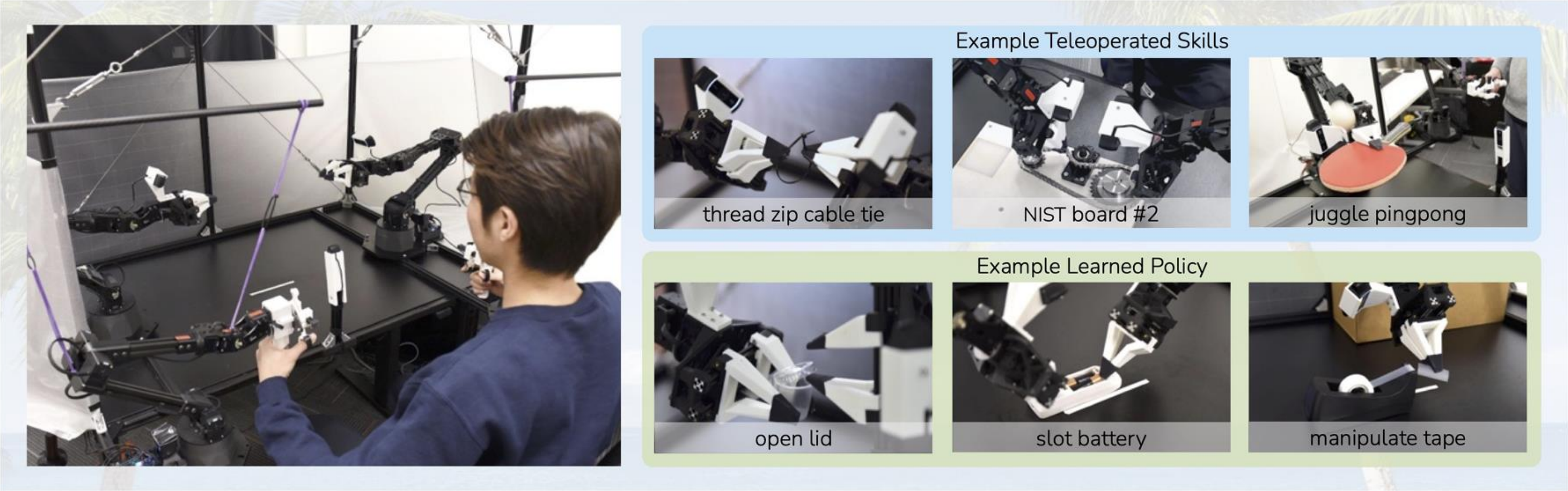


ALOHA

Presenter: Lu jinxuan, Hu Xinshu
5.14

仿生设计与学习实验室 Bionic Design & Learning Lab

main DL In Collaboration with MIT

AncoraSIR.com

SUSTech Southern University of Science and Technology

Motivation and Main Problem

ALOHA-significance

ALOHA：A Low-cost open-source hardware system for bimanual teleoperation

- ALOHA：A Low-cost open-source hardware system for bimanual teleoperation
- Action Chunking with Transformers (ACT): learns a generative model over action sequences

AncoraSIR.com

Motivation and Main Problem

ALOHA-significance

Learning Bimanual Manipulation with Low-Cost Whole-Body Teleoperation

AncoraSIR.com

Problem Setting

- **Low-cost:keep costs under $20,000**

- **Versatile: It can be applied to a wide range of fine manipulation tasks with real-world objects**

- **User-friendly: The system should be intuitive, reliable, and easy to use.**

- **Easy-to-build & Repairable**

AncoraSIR.com

Motivation and Main Problem

Learning Bimanual Manipulation with Low-Cost Whole-Body Teleoperation

- With the imitation learning algorithm Action Chunking with Transformers (ACT), learning precise closed-loop behaviors, achieving a success rate of 80% to 90% in fine motor skills only requires 10 minutes or 50 demonstration trajectories.

Context / Related Work / Limitations of Prior Work

Related Work

- Naoki Yokoyama, Alexander William Clegg, Eric Undersander, Sehoon Ha, Dhruv Batra, and Akshara Rai. Adaptive skill coordination for robotic mobile manipulation. arXiv preprint arXiv:2304.00410, 2023. 3
- Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In Proceedings of Robotics: Science and Systems (RSS), 2023. 1, 2, 3, 5, 8, 9

