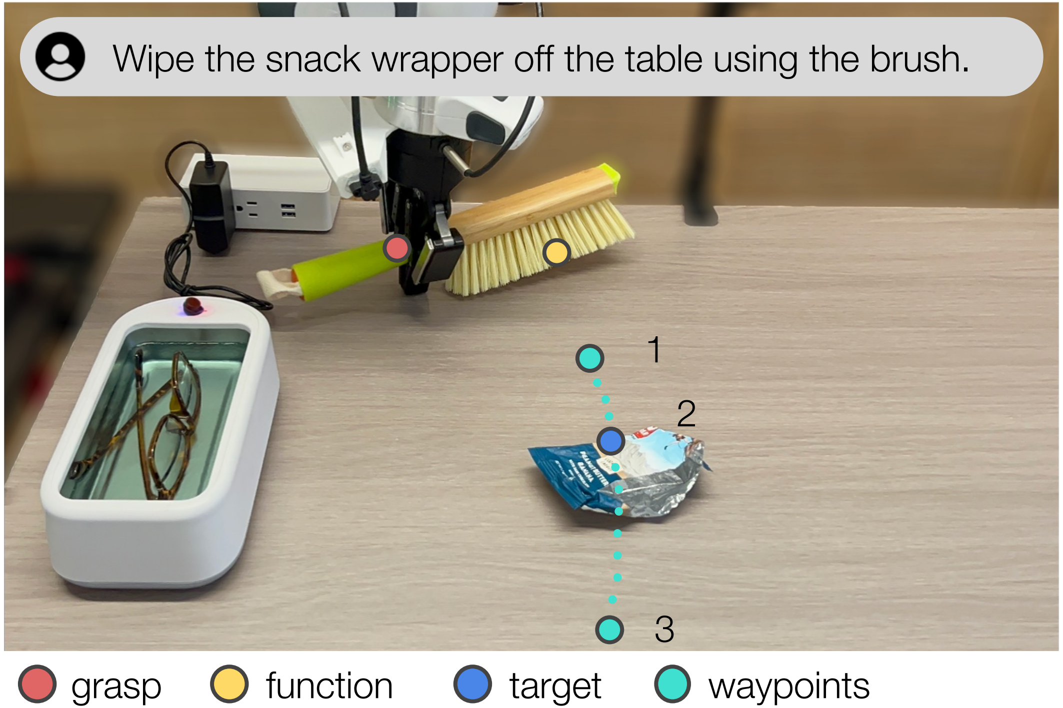
How can we leverage a pre-trained VLM for robotic control?

💡 **Key insight:** Convert motion planning into a series of **QA problems** that VLMs can solve.



MOKA: Marking Open-world Keypoint Affordances

Use a set of keypoints to specify the motion trajectory for solving the task.



✅ Separate semantics and motions

✅ Predictable on 2D images.

✅ Can specify diverse motions.

✅ Agnostic to the embodiment.

Fang*, Liu*, Abbeel, Levine. Multi-Task Domain Adaptation. RSS 2024

MOKA: Marking Open-world Keypoint Affordances

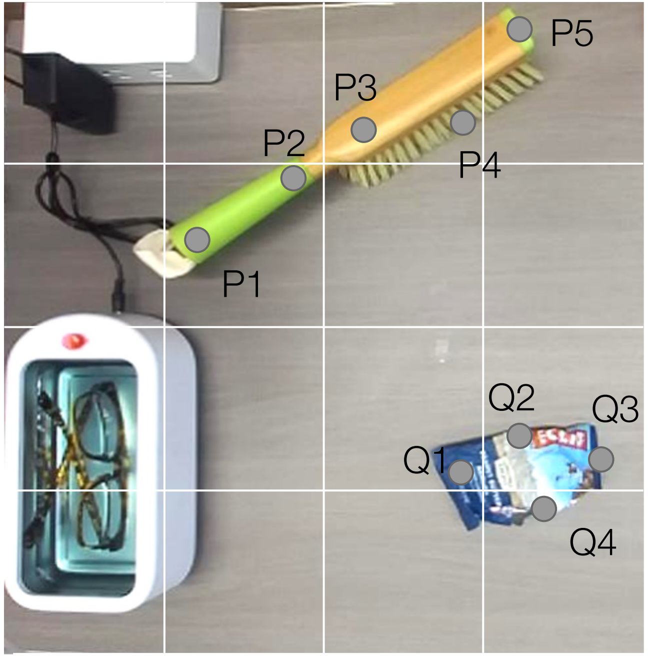
Challenge: Directly predicting keypoint coordinates requires fine-grained spatial reasoning.

grasp function target waypoints

Fang[*], Liu[*], Abbeel, Levine. Multi-Task Domain Adaptation. RSS 2024

MOKA: Marking Open-world Keypoint Affordances

To facilitate reasoning for the VLM, MOKA annotates a set of marks on the input image.



grasp function target waypoints

marks

Fang*, Liu*, Abbeel, Levine. Multi-Task Domain Adaptation. RSS 2024

MOKA: Marking Open-world Keypoint Affordances

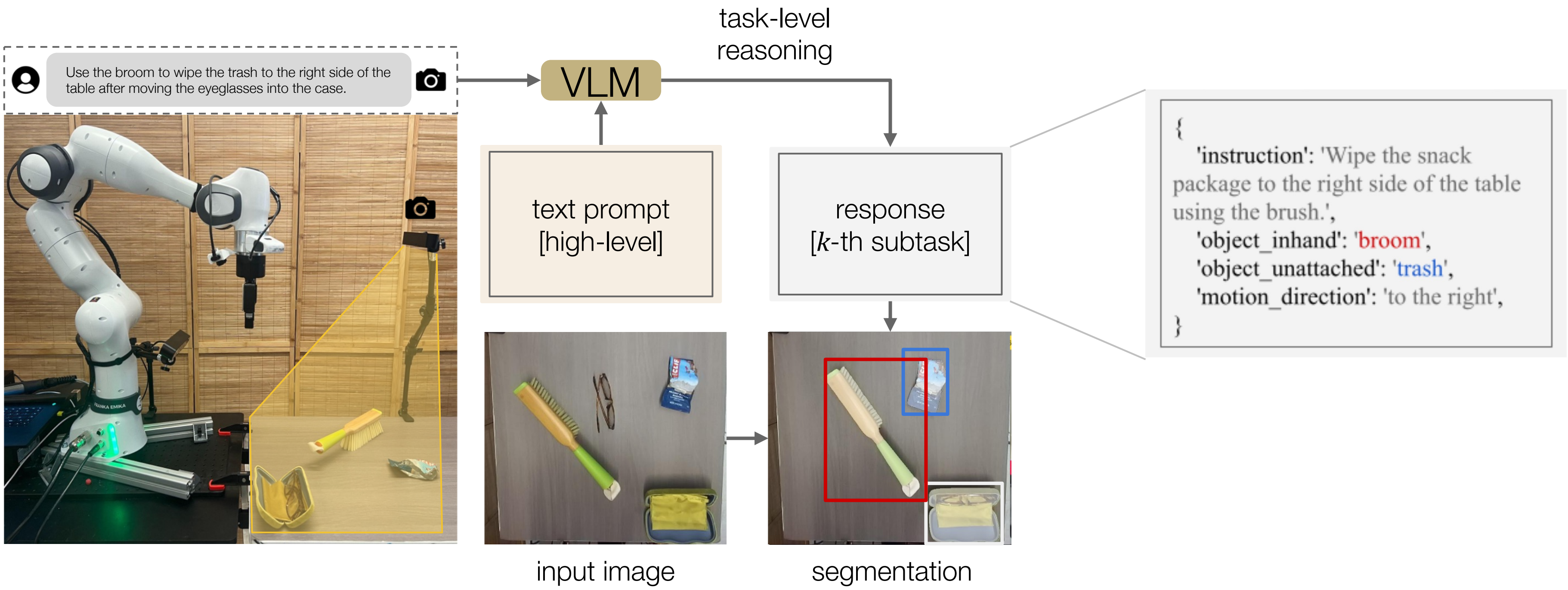
Without any training on any robot data, the VLM can solve the commanded manipulation task.

grasp function target waypoints

marks

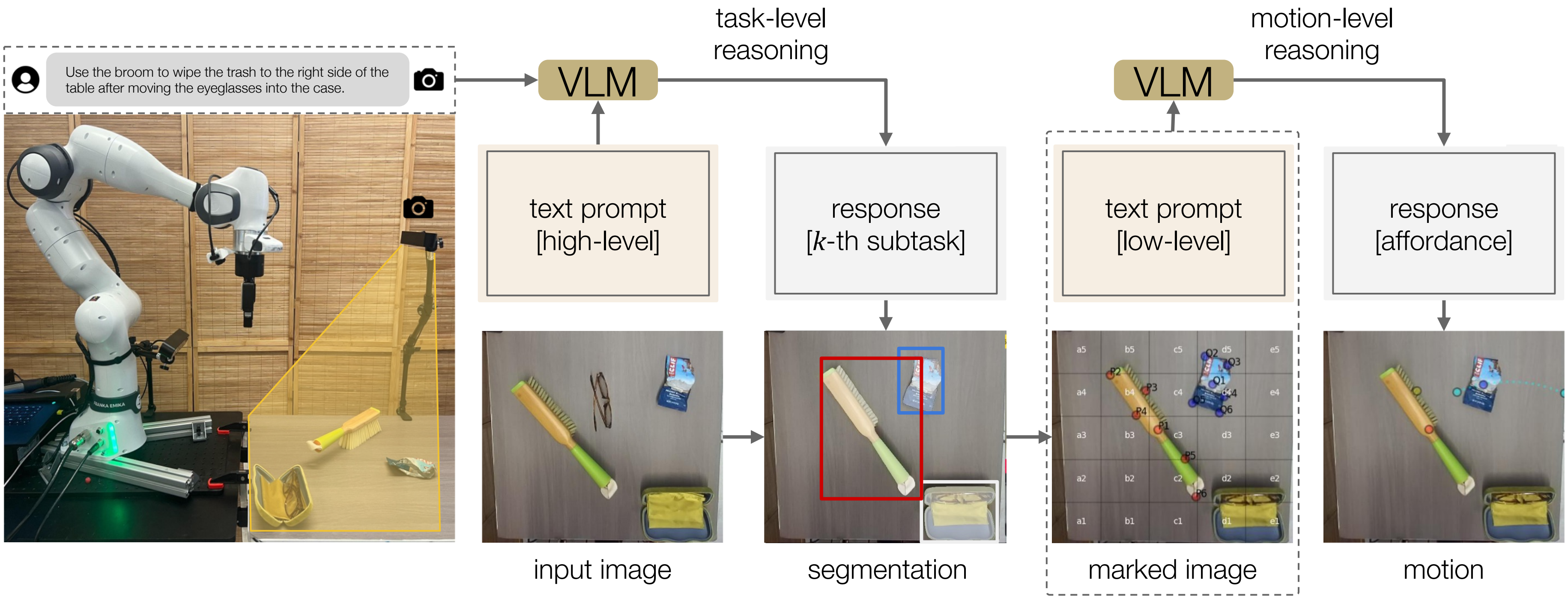
Fang*, Liu*, Abbeel, Levine. Multi-Task Domain Adaptation. RSS 2024

MOKA: Marking Open-world Keypoint Affordances



Fang*, Liu*, Abbeel, Levine. Multi-Task Domain Adaptation. RSS 2024

MOKA: Marking Open-world Keypoint Affordances

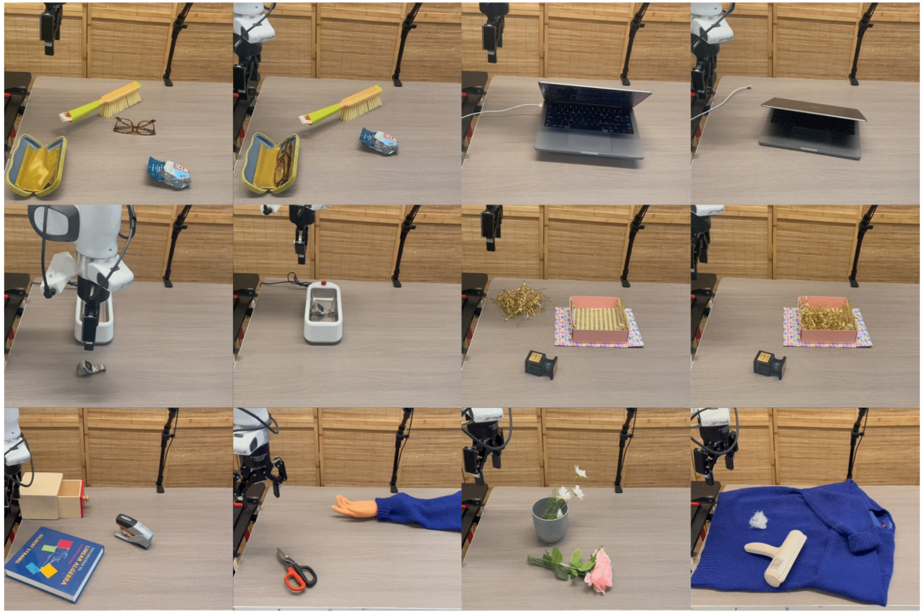


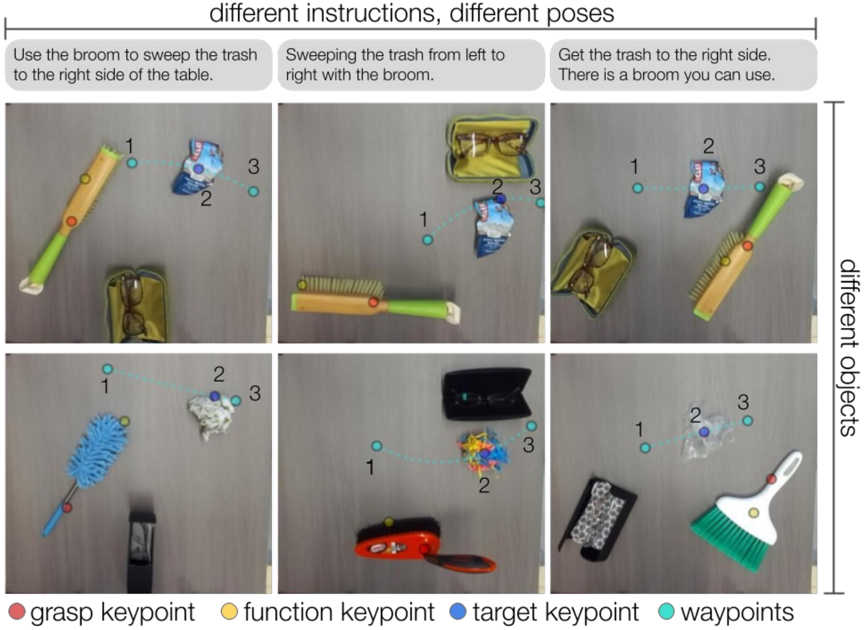
Fang*, Liu*, Abbeel, Levine. Multi-Task Domain Adaptation. RSS 2024

MOKA: Marking Open-world Keypoint Affordances

Without any training on any robot data, the VLM can solve the commanded manipulation task.

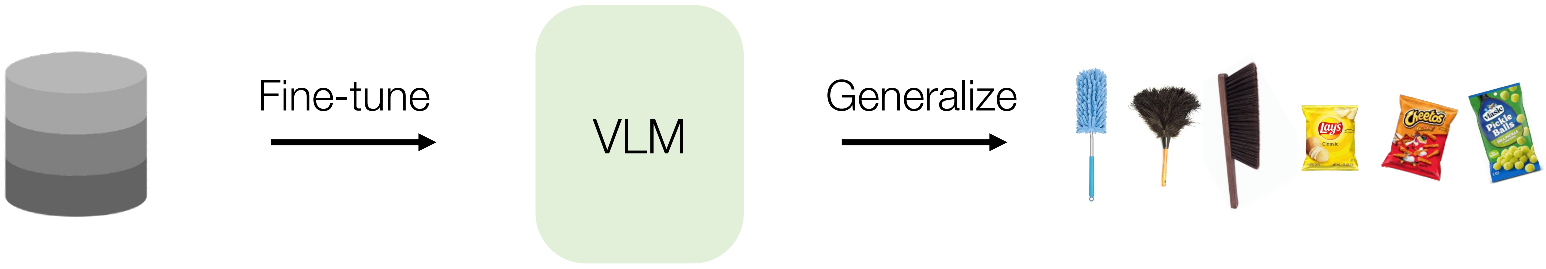
The prediction is robust to different instructions, poses, and objects.





