Autonomous Robotics

Robots and Foundation Models

Josiah Hanna
University of Wisconsin — Madison

Announcements

Final project due next week (Tuesday for extra credit opportunity).

Course Evaluation due May 1 (send screenshot for extra credit)

Learning Outcomes

After today's lecture, you will:

- Understand what it means for a robot's knowledge to be grounded.
- Understand how LLMs can be used in robotics and how they go beyond other learning approaches.
- Understand how robot and web data can be combined to train vision-language-action models.

Motivation for LLMs in Robotics

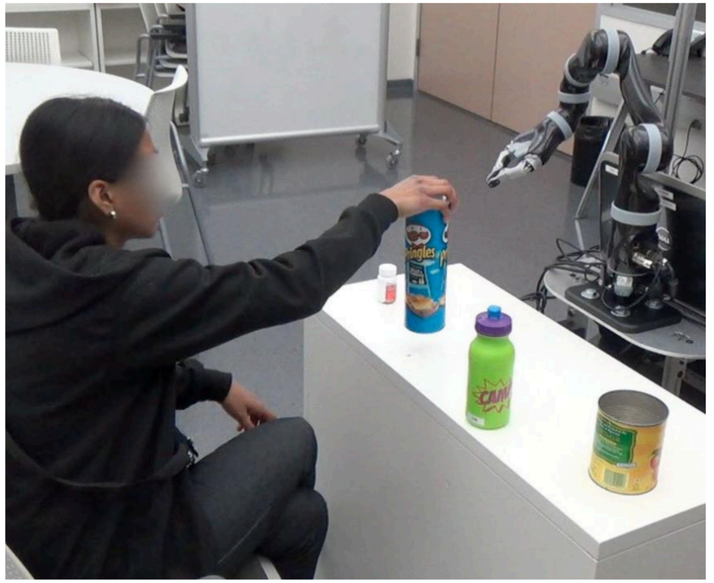
- Provide robots with common-sense reasoning.
- Allow robots to learn from repositories of human knowledge (i.e., the internet).

Grounding

- Knowledge is mapped to the physical senses and actions of a robot.
  - Example: “Pick up the red cup.”

What sequence of controls will accomplish this instruction?

What does a red cup look like in a camera image?



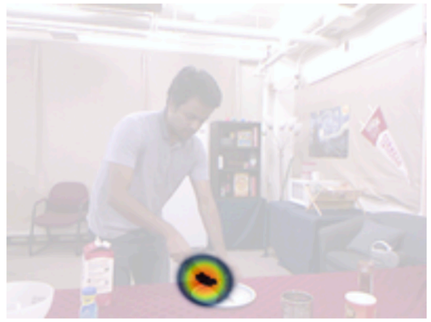
Prior Work (2014): RoboBrain

'Robo Brain' mines the Internet to teach robots

By Bill Steele

August 25, 2014

The **standing_human** can **cut** using a **knife** as shown in **$heatmap_2.**



Anticipation http://pr.cs.cornell.edu/anticipation/

236 3

Robo Brain: Large-Scale Knowledge Engine for Robots

Large Language Models

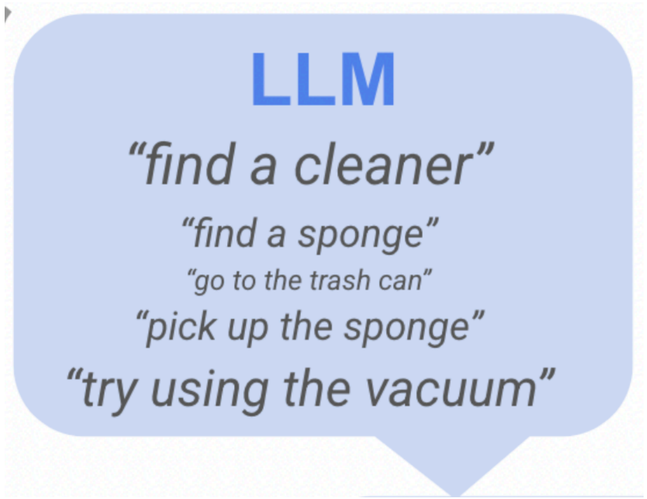
- Large neural networks that take text input and generate text output.
  - Formally, model $p(\texttt{nextword} \mid \texttt{prompt} + \texttt{previouswords})$.
  - Generate outputs one word at a time by sampling from this distribution.
  - The basic training procedure is self-supervised classification: take segments of words from a large text corpora and predict the next word that follows.
- Old idea but now **significantly** scaled up with internet scale data, datacenters of GPUs, and the transformer neural network architecture.

