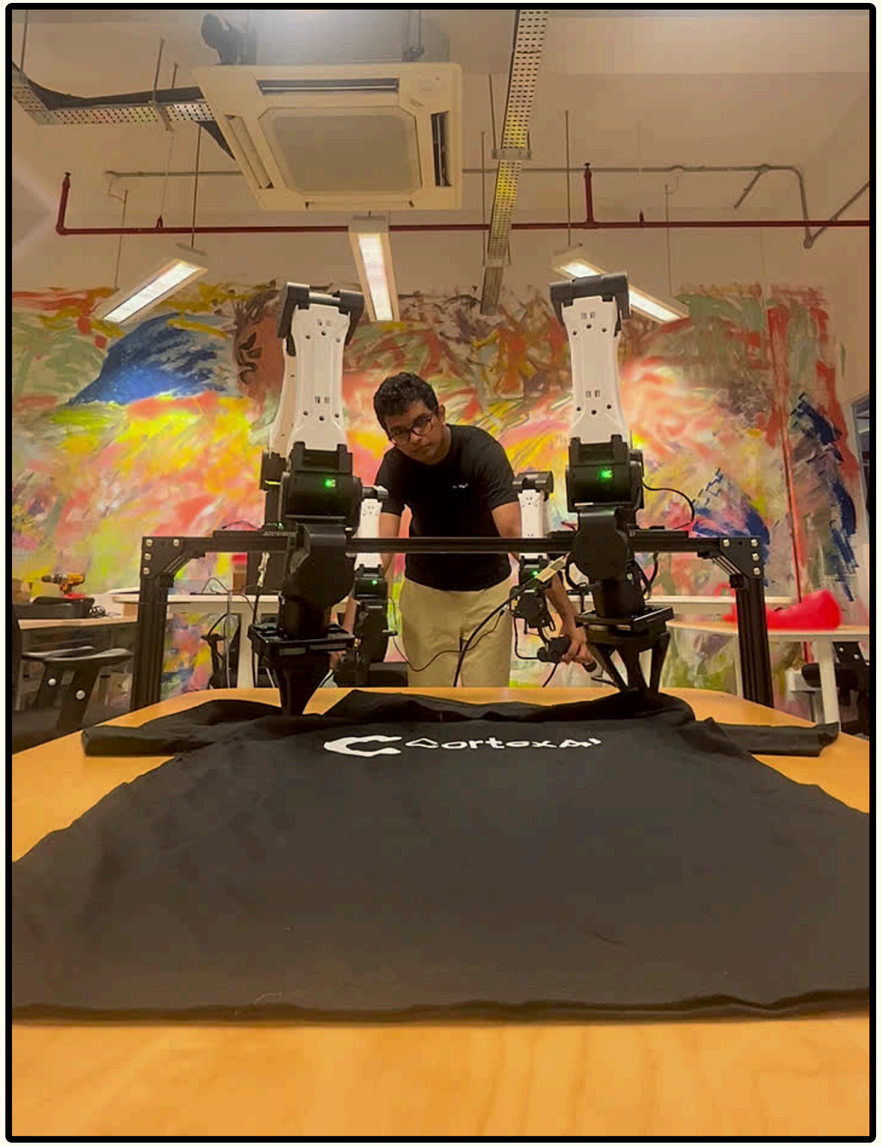
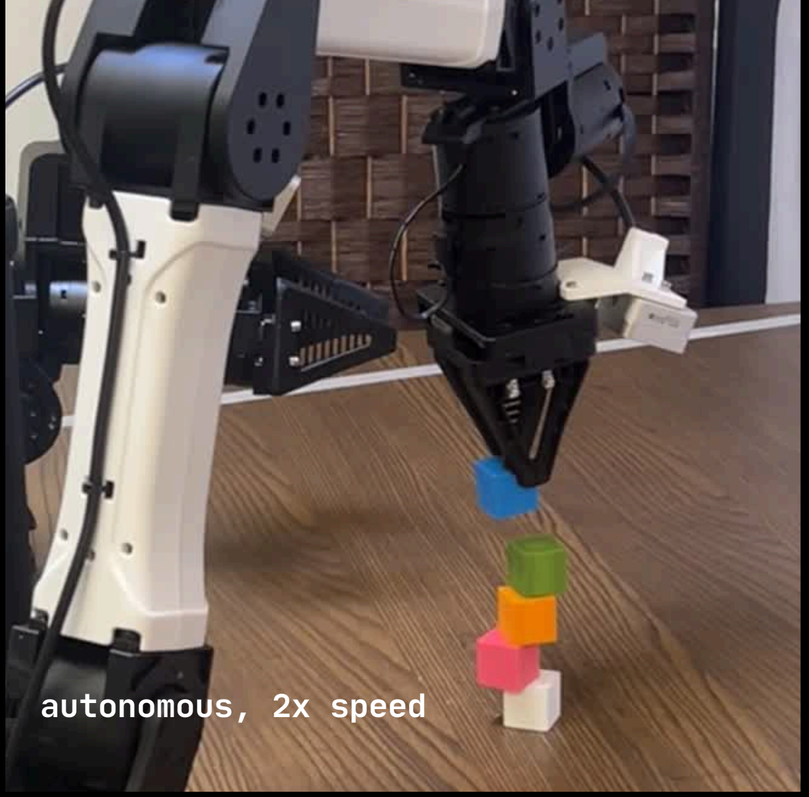
Full-Stack Robotics:

