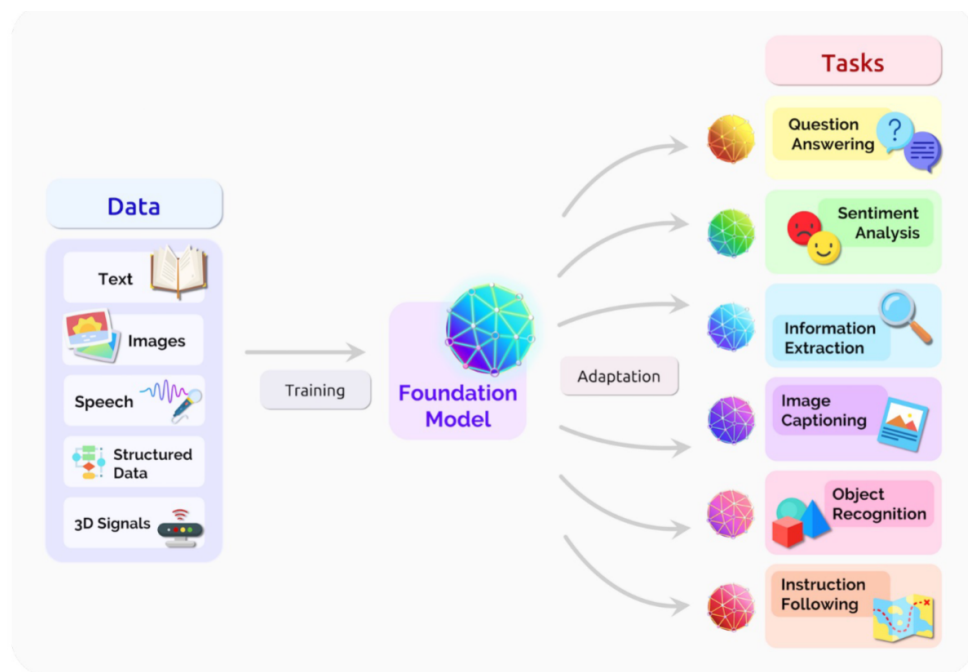# CS 188 Robotics Week 8

Aidan Jan

May 27, 2025
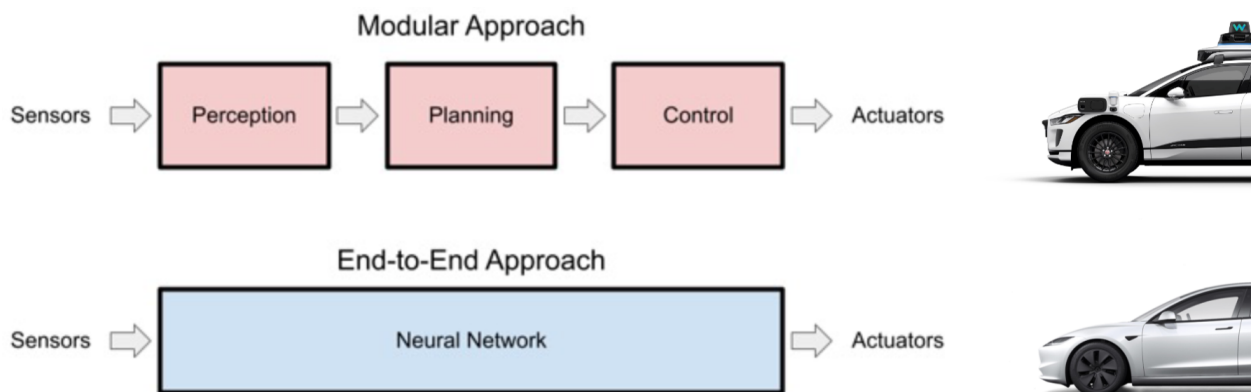
## Foundation Models

### The New Era of ML



Bommasani, Rishi, et al. **"On the opportunities and risks of foundation models."** (2021).

