ИНТЕЛЛЕКТ AIRI FusionBrain.Robotics

Lecture 4

Action Modality: Robots and Agents. Embodied AI

Zinkovich Viktoriia

Special thanks to **Vlad Shakhuro** for the slides and lecture content

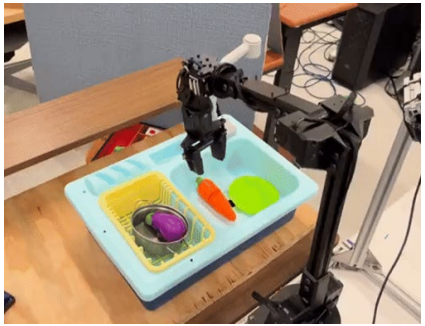
Introduction: Course Plan

Day 1
Image modality

Day 3
Data generation
in MLLMs

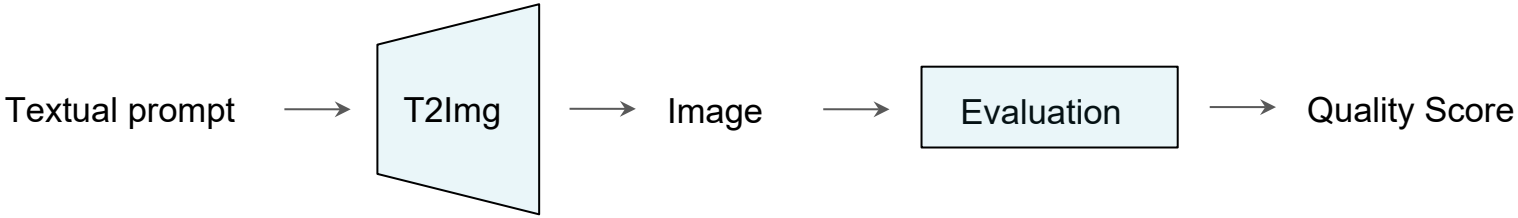
Day 5
3D models

Day 2
Video modality



Day 4
Action modality

Recap: Lecture #3

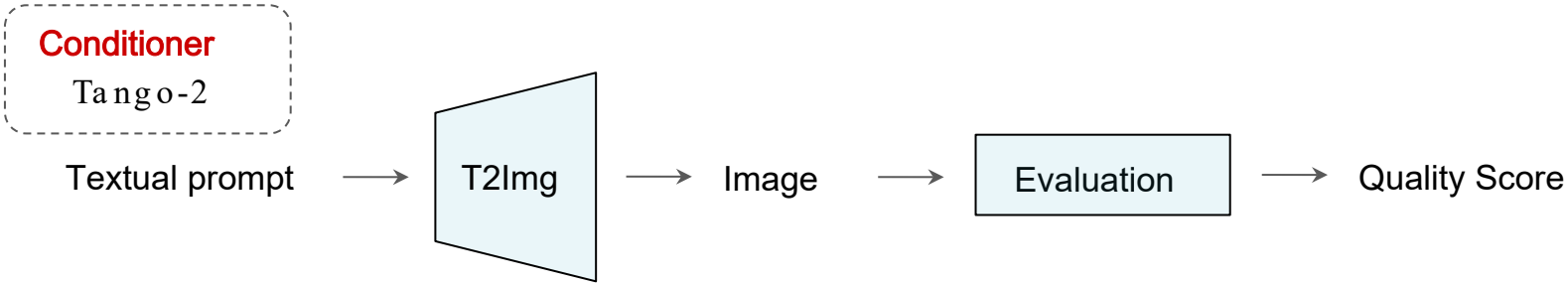


Dataset

Tool X
Tool X
Tool X

Tool X
Tool X
Tool X

Recap: Lecture #3

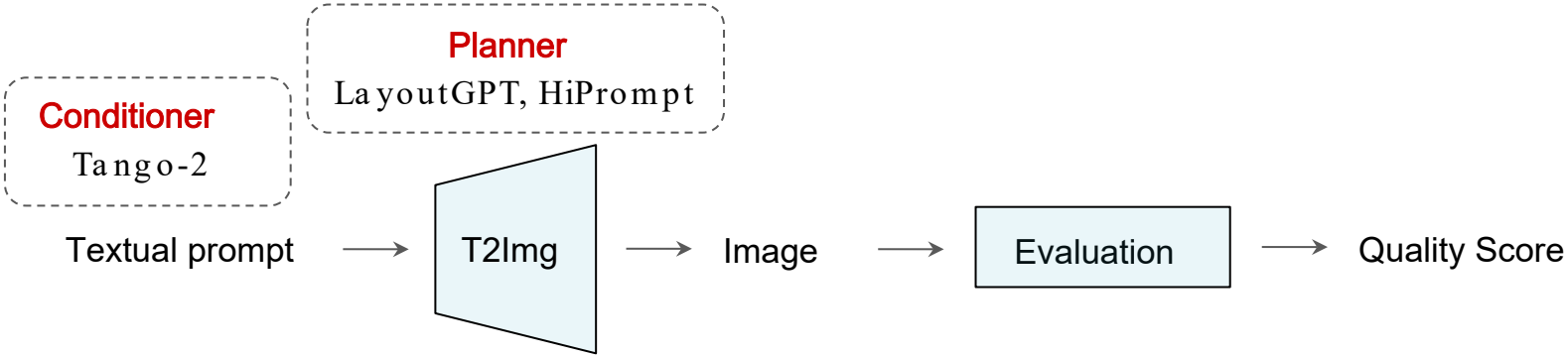


Dataset

Tool X
Tool X
Tool X

Tool X
Tool X
Tool X

Recap: Lecture #3

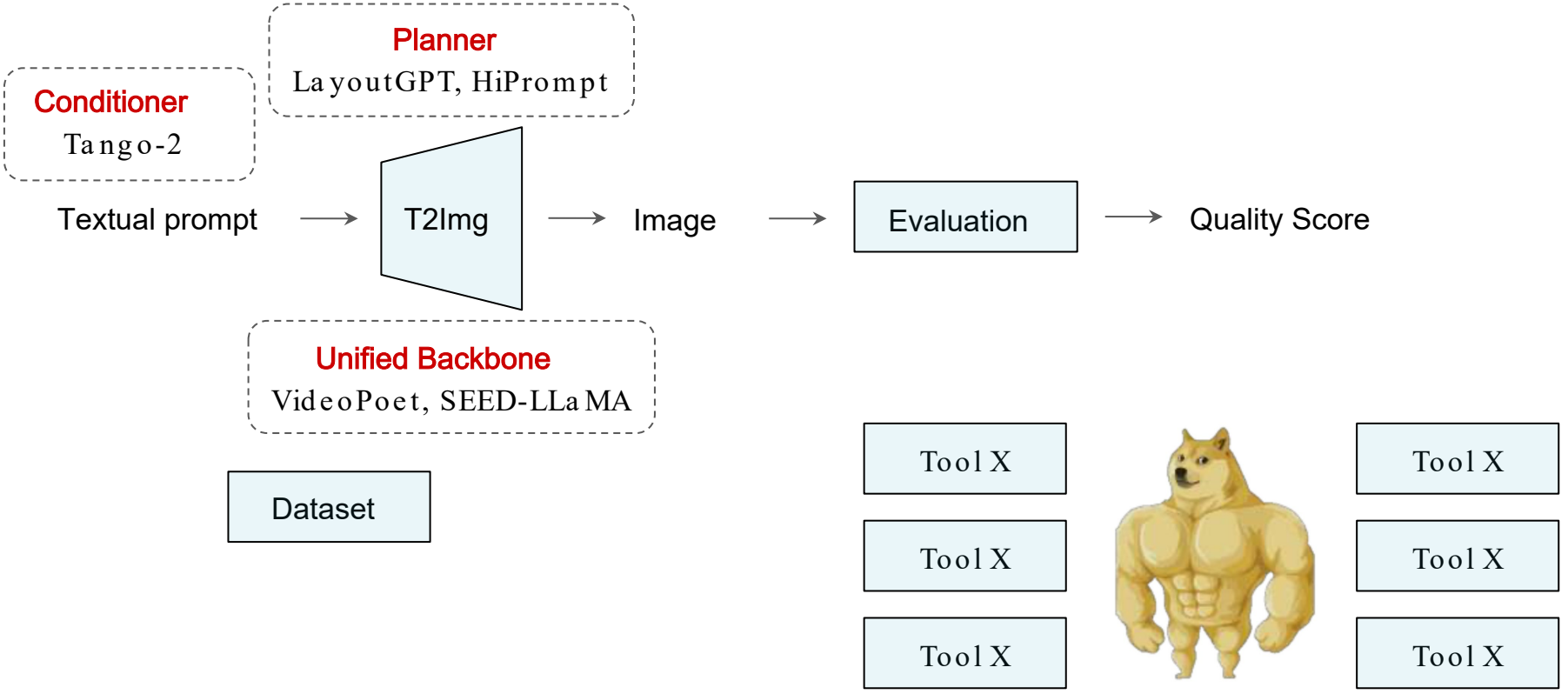


Dataset

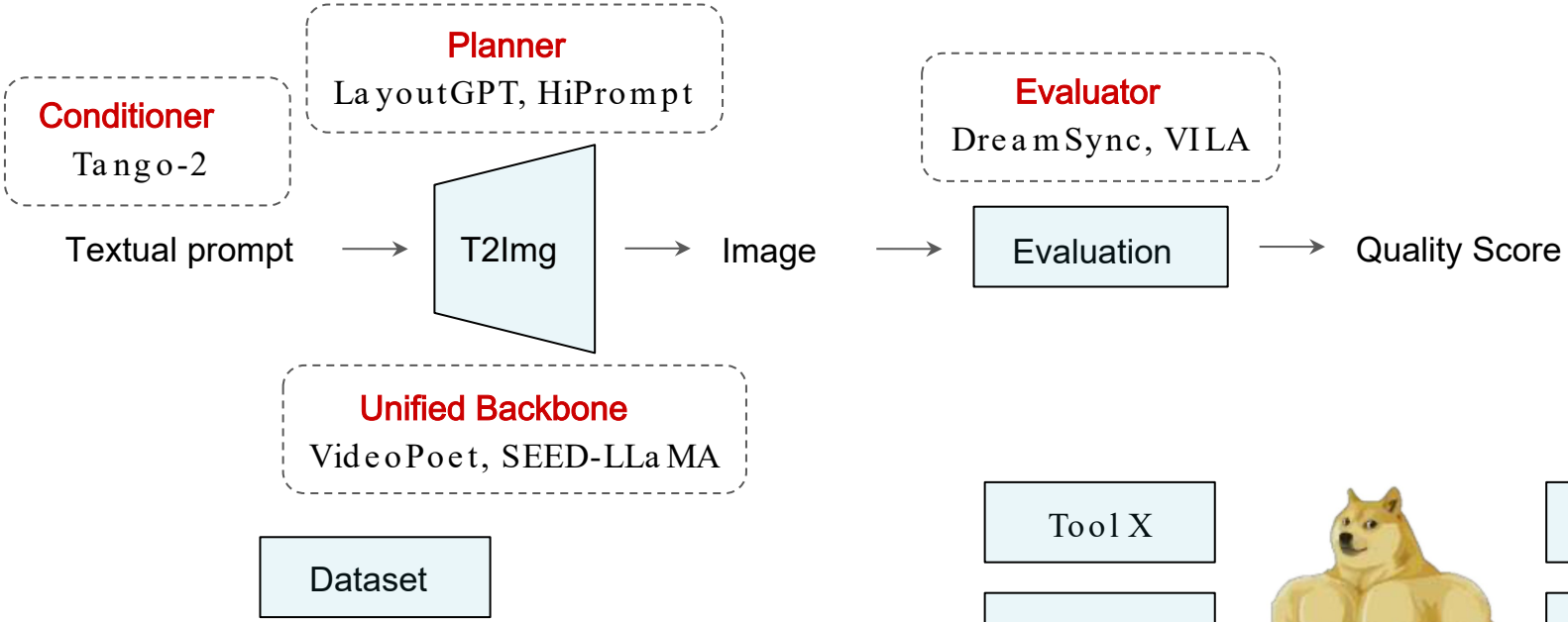
Tool X
Tool X
Tool X

Tool X
Tool X
Tool X

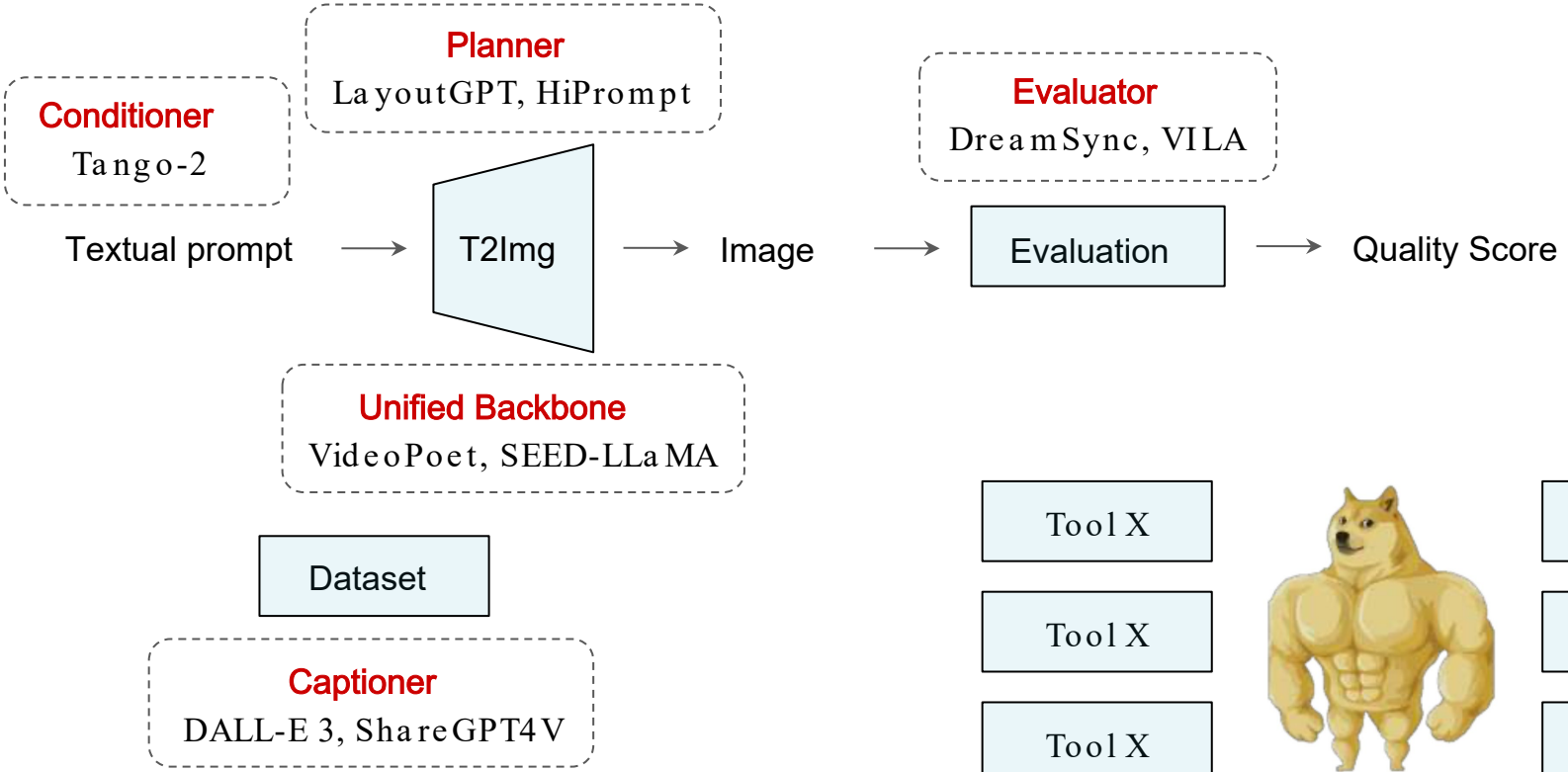
Recap: Lecture #3



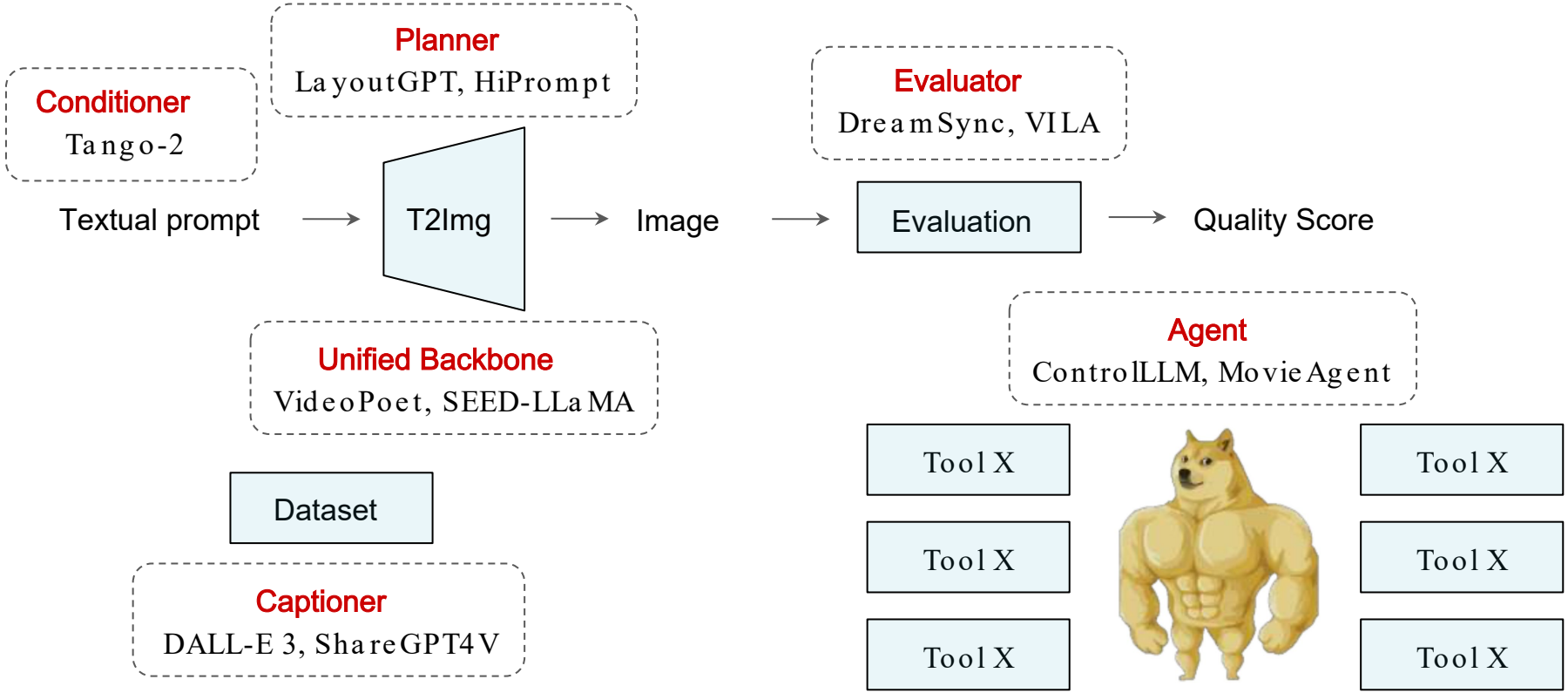
Recap: Lecture #3



Recap: Lecture #3

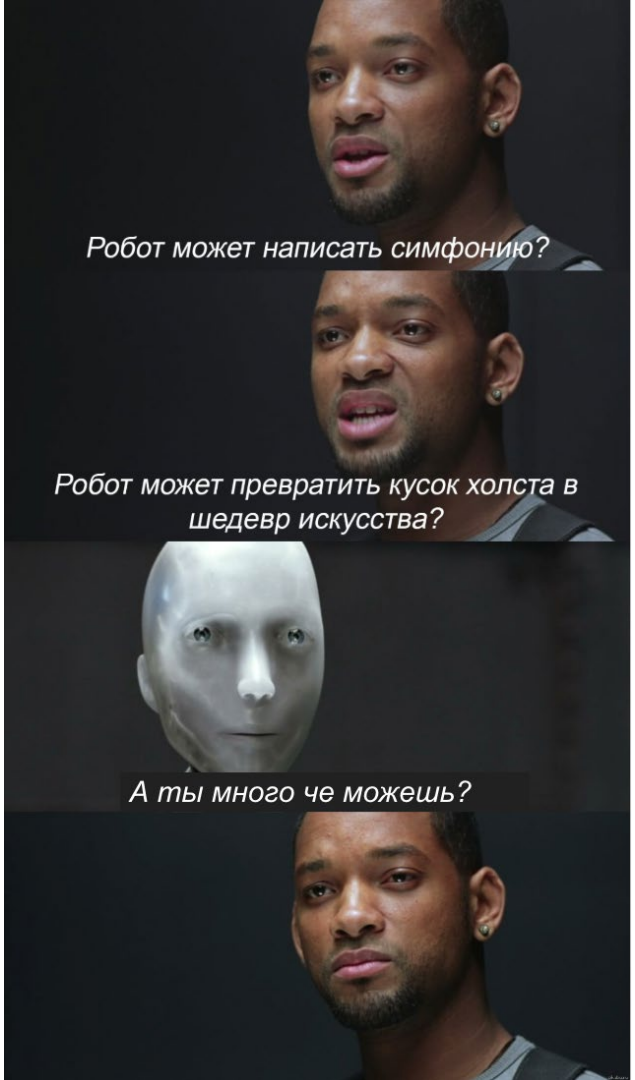


Recap: Lecture #3



Lecture Plan

1. What is Embodied AI?
2. **Evaluation** : sim and real
3. **Understanding** the world
4. **Planning**
5. **Acting** : manipulation & navigation



1

Introduction

What is Embodied AI?