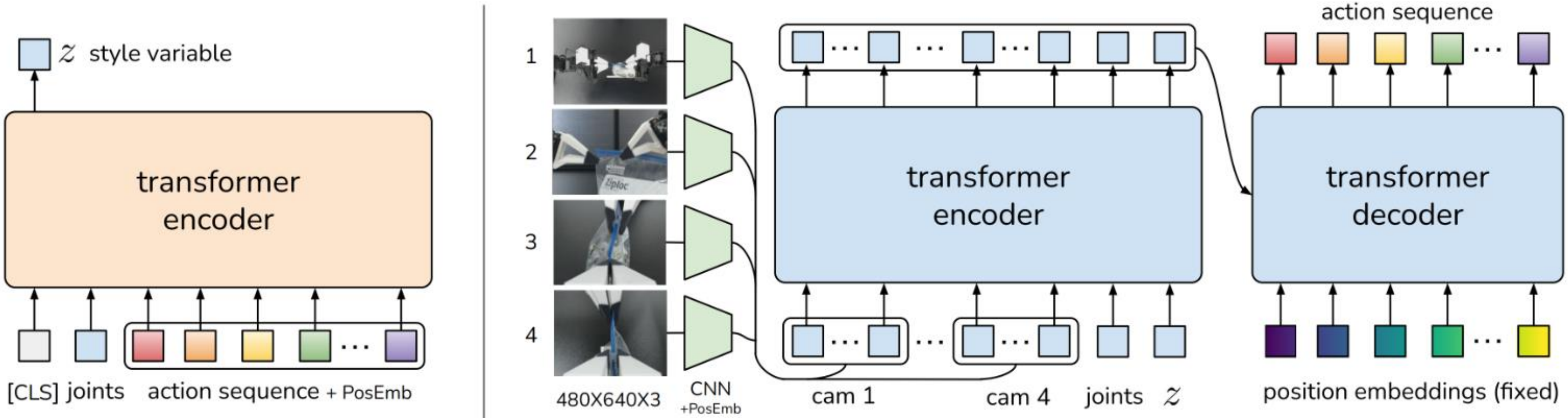
limitations

- **High precision often involves costly industrial robots and correspondingly demands higher levels of expertise and knowledge from users.**
- **everyday environments require whole-body coordination of both mobility and dexterous manipulation, rather than just individual mobility or manipulation behaviors.**

just individual mobility or

Theory

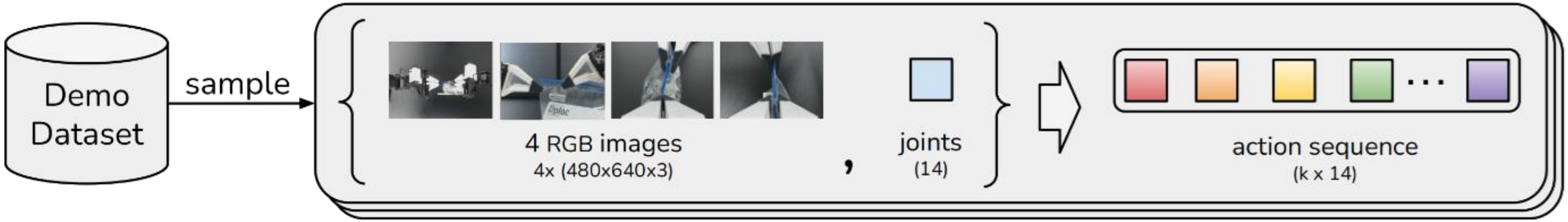
- start with ACT(Action Chunking with Transformers), the method introduced with ALOHA



