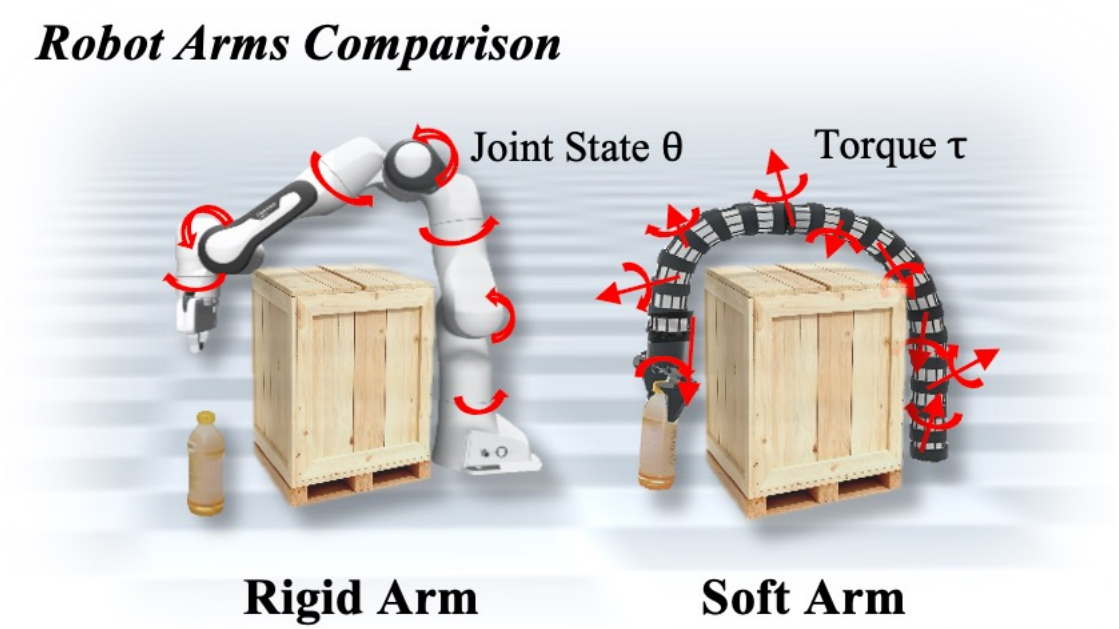
ManiSoft: Towards Vision-Language Manipulation for Soft Continuum Robotics

Ziyu Wei[12*] Luting Wang[12*] Chen Gao[13+] Li Wen[1+] Si Liu[12+]

+ Corresponding Author, 1 Beihang University,
2 Hangzhou Innovation Institute, 3 National University of Singapore.

Motivation



Advantages

- Compliance
- Adaptability
- Safe interaction
- High flexibility

Challenges

- Lack of suitable simulators.
- Low-level, high-dimensional control
- Difficulty in scaled data generation

There is a lack of *benchmark* and *expert data* to support research on soft-arm vision-language manipulation.

Benchmark Overview

ManiSoft is designed as an integrated benchmark for soft-arm manipulation.

A Simulation Framework

A Scalable Data Generation Pipeline

Asset Library

263 Objects

Scene Generation

#Clean
#Randomized

Trajectory Generation

High-Level Planning
Pre-defined Rules

Waypoints

Low-Level Control
RL Executor

Simulator

Quality Check

Failure Case

6,300 expert demonstration trajectories

Four manipulation tasks

Collection (COLL)

Stacking (STK)

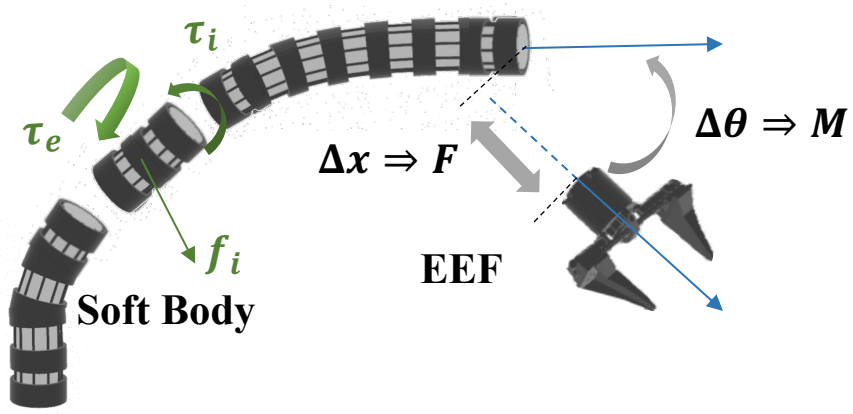
Alignment (ALN)