### Two Paradigms of Robotics

- The modular approach breaks down perception, planning, and control into three separate problems, each of which can be solved using some form of artificial intelligence.

  - Perception is the common computer vision problem of image recognition and classification.
  - Planning is the step to decide what the robot should do given its perception.
  - Control converts the plan into commands for hardware parts or actuators.
  - All of these sections can be solved with foundation models

- The end-to-end approach involves collecting data and throwing it into a neural network.

## Integrating FMs in Robotics

1. FM as a **Perception** or Representation Backbone

2. FM as a **Planner or Reasoning** Engine

3. FM as an **End-to-End** Action Policy Prior or Decision-Maker

## Computer Vision FMs

- **CLIP (Contrastive Language-Image Pretraining)** learns the relationship between **a whole sentence** and the **image** it describes
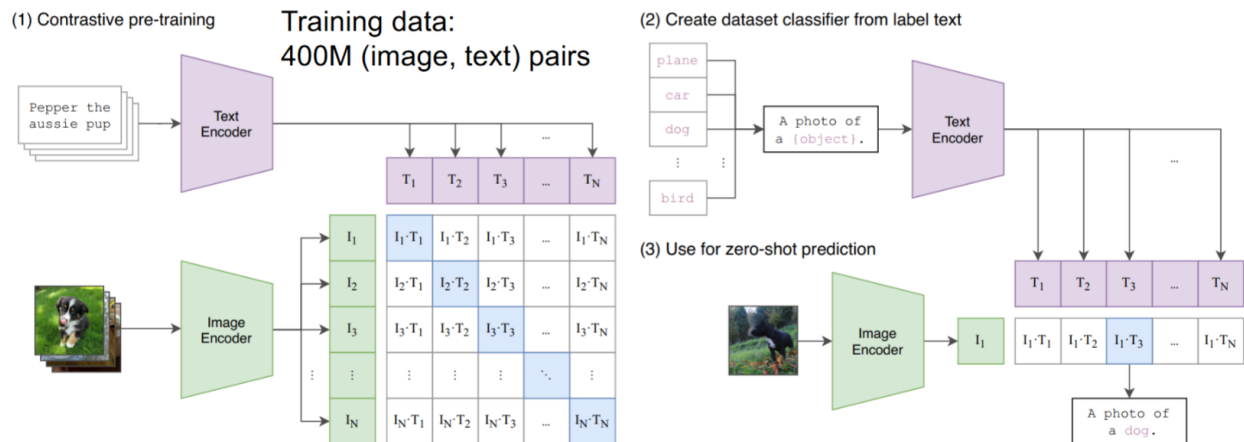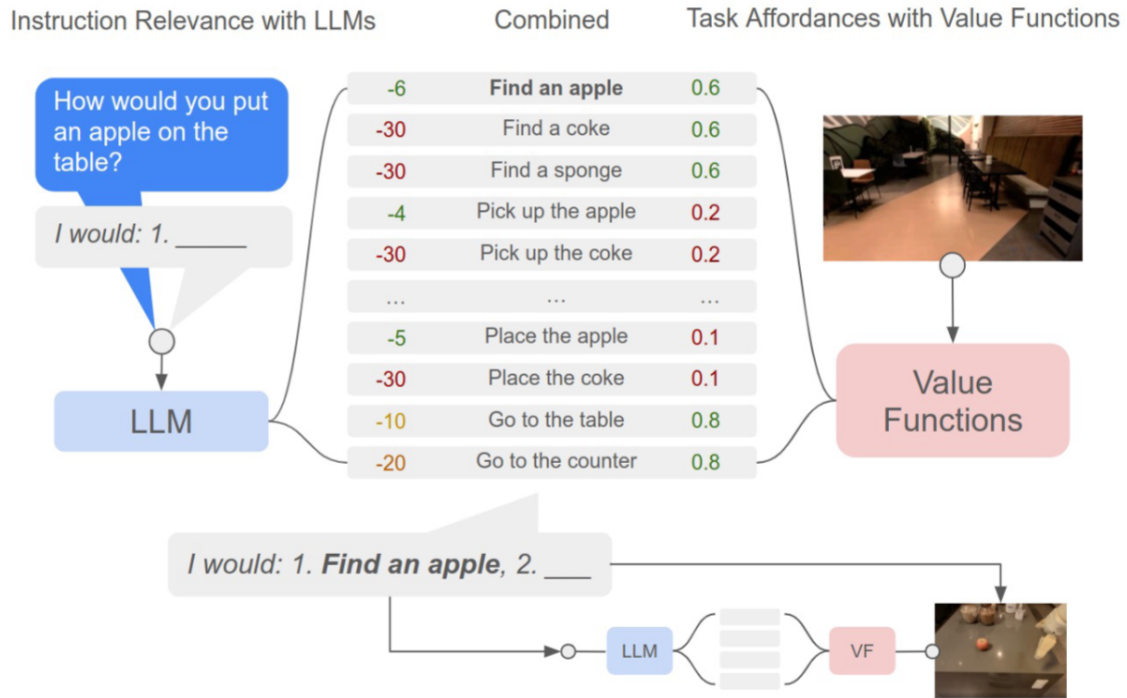


