看看具身智能都在做些啥

2024/06/19

许博深

Before Starting

- We **will** discuss the robotics that is more closely related to **Multi-Modal / Computer Vision / LLM**.
- We **will not** discuss the topics related to **Reinforcement Learning / Control / Autonomous Driving**.
- Without specification, the robot has the ability to **move** (How robot body moves without collision), achieve **motion planning** (How robot arm moves successfully), *etc.*

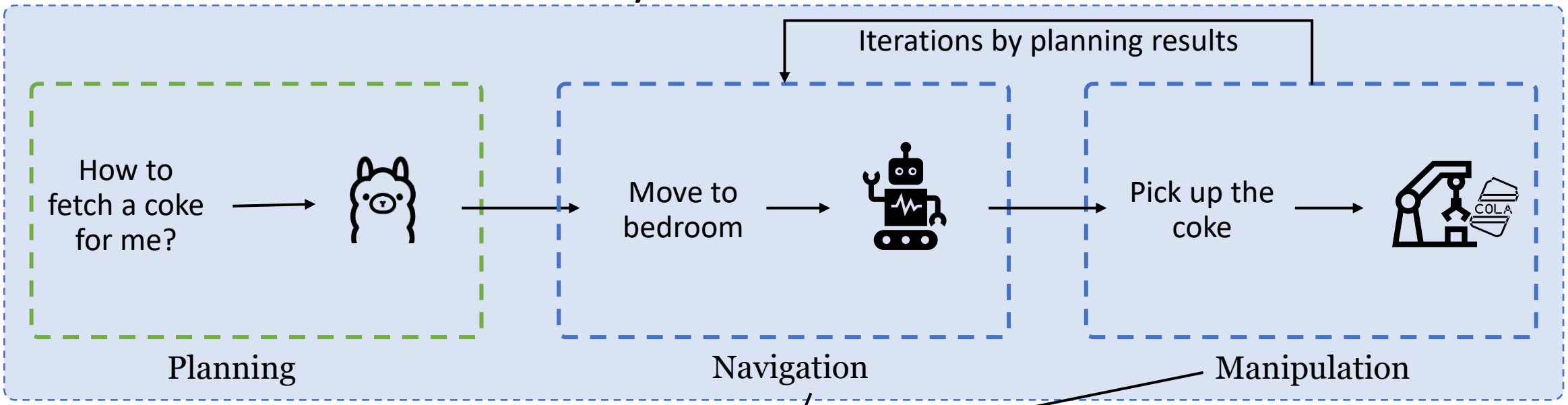
✓

This May Be How We Hope Robots Do:

Source from: Do As I Can, Not As I Say: Grounding Language in Robotic Affordances, Google Robotics

Rough Summary of Embodied AI

Task: difficulty & how to follow instruction



Data

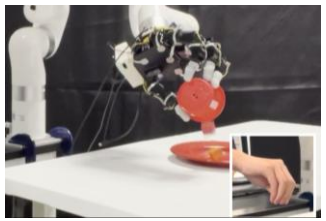
Data Type:how robot perceive and act?

embodiment

multi-modal

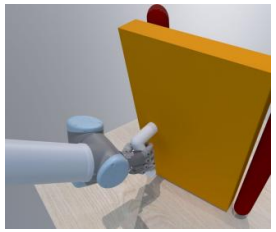
trajectory

Data Sources



real-world

human videos



simulators

Important Tasks and Solutions in Embodied AI

LLM-based planner

(SayCan [CoRL22] & LLM planner [ICML22])

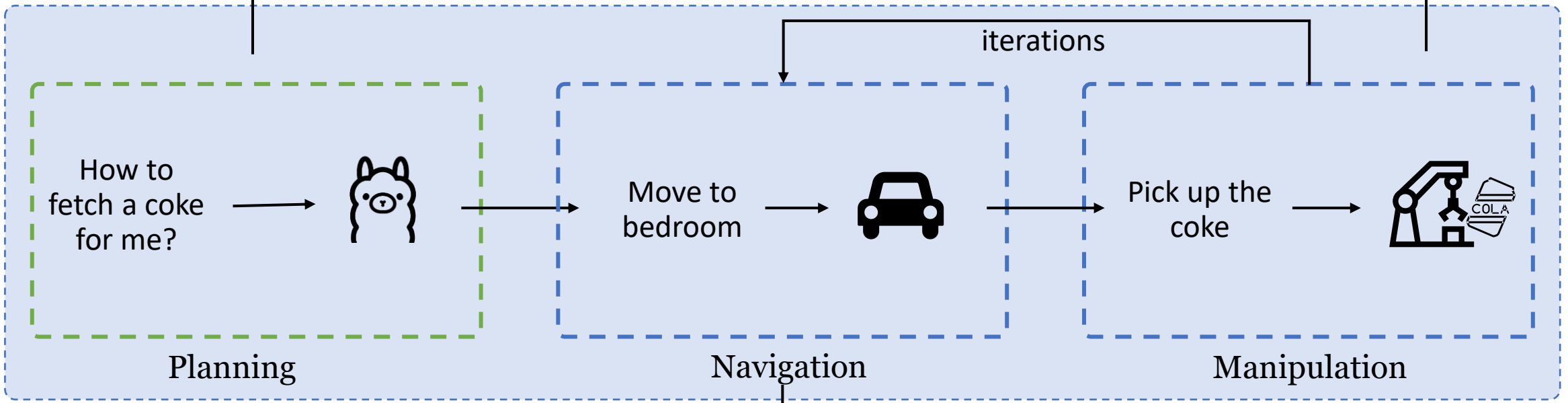
Step1: move to the bedroom;

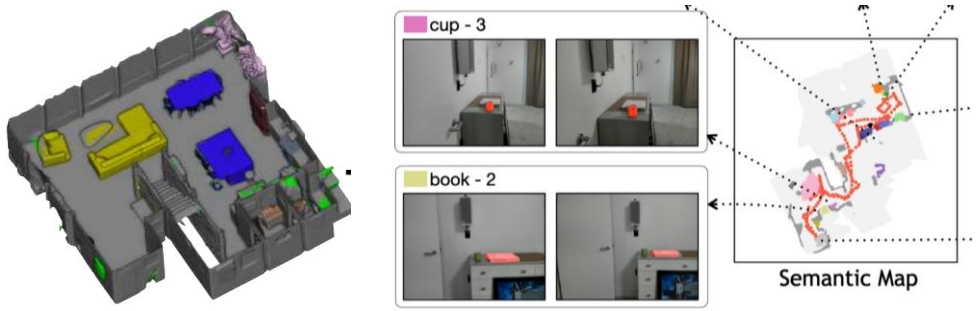
Step2: move to the table

Step3: pick up the coke …

Vision-based Solution

AnyGrasp [TRO23], RT-2 [Google]





Modular-based (core: map) for finding positions

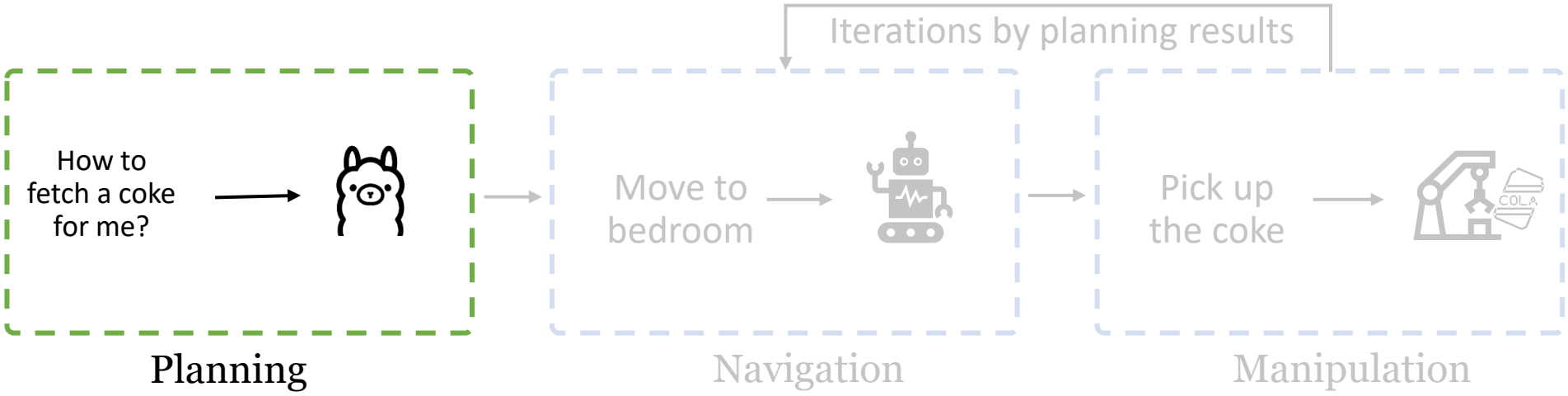
(3D Scene Graph[ICCV19] SemExp [NeurIPS20])

End-to-end (map-free) methods

(RIM[IROS23], NaVid[RSS24])

1 Planning – Language-Only Solution

LLM planner: Planning actionable steps for high-level task



Language Models as Zero-Shot Planners: Extracting Actionable Knowledge for Embodied Agents, ICML2022

Task and Assumption in Simulation

- Task: Given **Task** (in the form of language), output **action / skill sequences**.