Large Language Models (cont'd)

- Self-supervised pre-training doesn't produce a highly useful model.
- So typically follow with some type of fine-tuning:
  - Supervised fine-tuning: human annotators provide better outputs and model imitates those (imitation learning).
  - RLHF: human annotators rank responses, learn a reward function (IRL), and then use RL with the learned reward.
  - RLVF: Run RL against an automated correctness checker (e.g., check generated code passed tests)

Ungrounded LLMs

- No matter how sophisticated the outputs of an LLM are, they are not grounded in physical experience and ability.
- VLM: vision-language models ground language to images but not to robot capabilities.
- Example: Prompt LLM with “How should I clean up a spill?”



SayCan Approach

- LLMs can identify potential skills to complete a task.
- Value functions predict probability of a skill succeeding in a state.
- SayCan: combine both to identify skills with a high probability of task success.

SayCan

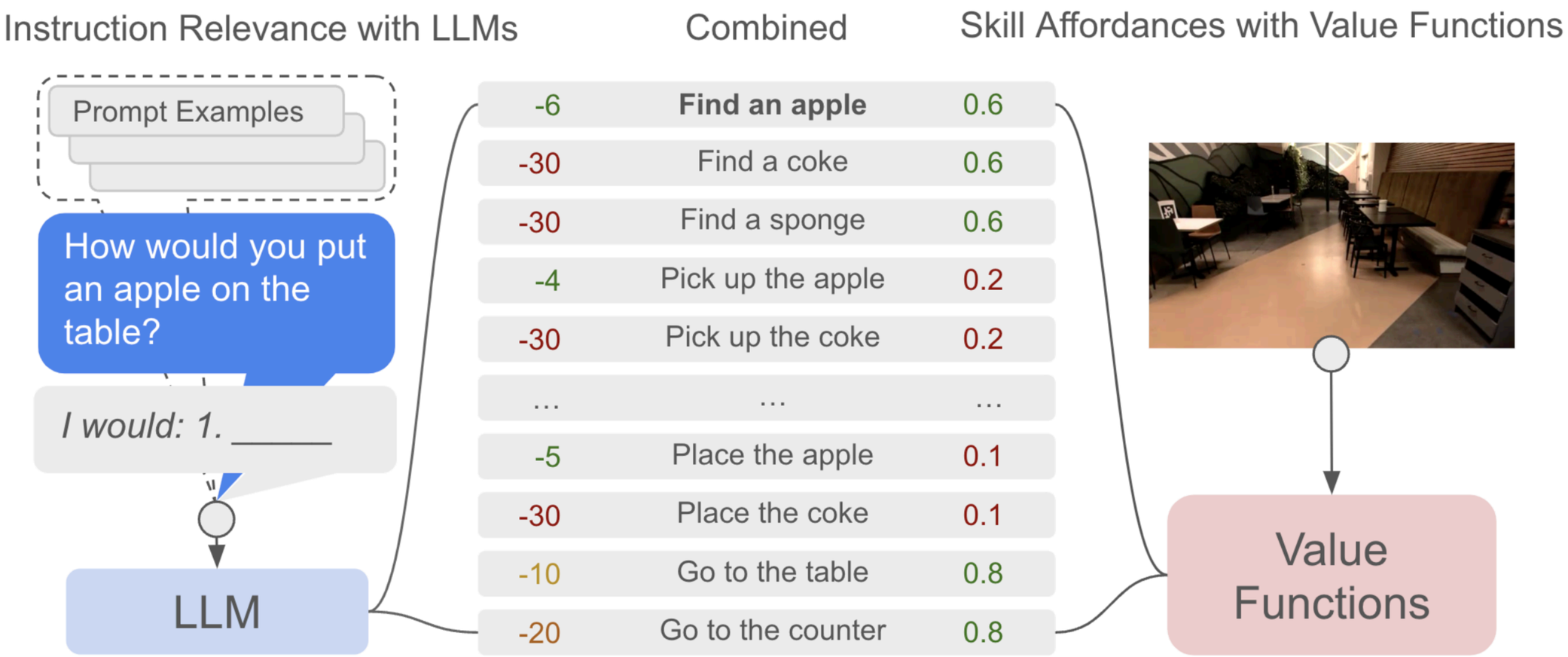
- First, assume that we already have a skill library, Π, and for $\pi \in \Pi$ we know the probability that π can be successfully executed, $p(\texttt{success} \mid s, \ell_\pi)$.
  - "If I ask robot to do ℓ_n, will it do it?"
  - Called a value or affordance function.
- When given a prompt, i, for a new task, the robot scores all skills with $p(\texttt{success} \mid s, \ell_\pi) p(\ell_\pi \mid i)$ and takes the skill most likely to succeed.

