



Robotics in the age of Generative AI

Vincent Vanhoucke
Distinguished Scientist, Google DeepMind

LLMs



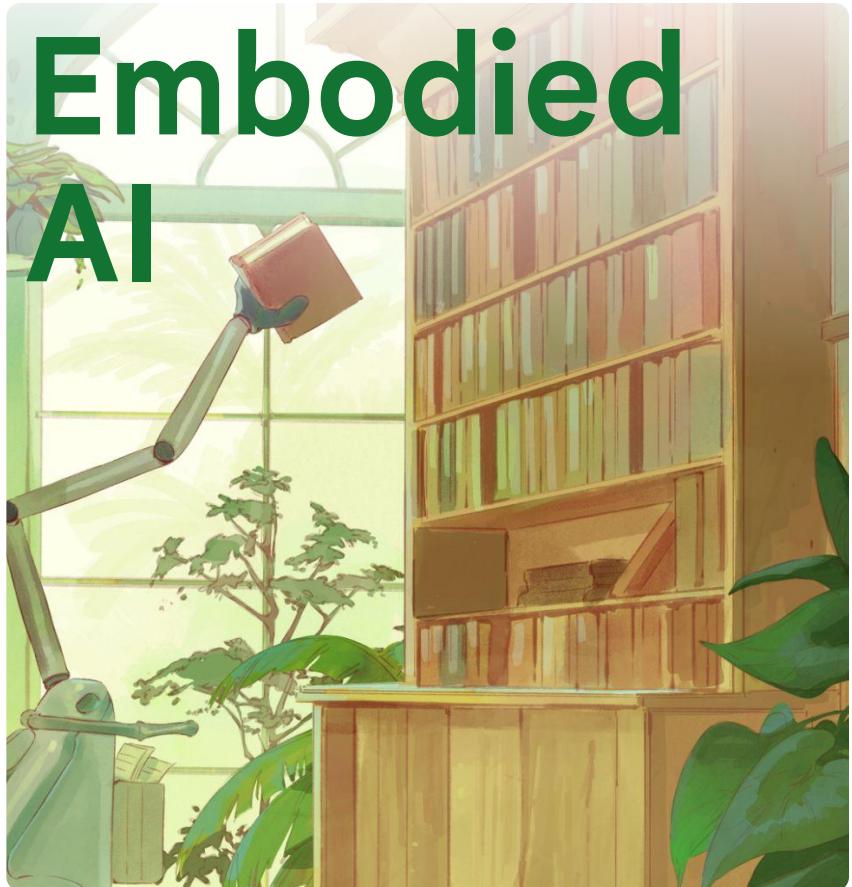
Embodied AI



LLMs



Embodied AI





I am a mobile robot on wheels with an arm. How would I go about making coffee? 



Making coffee with a mobile robot on wheels with an arm is a challenging but potentially rewarding task! Here's a breakdown of the possible steps:

1. Planning and Preparation:

- **Coffee brewing method:** Decide on the brewing method you want to use. Pour-over, capsule machines, and pod-based systems are generally easier for robots to handle compared to French press or stovetop espresso makers.
- **Ingredients and equipment:** Ensure you have all the necessary ingredients like coffee beans/grounds, water, milk (if desired), filters, cups, etc. You'll also need equipment like a grinder (for whole beans), kettle, brewing machine, and milk frother (if using milk).
- **Robot arm programming:** Program your robot arm's movements for each step of the coffee-making process. This includes picking up and manipulating objects, pouring liquids, and operating buttons or levers.

I spilled my drink, can you help?

LLM

- “Find a cleaner”
- “Find a sponge”
- “Go to the trash can”
- “Pick up the sponge”
- “Try using the vacuum”

I spilled my drink, can you help?

LLM

“Find a cleaner”
“Find a sponge”
“Go to the trash can”
“Pick up the sponge”
“Try using the vacuum”

Value Functions

“Find a cleaner”
“Find a sponge”
“Go to the trash can”
“Pick up the sponge”
“Try using the vacuum”

I spilled my drink, can you help?

LLM

“Find a cleaner”
“Find a sponge”
“Go to the trash can”
“Pick up the sponge”
“Try using the vacuum”

Value Functions

“Find a cleaner”
“Find a sponge”
“Go to the trash can”
“Pick up the sponge”
“Try using the vacuum”

SayCan

“Find a cleaner”
“Find a sponge”
“Go to the trash can”
“Pick up the sponge”
“Try using the vacuum”



I spilled my drink, can you help?

LLM

"Find a cleaner"
"Find a sponge"
"Go to the trash can"
"Pick up the sponge"
"Try using the vacuum"

Value Functions

"Find a cleaner"
"Find a sponge"
"Go to the trash can"
"Pick up teh sponge"
"Try using the vacuum"

SayCan

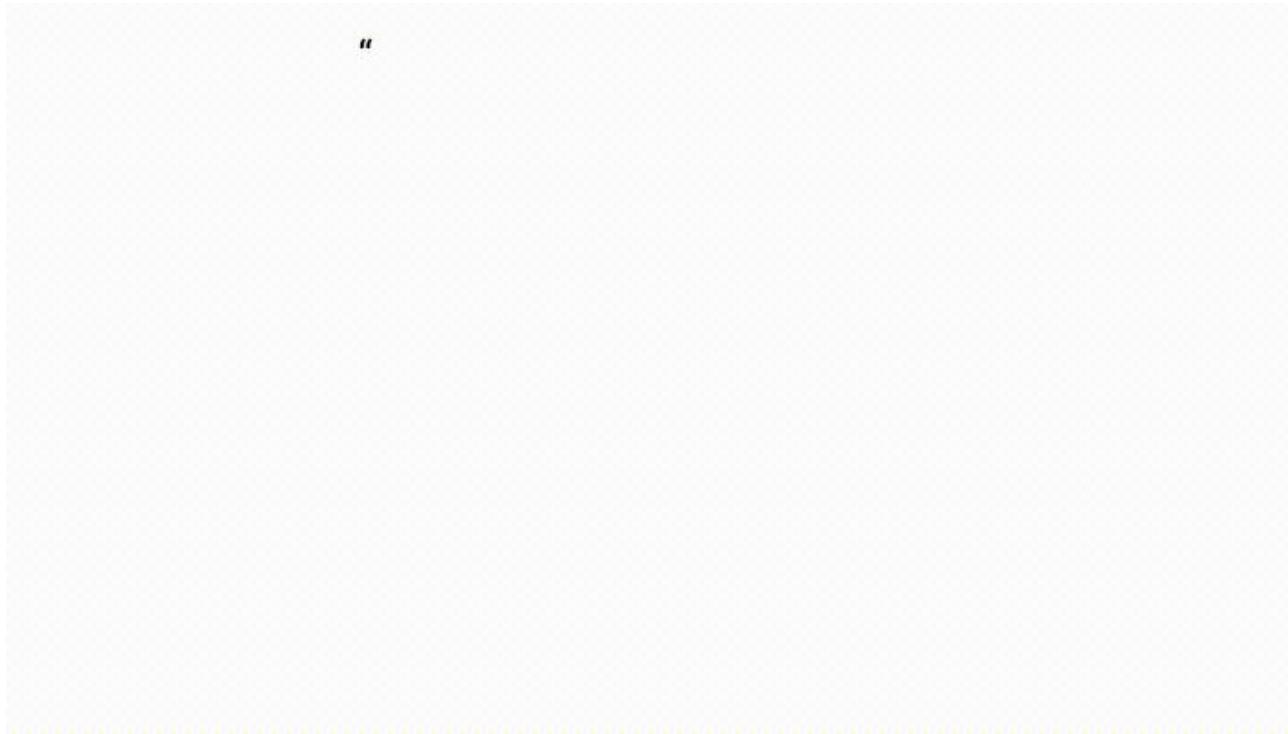
"Find a cleaner"
"Find a sponge"
"Go to the trash can"
"Pick up the sponge"
"Try using the vacuum"



I would:

1. Find a sponge
2. Pick up the sponge
3. Come to you
4. Put down the sponge
5. Done

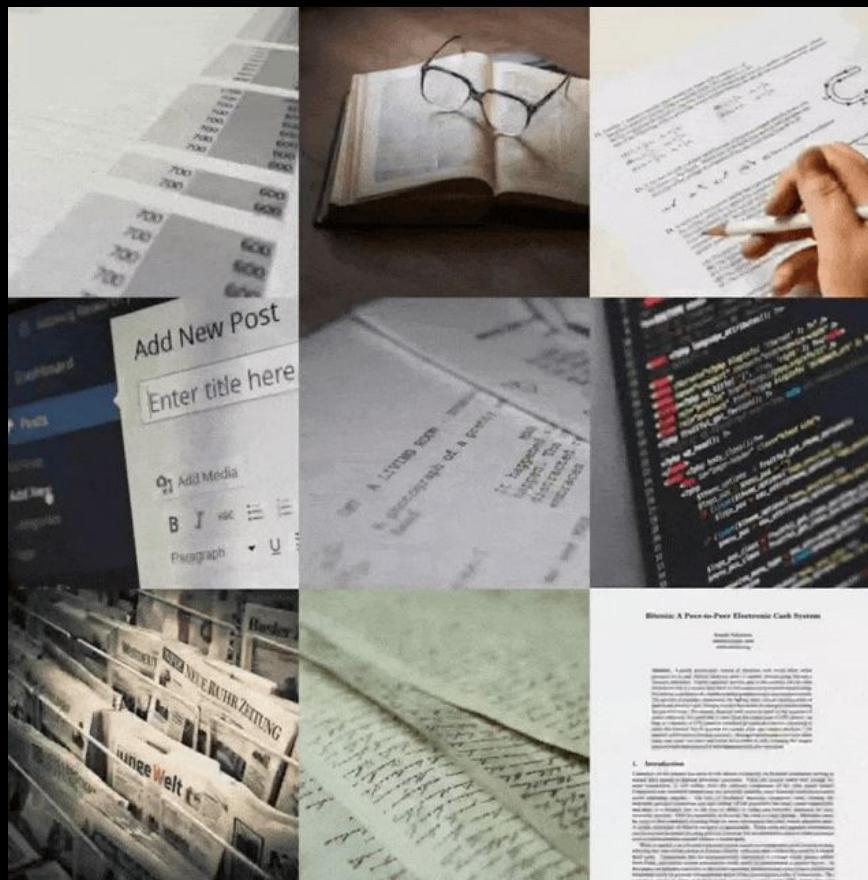
SayCan











Inner Monologue

SayCan



Language
Model



Robot Value
Functions

Inner Monologue



Robot



Human



Scene
Descriptor



Language
Model



Success
Detector

Inner Monologue



Robots that ask for help

Robot Planning & Human Interaction

Human



Place the bowl in the microwave, please.



Robot

Which one, plastic or metal?



Human



The plastic one, please.



Uncertainty Alignment with KnowNo

Environment Context

There is a microwave, a landfill bin, a recycling bin, and a compost bin.