- Assumptions in simulators

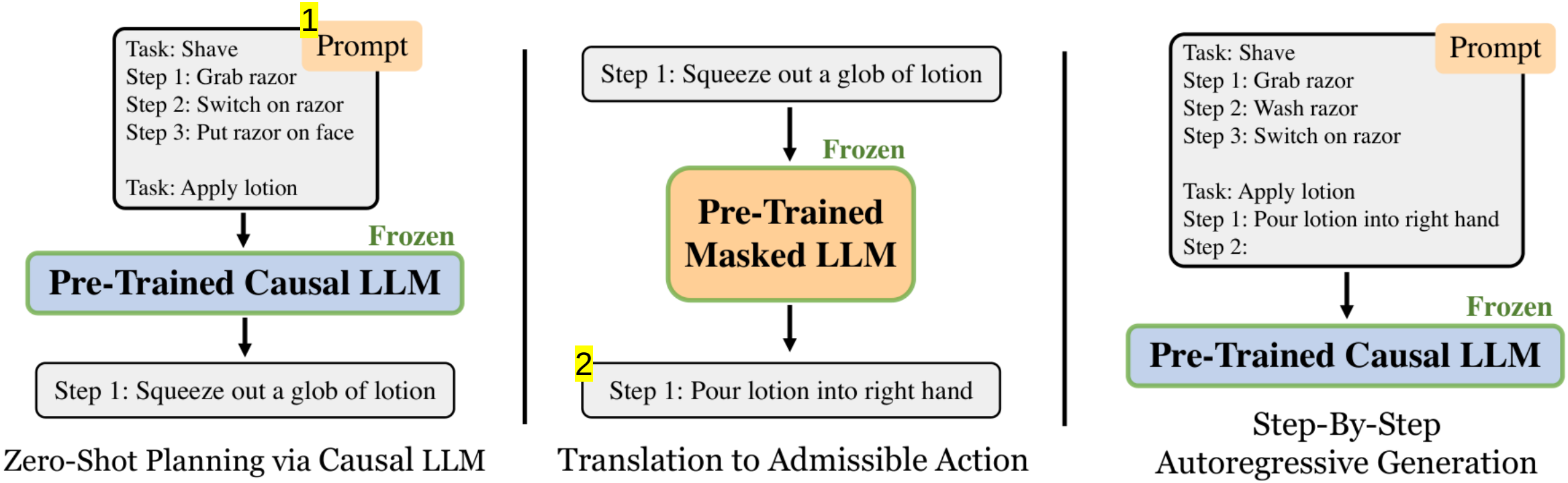
➢ Predefined skills: [Atomic action] <object> (id)

➢ Robot can 100% accomplish each skill.

Virtualhome: Simulating household activities via programs, CVPR2018

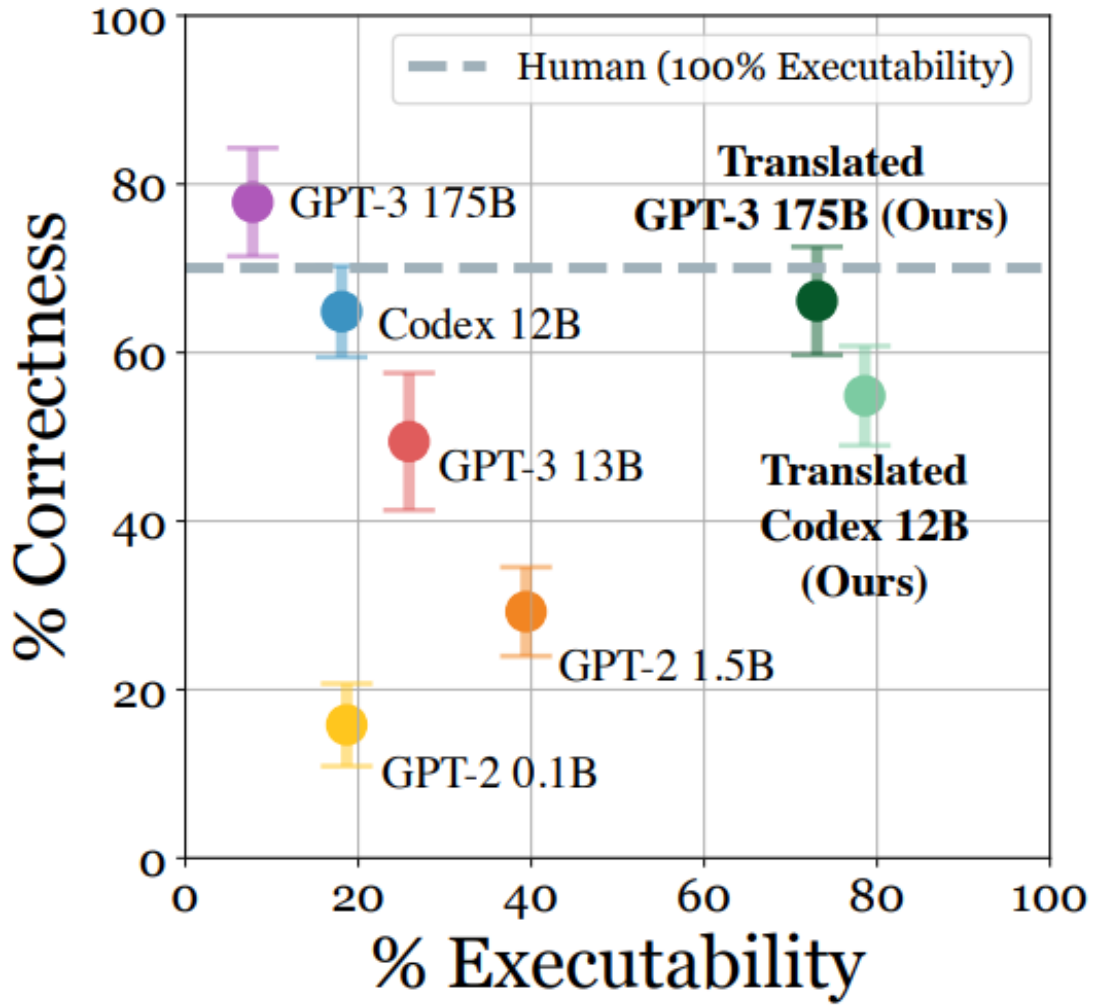
Autoregressive Planning and Mapping

1. For 88 eval tasks, find few-shot examples among 208 seen tasks (Apply lotion → Shave)
2. Get actionable skill from the skill sets (47k) by highest cosine similarities.