Arrangement (ARR)

ManiSoft Simulator

ManiSoft simulator is designed for soft arm vision-language manipulation, which needs stable interaction with the environment.

Soft Arm Modeling



- **Elastic force constraint** for soft body and end-effector coupling.

$$\mathbf{F} = -k_F \Delta \mathbf{x}, \quad \mathbf{M} = -k_M \Delta \boldsymbol{\theta}$$

- **Torque** based control method.

Dependencies

Elastica

- Deformable soft-body dynamics

MuJoCo

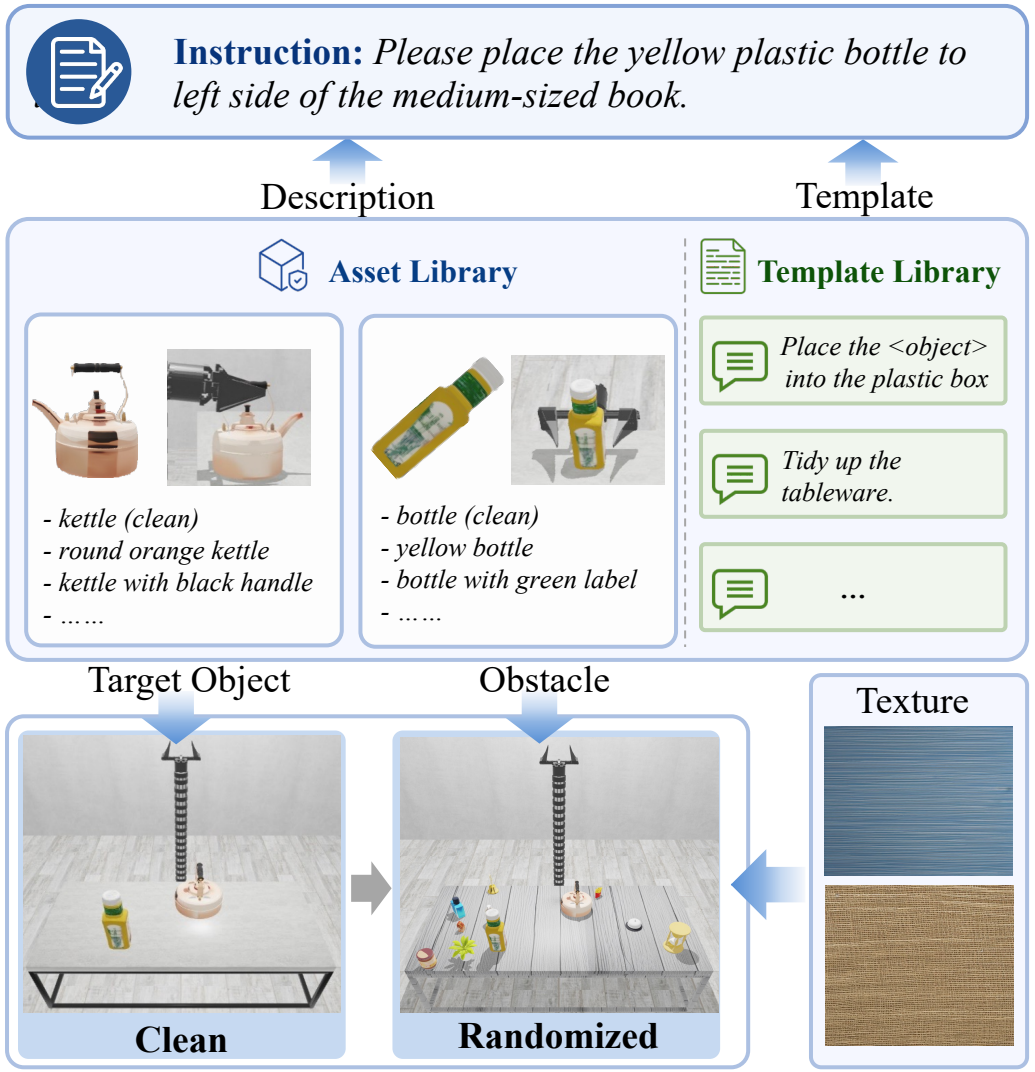
- Contact-rich environmental interaction