Do As I Can, Not As I Say: Grounding Language in Robotic Affordances (2022)

Value Function Training

- Skills in the skill library come from either behavior cloning or reinforcement learning.
  - Each skill is a policy ($\pi : S \rightarrow A$) and also has a short text description, ℓ_n.
- Now, given a skill, we need to learn $p(\texttt{success} \mid s, \ell_n)$.
  - Equivalent to $v_\pi(s)$ for an MDP where the reward is zero except for upon success.
  - Learn the success probability with temporal difference learning.

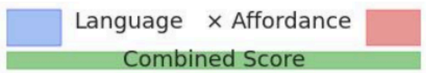
SayCan

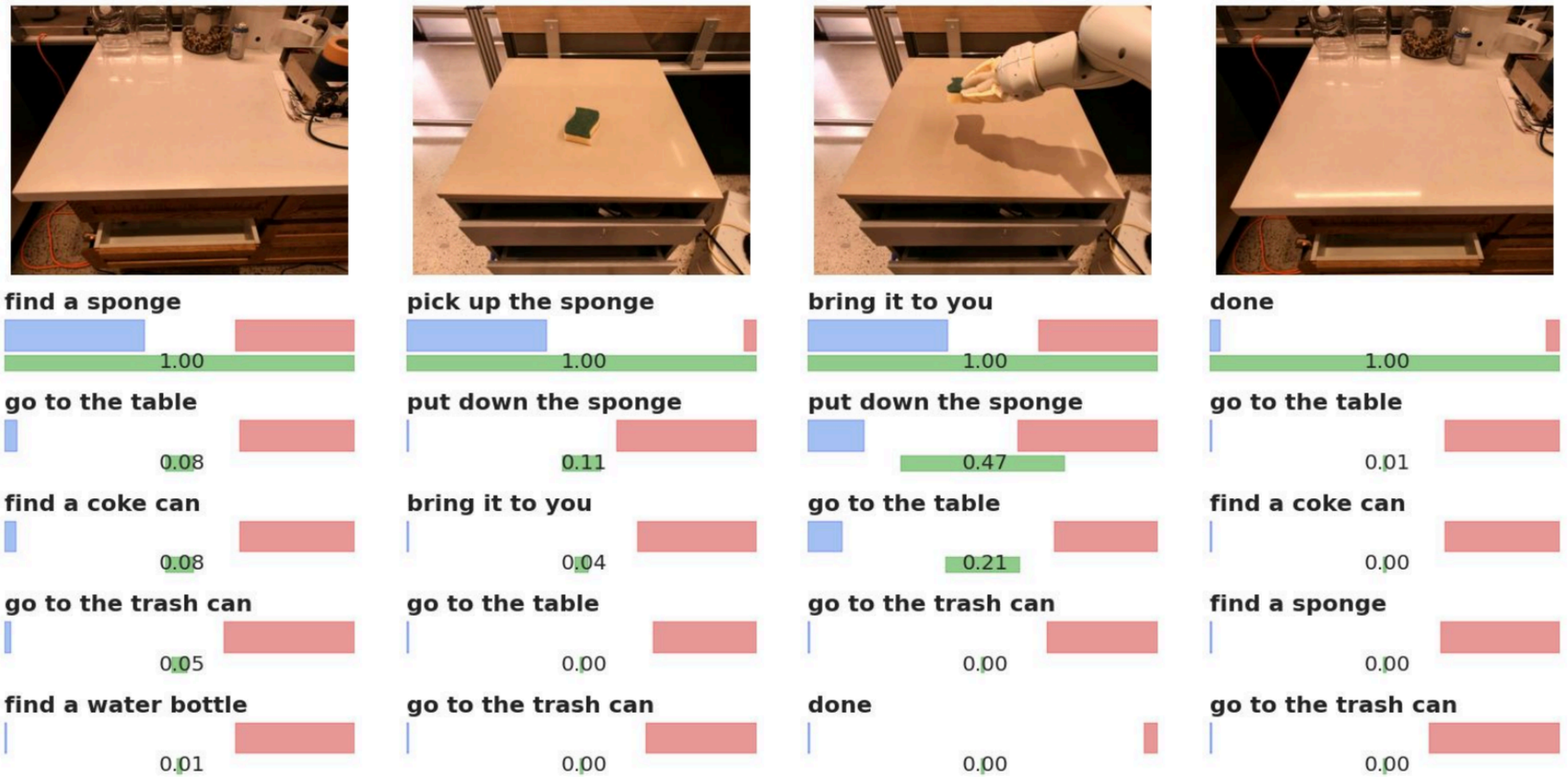

Instruction Relevance with LLMs
Combined
Skill Affordances with Value Functions
Prompt Examples
How would you put an apple on the table?
I would: 1. _____
LLM
-6 Find an apple 0.6
-30 Find a coke 0.6
-30 Find a sponge 0.6
