



Introduction to Embodied AI

Lecture 1 - Overview

Prof. He Wang



About Me

- 王鹤
- Assistant Professor, Center on Frontiers of Computing Studies (CFCS), School of CS
- Founder and CTO, Galbot
- lead Embodied AI center of BAAI, starting 2022
- Received Ph.D. from Stanford in 2021
- Received Bachelor from Tsinghua in 2014
- Our lab: *Embodied Perception and Interaction (EPIC) Lab*
- Research interest: Embodied AI
- Homepage: <https://hughw19.github.io/>

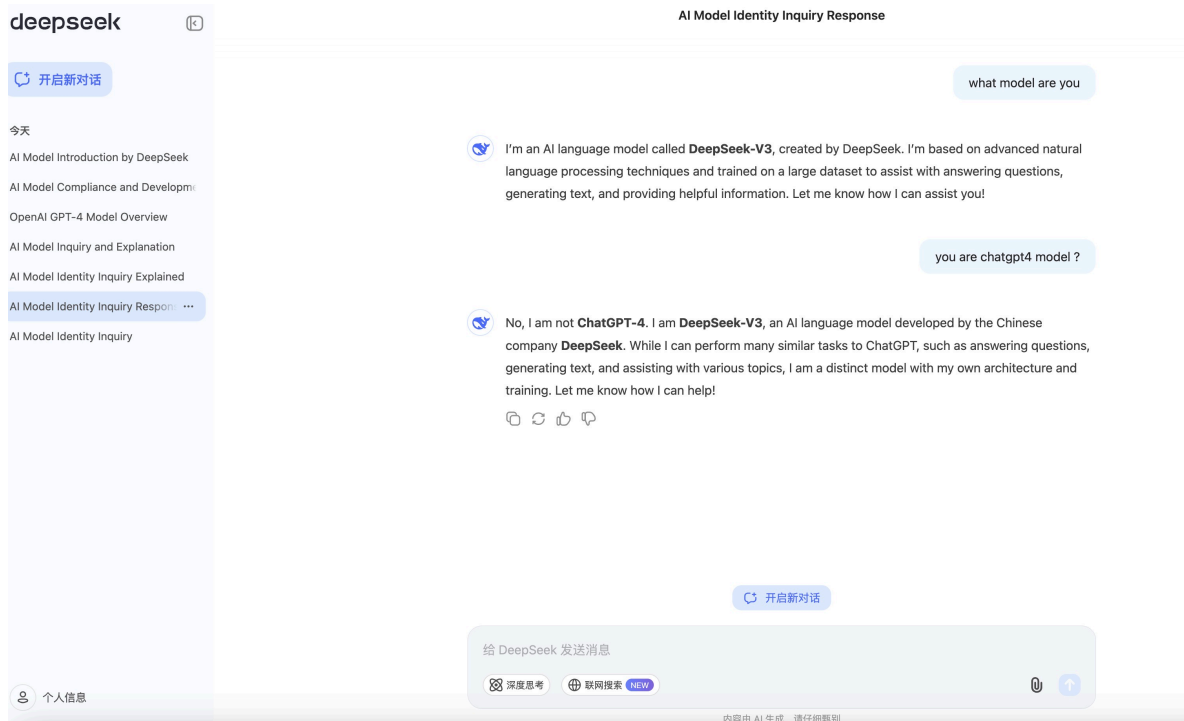


北京大学前沿计算研究中心
Center on Frontiers of Computing Studies, Peking University

GALBOT

BAAI

AI



- DeepSeek



- MidJourney

What do you want AI to do?

What is the best direction of AI?



Joanna Maciejewska—Snakebitten is on preorder now!

@AuthorJMac



You know what the biggest problem with pushing all-things-AI is? Wrong direction.

I want AI to do my laundry and dishes so that I can do art and writing, not for AI to do my art and writing so that I can do my laundry and dishes.

7:50 PM · Mar 29, 2024 · **3M** Views

23.2K Reposts **1,171** Quotes **102.4K** Likes **3,560** Bookmarks

We need AI to do physical labor works!

From Robotics to Embodied AI

What is a Robot?