blender

- Rendering pipeline for visual observations

Scene Generation

ManiSoft generates tabletop scenes with two settings: ***clean*** and ***randomized***.



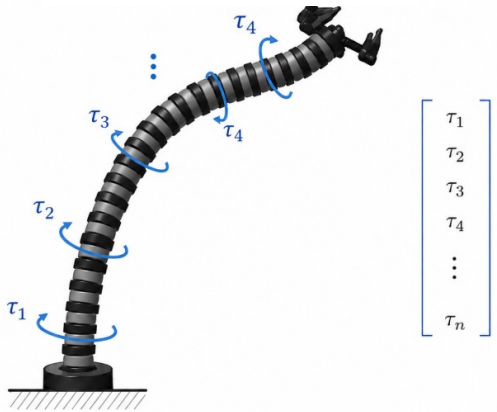
- Asset library
  - 3D objects
  - pre-annotated interaction poses
  - Descriptions
- Template library

- Randomized and collision-free object placement
- Template-based language instruction generation

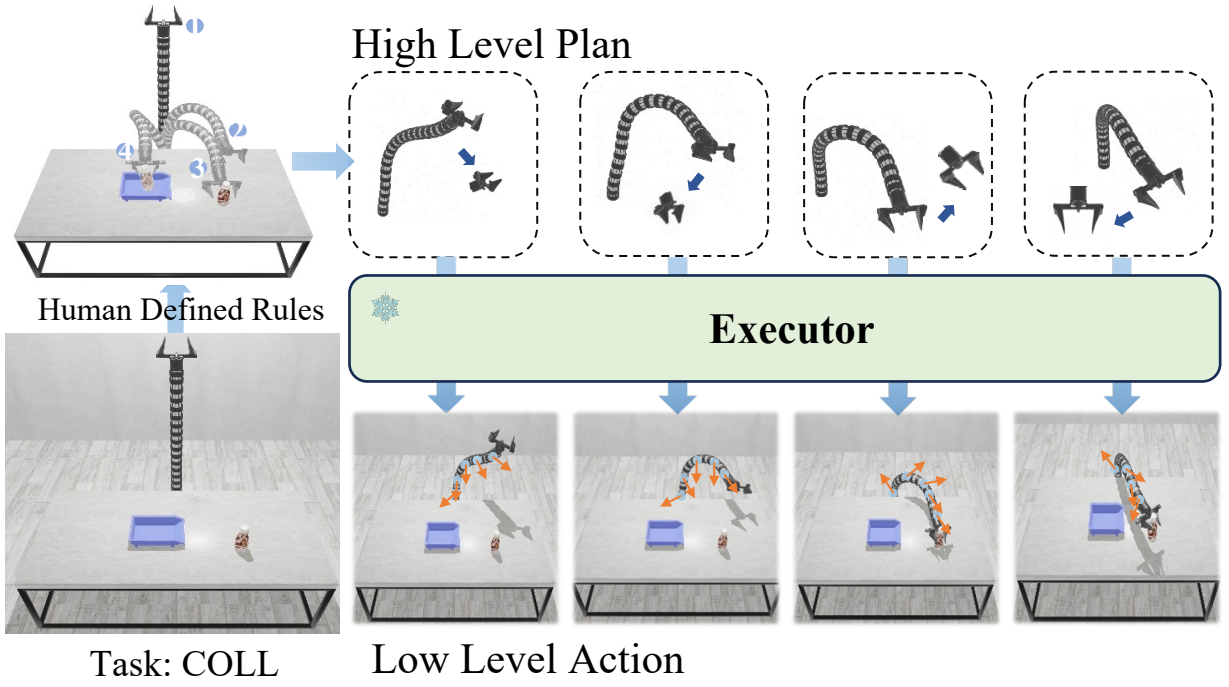
Expert Trajectory Generation

We adopt a hierarchical mechanism for accurate and stable trajectory generation.