Results

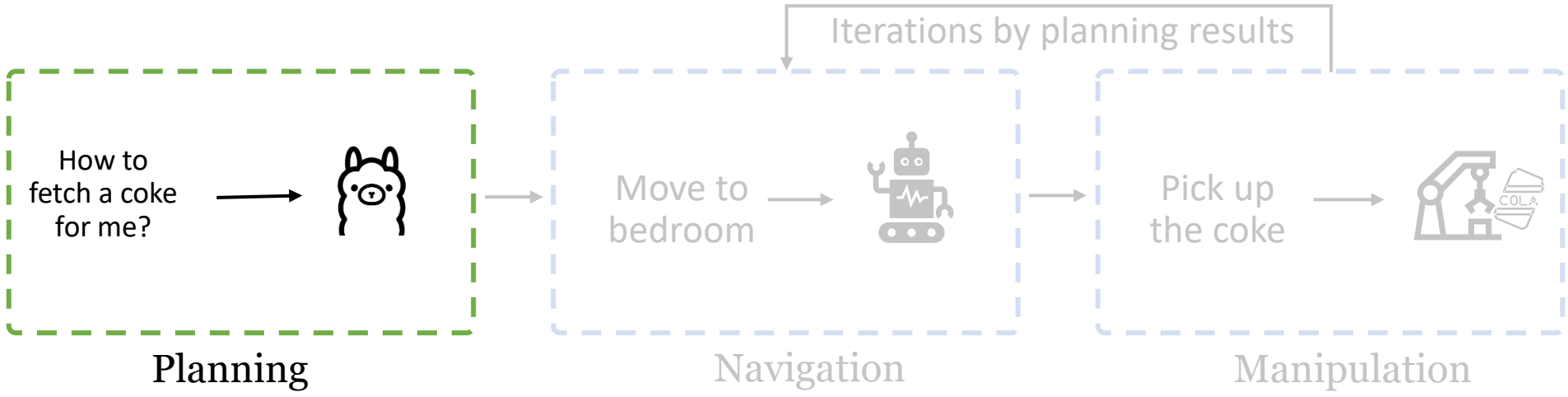
Human-eval



Env-eval

1 Planning - Grounding LLM in Real-World

SayCan: determing / grounding LLM's planning by visual feedback

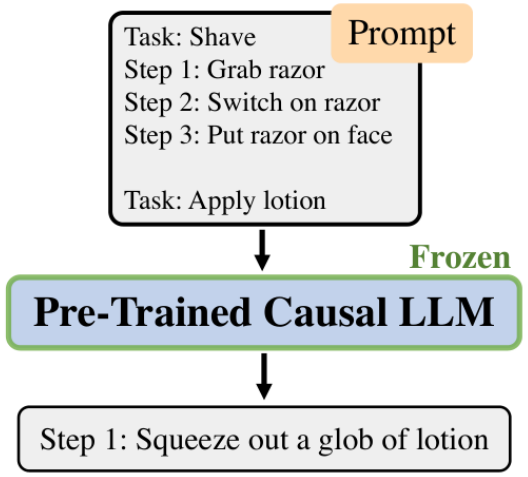


Language Models as Zero-Shot Planners: Extracting Actionable Knowledge for Embodied Agents, ICML2022

SayCan: Grounding Language with Visual State

- LLM planner: $a_t^* = \text{argmax}_a \, p(l_a | I, l_{a_{t-1}}, \ldots, l_{a_0})$
  - I: high-level language instruction, l: action language instruction, a: skill sets

- SayCan: $a_t^* = \text{argmax}_a \, p(c_a | l_a, \boldsymbol{s_t}) \cdot p(l_a | I, l_{a_{t-1}}, \ldots, l_{a_0})$
  - s_t: Image at t step; if do action l_a at image state $\boldsymbol{s_t}$, the value is c_a, value function trained by RL



LLM planner

SayCan

Do As I Can, Not As I Say: Grounding Language in Robotic Affordances, CoRL22

Planning is Easier, Execution is Harder

- SayCan: $a_t^* = \text{argmax}_a \, p(c_a|l_a, s_t) \cdot p(l_a|I, l_{a_{t-1}}, \ldots, l_{a_0})$

		Mock Kitchen		Kitchen		No Affordance		No LLM	
		PaLM-SayCan	**PaLM-SayCan**	**PaLM-SayCan**	**PaLM-SayCan**	**No VF**	**Gen.**	**BC NL**	**BC USE**
Family	**Num**	Plan	Execute	Plan	Execute	Plan	Plan	Execute	Execute
NL Single	15	100%	100%	93%	87%	73%	87%	0%	60%
NL Nouns	15	67%	47%	60%	40%	53%	53%	0%	0%
NL Verbs	15	100%	93%	93%	73%	87%	93%	0%	0%
Structured	15	93%	87%	93%	47%	93%	100%	0%	0%
Embodiment	11	64%	55%	64%	55%	18%	36%	0%	0%
Crowd Sourced	15	87%	87%	73%	60%	67%	80%	0%	0%
Long-Horizon	15	73%	47%	73%	47%	67%	60%	0%	0%
Total	101	84%	74%	81%	60%	67%	74%	0%	9%

- NL nouns: abstract noun (e.g. bring me a fruit)
- Embodiedment: change environment state (e.g. start)

Long-horizon Demo

Use input: I spilled my coke on the table, how would you throw it away and bring me something to help clean ?

Source from: Do As I Can, Not As I Say: Grounding Language in Robotic Affordances, Google Robotics

Brief Summary of Task Planning

- **Task planning**: $a_t^* = \text{argmax}_a \, p(c_a|l_a, s_t) \cdot p(l_a|I, l_{a_{t-1}}, \ldots, l_{a_0})$

Planner: LLM LLM planner [ICML22], Voyager [TMLR24], Code as Policies [ICRA23]	+ Value function SayCan [CoRL22] (visual feedback), SayPlan [CoRL23] (scene graph & tree search)
Planner: MLLM PaLM-E [ICML23], EmbodiedGPT [NeurIPS24]	+ Value function VLP [ICLR24] (t2v generative models & tree search)

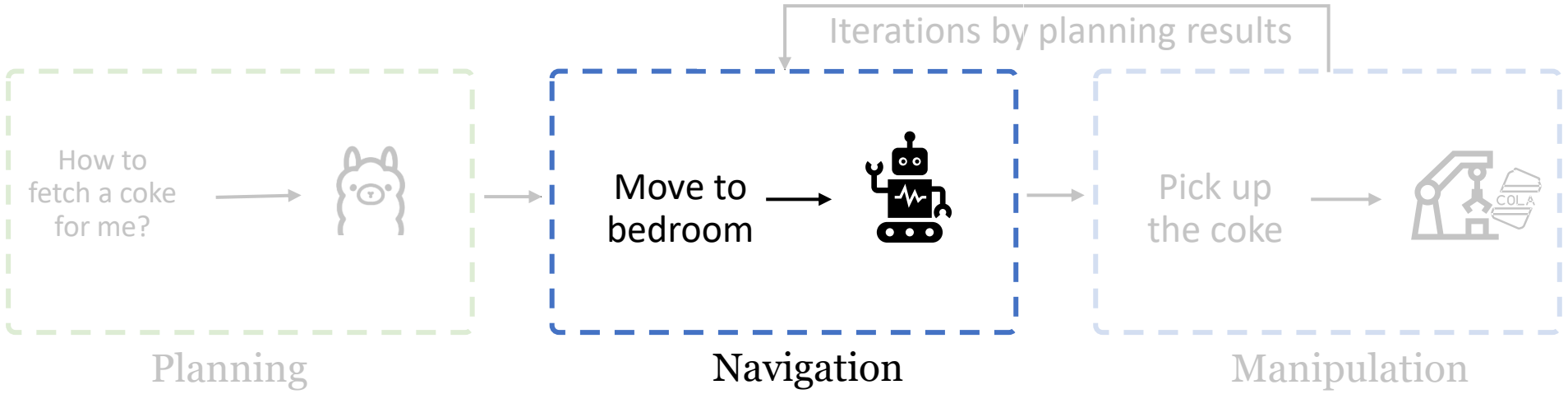
Key Problem: Grounding LLM in real-world

- LLM-Centered Research
- Value function is the key to ground (M)LLMs in the real-world, which leaves room for exploration.