Lessons from Building Across the Robotics Stack



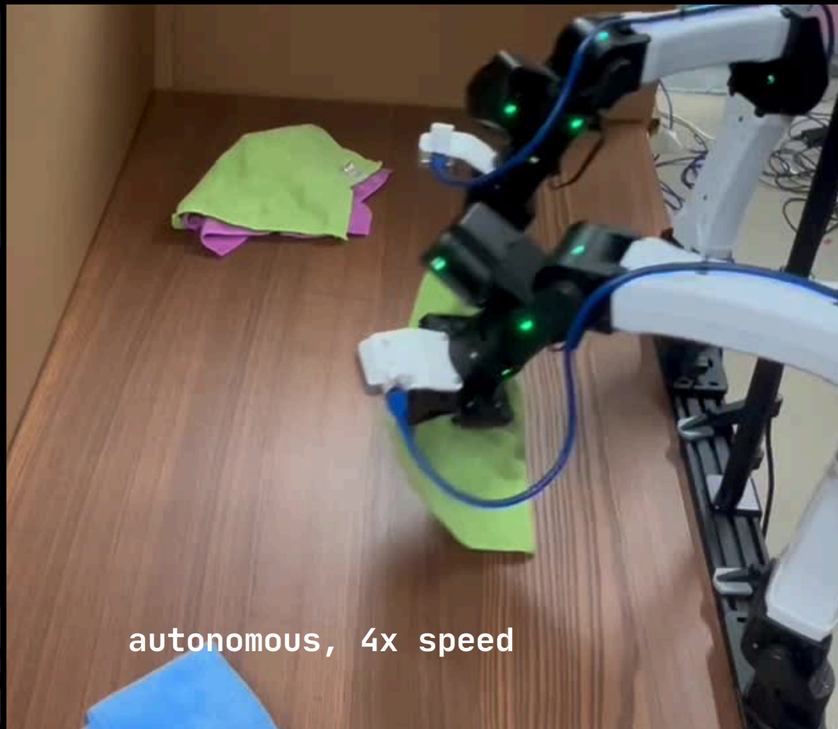
Suveen Ellawela
Founding Engineer @ Cortex AI


autonomous, 2x speed


autonomous, 4x speed


teleop, 2x speed


autonomous, 2x speed


autonomous, 4x speed

How did these robots get so smart?

Smart enough to even pour the last drop of milk into the cup?

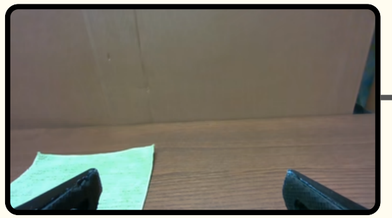
From pixels to actions

PIXELS

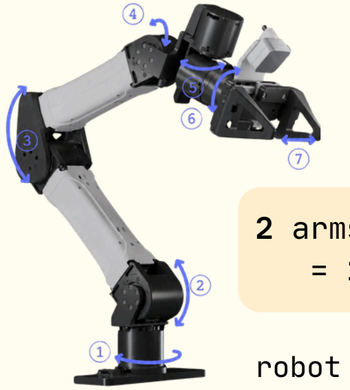
OBSERVATIONS

top camera

left wrist



right wrist



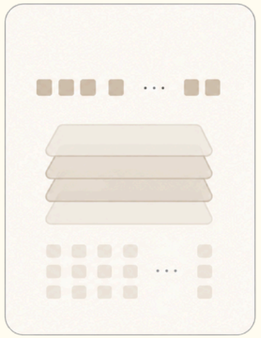
2 arms x **7** joints
= **14** joints

robot state

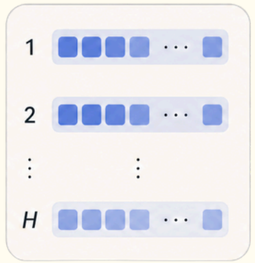
INSTRUCTION

“Fold the towel”