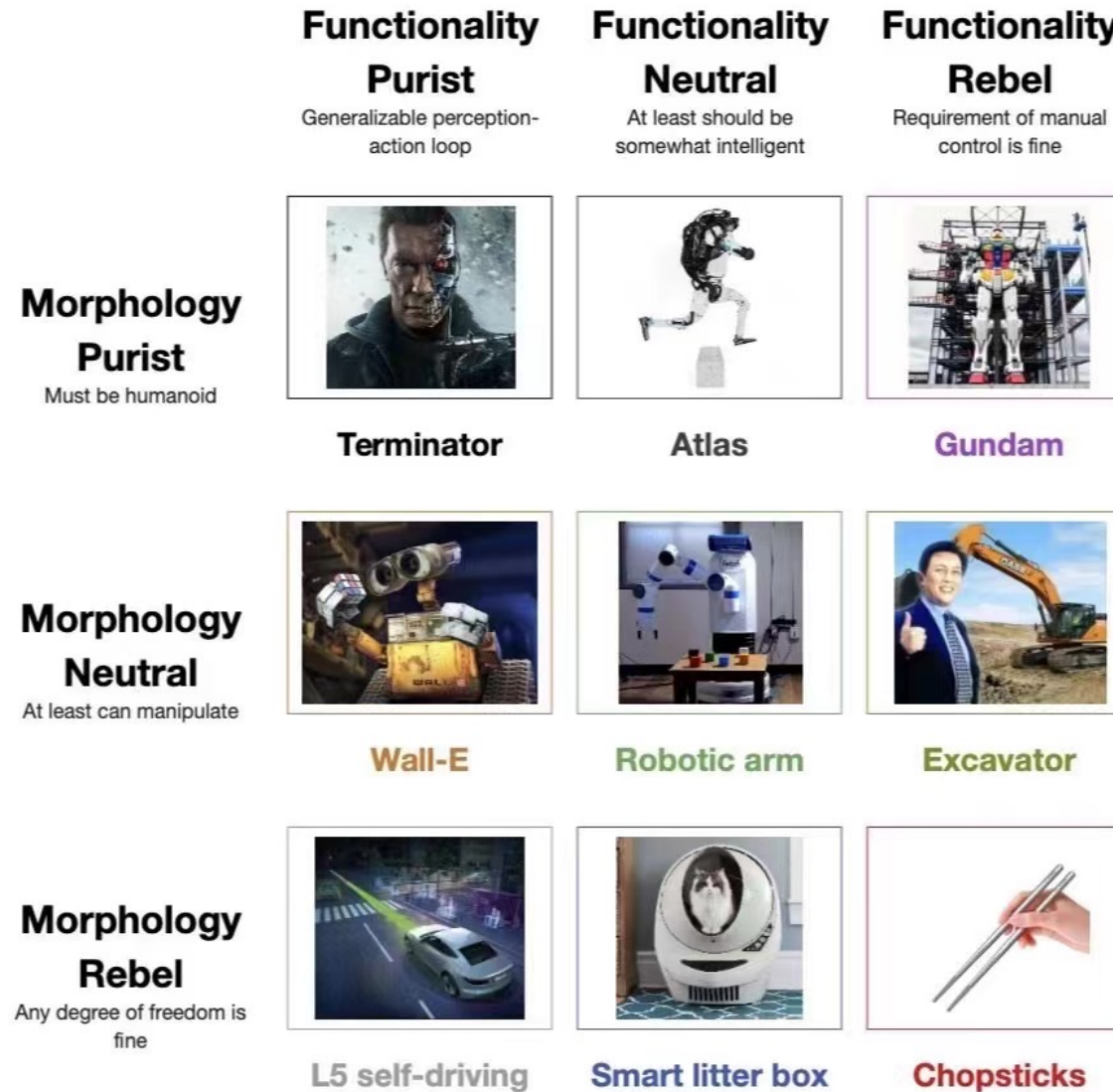
- A Robot is a machine capable of carrying out a complex series of actions automatically.
- Robots can be guided by an external control or attached within.
- Robots may be constructed to take on human form.



Rethink Robotics smart robot Baxter

What is a Robot?

@KuanFang



Types of Robots by Locomotion

- Stationary Robots - Cartesian, Articulated, Cylindrical, Spherical, Parallel.
- Wheeled Robots - Single, two-wheeled, three or more.
- Legged Robots - Bipedal, tripedal, quadrapedal, hexapod and more.
- Drones/Unmanned Aerial Vehicle(UAV) - Single rotor, quadrotor, Fixed wings
- ...