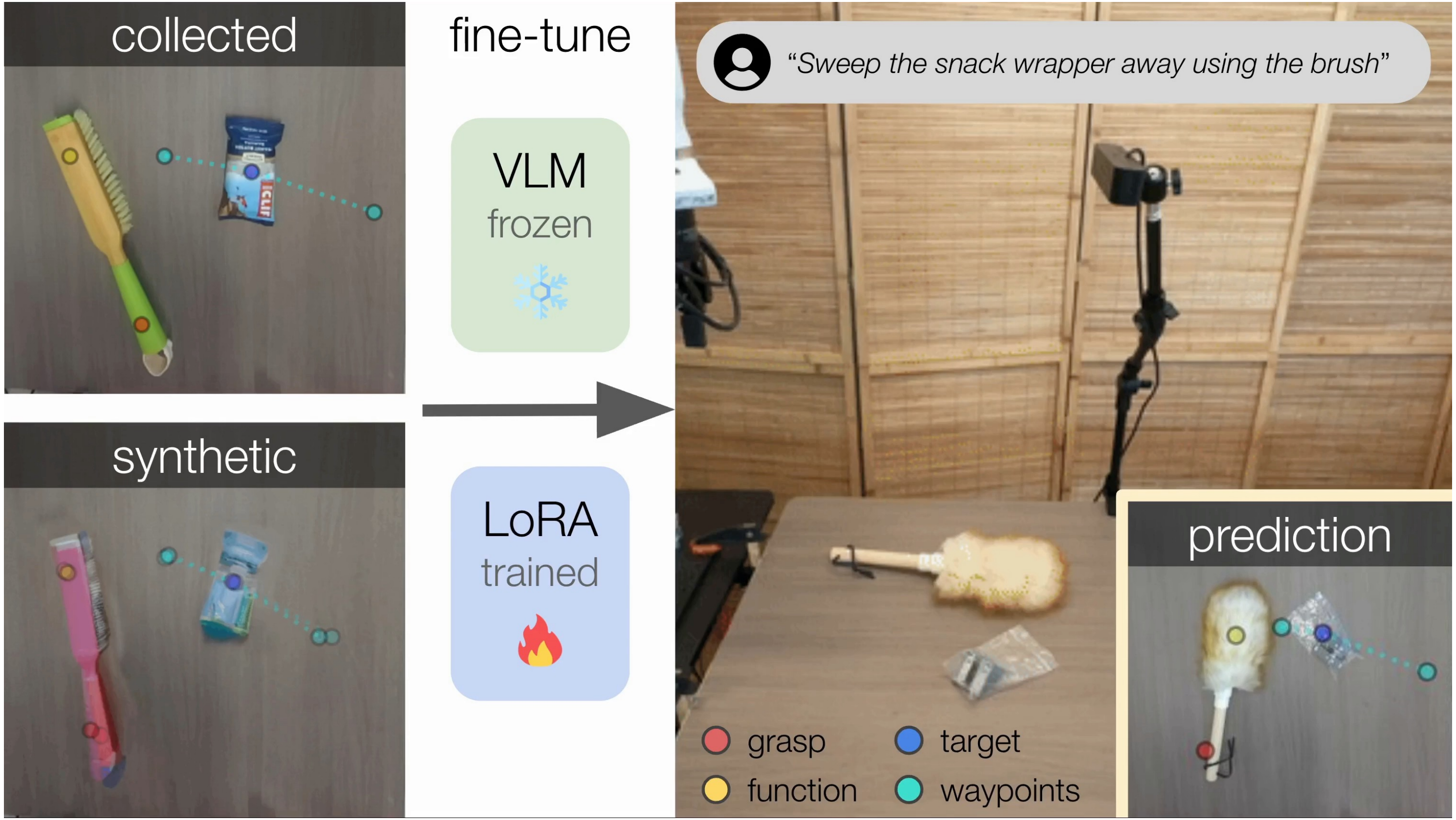
Fang*, Liu*, Abbeel, Levine. Multi-Task Domain Adaptation. RSS 2024

How to effectively fine-tune the VLM to improve generalization?



GPT-4 pre-training used
around 13 trillion tokens

KALIE: Keypoint Affordance Learning from Imagined Environments



Tang, Rajkumar, Zhou, Walke, Levine, **Fang**. Fine-Tuning Vision-Language Models for Open-World Manipulation without Robot Data. ICRA 2025

Challenge: How to generate physically consistent images?

Directly generating images from scratch or inpainting the images often lead to poor quality.

input

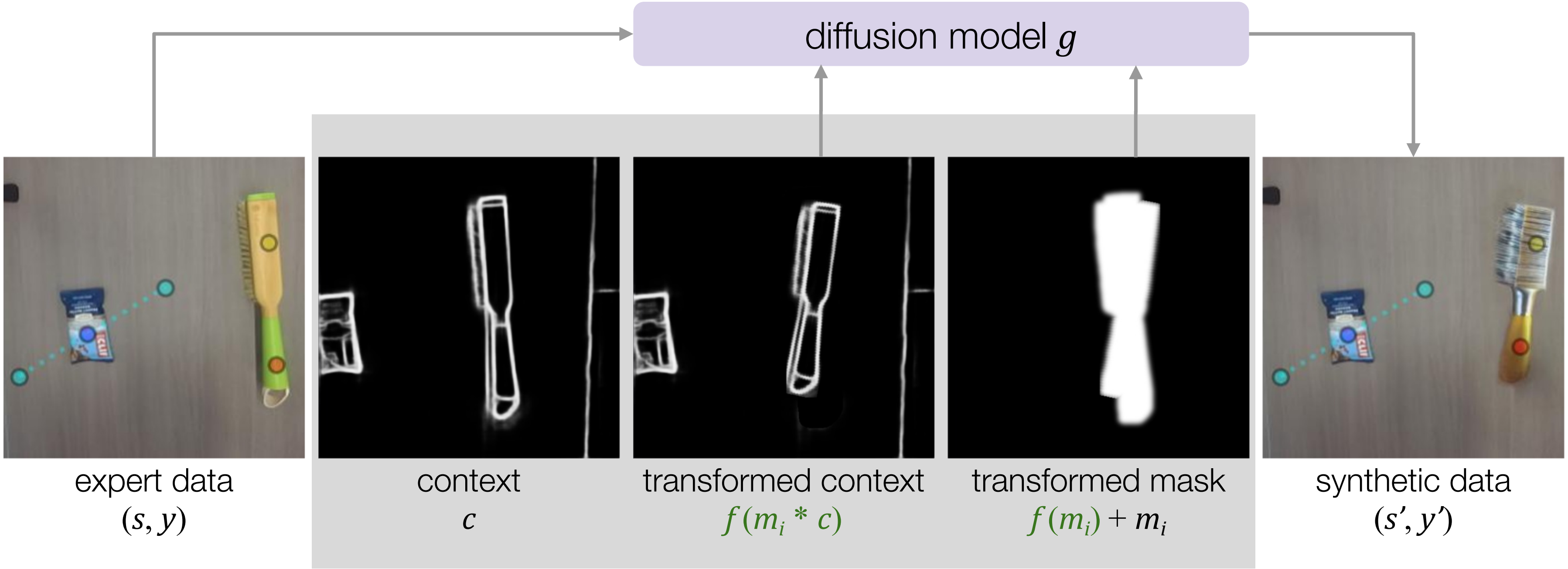
w/o original

w/o context

Tang, Rajkumar, Zhou, Walke, Levine, **Fang**. Fine-Tuning Vision-Language Models for Open-World Manipulation without Robot Data. ICRA 2025

Affordance-Aware Object Diversification

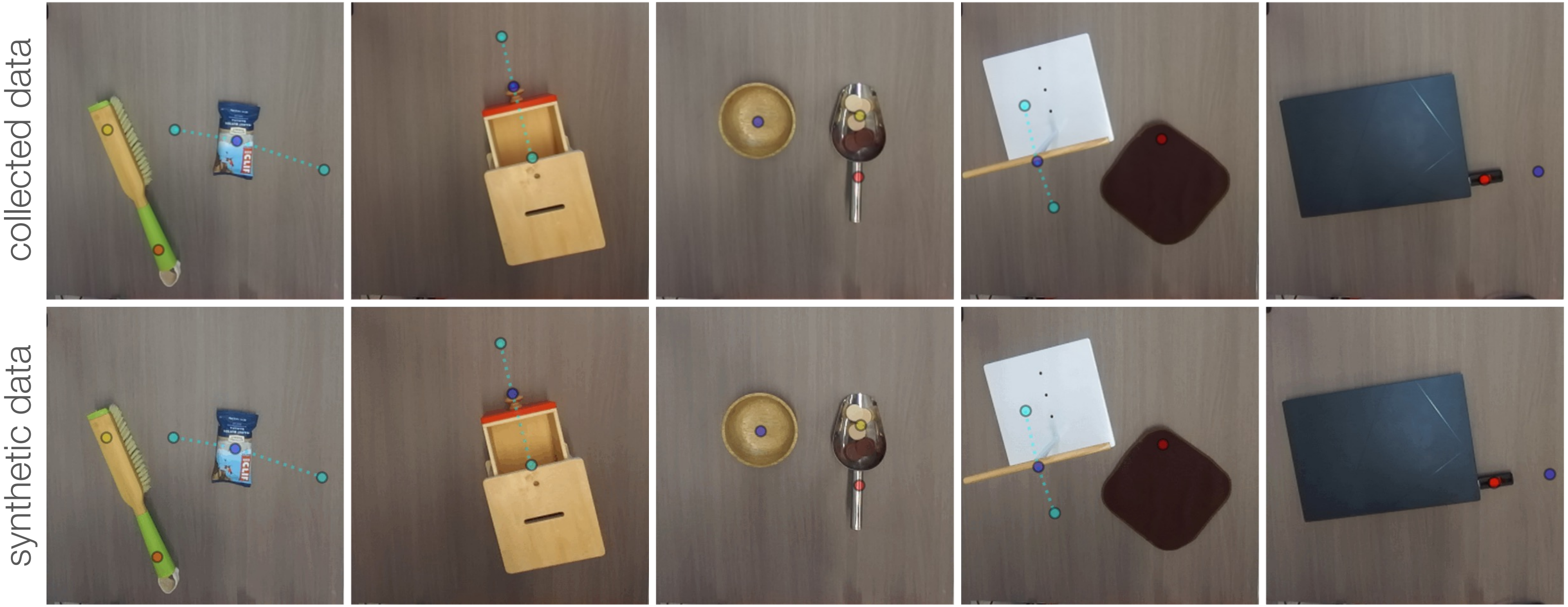
KALIE uses a **context image** as additional inputs to the diffusion model, which specifies the geometric properties of the object to be inpainted.



Tang, Rajkumar, Zhou, Walke, Levine, **Fang**. Fine-Tuning Vision-Language Models for Open-World Manipulation without Robot Data. ICRA 2025

Generated Data

- Employ conditional diffusion models to **diversify** the training data.
- **Fine-tune** the VLM to predict affordances through low-rank adaptation.



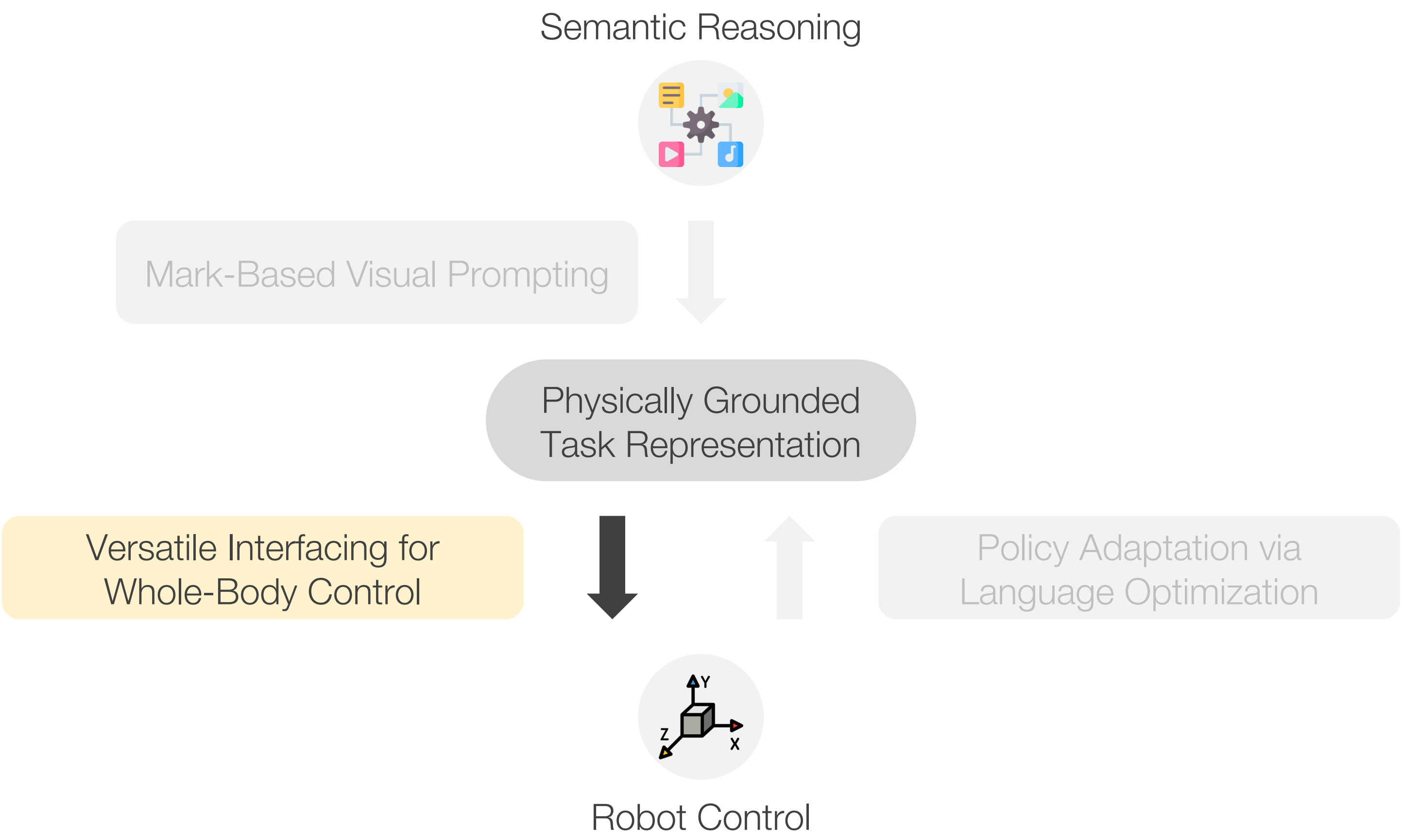
Tang, Rajkumar, Zhou, Walke, Levine, **Fang**. Fine-Tuning Vision-Language Models for Open-World Manipulation without Robot Data. ICRA 2025

Performance

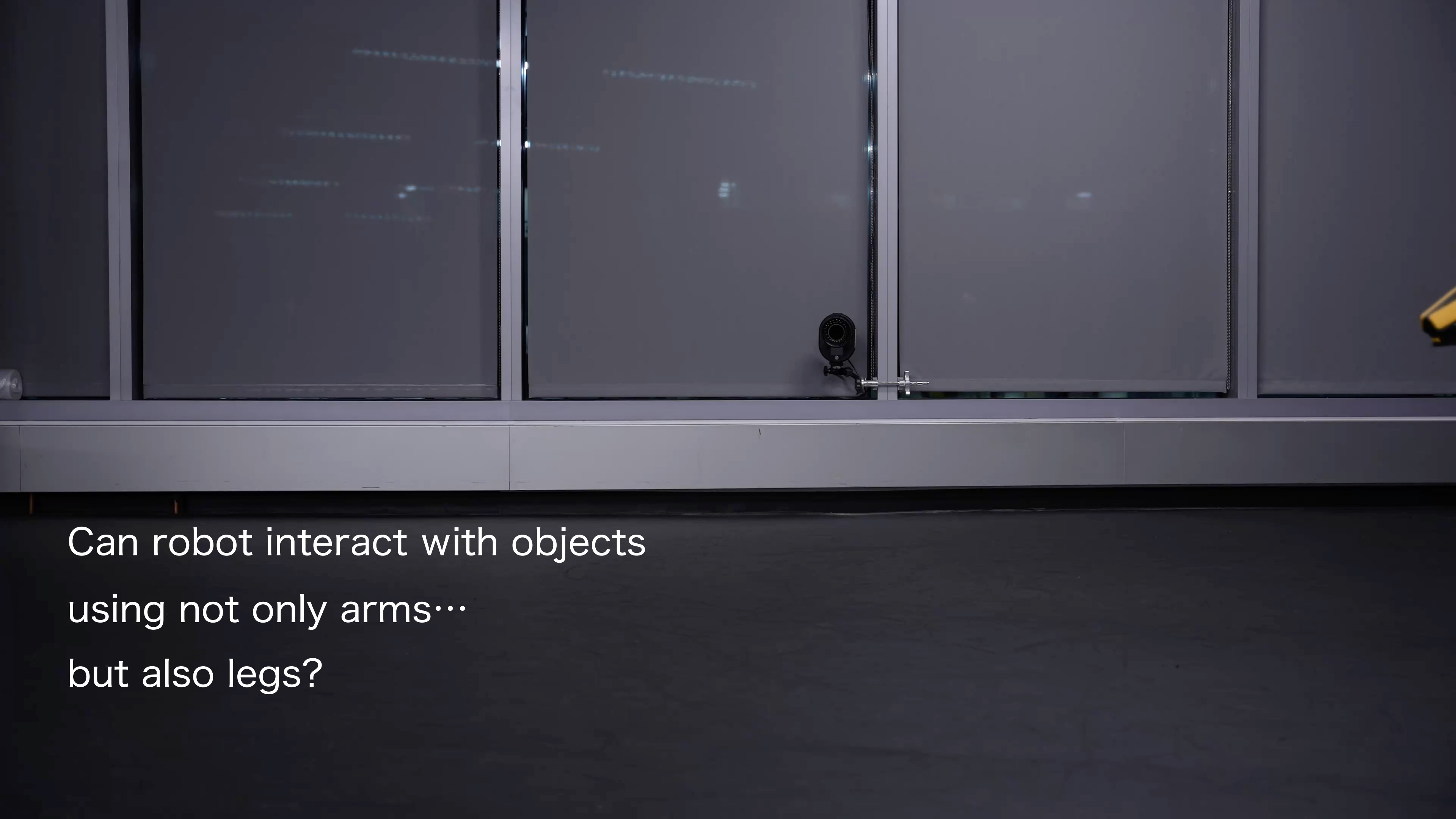
KALIE robustly solves these tasks and consistently achieves superior performances compared to baselines.