Challenges



- High-dimensional, high-order, and non-intuitive control
- Inaccurate body-state perception
- Difficult to directly generate trajectories through pose planning and inverse kinematics



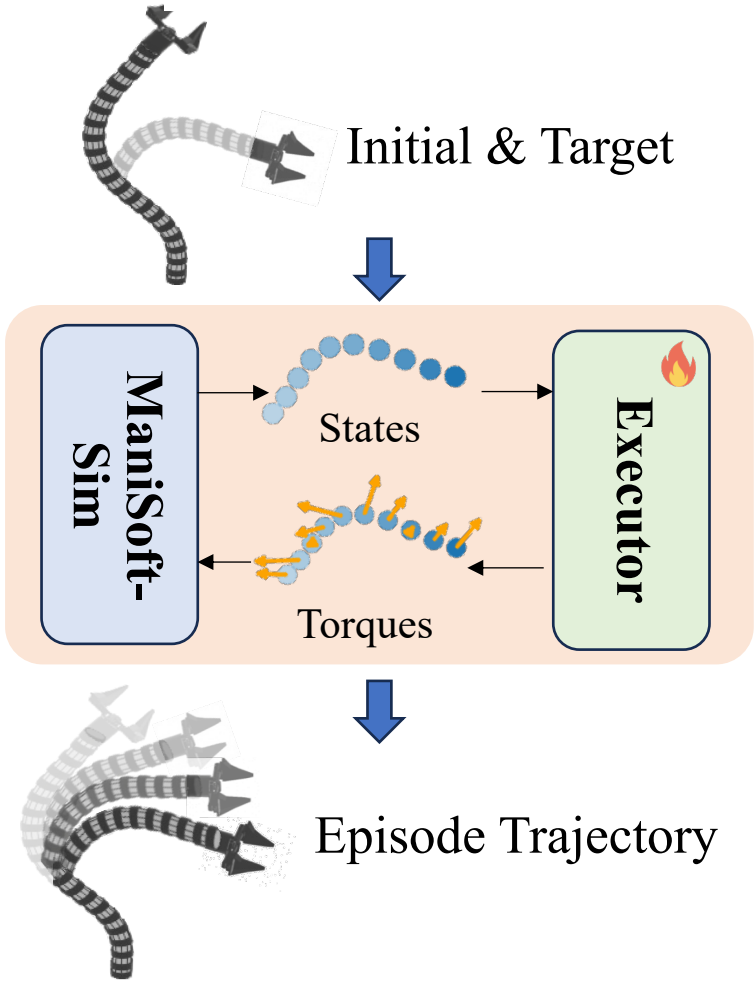
Advantages

- Avoids long-horizon torque-sequence planning
- Produces stable demonstrations across diverse scenes

Expert Trajectory Generation

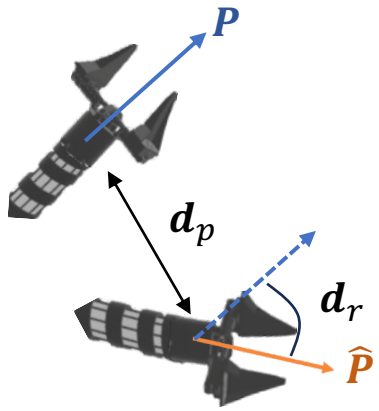
We train an executor to convert high-level waypoints into low-level torques.

