Transferring Foundation Models for Generalizable Robotic Manipulation

Jiange Yang[1] Wenhui Tan[2] Chuhao Jin[2] **Keling Yao**[3]
Bei Liu[4] Jianlong Fu[4] Ruihua Song[2] Gangshan Wu[1] Limin Wang[1,5*]

[1] State Key Laboratory for Novel Software Technology, Nanjing University, China
[2] Renmin University of China
[3] The Chinese University of Hong Kong, Shenzhen
[4] Microsoft Research
[5] Shanghai AI Lab

What are Universal Robotic Agents?

Objectives: Creating a robotic agent that is

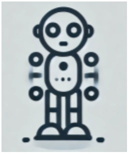
1) **general-purposed**
2) performing **diverse tasks**
3) fulfill **human daily needs**
4) in **real-world**

Motivated Toy Example:

Initial Scene

User instruction: “I want to take a shower”

Robot Agents: “You need the towel”



Perception

Reasoning

Decision

Execution

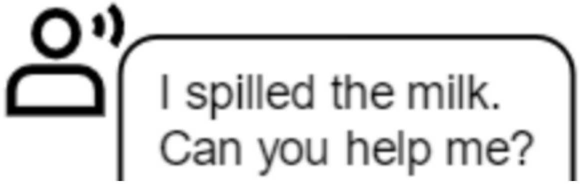
Challenges:

- Effectively converting abstract task instructions into specific robot inputs
- Enhancing the generalization capabilities of a single robot model to handle multiple tasks

✦ Current Task condition forms

task identifiers, goal images, human videos, natural languages

Language is natural, scalable, but under-specified and ambiguous



Challenges:

- Effectively converting abstract task instructions into specific robot inputs
- Enhancing the generalization capabilities of a single robot model to handle multiple tasks

✦ Current Robot data-driven methods

RT-1, Octo, OpenVLA, π_0

Data collection is expensive and time consuming

Exhibits limitation to compositional generalization, struggling with unseen objects and environments

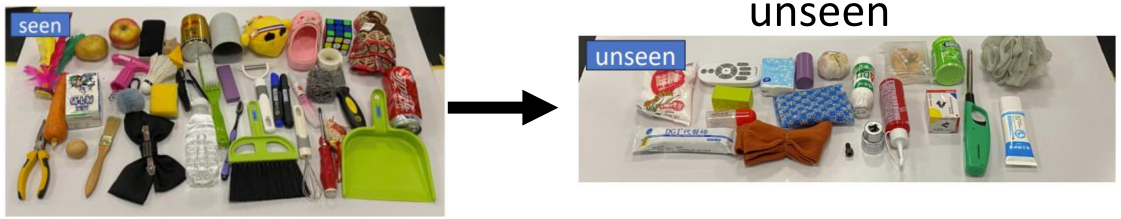
RT-1 Dataset Collection

Environments:

seen unseen

Objects:

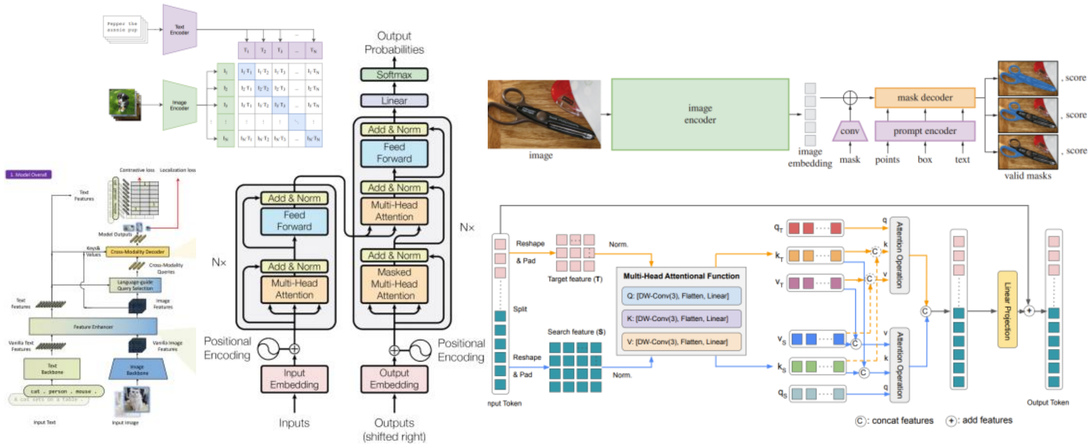
seen unseen



Introduction

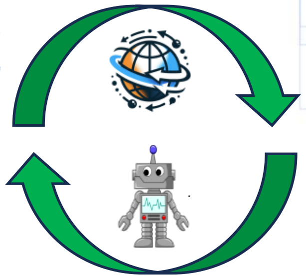
- **Internet-scale foundational models**