Methods	Table Sweeping	Drawer Closing	Towel Hanging	Trowel Pouring	USB Unplugging
VoxPoser [13]	3/15	8/15	1/15	0/15	0/15
MOKA [10]	9/15	9/15	5/15	7/15	2/15
KALIE (Ours)	**14**/15	**15**/15	**13**/15	**13**/15	**9**/15

Tang, Rajkumar, Zhou, Walke, Levine, Fang. Fine-Tuning Vision-Language Models for Open-World Manipulation without Robot Data. ICRA 2025


Semantic Reasoning
Mark-Based Visual Prompting
Physically Grounded
Task Representation
Versatile Interfacing for
Whole-Body Control
Policy Adaptation via
Language Optimization
Y
Z
X
Robot Control


Can robot dog perform task with both arm and legs?

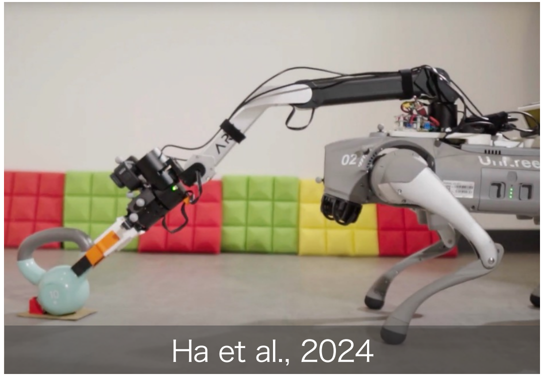

Can robot interact with objects
using not only arms…
but also legs?

Quadruped Loco-Manipulation with Arms and Legs

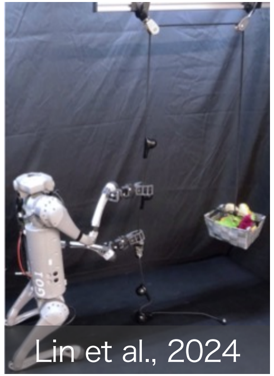
Manipulate with only the arm

Repurpose legs for manipulation


Liu et al., 2024


Ha et al., 2024


Cheng et al., 2023


Lin et al., 2024

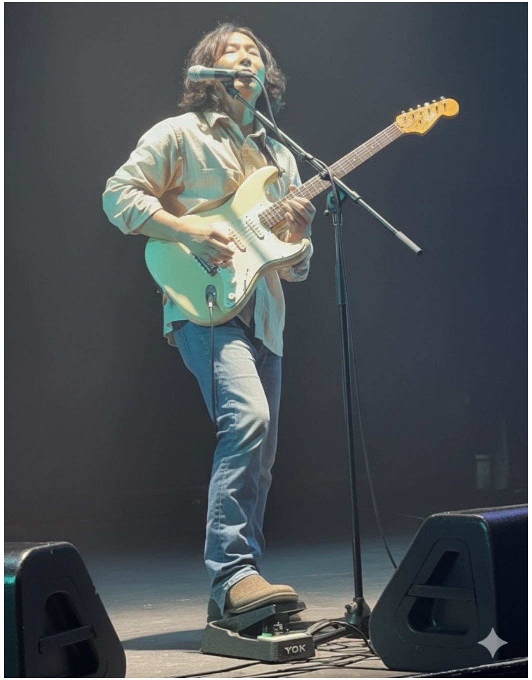
Fixed limb roles

Static limb coordination

Task-specific designs

Human Interlimb Coordination

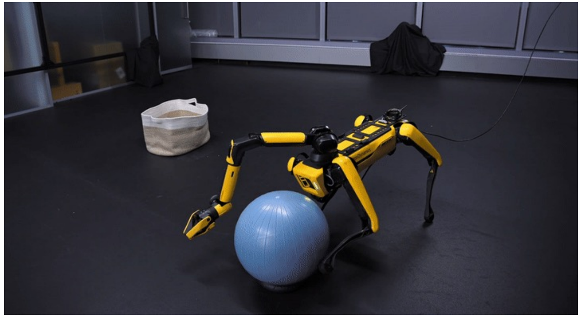
Humans can perform complex tasks by *jointly using multiple limbs*.



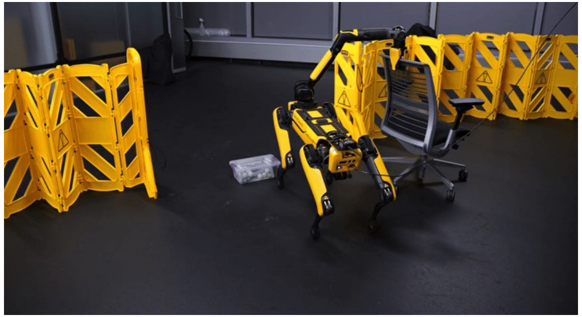
Images created with GenAI

Loco-Manipulation via Interlimb Coordination

By coordinating the arm and legs, we aim to enable the robot to:



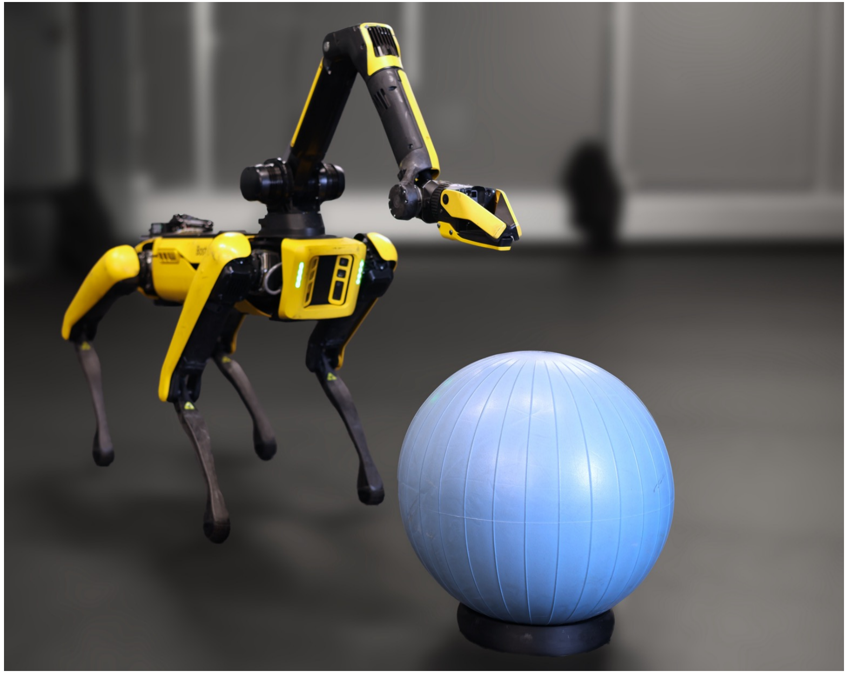
Manipulate with arm and leg while walking



Manipulate with arm and leg while standing

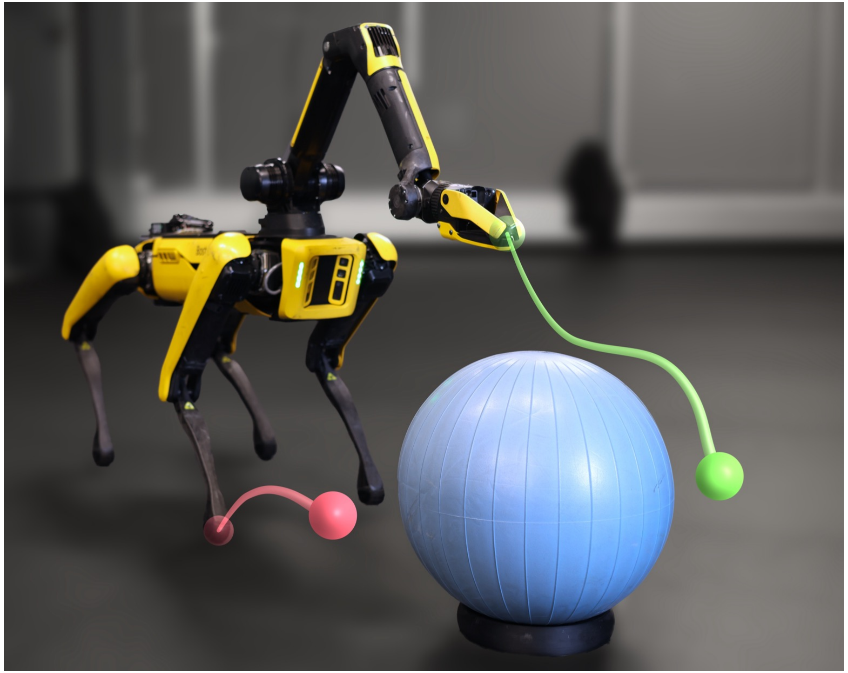
Assist or accelerate multi-step tasks with legs

Loco-Manipulation via Interlimb Coordination



Task: Transporting the yoga ball to the other side of the room

Loco-Manipulation via Interlimb Coordination



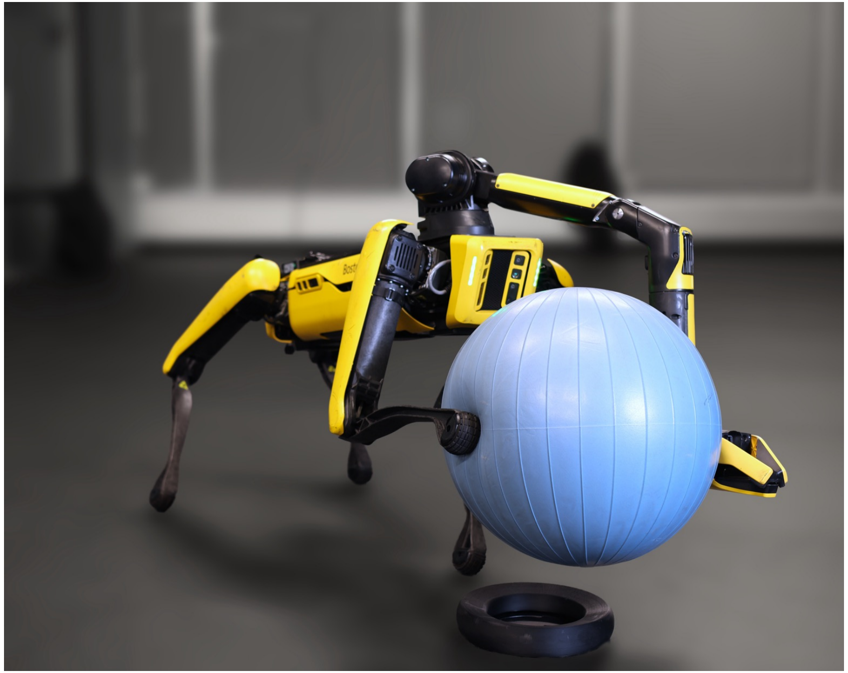
Given assigned roles of limbs and their target trajectories

Loco-Manipulation via Interlimb Coordination

Given assigned roles of limbs and their target trajectories,

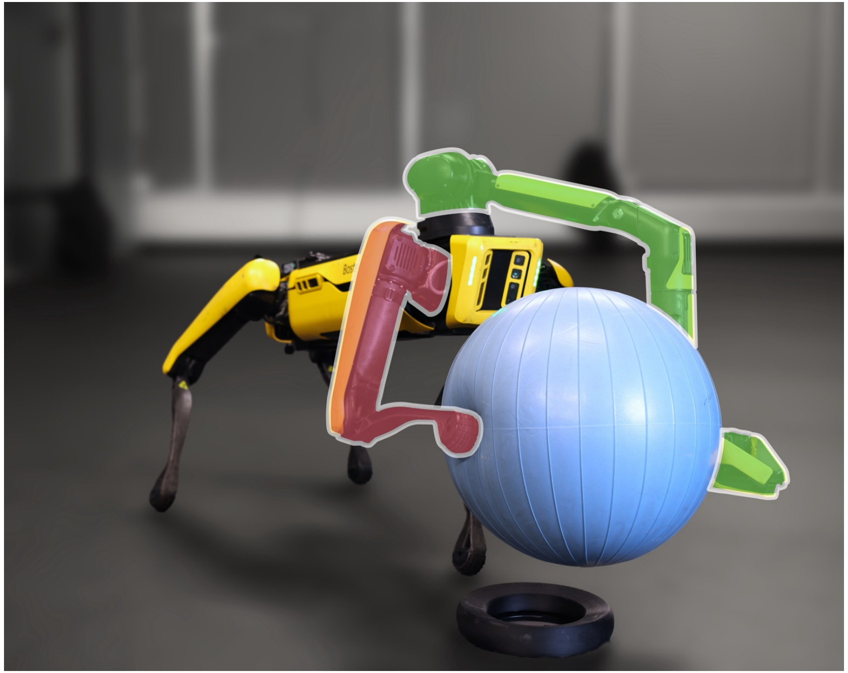
jointly control the arm and legs to solve the task

Loco-Manipulation via Interlimb Coordination



Key Challenge

Loco-Manipulation via Interlimb Coordination



Key Challenge

Precisely perform manipulation with the arm and a selected leg

Loco-Manipulation via Interlimb Coordination

Key Challenge

Precisely perform manipulation with the arm and a selected leg,

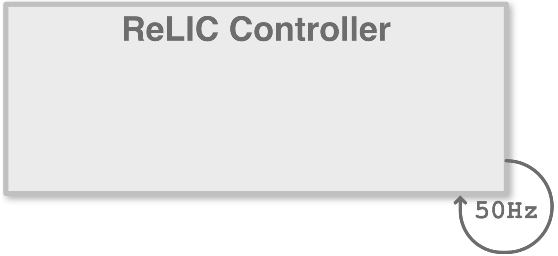
while maintaining stable locomotion with the remaining limbs

Loco-Manipulation via Interlimb Coordination

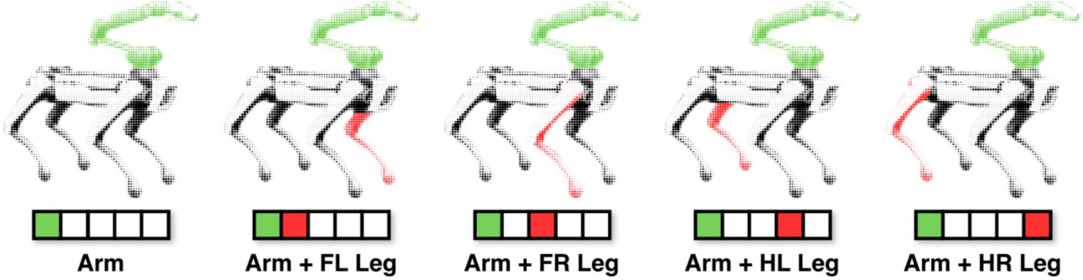
Our Aims

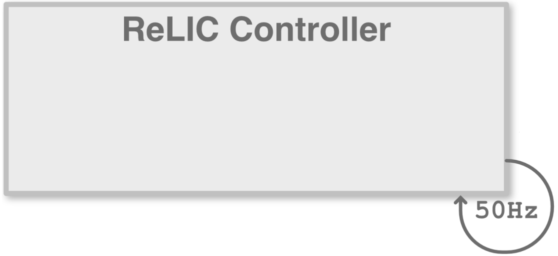
- ✓ Flexible coordination strategies
- ✓ Dynamic limb assignments
- ✓ Versatile task specifications

Reinforcement Learning for Interlimb Coordination (ReLIC)