MODEL



ACTION CHUNK



$H \times d_{action}$

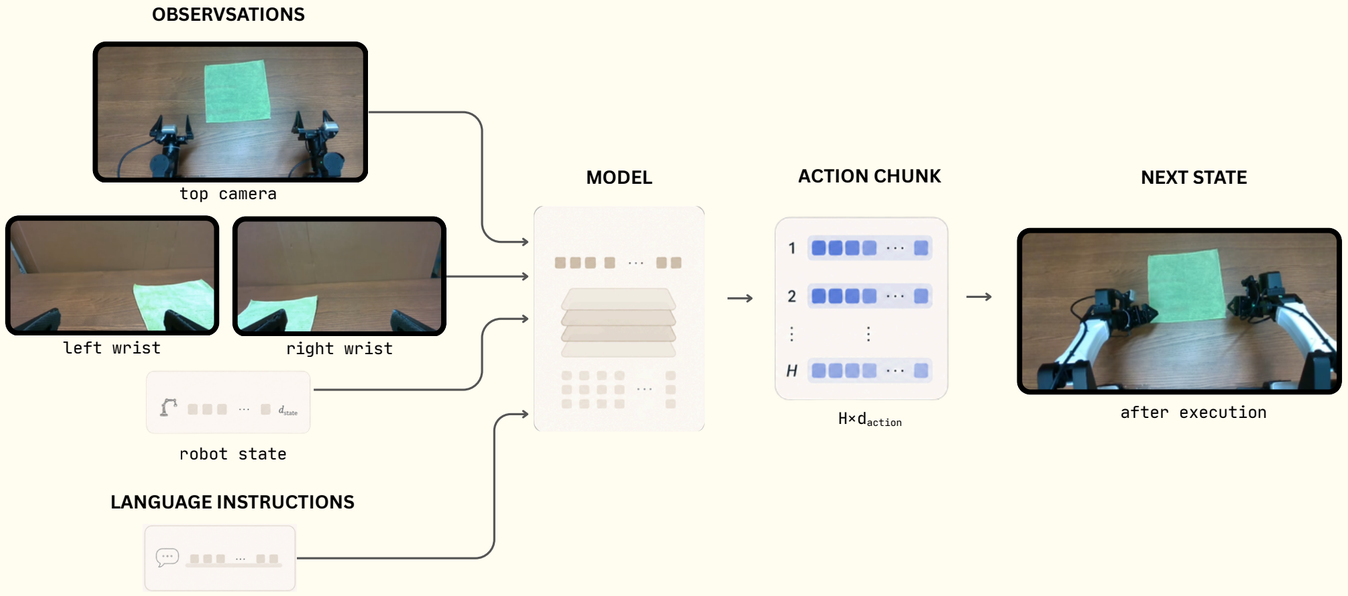
NEXT STATE

after execution

ACTIONS

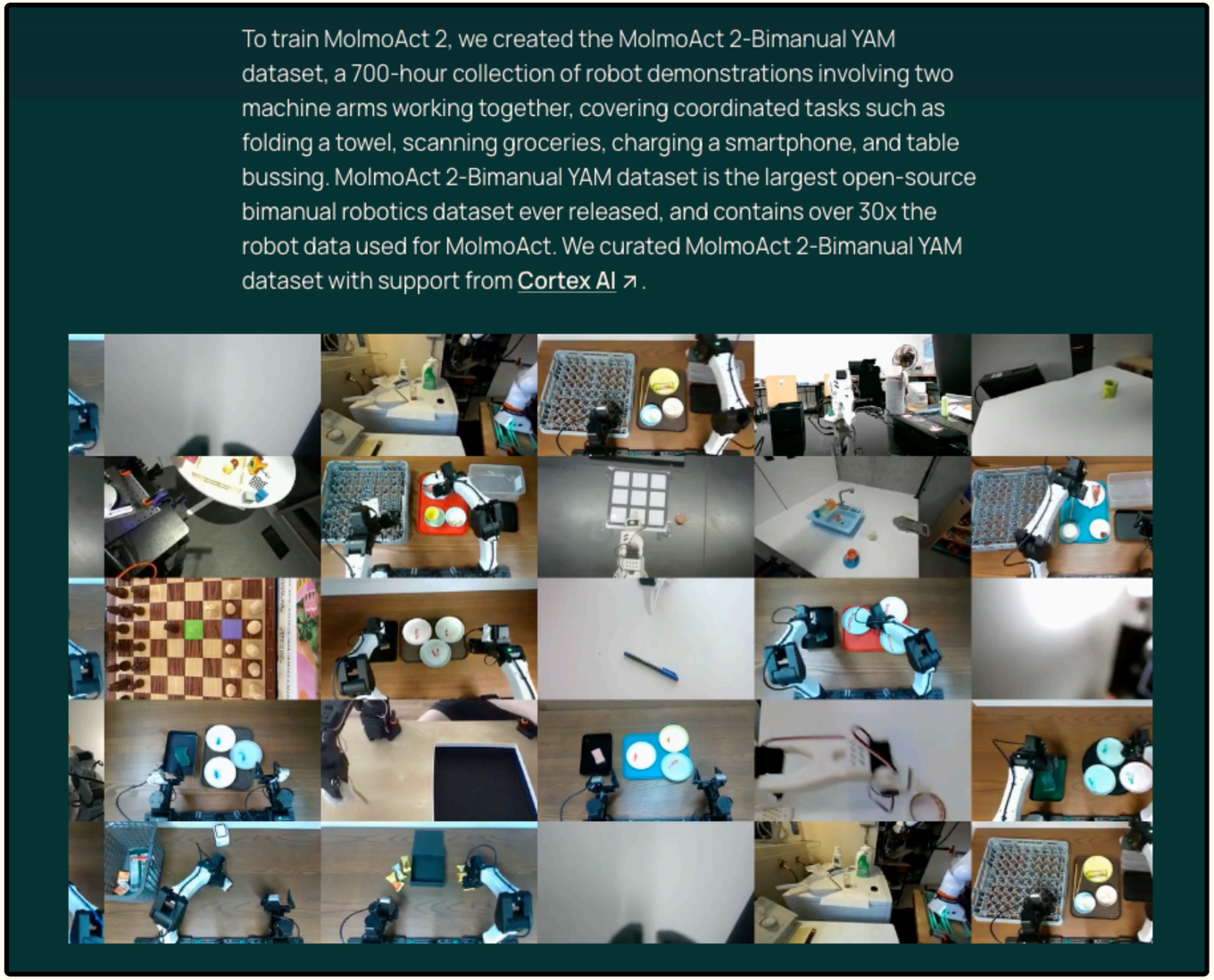
YAM arm - https://i2rt.com/collections/yam-arm

THE ROBOTICS STACK



data
software
hardware
models
evaluation
operations

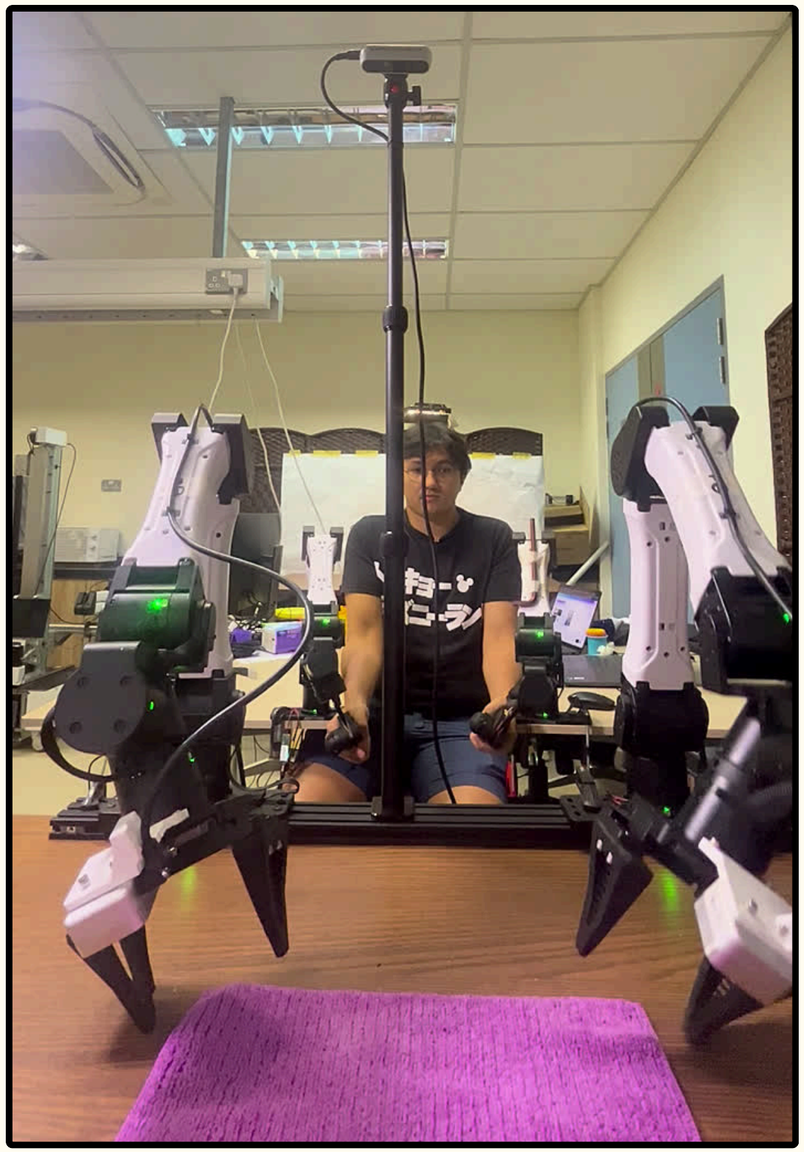
Robotics is also a data and operations problem