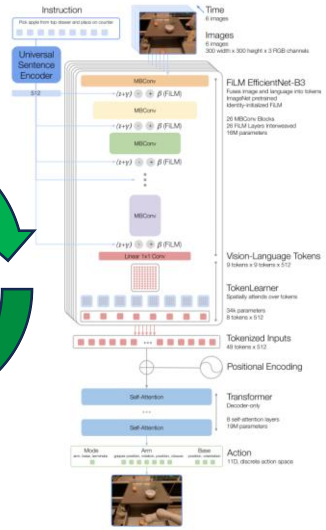
Rich data sources but lack physics

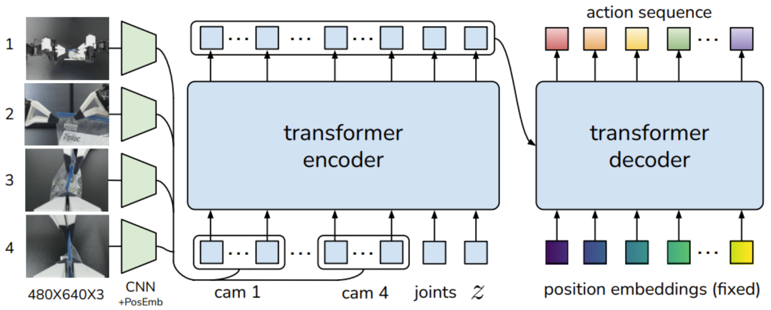


- **Standard behavior cloning from pixels to actions**

Lower sample efficiency but can learn form demos



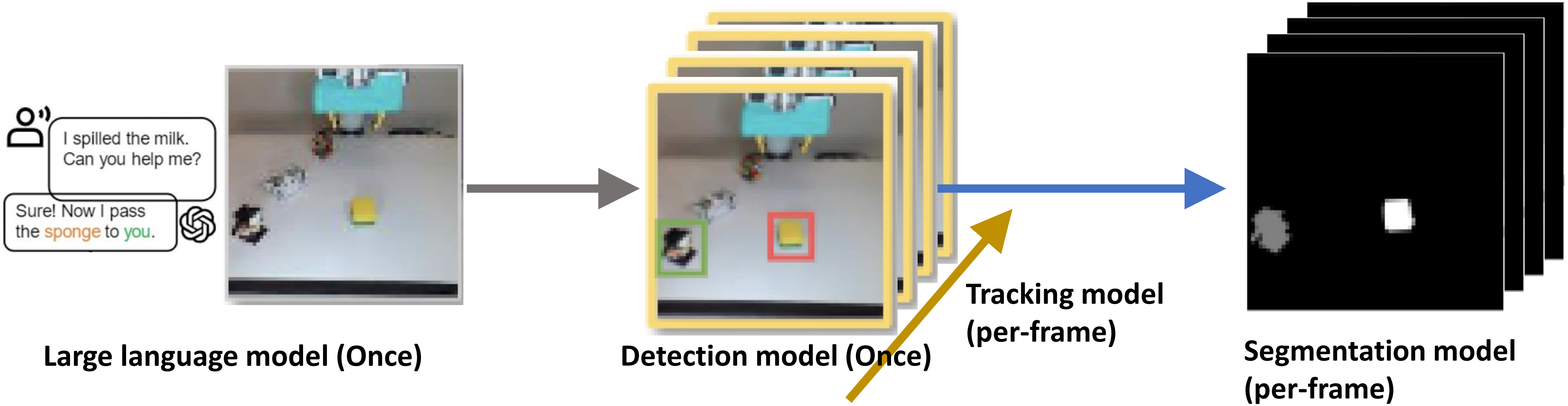




Combine Internet-scale foundational models with Behavior Cloning Policy Model!