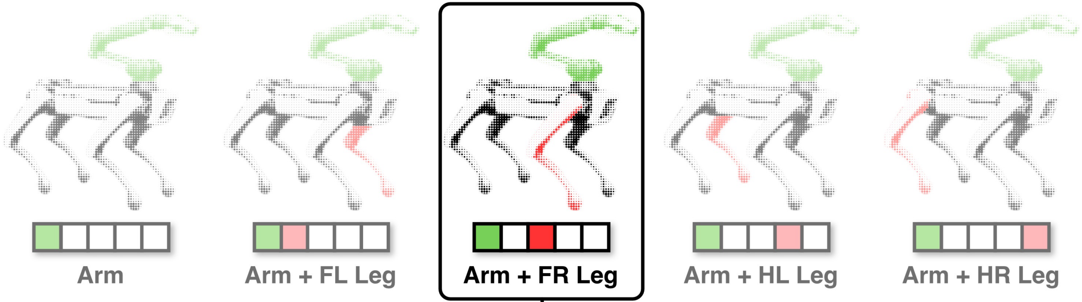


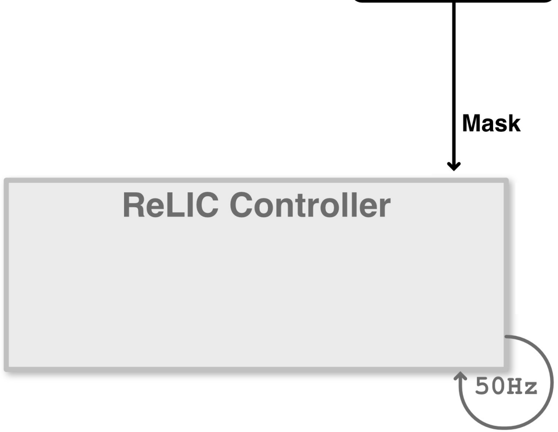
Reinforcement Learning for Interlimb Coordination (ReLIC)



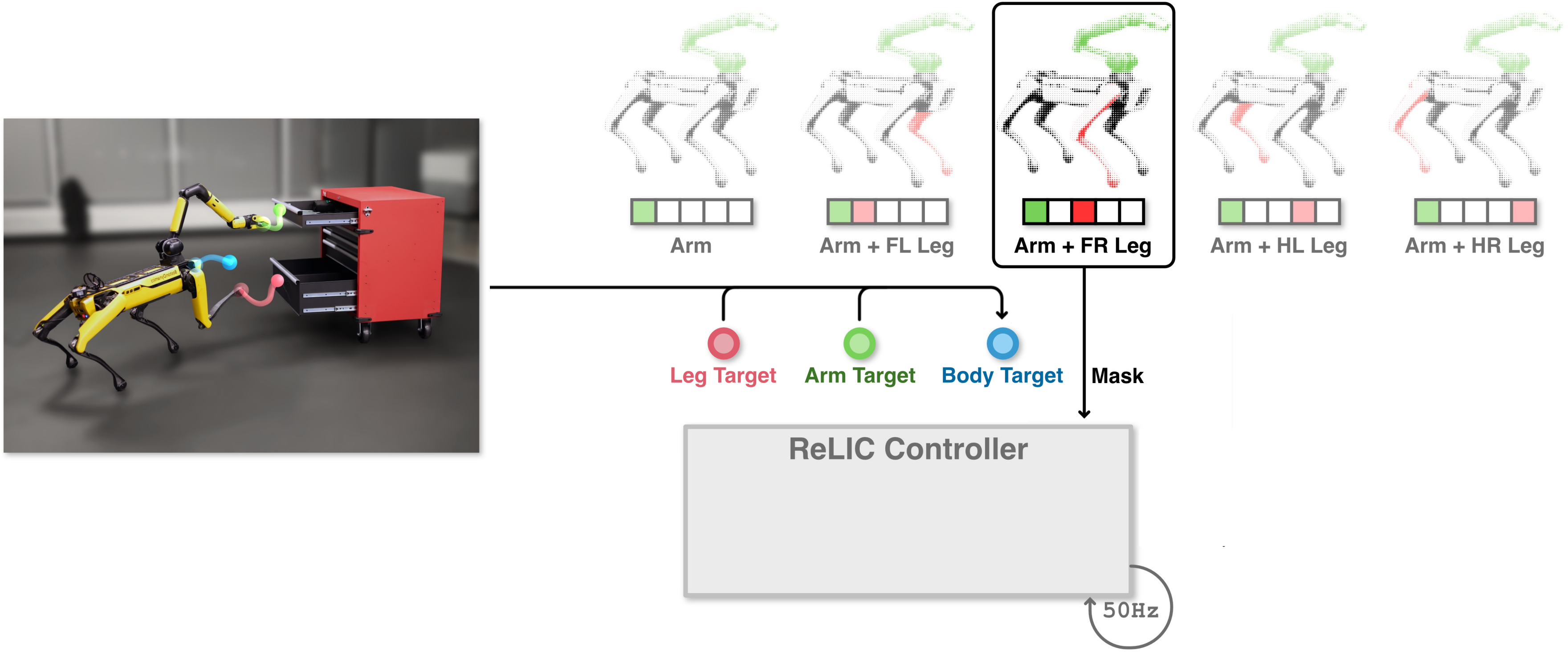


Reinforcement Learning for Interlimb Coordination (ReLIC)

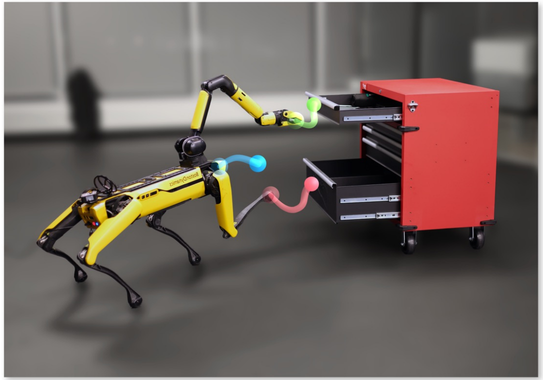


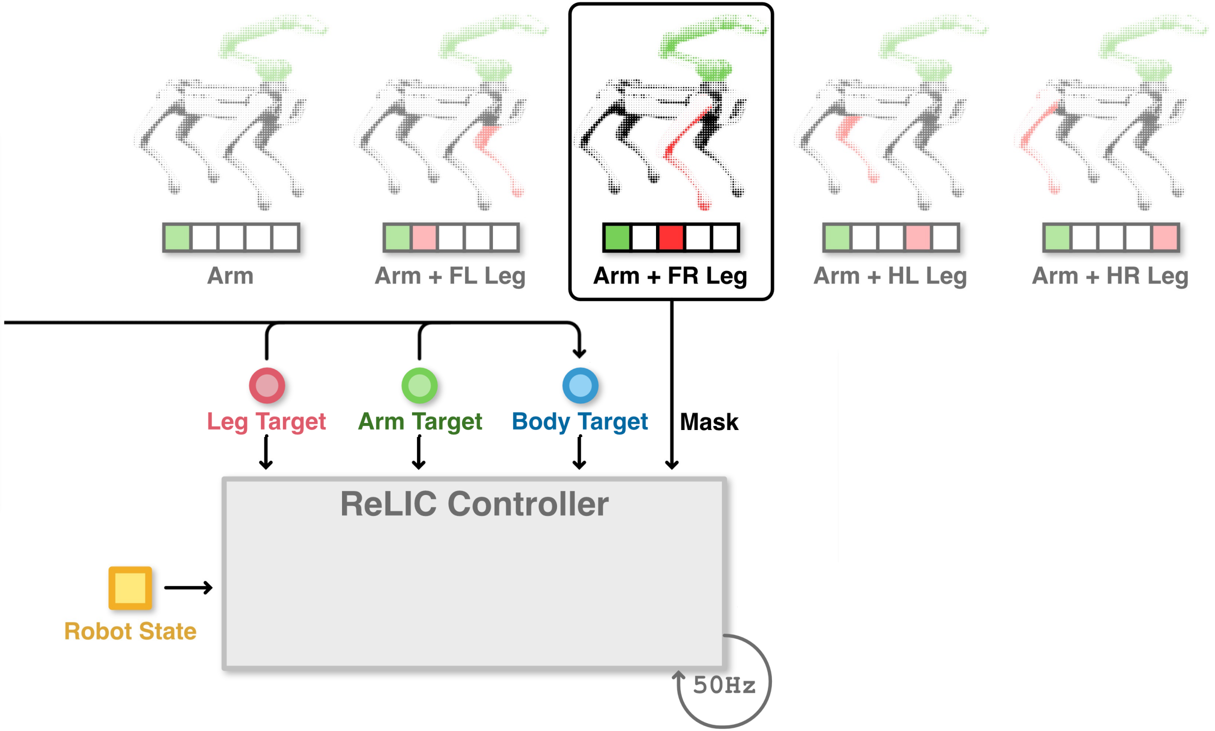


Reinforcement Learning for Interlimb Coordination (ReLIC)

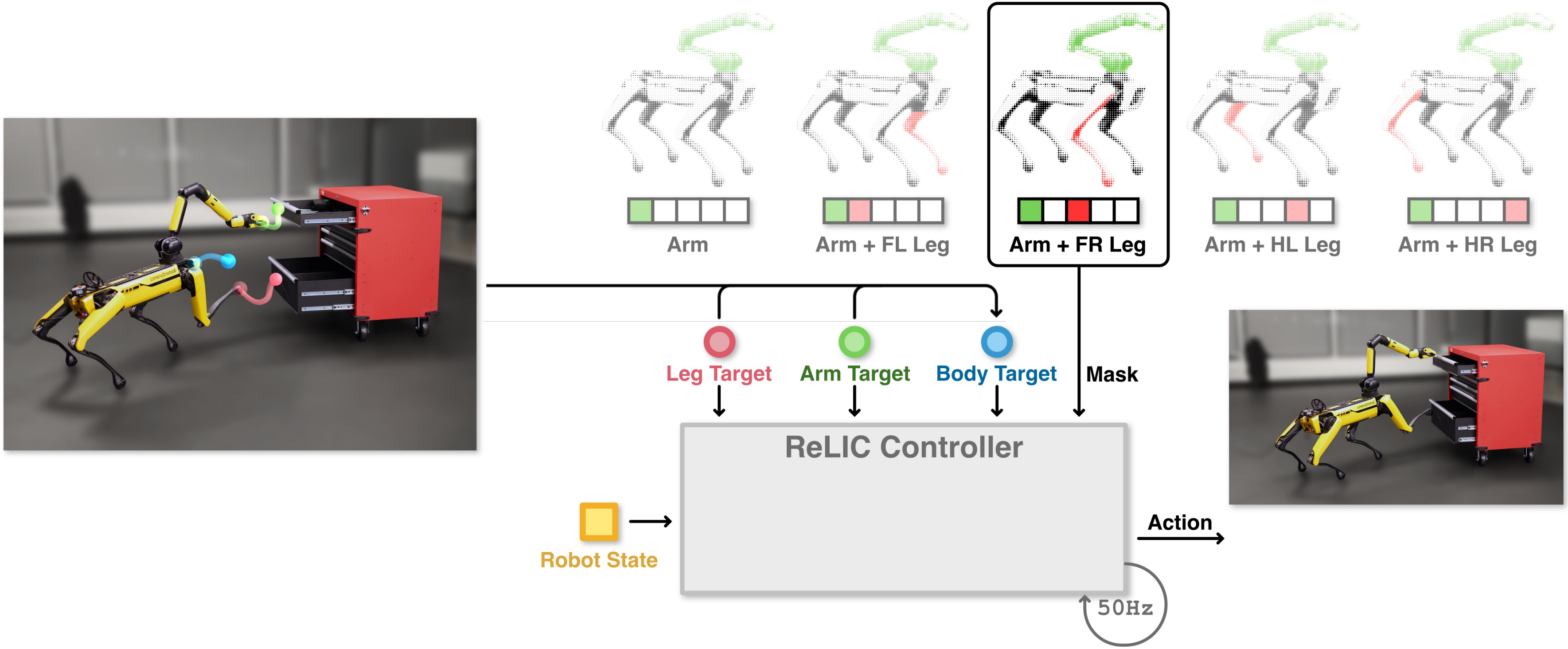


Reinforcement Learning for Interlimb Coordination (ReLIC)

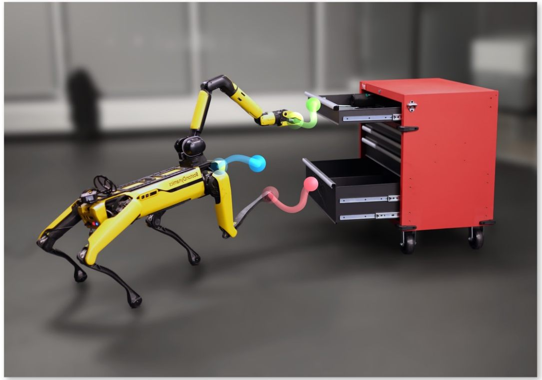


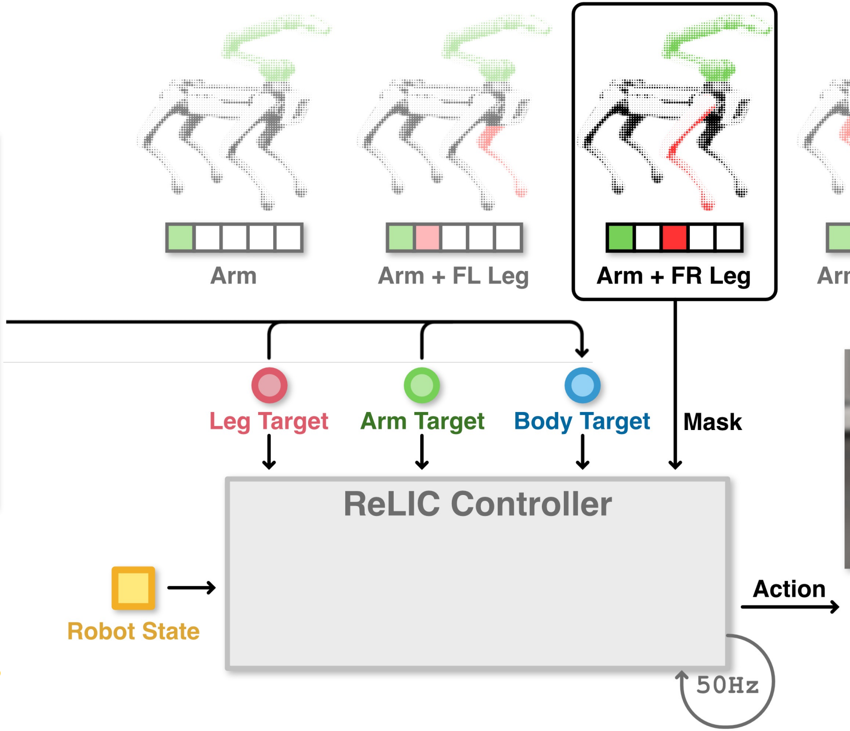


Reinforcement Learning for Interlimb Coordination (ReLIC)



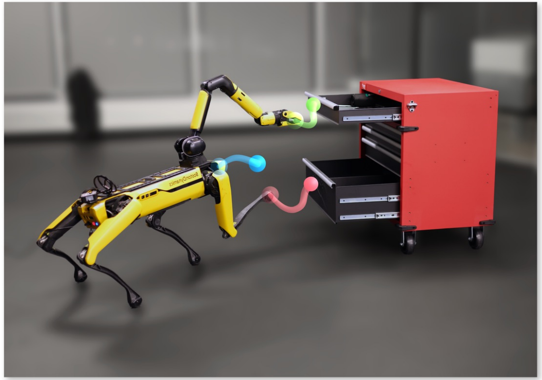
Reinforcement Learning for Interlimb Coordination (ReLIC)