Introduction: Embodied AI

Create intelligent agents (i.e. robots) with physical **embodiment** that can solve challenging tasks

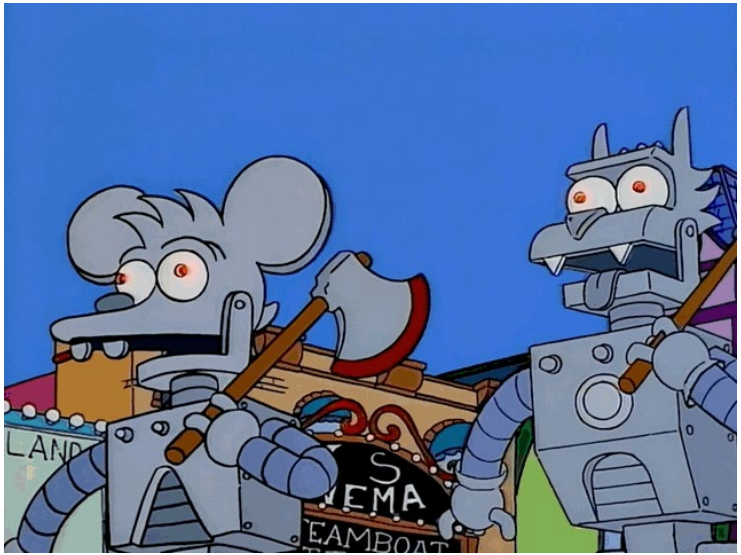
Such agents should be able to:

→**Perceive** —see, listen using various sensors

→**Talk** —natural dialog

→**Reason** —long-term consequences of actions

→**Act** —navigate and interact



Introduction: General-purpose Robots

fetch robot
with spring

Introduction: General-purpose Robots

fetch robot
with spring

robot dog
UniTree

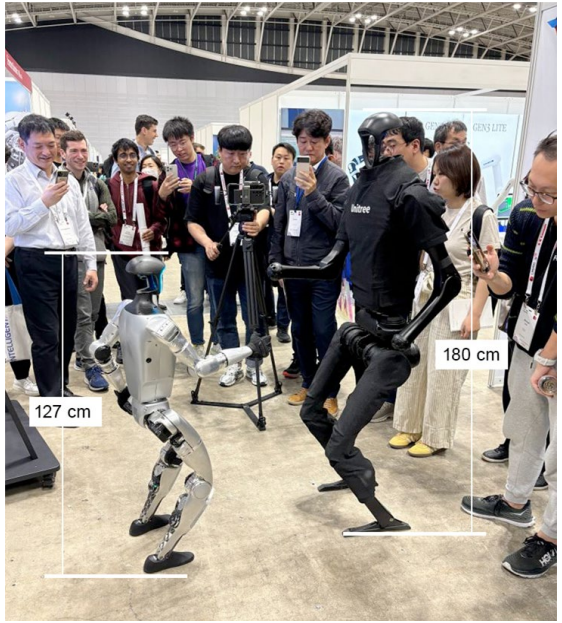
Introduction: General-purpose Robots

fetch robot
with spring

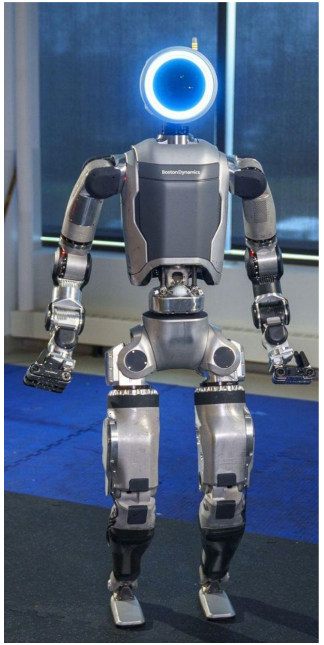
robot dog
UniTree

robot dog with hybrid locomotion
UniTree B2-W

Introduction: General-purpose Robots



UniTree G1

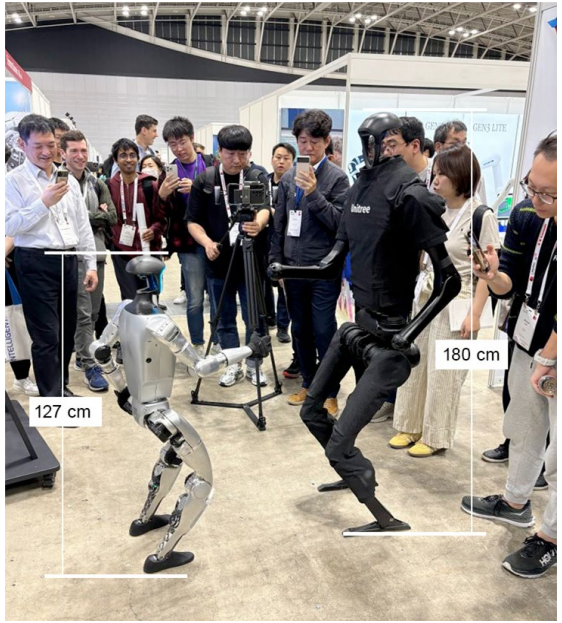


Atlas
Boston Dynamics

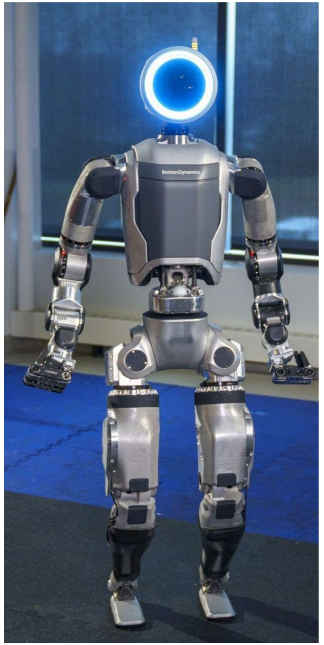
Humanoid robots (knees …)

- Embodiment **is not limited** to that list (drones, autonomic vehicles…)

Introduction: General-purpose Robots



UniTree G1



Atlas
Boston Dynamics

Humanoid robots

- Embodiment **is not limited** to that list (drones, autonomic vehicles…)

What could be **automated by a robot** ?

Introduction: General-purpose Robots

Thermonator, **Flame-Throwing** Robot Dog

9,420 $

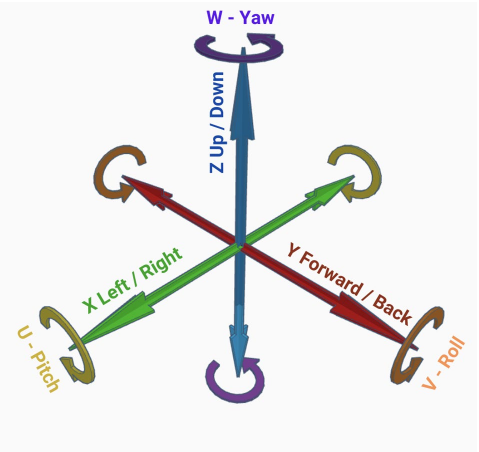
Humanoid robots

- Embodiment **is not limited** to that list (drones, autonomic vehicles…)

What could be **automated by a robot** ?

Introduction: General-purpose Robots

Each embodiment is characterized by **how we can control it**



robot with gripper
7 Degrees of Freedom

autonomic vehicle
steering and acceleration

2

Evaluation

Sim Vs Real. Simulators.

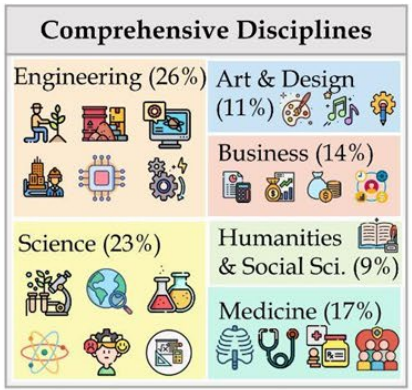
Real-world evaluation

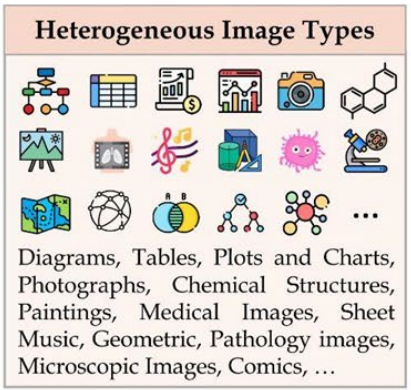
Evaluation: CV/NLP Benchmarks

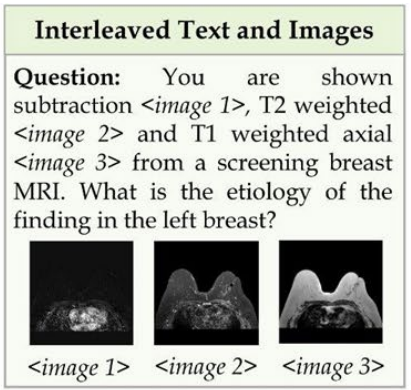
A lot of **CV/NLP benchmarks** are available for various tasks:

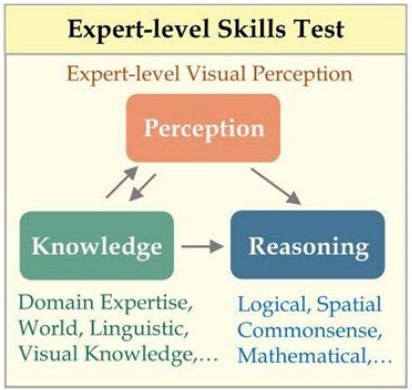
- → usually very **fast evaluation**
- → **reproducible**
- → **cheap**

- → have **big domain shift** in relation to robotics data
- → may be only a **proxy metric/guidance** for choosing baseline models







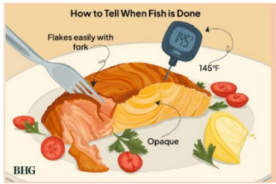


Yue et al. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. CVPR 2024

Evaluation: Arena

Rules

- Chat with two anonymous models
- Continue to chat until you identify a winner
- Vote for the better one with reason



What English words are on the lower right side of the fish meat?

Model A

On the lower right side of the cooked fish, the word "Opaque" is labeled.

Model B

The English word on the lower right side of the fish meat is "Opaque."

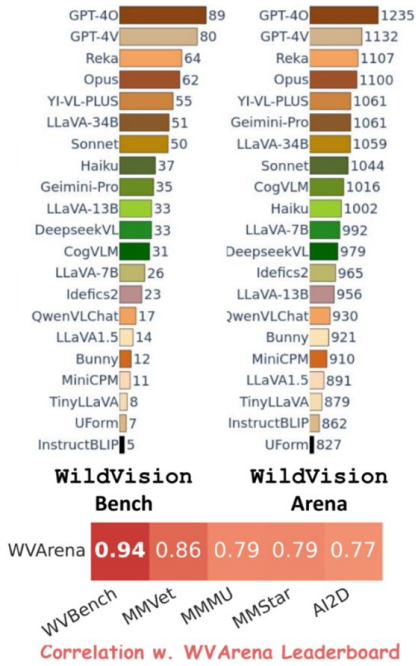
Reason Both Model A and Model B answer correctly regarding the text.

Vote A is Better | B is Better | Tie | Both are bad

Model A: Claude-3-Sonnet, Model B: GPT-4V

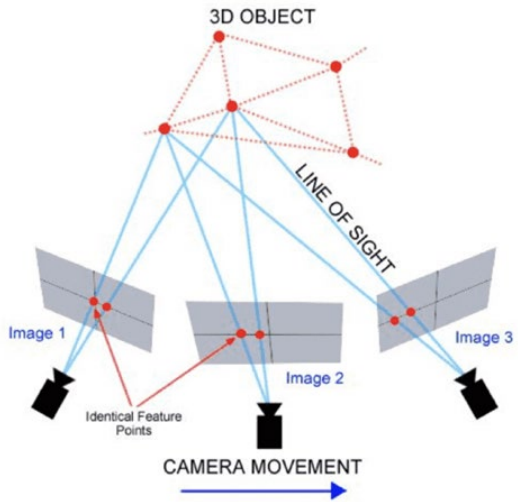
WVArena Elo Ratings

Submit



Evaluation: Simulators

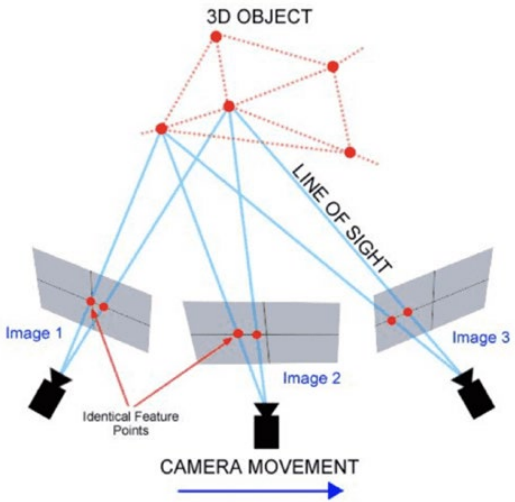
1 Scan-based Simulators —static 3D reconstructions

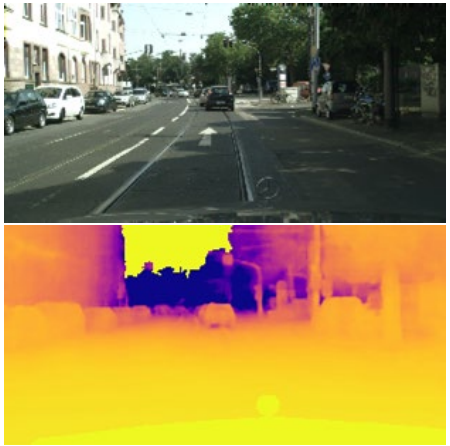


Szot et al. Habitat 2.0: Training Home Assistants to Rearrange their Habitat. NeurIPS 2021.

Evaluation: Simulators

1 Scan-based Simulators —static 3D reconstructions



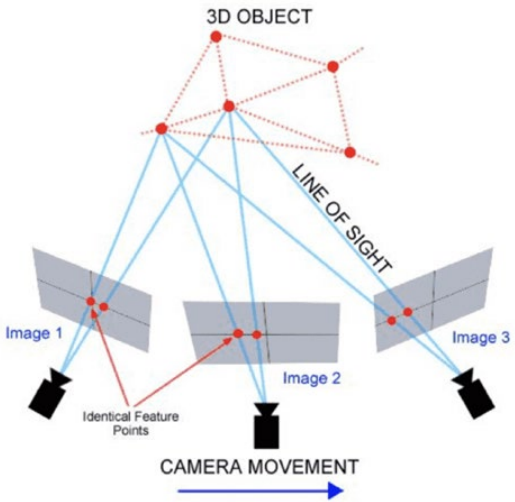


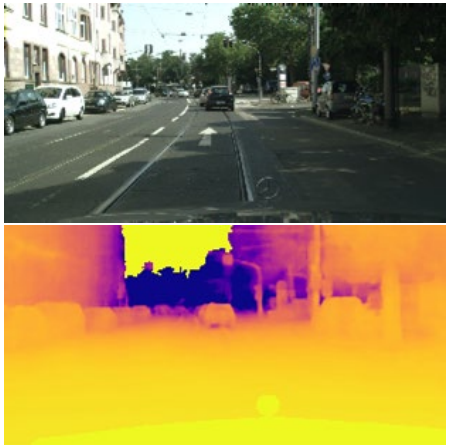
RGB + Depth

Szot et al. Habitat 2.0: Training Home Assistants to Rearrange their Habitat. NeurIPS 2021.

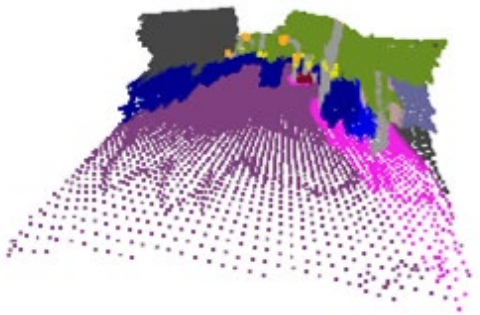
Evaluation: Simulators

1 Scan-based Simulators — static 3D reconstructions





RGB + Depth



3D point cloud

Szot et al. Habitat 2.0: Training Home Assistants to Rearrange their Habitat. NeurIPS 2021.

Evaluation: Simulators

1 Scan-based Simulators — static 3D reconstructions

Import this point cloud into a **graphics engine** – draws you the scene from any new camera viewpoint

Szot et al. Habitat 2.0: Training Home Assistants to Rearrange their Habitat. NeurIPS 2021.

Evaluation: Simulators

2 CAD-based Simulators —hand-crafted 3D models (CAD = Computer-Aided Design)

- **3D designers** are needed
- textures / set lighting takes **human labor**
- computer graphics are more **complex**
- **GAP** in physics

Szot et al. Habitat 2.0: Training Home Assistants to Rearrange their Habitat. NeurIPS 2021.

Evaluation: Simulators

1 **Scan-based** env:

- → relatively fast to collect
- → realistic look
- → very fast to render
- → limited: frozen point clouds

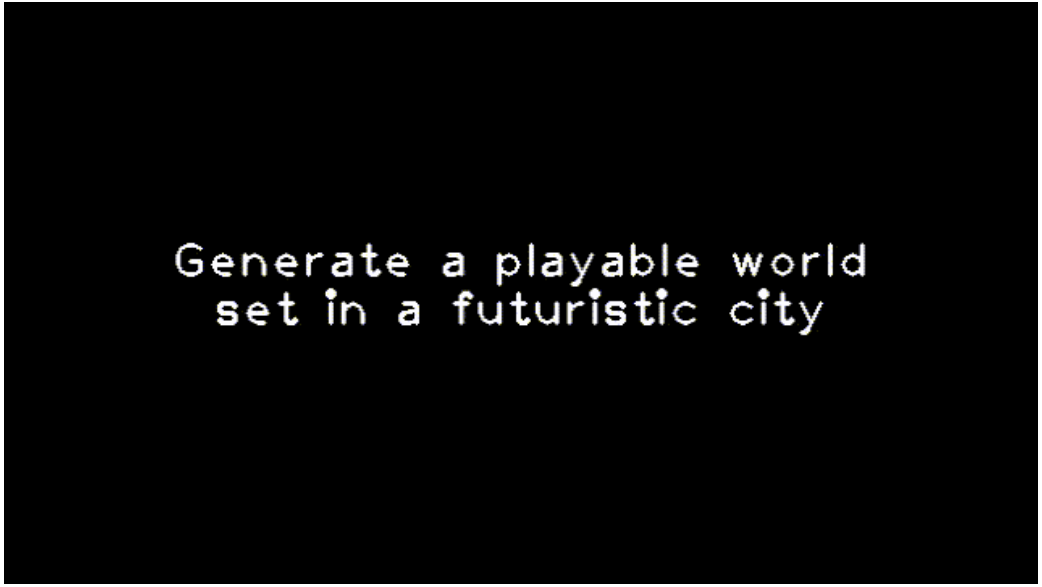
2 **CAD-based** env:

- → hard to prepare
- → realistic look requires a lot of effort
- → may be challenging to render
- → all tasks are supported

Simulators: World Model

World model is a special case of simulators. Instead of painstakingly modeling every object and light ray, the world model **learns directly from data**

Action???