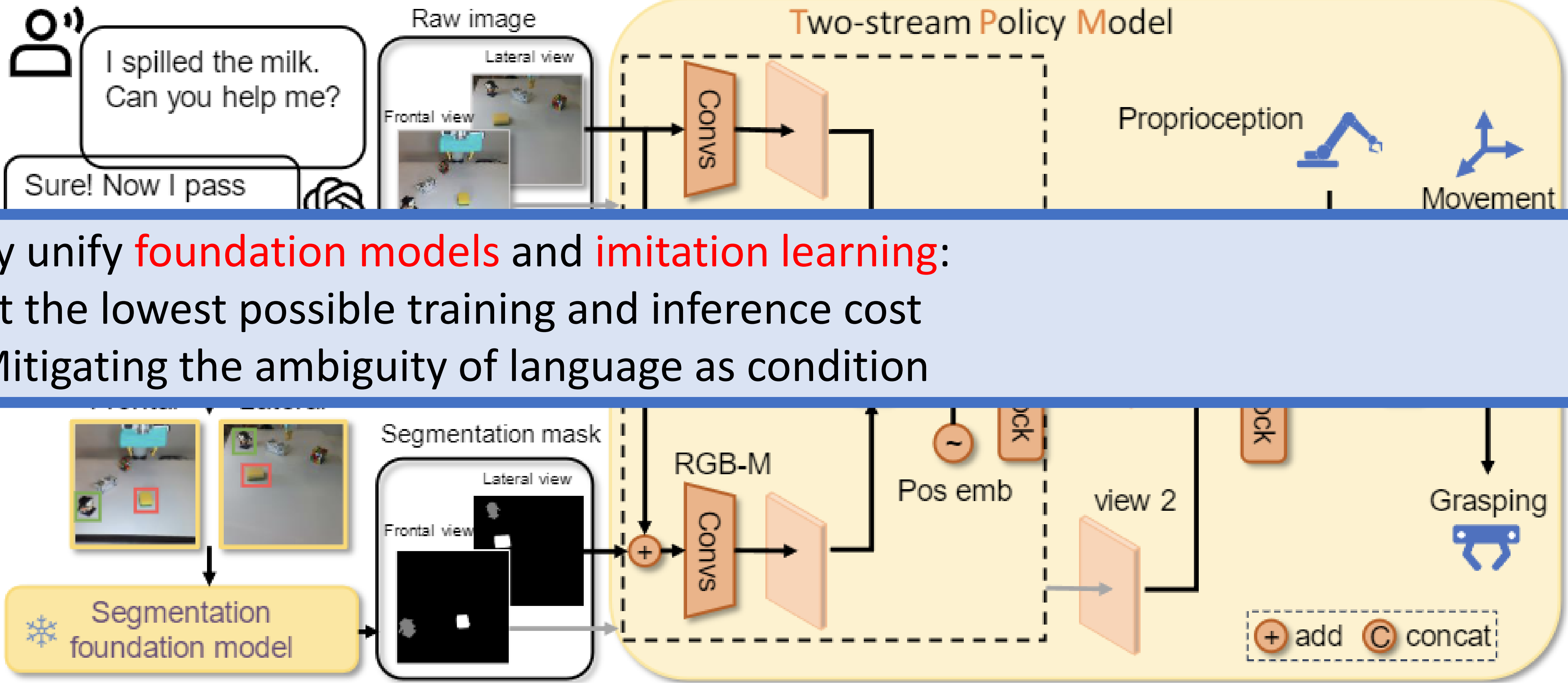
Universal language-reasoning mask

- **Large language models:** Reasoning and Planning (GPT-4)
- **Detection models:** Semantic recognition (Grounding Dino)
- **Tracking models**: Temporal correlation (MixFormer)
- **Segmentation models:** Spatial geometry (SAM)



We propose to achieve sample-efficient generalization for robotic manipulation by introducing **language-reasoning mask** modality containing semantics, geometry, and temporal correlation priors inherent from internet-scale vision foundation models into an end-to-end policy model

Our Paradigm



Fully unify foundation models and imitation learning:

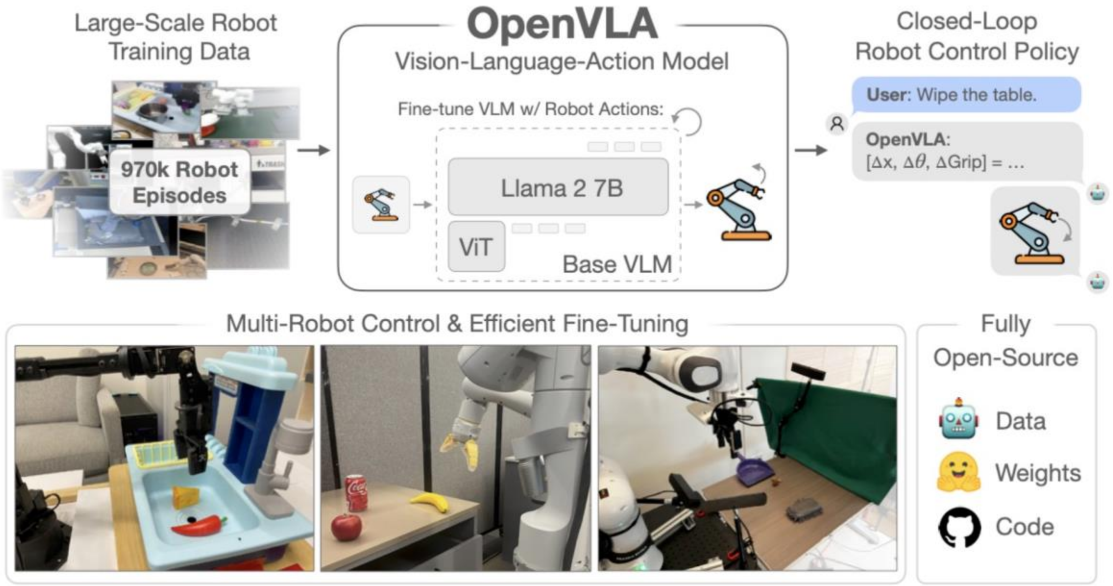
1. At the lowest possible training and inference cost
2. Mitigating the ambiguity of language as condition

Universal language-reasoning mask generator

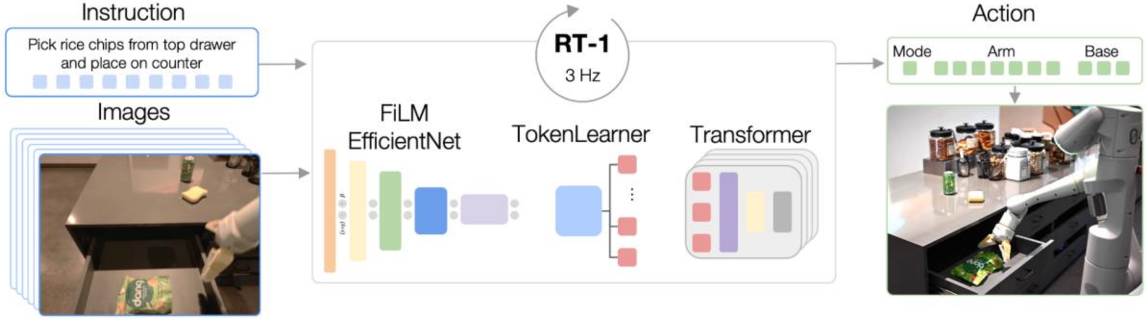
Two-stream architecture Policy Model (TPM)

$$\pi_\theta(p, (o_1, m_1), (o_2, m_2))$$

Compared with E2E policy model: (RT-1, *OpenVLA*, π_0 etc.)



OpenVLA

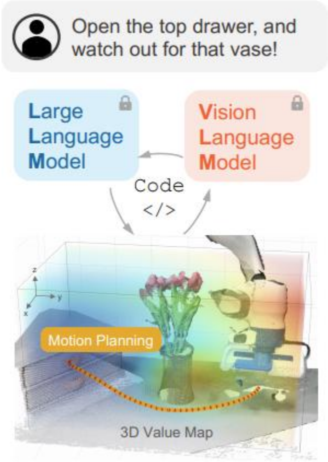


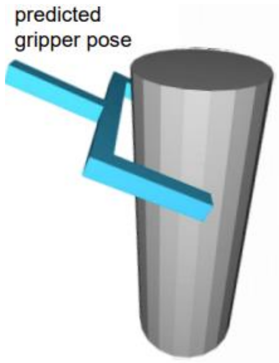
RT-1

- Our paradigm only inference ONCE for Large-Language-Model

- Frozen all foundation models, achieving Sample-Efficient Training and Resource-Efficient Training

Compared with multi-stage non-training methods: (voxposer, anygrasp, etc.)