2 Navigation-Modular System

GOAT: Go to any thing by modular system

(perception, planning by memory & goal, control)



GOAT: Modular System for Navigation

Multimodal Goals

Language Goal

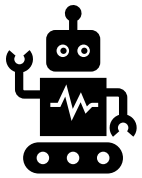
Brown office chair
Bed with white sheets
Cup on the kitchen counter

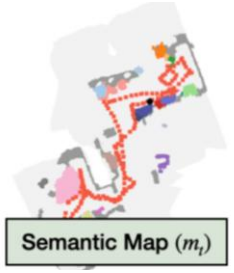
Image Goal

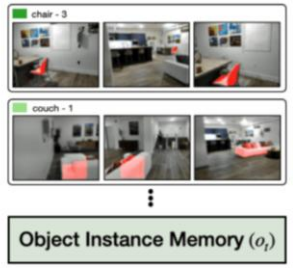
Category Goal

CHAIR
BOOK
CUP

Global Policy
Specify goal
(x, y)

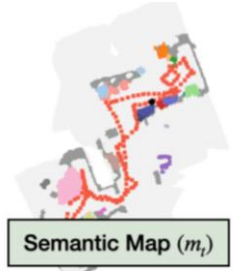


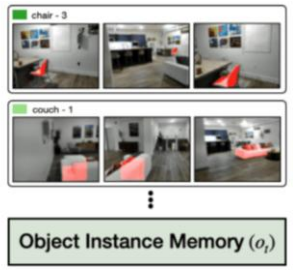




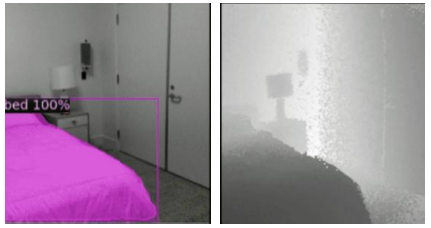
Maps & memory

update





Perception
RGBD+Seg



Goal?

no

yes

terminate

Local Policy
Plan Action & execute
(e.g. Move forward)