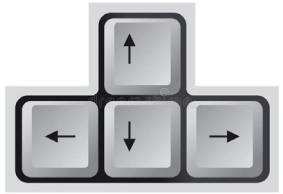


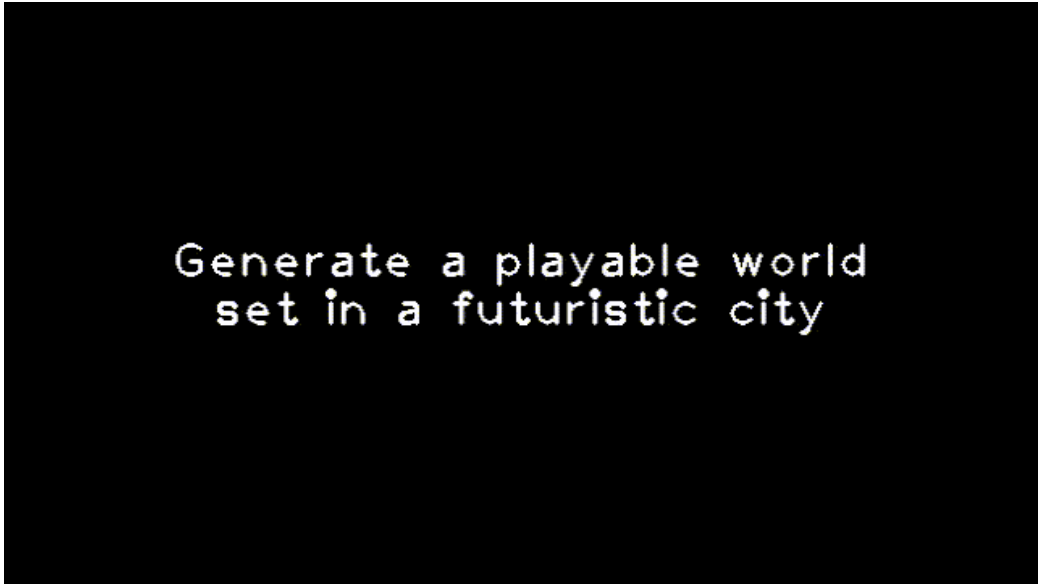
Bruce et al. Genie: Generative Interactive Environments. arXiv:2402.15391

Simulators: World Model

World model is a special case of simulators. Instead of painstakingly modeling every object and light ray, the world model **learns directly from data**

Action





Bruce et al. Genie: Generative Interactive Environments. arXiv:2402.15391

Simulators: World Model

- **Speed** – real-time 3D rendering plus physics is slow
- **Flexibility** – don't need CAD files of every object in every room
- **Data-driven** – world models adapt to whatever domain you train them on

world model learns directly from data

Simulators: World Model

Limited application – you already have massive recorded interactions (gameplay video, driving logs, lab trials)

Bruce et al. Genie: Generative Interactive Environments. arXiv:2402.15391

Evaluation: Real World

→ the only evaluation that really matters

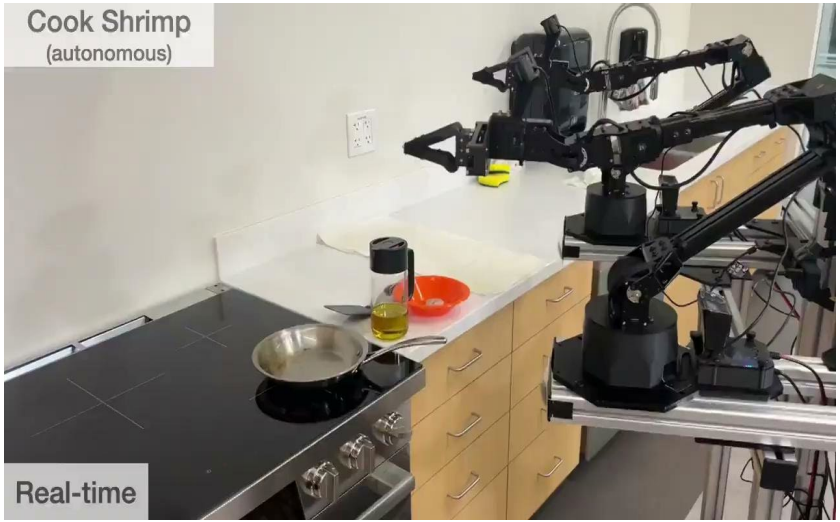
→ very slow (**initialization** of the scene)

→ very expensive and technically complex

→ scales badly (if **different kitchen** ?)

Fu et al. Mobile ALOHA: Learning Bimanual Mobile Manipulation with Low -Cost Whole -Body Teleoperation. arXiv:2401.02117

Evaluation: Real World



Fu et al. Mobile ALOHA: Learning Bimanual Mobile Manipulation with Low -Cost Whole -Body Teleoperation. arXiv:2401.02117

Evaluation: Real World

Fu et al. Mobile ALOHA: Learning Bimanual Mobile Manipulation with Low -Cost Whole -Body Teleoperation. arXiv:2401.02117

Evaluation: Comparison

Cheap benchmarks, but remote proxy metrics; **relevant** benchmarks, but complex and time consuming

	Static benchmarks	Arena	End-to-end sim	End-to-end real
Relevance	Low	Medium	Medium	High
Safety	High	High	High	Low
Speed	High	Medium	Medium	Low
Cheapness	High	Medium	Medium	Low
Reproducibility	High	High	Medium	Low

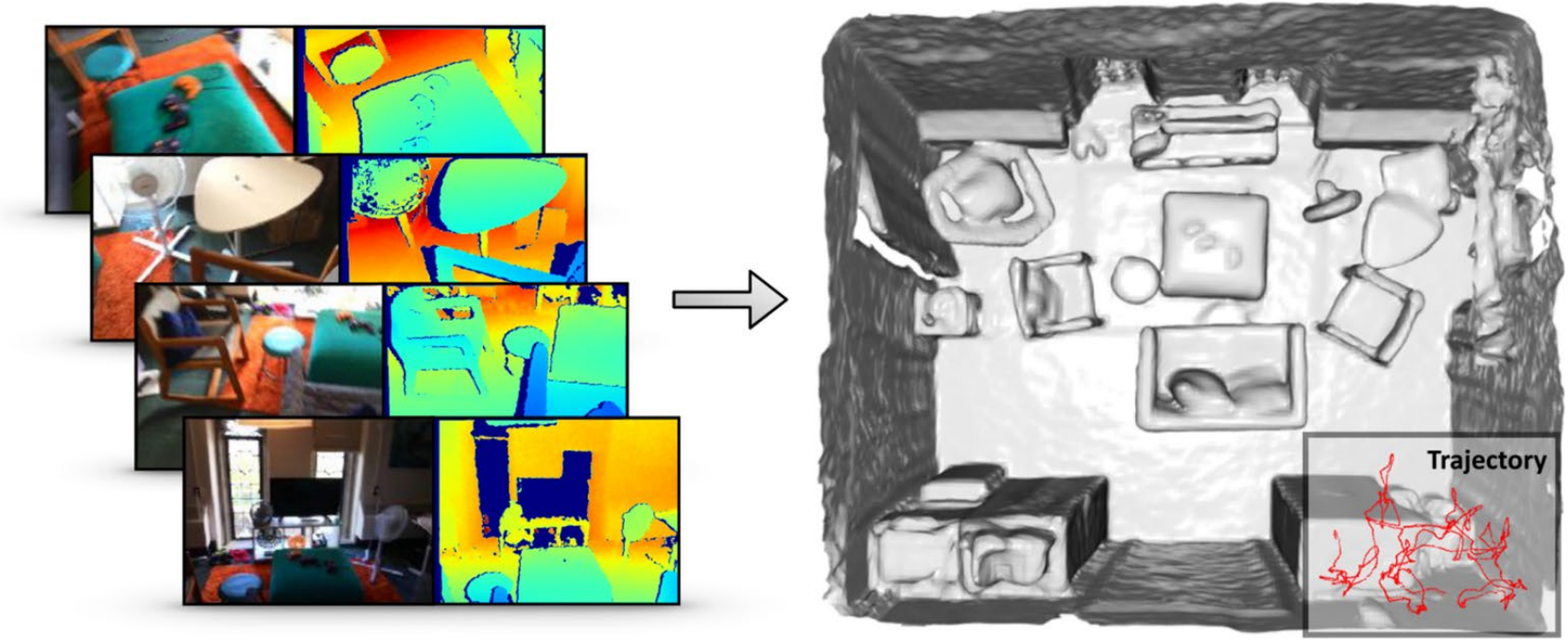
3

Understanding

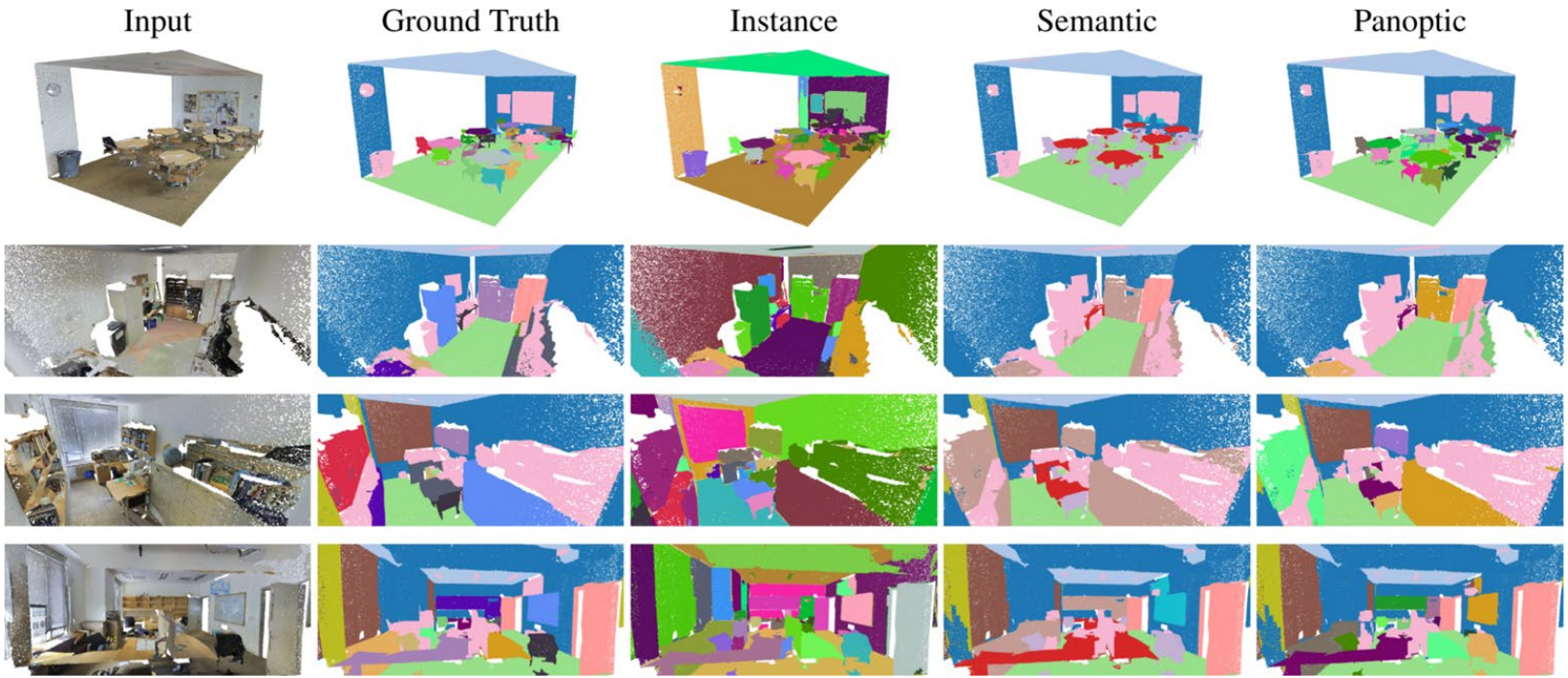
Understanding the World.

Embodied Question Answering

Understanding: 3D Reconstruction



Understanding: 3D Segmentation

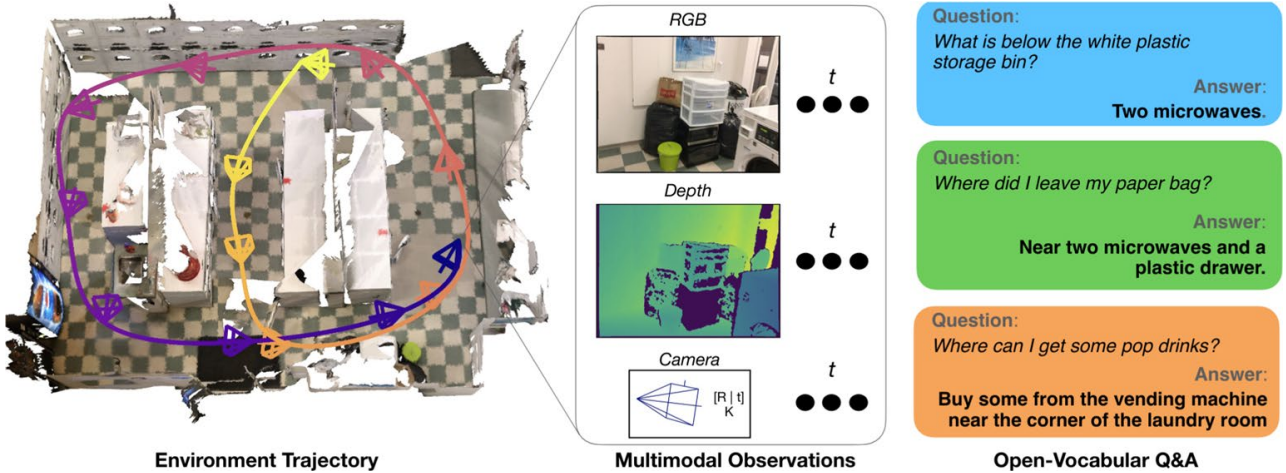


Kolodiazhnyi et al. OneFormer3D: One Transformer for Unified Point Cloud Segmentation. CVPR 2024

Understanding: Embodied QA

Evaluation of answers is done with GPT -4 or via human evaluation

→ **pre-recorded** video stream
→ fully interactive mode: agent can **freely explore** environment

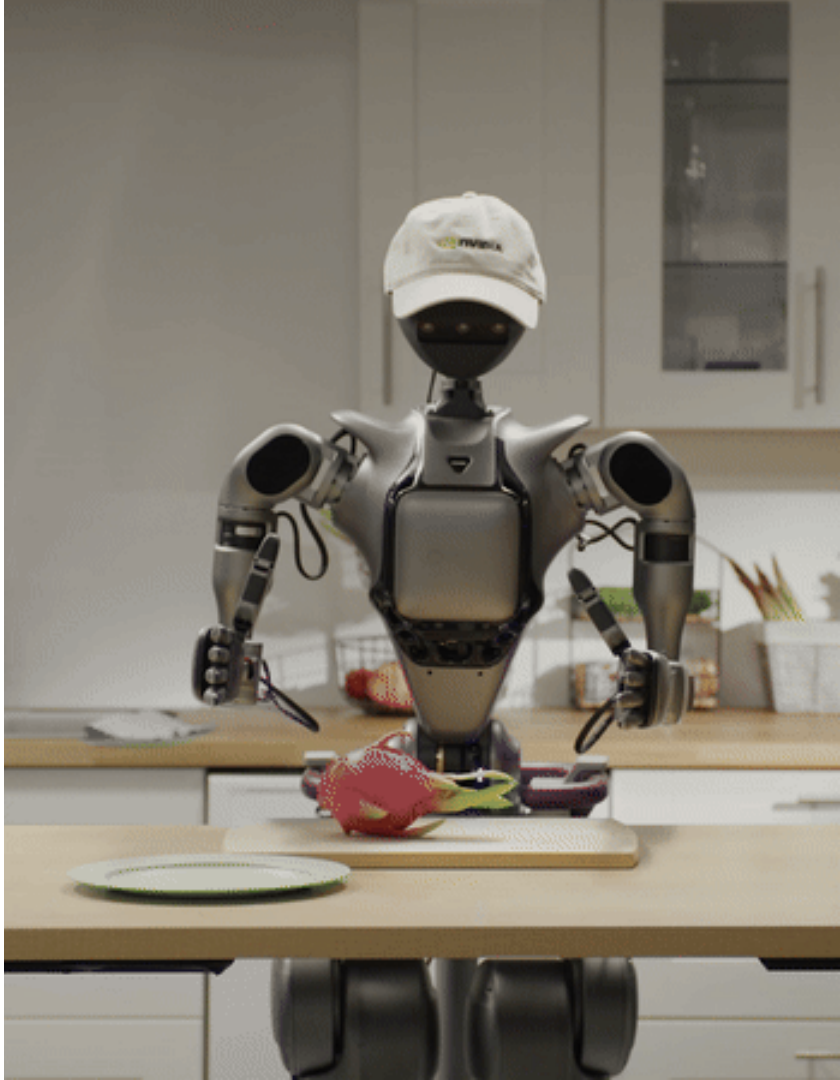


Majumdar et al. OpenEQA: Embodied Question Answering in the Era of Foundation Models. CVPR 2024.

4

Planning

Open-world planning.
Dividing instruction in small substacks

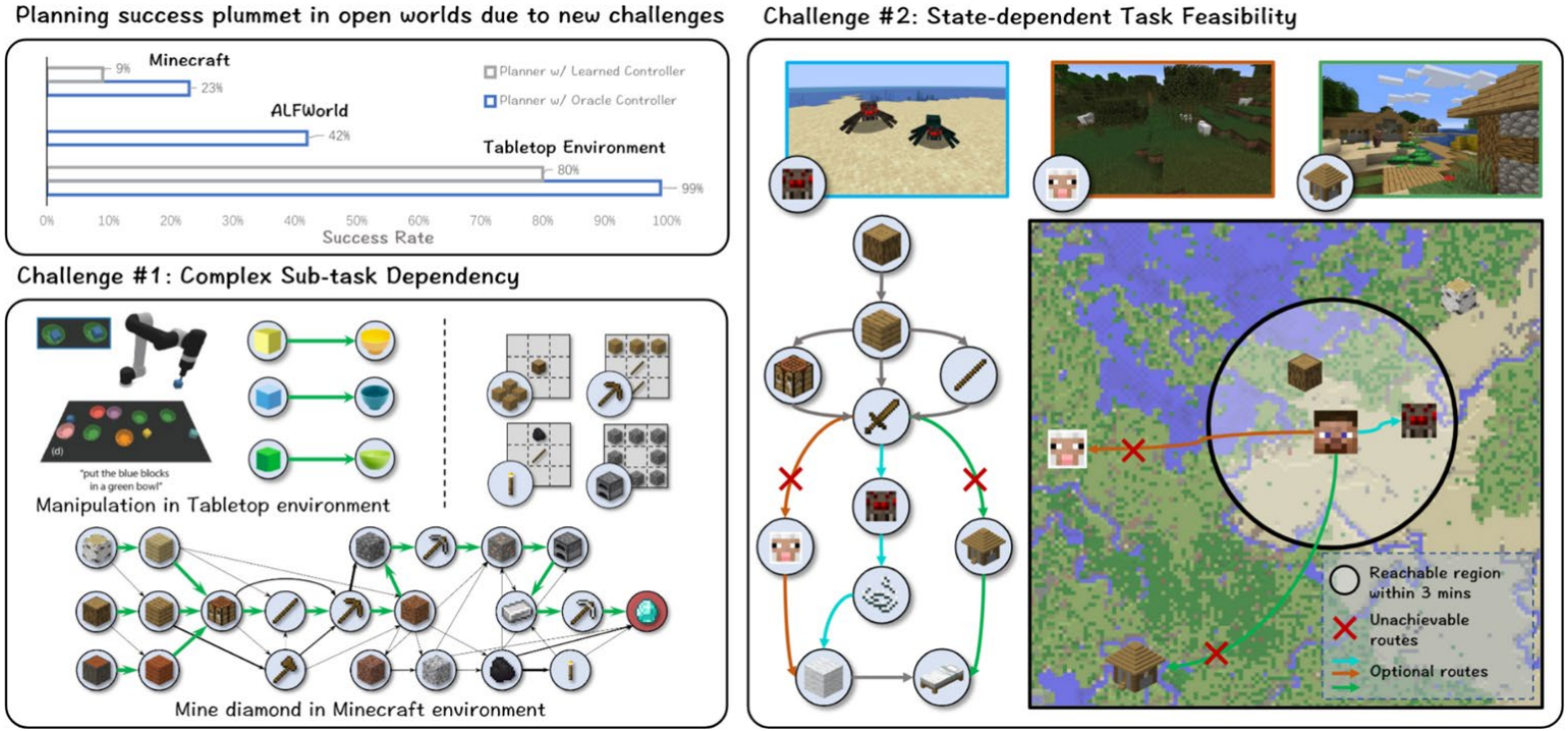


Planning: Embodied QA

Planning is **decomposing** high-level task into **sequence of sub-tasks**

Zhang et al. MFE-ETP: A Comprehensive Evaluation Benchmark for Multi-modal Foundation Models on Embodied Task Planning. arXiv:2407.05047