Robot Observations

Observations: I see a metal bowl and a plastic bowl on the counter.



LLM Next Step Prediction with Confidence

Possible next steps:

- 0.44 - Put plastic bowl in microwave.
- 0.41 - Put metal bowl in microwave.
- 0.03 - Put metal bowl in landfill bin
- 0.08 - Put plastic bowl in recycling bin.



Prediction Set from Conformal Prediction

Conformal prediction threshold: 0.21

Steps with scores above threshold:

- 0.44 - Put plastic bowl in microwave.
- 0.41 - Put metal bowl in microwave.



Trigger Human Help

Prediction size 2 > 1 → ask for help.



LLM Generates Question

Question: Which one, plastic or metal?

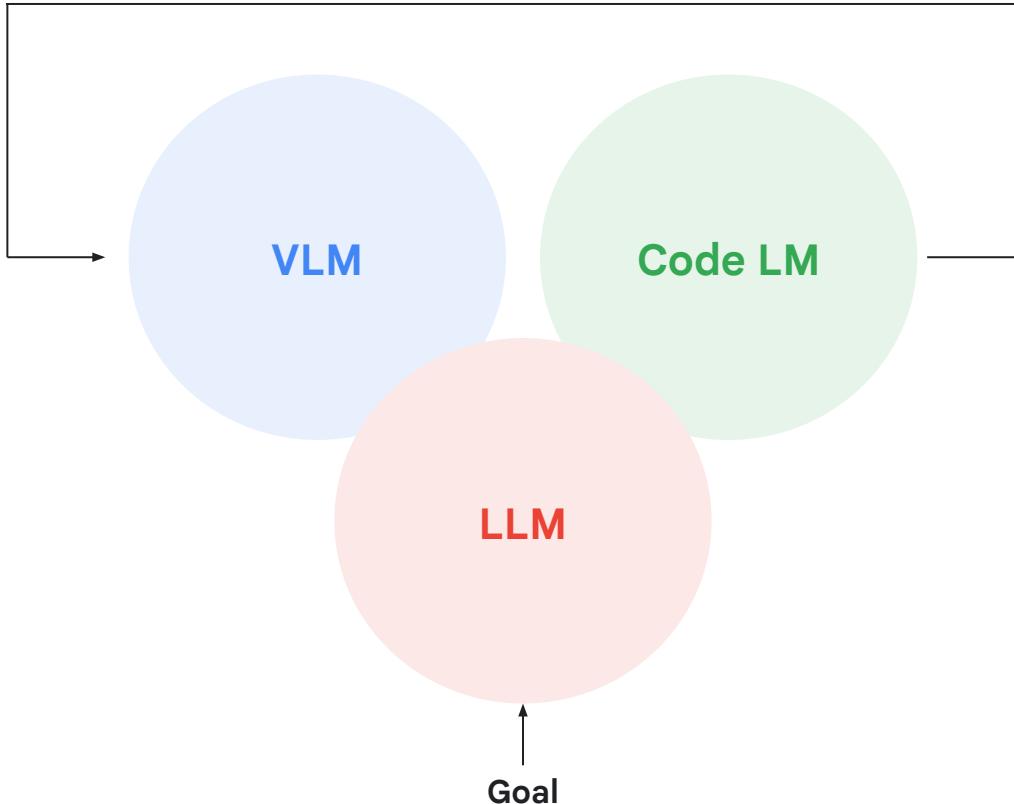


Language driven exploration

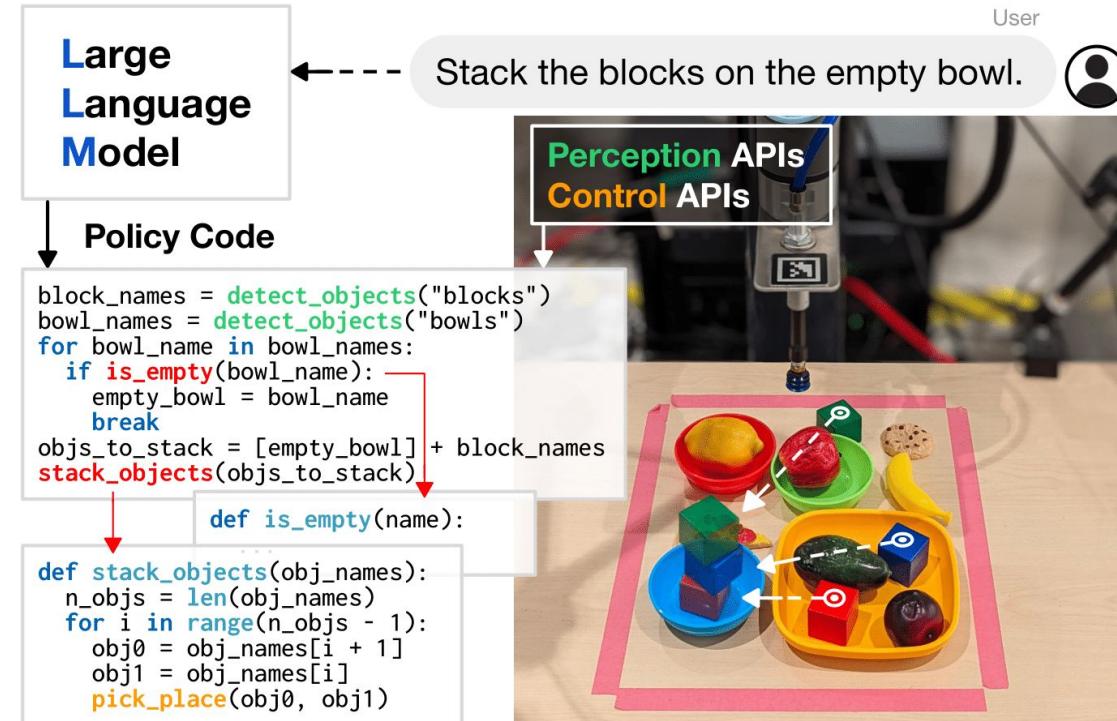
auto-rt.github.io



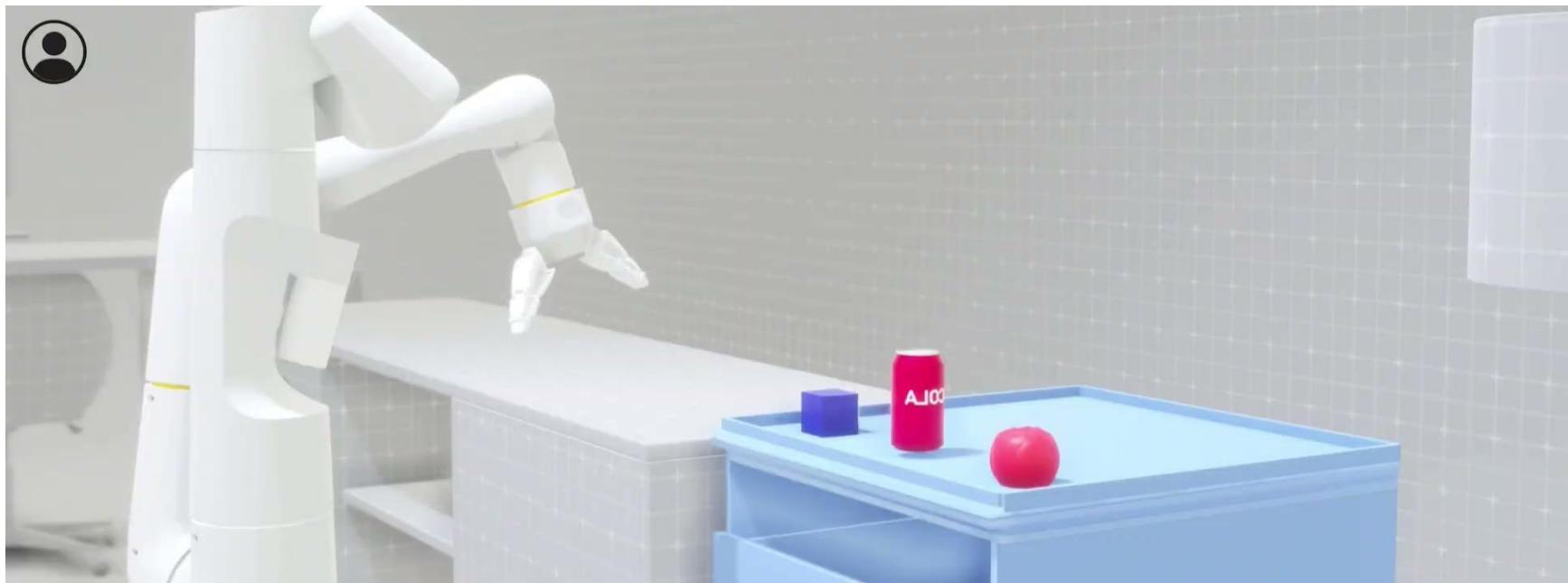




Code as policies



Non-expert teaching

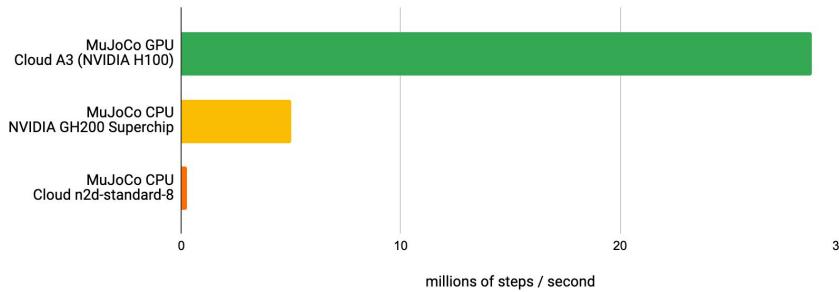


Non-expert teaching



MuJoCo 3

Accelerated physics with MuJoco XLA in JAX (MJX)



Language Model Predictive Control

Fast Adaptation via In-Context Learning

User Language Inputs

please sit down like a dog

LLM Robot Code - Rewards & Costs

```
# Reward function code.
import numpy as np
set_torso(0.2, np.deg2rad(-45))
```



no, tilt the other way

```
set_torso(0.2, np.deg2rad(45))
```



great, now give me a high five!