720h MolmoAct2 Bimanual YAM dataset

The largest open bimanual dataset to date

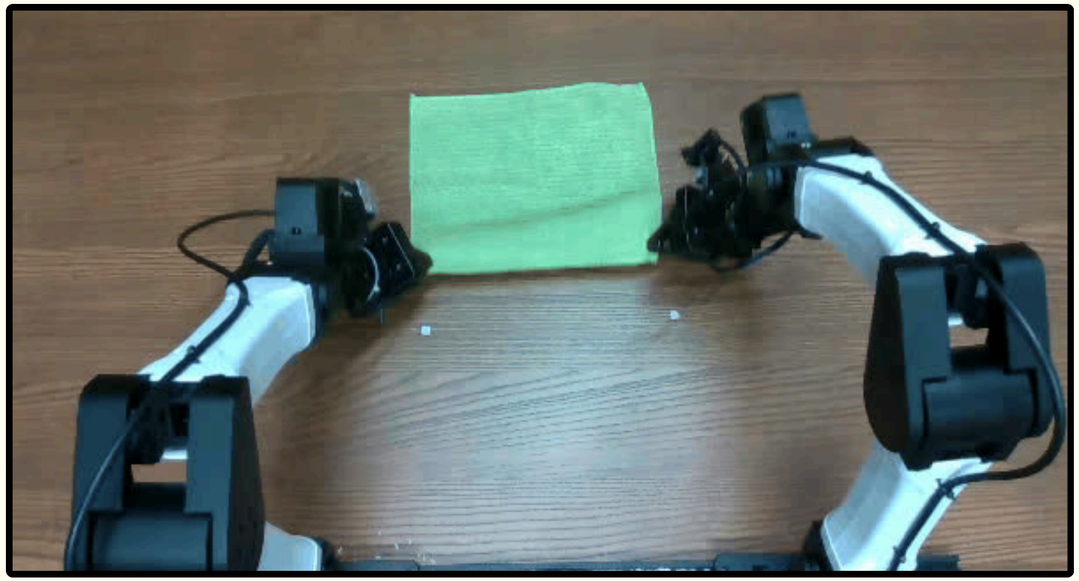
https://huggingface.co/collections/allenai/molmoact2-bimanualyam-dataset



Teleoperating Robots is Hard!

Human intuition does not transfer perfectly to a new body.

High-quality data starts before recording



1

Plan the task

Define task steps and success conditions

2

Practice the Motion

Make the behavior smooth and repeatable

3

Collect Consistently

Reduce variation across episodes and operators

Small workflow changes compound

FRICTION	WHAT WE CHANGED
Waiting for every episode to encode	Encode after the session
Encoding blocks the setup	Disconnect hardware and start new session

RESULT

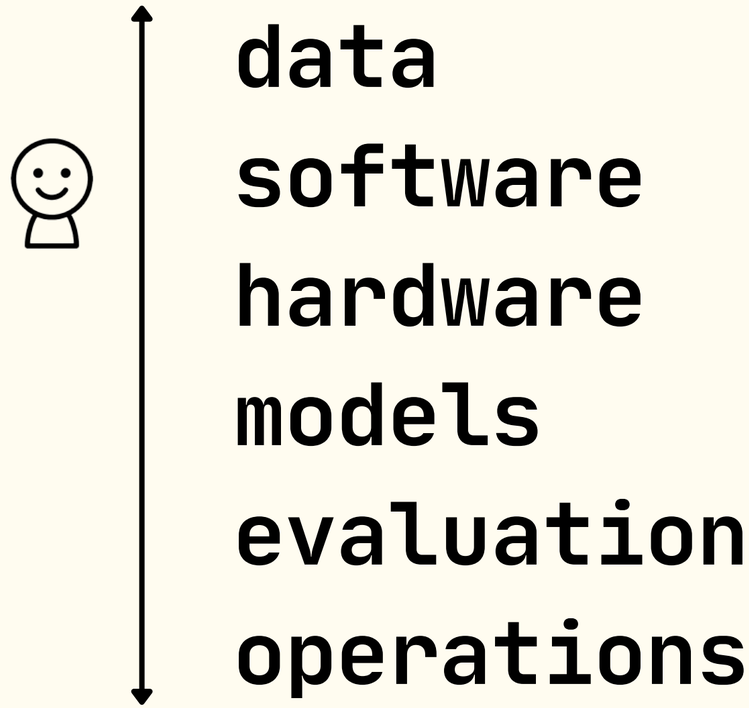
DATA RECORDING

ENCODING/UPLOADING

Completely decoupled!