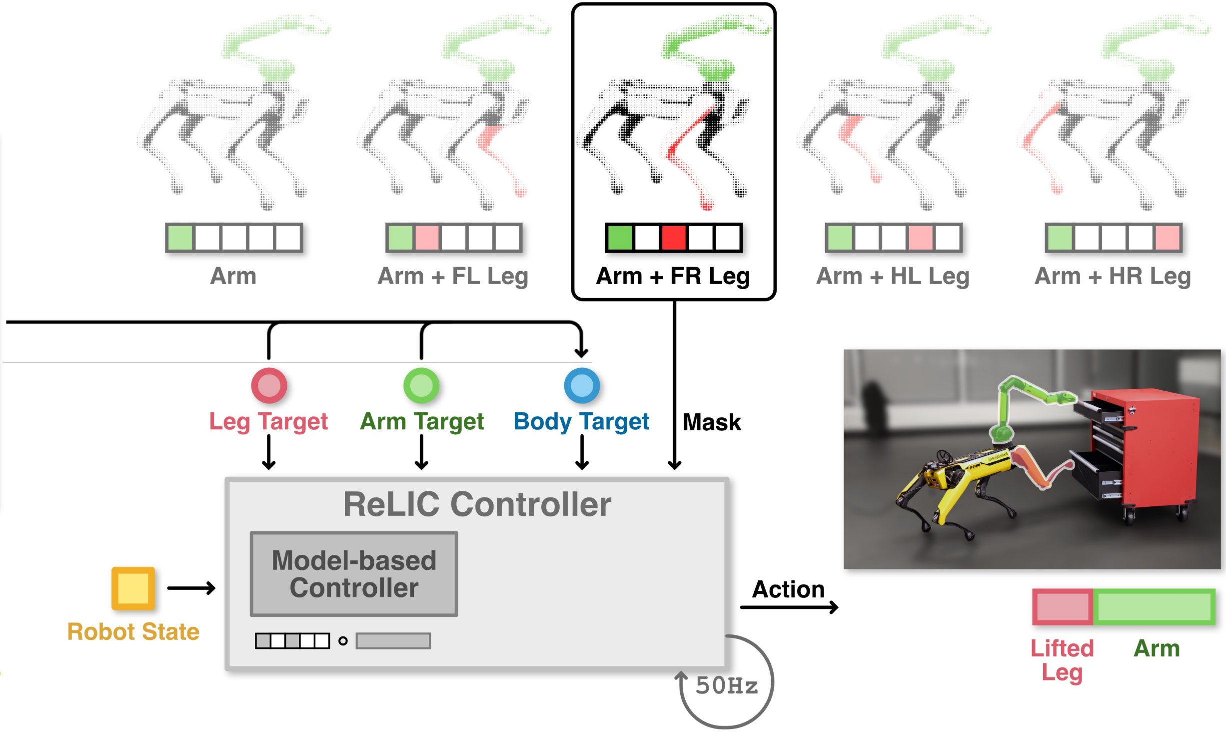




💡 Generate actions via the interplay of two modules

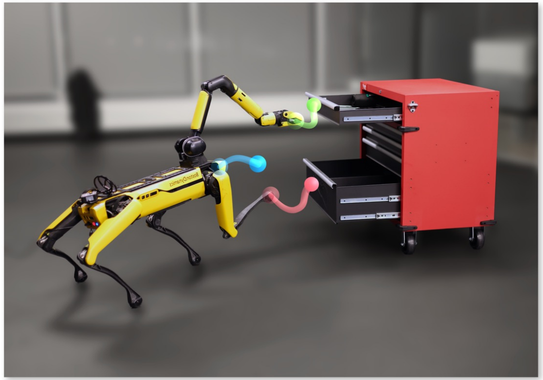
Reinforcement Learning for Interlimb Coordination (ReLIC)

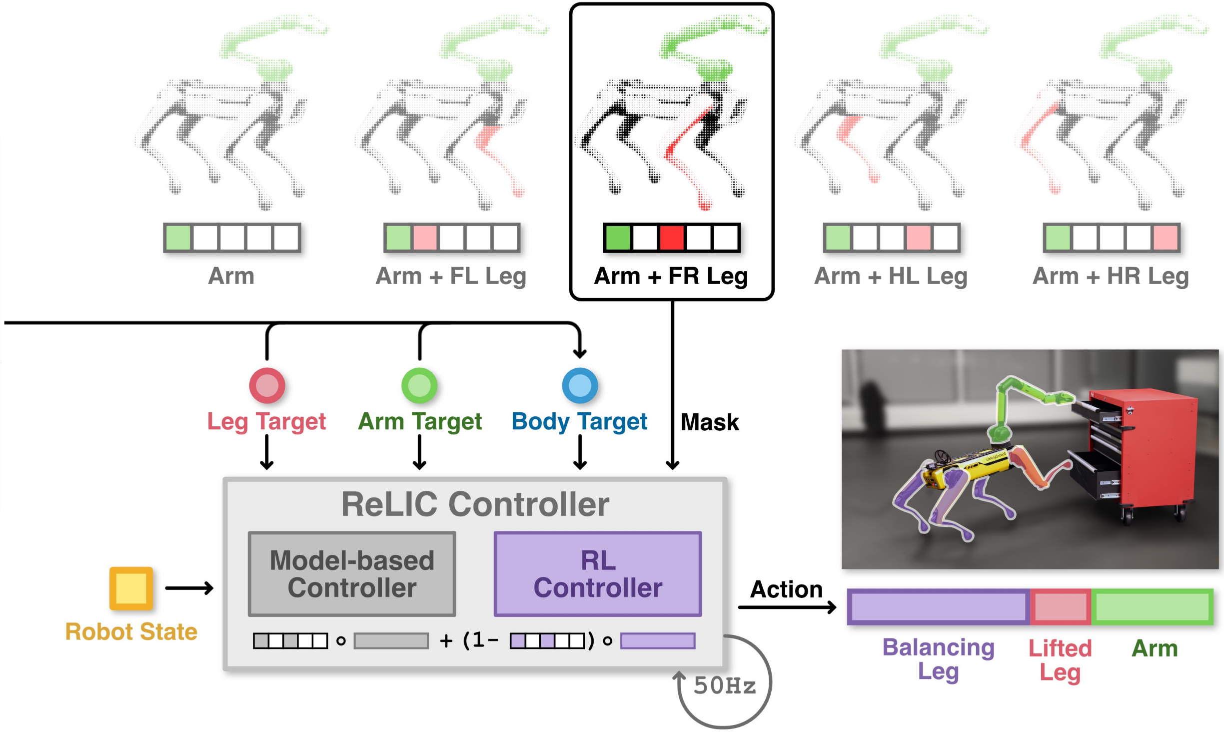




Generate actions via the interplay of two modules

Reinforcement Learning for Interlimb Coordination (ReLIC)

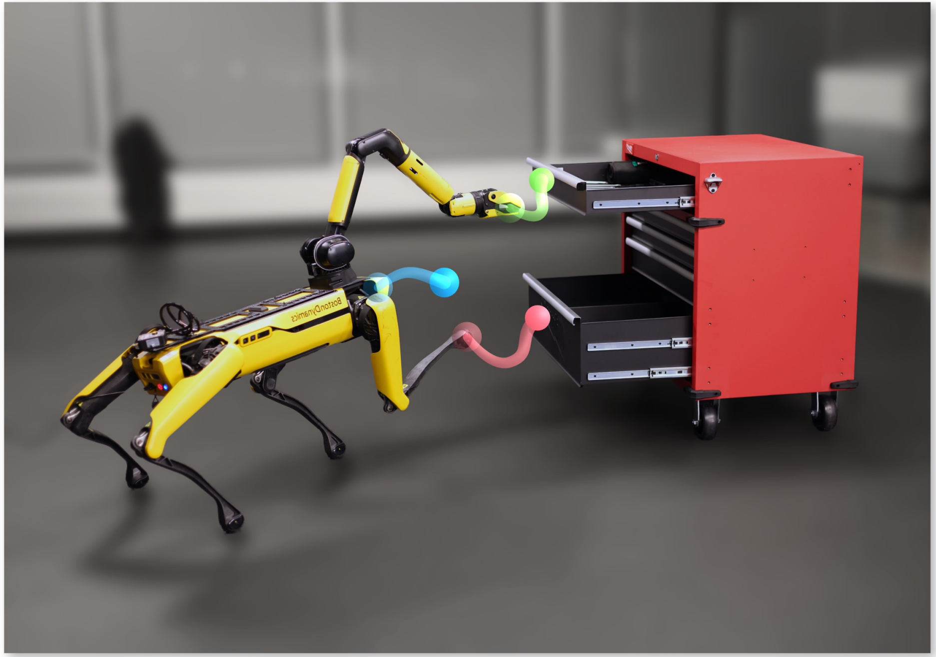


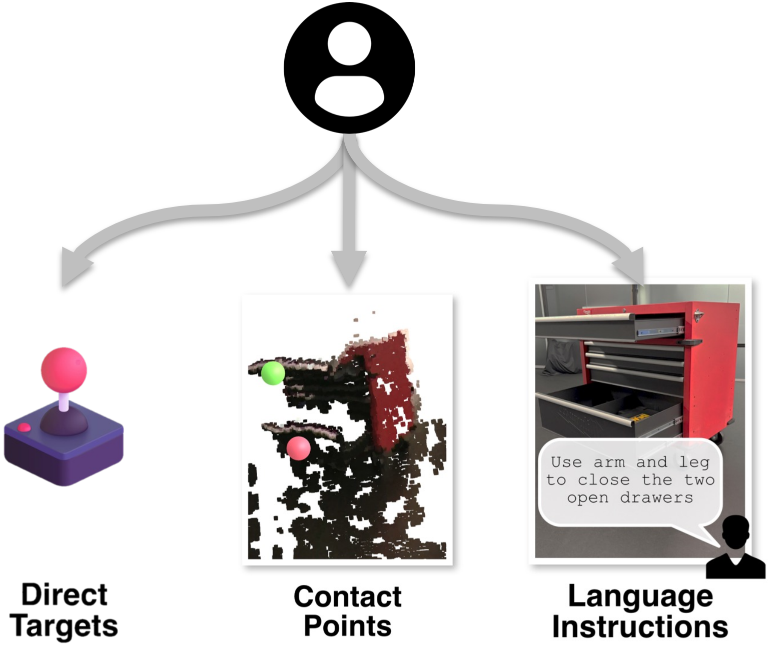


💡 Generate actions via the interplay of two modules

ReLIC
Task Interfaces

ReLIC can be interfaced with various types of user commands.





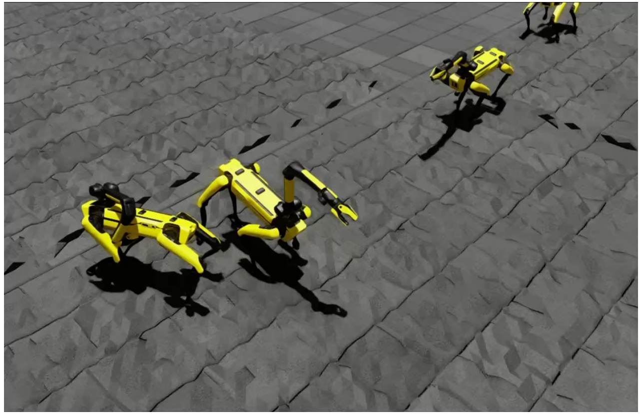

Task Interfaces
Direct Targets
BostonDynamics

Task Interfaces
Contact Points


Task Interfaces
Language Instructions
BostonDynamics

ReLIC
Learning Transferrable Policy in Simulation

Training in simulation



Deployment in the real world

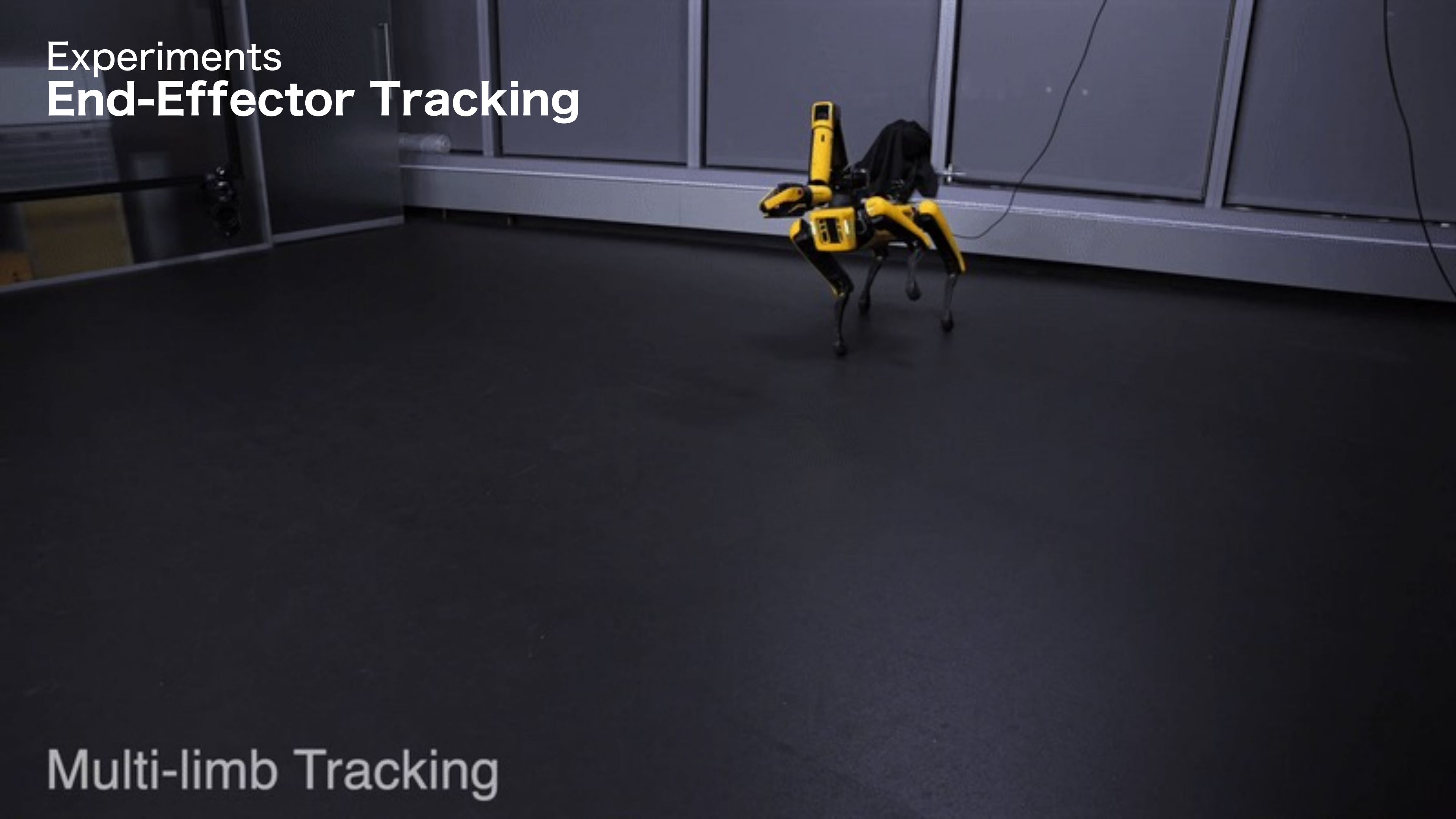
Motor calibration: Optimizes torque limits with CMA-ES[1] close the sim-to-real gap.

Gait regularization: Constraining contact-time patterns to stabilize locomotion.