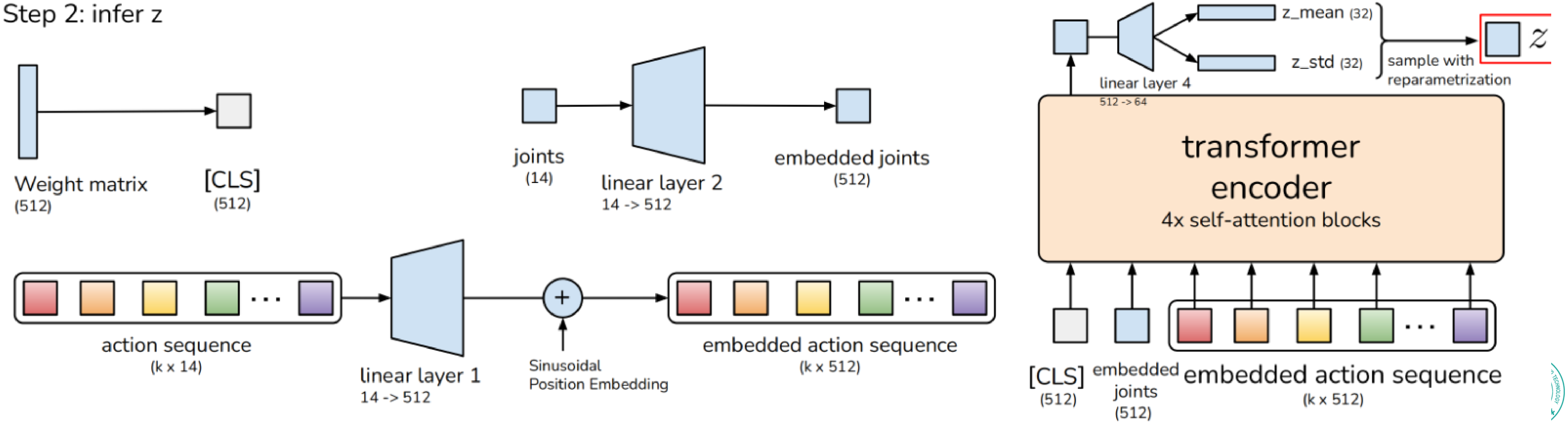
AncoraSIR.com

Theory

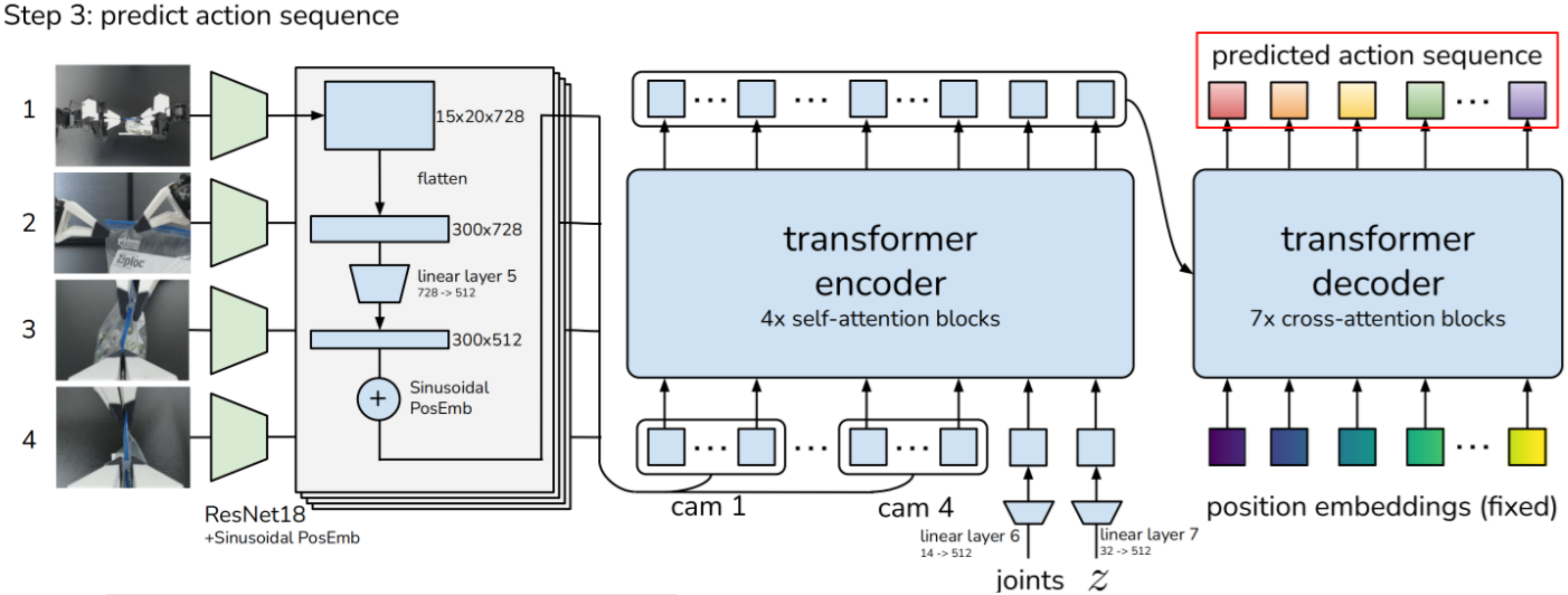
Step 1: sample data

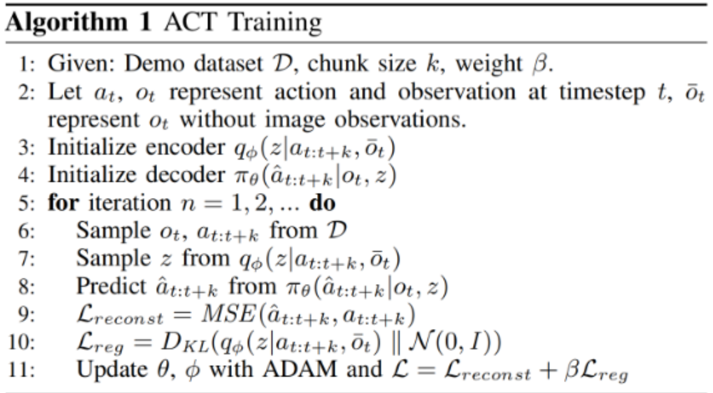


Step 2: infer z



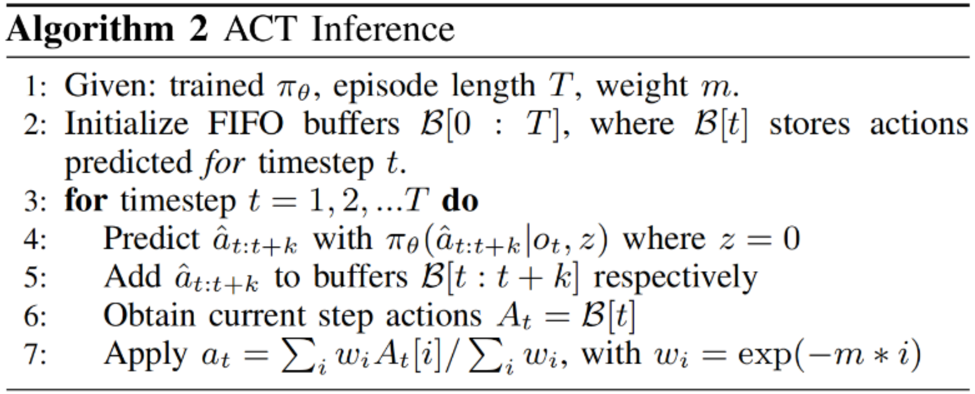
Theory





Algorithm 1 ACT Training

```
Given: Demo dataset $\mathcal{D}$, chunk size $k$, weight $\beta$.
Let $a_t$, $o_t$ represent action and observation at timestep $t$, $\bar{o}_t$ represent $o_t$ without image observations.
Initialize encoder $q_\phi(z|a_{t:t+k}, \bar{o}_t)$
Initialize decoder $\pi_\theta(\hat{a}_{t:t+k}|o_t, z)$
for iteration $n = 1, 2, \ldots$ do
    Sample $o_t$, $a_{t:t+k}$ from $\mathcal{D}$
    Sample $z$ from $q_\phi(z|a_{t:t+k}, \bar{o}_t)$
    Predict $\hat{a}_{t:t+k}$ from $\pi_\theta(\hat{a}_{t:t+k}|o_t, z)$
    $\mathcal{L}_{reconst} = MSE(\hat{a}_{t:t+k}, a_{t:t+k})$
    $\mathcal{L}_{reg} = D_{KL}(q_\phi(z|a_{t:t+k}, \bar{o}_t) \parallel \mathcal{N}(0, I))$
    Update $\theta$, $\phi$ with ADAM and $\mathcal{L} = \mathcal{L}_{reconst} + \beta\mathcal{L}_{reg}$
```