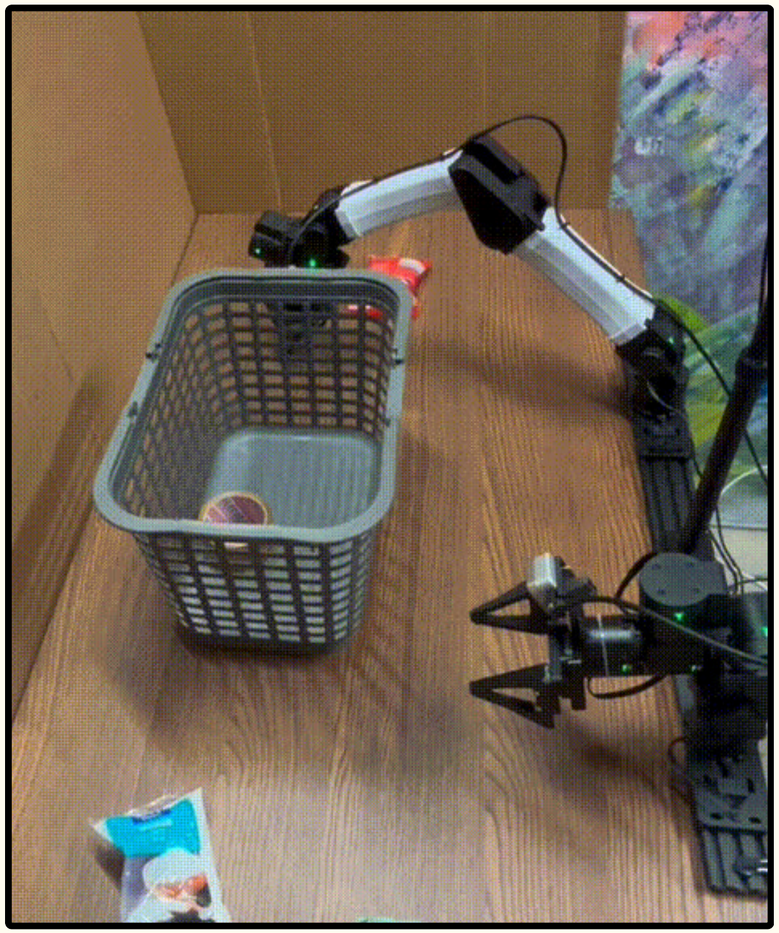
Breadth matters more than ever in Robotics today

OPERATING ACROSS THE STACK



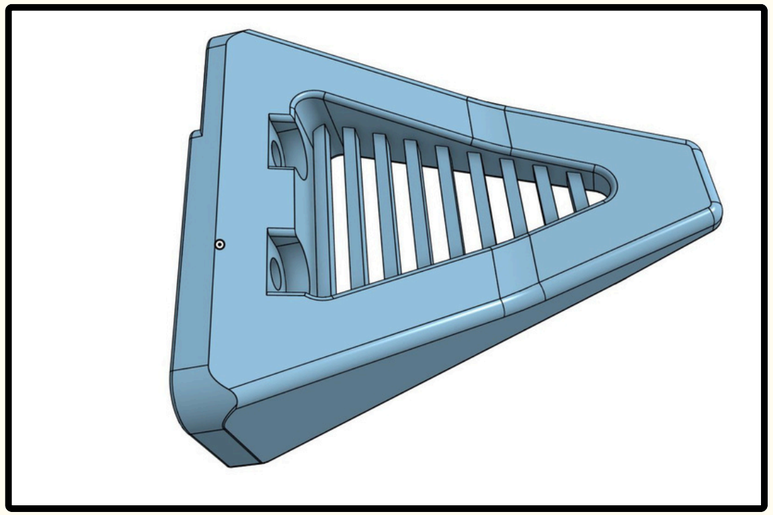
data
software
hardware
models
evaluation
operations