### Example: SAM

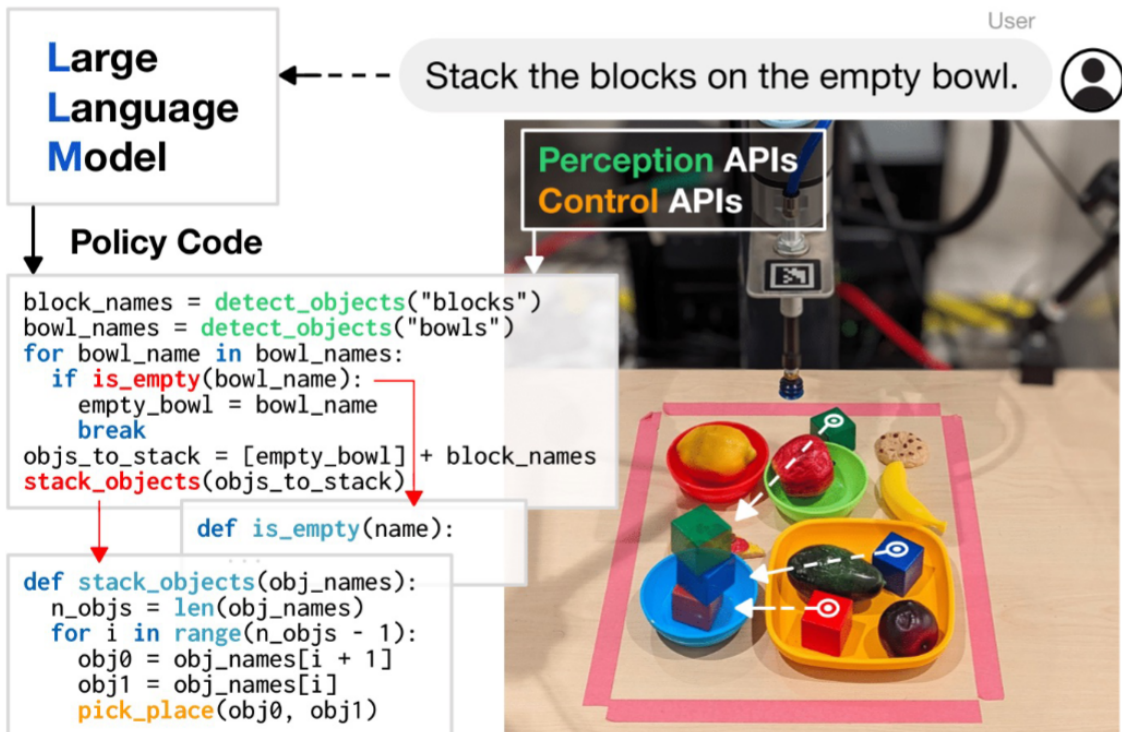SAM: Segment Anything Model

- Website: https://segment-anything.com/demo

- Documentation: https://docs.ultralytics.com/models/sam/

**LLM as Planner**

SayCan



**LLM / VLM Code Generation**

Code As Policies (https://code-as-policies.github.io/)