-4 Pick up the apple 0.2
-30 Pick up the coke 0.2
… … …
-5 Place the apple 0.1
-30 Place the coke 0.1
-10 Go to the table 0.8
-20 Go to the counter 0.8
Value Functions

SayCan

Human: I spilled my coke, can you bring me something to clean it up?

Robot: I would
1. Find a sponge
2. Pick up the sponge
3. Bring it to you
4. Done



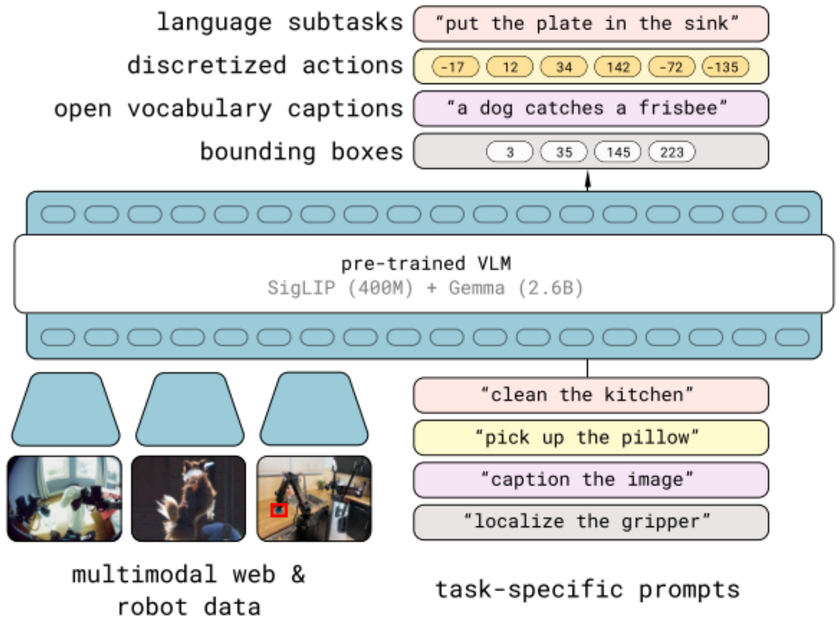


LLMs and Robots

- Strengths:
  - Take advantage of internet-scale data to help robots understand the world.
  - LLMs have demonstrated remarkable capabilities resembling advanced reasoning and problem solving —> use to inform robot action.
- Weaknesses & open questions:
  - Large models have a high storage footprint and inference cost.
  - Models may hallucinate.
  - Robots have a different embodiment than humans; knowledge may not transfer.
  - Continual learning?