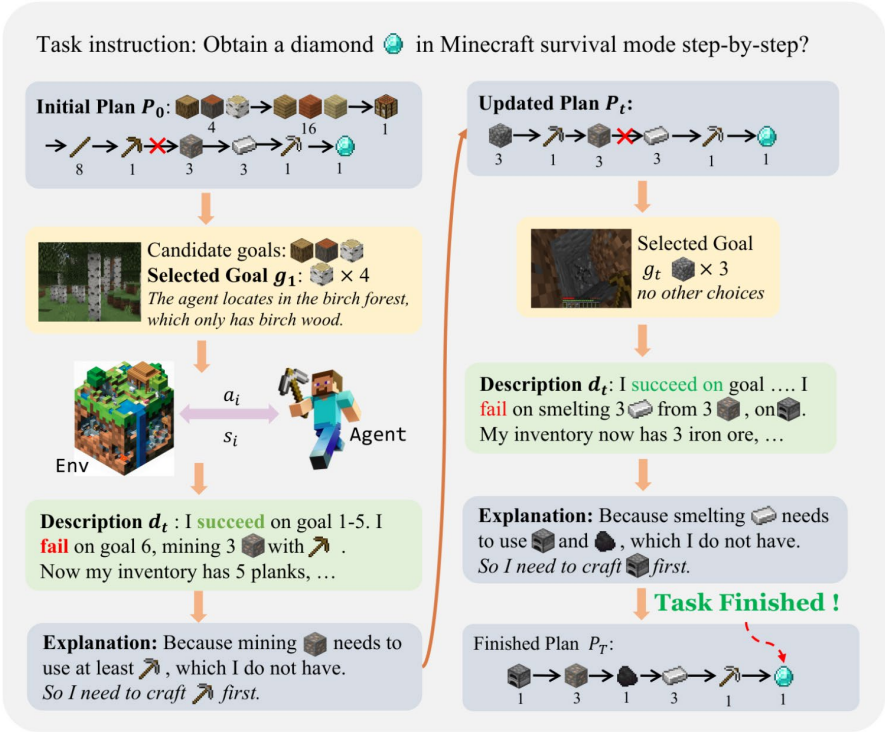
Planning: Open World Planning



Wang et al. Describe, Explain, Plan and Select: Interactive Planning with Large Language Models Enables Open -World Multi -Task Ag ents. NeurIPS 2023

Planning: Open World Planning

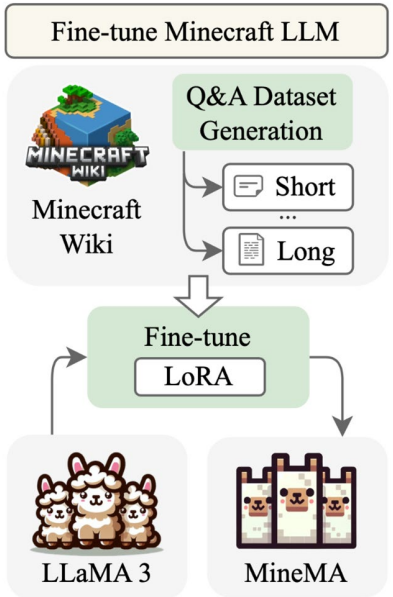
- we **can't benchmark** planning on fixed ground truth sequences
- running in simulator and measuring **success rate (SR)** is the only option
- the planning module has to be adaptable and be able to **modify the plan**



Wang et al. Describe, Explain, Plan and Select: Interactive Planning with Large Language Models Enables Open -World Multi -Task Ag ents. NeurIPS 2023

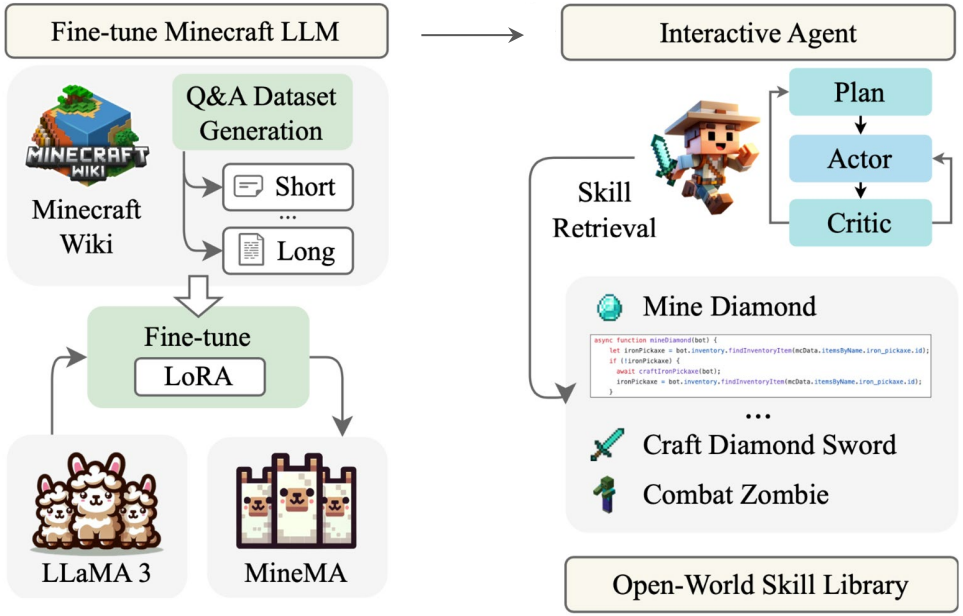
Planning: Odyssey

Odyssey —new framework that empowers **Large Language Model (LLM)** **-based agents** with open-world skills to explore the vast Minecraft world



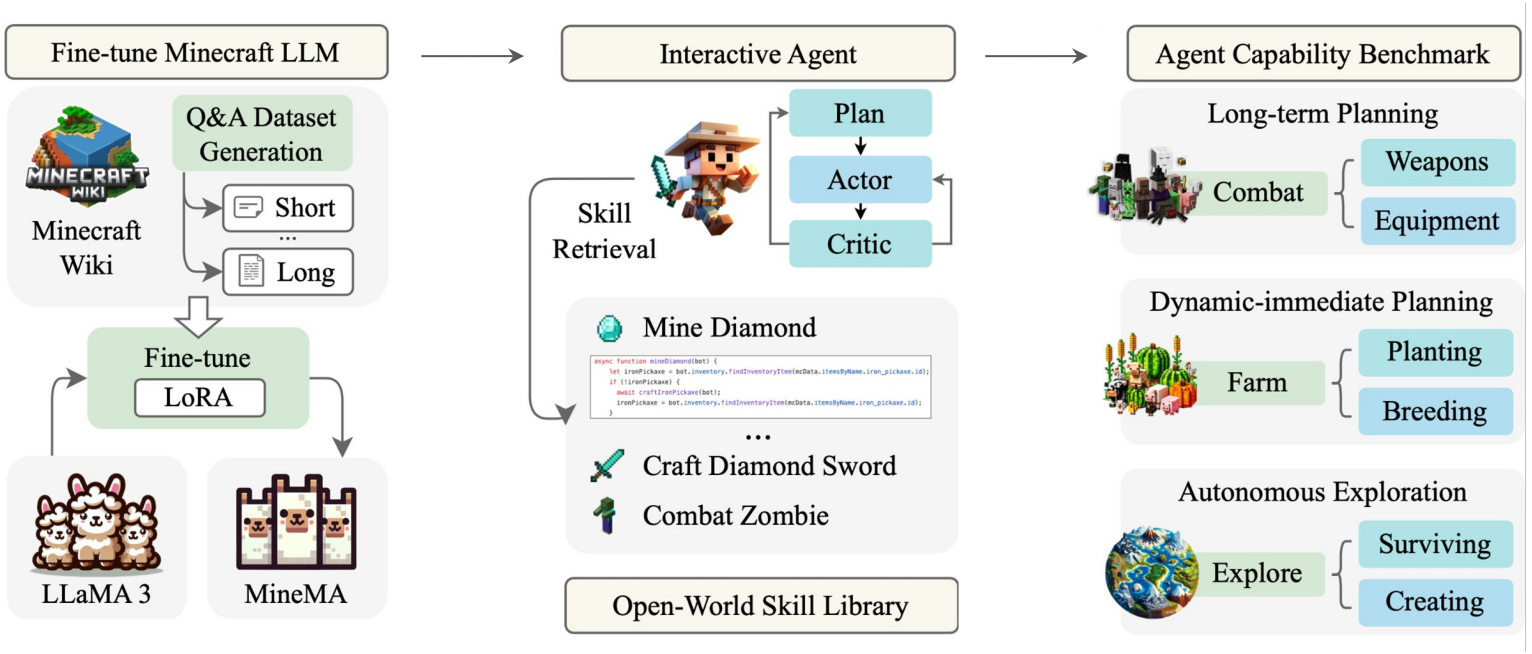
Planning: Odyssey

Odyssey —new framework that empowers **Large Language Model (LLM) -based agents** with open-world skills to explore the vast Minecraft world



Planning: Odyssey

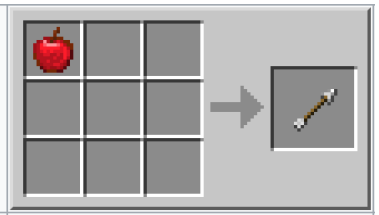
Odyssey —new framework that empowers **Large Language Model (LLM) -based agents** with open-world skills to explore the vast Minecraft world

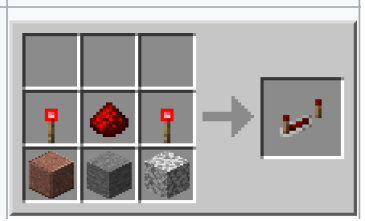


Odyssey: Minecraft Wiki

To improve agent performance in Minecraft, we fine-tune the **LLaMA-3 model** using a large-scale Q&A dataset with **390k+ instruction entries** sourced from the **Minecraft Wiki**

Crafting:

Apple	
Block of Gold + Apple	

Stone Bricks	Any Stone	
Redstone Repeater	Redstone Torch + Redstone Dust + Any Stone	

Odyssey: Minecraft Wiki

To improve agent performance in Minecraft, we fine-tune the **LLaMA-3 model** using a large-scale Q&A dataset with **390k+ instruction entries** sourced from the **Minecraft Wiki**

Passive mobs:

Hostile mobs:

Odyssey: Interactive Agent

Efficient **retrieval of skills** is provided by generating a description for each skill by calling the LLM – **Sentence Transformer** to encode each skill

collectItem.js

```
if (!mob) {
    bot.chat("Could not find a mob.");
    return false;
}
// kill the mob using the sword
await equipBestTool(bot, tool);
await killMob(bot, mob.name, 300);
// collect the dropped items
await bot.pathfinder.goto(new GoalBlock(mob.position.x,
mob.position.y, mob.position.z));
bot.chat("Collected dropped items.");
```

LLM-based agent employs a **planner - actor - critic** architecture to define which actions to do

- 40 primitive skills
- 183 compositional skills

Odyssey: Interactive Agent

1

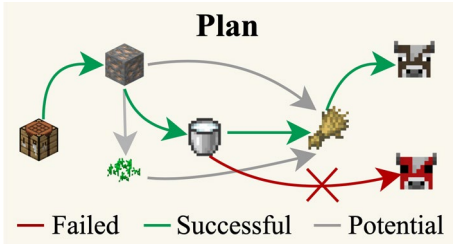
LLM Planner — breaks down high-level goals into specific low-level subgoals

a) **Ultimate goal** = I want to breed cow and collect items from it.

b) **State of the agent**

[**Position**]: x=2134.5, y=69.0, z=769.5
[**Time**] day
[**Nearby bocks**] dirt, grass, oak_log
[**Nearby entities**] horse, pig

c) **Achievements**



Odyssey: Interactive Agent

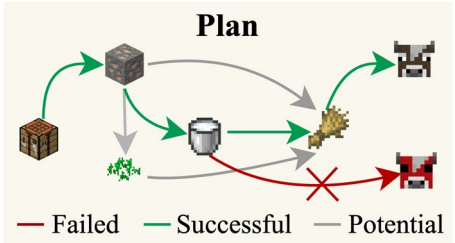
2

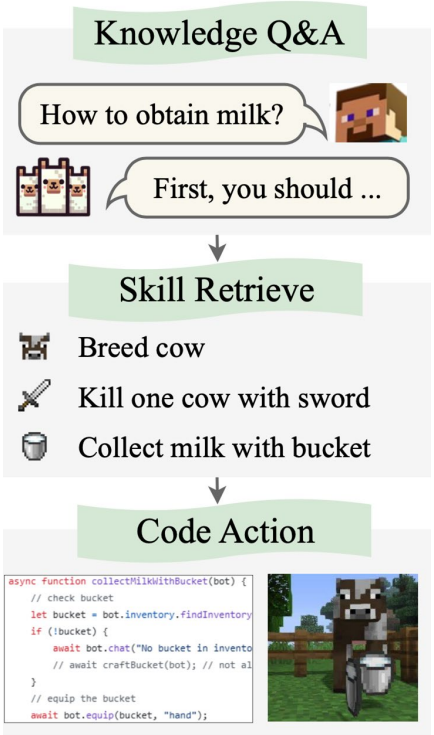
LLM Actor — invoked to sequentially execute the subgoals generated by the LLM planner within the Minecraft environment

a) Query context

b) Similarity matching

c) Skill Selection

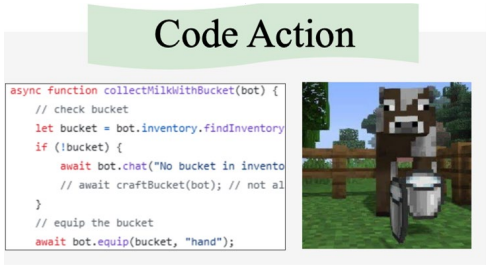


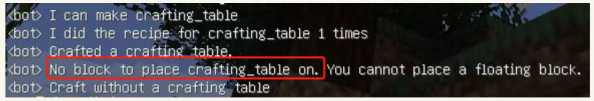


Odyssey: Interactive Agent

3 **LLM Critic** — noting successful outcomes and failure points, **refine strategy**

a) Execution Feedback
b) Self-validation
c) Self-reflection





Execution Feedback
[Lack of pre-requirements]
I cannot collect milk without a .
[Environment feedback]
I could not find a to collect milk.

Self-validation:

Observation
My subgoal is to:
collect milk
last_inventory (16/36): ...
cur_inventory (18/36): ...

Thought
Based on changes of my inventory, is my subgoal successful?

Self-reflection:

Rethink
You should analysis the reason why my subgoal is failed based on the logs provided.

Critic
Since you only have , you might need the to attract a for milk.

Odyssey: Examples

- Use **GPT-3.5 and GPT-4** for initial **data** generation
- All experiments are conducted with the open-source **LLaMA-3 model**
- Simulation environment is built on top of **Voyager**

Shear a Sheep

Odyssey: Examples

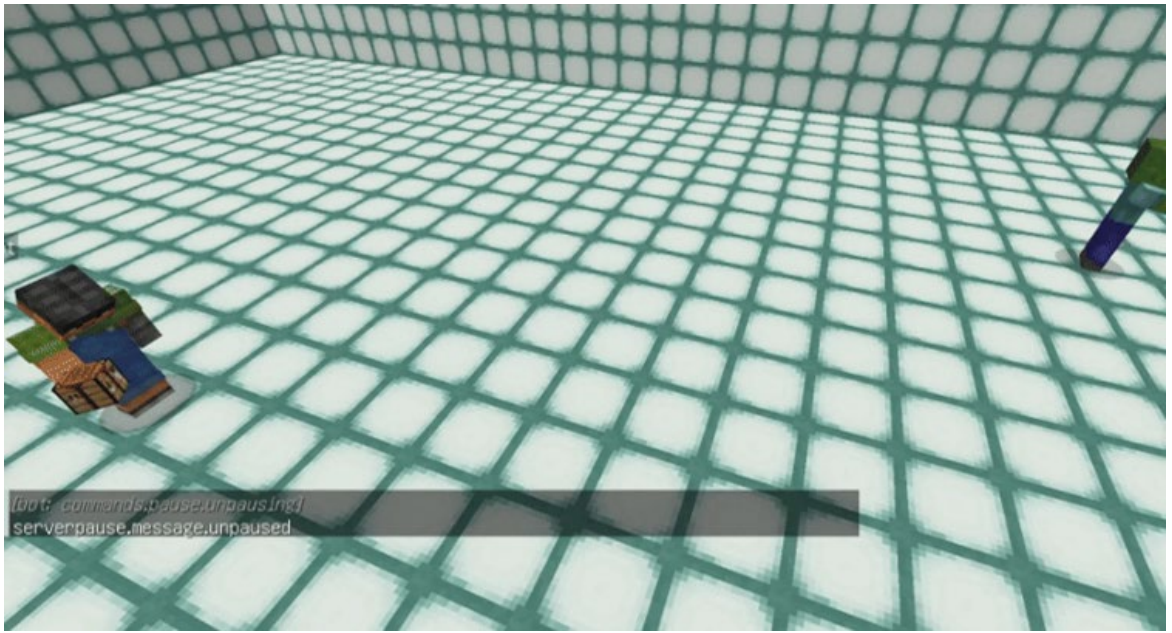
- Use **GPT-3.5 and GPT-4** for initial **data** generation
- All experiments are conducted with the open-source **LLaMA-3 model**
- Simulation environment is built on top of **Voyager**

Mining diamonds from scratch

Odyssey: Examples

- Use **GPT-3.5 and GPT-4** for initial **data** generation
- All experiments are conducted with the open-source **LLaMA-3 model**
- Simulation environment is built on top of **Voyager**

Craft sword and Combat a zombie



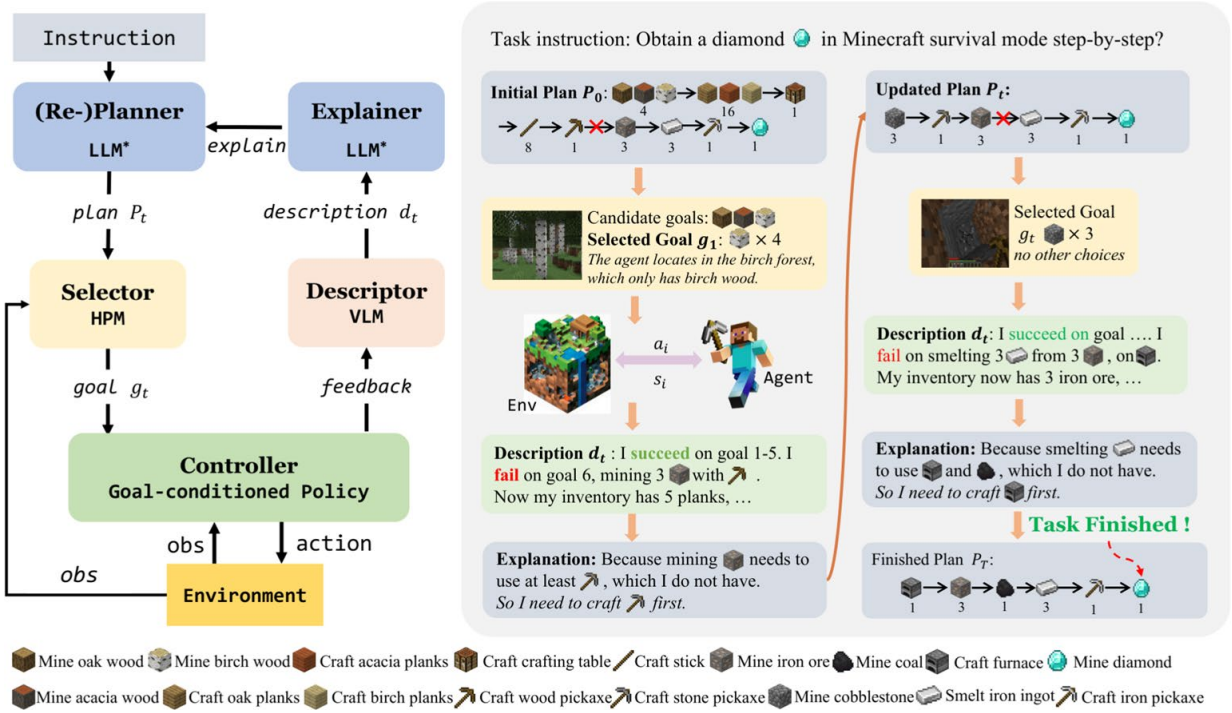
Odyssey: Examples

- Use **GPT-3.5 and GPT-4** for initial **data** generation
- All experiments are conducted with the open-source **LLaMA-3 model**
- Simulation environment is built on top of **Voyager**

Task	Time (min)	2min	5min	10min	15min
	0.59 ± 0.79	95.8%	99.2%	100.0%	100.0%
	0.95 ± 0.80	92.5%	99.2%	100.0%	100.0%
	1.48 ± 0.96	85.0%	97.5%	100.0%	100.0%
	4.43 ± 1.48	0.0%	76.7%	100.0%	100.0%
	6.48 ± 2.02	0.0%	21.7%	92.5%	100.0%

Planning: DEPS

Dynamic error recovery: DEPS doesn't rely on a rigid skill library; it introspects and corrects mistakes in realtime



Planning: DEPS

Meta	Name	Number	Example Task	Max. Steps	Initial Inventory	Given Tool
MT1	Basic	14	Make a wooden door.	3000	Empty	Axe
MT2	Tool (Simple)	12	Make a stone pickaxe.	3000	Empty	Axe
MT3	Hunt and Food	7	Cook the beef.	6000	Empty	Axe
MT4	Dig-Down	6	Mine coal.	3000	Empty	Axe
MT5	Equipment	9	Equip the leather helmet.	6000	Empty	Axe
MT6	Tool (Complex)	7	Make shears and bucket.	6000	Empty	Axe
MT7	IronStage	13	Obtain an iron sword.	6000	Empty	Axe
MT8	Challenge	1	Obtain a diamond!	12000	Empty	Axe

Methods	MT1	MT2	MT3	MT4	MT5	MT6	MT7	MT8	AVG
GPT[16, 32]	25.85±24.8	47.88±31.5	10.78±14.6	7.14±9.0	1.98±5.9	0.0±0.0	0.0±0.0	0.0±0.0	15.42
PP[42]	30.61±23.6	40.09±30.6	17.13±19.1	16.00±17.3	3.21±4.9	0.47±1.3	0.60±2.2	0.0±0.0	16.88
CoT[45]	40.24±30.8	55.21±26.8	6.82±11.6	4.76±8.2	1.73±5.2	0.0±0.0	0.0±0.0	0.0±0.0	18.89
IM[17]	46.89±31.4	53.73±20.8	3.64±6.9	18.41±17.4	4.57±7.4	0.64±1.7	1.02±2.5	0.0±0.0	21.64
CaP[20]	60.08±17.3	60.11±20.24	8.72±9.7	20.33± 21.0	2.84±4.6	0.63±1.3	0.60±2.2	0.0±0.0	25.77
DEP	75.70±10.4	66.13±13.4	45.69±16.2	43.35±20.2	15.93±13.9	5.71±3.7	4.60±7.1	0.50±0.5	**39.36**
DEPS	79.77±8.5	79.46±10.6	62.40±17.9	53.32±29.3	29.24±27.3	13.80±8.0	12.56±13.3	0.59±0.5	**48.56**

Planning: DEPS

Examples of planning by the agent

5.1

Acting: Manipulation

Manipulation task.

