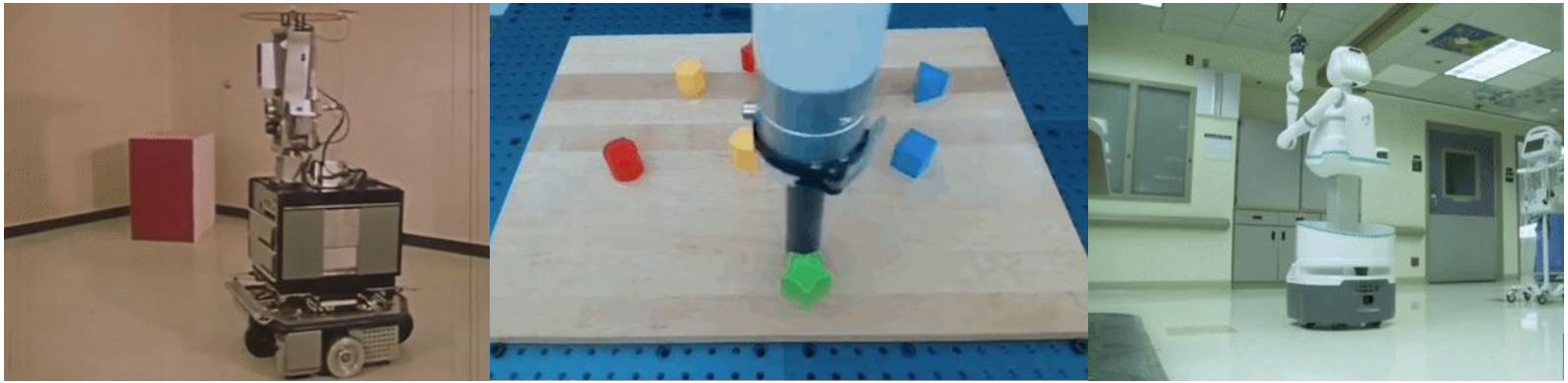
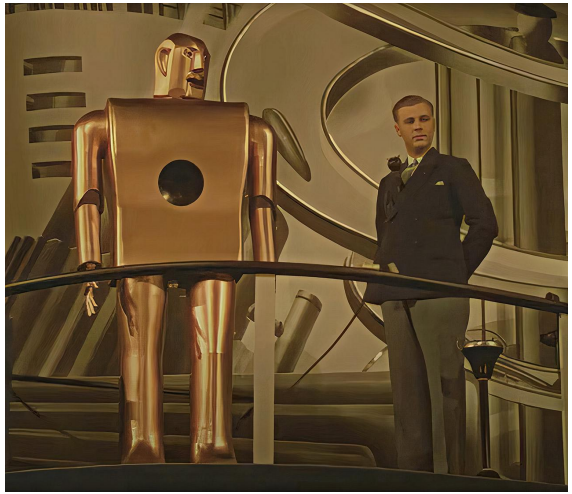
LLMs and Robotics



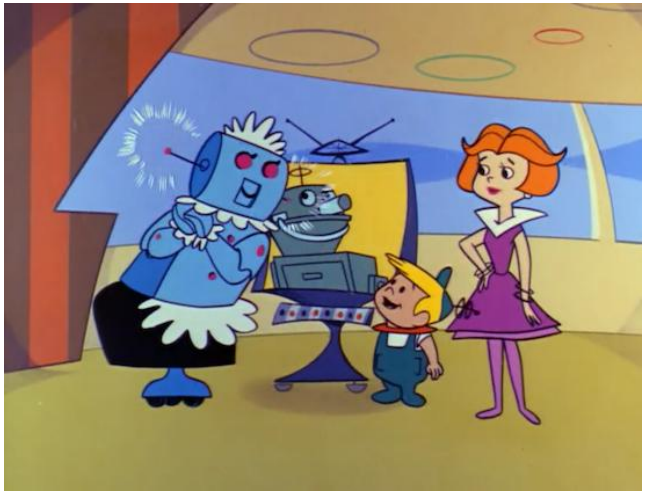
CMU 11-766
Spring 2026
Leena Mathur

Robot Dreams

Rossum's Universal Robots
Karel Čapek (1920)

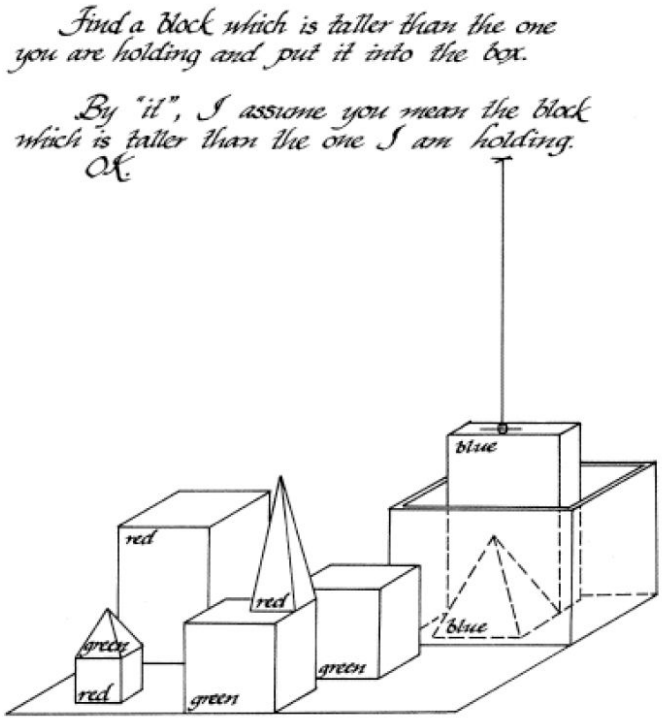


Elektro the robot
World's Fair (1939)



Rosey the Robot
The Jetsons (~1960s)

SHRDLU (Terry Winograd, 1971)



9. can the table pick up blocks?

NO.

The semantic rules make this a nonsensical question which therefore must be false.

10. can a pyramid be supported by a block?

YES.

The deductive system finds an actual example, so it knows this is possible.

11. can a pyramid support a pyramid?

I DON'T KNOW.

The question makes sense, but there is no example in the scene, so there is no easy way to tell without being able to examine the robot's programs.

12. stack up two pyramids.

I CAN'T.

The robot tries and fails.

13. the blue pyramid is nice.

I UNDERSTAND.

This data is added to the data base.

14. I like blocks which are not red, but I don't like anything which supports a pyramid.

I UNDERSTAND.

Two new theorems are created for proving things about "liking".

15. do I like the box?

NO.

The box supports whatever it contains, including a pyramid.

$\pi_{0.5}$ model demo (Physical Intelligence, 2025)

Today's Talk

- What makes robotics challenging?
- (Brief) history to understand LLMs for robotics
- VLA Training and Insights
- Data Landscape
- Evaluation and Generalization
- Open Questions

Scope + Terms

This talk will focus on LLMs and ideas you've been learning about all semester applied to **robotics**

Robotics → physical hardware

Embodied AI → broader category of agents that can take actions in environments

https://ai.meta.com/research/seamless-interaction/

Why is robotics hard?

From language to physical action

“Make me a cup of tea”

Imagine you are a robot viewing this kitchen

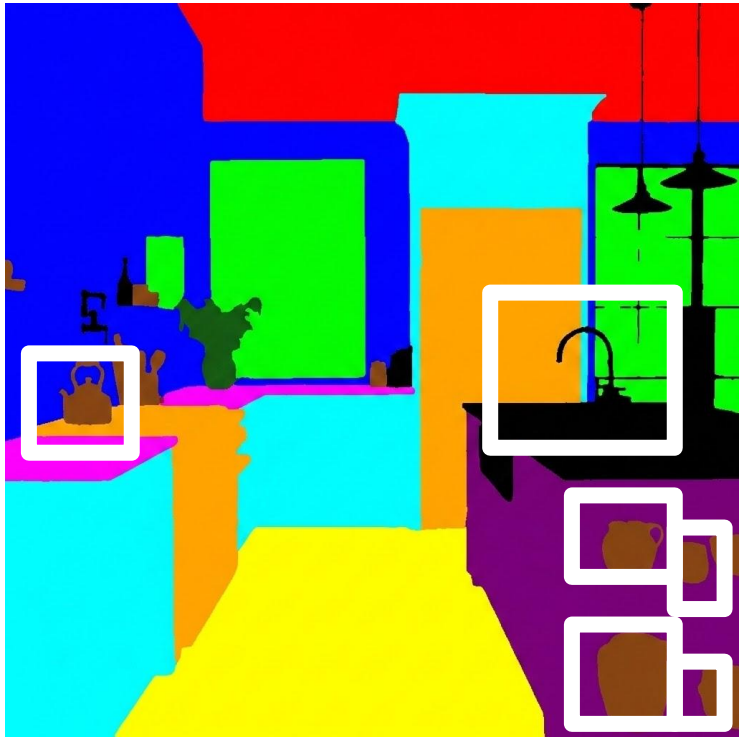
“Make me a cup of tea”

Language Understanding
What does “make tea” mean?