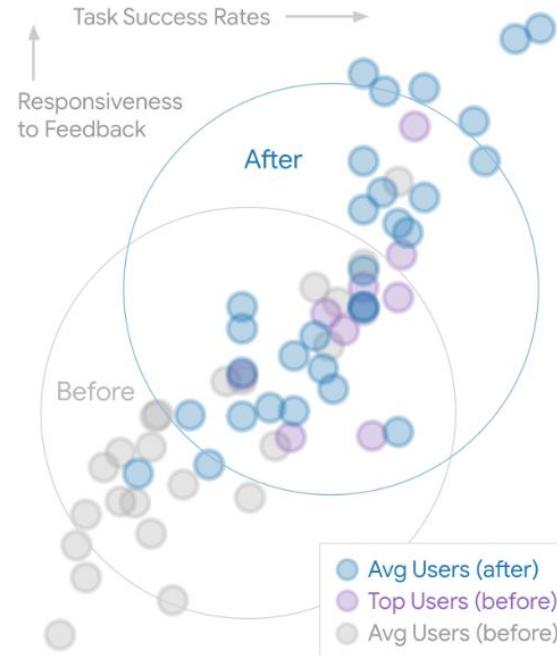
```
set_torso(0.2, np.deg2rad(-45))
set_feet('front_left', 0.3, 0.1)
```



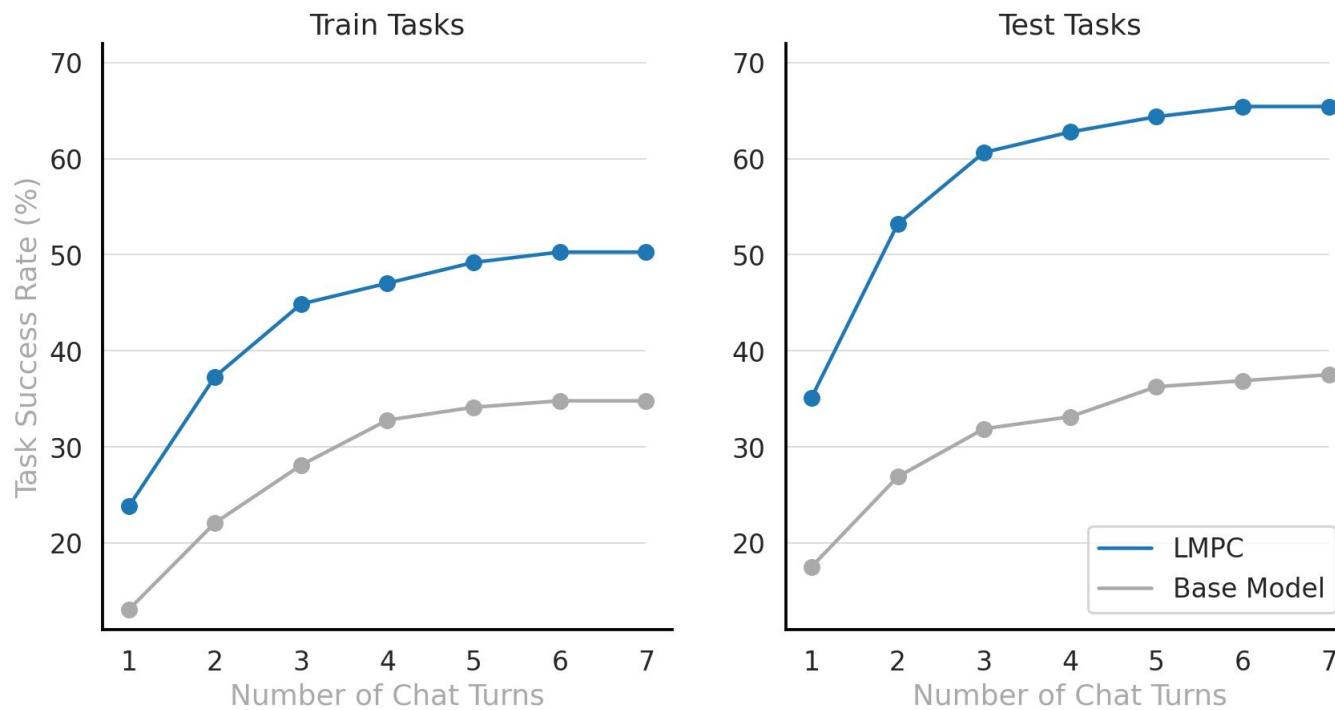
a bit too low, raise it higher

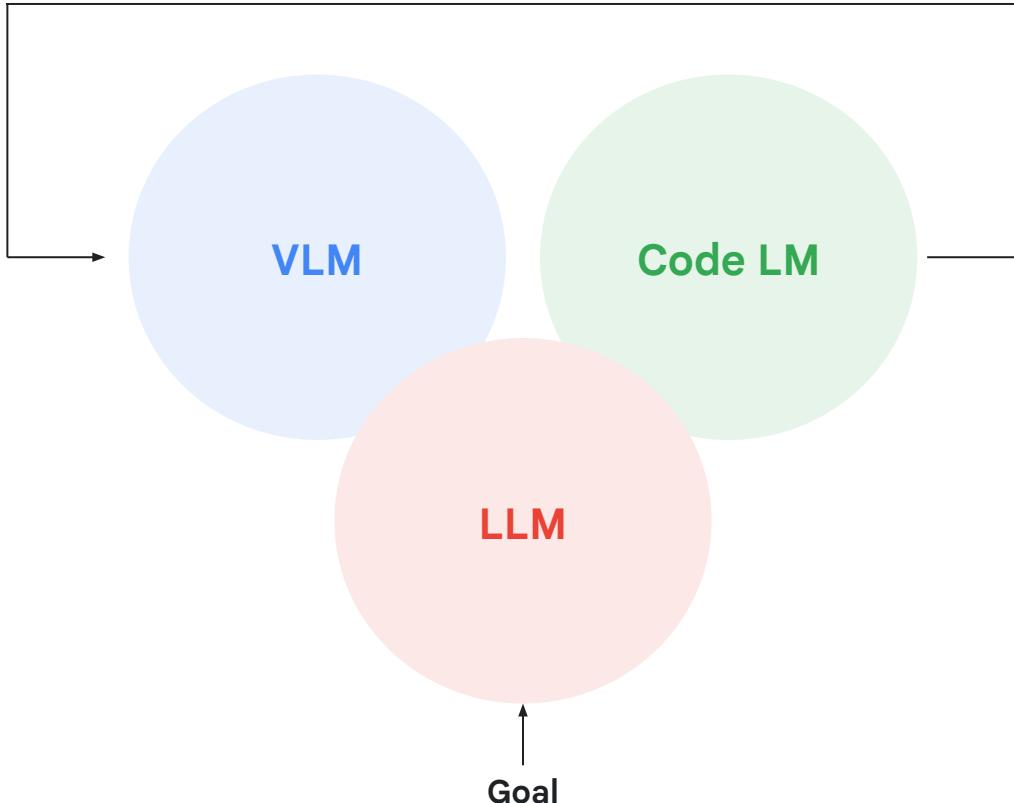


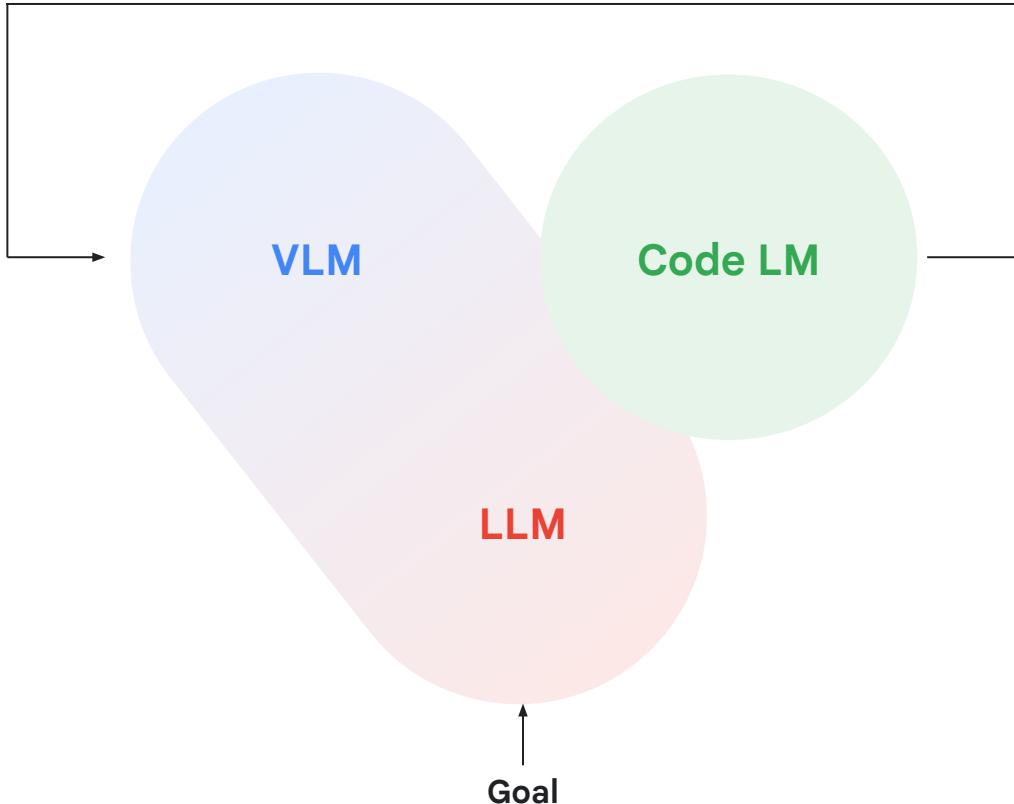
Slow Adaptation via Model Fine-Tuning



Language Model Predictive Control

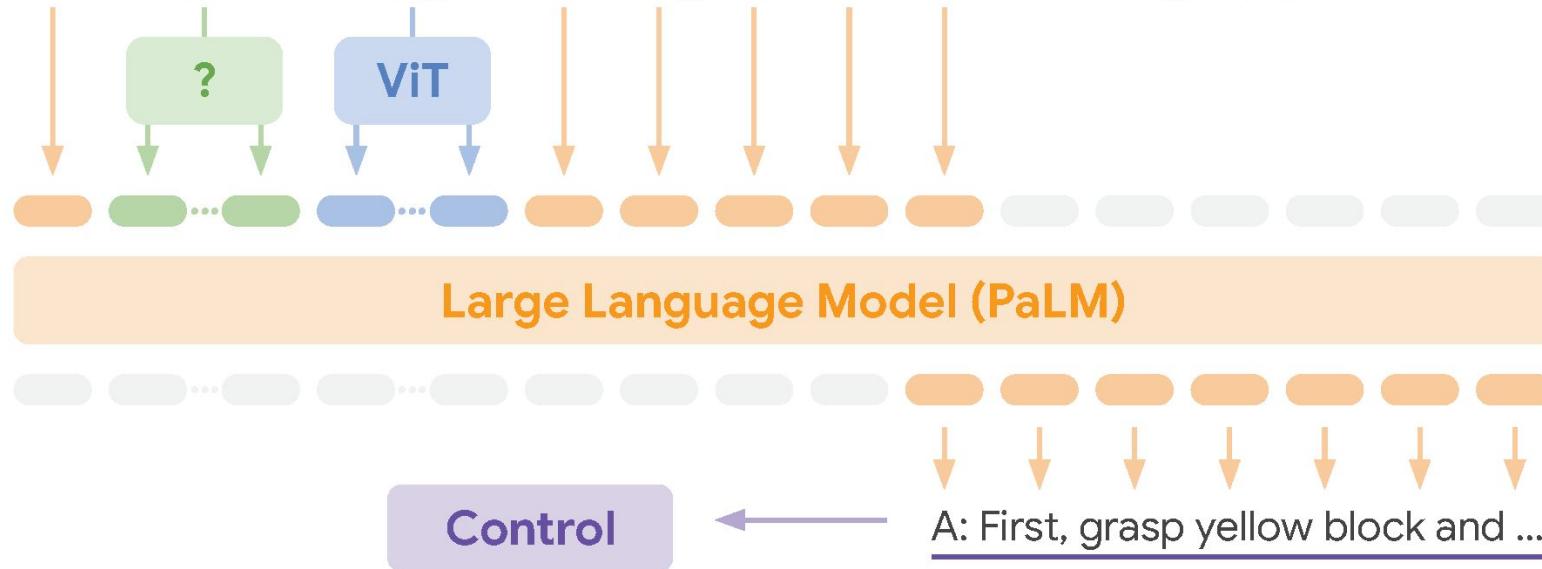






PaLM-E: An embodied multimodal language model

Given <emb> ... Q: How to grasp blue block? A: First, grasp yellow block



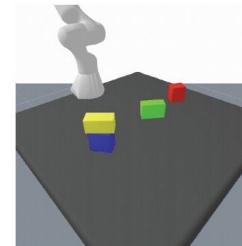
PaLM-E: An embodied multimodal language model