Breaking Grippers



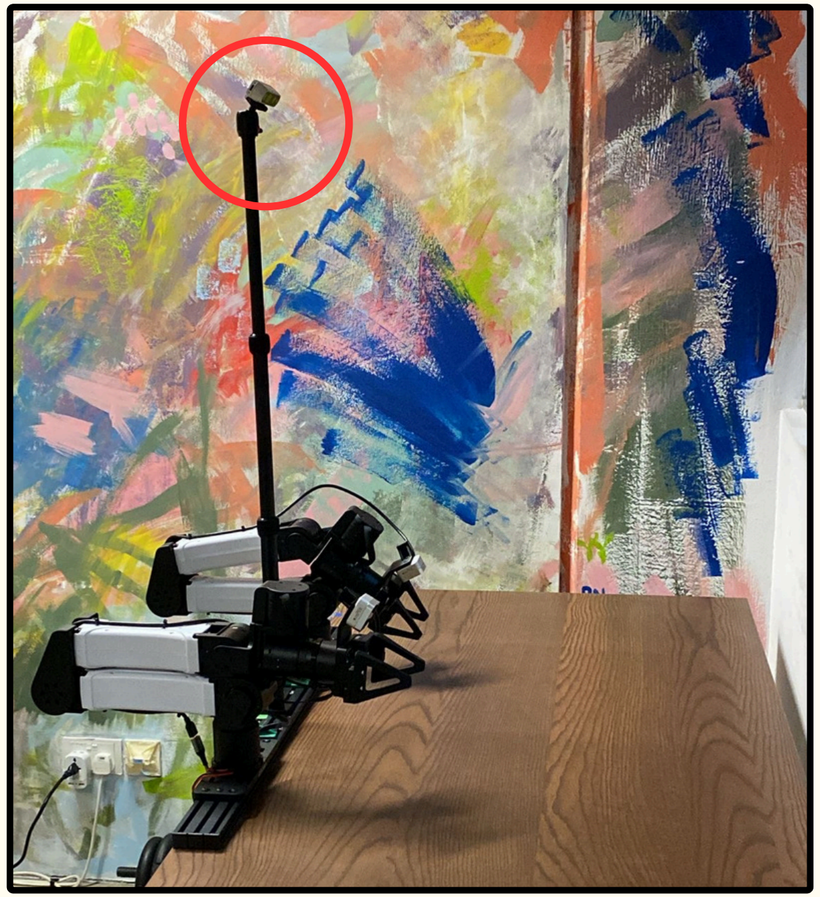
```
self.gripper.set_max_force(x)
```

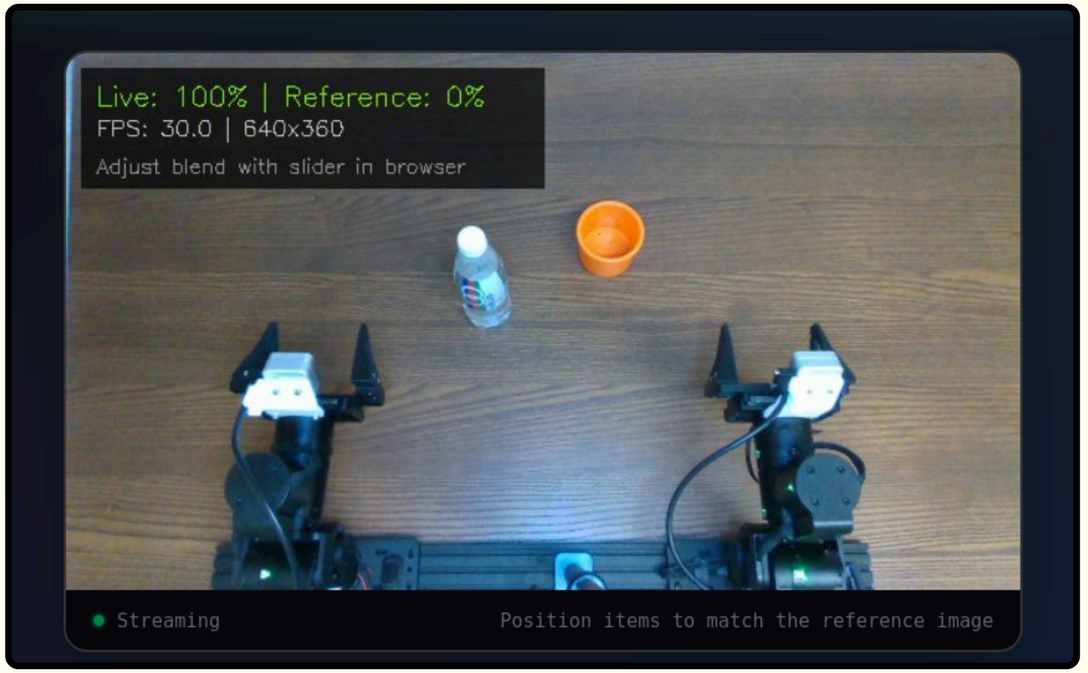
↓



Hardware solution for a **Software problem!**

Moving Cameras





Software solution for a **Hardware problem!**

Evaluation is still one of robotics’ biggest bottlenecks

Evaluating Robots

SIMULATION

VS

REAL-WORLD

Simulations are useful but diverse real-world environments capture true reliability and edge cases.

To validate MolmoAct 2 beyond our own lab evaluations, we retained Cortex AI, a robotics data and evaluation company, to conduct a third-party benchmark of MolmoAct 2's real-world fine-tuning performance. Cortex AI evaluated five robotics policies including MolmoAct 2 across multiple bimanual tasks using a systematic many-trial setup.

MolmoAct 2 achieved the highest average score at 0.51, ahead of OpenVLA-OFT at 0.36, π0.5 at 0.32, Cosmos Policy at 0.16, and X-VLA at 0.05. It also had the strongest task-level showing, ranking first on 7 of 8 tasks, including returning a test tube to a tray, storing candy, putting tools away, putting toys away, storing cups, preparing the pipette tip, and making popcorn.