“Make me a cup of tea”

Language Understanding
What does “make tea” mean?

Scene Understanding
What is in front of me?

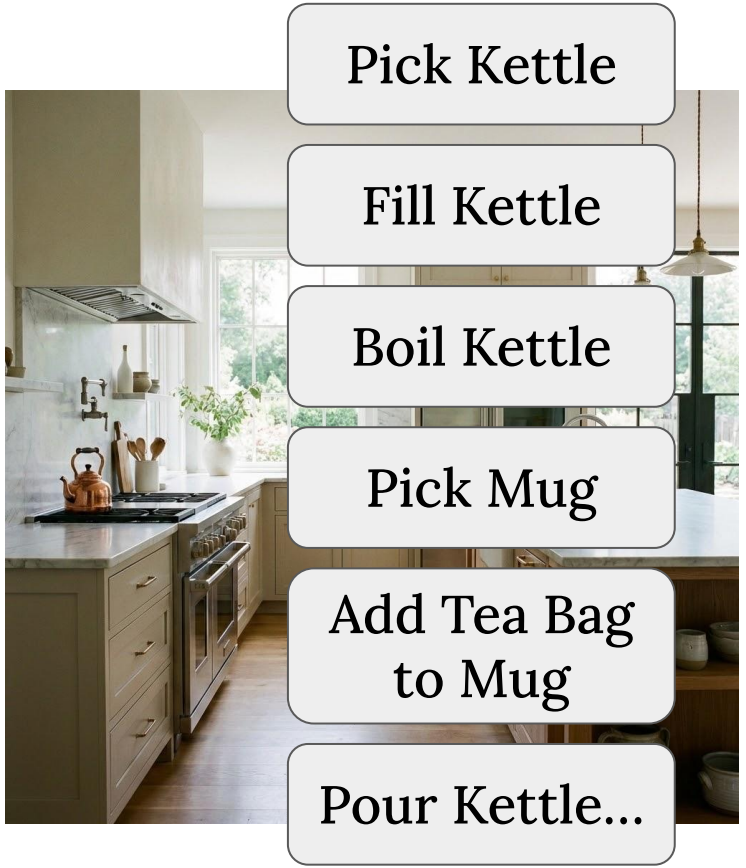


“Make me a cup of tea”

Language Understanding
What does “make tea” mean?

Scene Understanding
What is in front of me?

Planning
Given what I understand/see, what steps do I take?



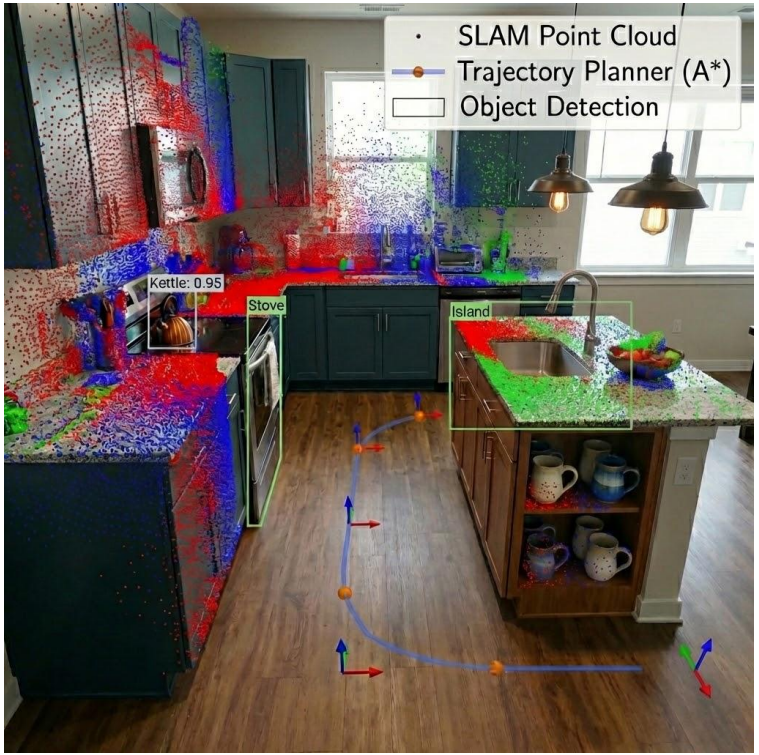
“Make me a cup of tea”

Language Understanding
What does “make tea” mean?

Scene Understanding
What is in front of me?

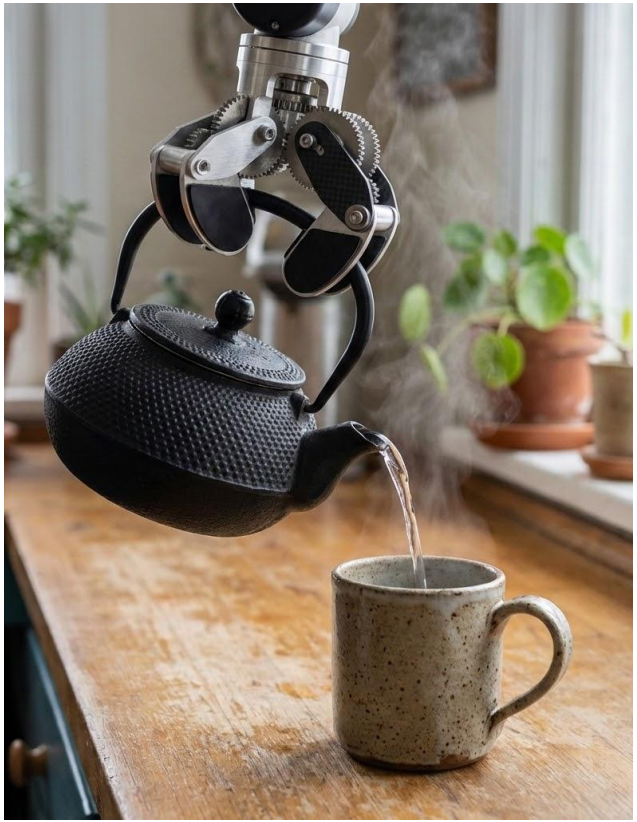
Planning
Given what I understand/see, what steps do I take?

Navigation
Where am I? Where do I go?



“Make me a cup of tea”

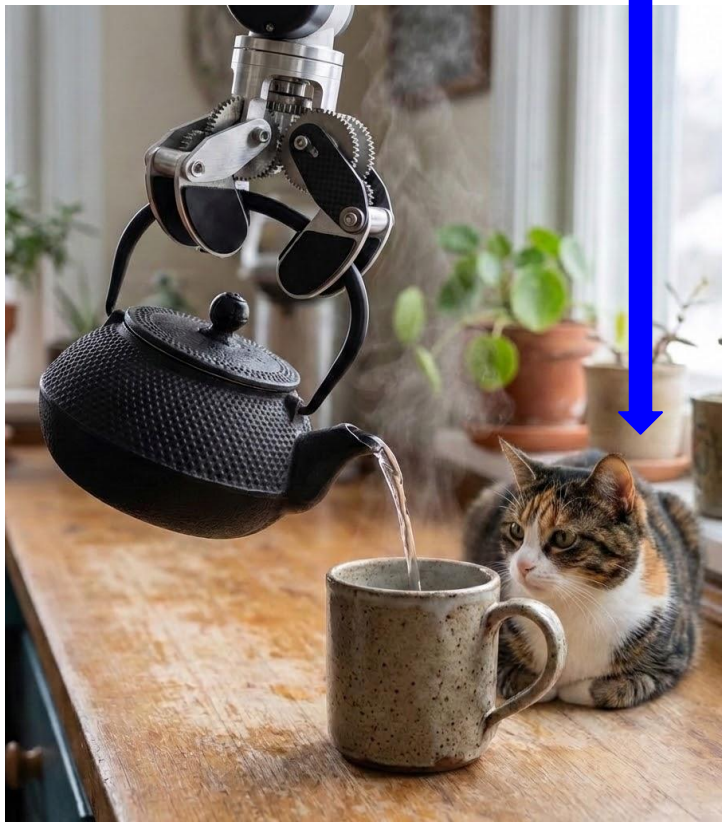
Manipulation
How do I pick up objects? Where do I grasp? How hard do I grip and tilt?