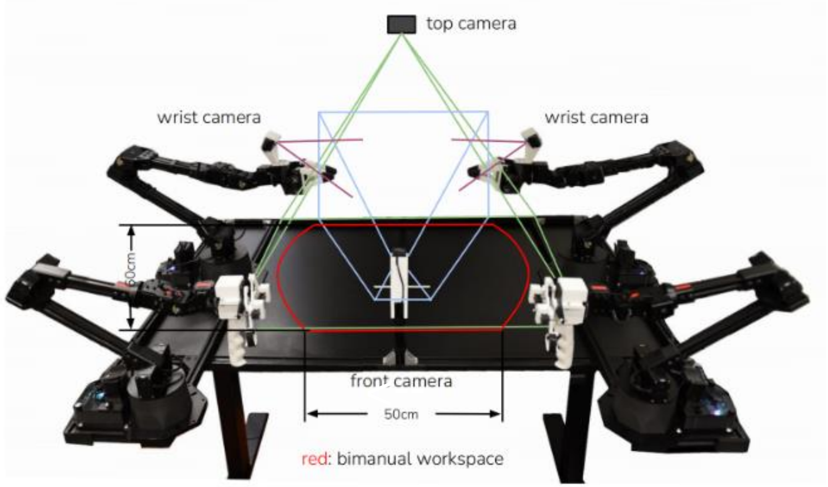


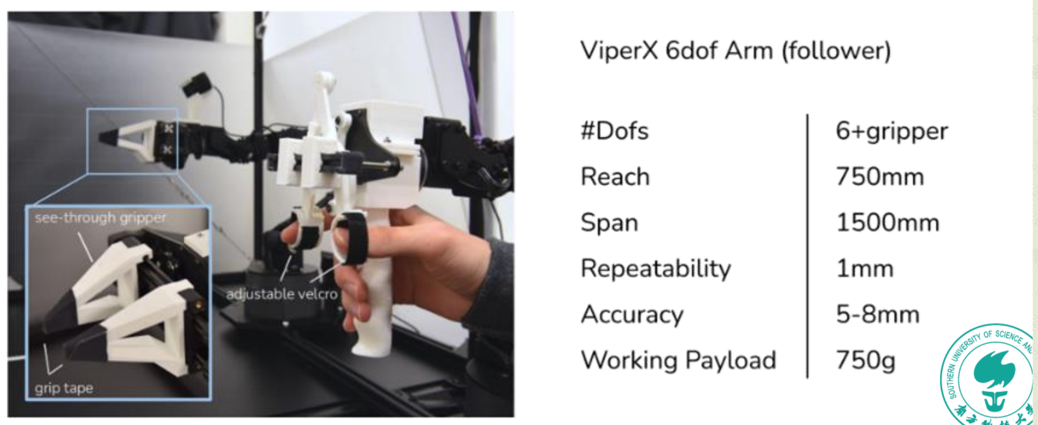
Algorithm 2 ACT Inference

```
Given: trained $\pi_\theta$, episode length $T$, weight $m$.
Initialize FIFO buffers $\mathcal{B}[0 : T]$, where $\mathcal{B}[t]$ stores actions predicted for timestep $t$.
for timestep $t = 1, 2, \ldots T$ do
    Predict $\hat{a}_{t:t+k}$ with $\pi_\theta(\hat{a}_{t:t+k}|o_t, z)$ where $z = 0$
    Add $\hat{a}_{t:t+k}$ to buffers $\mathcal{B}[t : t + k]$ respectively
    Obtain current step actions $A_t = \mathcal{B}[t]$
    Apply $a_t = \sum_i w_i A_t[i] / \sum_i w_i$, with $w_i = \exp(-m * i)$
```

Proposed Approach / Algorithm / Method

- start with ACT, the method introduced with ALOHA
- **Innovation：**
- Developing a robotic arm capable of precision operations while maintaining costs below $20,000
- Achieving high precision, coordination, and closed-loop visual feedback.
- Minimizing the accumulation of errors and mistakes over time during training





ViperX 6dof Arm (follower)

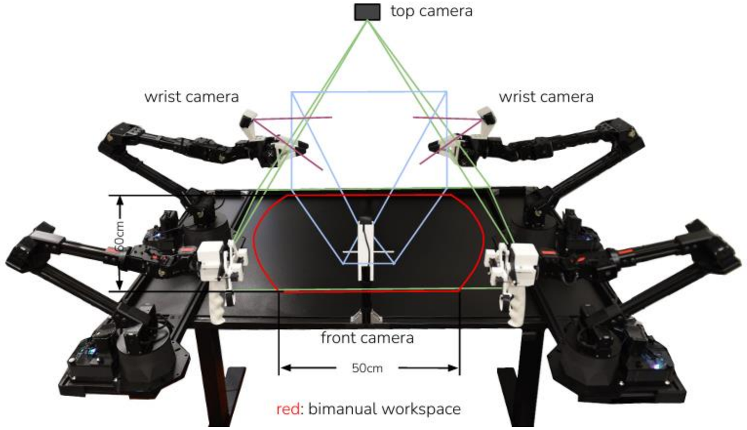
#Dofs	6+gripper
Reach	750mm
Span	1500mm
Repeatability	1mm
Accuracy	5-8mm
Working Payload	750g