Cortex AI YAM bimanual benchmark results

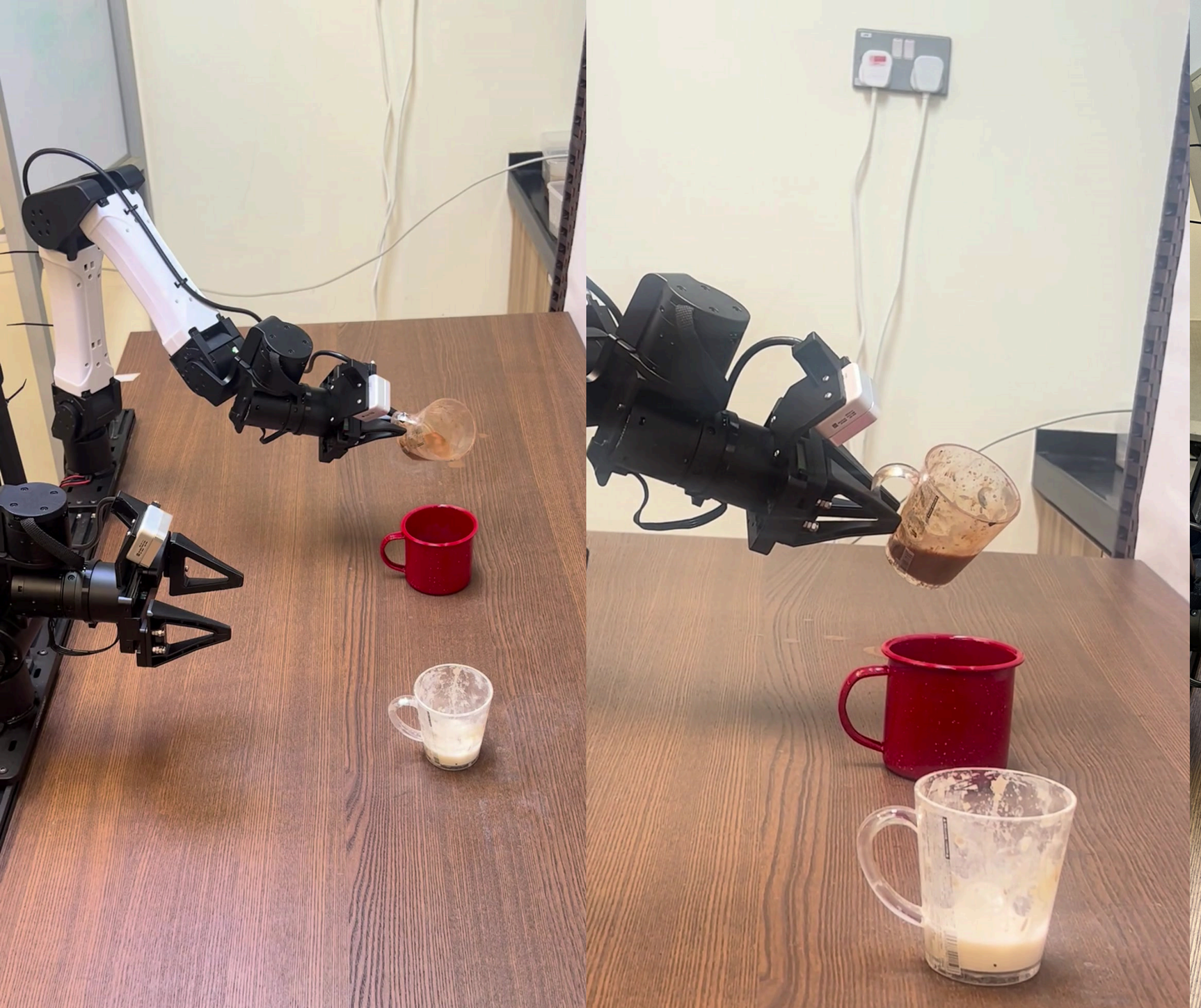
Task scores across Cortex AI's 8-task YAM bimanual real-world evaluation, rounded to two decimal places. Best score in each row is bold and underlined.

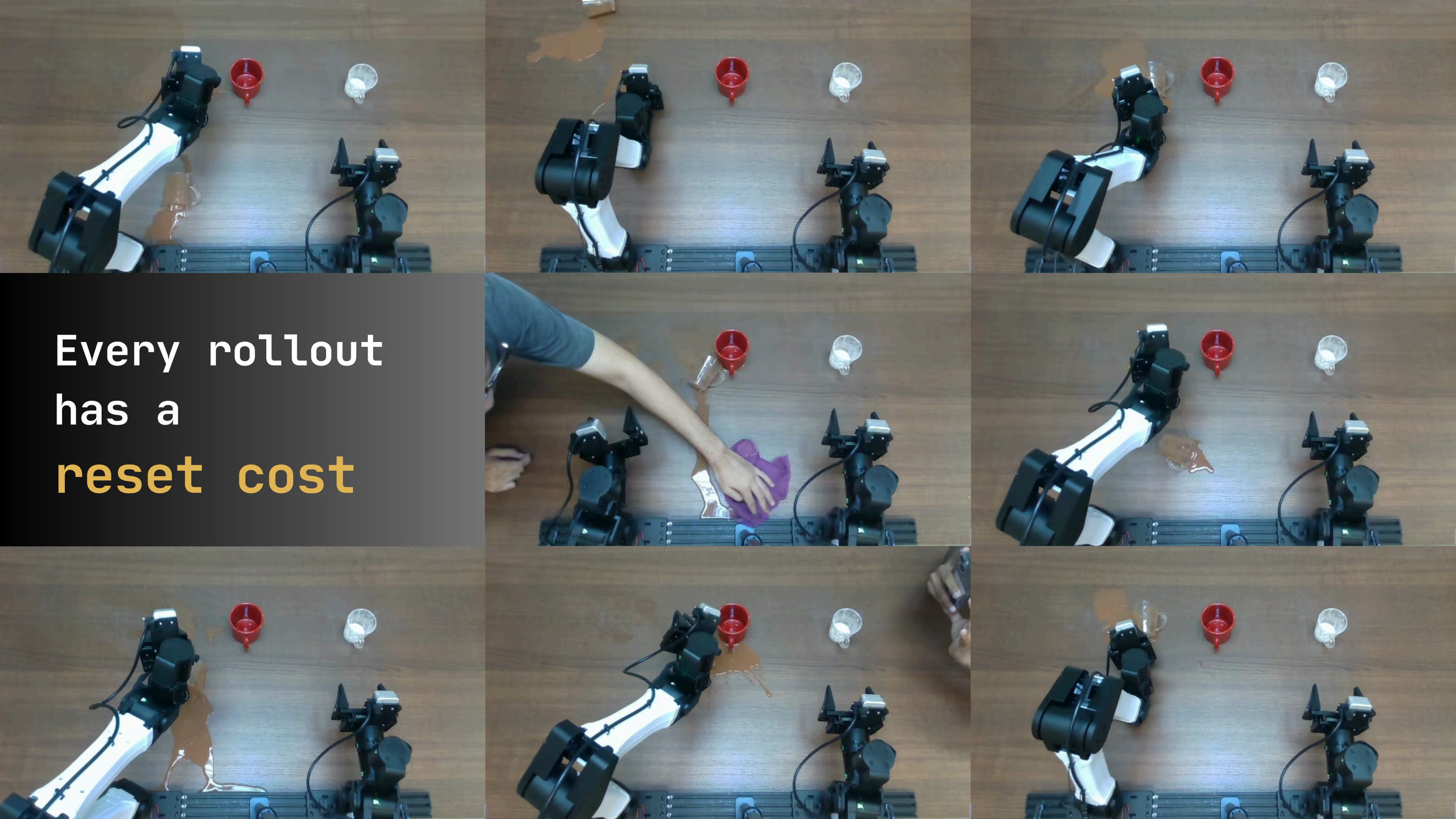
TASK	π0.5	COSMOS POLICY	X-VLA	OPENVLA-OFT	MOLMOACT 2
Cups stacking	0.62	**0.65**	0.06	0.61	0.54
Put the test tube back	0.17	0.11	0.00	0.42	**0.67**
Store candy	0.26	0.00	0.04	0.27	**0.41**
Put tools away	0.02	0.00	0.03	0.05	**0.14**
Put toys away	0.26	0.07	0.03	0.02	**0.32**
Store cups	0.86	0.00	0.08	0.96	**1.00**
Prepare the pipette tip	0.28	0.32	0.12	0.29	**0.33**
Make popcorn with LinearBot (LinearBot)	0.12	0.13	0.05	0.24	**0.64**
Average	**0.32**	**0.16**	**0.05**	**0.36**	**0.51**