Collecting real data.

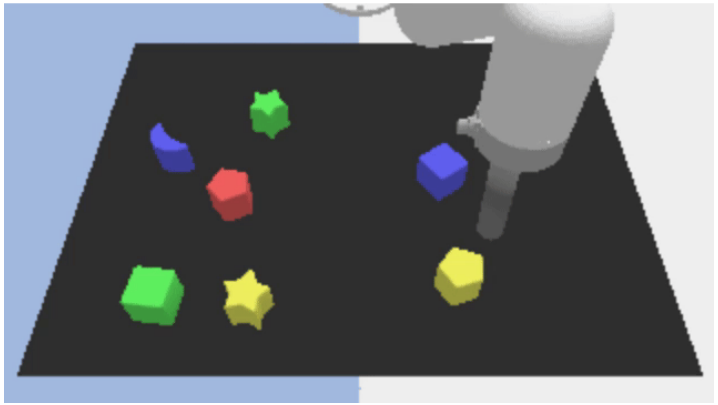
Policies

Manipultaion: Action

Now given **language instruction** and **observations** from sensors, output **action / sequence of actions**. Modern notable approaches are based on:

- → hybrid models
- → diffusion models
- → transformers
- → VLLMs

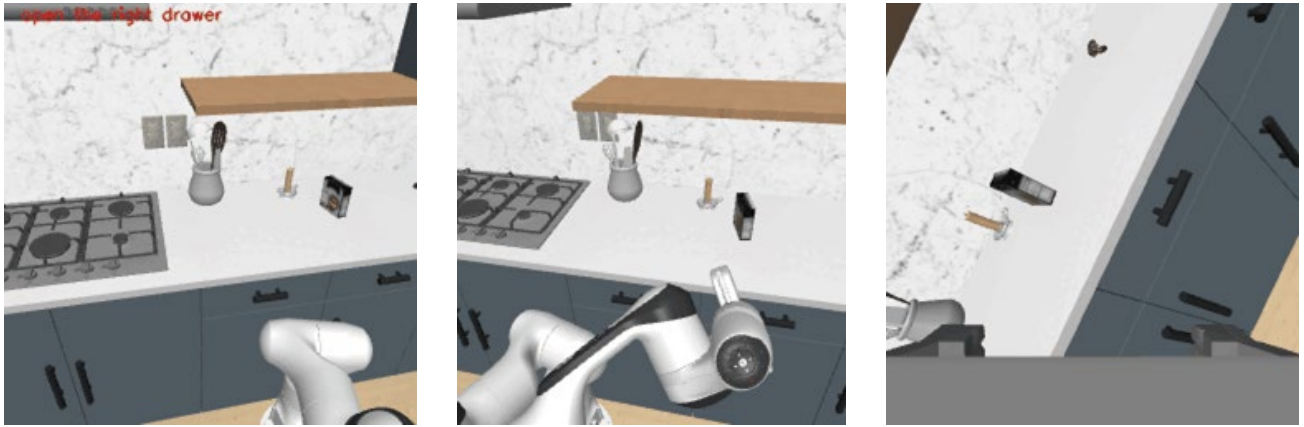
instruction: slide the green star next to the red moon



Manipulation: Action

Now given **language instruction** and **observations** from sensors, output **action / sequence of actions**. Modern notable approaches are based on:

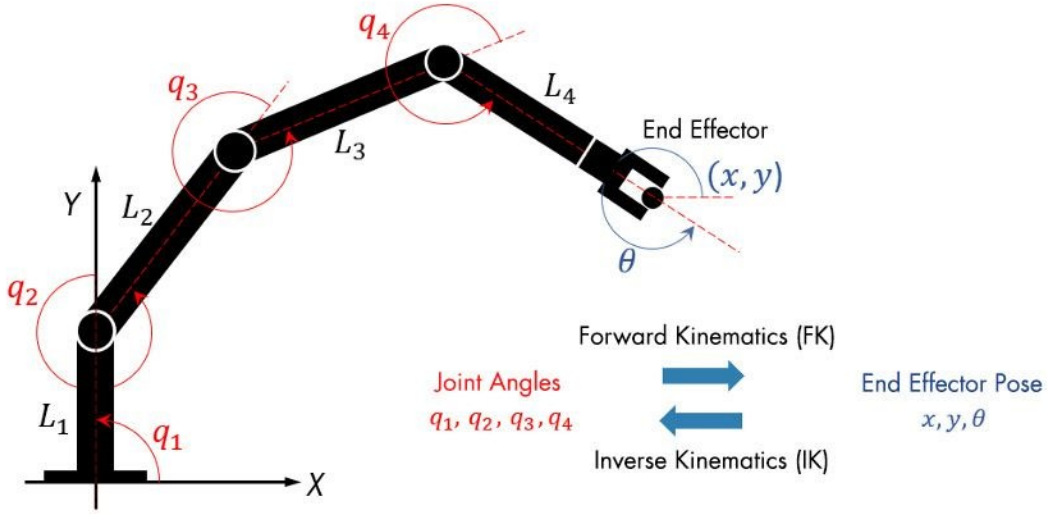
- → hybrid models
- → diffusion models
- → transformers
- → VLLMs



view from three cameras

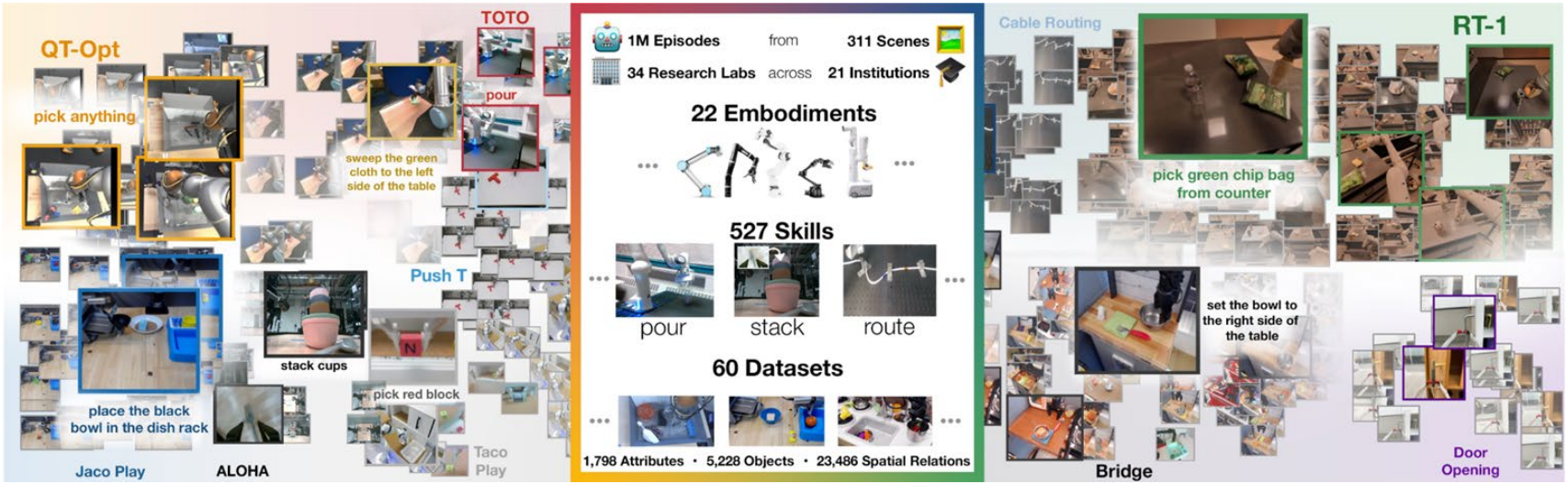
Manipulation: Inverse Kinematics

How the rest of the joints move is the **inverse kinematics** that gives out these joints



Manipulation: Open X-Embodiment

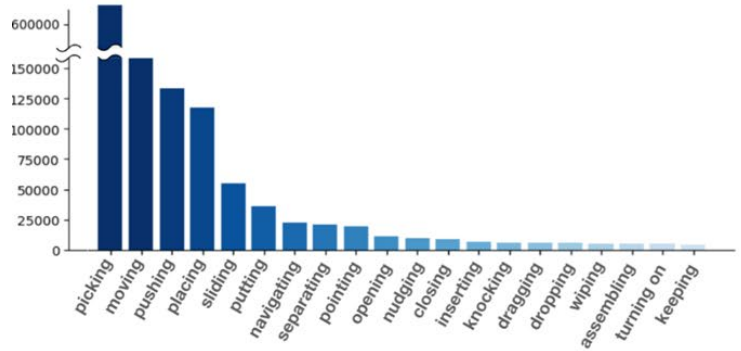
ImageNet = **1000** classes and pictures; in **Robotics** = if **1000 tasks**, 1 trajectory is **250 samples** (250 pictures is 1 sample = pictures + coordinates)



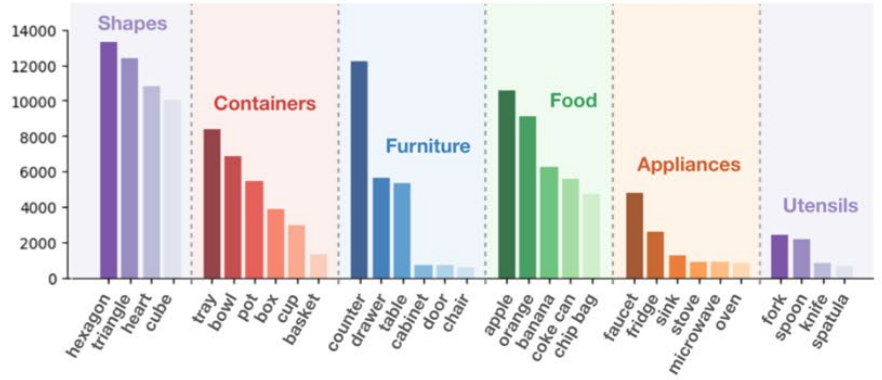
Open X-Embodiment Collaboration. Open X-Embodiment: Robotic Learning Datasets and RT-X Models. arXiv:2310.08864

Manipulation: Open X-Embodiment

Breakdown by tasks — **skills** (take, move) and **tasks** (what to move)



(d) Common Dataset Skills

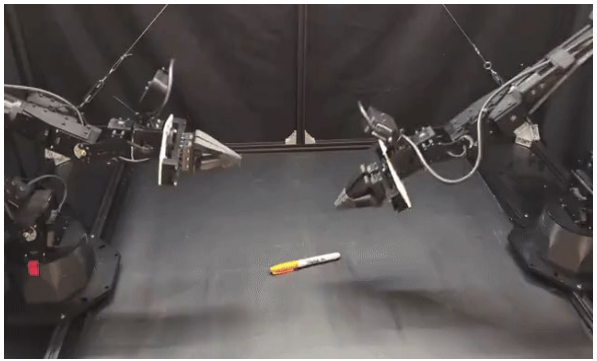


(e) Common Dataset Objects

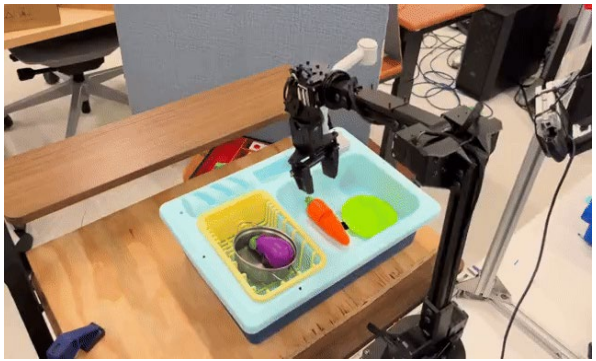
Open X-Embodiment Collaboration. Open X-Embodiment: Robotic Learning Datasets and RT-X Models. arXiv:2310.08864

Manipulation: Open X-Embodiment

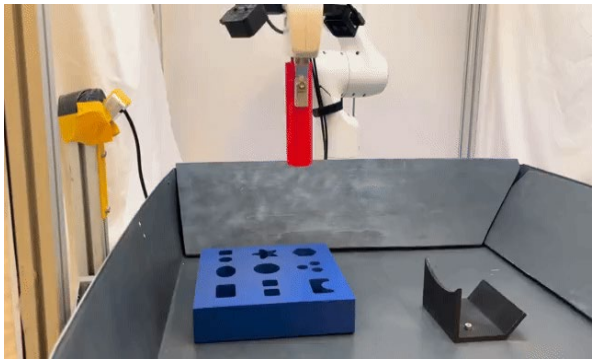
Examples of collected trajectories (22 embodiments)



dual-arm robot, Stanford
Aloha



part of the Bridge Dataset
WidowX BridgeV2



Functional manipulation benchmark
Berkley Peg insertion

Open X-Embodiment Collaboration. Open X -Embodiment: Robotic Learning Datasets and RT -X Models. arXiv:2310.08864

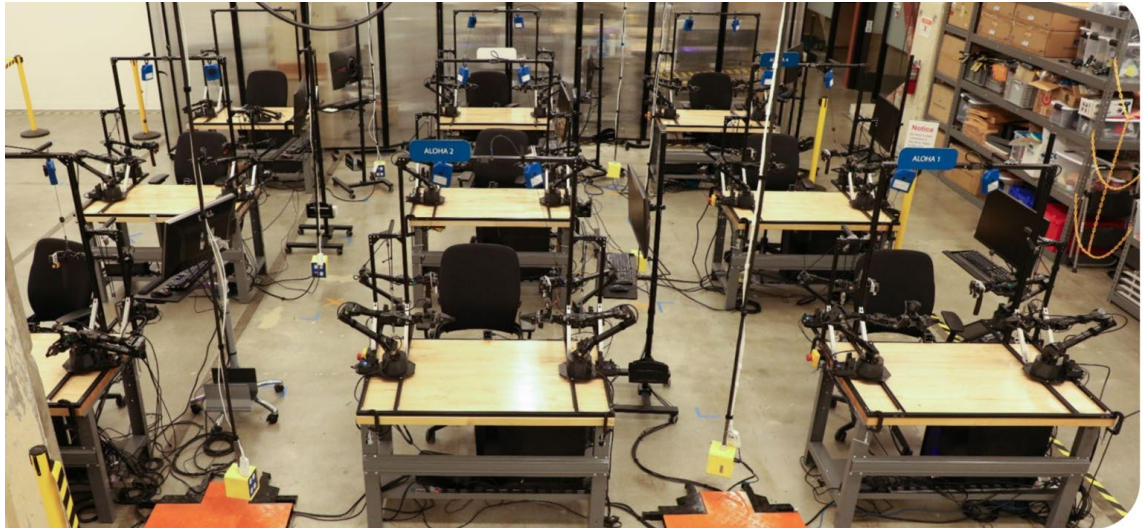
Collecting Data: ALOHA

Gathering trajectories with **ALOHA** – can be difficult task due to mirroring

Aldaco et al. ALOHA 2: An Enhanced Low -Cost Hardware for Bimanual Teleoperation

Collecting Data: ALOHA

- Data collection room, Google
- Buy a bunch of ALOHAs (3D printer)
- Put a bunch of people to collect data
- Poorly scalable process
- Installation next to the dining room :)



Aldaco et al. ALOHA 2: An Enhanced Low -Cost Hardware for Bimanual Teleoperation

Collecting Data: Teleoperation

Instead of the robot acting on its own, a person —sometimes sitting at a **console**, sometimes wearing a **VR headset** or **data-glove**

- **EgoView** (two-fingers move)
- Separate **navigation** and **manipulation**

Fu et al. Mobile ALOHA: Learning Bimanual Mobile Manipulation with Low -Cost Whole -Body Teleoperation. arXiv:2401.02117

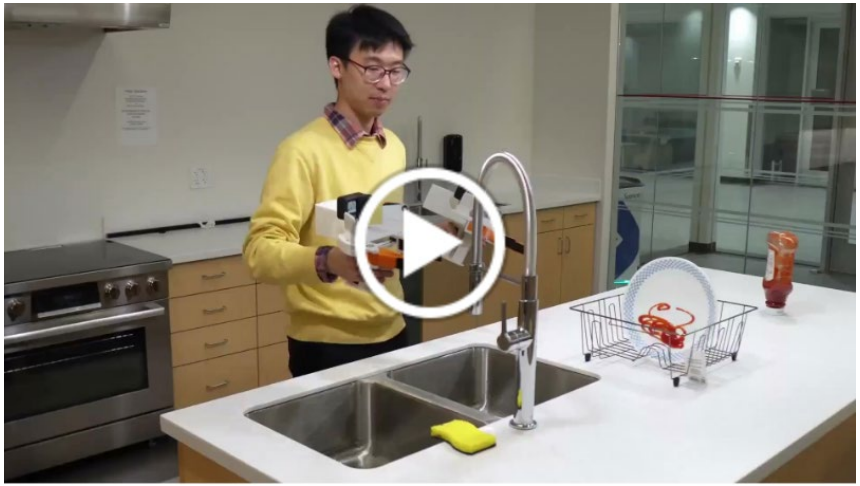
Collecting Data: UMI

3D-printed, trigger-activated parallel-jaw gripper (the "UMI gripper") in your hand, with a GoPro (or similar) mounted on

Do not need actual robot!