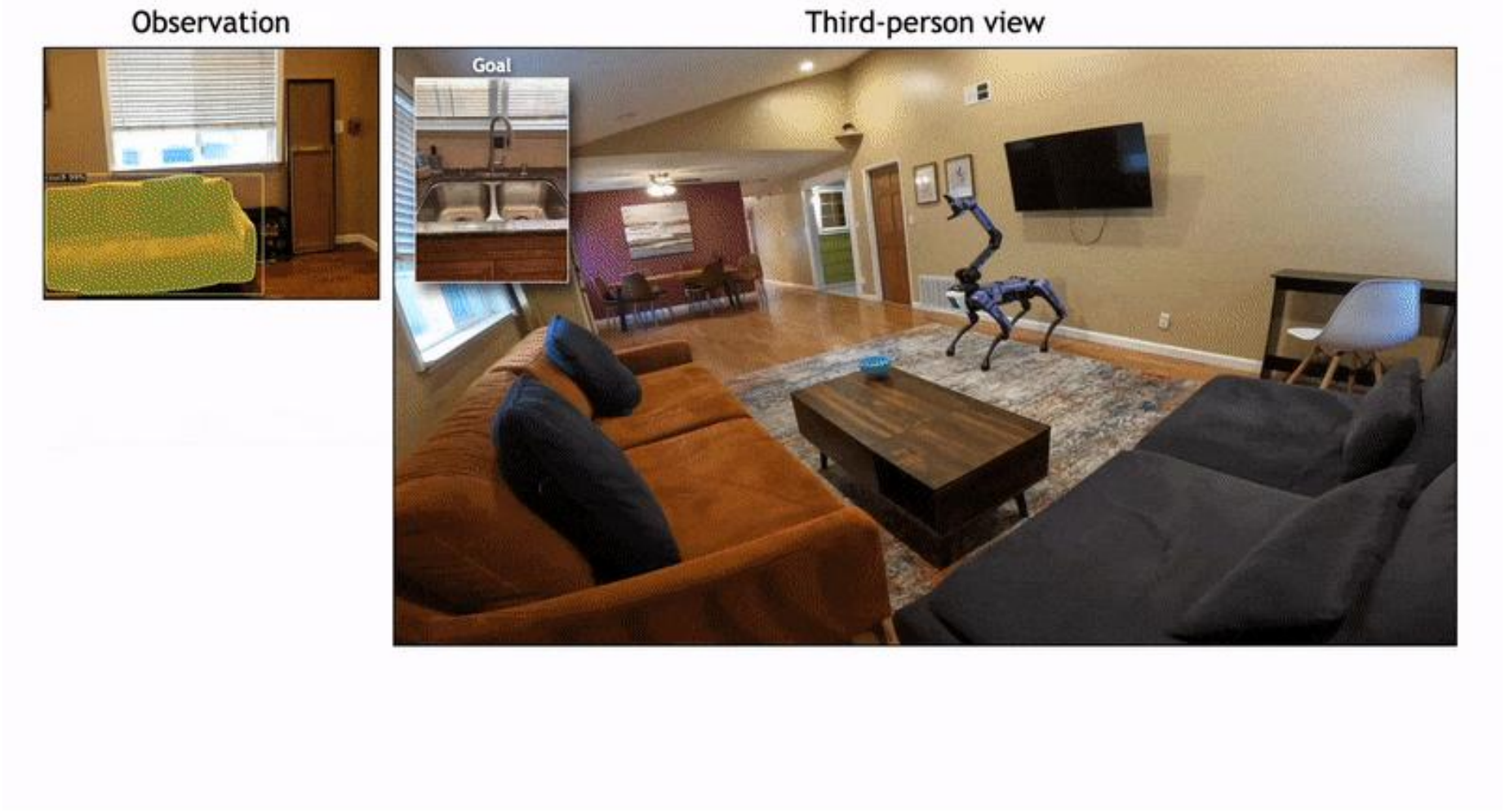
goal (x, y)
& current position

GOAT: GO to Any Thing, arxiv23, CVPR24, Science Robotics23

By Boshen Xu

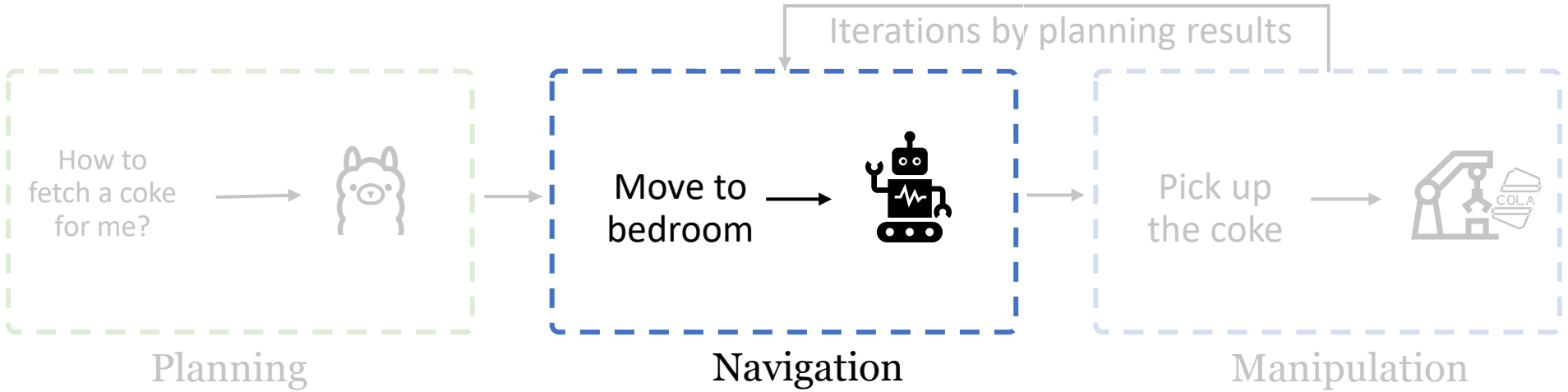
Superior Results of GOAT

- **83%** success rate in 9 unseen homes; 675 goals with 200+ object instances

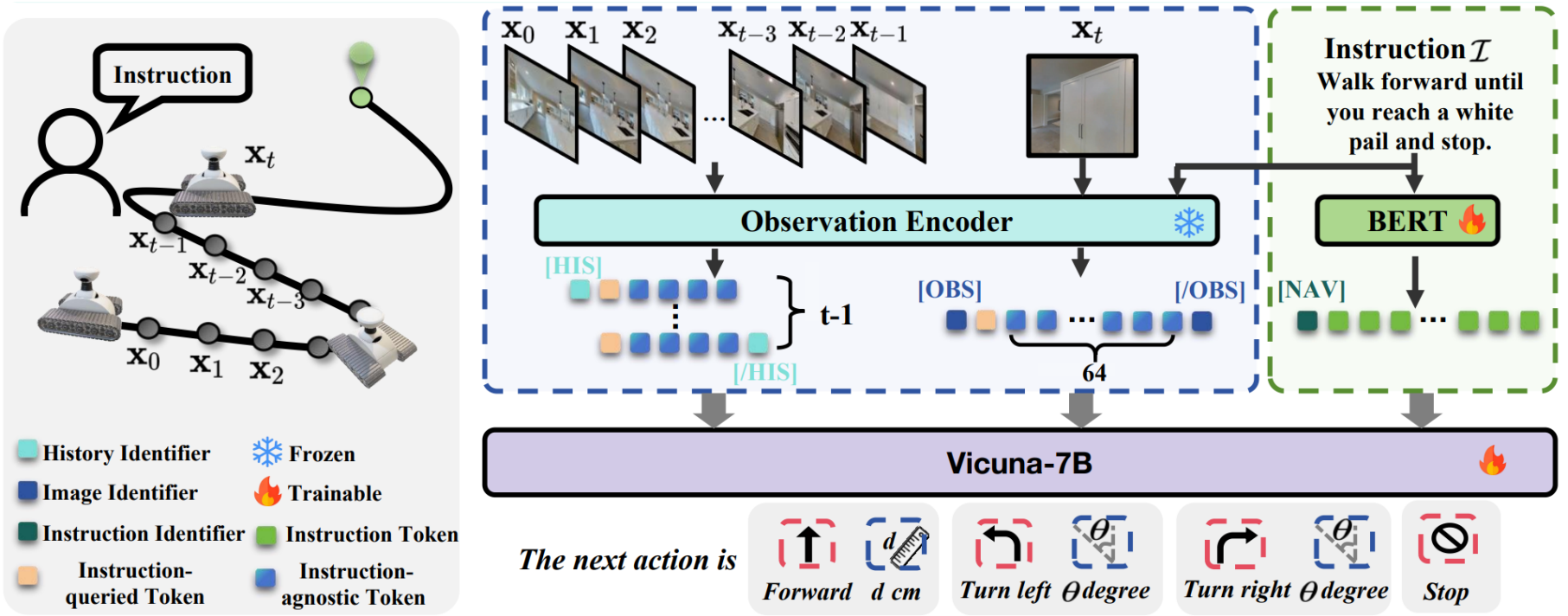


2 Navigation: Vision-Language-Action Model

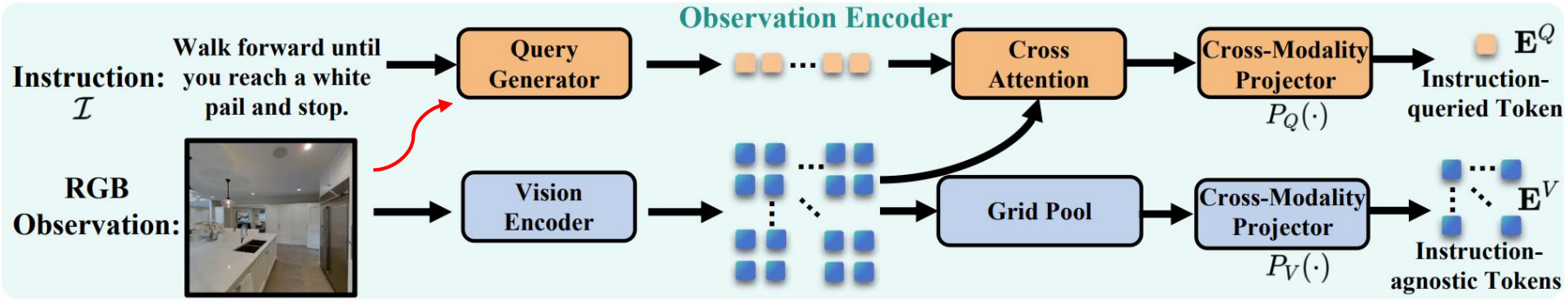
NaViD: navigation via Multimodal Large Language Models



NaVid: Large Vision-and-Language Model



Q-former from InstructBLIP



Similar to LLaMA-VID

Results of NaVid

	Observation				VLN-CE R2R Val-Unseen				
	Pan.	S.RGB	Depth	Odo.	TL	**NE↓**	**OS↑**	**SR↑**	**SPL↑**
HPN+DN* [47]	✓		✓	✓	7.62	6.31	40.0	36.0	34.0
CMA*† [37]	✓		✓	✓	10.90	6.20	52.0	41.0	36.0
VLN↻BERT*† [37]	✓		✓	✓	12.23	5.74	53.0	44.0	39.0
Sim2Sim* [44]	✓		✓	✓	10.69	6.07	52.0	43.0	36.0
GridMM*† [102]	✓		✓	✓	13.36	5.11	61.0	49.0	41.0
HAMT*†‡ [103]	✓		✓	✓	–	4.80	–	55.0	51.0
ETPNav* [4]	✓		✓	✓	11.99	4.71	65.0	57.0	49.0
AG-CMTP [15]	✓		✓	✓	–	7.90	39.2	23.1	19.1
R2R-CMTP [15]	✓		✓	✓	–	7.90	38.0	26.4	22.7
LAW [77]		✓	✓	✓	8.89	6.83	44.0	35.0	31.0
CM2 [31]		✓	✓	✓	11.54	7.02	41.5	34.3	27.6
WS-MGMap [16]		✓	✓	✓	10.00	6.28	47.6	**38.9**	34.3
Seq2Seq [45]		✓	✓		9.30	7.77	37.0	25.0	22.0
CMA [45]		✓	✓		8.64	7.37	40.0	32.0	30.0
RGB-Seq2Seq		✓			4.86	10.1	8.10	0.00	0.00
RGB-CMA		✓			6.28	9.55	10.8	5.00	4.43
Ours		✓			7.63	**5.47**	**49.1**	37.4	**35.9**

in simulation

	Meeting Room				Office			
	Simple I.F.		Complex I.F.		Simple I.F.		Complex I.F.	
	SR↑	NE↓	SR↑	NE	SR↑	NE↓	SR↑	NE↓
Seq2Seq [45]	4%	4.45	0%	7.21	0%	4.28	0%	6.92
CMA [45]	0%	4.27	0%	7.30	8%	4.62	0%	5.71
WS-MGMap [16]	52%	1.18	24%	2.20	60%	0.96	20%	2.94
Ours	**92%**	**0.55**	**56%**	**0.98**	**84%**	**0.63**	**48%**	**0.71**

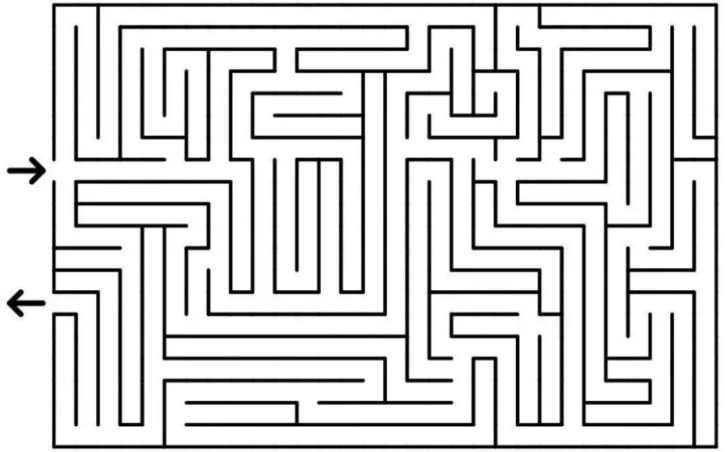
Zero-shot Real-world

Brief Summary of Navigation

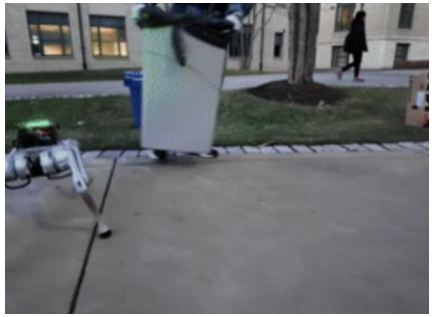
- **Need Map?**

➢ Map-based methods:
  - include all previous SLAM methods;
  - SLAM+semantic (*e.g.* CLIP)