Mobile Manipulation



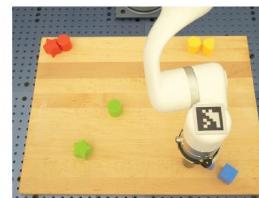
Human: Bring me the rice chips from the drawer. Robot: 1. Go to the drawers, 2. Open top drawer. I see 3. Pick the green rice chip bag from the drawer and place it on the counter.

Task and Motion Planning



Given Q: How to grasp blue block?
A: First grasp yellow block and place it on the table, then grasp the blue block.

Tabletop Manipulation



Given Task: Sort colors into corners.
Step 1. Push the green star to the bottom left.
Step 2. Push the green circle to the green star.

PaLM-E: An embodied multimodal language model

Visual Q&A, Captioning ...



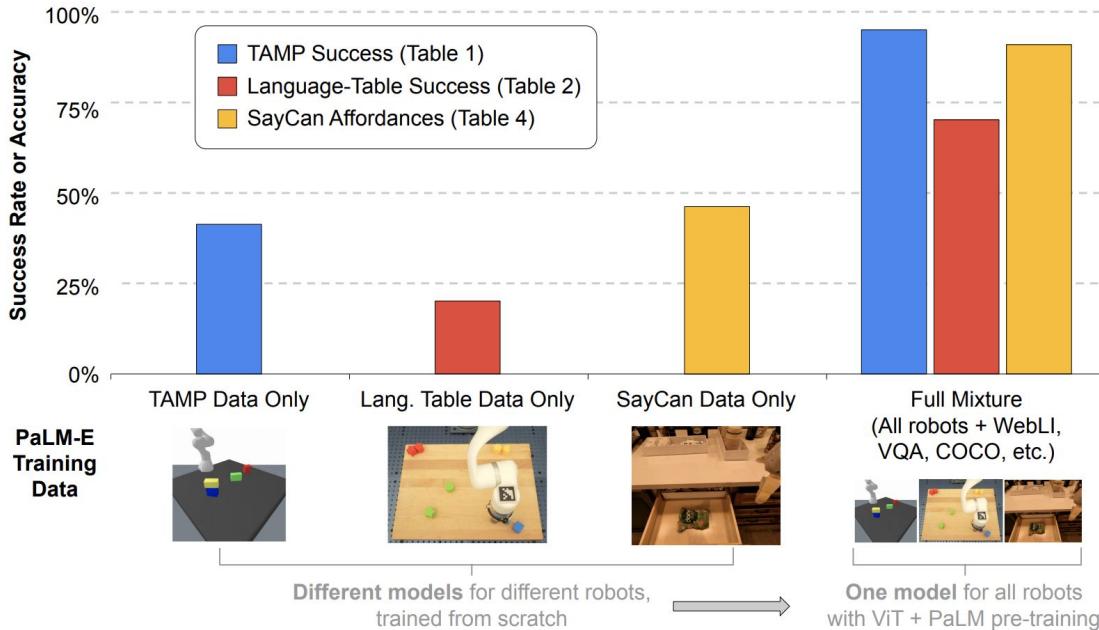
Given . Q: What's in the image? Answer in emojis.

A: 🍎 🍌 🍇 🍐 🍑 🍃 🍅



Describe the following :
A dog jumping over a hurdle at a dog show.

PaLM-E: An embodied multimodal language model



PaLM-E is massive
(562B params)

Yet we observe
positive transfer
across robots using
little robot data.