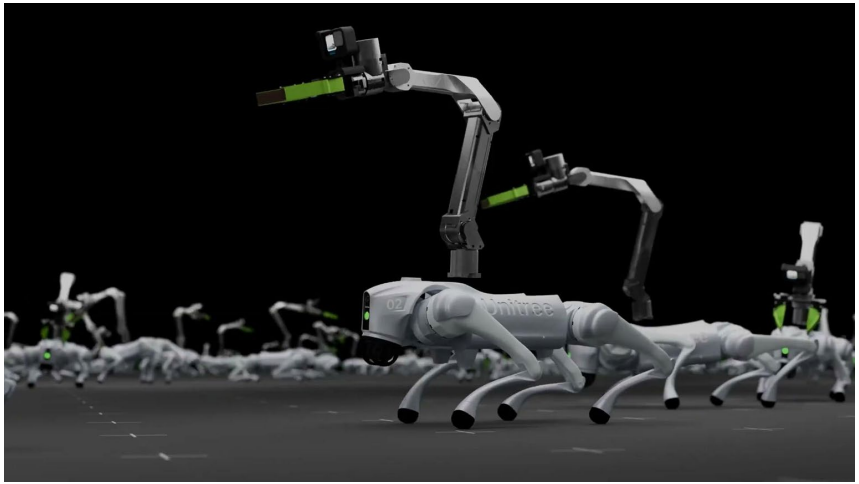
Ha et al. UMI on Legs: Making Manipulation Policies Mobile with Manipulation-Centric Whole-body Controllers. arXiv:2407.10353

Collecting Data: UMI



Ha et al. UMI on Legs: Making Manipulation Policies Mobile with Manipulation -Centric Whole -body Controllers. arXiv:2407.10353

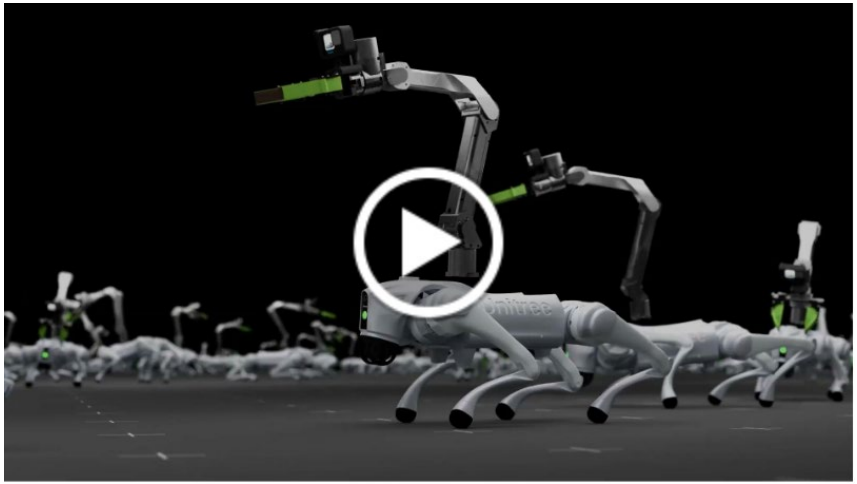
Collecting Data: UMI on Legs

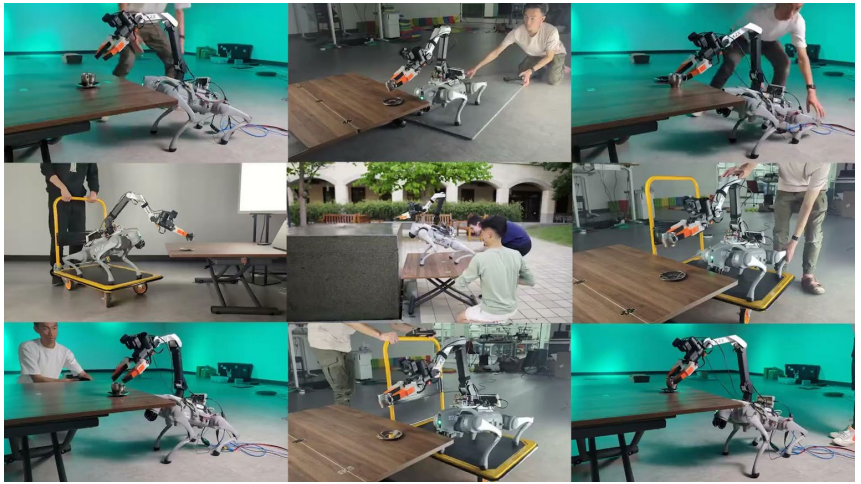


A bunch of dogs are learning how to stabilize their hand using reward in a simulation

Ha et al. UMI on Legs: Making Manipulation Policies Mobile with Manipulation -Centric Whole -body Controllers. arXiv:2407.10353

Collecting Data: UMI on Legs





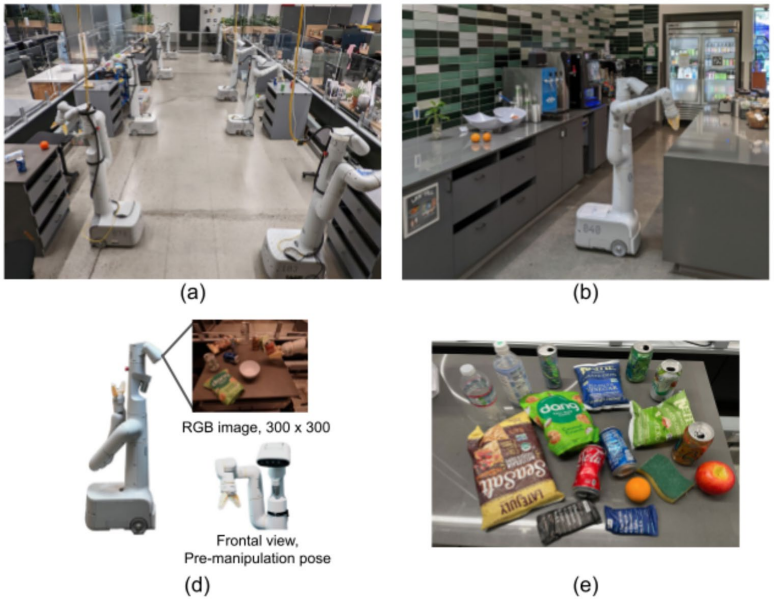
Ha et al. UMI on Legs: Making Manipulation Policies Mobile with Manipulation -Centric Whole -body Controllers. arXiv:2407.10353

Hybrid Policy: RT-1

Training data:

- 130k episodes
- 700 tasks
- collected with 13 robots over 17 months

Inference time is 100ms, overall system works at 3 Hz

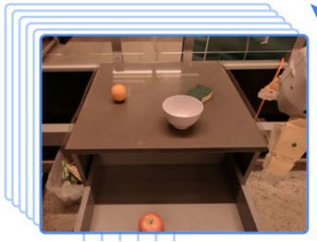


robot classroom where data was collected

Hybrid Policy: RT-1

Instruction

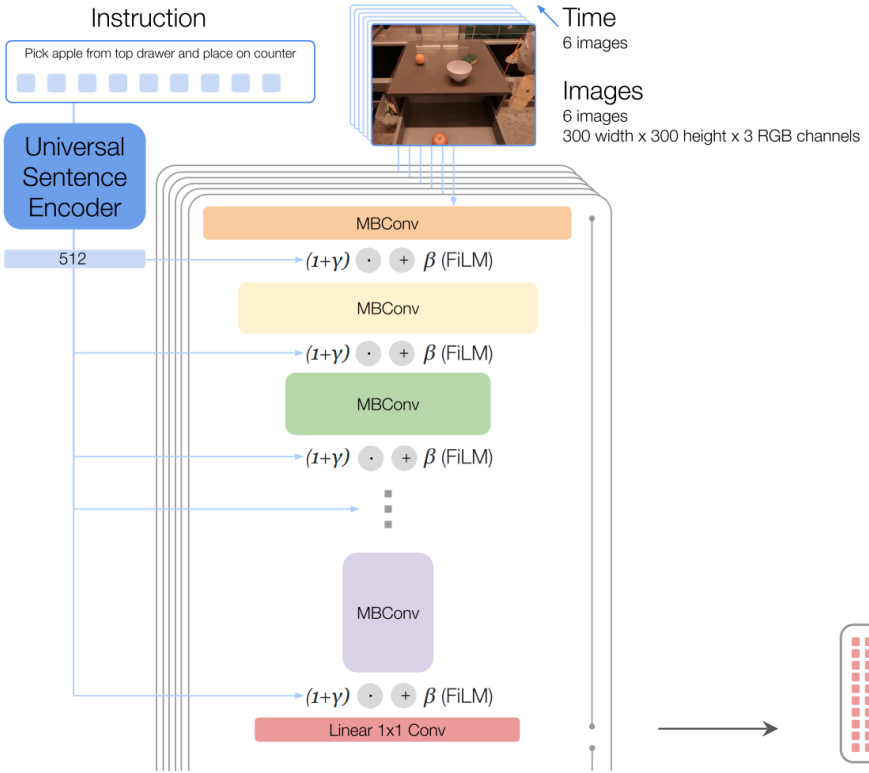
Pick apple from top drawer and place on counter