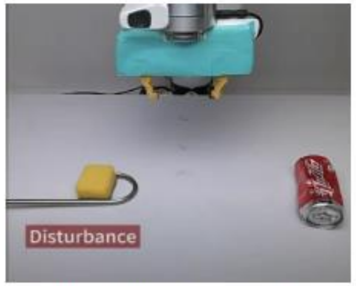




achieve visual perception → Conduct Motion Planning

- Our paradigm dynamically receive raw image as input and output continuous action in a close-loop manner, which does not rely on depth calibration and completely accurate object masks.

- We can deal with multiple skills, transparent and disturbed objects, as well as unstructured environments with collision situation.

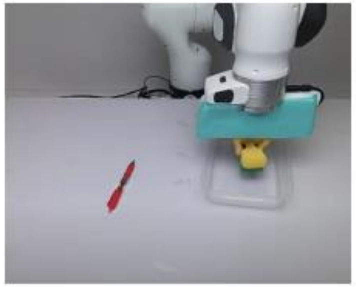


Pick and Place near (seen)

Open drawer (seen)

Open drawer (unseen)

Pick and Place on (unseen)



Pick and Place inside (seen)

Pick and Place on (seen)

Experiments

Figure 3. (a): Overview of our workstation, which has a Franka robot arm, a frontal view camera, and a lateral view camera. (b): Seen and unseen objects in the experiments. (c): Three backgrounds in the training data. (d): A challenging background with complex texture for new background evaluation.

Experiments

Scenario	Seen	Unseen	Average
Standard	82.5	80.0	81.25
New background	65.0	55.0	60.0
More distractors	75.0	70.0	72.5

Table 1. Experimental results evaluated on different scenarios.

Method	Seen	Unseen	New background	More distractors	Average
Ours	82.5	80.0	65.0	75.0	**75.625**
-MOO-like [65]	50.0	42.5	27.5	35.0	38.75
-RT-1-like [4]	65.0	0.0	20.0	60.0	36.25
-replace mask with bbox	50.0	40.0	25.0	30.0	36.25
-w/o tracking	70.0	50.0	55.0	70.0	61.25
-single view	65.0	80.0	20.0	70.0	58.75
-RGB-M only	85.0	70.0	50.0	70.0	68.75

Table 2. Comparison of our method and its variants on various settings.

Experiments

	GPT-4 [52]	DetGPT [56]	MiniGPT-4 [80]
Success Rate	**0.95**	0.75	0.2

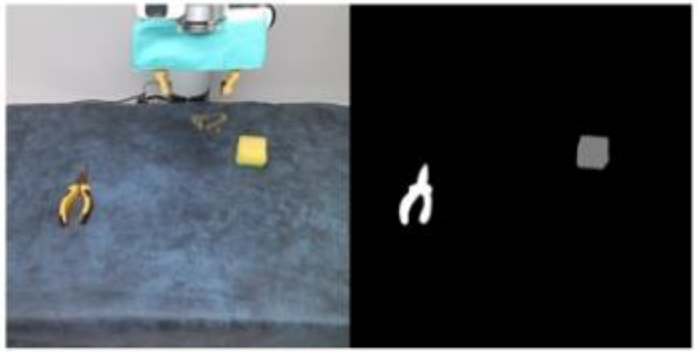
Table 3. The reasoning performance comparison of LLMs.

	GroundingDINO-B	Mixforemr-B	MixformerV2-S	SAM-B	SAM-T	TPM
Inference Time (ms)	148.6	103.4	17.0	18.2	10.1	34.8

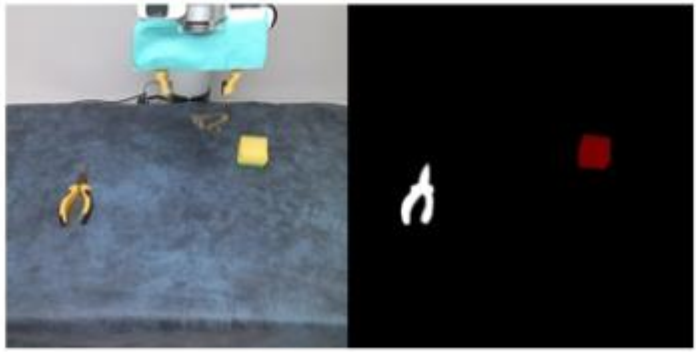
Table 4. The inference time for different modules and model sizes.