Cortex AI YAM Bimanual Benchmark

1000s of real world evaluation rollouts across 5 policies

https://allenai.org/blog/molmoact2




Every rollout
has a
reset cost

When a robot fails, where did it actually fail?

setup · data · training
robot execution · evaluation

Robot failures can come from many layers

DATA

·inconsistent demos ·task strategy mismatch

TRAINING

·LoRA vs Full fine-tuning

SETUP

·model loading issue ·wrong action chunk size

EVALUATION

·object placement

and more....

Safety is part of the stack

Every stage has a human safety surface.

DATA COLLECTION

Leader arm failure → operator suddenly bears the load

EVALUATION

Evaluation failures → physical hazards
Stale action queues → sudden arm movements

and more....

Running AI-written code on Robots

https://www.anthropic.com/research/project-fetch-robot-dog

Run it like it can fail

1 Offline checks

type checks, unit tests, manually check if business logic is correct

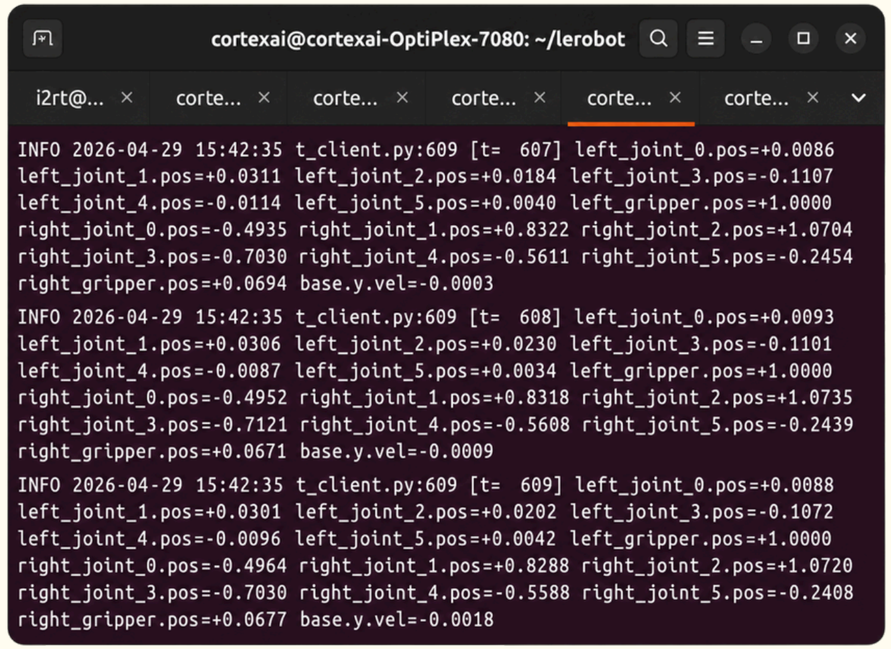
2 Replay/ Simulation

test on logs or in simulation when possible

3 Supervised robot runs

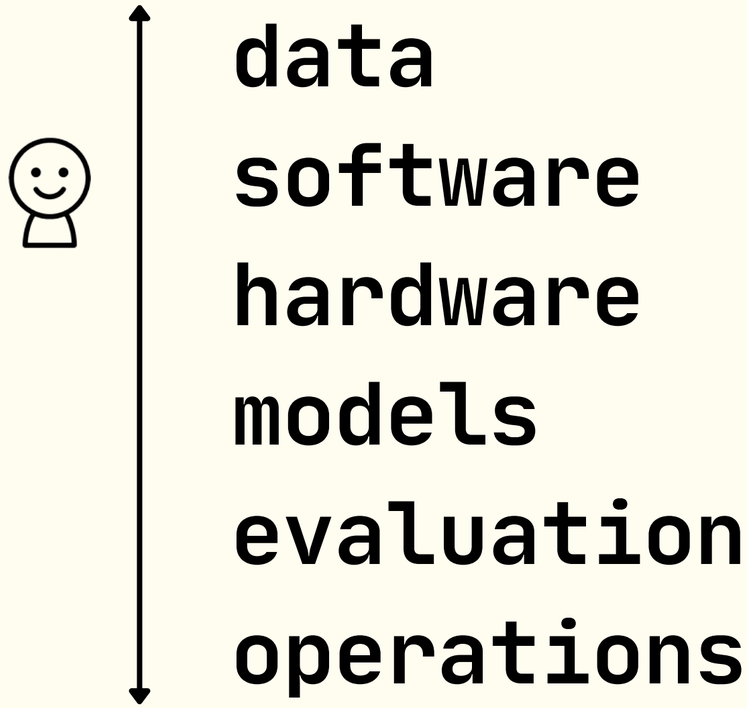
slow speed, move a single joint first

operator + E-stop



Biggest takeaway

The faster you can move across the stack, the faster you can solve problems.



We’re hiring people who like working across the robotics stack.
cortexrobot.ai/careers

Thank you!

cortexrobot.ai

@suveene