“Make me a cup of tea”

Manipulation
How do I pick up objects?
How hard do I grip and tilt?

Nonstationary Environments
The world is changing!

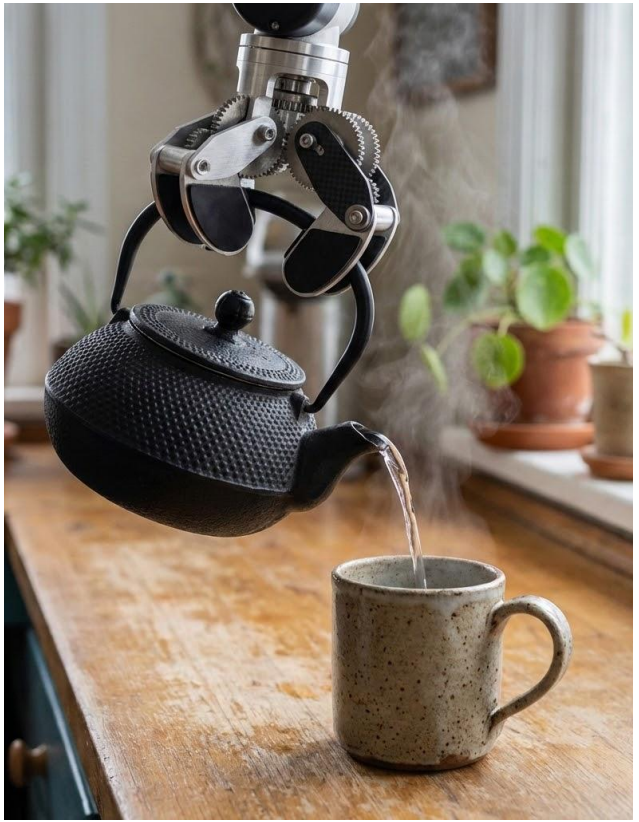


“Make me a cup of tea”

Manipulation
How do I pick up objects?
How hard do I grip and tilt?

Nonstationary Environments
The world is changing!

Long-Horizon Problem
5 min = 84,000 *motor commands*



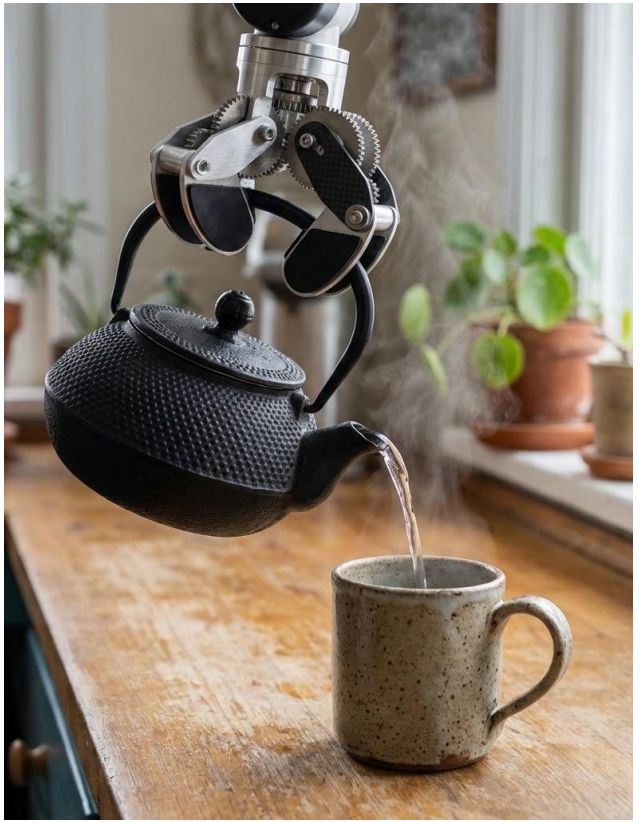
“Make me a cup of tea”

Manipulation
How do I pick up objects?
How hard do I grip and tilt?

Nonstationary Environments
The world is changing!

Long-Horizon Problem
5 min = 84,000 *motor commands*

Reliability
Needs to work 99.X% of the time!



What robotics challenges can LLMs help address?

Language Understanding

Scene Understanding

Planning

Navigation

Manipulation

Nonstationary World

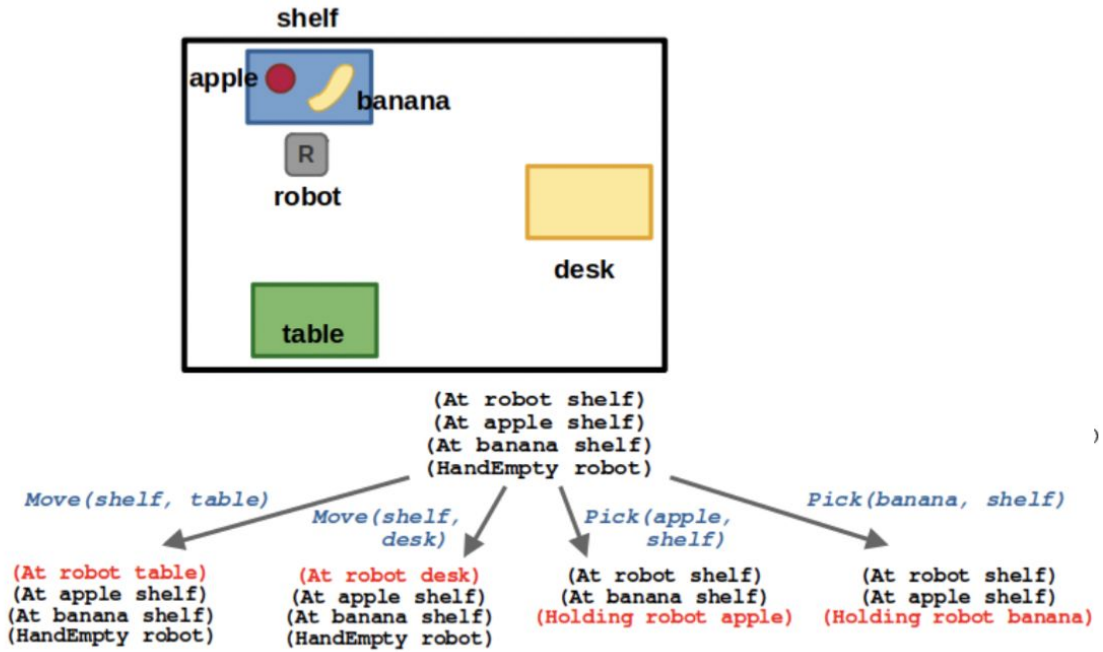
Long-Horizon Tasks

Reliability

A (Brief) History

When and why did roboticists start using LLMs for robots?