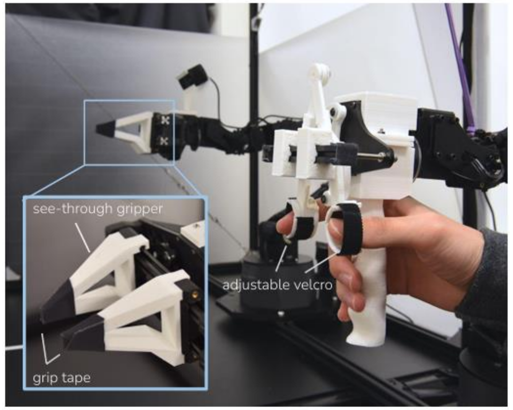
Experimental Setup

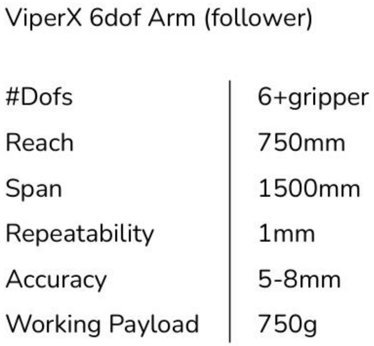
Hardware

ALOHA：A Low-cost open-source hardware system for bimanual teleoperation

- Two ViperX 6-DoF robot arms
- With 3D printed “see-through” fingers
- A teleoperation system: WidowX







ViperX 6dof Arm (follower)

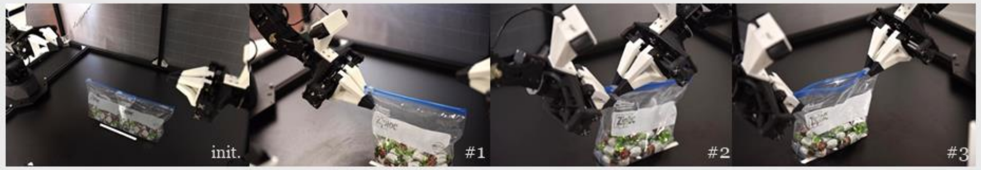
#Dofs	6+gripper
Reach	750mm
Span	1500mm
Repeatability	1mm
Accuracy	5-8mm
Working Payload	750g

AncoraSIR.com

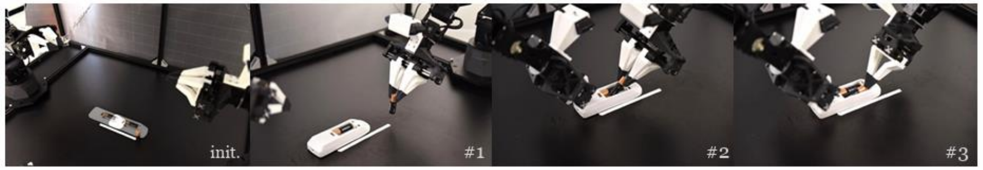
Experimental Setup

Tesk and Data Collection

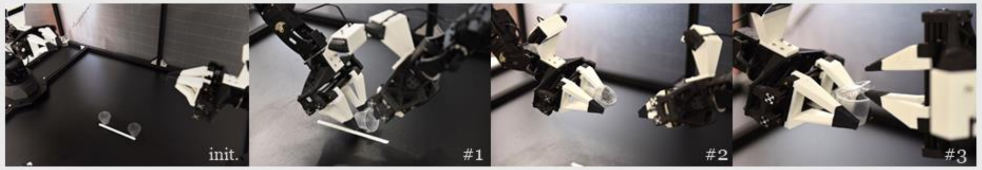
6 real-world tasks:



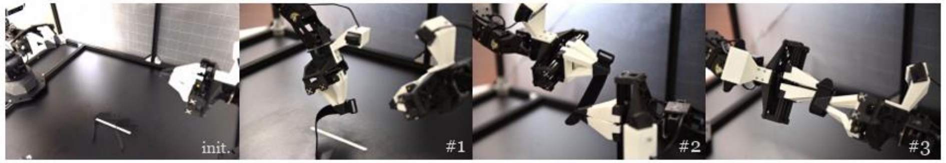
Slide Ziploc: Open the ziploc bag that is standing upright on the table. The bag is randomized along the 15cm white line. It is dropped from ~5cm above the table to randomize the deformation, which affects the height and appearance of the bag. The left arm first grasps the bag body *(Subtask#1 Grasp)* followed by the right arm pinching the slider *(Subtask #2 Pinch)*. Then the right arm moves right to unzip the bag *(Subtask #3 Open)*.



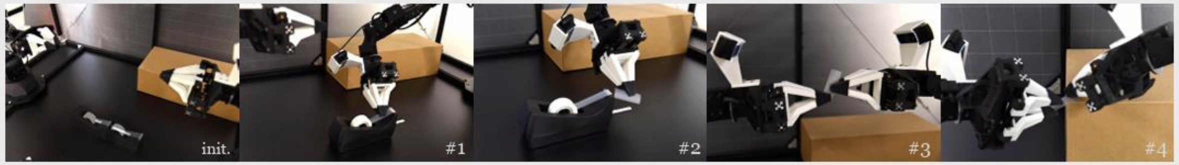
Slot Battery: Insert the battery into the remote controller. The controller is randomized along the 15cm white line. The battery is initialized in roughly the same position with different rotations. The right arm first grasps the battery *(Subtask#1 Grasp)* then places it into the slot *(Subtask#2 Place)*. The left arm presses onto the remote to prevent it from sliding, while the right arm pushes in the battery *(Subtask#3 Insert)*.