RL Executor



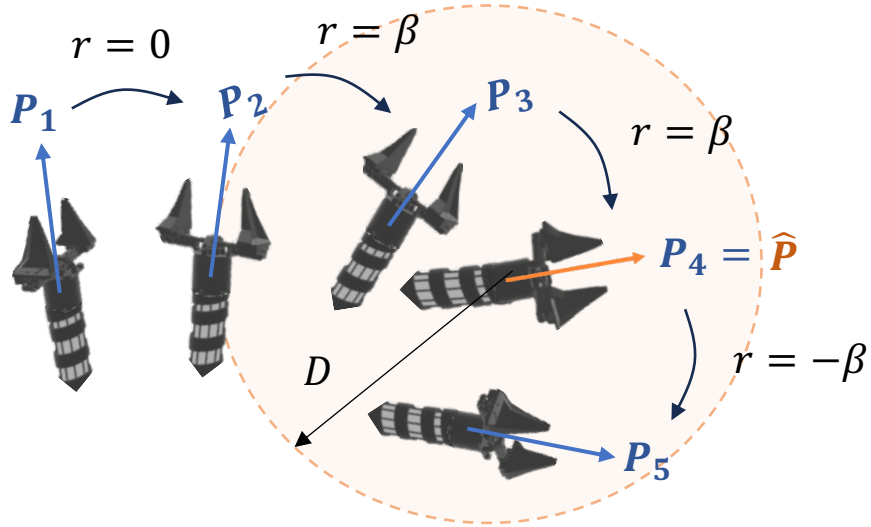
Rewards

Distance Reward



$$R_d = -d + k_1 \mathbf{1}_{\{\mathbf{d<d_1}\}} + k_2 \mathbf{1}_{\{\mathbf{d<d_2}\}}$$

Stable Reward



$$R_s = \begin{cases} -sgn\left(\frac{\partial d}{\partial t}\right) \cdot \beta, & d < D \\ 0, & d \geq D \end{cases}$$