Classical Planning



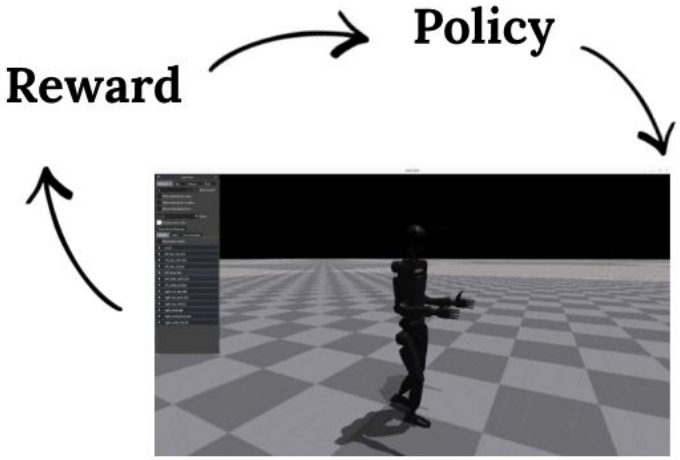

```
(define (domain vehicle)
 (:requirements :strips :typing)
 (:types vehicle location fuel-level)
 (:predicates (at ?v - vehicle ?p - location)
              (fuel ?v - vehicle ?f - fuel-level)
              (accessible ?v - vehicle ?p1 ?p2 - location)
              (next ?f1 ?f2 - fuel-level))

 (:action drive
   :parameters (?v - vehicle ?from ?to - location
                ?fbefore ?fafter - fuel-level)
   :precondition (and (at ?v ?from)
                      (accessible ?v ?from ?to)
                      (fuel ?v ?fbefore)
                      (next ?fbefore ?fafter))
   :effect (and (not (at ?v ?from))
                (at ?v ?to)
                (not (fuel ?v ?fbefore))
                (fuel ?v ?fafter))
 )
)
```

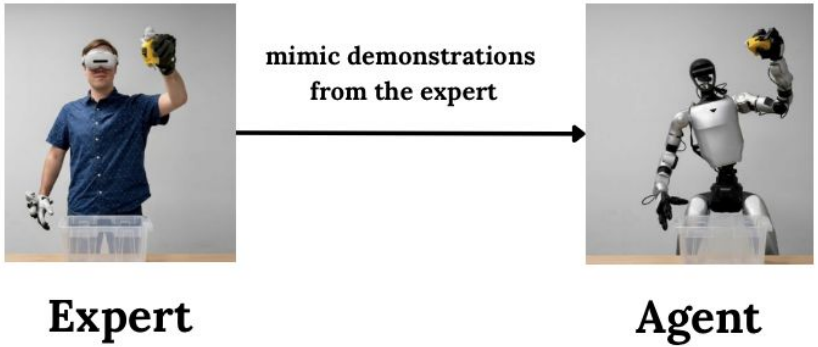
Figure 2: A domain description in PDDL.

Learning-Based Approaches

Reinforcement Learning



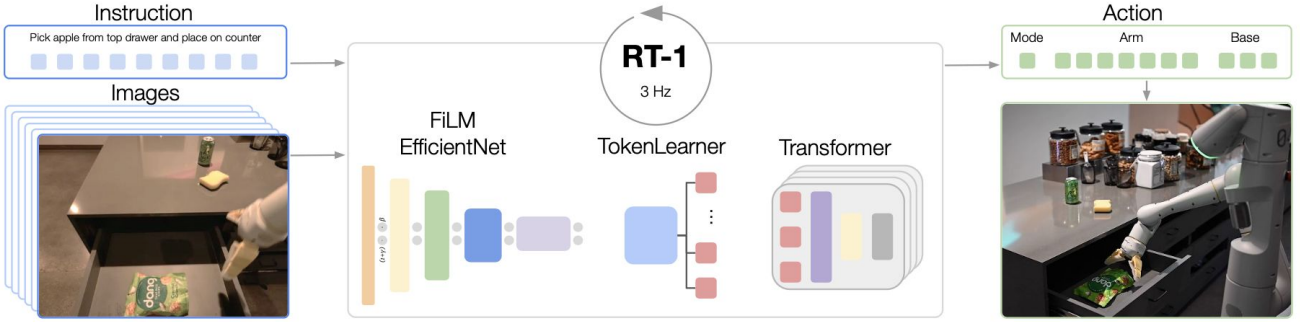
Imitation Learning



https://www.manus-meta.com/blog/nvidia-launches-isaac-teleop-at-gtc-2026-with-manus-as-the-official-data-glove

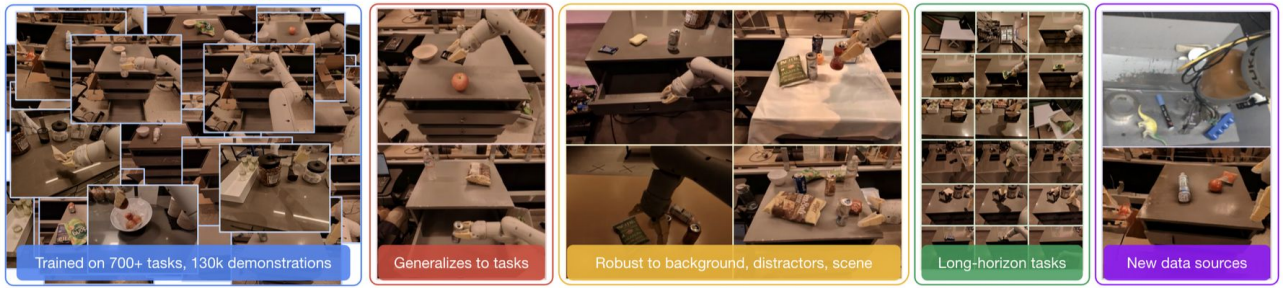
RT-1 Robotics Transformer (2022)

Transformer trained on 130,000 demonstrations across 700 tasks



(a) RT-1 takes images and natural language instructions and outputs discretized base and arm actions. Despite its size (35M parameters), it does this at 3 Hz, due to its efficient yet high-capacity architecture: a FiLM (Perez et al., 2018) conditioned EfficientNet (Tan & Le, 2019), a TokenLearner (Ryoo et al., 2021), and a Transformer (Vaswani et al., 2017).

Trained with **imitation learning** on demonstrations



RT-1: Robotics Transformer for Real-World Control at Scale, https://arxiv.org/abs/2212.06817

Gaps in 2022

What worked (in controlled labs):

- Perception (detection and segmentation)
- Short-horizon manipulation
- Task-specific learned policies via RL or imitation learning

Gaps in 2022

What worked (in controlled labs):

- Perception (detection and segmentation)
- Short-horizon manipulation
- Task-specific learned policies via RL or imitation learning

What was missing?

- Commonsense world knowledge
- Open-ended language understanding
- Task decomposition
- Long-horizon reasoning

Gaps in 2022

What worked (in controlled labs):

- Perception (detection and segmentation)
- Short-horizon manipulation
- Task-specific learned policies via RL or imitation learning

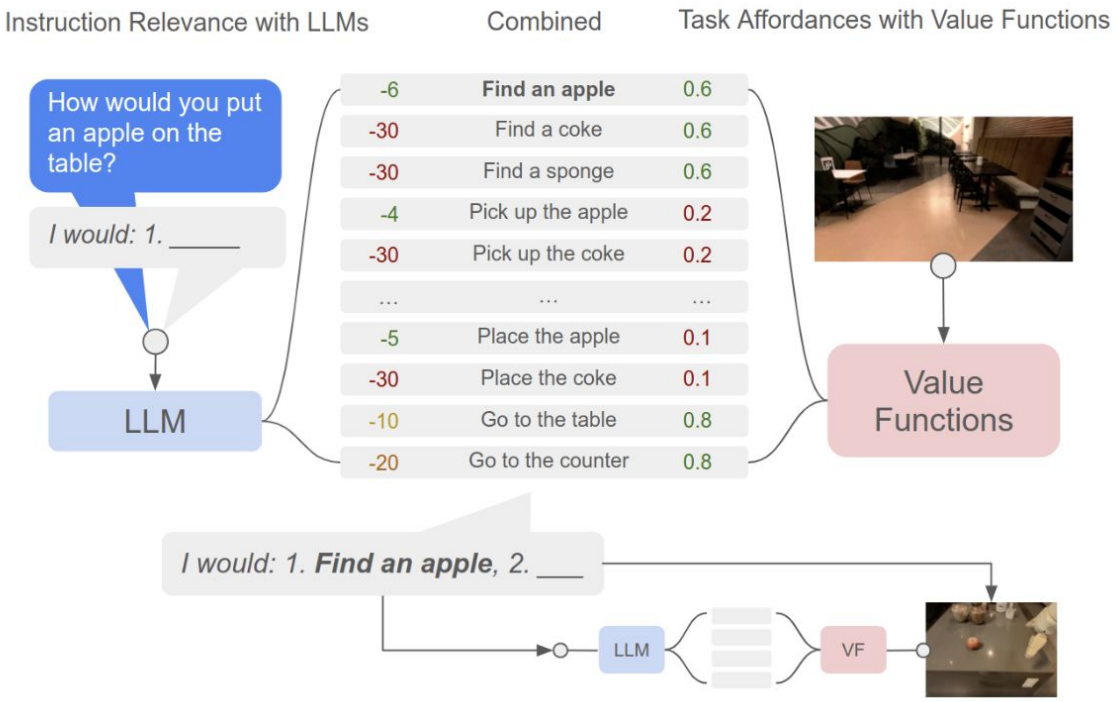
What was missing?