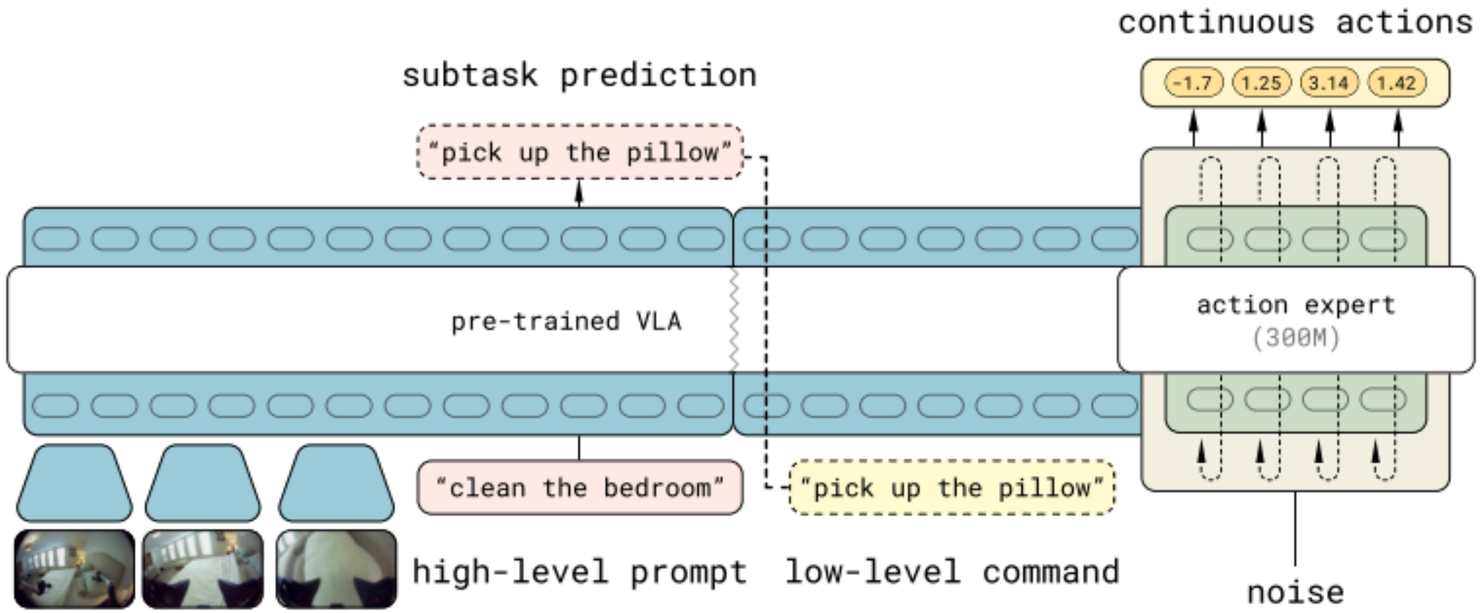
Robot Foundation Models

- VLA: vision-language-action models
- Train single model with language, images, and robot actions.
  - Need to tokenize action outputs (different schemes exist).



Post-Training

- After pre-training on all data, perform additional training with only robot data.
- Action expert: produces continuous actions for finer grained control.



Data

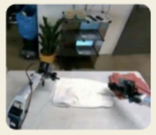
Pre-training

Laboratory cross-embodiment

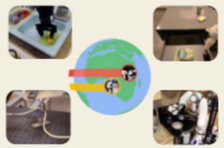

Sort drawer


Pack bottles

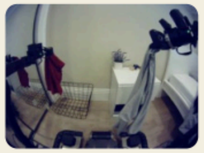

Sweep table


Fold laundry


Bus table


Open X-Embodiment

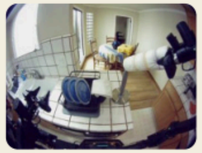
Diverse mobile manipulator


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

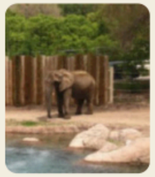

How would you clean the bedroom?

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


How would you clean the kitchen?