Video Language Planning

Video Language Planning

Video Generation

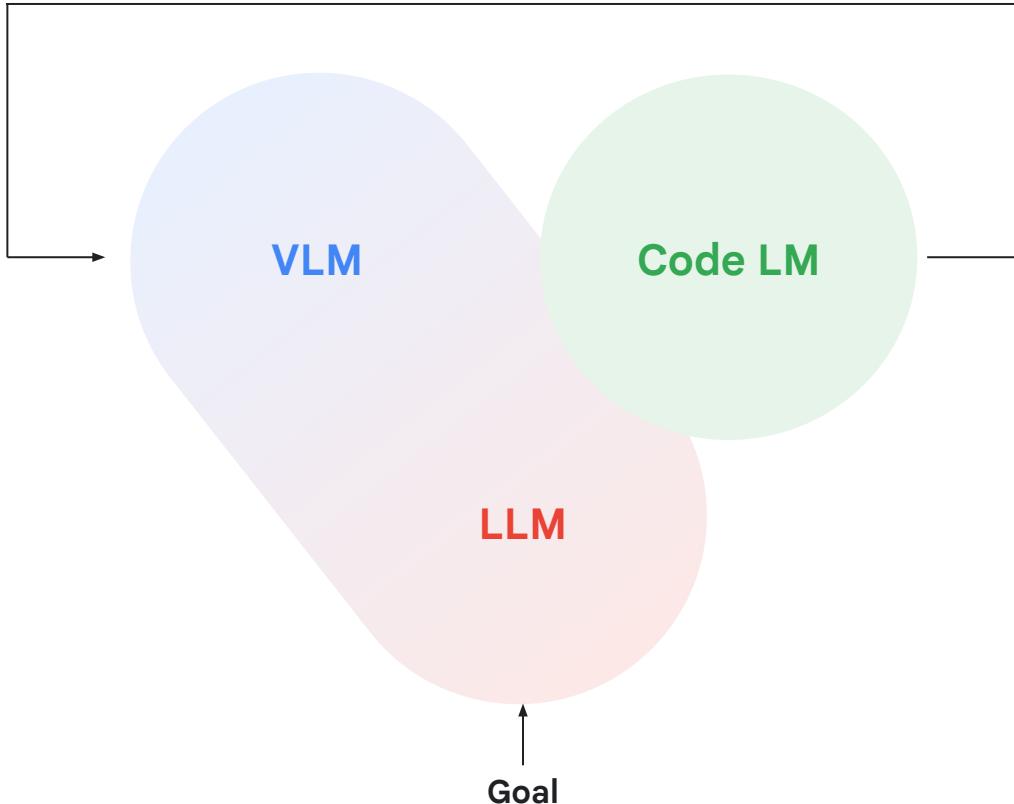
Real Execution

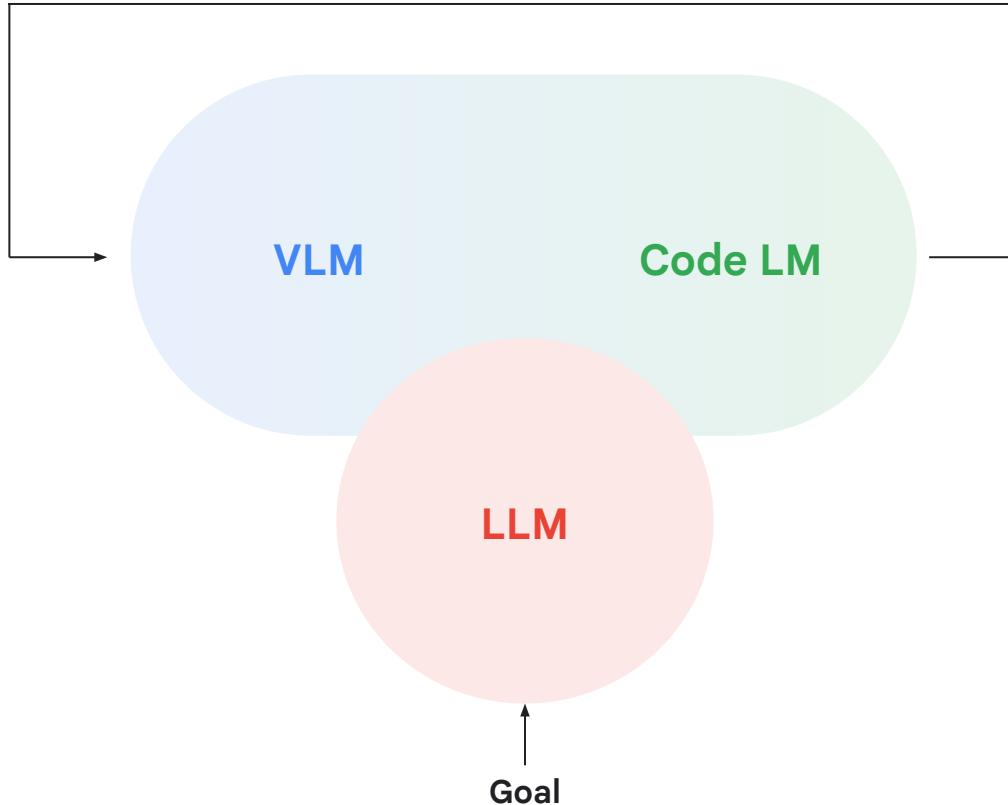
Put the fruits into
the top drawer



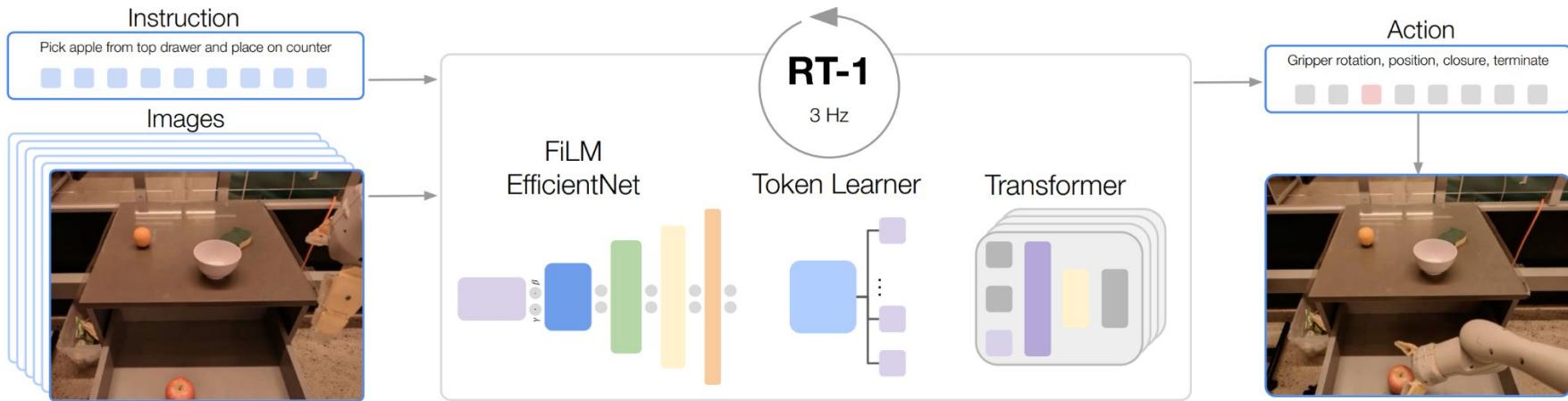
Execution







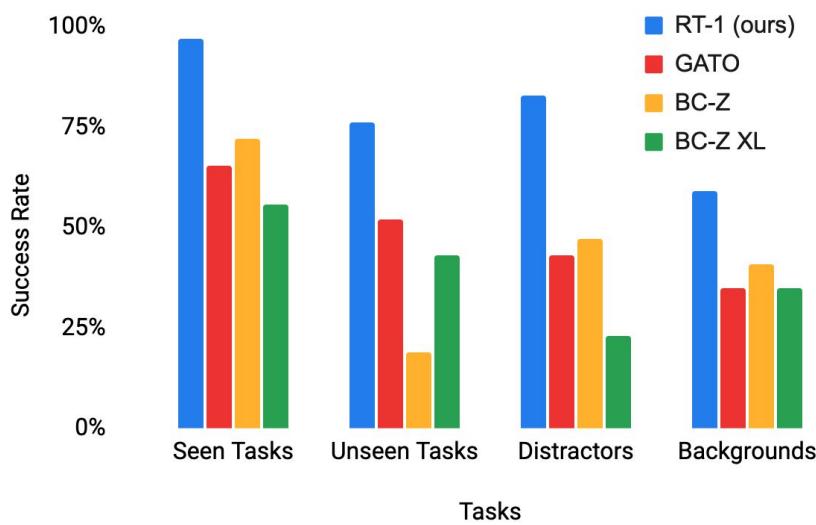
RT-1: Robotics Transformer v1



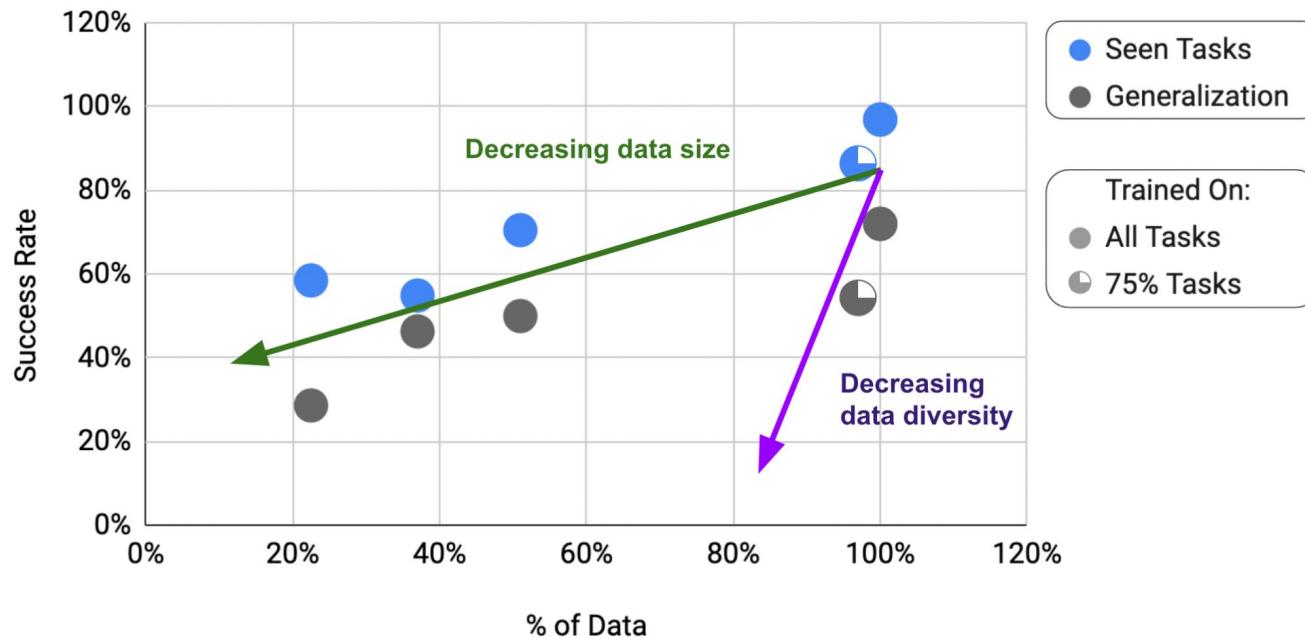
RT-1: Robotics Transformer v1

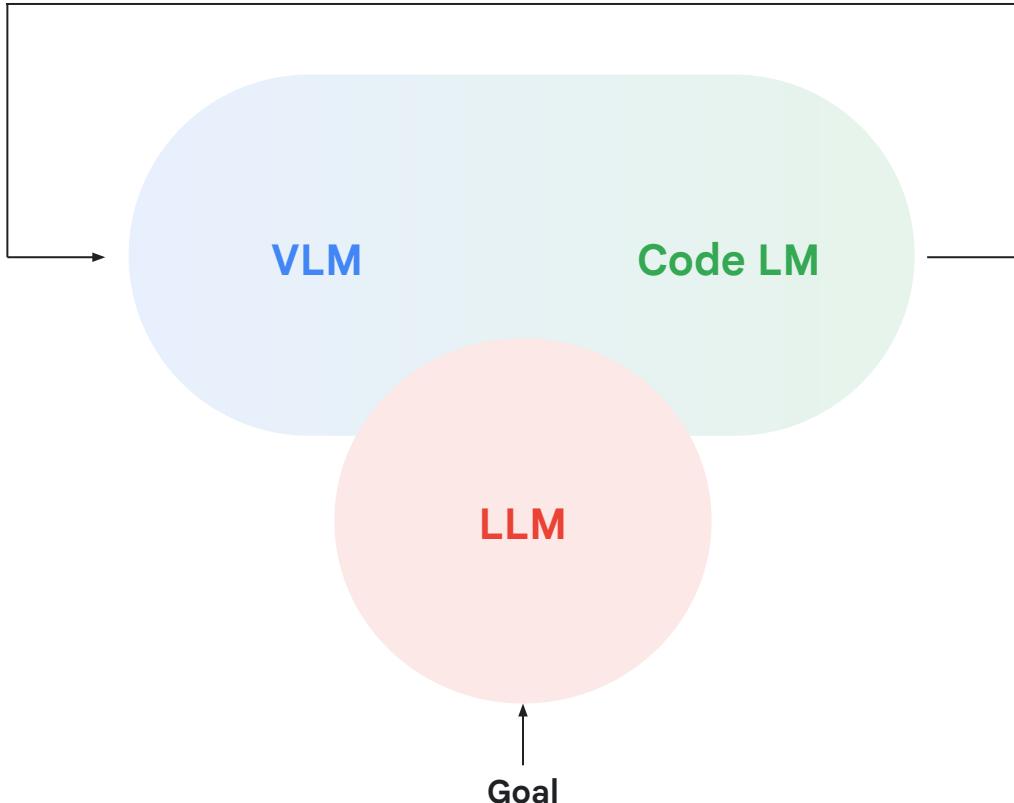
RT-1 is able to reach ~100% performance on seen tasks,

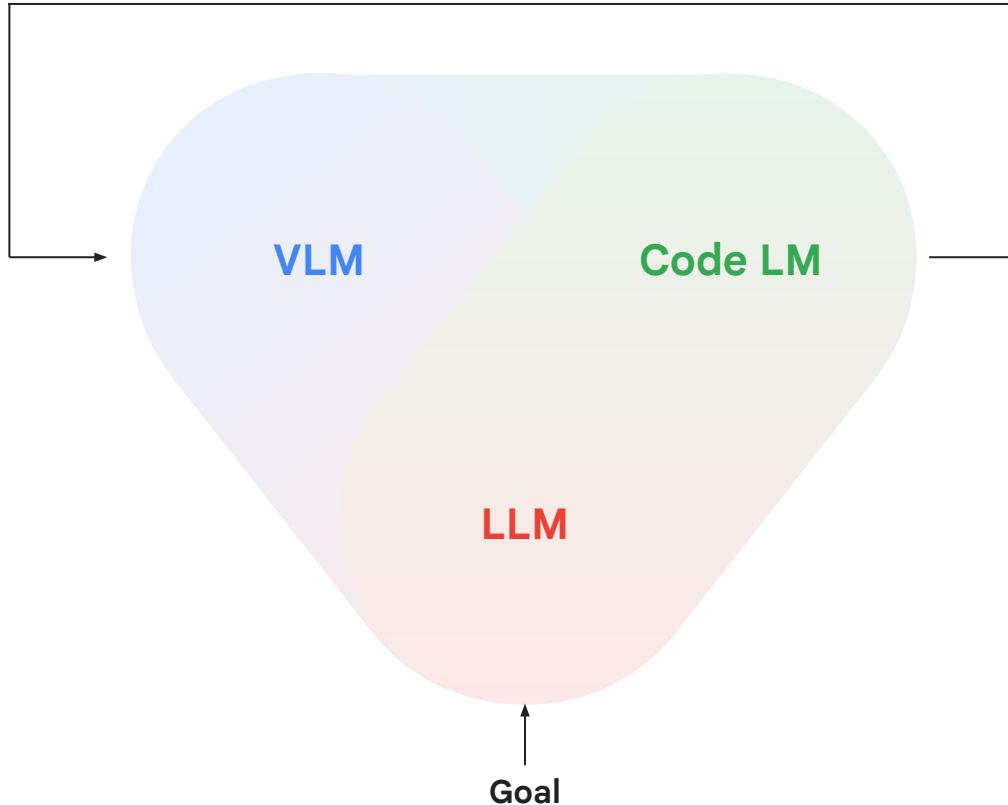
While maintaining better robustness to unseen variability.