➢ Map-free methods:
  - seems not so reliable, while leaving room for research



- More autonomous & intelligent agent: **System or End-to-end?**
- Navigation planning is more crucial than **embodiment**, except for some special requirements



Agile But Safe, RSS24

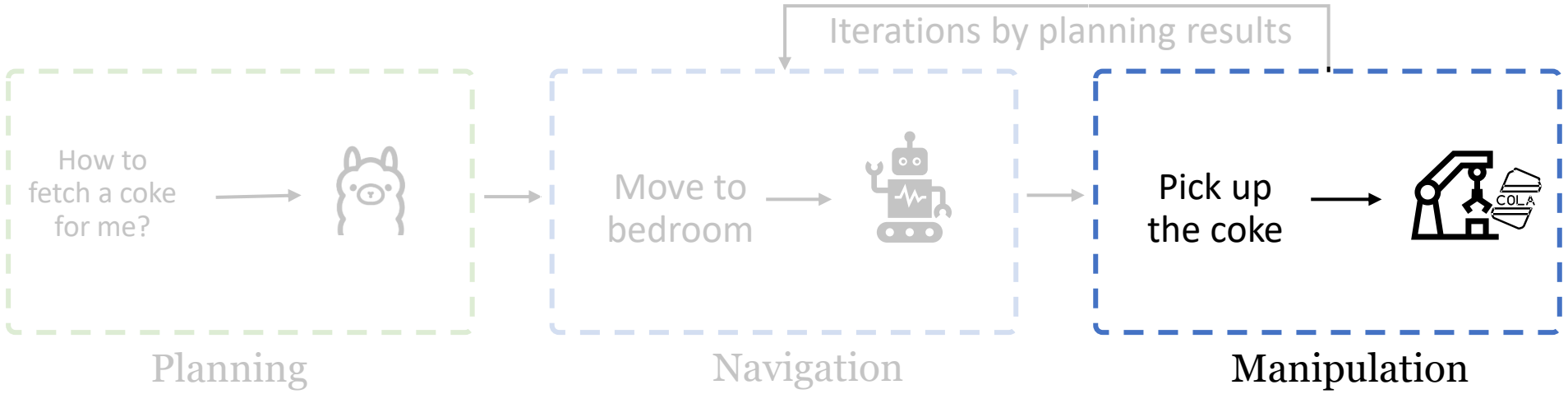
Animal Imitator, CoRL-W23

Robot Parkour, CoRL23

3 Manipulation: Vision Representation for Robotics

GraspNet: Pick up any objects with any shapes

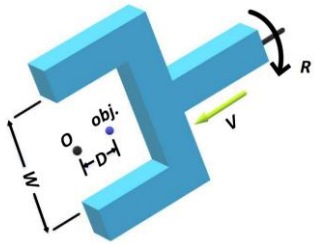
AnyGrasp: a data-driven approach for 6DoF pose estimation



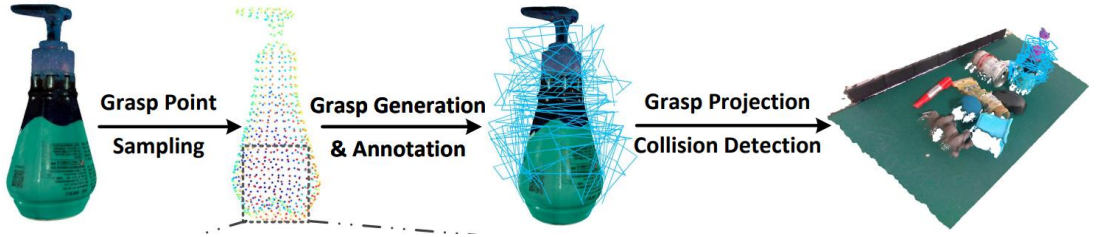
GraspNet: Scene-level General Grasping

Problem: Scene-level grasping

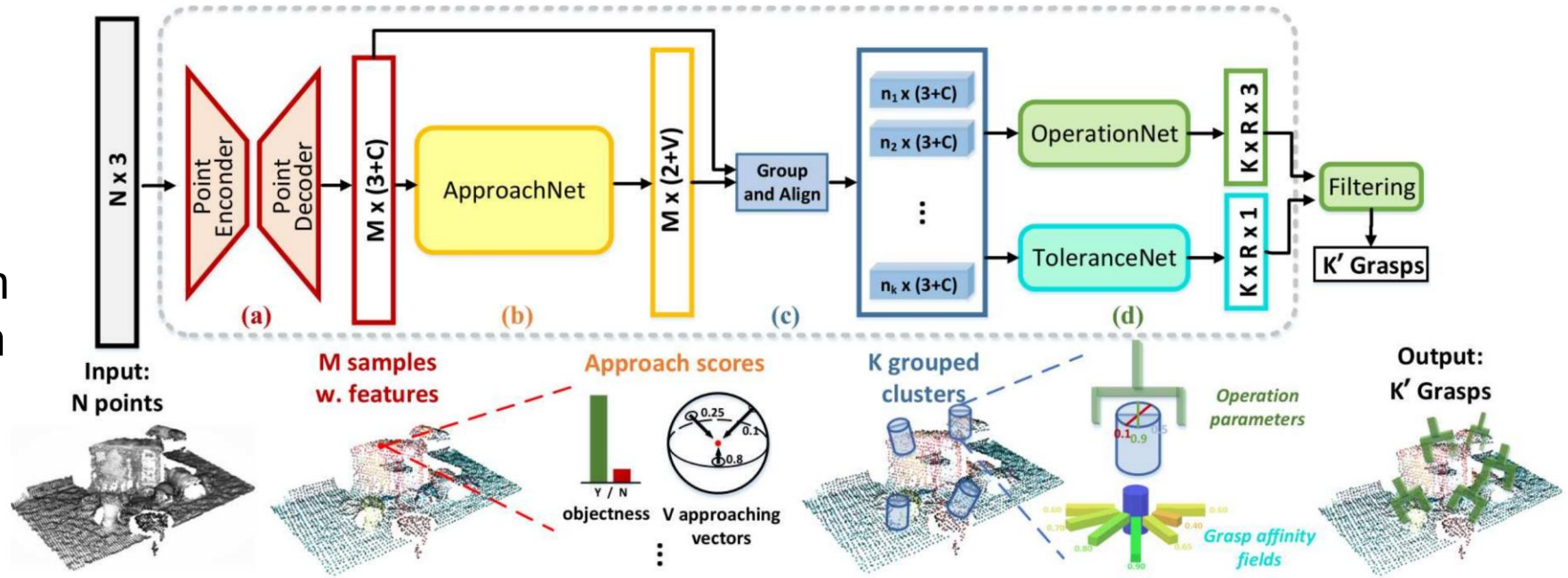
input: RGBD output: poses for points 6DoF pose