- Commonsense world knowledge
- Open-ended language understanding
- Task decomposition
- Long-horizon reasoning

LLMs can help here!

How can we connect the knowledge in LLMs to a policy controlling a robot body?

SayCan (LLM never sees robot data)



Frozen LLM proposes candidate steps

Affordance function scores each step (can the robot do this, given its current state)

Highest-scoring feasible action gets executed

Do As I Can, Not As I Say: Grounding Language in Robotic Affordances, https://arxiv.org/abs/2204.01691

Training-Free Approaches

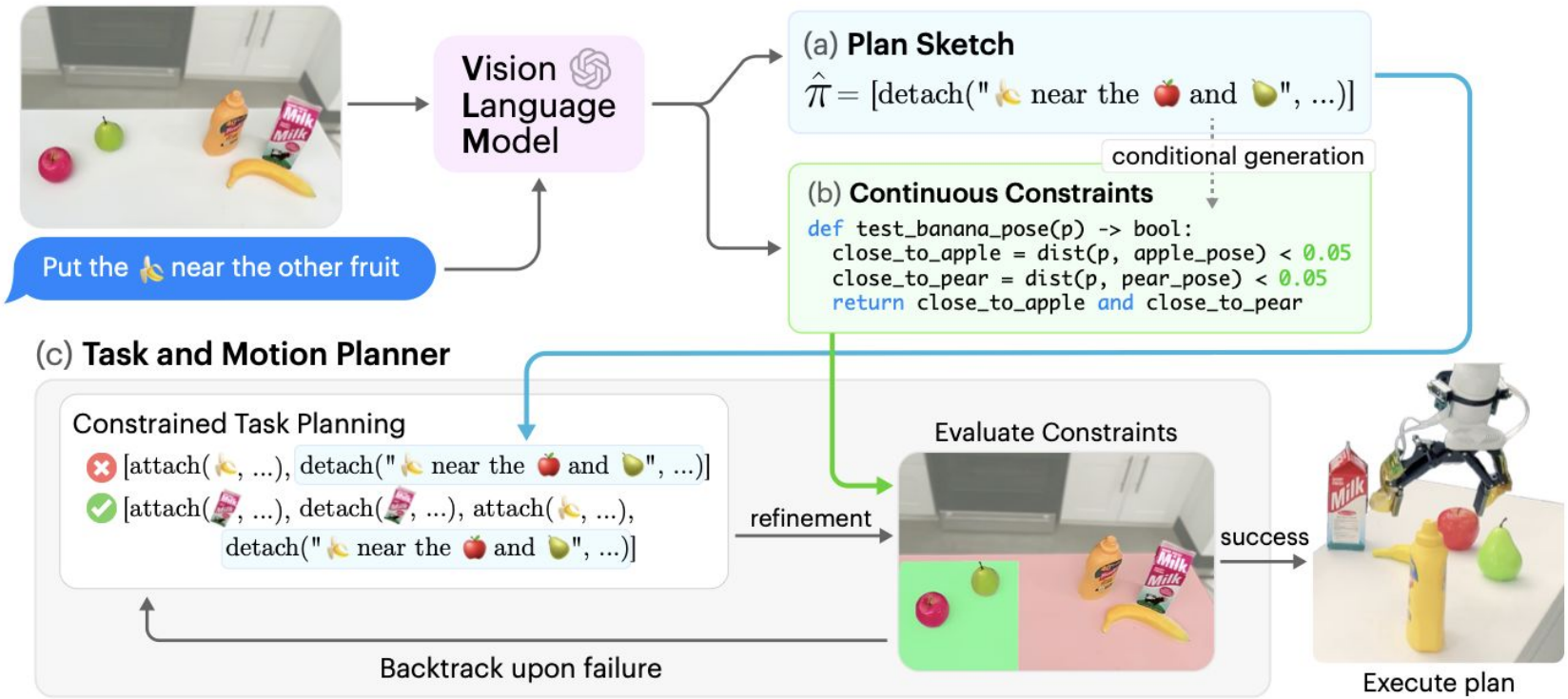
- Code as Policies (2022)
- ProgPrompt (2023)
- VoxPoser (2023)
- ProVox (2025)

...

- OWL-TAMP (2026)

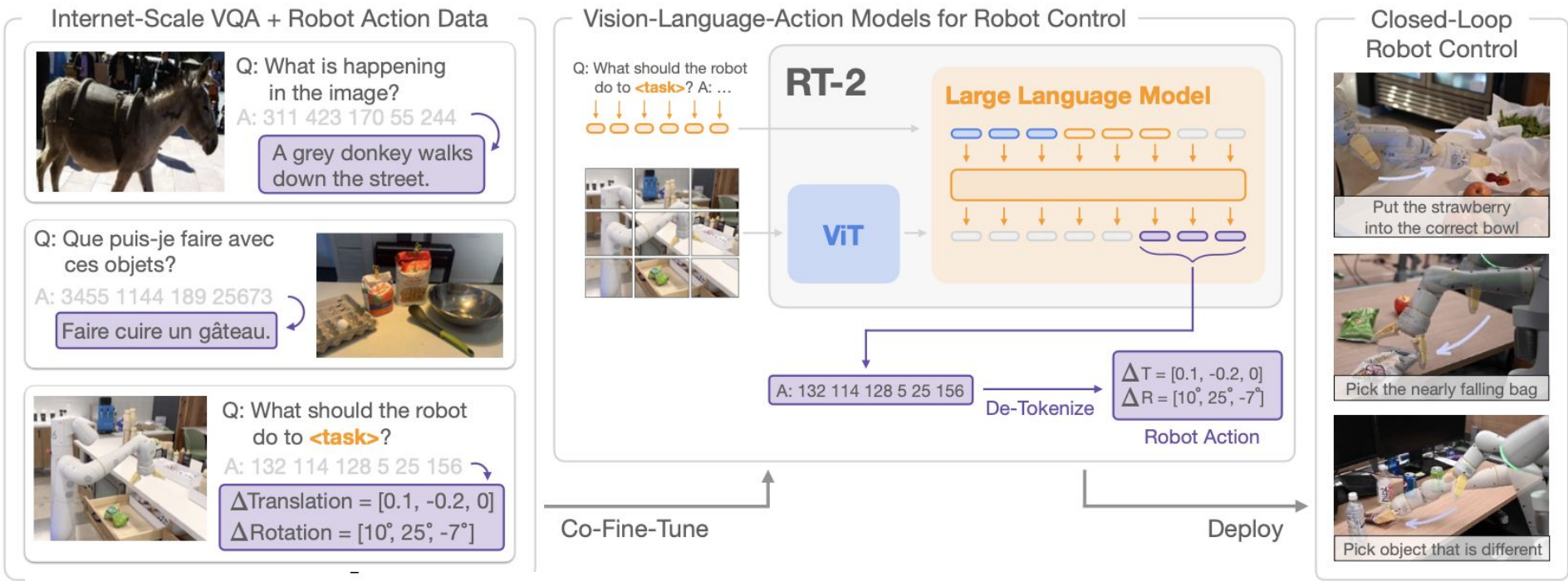
Philosophy: Compose frozen foundation models with existing controllers, planners, and/or verifiers at inference time!