[1]Nomura and Shibata. 2024

Experiments

End-Effector Tracking



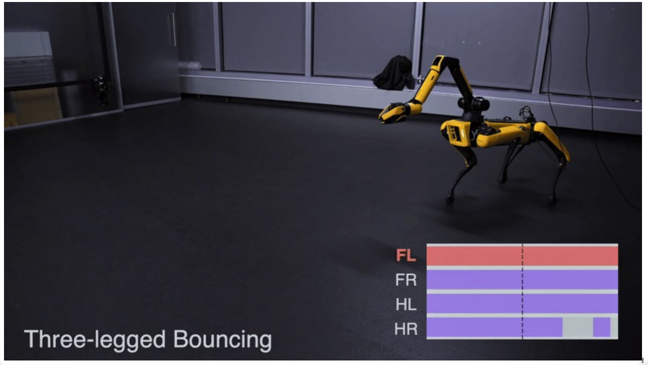
ReLIC

No Motor Calibration

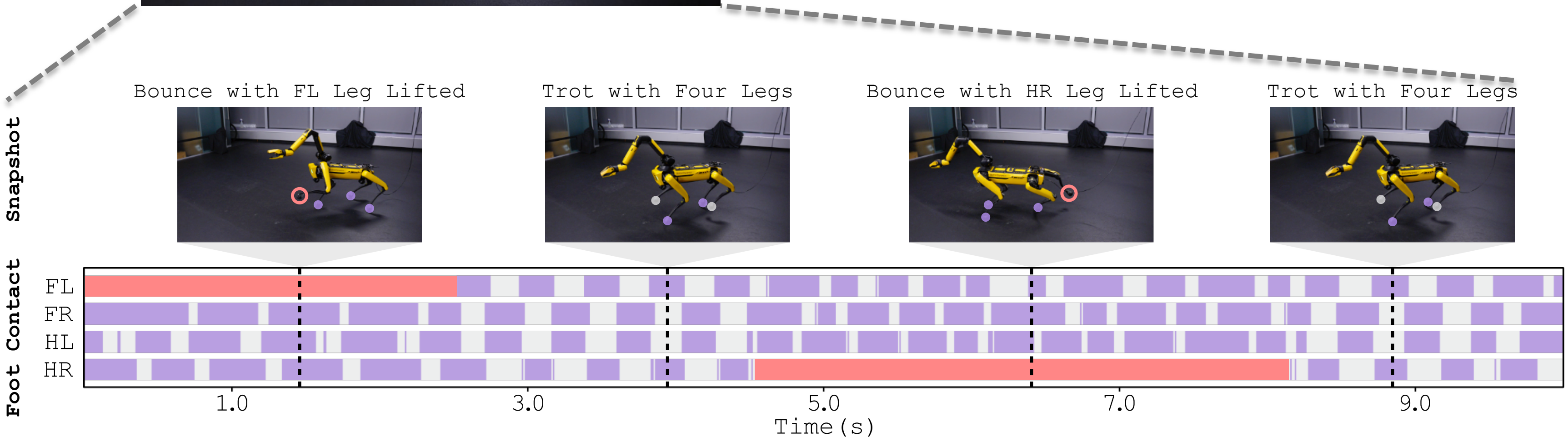
No Gait Regularization


Experiments
Gait Transitions
FL
FR
HL
HR
Three-legged Bouncing

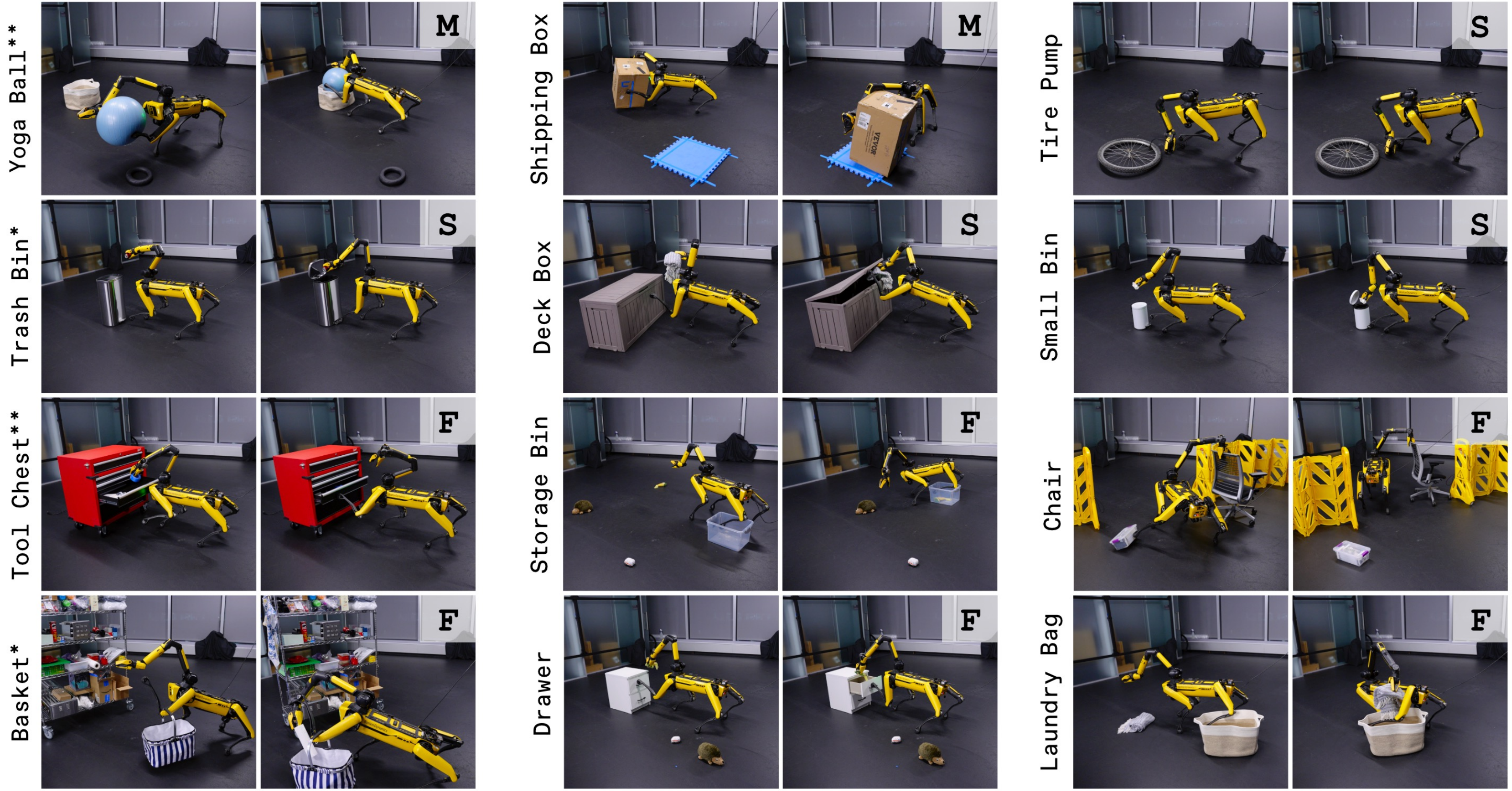
Experiments
Gait Transitions



Smoothly switching between different limb assignments ***without pausing or failing***



Experiments
Tasks

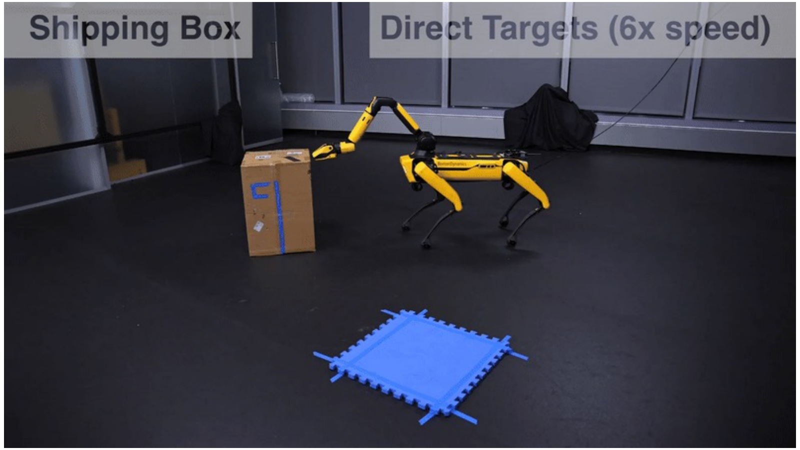


Experiments
Tasks: Mobile Interlimb Coordination

Manipulate with **arm and leg** while **walking**

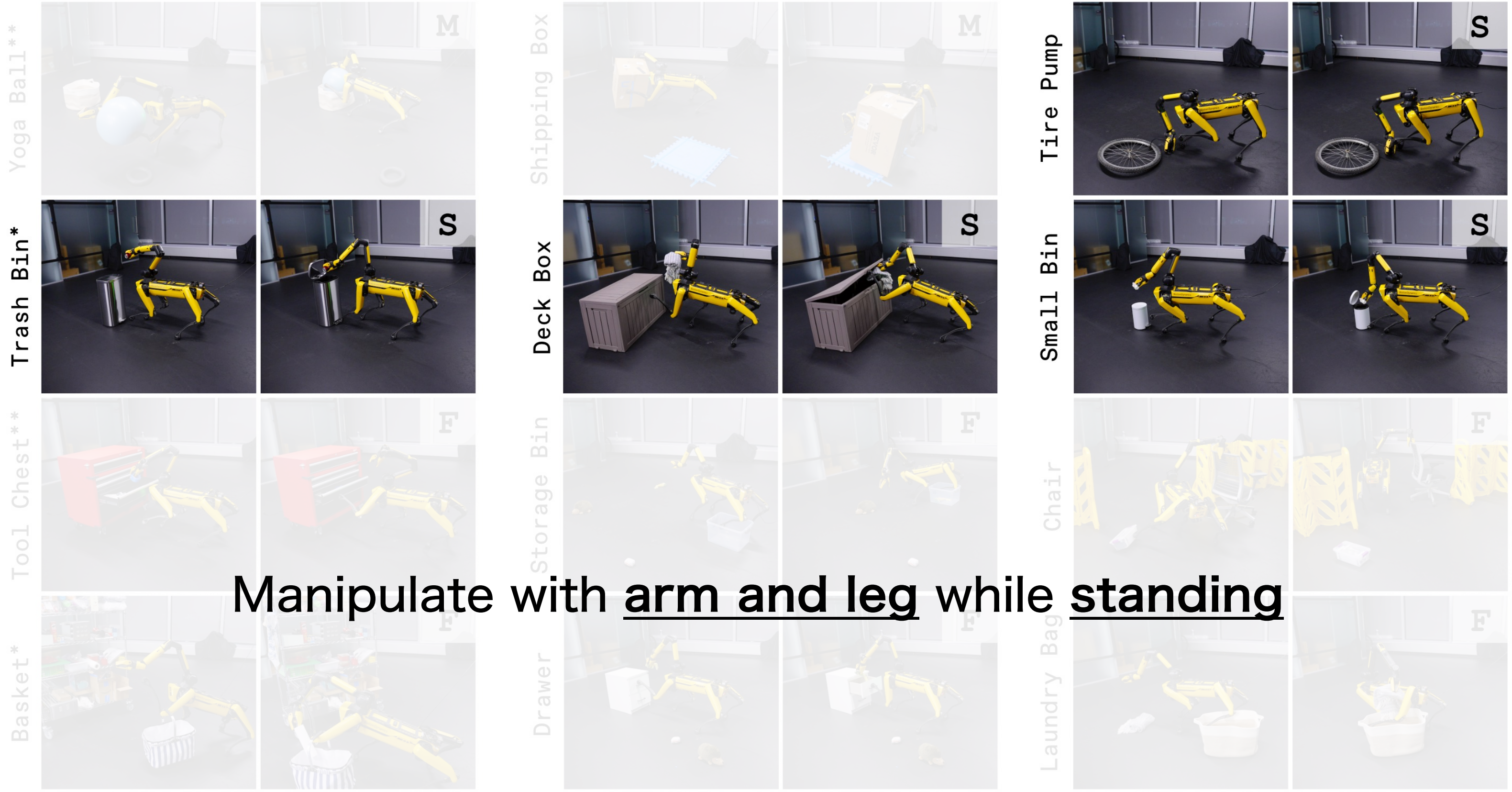
Experiments

Tasks: Mobile Interlimb Coordination



Experiments

Tasks: Stationary Interlimb Coordination



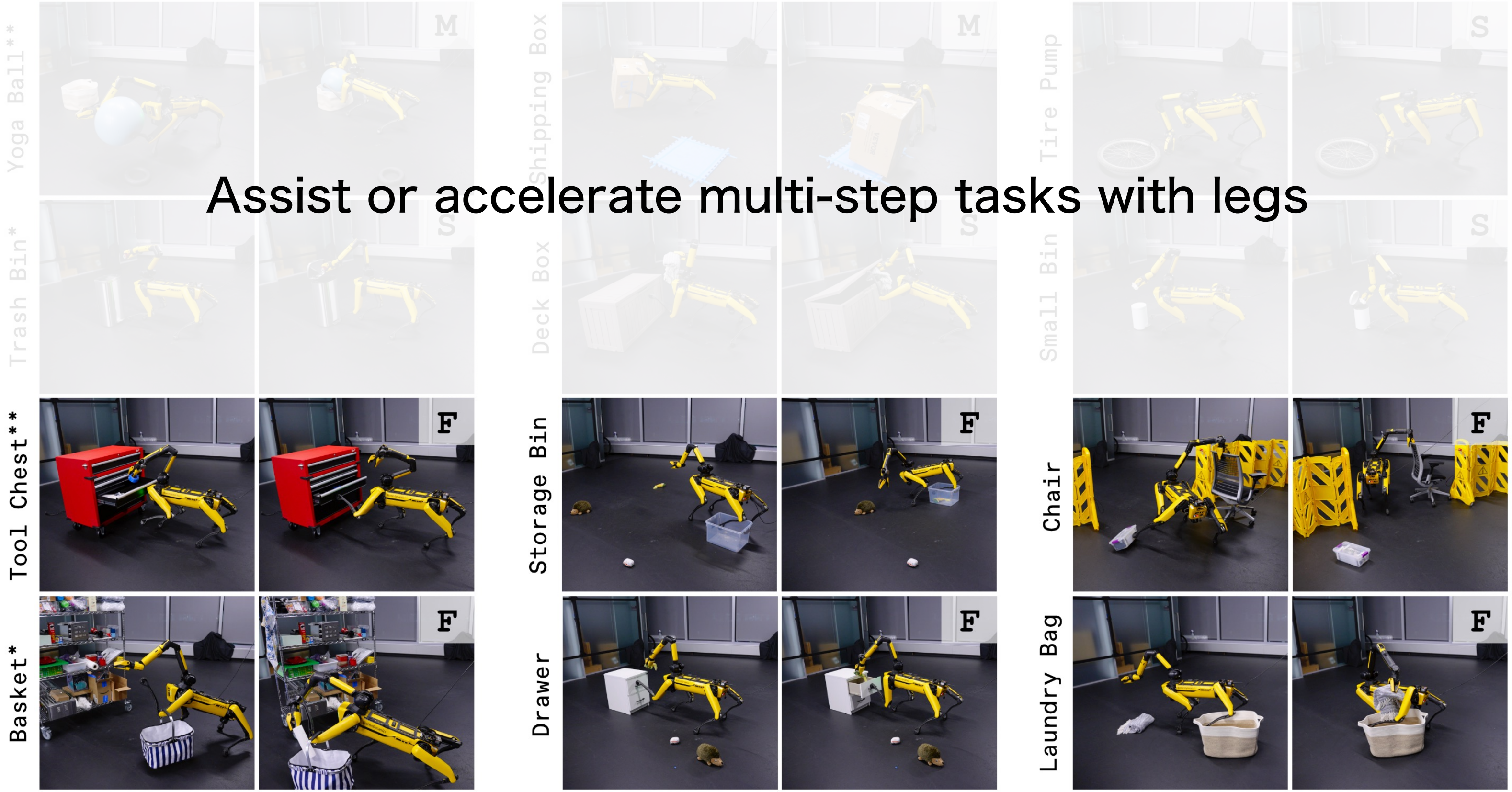
Manipulate with **arm and leg** while **standing**

Experiments

Tasks: Stationary Interlimb Coordination

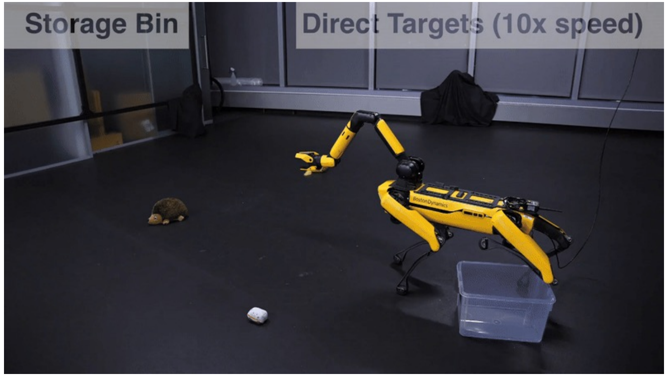
Experiments

Tasks: Foot-Assisted Manipulation



Assist or accelerate multi-step tasks with legs

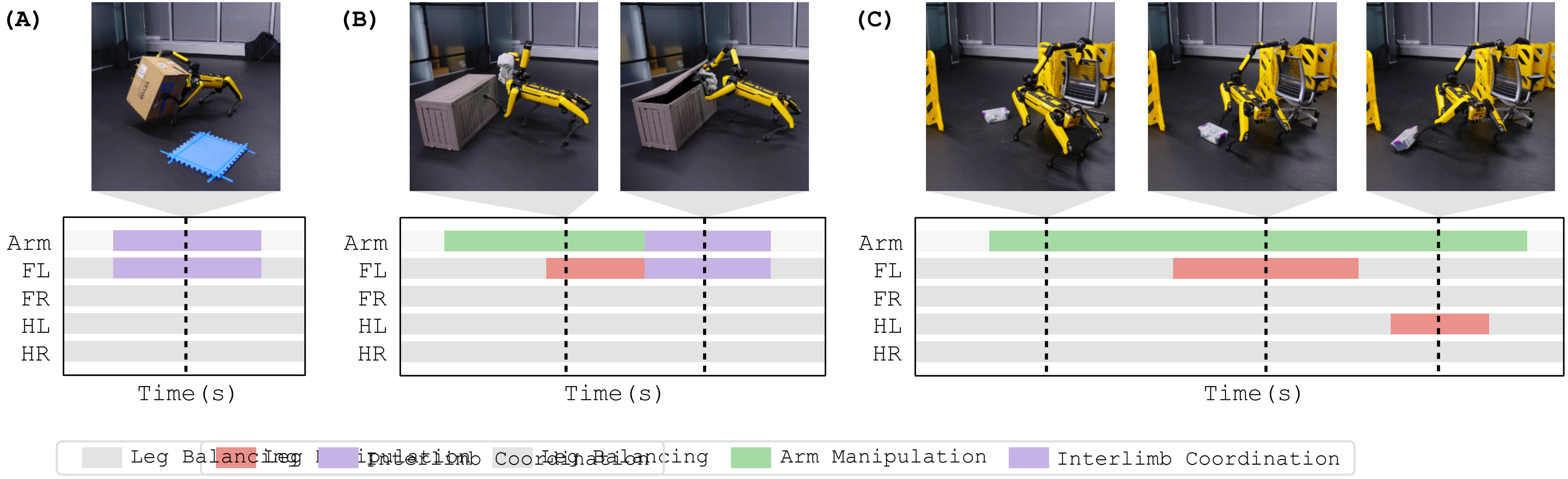
Experiments
Tasks: Foot-Assisted Manipulation



Experiments

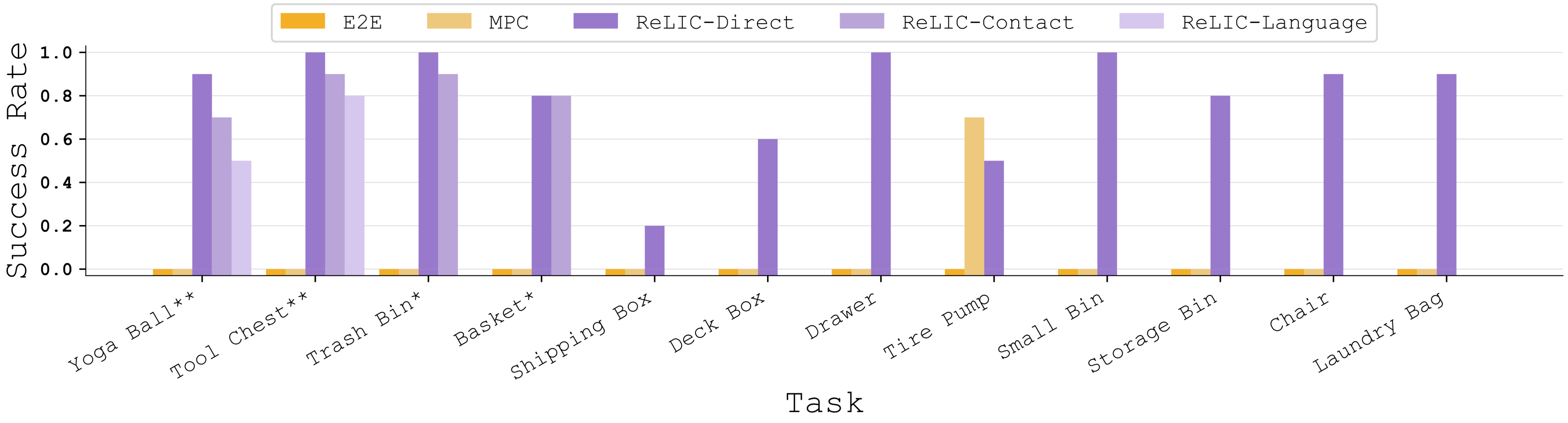
Dynamics Limb Assignments

Diverse assignment patterns are supported by ReLIC in these tasks.