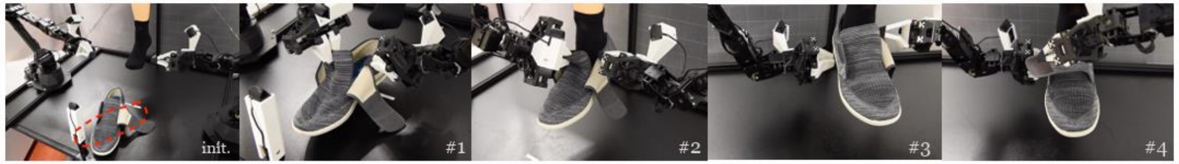
Open Cup: Pick up and open the lid of a translucent condiment cup. The cup is randomized along the 15cm white line. Both arms approach the cup, and the right gripper gently tips over the cup *(Subtask#1 Tip Over)* and pushes it into the gripper of the left arm. The left arm then gently closes its gripper and lifts the cup off the table *(Subtask#2 Grasp)*. Next, the right gripper approaches the cup lid from below and prys open the lid.



Thread Velcro: Pick up the velcro cable tie and insert one end into the small loop on the other end. The velcro tie is randomized along the 15cm white line. The left arm first picks up the velcro tie by pinching near the plastic loop *(Subtask#1 Lift)*. The right arm grasps the tail of the velcro tie mid-air *(Subtask#2 Grasp)*. Next, both arms coordinate to deform the velcro tie and insert one end of it into the plastic loop on the other end.



Prep Tape: Hang a short segment of tape on the edge of the box. The tape dispenser is randomized along the 15cm white line. First, the right gripper grasps the tape from the side *(Subtask#1 Grasp)*. It then lifts the tape and pulls to unroll it, followed by cutting it with the dispenser blade *(Subtask#2 Cut)*. Next, the right gripper hands the tape segment to the left gripper in mid-air *(Subtask#3 Handover)*, and both arms move toward the corner of the stationery cardboard box. The left arm then lays the tape segment flat on the surface of the box while the right gripper pushes down on the tape to prevent slipping. The left arm then opens its gripper to release the tape *(Subtask#4 Hang)*.

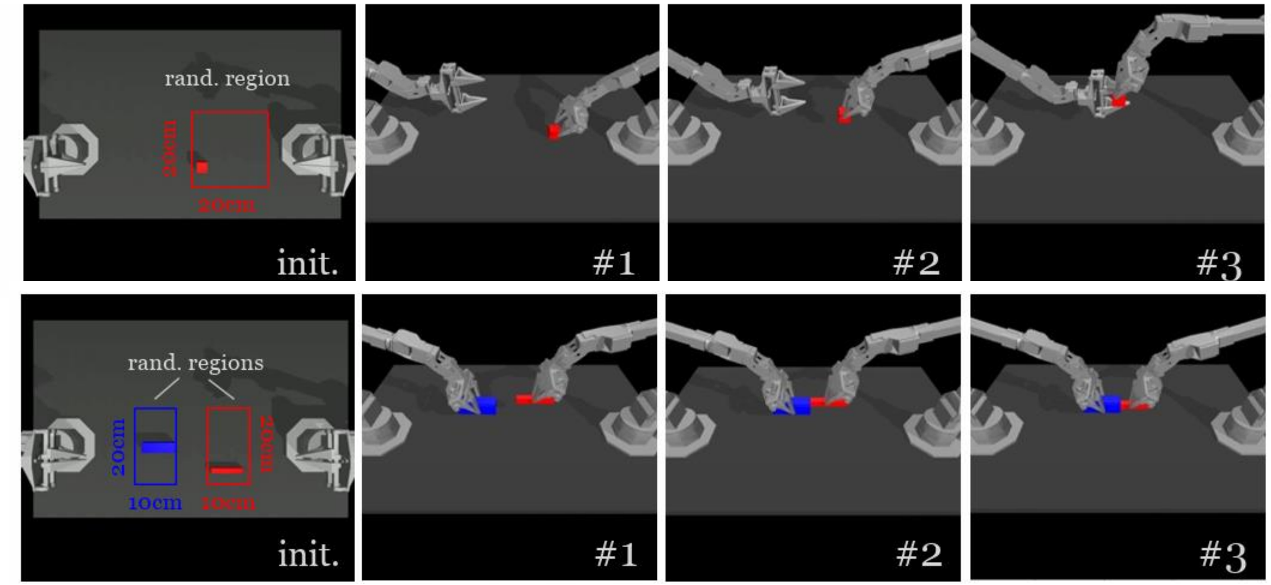


Put On Shoe: Put a velcro-strap shoe on a fixed manniquin foot. The shoe pose is randomized along the 15cm white line. First, both left and right grippers pick up the shoe *(Subtask#1 Lift)*. Then both arms coordinate to put it on, with the heel touching the heel counter *(Subtask#2 Insert)*. Next, the left arm moves to support the shoe *(Subtask#3 Support)*, followed by the right arm securing the velcro strap *(Subtask#4 Secure)*.