OWL-TAMP



Open-World Task and Motion Planning via Vision-Language Model Generated Constraints, https://arxiv.org/abs/2411.08253

RT-2 Robotics Transformer

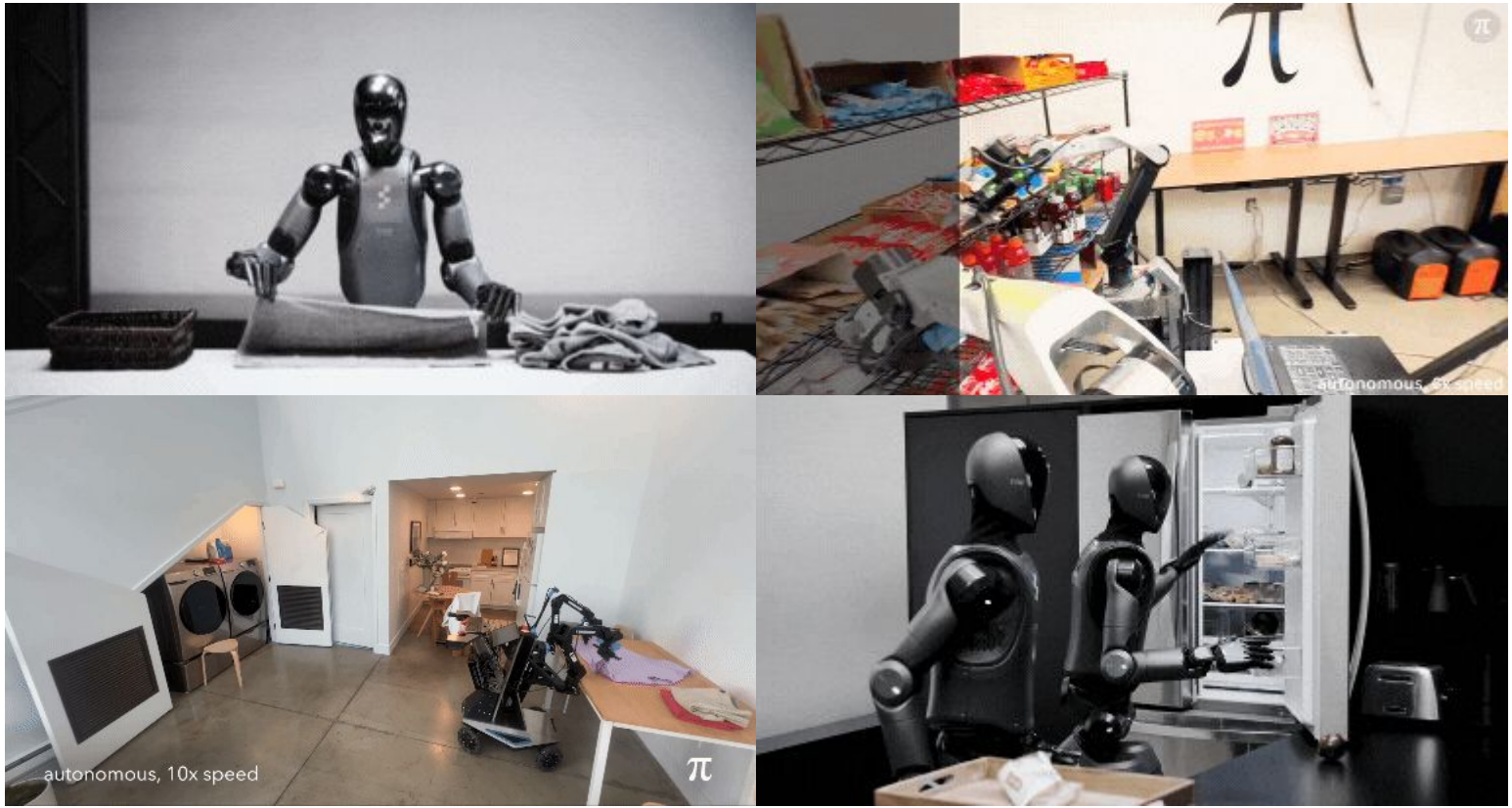


"terminate Δpos_x Δpos_y Δpos_z Δrot_x Δrot_y Δrot_z gripper_extension" "1 128 91 241 5 101 127"

RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control, https://arxiv.org/abs/2307.15818

What issues (if any) might arise by treating motor commands as discrete language tokens?

Vision-Language-Action Models ~2025

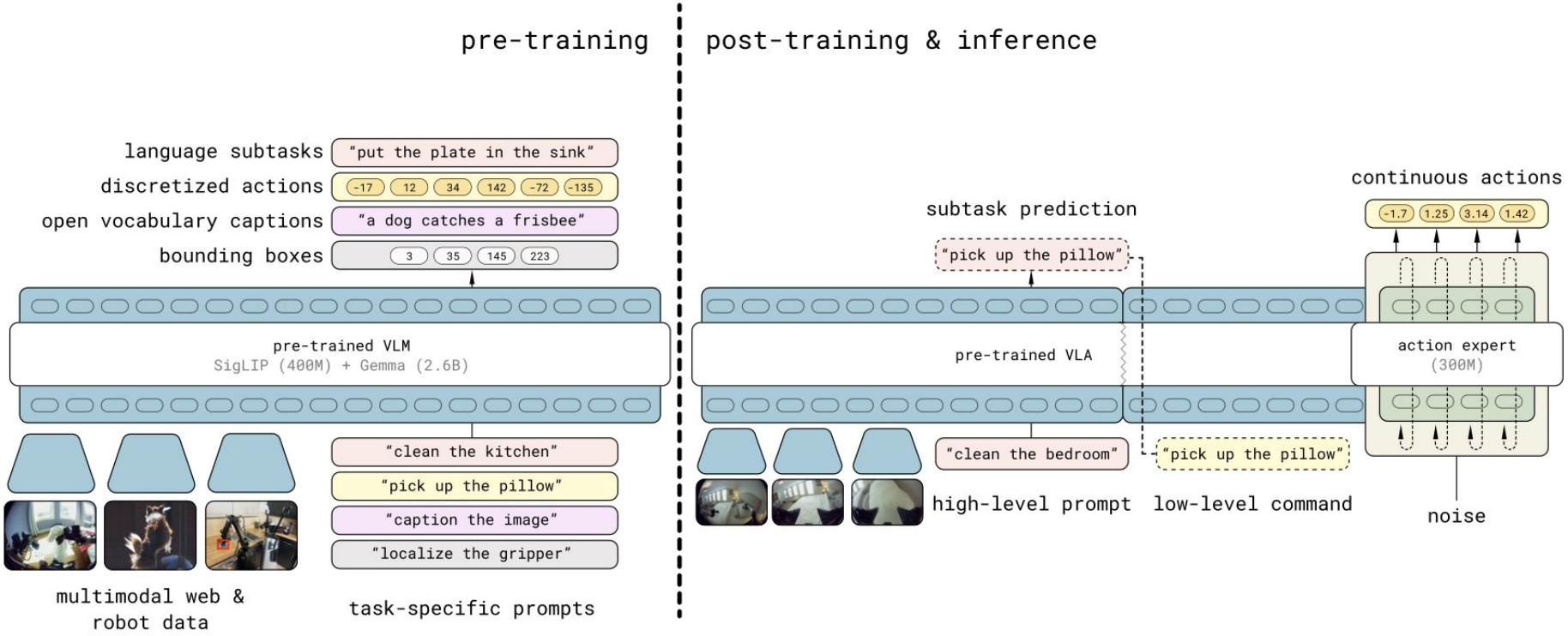


Physical Intelligence (https://www.pi.website) and Figure AI (https://www.figure.ai)

Vision-Language-Action Models

How do you turn a language model into a robot controller?

VLA Components



π0.5: a Vision-Language-Action Model with Open-World Generalization, https://arxiv.org/abs/2504.16054

Pre-training

Laboratory cross-embodiment

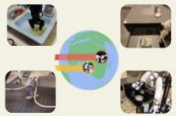

Sort drawer


Pack bottles


Sweep table


Fold laundry


Bus table


Open X-Embodiment

Diverse mobile manipulator


Shirt in basket


Spatula in holder


Wipe plate


Hang dress


Tissue on stand


Dish in sink


Make bed

Diverse non-mobile manipulator


Item in drawer


Fold linen


Tidy table


Cabinet putaway


Kettle on base


Towel on oven handle

High-level subtask

How would you clean the bedroom?

Bounding boxes:
<loc0405><loc0011><loc0911><loc0197>closet
Subtask: move to closet

How would you clean the kitchen?

Bounding boxes:
<loc0571><loc0376><loc0815><loc0484>mitten
<loc0787><loc0346><loc1003><loc0490>drawer
Subtask: move left arm forward and pick up mitten

Multi-modal web data

Describe this region:
<loc0470><loc0390><loc0605><loc0484>
Front legs of elephant

What kind of pie is this?
This is a delicious-looking pecan pie. The image shows a classic pecan pie with its characteristic dark brown filling studded with pecans.