Experiments

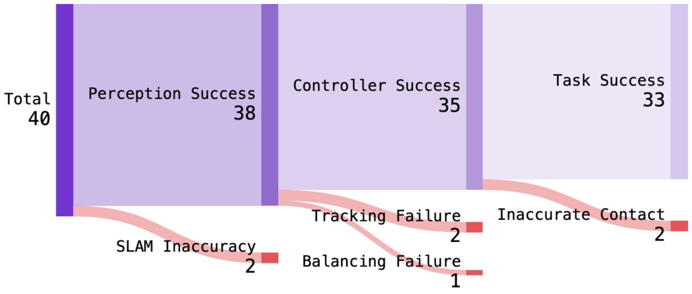
Comparative Results

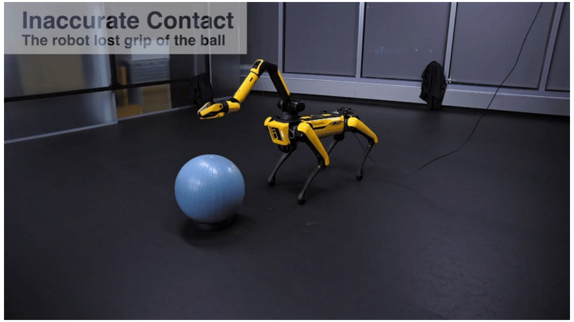


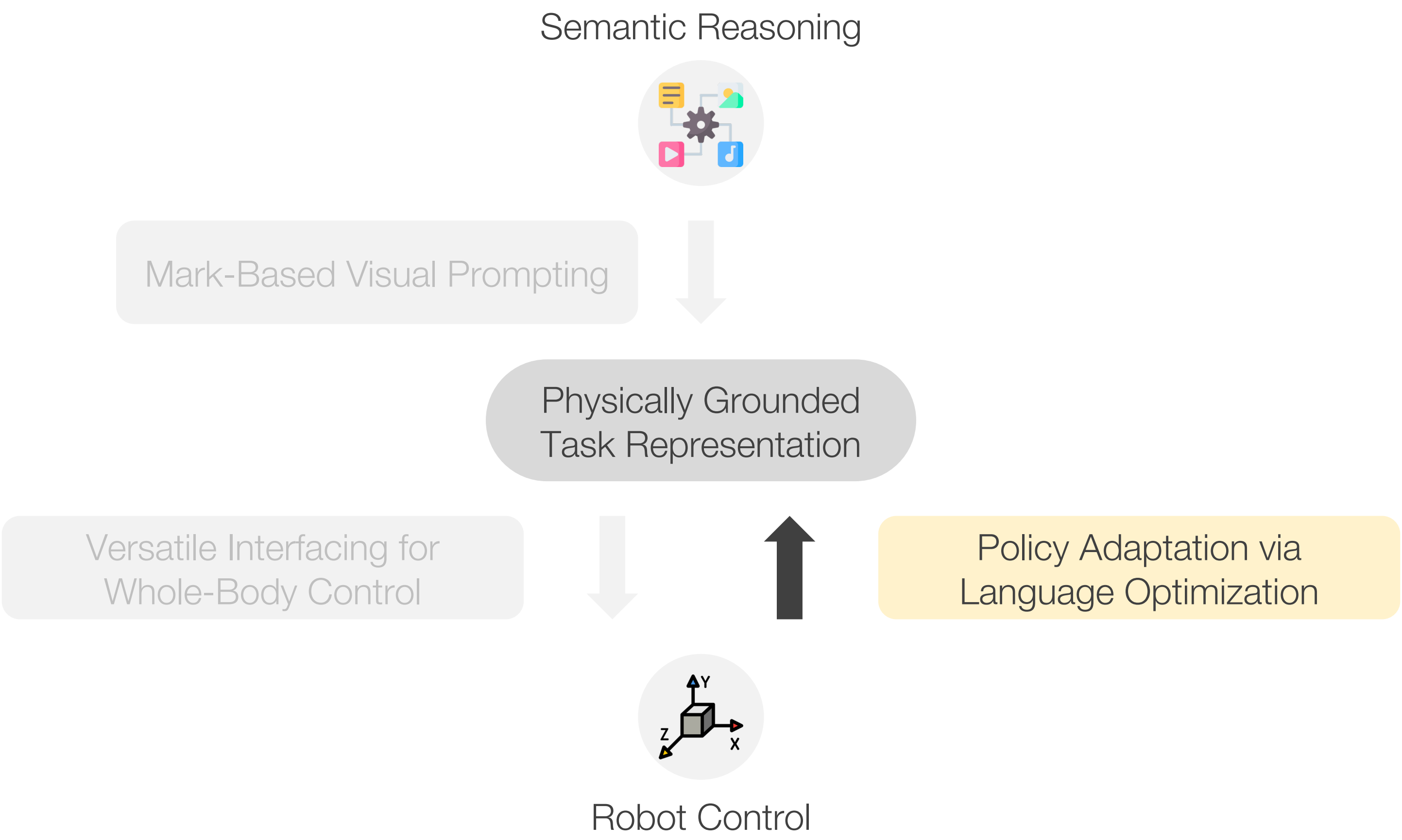
ReLIC achieves high success rates in most tasks, outperforming the end-to-end and MPC baselines

Experiments
Failure Analysis

We summarize failure cases in three categories:






Semantic Reasoning
Mark-Based Visual Prompting
Physically Grounded Task Representation
Versatile Interfacing for Whole-Body Control
Policy Adaptation via Language Optimization
Y
Z
X
Robot Control

Adaptation to new instruction-following tasks

Pour the coffee beans into the container

Massive offline dataset

pre-train

Pre-trained VLA policy

$$\pi(a|s, l; \theta)$$

Data for the new task

Fine-tune over policy parameters

Fine-tuning on each new task usually require 10^2 - 10^3 successful demos

Walke, Black, Lee, ..., **Fang**, Finn, Levine. BridgeData V2: A Dataset for Robot Learning at Scale. CoRL 2023

Adaptation to new instruction-following tasks

Pour the coffee beans into the container

Massive offline dataset

pre-train

Pre-trained VLA policy

$$\pi(a|s, l; \theta)$$

Data for the new task

💡 What if we instead adapt the instruction?

Walke, Black, Lee, ..., **Fang**, Finn, Levine. BridgeData V2: A Dataset for Robot Learning at Scale. CoRL 2023

Adaptation to new instruction-following tasks

Pre-trained VLA policy

$$\pi(a|s, l; \theta)$$

Pour the coffee beans into the container

Reach to the wooden tool on the table

Close the fingers

Lift the gripper upward by 10 cm

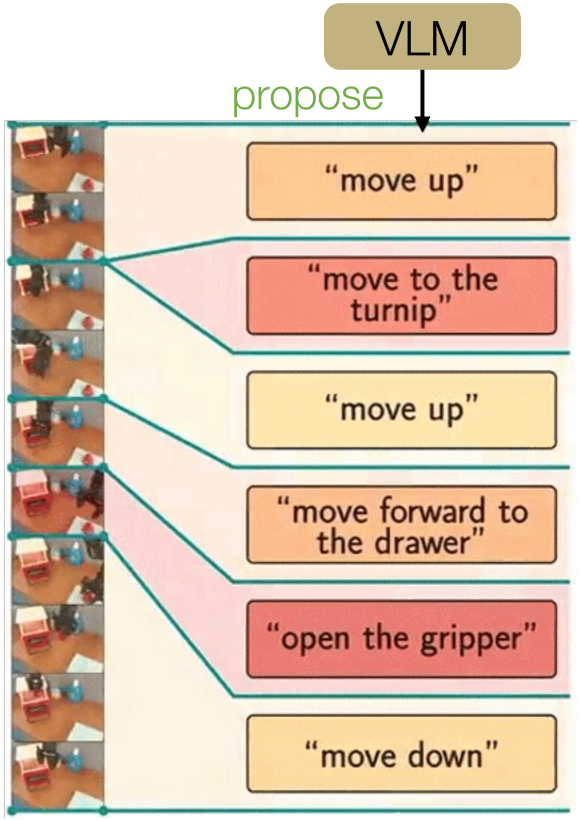
Move toward the blue bowl

Rotate the gripper by 30 degrees counterclockwise

Pick up the shovel

The phrasing of the instruction matters!

PALO: Policy Adaptation via Language Optimization



$$\pi(\hat{a}_t | s_t, c; \theta)$$

c_1 $\hat{a}_1$

c_2 $\hat{a}_2$

c_3 $\hat{a}_3$

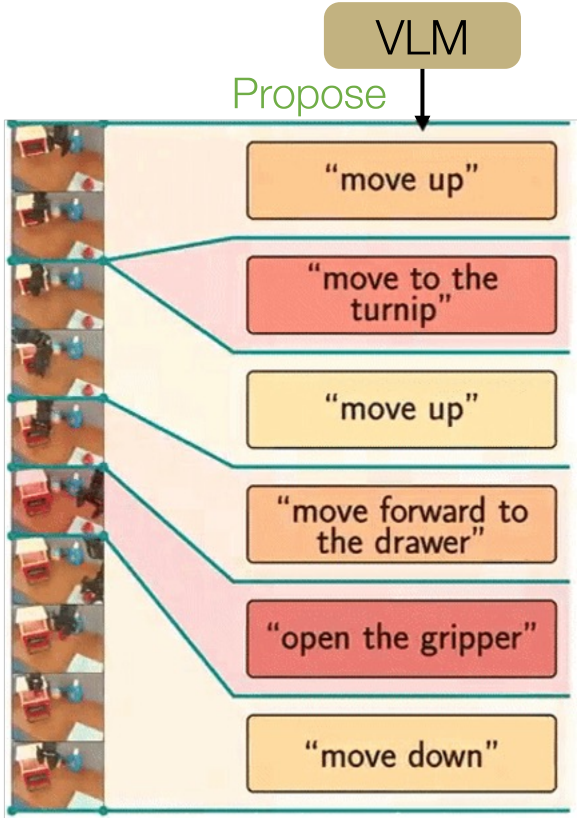
c_4 $\hat{a}_4$

c_5 $\hat{a}_5$

c_6 $\hat{a}_6$

Myers*, Zheng*, Mees, Levine†, **Fang**†. Policy Adaptation via Language Optimization: Decomposing Tasks for Few-Shot Imitation. CoRL 2024

PALO: Policy Adaptation via Language Optimization



c_1 c_2 c_3 c_4 c_5 c_6

$$\pi(\hat{a}_t | s_t, c; \theta)$$

Freeze

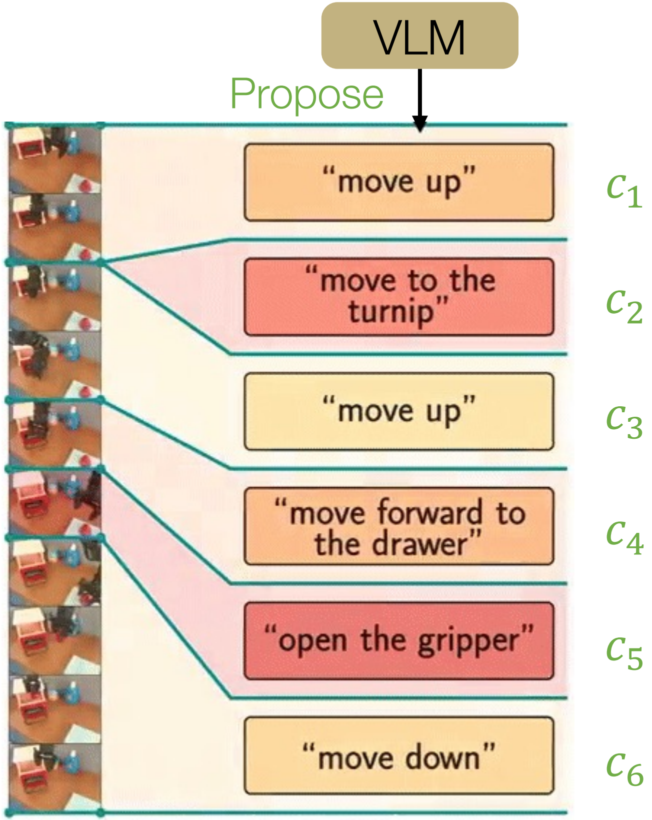
$\hat{a}_1$ $\hat{a}_2$ $\hat{a}_3$ $\hat{a}_4$ $\hat{a}_5$ $\hat{a}_6$

Optimize instruction sequences using behavior cloning loss

$$c^* = \arg\min_c \sum_t \|\hat{a}_t - a_t\|^2$$

Myers*, Zheng*, Mees, Levine†, **Fang**†. Policy Adaptation via Language Optimization: Decomposing Tasks for Few-Shot Imitation. CoRL 2024

PALO: Policy Adaptation via Language Optimization



$$\pi(\hat{a}_t | s_t, c; \theta)$$

Freeze

u: Subtask segmentation

$\hat{a}_1$

$\hat{a}_2$

$\hat{a}_3$

$\hat{a}_4$

$\hat{a}_5$

$\hat{a}_6$

Optimize instruction sequences using behavior cloning loss

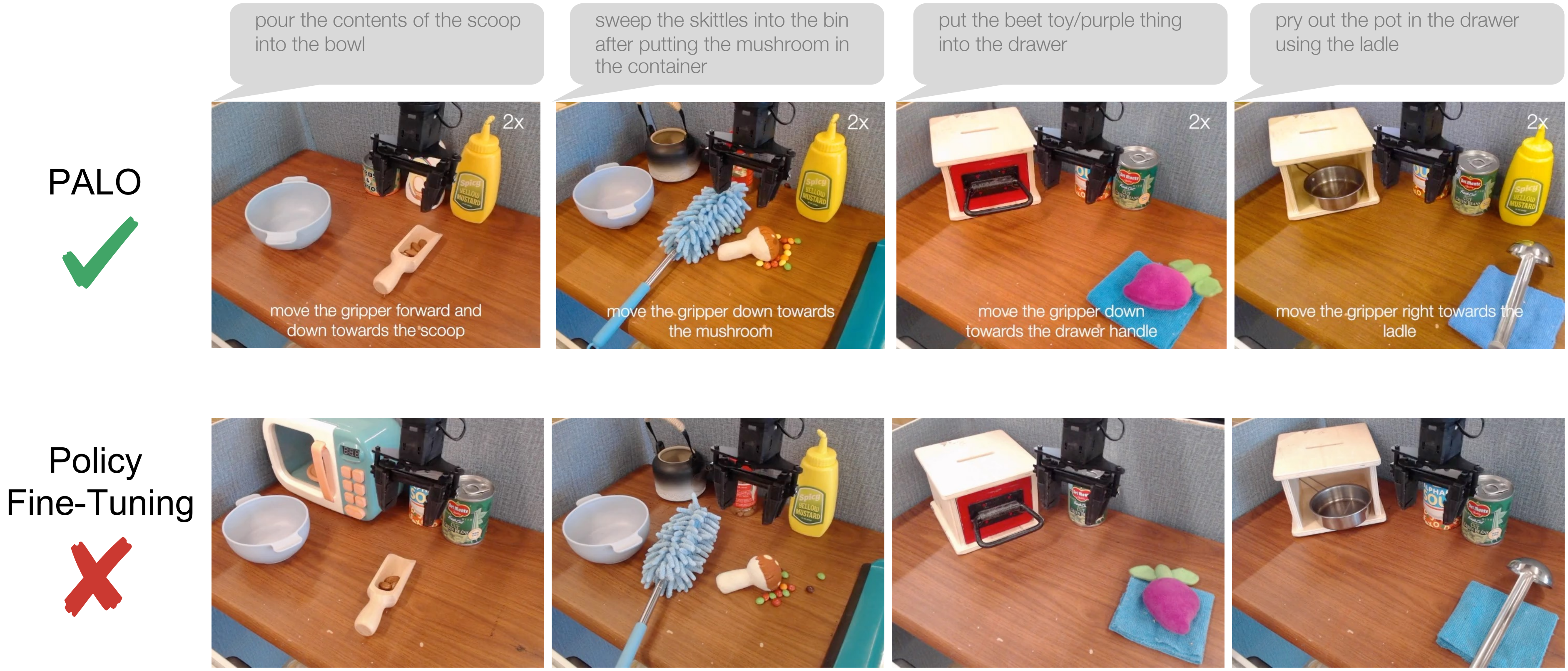
$$c^*, u^* = \arg\min_{c,u} \sum_t \|\hat{a}_t - a_t\|^2$$