Experimental Setup

Tesk and Data Collection

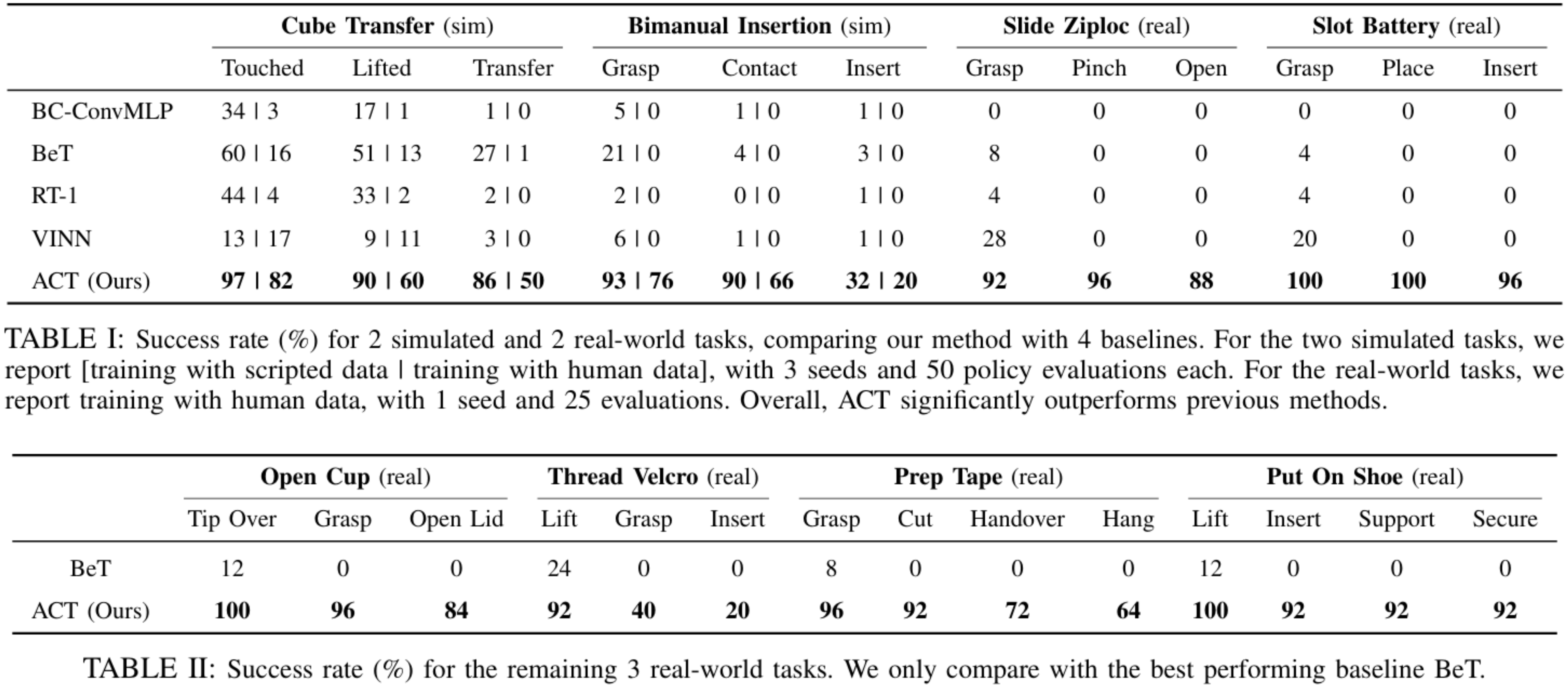
2 simulated fine manipulation tasks in MuJoCo:



Left: Cube Transfer. Transfer the red cube to the other arm. The right arm touches *(#1)* and grasps *(#2)* the red cube, then hands it to the left arm.
Right: Bimanual Insertion. Insert the red peg into the blue socket. Both arms grasp *(#1)*, let socket and peg make contact *(#2)* and insertion.

Experimental Results

Compare ACT with four prior mitation learning methods



	Cube Transfer (sim)			**Bimanual Insertion** (sim)			**Slide Ziploc** (real)			**Slot Battery** (real)		
	Touched	Lifted	Transfer	Grasp	Contact	Insert	Grasp	Pinch	Open	Grasp	Place	Insert
BC-ConvMLP	34 | 3	17 | 1	1 | 0	5 | 0	1 | 0	1 | 0	0	0	0	0	0	0
BeT	60 | 16	51 | 13	27 | 1	21 | 0	4 | 0	3 | 0	8	0	0	4	0	0
RT-1	44 | 4	33 | 2	2 | 0	2 | 0	0 | 0	1 | 0	4	0	0	4	0	0
VINN	13 | 17	9 | 11	3 | 0	6 | 0	1 | 0	1 | 0	28	0	0	20	0	0
ACT (Ours)	**97 | 82**	**90 | 60**	**86 | 50**	**93 | 76**	**90 | 66**	**32 | 20**	**92**	**96**	**88**	**100**	**100**	**96**

TABLE I: Success rate (%) for 2 simulated and 2 real-world tasks, comparing our method with 4 baselines. For the two simulated tasks, we report [training with scripted data | training with human data], with 3 seeds and 50 policy evaluations each. For the real-world tasks, we report training with human data, with 1 seed and 25 evaluations. Overall, ACT significantly outperforms previous methods.

	Open Cup (real)			**Thread Velcro** (real)			**Prep Tape** (real)				**Put On Shoe** (real)			
	Tip Over	Grasp	Open Lid	Lift	Grasp	Insert	Grasp	Cut	Handover	Hang	Lift	Insert	Support	Secure
BeT	12	0	0	24	0	0	8	0	0	0	12	0	0	0
ACT (Ours)	**100**	**96**	**84**	**92**	**40**	**20**	**96**	**92**	**72**	**64**	**100**	**92**	**92**	**92**

TABLE II: Success rate (%) for the remaining 3 real-world tasks. We only compare with the best performing baseline BeT.

Discussion of Results

For simulated tasks:

Average performance across 3random seeds with 50 trials each

- ACT outperforms the best previous method in success rate by 59%, 49%, 29% and 20%

For real-word tasks:

Run one seed and evaluate with 25 trials.

- ACT achieves 88% and 96% final success rates respectively
- Other methods making no progress
- compounding errors and non-Markovian behavior, ACT mitigates both issues with action chunking.