Verbal instruction


Policy: put plate in sink
Relabeled: put plate on rack


Policy: push the top drawer
Relabeled: pick up blue shirt

Post-training

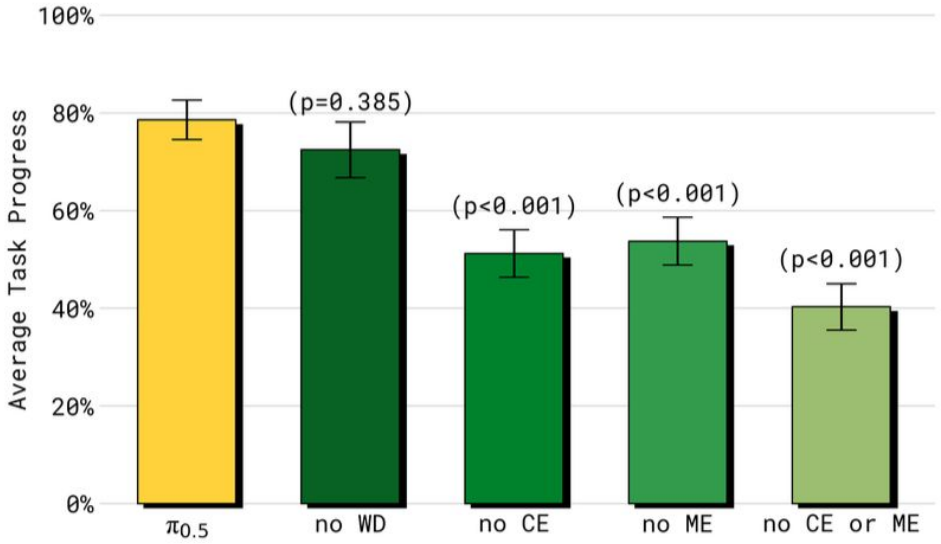
Training Recipe Ablations

No WD = no web data

No CE = cross-embodiment data

No ME = multi-environment non-mobile data

No CE or ME = excludes both data sources from other robots



Action Tokens and Flow Matching

RT-2 predicting a_t (one action at 3 Hz)

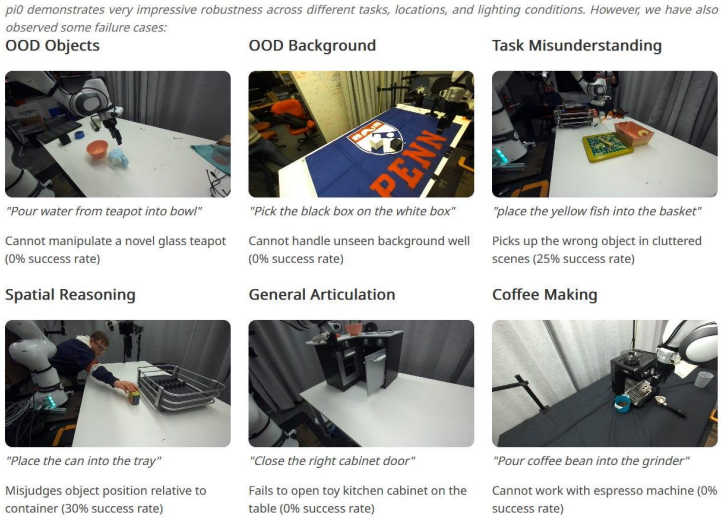
$\pi_{0.5}$ predicting $a_{t:t+50}$ (50 actions, re-plan)

Flow matching:

- Mix the real trajectory with random noise
- Train the model to predict the direction from where it is back to the clean trajectory
- At inference, start from noise and take 10 steps in the predicted direction for smoother motion

Limitations of VLAs

- VLAs do not always generalize well after fine-tuning
- Evidence of memorizing action chunks in training data
- Evidence of memorizing latent token embeddings

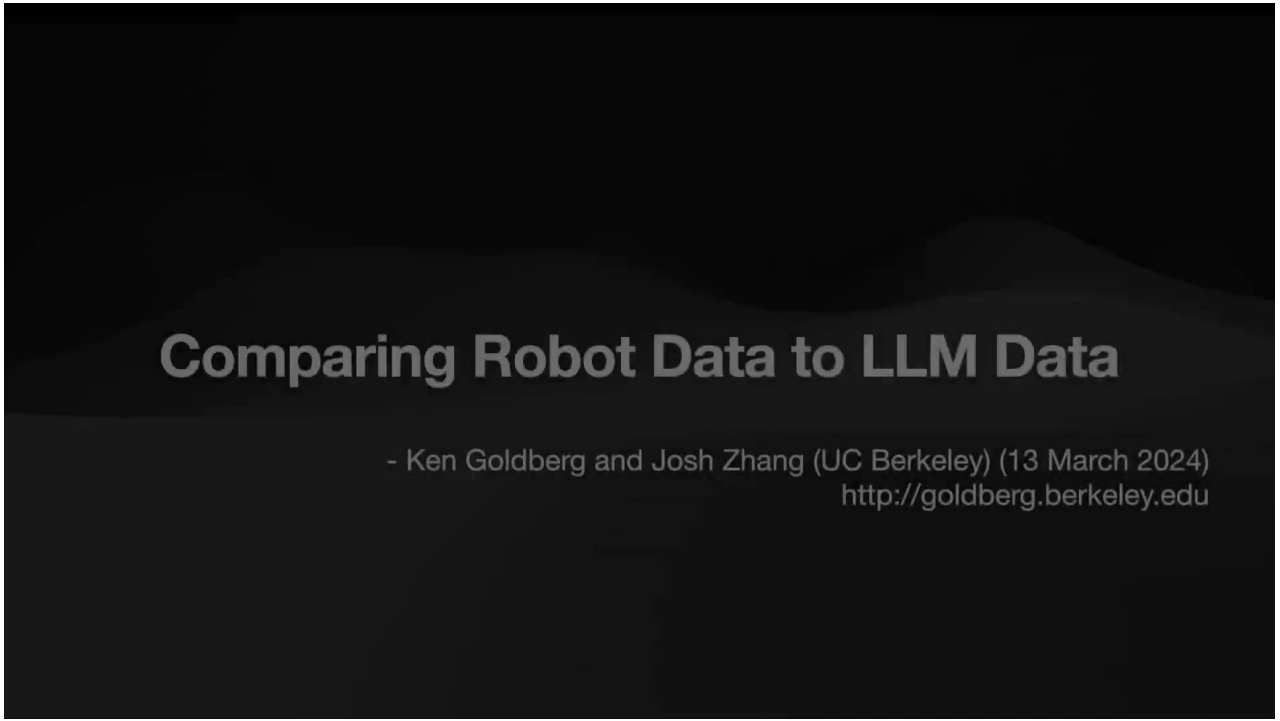


Sparse Autoencoders Reveal Interpretable and Steerable Features in VLA Models, https://arxiv.org/abs/2603.19183
VLA Models Are More Generalizable Than You Think: Revisiting Physical and Spatial Modeling, https://arxiv.org/abs/2512.02902
SAFE: Multitask Failure Detection for Vision-Language-Action Models, https://arxiv.org/abs/2506.09937
Image from https://itcanthink.substack.com/p/vision-language-action-models-andfrom Jie Wang et al. at GRASP Lab

Data Landscape

Sources of training data

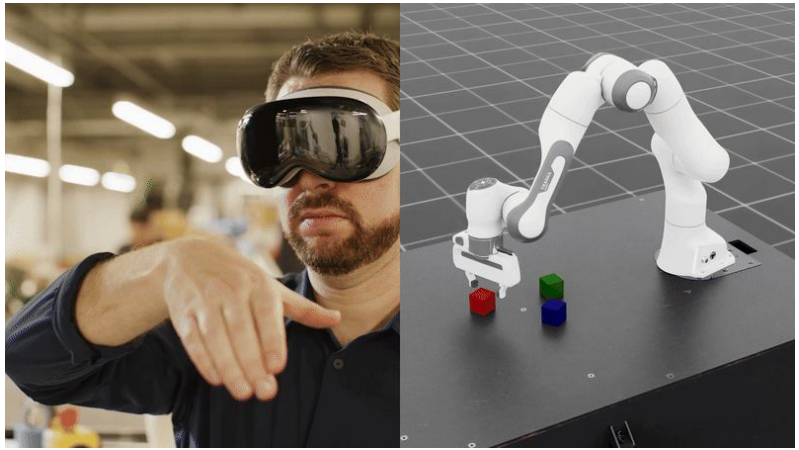
Scarcity of Robot Data

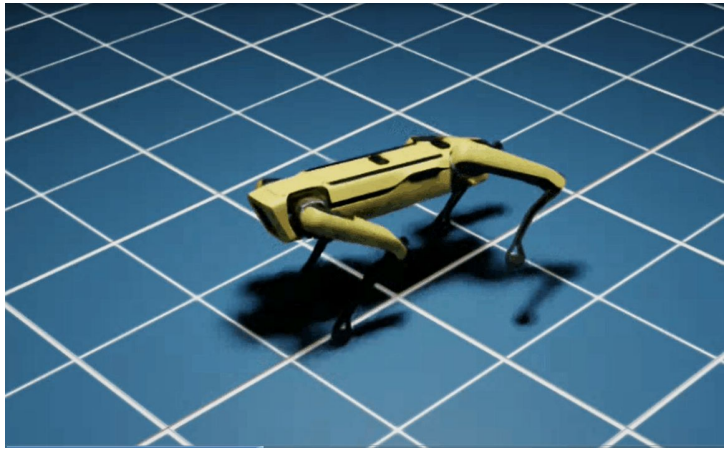


K. Goldberg, J. Zhang, Comparing the size of robot data to LLM data

Sources of Robot Data

- Simulation
- Synthetic Data Generation
- Human teleoperation
- Repurposing human data
- Repurposing web data