Jointly optimize the temporal segmentation

similar to prompt tuning in NLP

Myers*, Zheng*, Mees, Levine†, **Fang**†. Policy Adaptation via Language Optimization: Decomposing Tasks for Few-Shot Imitation. CoRL 2024

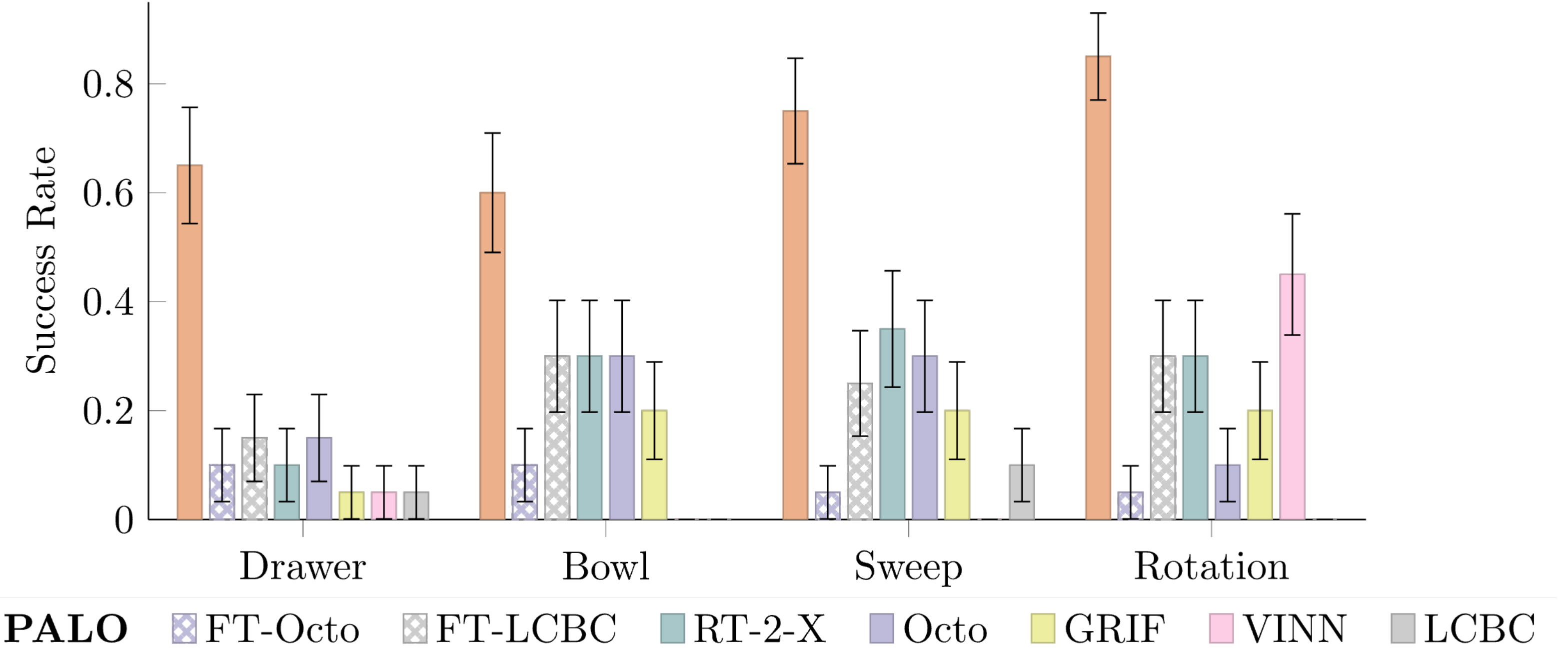
Given only 5 demos, PALO is able to robustly solve unseen, temporally extended tasks.



Myers*, Zheng*, Mees, Levine†, **Fang**†. Policy Adaptation via Language Optimization: Decomposing Tasks for Few-Shot Imitation. CoRL 2024

Comparative Results

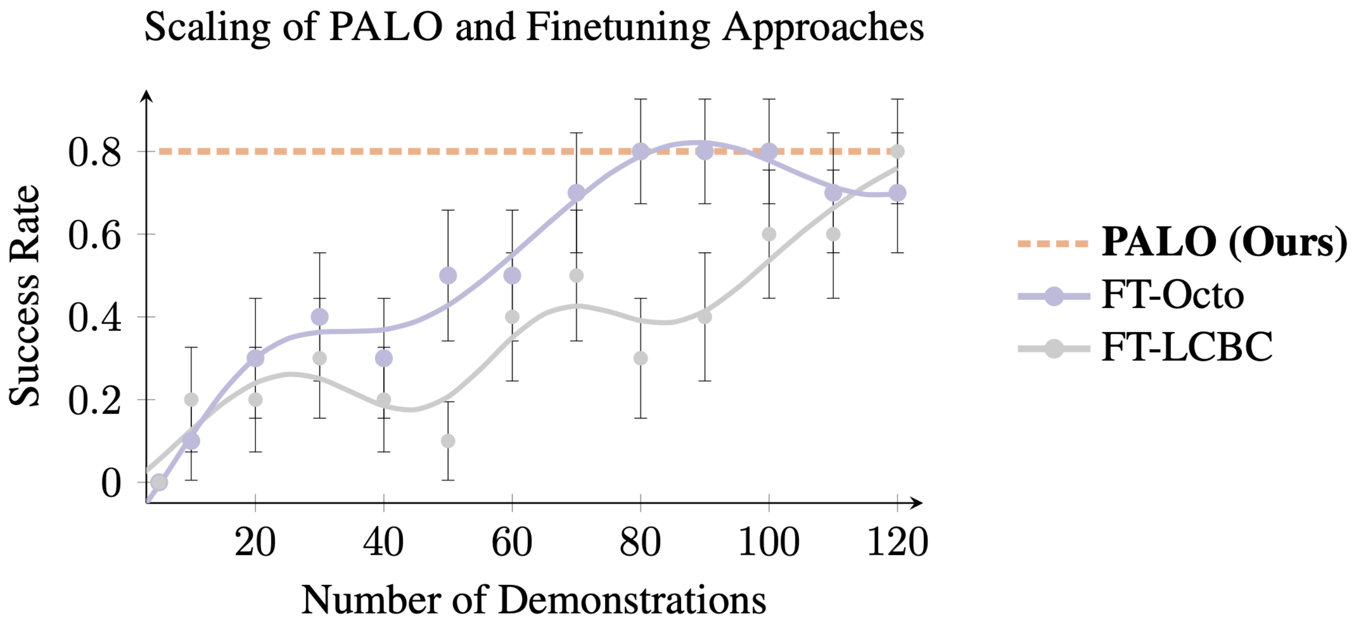
Evaluating on long-horizon and unseen skills tasks, PALO outperforms all conventional zero-shot generalization methods **by 3x** in terms success rate.



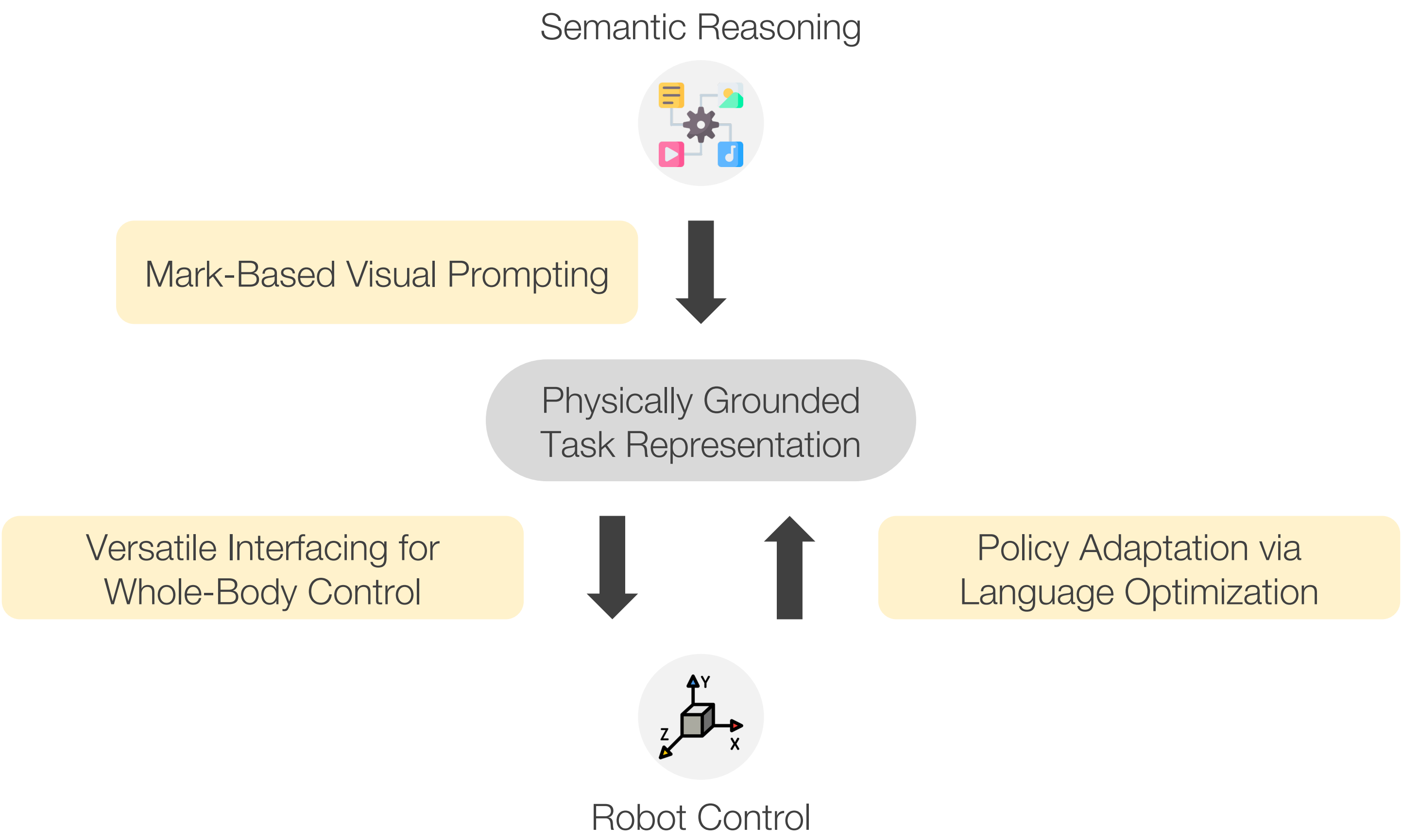
Myers*, Zheng*, Mees, Levine†, **Fang**†. Policy Adaptation via Language Optimization: Decomposing Tasks for Few-Shot Imitation. CoRL 2024

Comparative Results

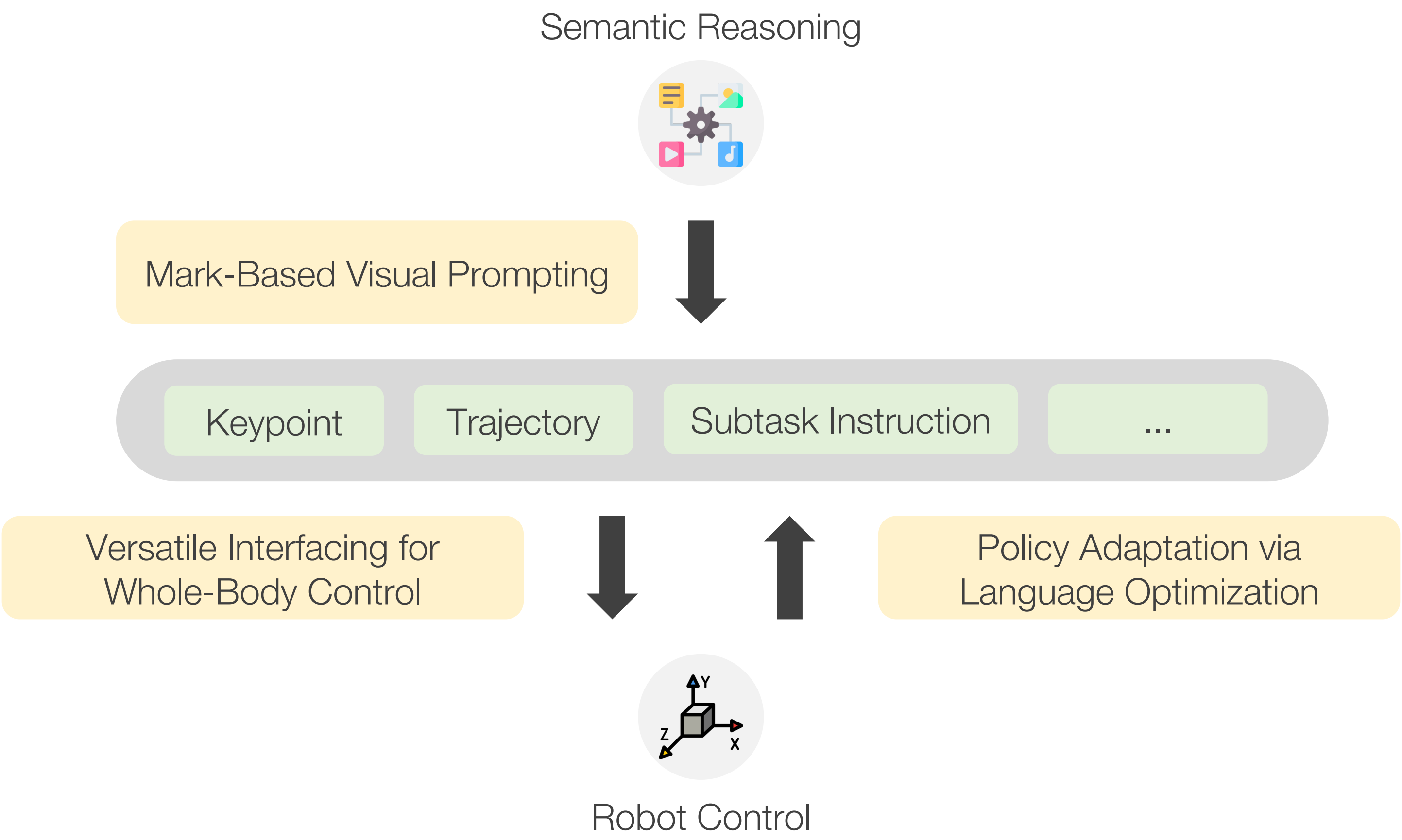
Performance of PALO with 5 demonstrations compared to finetuning the Octo model on different number of demonstrations.



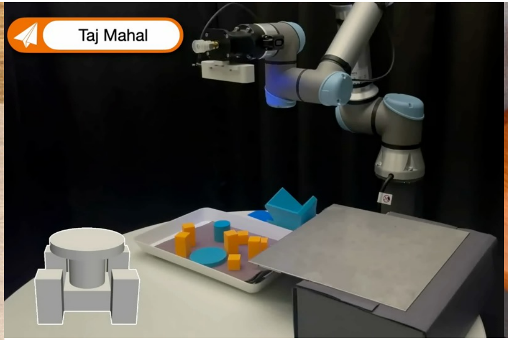
Myers*, Zheng*, Mees, Levine†, **Fang**†. Policy Adaptation via Language Optimization: Decomposing Tasks for Few-Shot Imitation. CoRL 2024



Semantic Reasoning
Mark-Based Visual Prompting
Physically Grounded Task Representation
Versatile Interfacing for Whole-Body Control
Policy Adaptation via Language Optimization
Y
Z
X
Robot Control


Semantic Reasoning
Mark-Based Visual Prompting
Keypoint
Trajectory
Subtask Instruction
...
Versatile Interfacing for Whole-Body Control
Policy Adaptation via Language Optimization
Y
Z
X
Robot Control

Question?



Kuan Fang
Department of Computer Science
Cornell University

Cornell Bowers C·IS
College of Computing
and Information Science