Time
6 images

Images
6 images
300 width x 300 height x 3 RGB channels

Hybrid Policy: RT-1



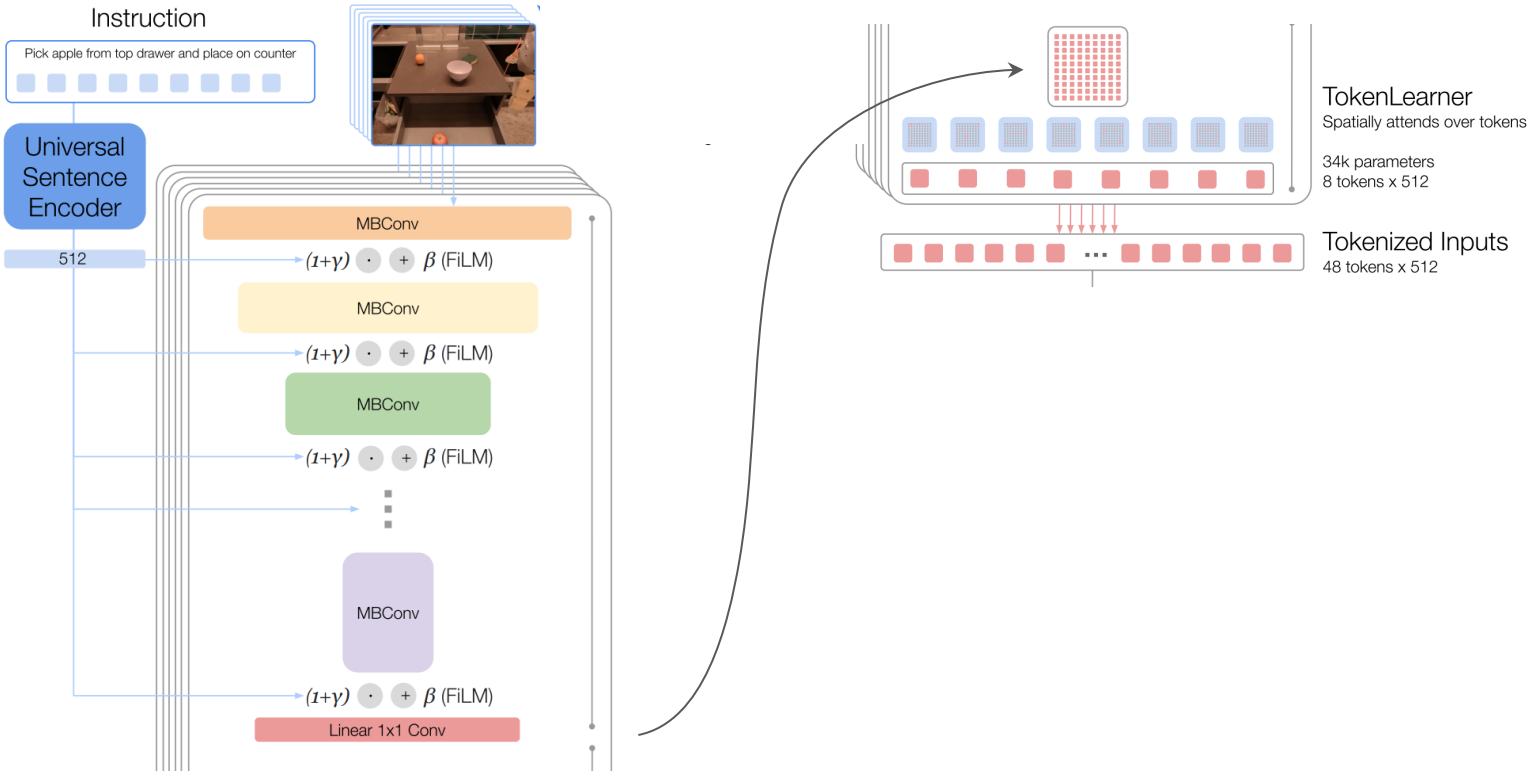
FiLM EfficientNet -B3

Fuses language instruction with image embeddings

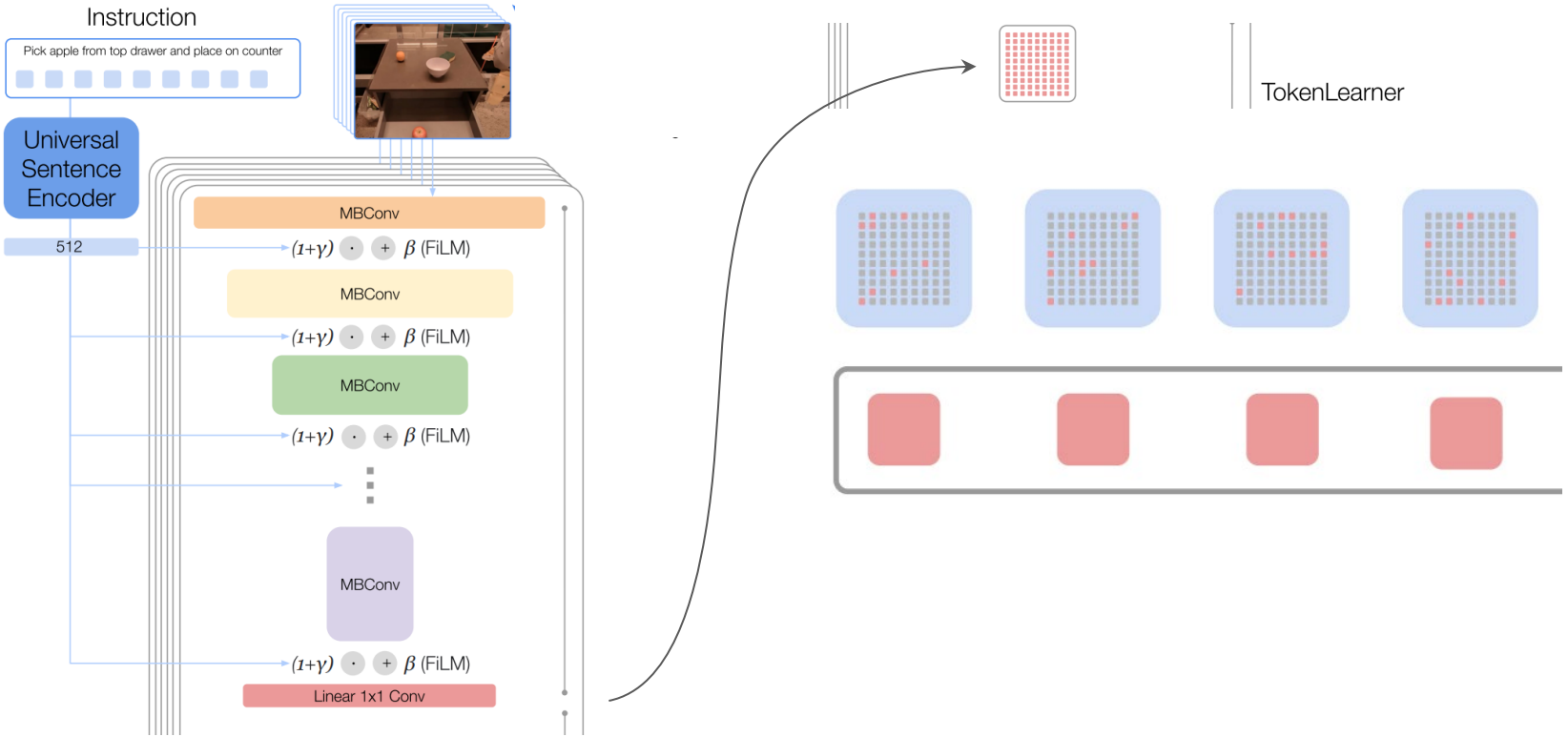
Feature map

9 x 9 x 512

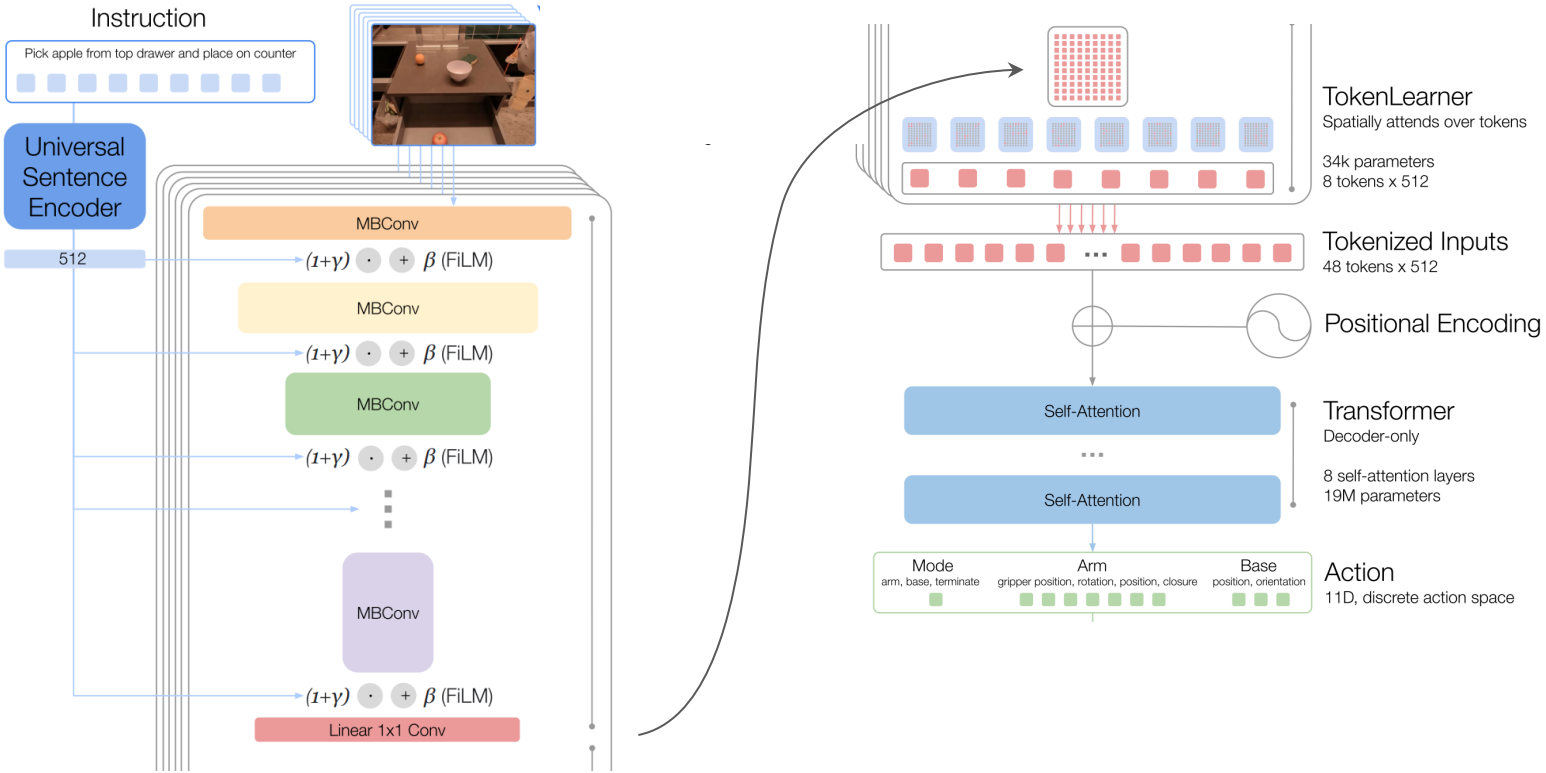
Hybrid Policy: RT-1



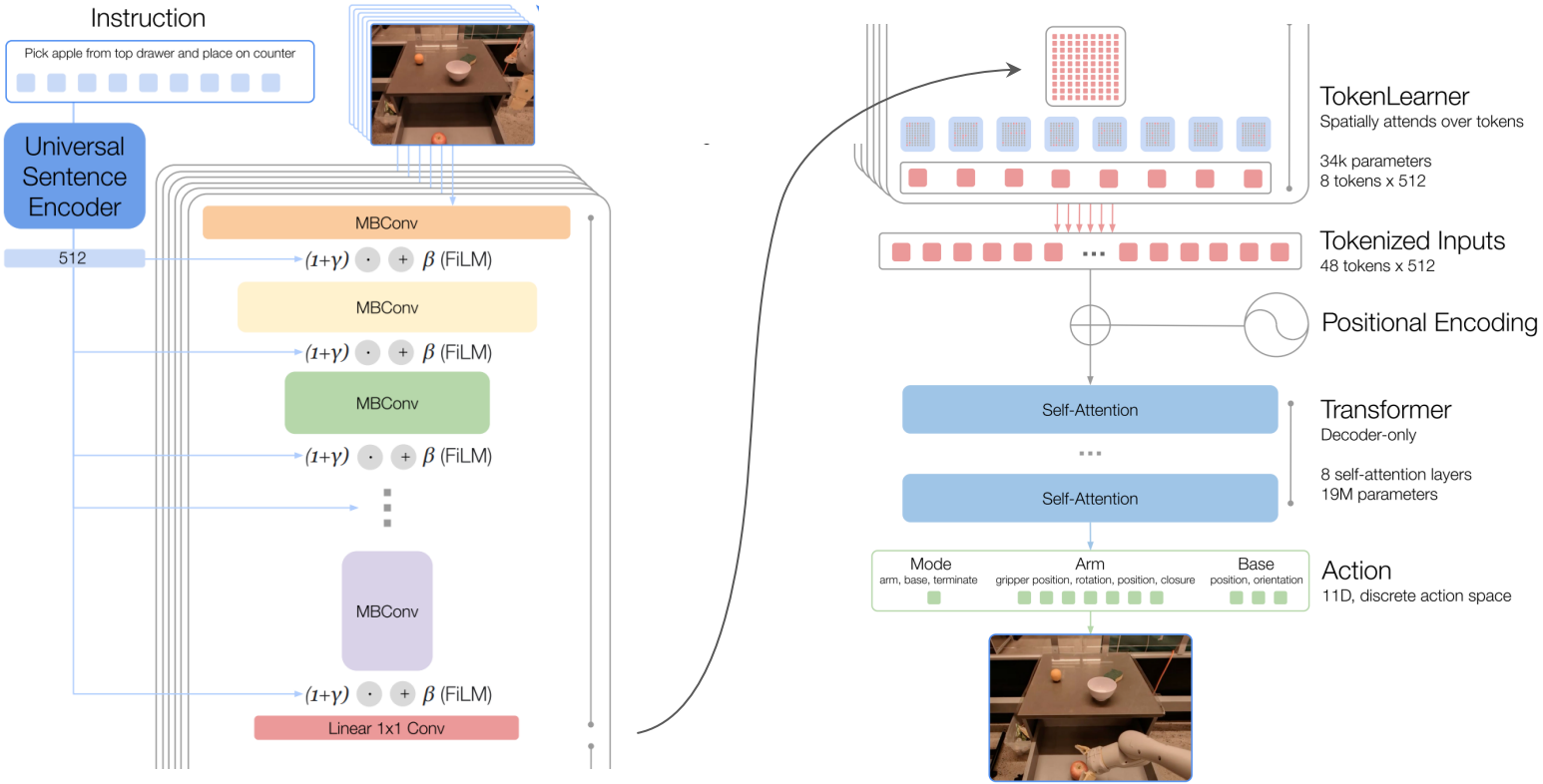
Hybrid Policy: RT-1



Hybrid Policy: RT-1



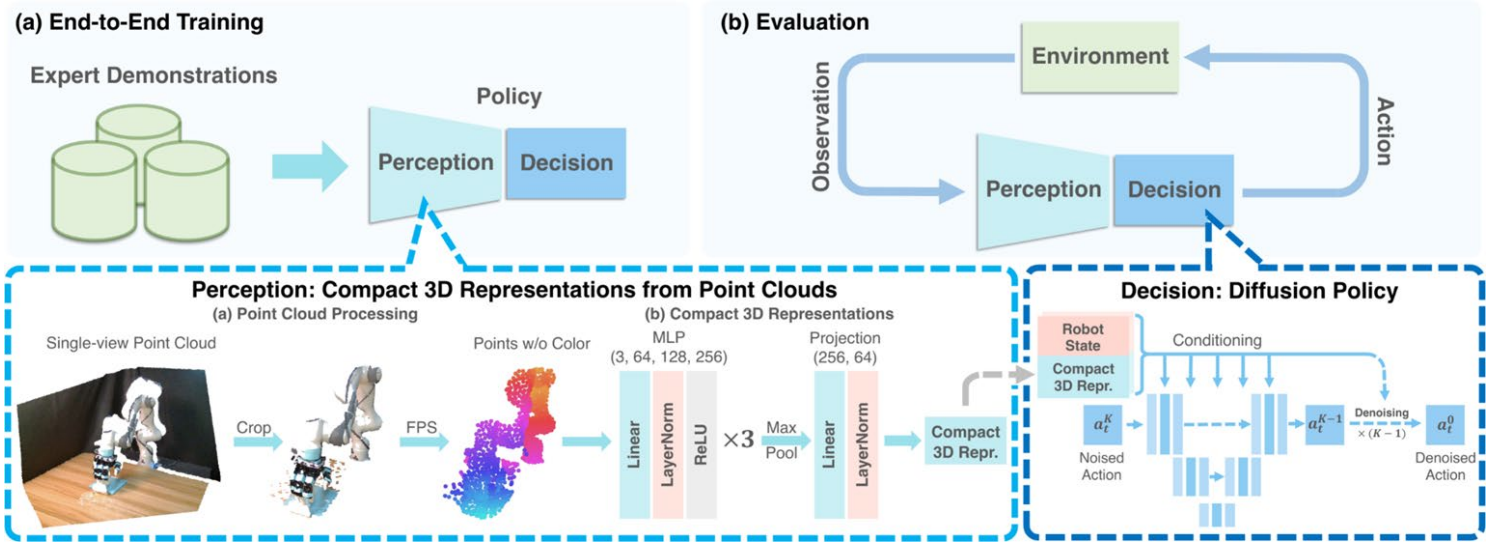
Hybrid Policy: RT-1



Diffusion-based Policy

Transformers → with **diffusion** can model various complex distributions (same action differently)

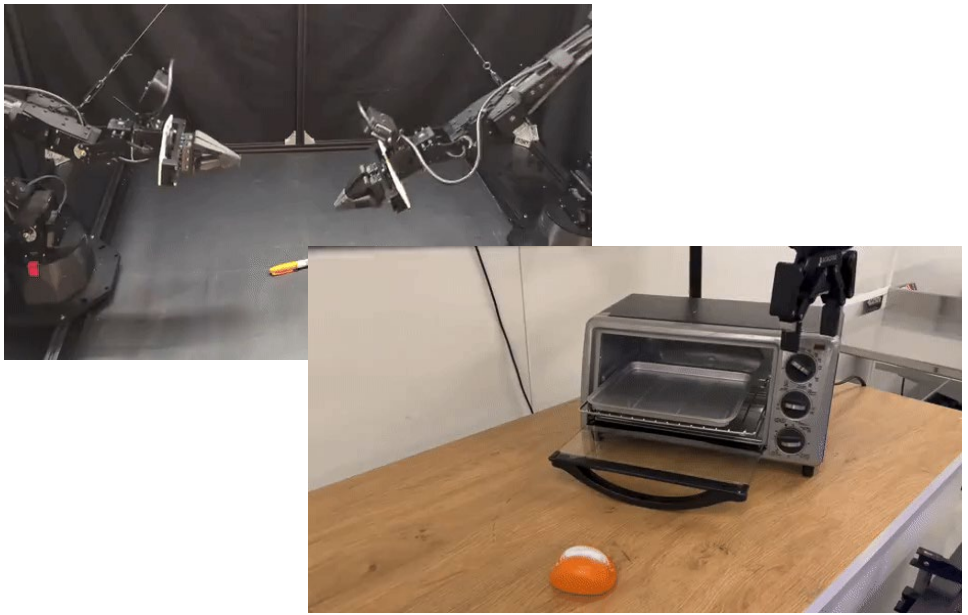
Diffusion **conditions** on the textual prompt



Ze et al. 3D Diffusion Policy: Generalizable Visuomotor Policy Learning via Simple 3D Representations. RSS 2024

Transformer-based: Octo

- Pretrained on 128 TPUv4 (≈ 200 A100) for 14 hours, **much longer than on ImageNet**
- May be finetuned on 3090 in 4 hours
- Best starting point for training own manipulation policies

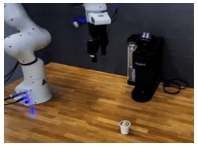


Dibya Ghosh et al. **Octo:** An Open-Source Generalist Robot Policy. **RSS, Oct 2024**. [link]

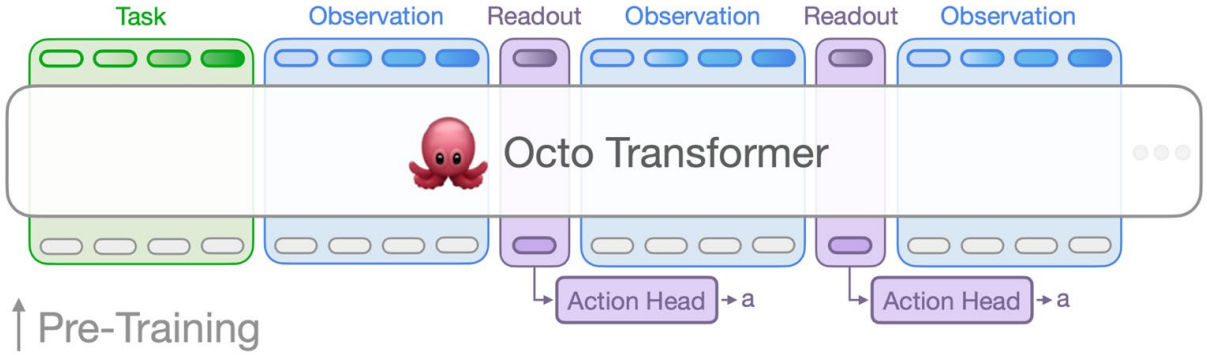
Transformer-based: Octo

Core model is **transformer architecture** that maps arbitrary input tokens (created from observations and tasks) to output tokens (then decoded into actions)

Make coffee

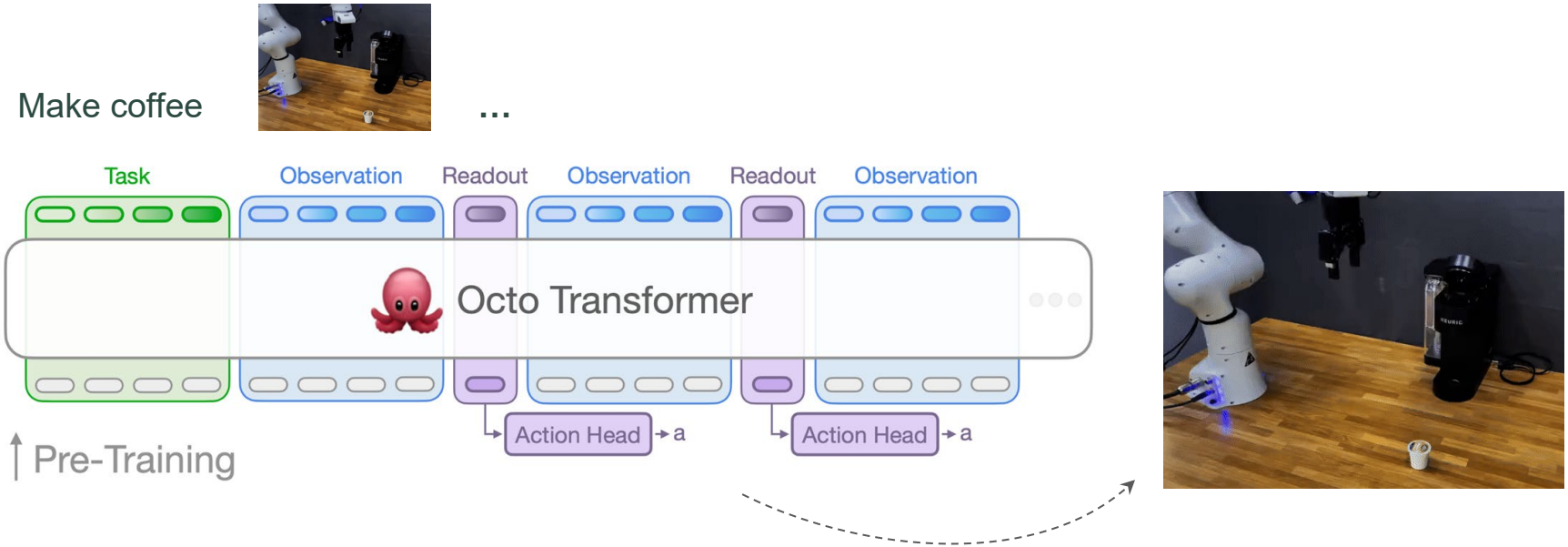


...



Transformer-based: Octo

Core model is **transformer architecture** that maps arbitrary input tokens (created from observations and tasks) to output tokens (then decoded into actions)

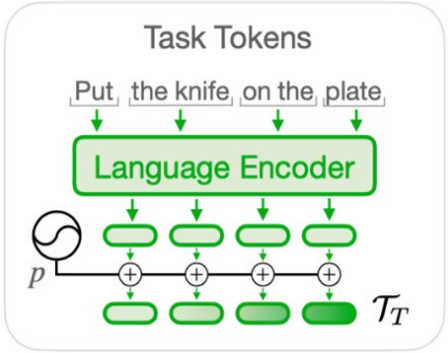


Transformer-based: Octo

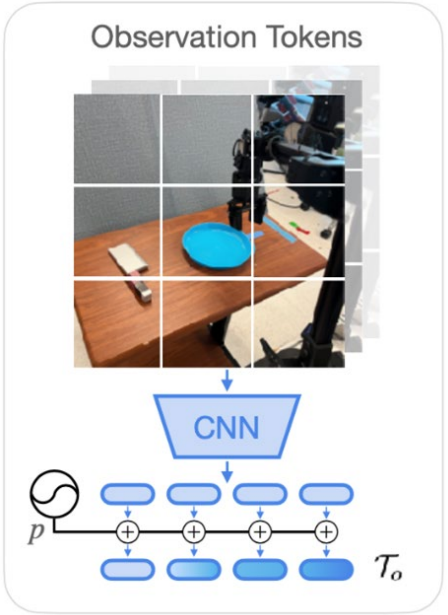
Obtain tokens through **modality-specific tokenizers** ($[\mathcal{T}_T, \mathcal{T}_{o,1}, \mathcal{T}_{o,2}, \ldots]$ usually "embeddings", not "tokens") and arrange all them sequentially

t5-base [111M]

passed through a pretrained transformer that produces a sequence of **language embedding tokens**

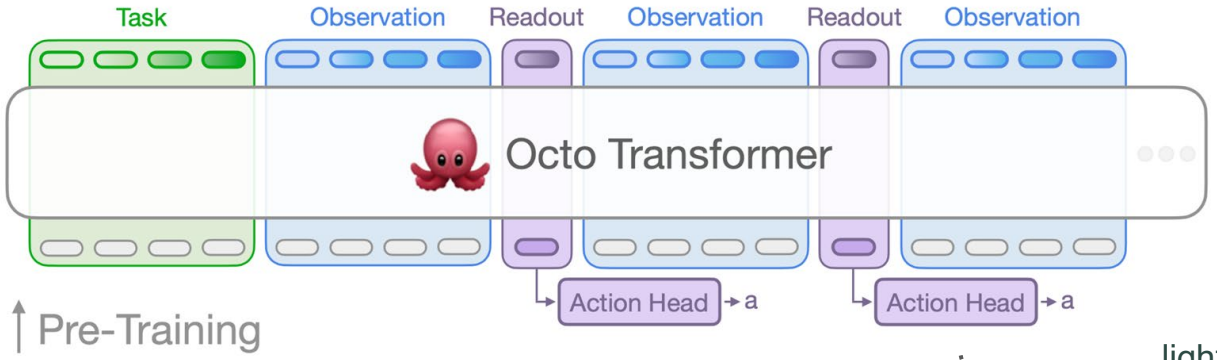


shallow CNN, **flattened patches**



Transformer-based: Octo

observation tokens can only attend causally to tokens **from the same or earlier time steps** as well as task tokens

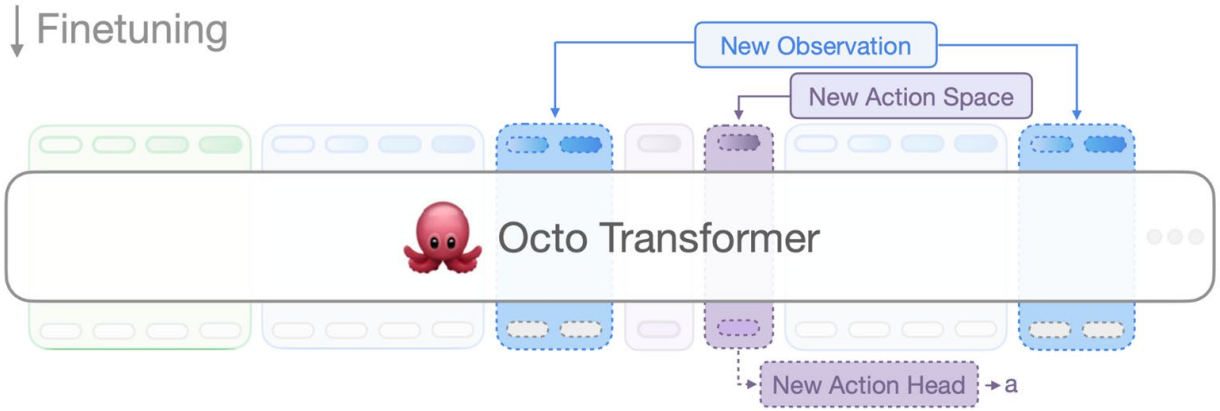


readout tokens is a compact vector embedding of the observation

lightweight **"action head"** that implements the **diffusion process** is applied to the embeddings for the readout tokens

Transformer-based: Octo

When adding new task, observations, loss functions, we can **fully retain the pretrained weights** of the transformer

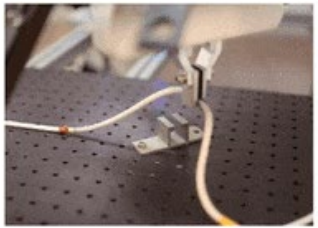


Only adding new encoders or parameters in new head → **generalist model**

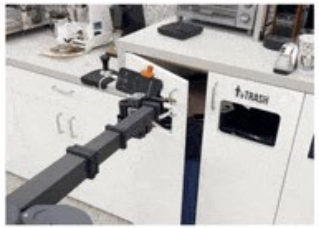
Transformer-based: Octo

Octo was trained on **800k trajectories** from the Open X - Embodiment dataset *(from 1.5M robot episodes)*

CLVR, USC

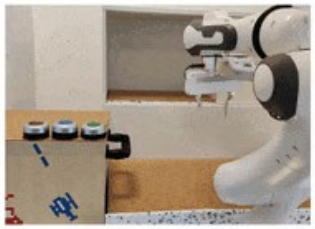


RAIL, UC Berkeley



CILVR, NYU

AUTOLab, UC Berkeley



AiS, University of Freiburg

25 datasets

several robot **embodiments** , scenes

sensors and labels

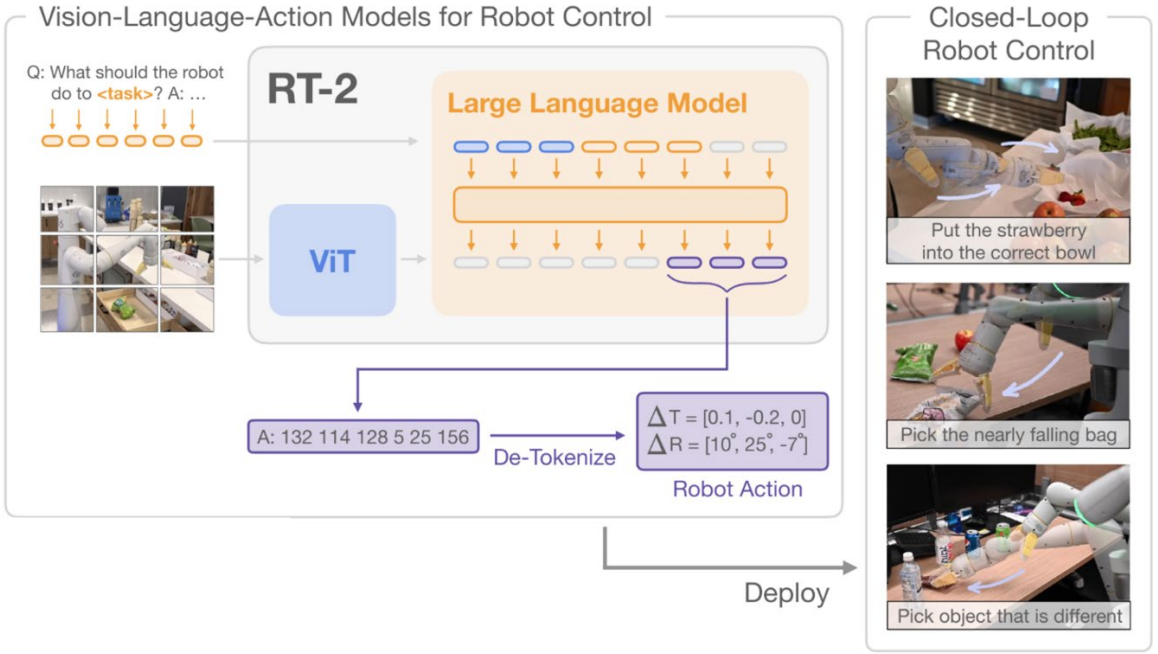
remove **repetitive** datasets

remove **low** image resolution

diverse dataset

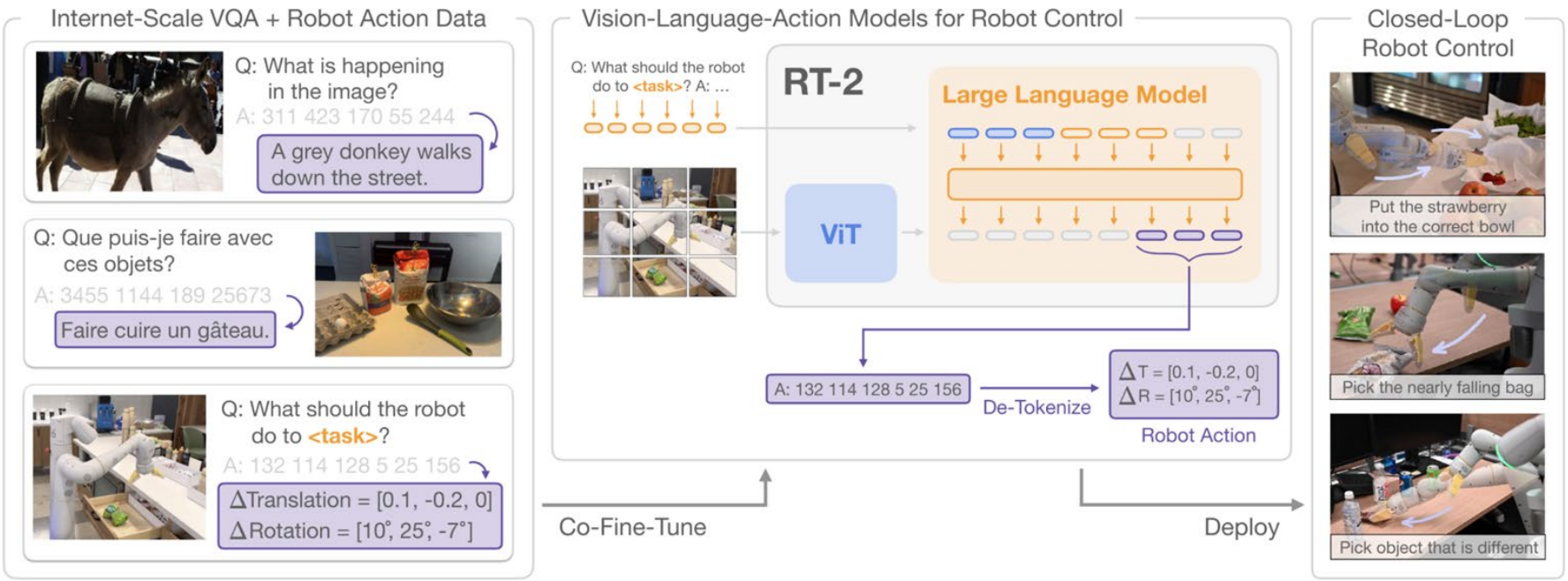
VLLMs Policies: RT-2

- **Reusing knowledge** from LLM: easier generalizability
- Manipulating objects that it has not seen
- Action questions + VQA for avoiding catastrophic forgetting