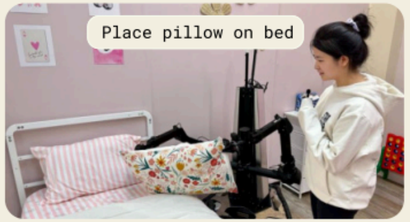
Bounding boxes:
<loc0571><loc0376><loc0815><loc0484>mitten
<loc0787><loc0346><loc1003><loc0490>drawer
Subtask: move left arm forward and pick up mitten

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

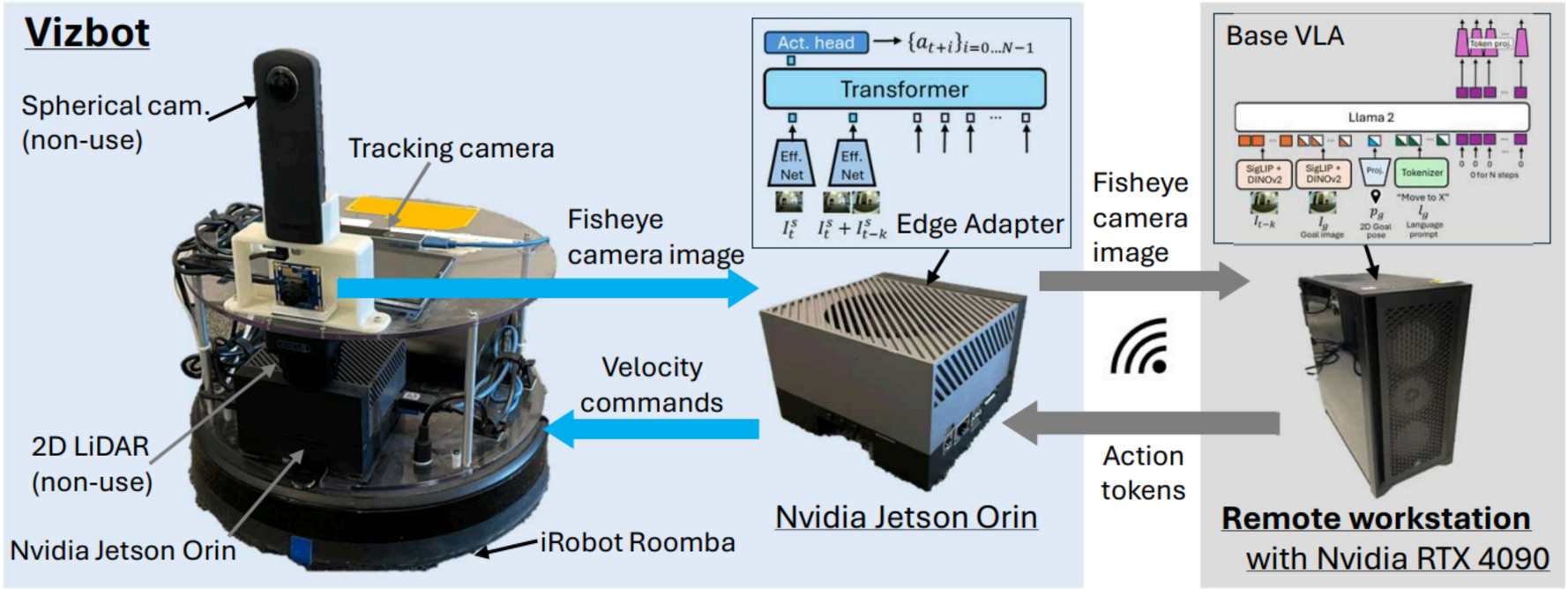
Training Objectives

- Pre-training:
  - Standard self-supervised classification except over text, object locations, and actions.
- Post-training: action expert is a generative model trained with flow matching (similar to diffusion).
  - f_θ^a is used to iteratively denoise a noisy action conditioned on state and langague.