Critique / Limitations / Open Issues

Hardware Limitation

1. struggle with tasks that require multiple fingers
2. struggles with tasks that require high amount of forces
3. Tasks that requires finger nails are also difficult

Policy Limitation

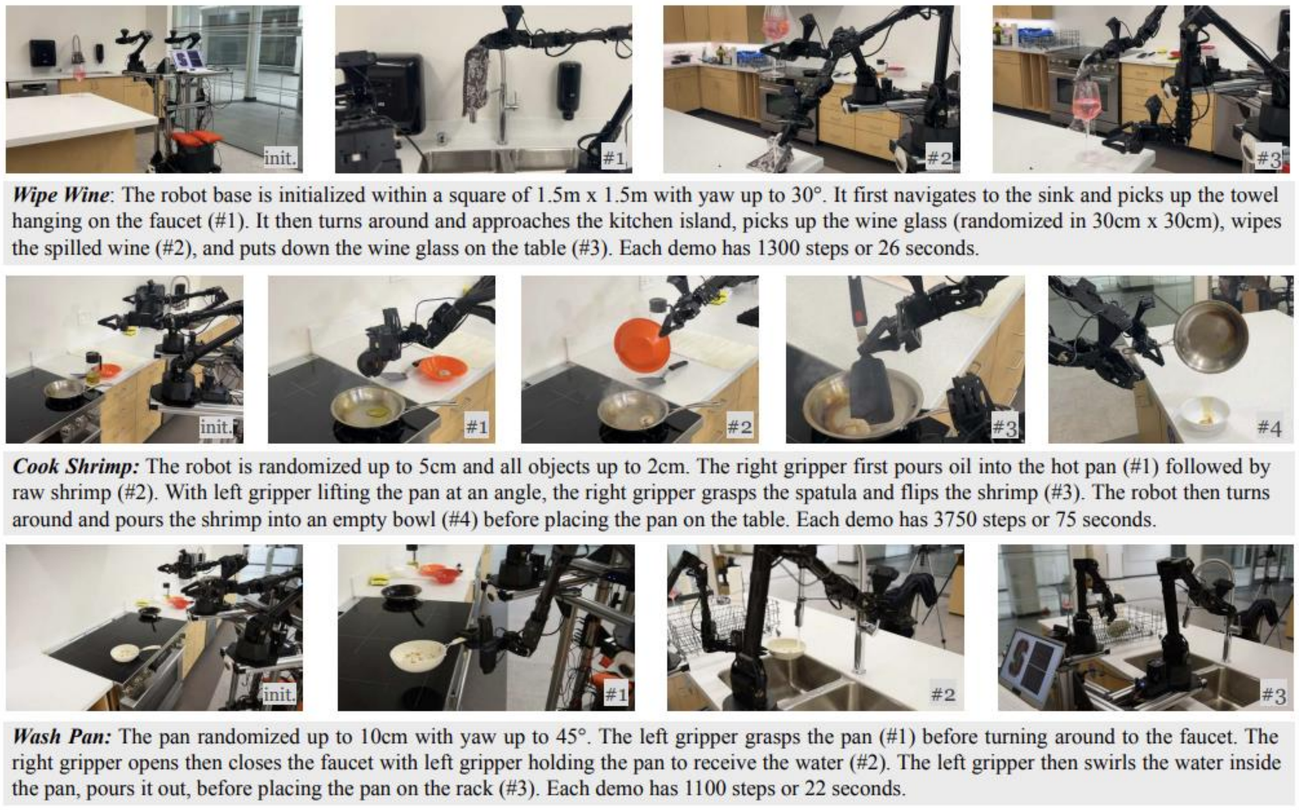
ACT struggles with tasks like opening a candy and opening a small ziploc bag laying flat on the table, due to the difficulty of acurate way of perception and the insuffcent of training data.

AncoraSIR.com

Future Work for Paper / Reading

Further explaination of the title with supporting evidence

Mobile Aloha



Wipe Wine: The robot base is initialized within a square of 1.5m x 1.5m with yaw up to 30°. It first navigates to the sink and picks up the towel hanging on the faucet (#1). It then turns around and approaches the kitchen island, picks up the wine glass (randomized in 30cm x 30cm), wipes the spilled wine (#2), and puts down the wine glass on the table (#3). Each demo has 1300 steps or 26 seconds.

Cook Shrimp: The robot is randomized up to 5cm and all objects up to 2cm. The right gripper first pours oil into the hot pan (#1) followed by raw shrimp (#2). With left gripper lifting the pan at an angle, the right gripper grasps the spatula and flips the shrimp (#3). The robot then turns around and pours the shrimp into an empty bowl (#4) before placing the pan on the table. Each demo has 3750 steps or 75 seconds.

Wash Pan: The pan randomized up to 10cm with yaw up to 45°. The left gripper grasps the pan (#1) before turning around to the faucet. The right gripper opens then closes the faucet with left gripper holding the pan to receive the water (#2). The left gripper then swirls the water inside the pan, pours it out, before placing the pan on the rack (#3). Each demo has 1100 steps or 22 seconds.

AncoraSIR.com

Extended Readings

Further explaination of the title with supporting evidence

1. 2023CoRL – “Waypoint-Based limitation Learing for Robotic Manipulation” https://arxiv.org/abs/2302.12766
2. 2023CoRL – “Language-Driven Reprentation Learing for Robotics”
3. Yell at your Robot – https://arxiv.org/pdf/2403.12910

Summary

- **A low-cost** system for fine manipulation, comprising a teleoperation system **ALOHA** and a novel imitation learing algorithm **ACT.**
- To learn fine manipulation skills directly in the real-world, such as opening a translucent condiment cup and slotting a battery with a 80-90% success rate and around 10 min of demonstrations.
- ACT outperforms previous method, BC-ConvMLP, BeT etc.

Q&A

Team5