Other skills

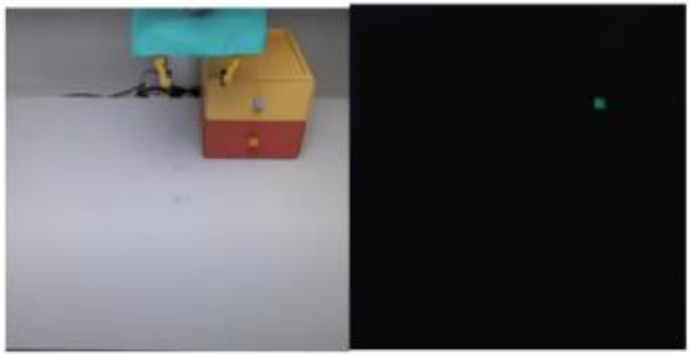
Our method can be flexibly extended to some common skills by transferring language instruction of skills to mask values of manipulated objects



Pick and Place near



Pick and Place on



Open

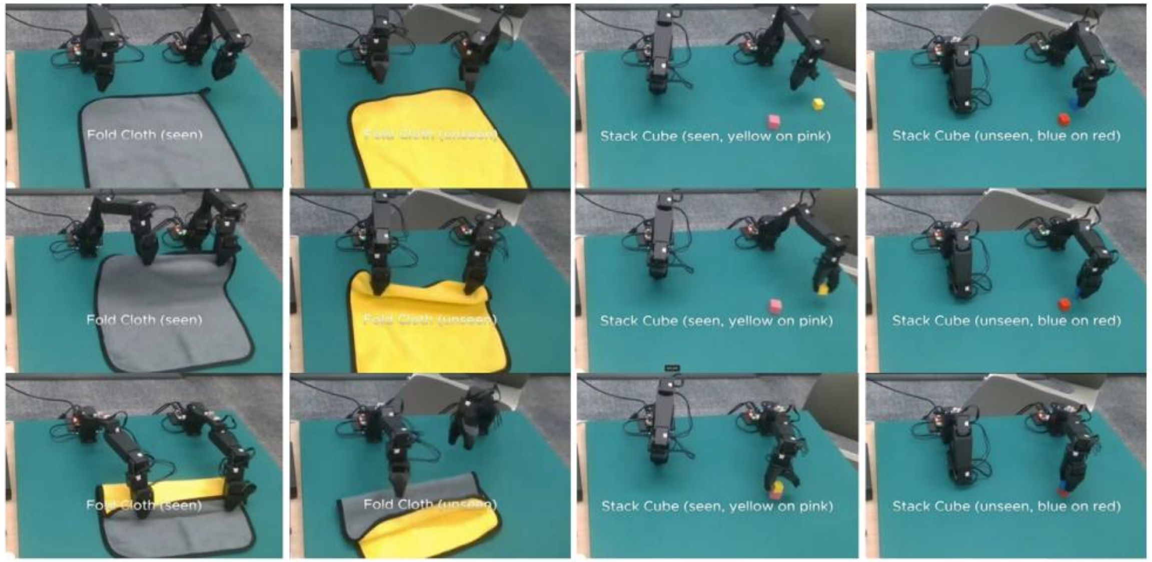


Figure 1. The demos of folding cloth and stacking cube skills.

Limitation & Future!

Which path let to Universal Robotic Agent

- Each new skill corresponds to a new mask value
- Predefined prompt are need for LLM

1

Perception ↓ **Reasoning** ↓ **Decision** ↓ **Execution**

OpenVLA,
Π_0,
RT-1,
Etc.

Unified E2E policy Model

2

Perception ↓ **Reasoning**

↓

Decision ↓ **Execution**

cerebrum

cerebellum

Ours,
Hi Robot,
Helix,
Etc.

cerebrum & cerebellum Model

Thanks!

demos

code