Data Statistic

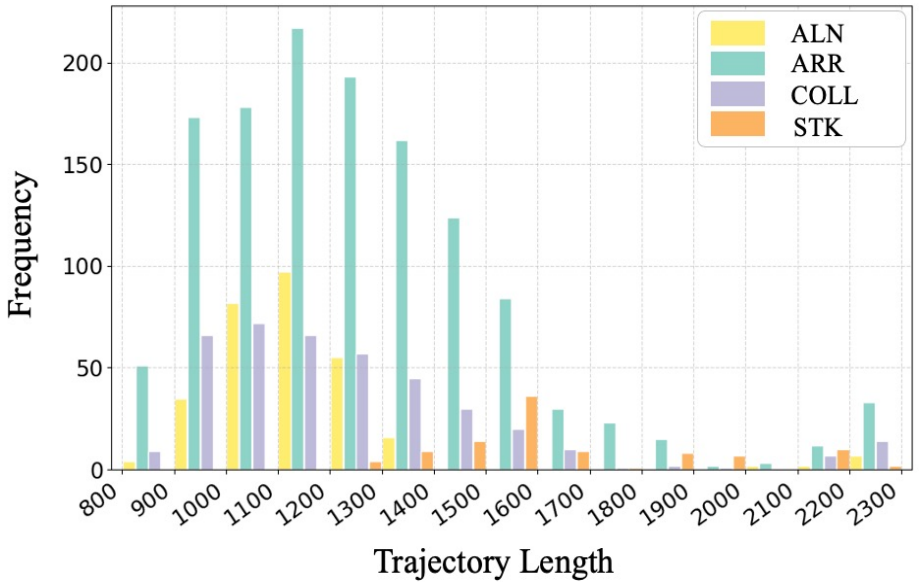
6,300 scene–trajectory pairs

40 language instructions per scene

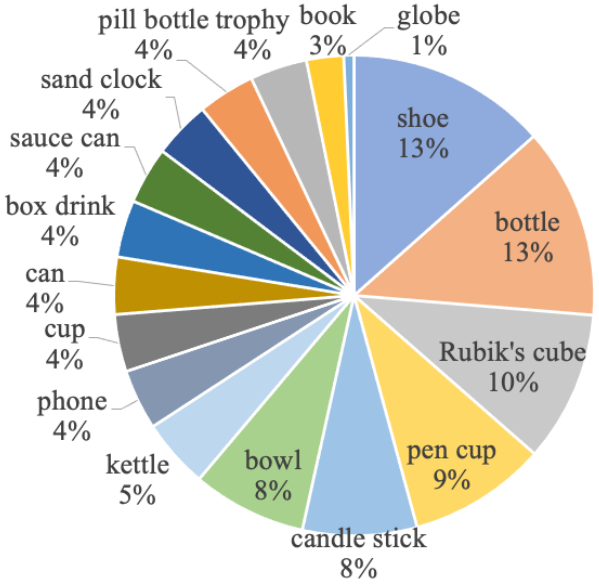
109 manipulable objects across **17** object categories

154 obstacles across **35** obstacle categories

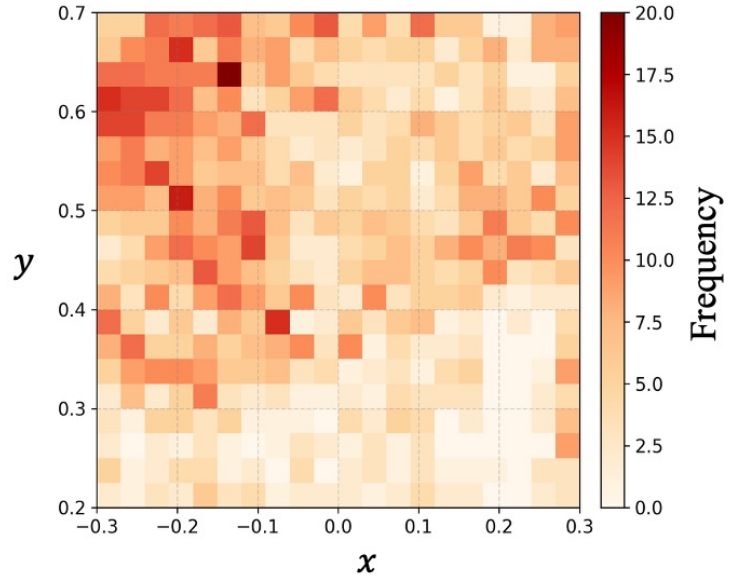
Length Distribution for Difference Tasks



Rich and Diverse Objects



Wide Spatial Coverage



Main Results

We evaluate three representative policy models: Diffusion Policy, RDT, and OpenVLA-OFT.

Method	COLL		ALN		STK		ARR		Average	
	ACC(%)	#Steps	ACC(%)	#Steps	ACC(%)	#Steps	ACC(%)	#Steps	ACC(%)	#Steps
Clean										
DP (Chi et al., 2025)	**63.0**	547	18.3	**442**	15.0	517	30.0	573	**31.6**	520
RDT (Liu et al., 2025b)	13.8	**509**	11.7	463	10.0	803	1.3	**210**	9.2	**496**
OpenVLA-OFT (Kim et al., 2025)	45.4	565	**25.0**	472	**20.0**	**492**	**31.3**	578	30.4	527
Randomized										
DP (Chi et al., 2025)	3.8	**521**	1.7	**324**	2.5	818	0.6	790	2.2	613
RDT (Liu et al., 2025b)	1.2	487	4.2	379	0.0	-	1.3	**238**	1.6	**368**
OpenVLA-OFT (Kim et al., 2025)	**32.7**	601	**26.7**	489	**35.0**	563	**13.7**	563	**27.0**	554

We further show the policy performance on each object category in the ARR task.