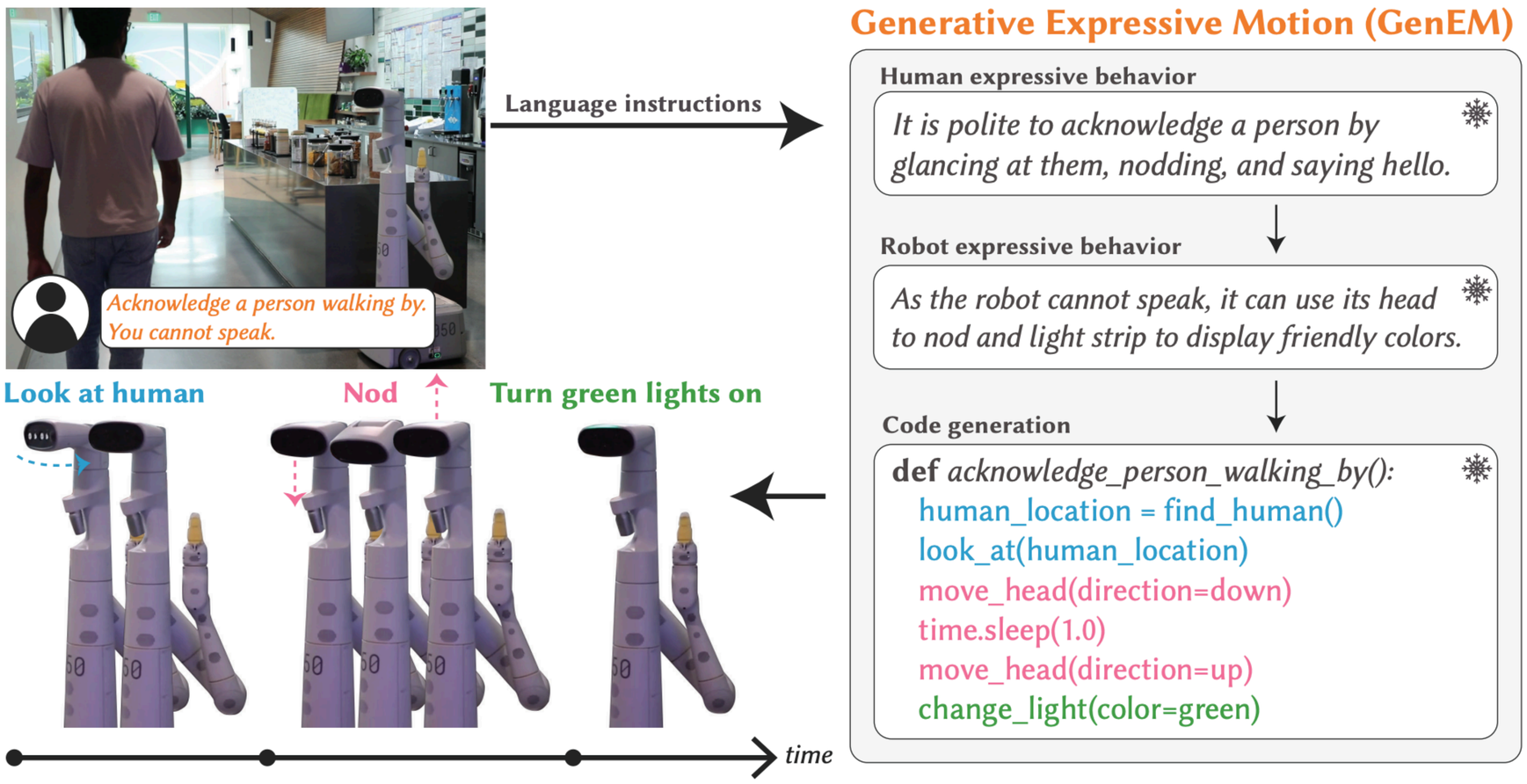
$$\mathbb{E}_{\mathcal{D},\tau,\omega}\Big[\, H\big(x_{1:M}, f_\theta^\ell(\mathbf{o}_t, \ell)\big) + \alpha \left\| \omega - \mathbf{a}_{t:t+H} - f_\theta^a(\mathbf{a}_{t:t+H}^{\tau,\omega}, \mathbf{o}_t, \ell) \right\|^2 \Big],$$

Async Deployment

- Size of modern VLAs makes them difficult to deploy at the edge.
- One solution: use VLA when you can and on-board compute when you can’t.



Code Generation for Robots



Generative Expressive Robot Behaviors using Large Language Models

Summary

Today we covered:

1. What it means for language to be grounded.

2. The SayCan approach as an example method for grounding advanced LLMs.

3. An example VLA and how it is trained on web and robot data.

Action Items

Impact reading for next week.

Complete final project.