Brohan et al. RT -2: Vision -Language -Action Models Transfer Web Knowledge to Robotic Control. arXiv:2307.15818

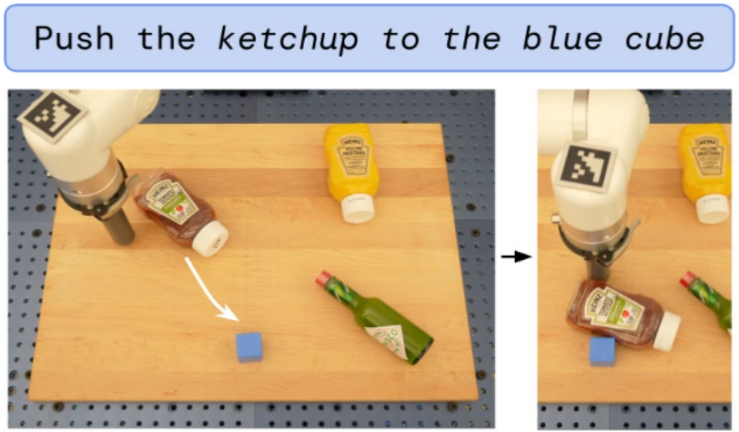
VLLMs Policies: RT-2



Brohan et al. RT -2: Vision -Language -Action Models Transfer Web Knowledge to Robotic Control. arXiv:2307.15818

VLLMs Policies: RT-2

→ better generalization to unseen objects

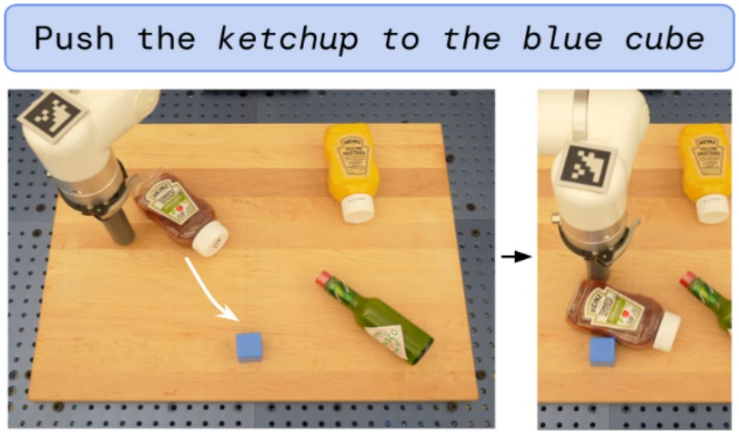


Brohan et al. RT -2: Vision -Language -Action Models Transfer Web Knowledge to Robotic Control. arXiv:2307.15818

VLLMs Policies: RT-2

→ better generalization to unseen objects

→ reasoning

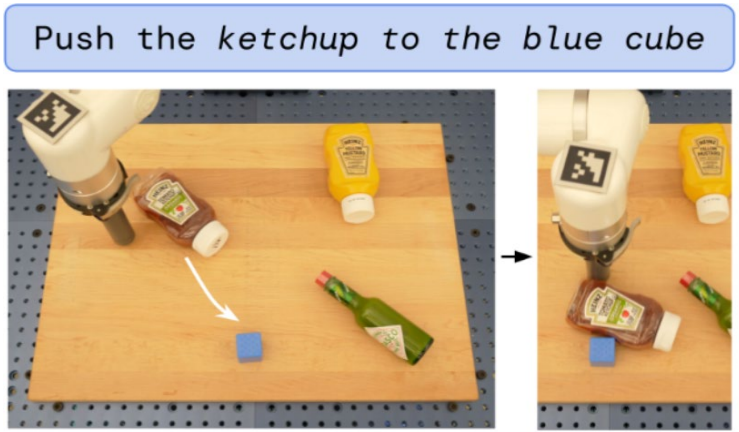


move banana to the sum of two plus one

Brohan et al. RT -2: Vision -Language -Action Models Transfer Web Knowledge to Robotic Control. arXiv:2307.15818

VLLMs Policies: RT-2

→ better generalization to unseen objects → reasoning → human recognition



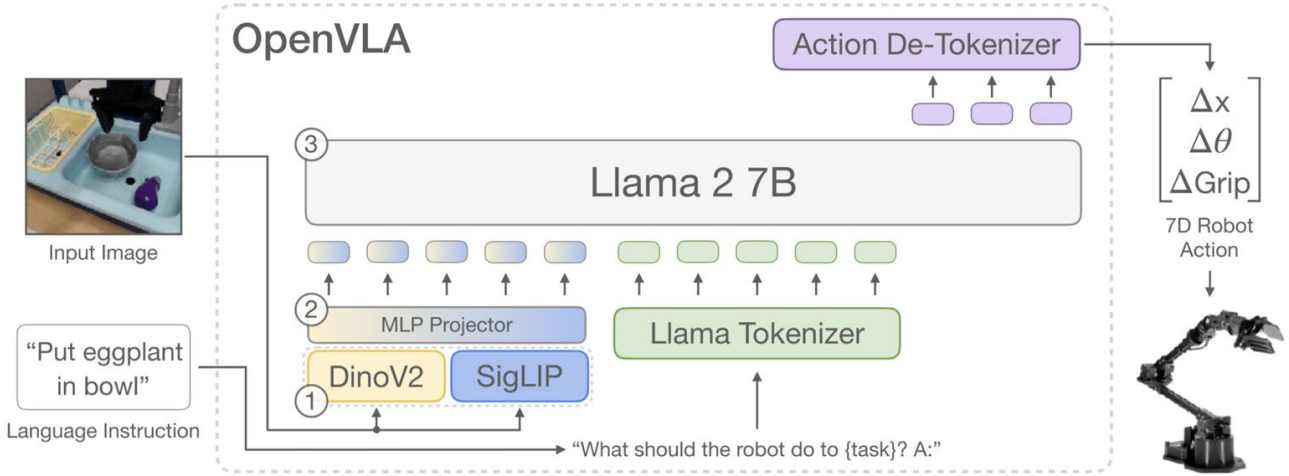
move banana to the sum of two plus one

move coke can to Taylor Swift

Brohan et al. RT -2: Vision -Language -Action Models Transfer Web Knowledge to Robotic Control. arXiv:2307.15818

VLLMs Policies: OpenVLA

Pretrained on 64 A100 for 14 days, finetuned on 8 A100 for 5 -15 hours



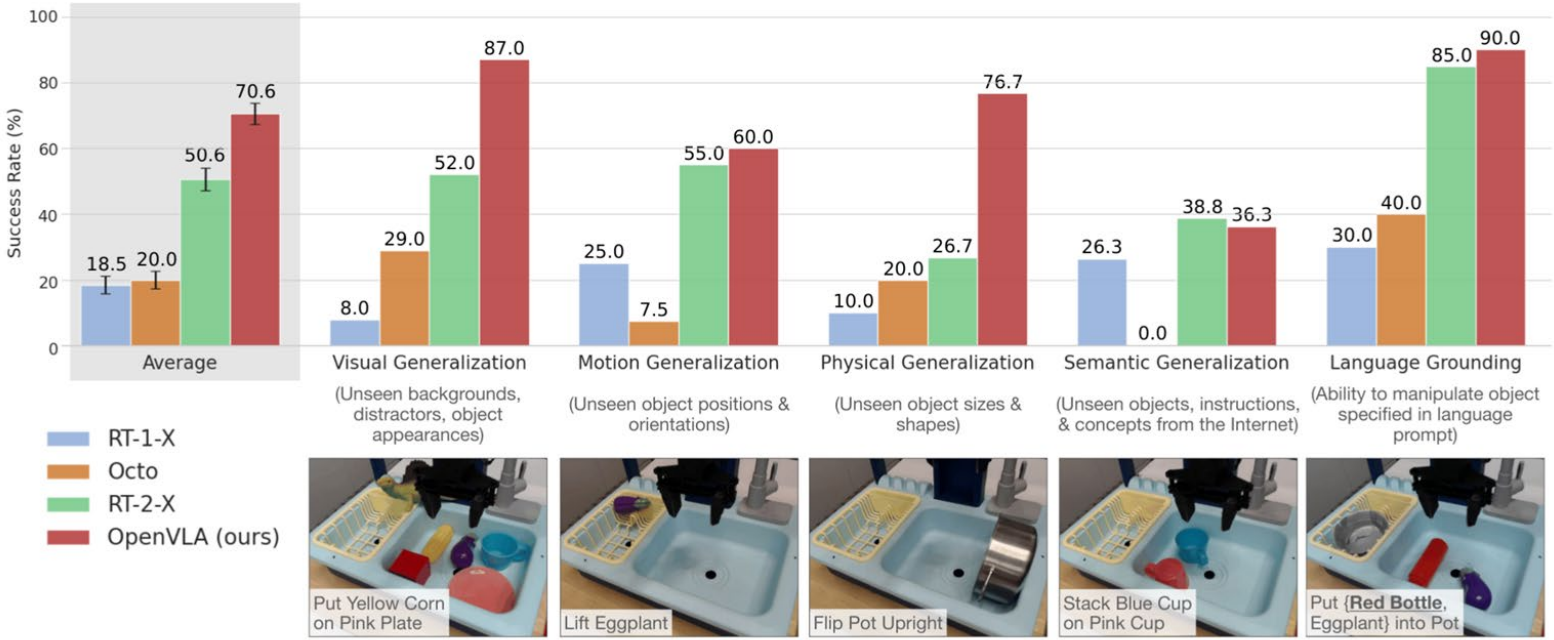
LLM – LLaVA-2

Dino – spatial understanding is better

SigLIP – about general concepts

Moo Jin Kim et al. OpenVLA: An Open -Source Vision -Language -Action Model. 2024

VLLMs Policies: OpenVLA



Moo Jin Kim et al. OpenVLA: An Open -Source Vision -Language -Action Model. 2024

Future Research

- → VLA model with multi - image/videos and **depth observations**
- → Performance **improvement** (now SR < 90%)
- → Co-training for VQA and action prediction is to be explored — **resource intensive** now

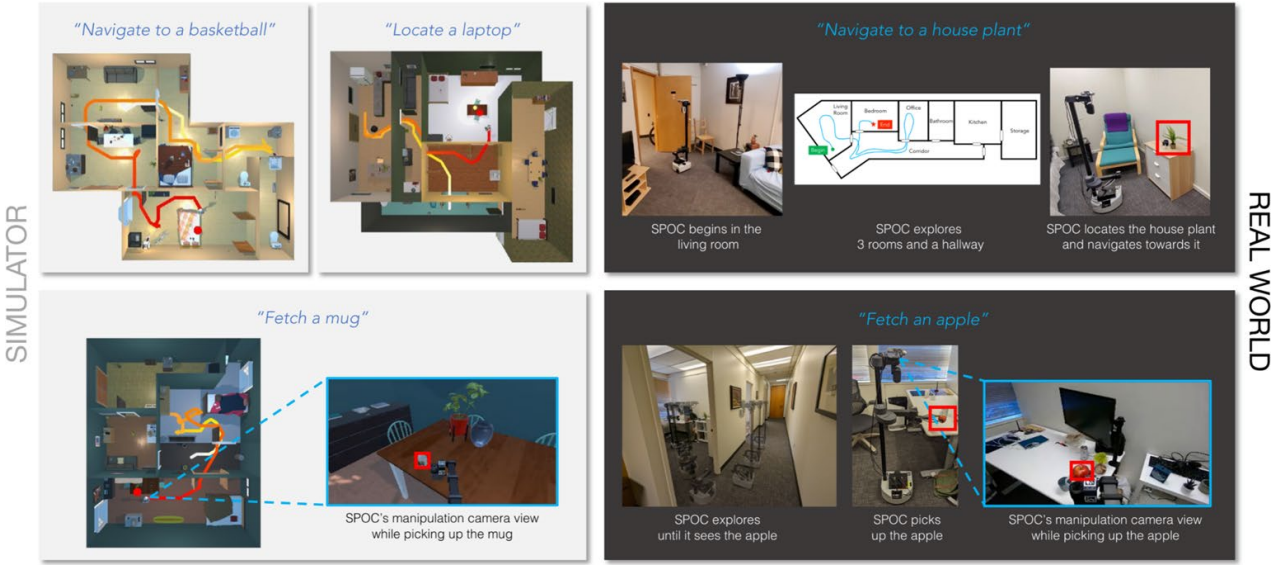
5.2

Acting: Navigation

How to find an optimal path, base movement

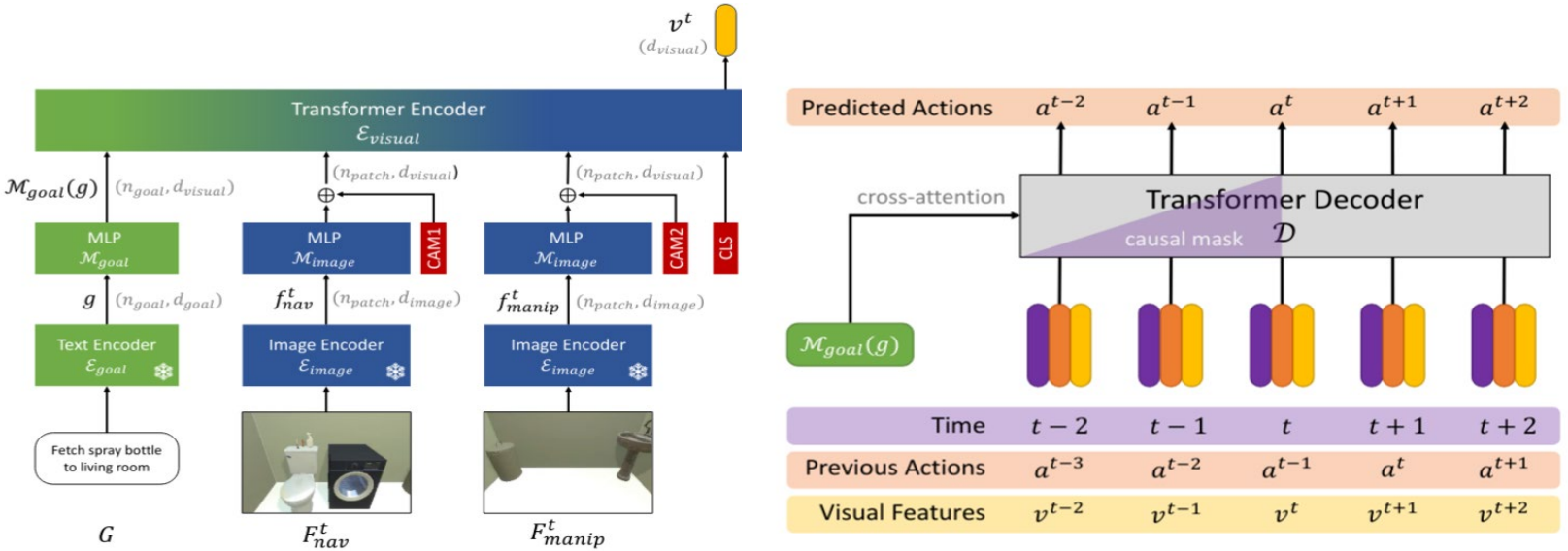
Navigation: SPOC

The robot arm can ride up and down + base movements



Navigation: SPOC

Actions: move base (±20cm), rotate base (±6°,±30°), move arm (x, z) (±2cm, ±10cm)



Ehsani et al. SPOC: Imitating Shortest Paths in Simulation Enables Effective Navigation and Manipulation in the Real World. C VPR 2024

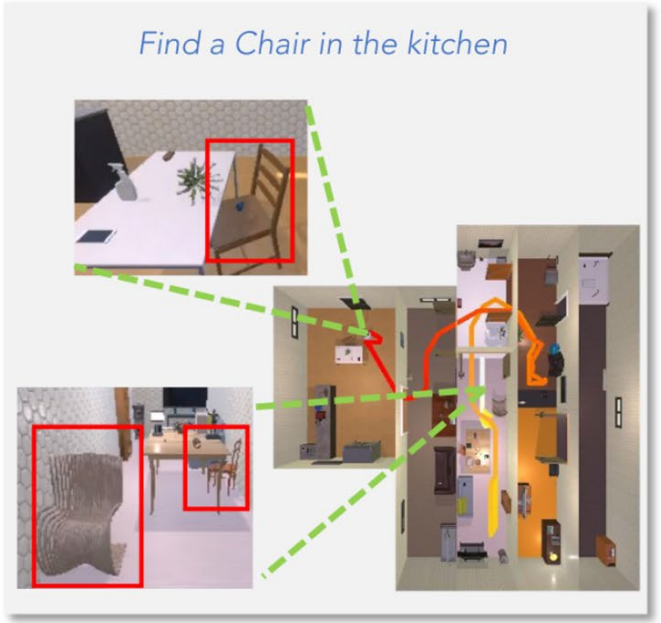
Navigation: SPOC

Trajectories:

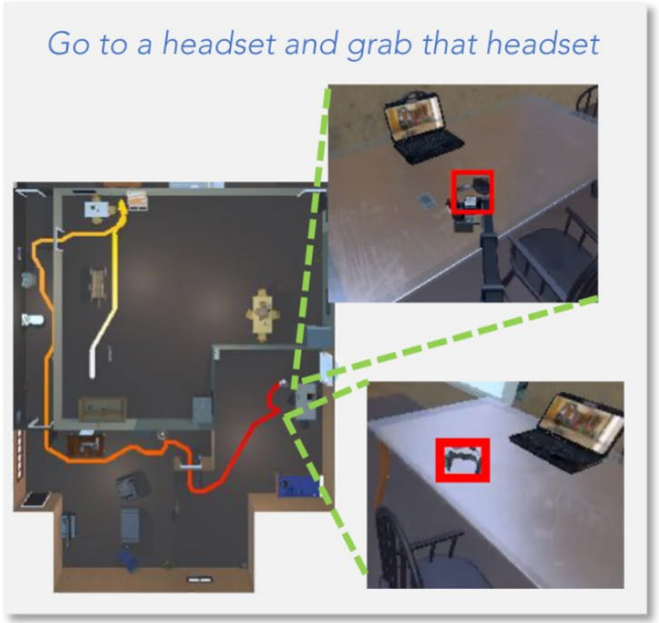
- → **Navigation:** go to target using approximation of shortest path
- → **Room visitation:** calculate center of house, then navigate to all rooms via shortest paths

Navigation: SPOC, Examples

skipping all chairs to find one in the kitchen



repositioning itself to find a location where **headset is reachable**



Open Questions

- → Unified and effective **evaluation**
- → **Sim-to-real** gap
- → Efficient **collection** of human demonstration data
- → **High inference time** of foundation models
- → **Long-horizon** task planning
- → Ensuring **robustness** and safety of deployed models

Conclusions

1. Embodied AI is a research area at the intersection of **NLP, CV and RL**

2. **Evaluation of EAI models** in general is a very challenging task

3. Despite a **decade of the rapid progress** in NLP and CV, EAI systems (understanding the world, planning and acting) are in the beginning of their development