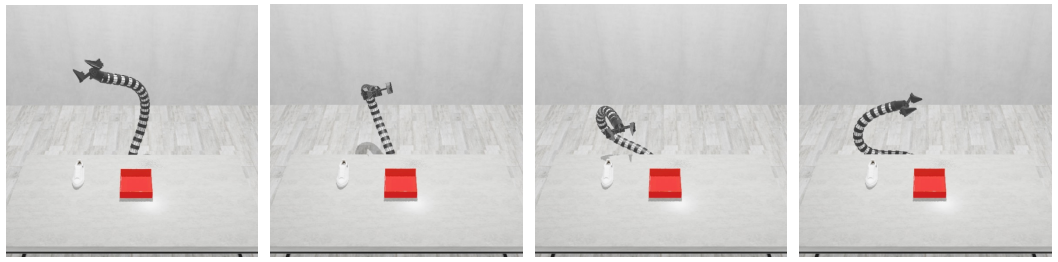
Method	Rubik's Cube		Bottle		Pen Cup		Shoe		Average	
	ACC(%)	#Steps	ACC(%)	#Steps	ACC(%)	#Steps	ACC(%)	#Steps	ACC(%)	#Steps
Clean										
DP (Chi et al., 2025)	**50.0**	705	25.0	454	25.0	602	20.0	534	30.0	573
RDT (Liu et al., 2025b)	5.0	210	0.0	-	0.0	-	0.0	-	1.3	210
OpenVLA-OFT (Kim et al., 2025)	40.0	667	**30.0**	430	**35.0**	525	**20.0**	690	**31.3**	578
Randomized										
DP (Chi et al., 2025)	0.0	-	0.0	324	-	818	2.5	790	0.6	790
RDT (Liu et al., 2025b)	0.0	-	2.5	174	2.5	302	0.0	-	1.3	238
OpenVLA-OFT (Kim et al., 2025)	**15.0**	728	**7.5**	482	**25.0**	507	**7.5**	536	**13.7**	563

Results and Analyse

ICML International Conference On Machine Learning

Common failure cases of rigid-arm policies on ManiSoft

Proprioceptive State Ambiguity



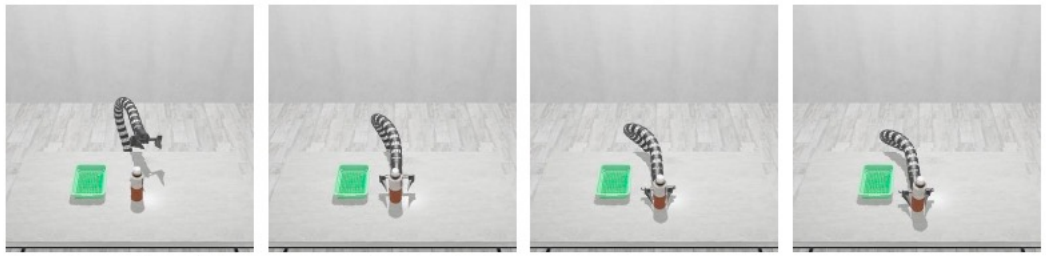
Instruction: *Set the shoe into the plastic container, ensuring it stays stable.*

Unable to Leverage Compliance

Instruction: *Put the blue pencup with rounded edges immediately to the left of the rectangular book.*

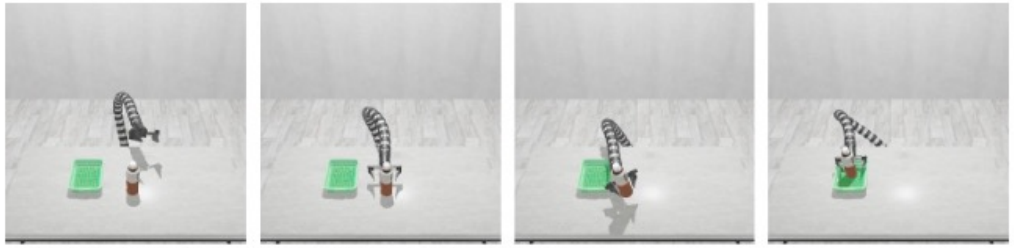
Stop-Moving Behavior

OpenVLA-OFT



Instruction: Tuck away the bottle into storage.

DP



Instruction: *Tuck away the bottle into storage.*

Thanks!

***ManiSoft*: Towards Vision-Language Manipulation for Soft Continuum Robotics**

Paper (arXiv)

Scan to view the paper on arXiv.

Our Laboratory!

Scan to learn more about our laboratory.

Project Website / Code

Scan to visit our project website and code repository.

ColaLab