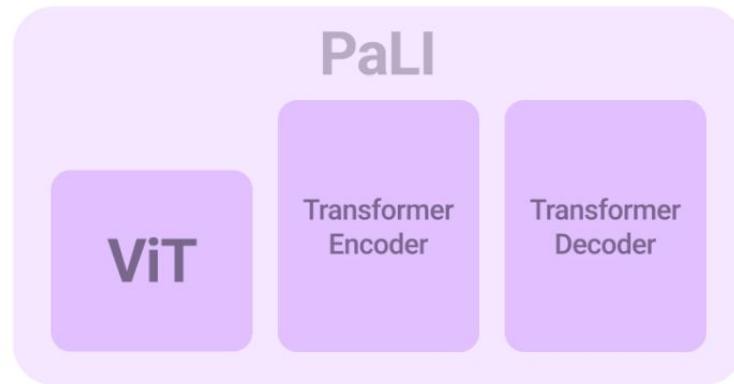
RT-1: Diversity is all you need?



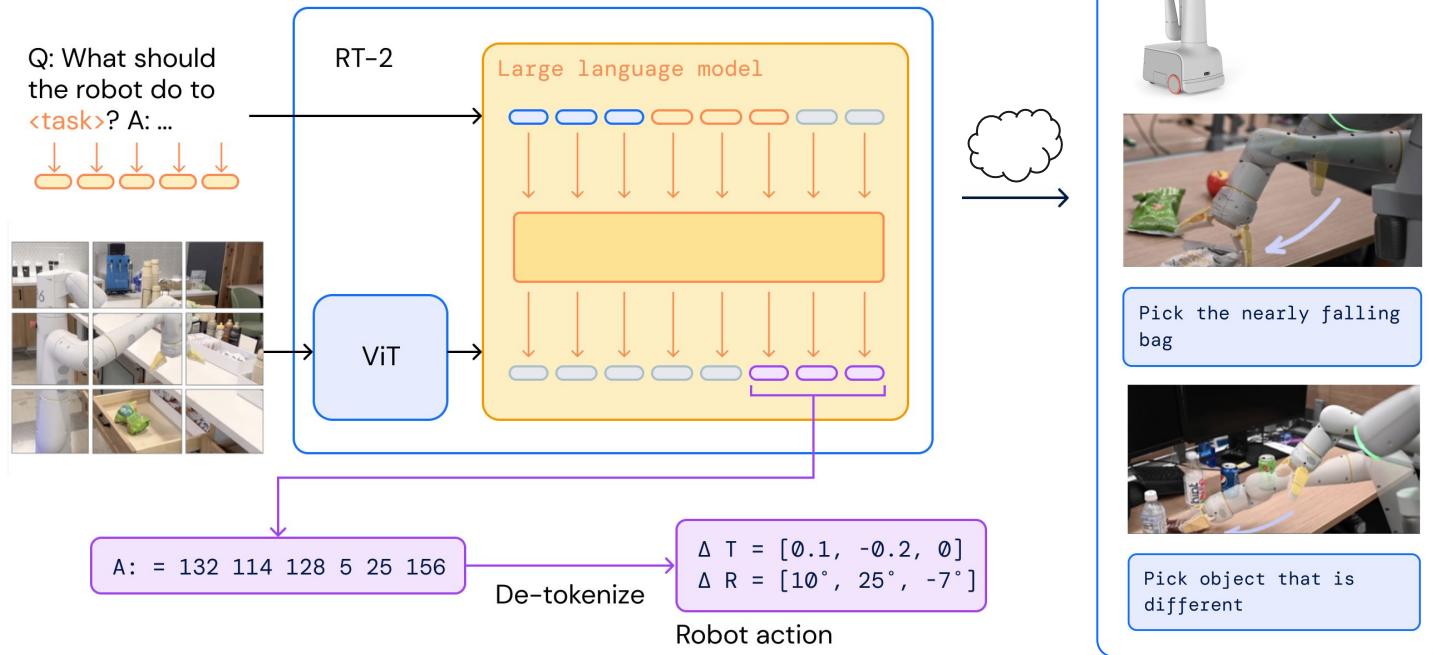




RT-2: Making VLMs ‘speak robot’



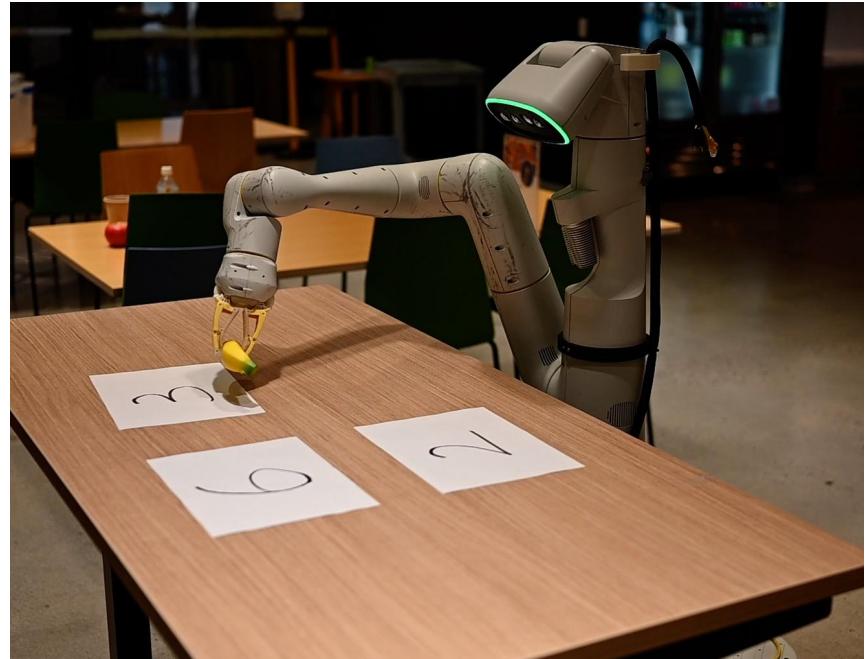
RT-2: Making VLMs ‘speak robot’



RT-2: Emergent transfer

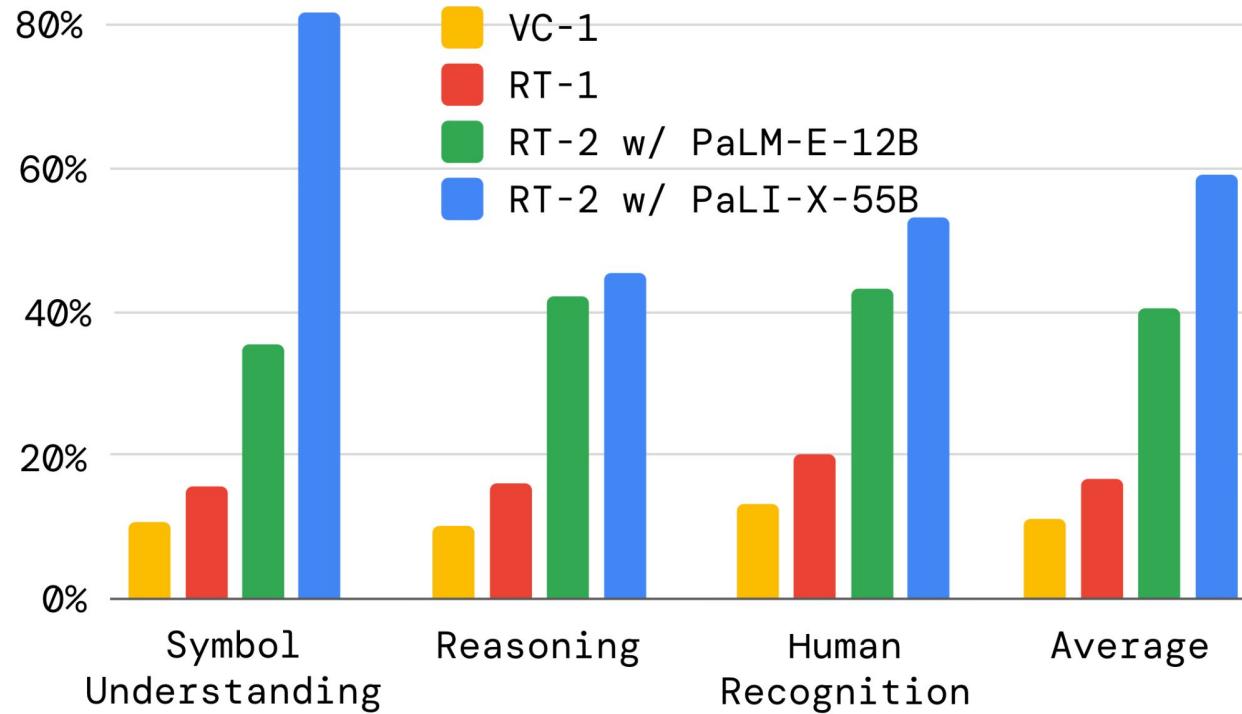


“Move coke can to Taylor Swift”

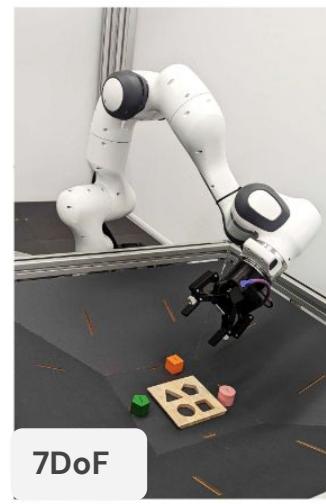
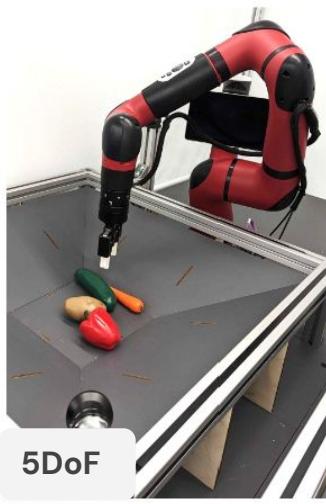
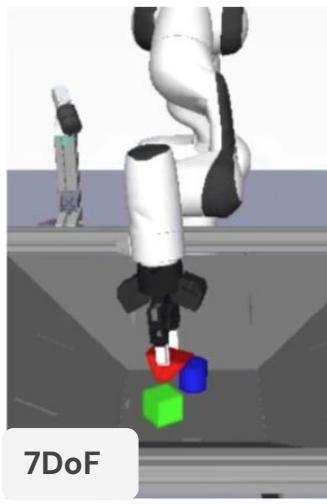
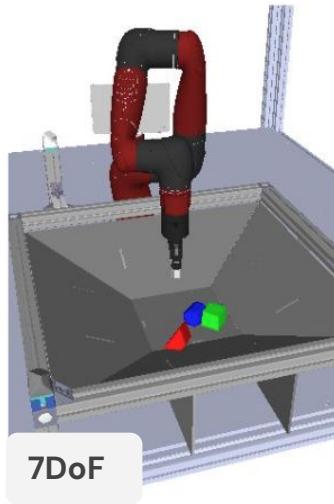


“Move the banana to the sum of $2 + 1$ ”