Data annotation: 1B+ Poses on 144 objects



Model:
End-to-end with modular design

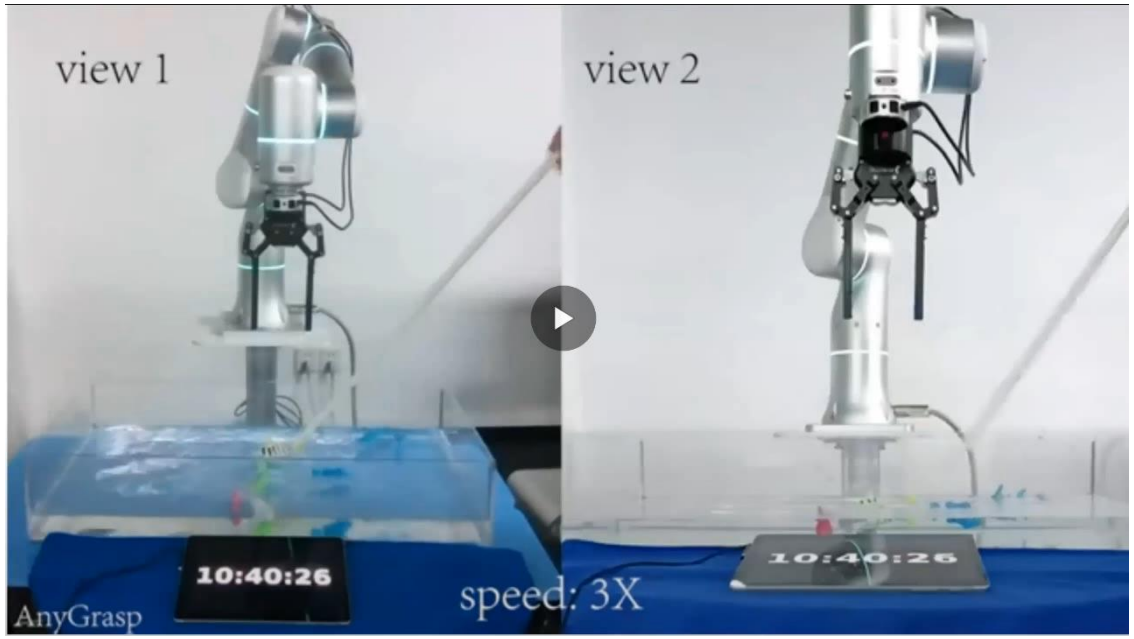


GraspNet-1Billion: A Large-Scale Benchmark for General Object Grasping, Fang *et al*, CVPR2020

Grasp Any Objects with 93%+ Success Rates

- Surpassing previous sota by improvement of 21%+

robot picking 300+ unseen objects



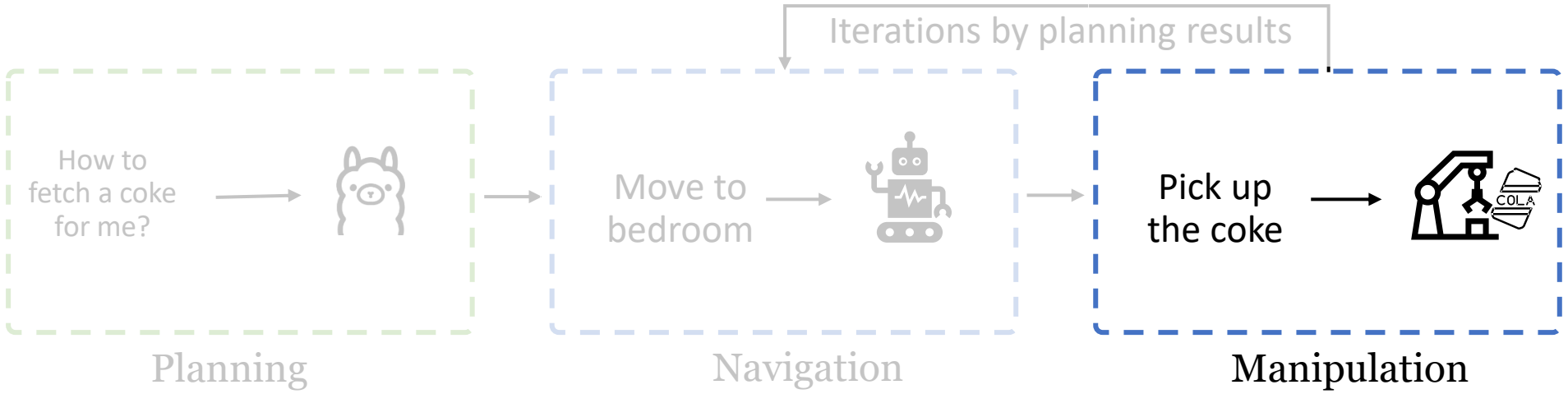
Dynamic fish catching

AnyGrasp: Robust and Efficient Grasp Perception in Spatial and Temporal Domain , Fang *et al*, T-RO2023

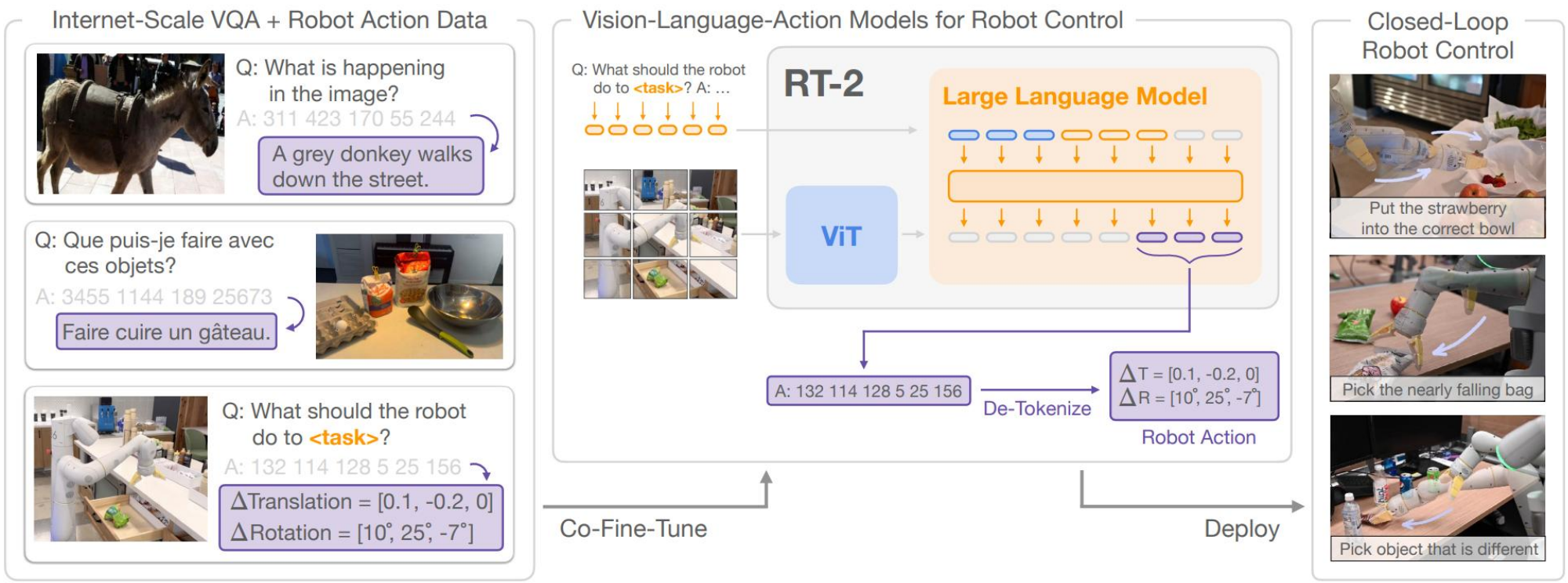
By Boshen Xu

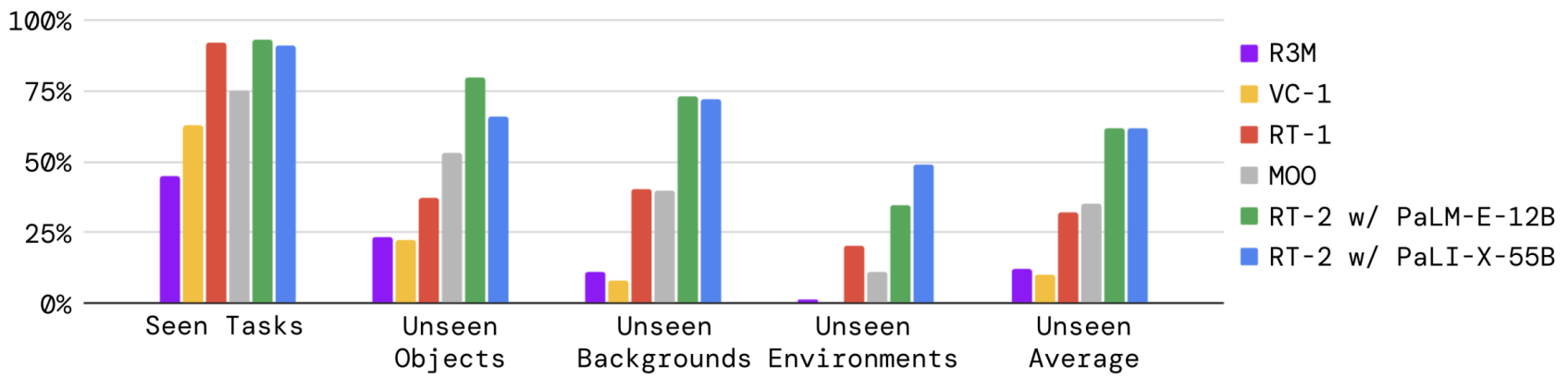
3 Manipulation: Vision-Language-Action Model