Evaluation

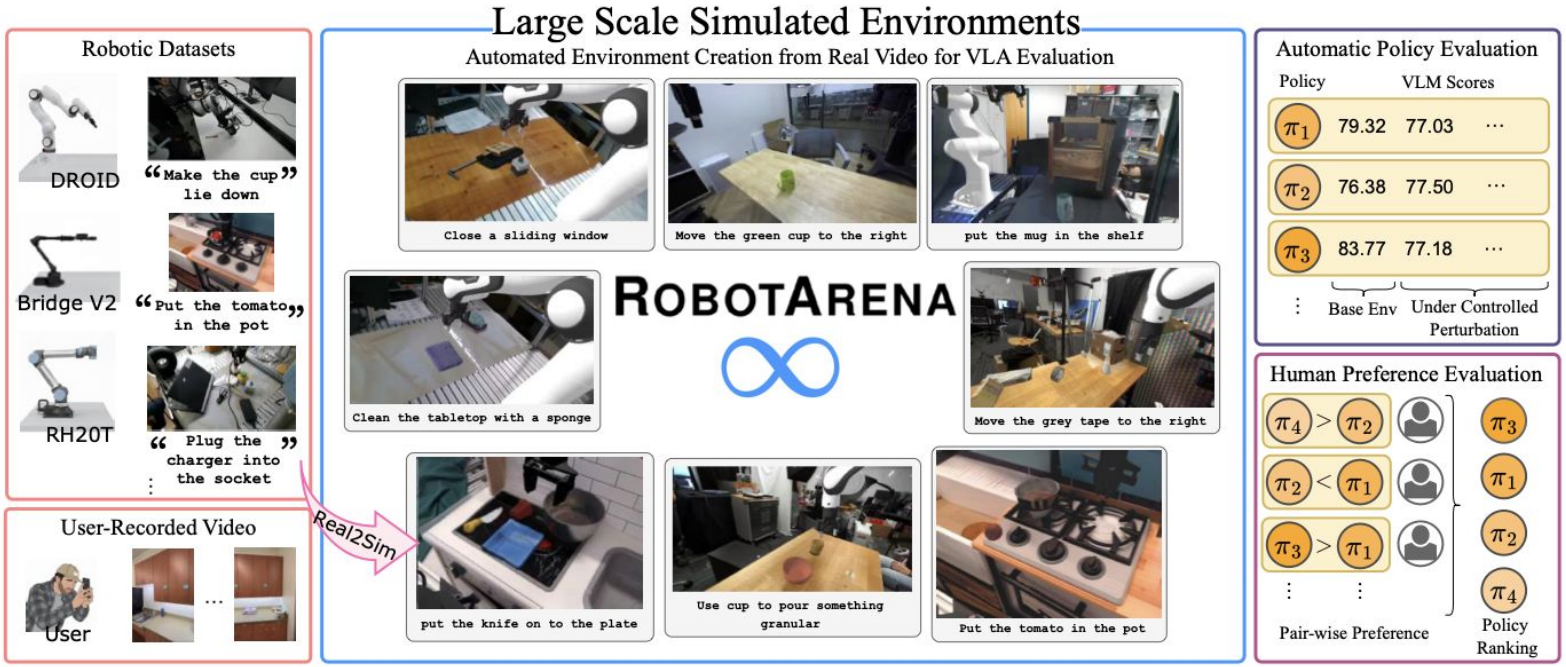
How can LLM and VLA evaluation paradigms improve for robotics applications?

Evaluation Crisis

- True: $\pi_{0.5}$ can clean kitchens it has never seen
- Also true: π models (and all VLAs) can fail if the lighting is different, drawer handle is slippery, etc
- Benchmarking is difficult in robotics!

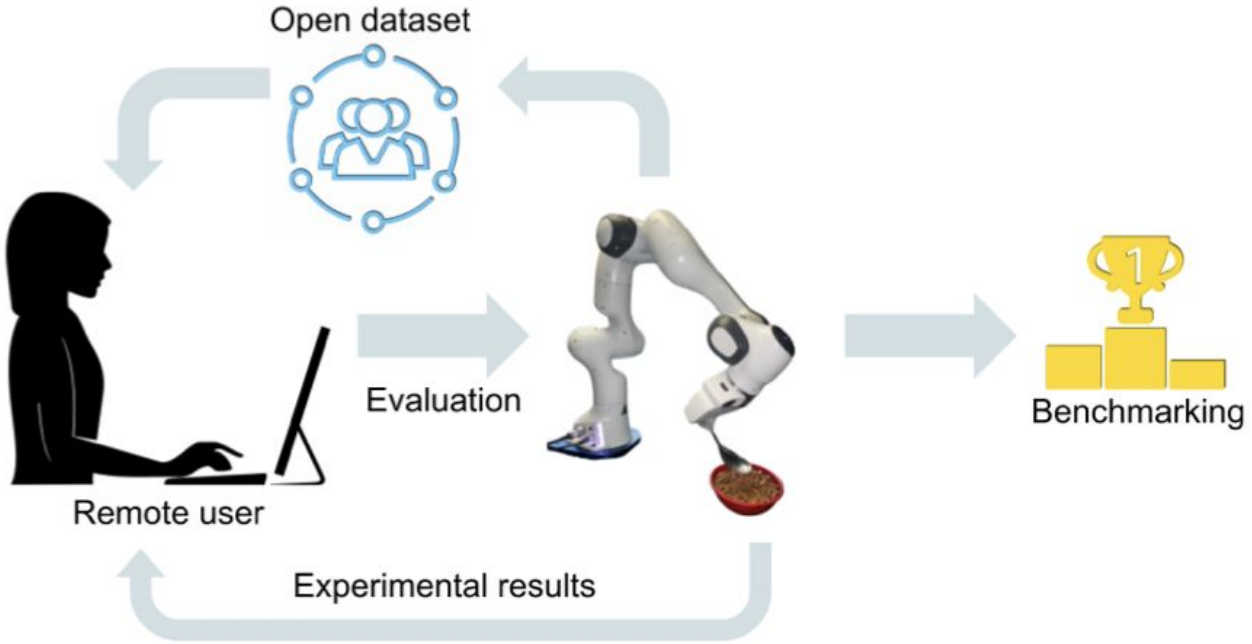
https://itcanthink.substack.com/p/how-do-we-quantify-progress-in-robotics

RobotArena for Evaluation



RobotArena ∞: Scalable Robot Benchmarking via Real-to-Sim Translation (https://arxiv.org/abs/2510.23571)

Remote Real-World Benchmarks



Train Offline, Test Online: A Real Robot Learning Benchmark (https://arxiv.org/abs/2306.00942)

Evaluation via Demos

https://x.com/SnehalJauhri/status/1937577400541351956?s=20

Given your experience so far in LLM evaluation, how would you evaluate robot policies?

Takeaways

Takeaways

- Robotics is not another modality for LLMs
- Data is a key bottleneck for robot learning and training LLM-based models for robotics
- Training-free approaches vs end-to-end approaches have distinct strengths and weaknesses
- Need standardized evaluation to make more rigorous claims about LLM effectiveness for robotics

Future Directions

- Steering VLA representations to improve grounding
- Test-time adaptation and online correction
- Latent plan representations beyond language
- Long-context LLMs, VLAs, and memory for robotics

Thank you!