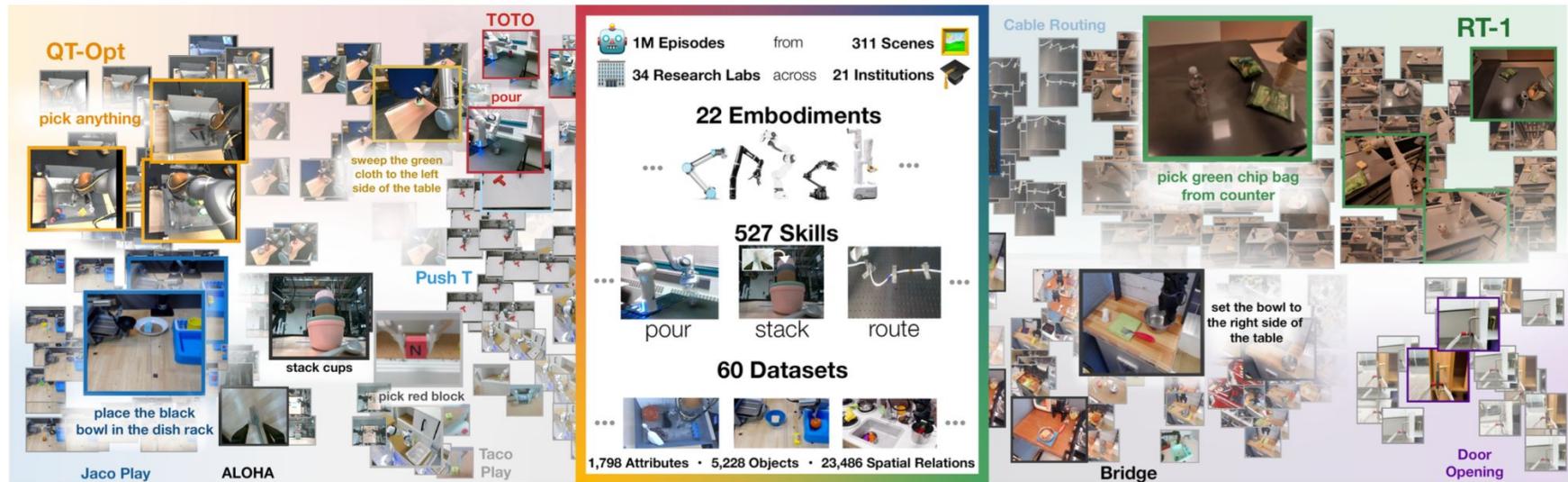
RT-2: Scaling

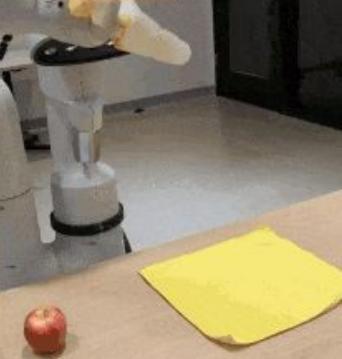
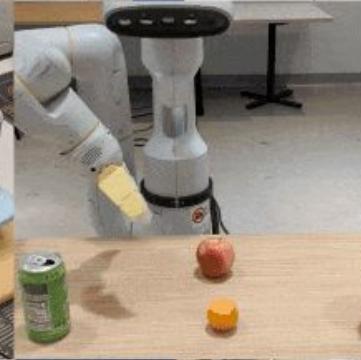
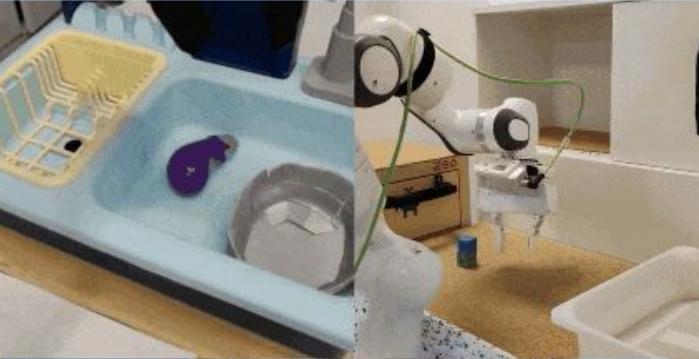
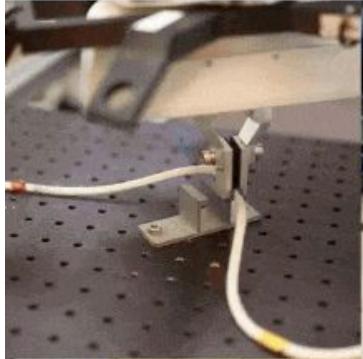


RoboCat: scaling across robots



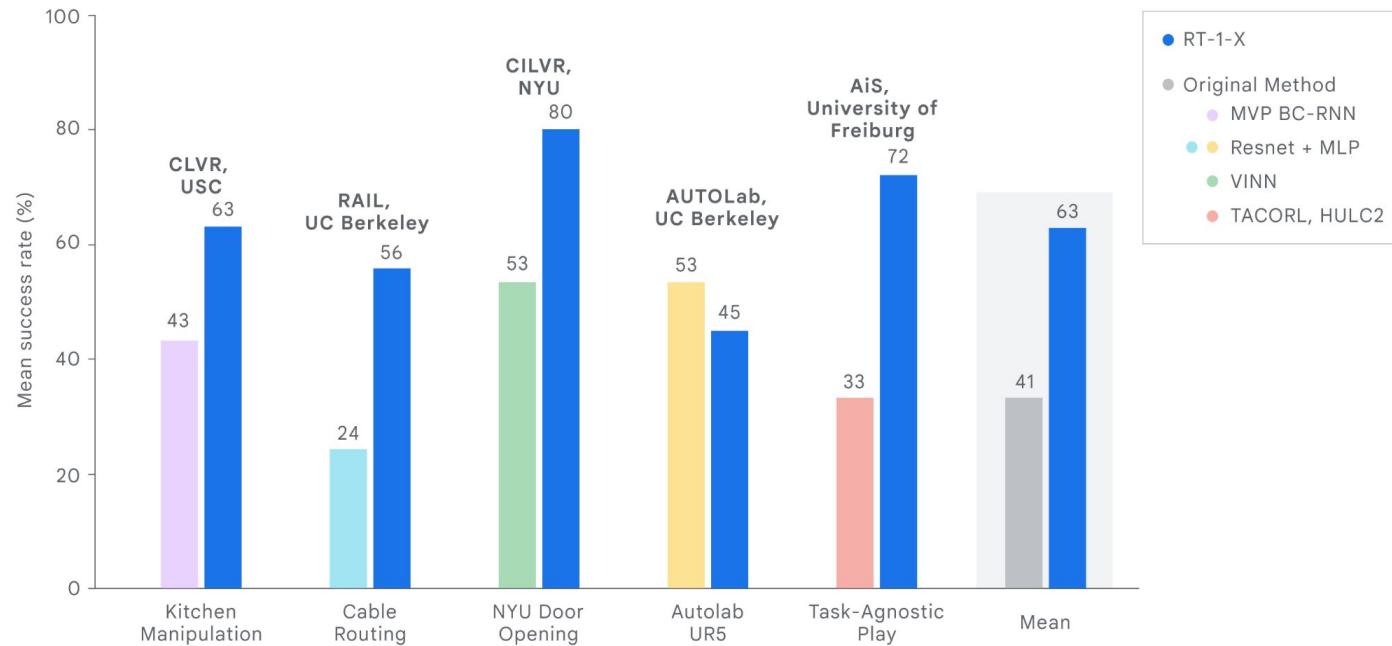
Open X-Embodiment: Open foundations for robotics





Open X-Embodiment: RT-1-X

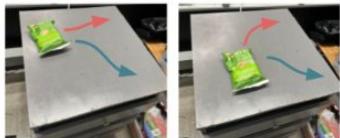
RT-1-X mean success rate is 50% higher than that of the Original Method



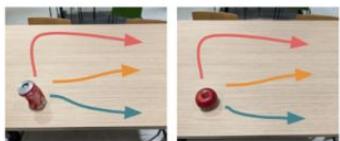
Open X-Embodiment: RT-2-X

(a) Absolute Motion

*move the chip bag to the
top / bottom right of the counter*



*move to top right /
right / bottom right*



(b) Object-Relative Motion

*move apple between coke and cup /
coke and sponge / cup and sponge*



(c) Preposition Alters Behavior

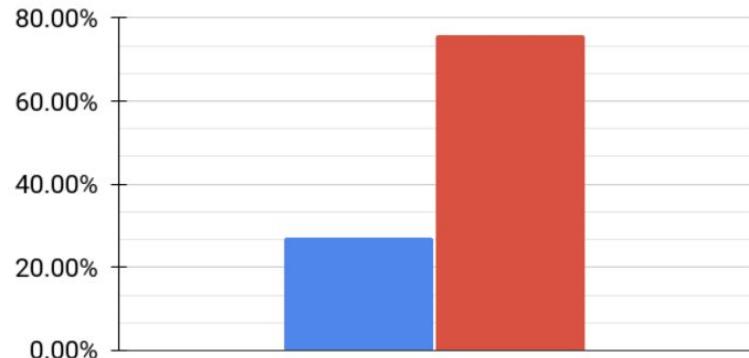
*put apple on cloth /
move apple near cloth*



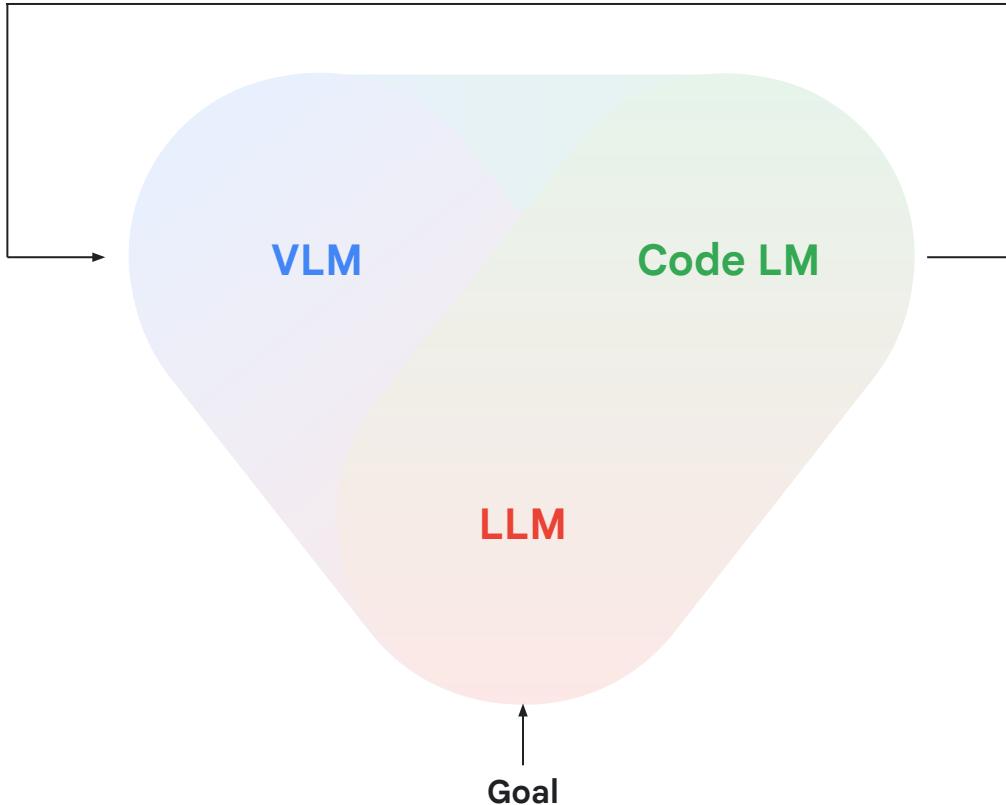
*put orange into the pot /
move orange near pot*