**Vision Language Action Models (VLAs)**

- Multi-task

- Language-conditioned



- https://openvla.github.io/
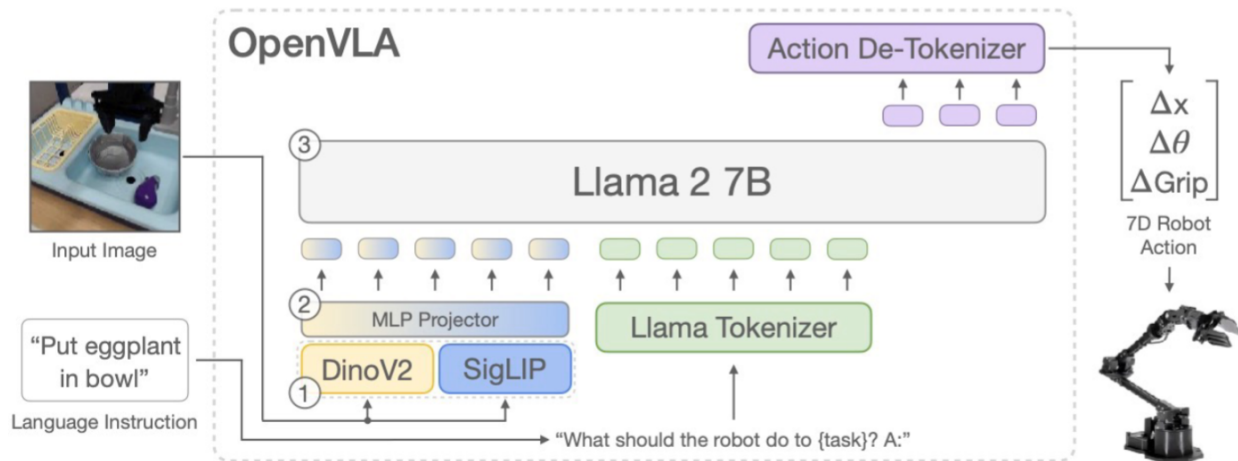
## Challenges?

### Challenges in Embodiment and Grounding

- FMs lack embodied experience; they process abstract data (e.g., text, images), not the physical effects of actions.

- Aligning abstract representations (like "grasp the cup") with real motor commands and sensor feedback.

- A model that understands "grasping" in language may still not control a gripper effectively.

### Challenges in Data Efficiency and Cost

- Robotic data is expensive and slow to collect compared to web-scale text/image data.

- FMs need to learn or adapt using relatively little robotic data.

- Scaling up robotic training like we do with vision/language is not yet feasible.

### Challenges in Real-Time Constraints

- FMs are large and computationally intensive.

- Running inference fast enough for real-time control on embedded hardware.

- Robots often need millisecond-level decision-making; FMs often can't meet these latencies natively.

### Challenges in Safety and Reliability

- FMs can be unpredictable or "hallucinate."

- Ensuring safe, explainable, and fail-safe actions in physical environments.

- Mistakes can cause real-world damage or harm.

**Challenges in Multimodal Integration**

- FMs typically operate on specific modalities (e.g., text, vision).

- Fusing multimodal sensory input (e.g., vision, proprioception, tactile) with high-level reasoning.

- Effective robotic control requires coordinated understanding across many input streams.

**Challenges in Generalization**

- FMs don't always generalize across different embodiments or tasks.

- Making a single model work across varied robots (arms, legs, drones) and diverse manipulation goals.

- One of the key promises of FMs is general-purpose utility, but robotics tasks are often highly specific.