Google Robotics / Deepmind: RT-1, RT-2, RT-X, …



RT-2: Large VLA Model for Robot Control



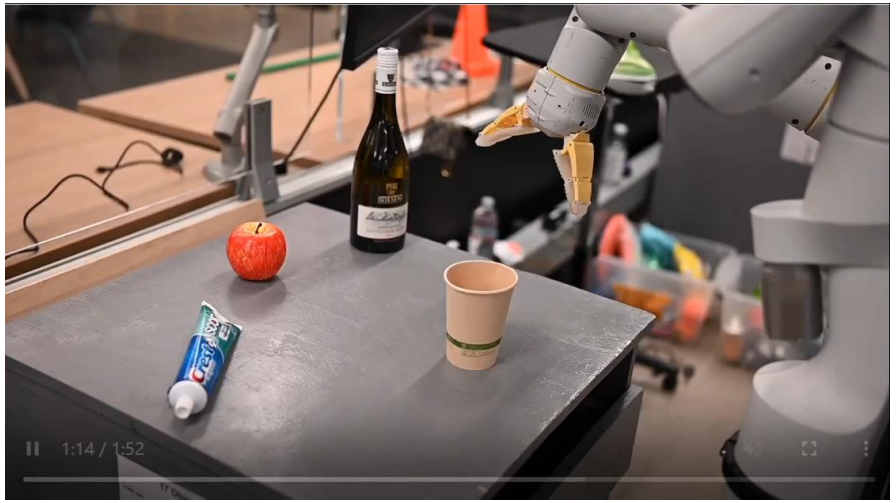


Results Seems Promising, However…

failure case

unsafe

8×



hallucination

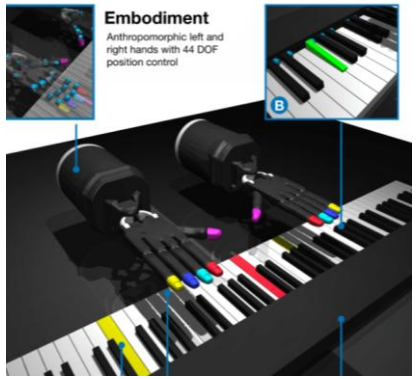
RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control, Google DeepMind, 2023

Brief Summary of Manipulation

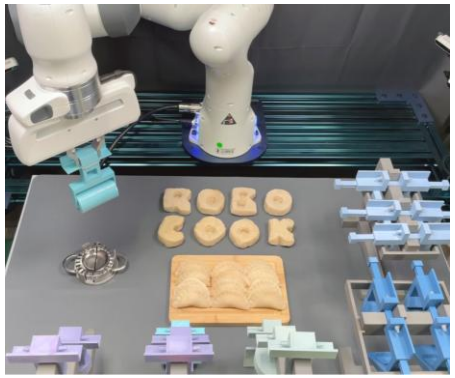
- Mainstream research: picking up and placing down of grippers (pick up and place).
  - ➢ End-to-End 2D-VLA model seems a bad idea, vision-based solutions (separating robot) are brilliant.
- What makes pick up & other directions different? Data-driven approach is easier for pick up.
  - ➢ In the future, they may all be data-driven as long as the data collection becomes easier:

Robot Synesthesia, ICRA24

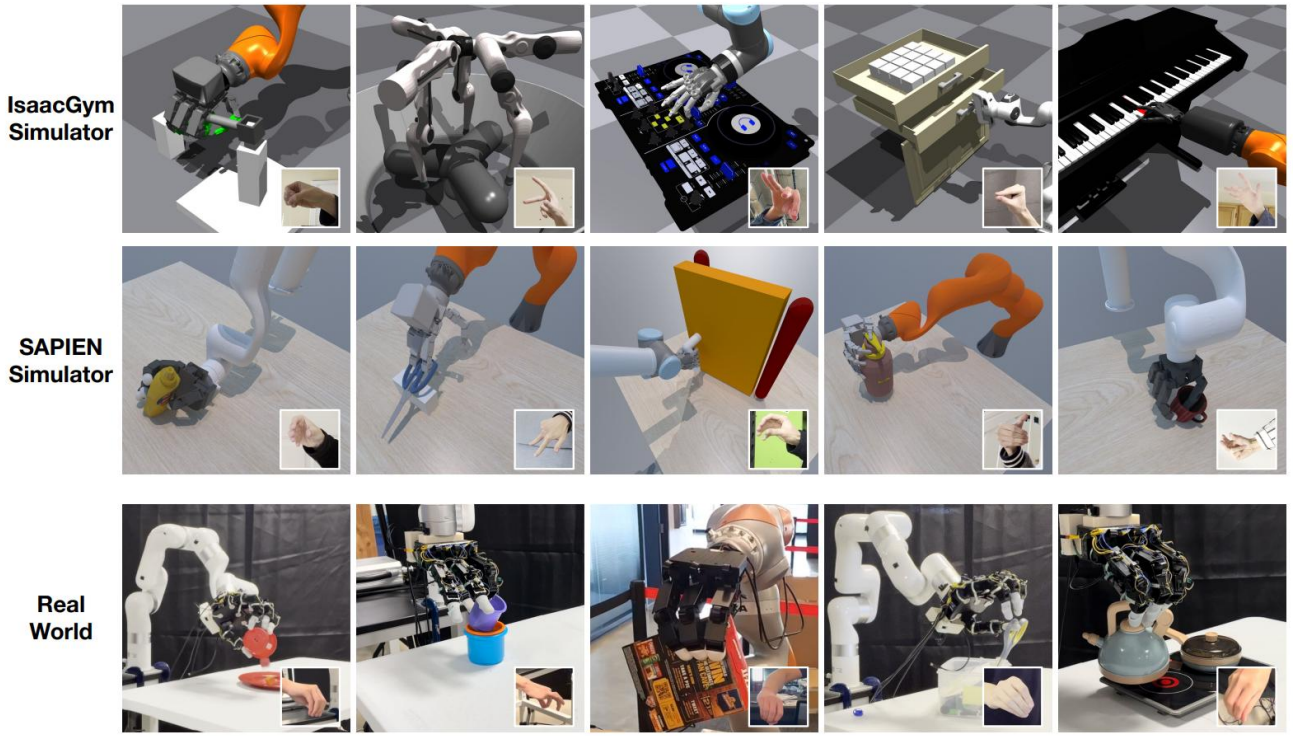
Eureka, ICLR24



RoboPianist, CoRL23



RoboCook, CoRL23



AnyTeleop, RSS23

Summary of Embodied AI

- LLM is the key of **planning**, while grounding LLM in the real-world is a key problem.
- **Navigation** steps towards autonomous driving agents.
- Data-driven approach works extremely well for pick and place in **manipulation**.

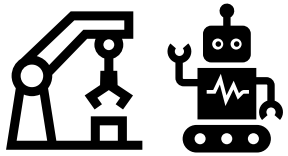
- Key Challenge for Data-driven Embodied AI:

➢ Q: For what type of **robot**, by **what way** to collect what kind of **data**, to train what kind of **model**?

➢ A: For **chosen robot**, collect **data** for **certain type abstract task** to train a **3D-vision-based model**.

Controlling body is earlier to learn than thinking & speaking with language.

for human/animals



Controlling body is harder than thinking & speaking with language.

for robot


Thank You for Listening!
In the pursuit of generally intelligent machines...
Build decision-making agents that interact with the world and improve themselves over time