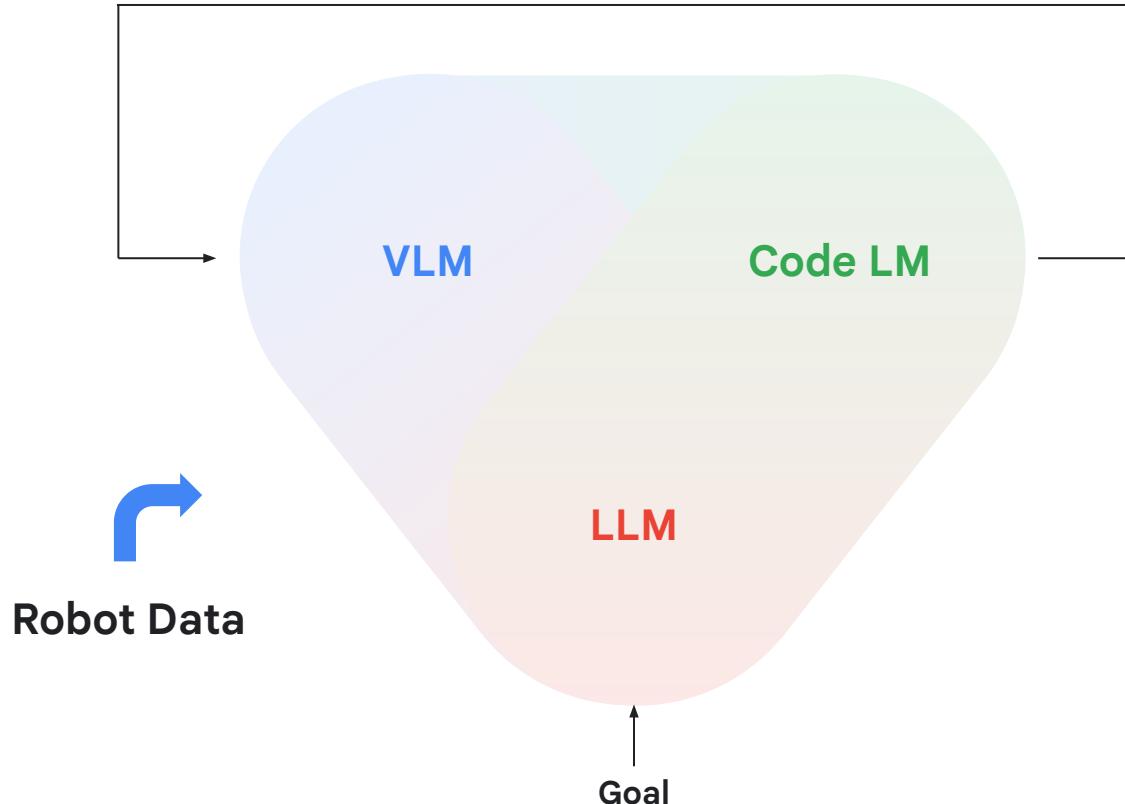


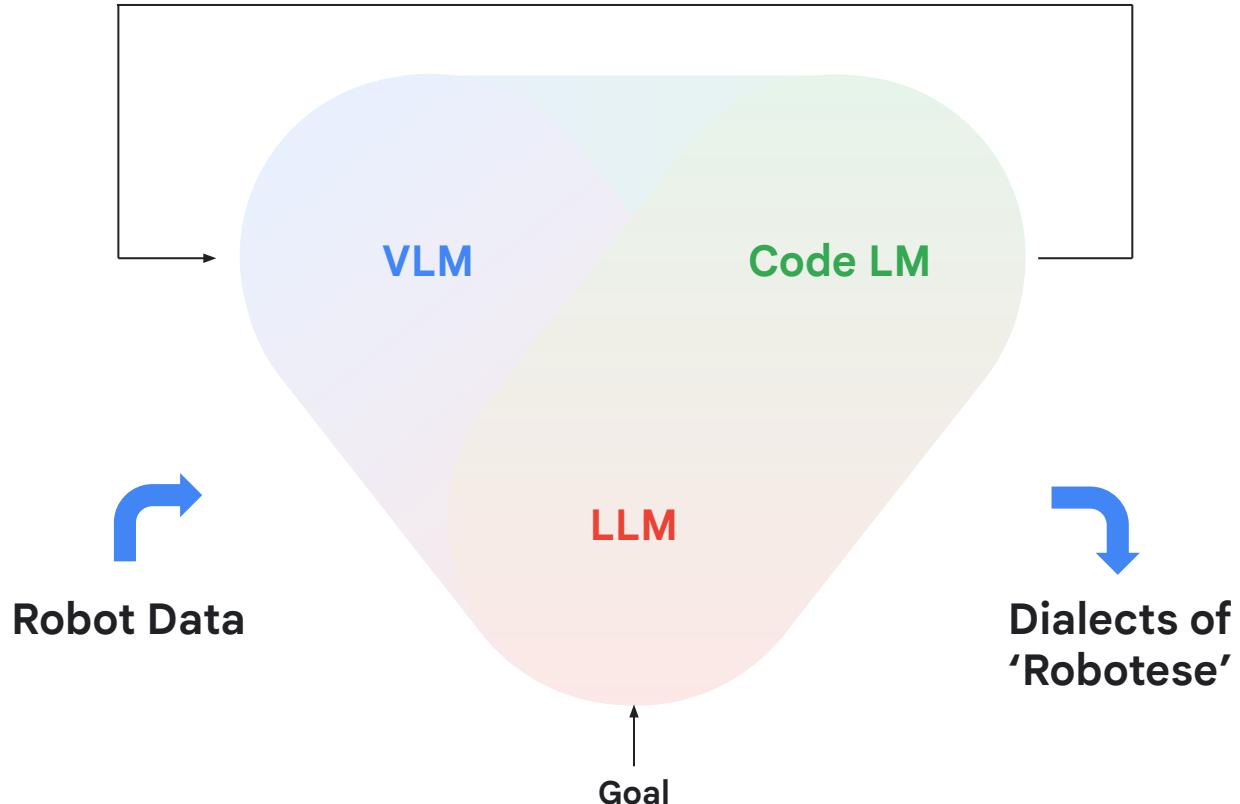
■ RT-2 ■ RT-2-X



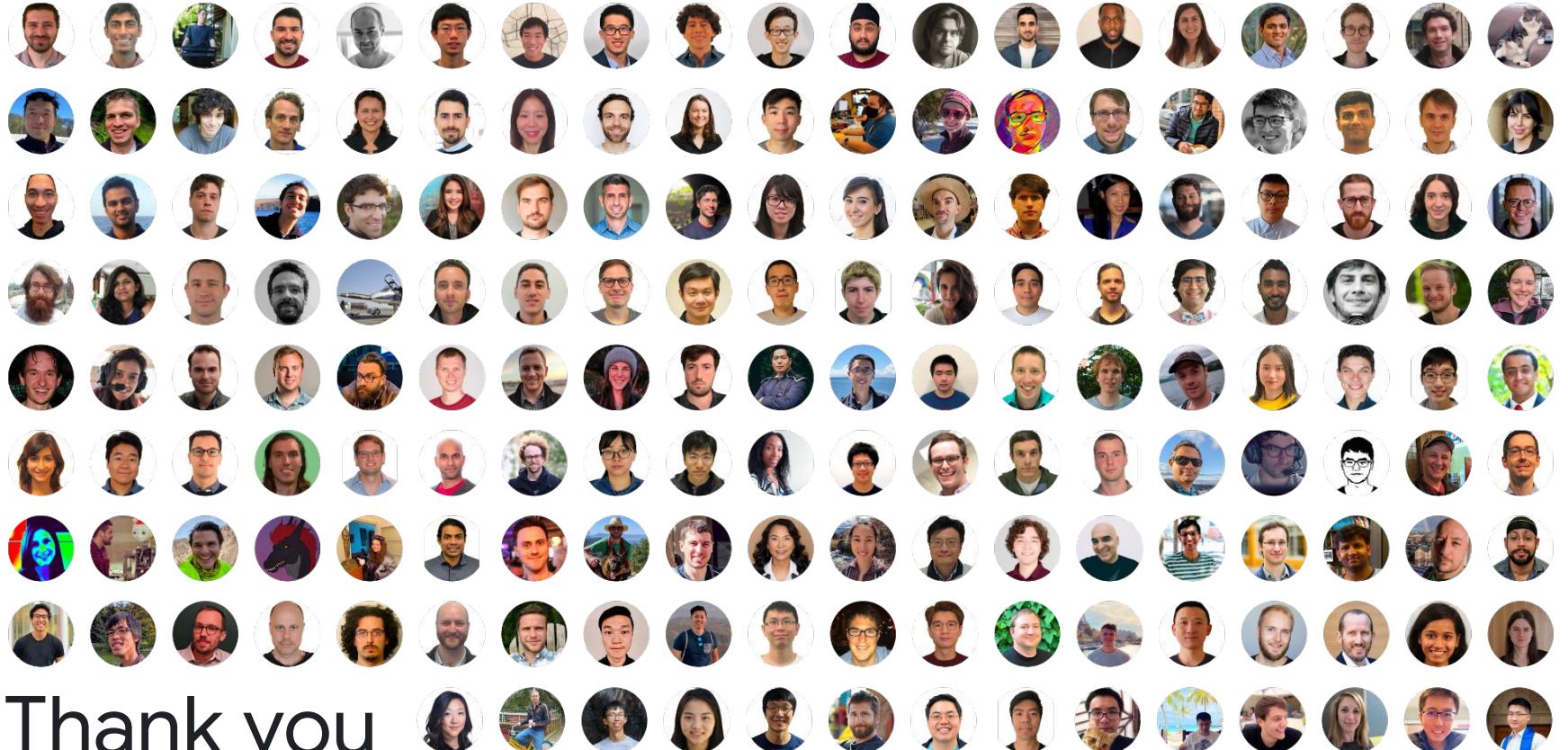
RT-2-X outperforms RT-2 by 3x in emergent skill evaluation











Thank you