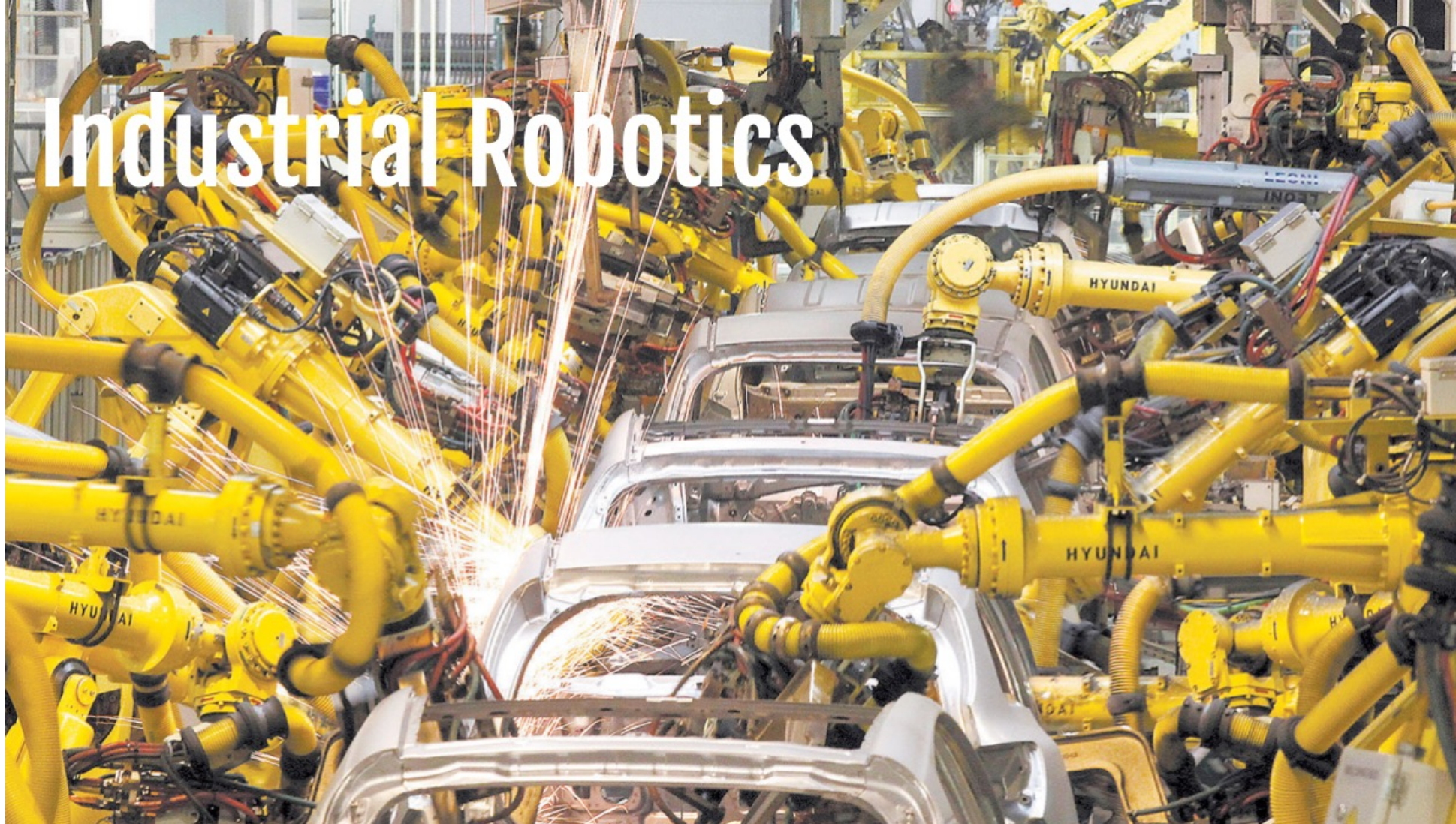
What is Robotics?

- Robotics is a branch of engineering that involves the conception, design, manufacture and operation of [robots](#).
- The objective of the robotics field is to create intelligent machines that can assist humans in a variety of ways.

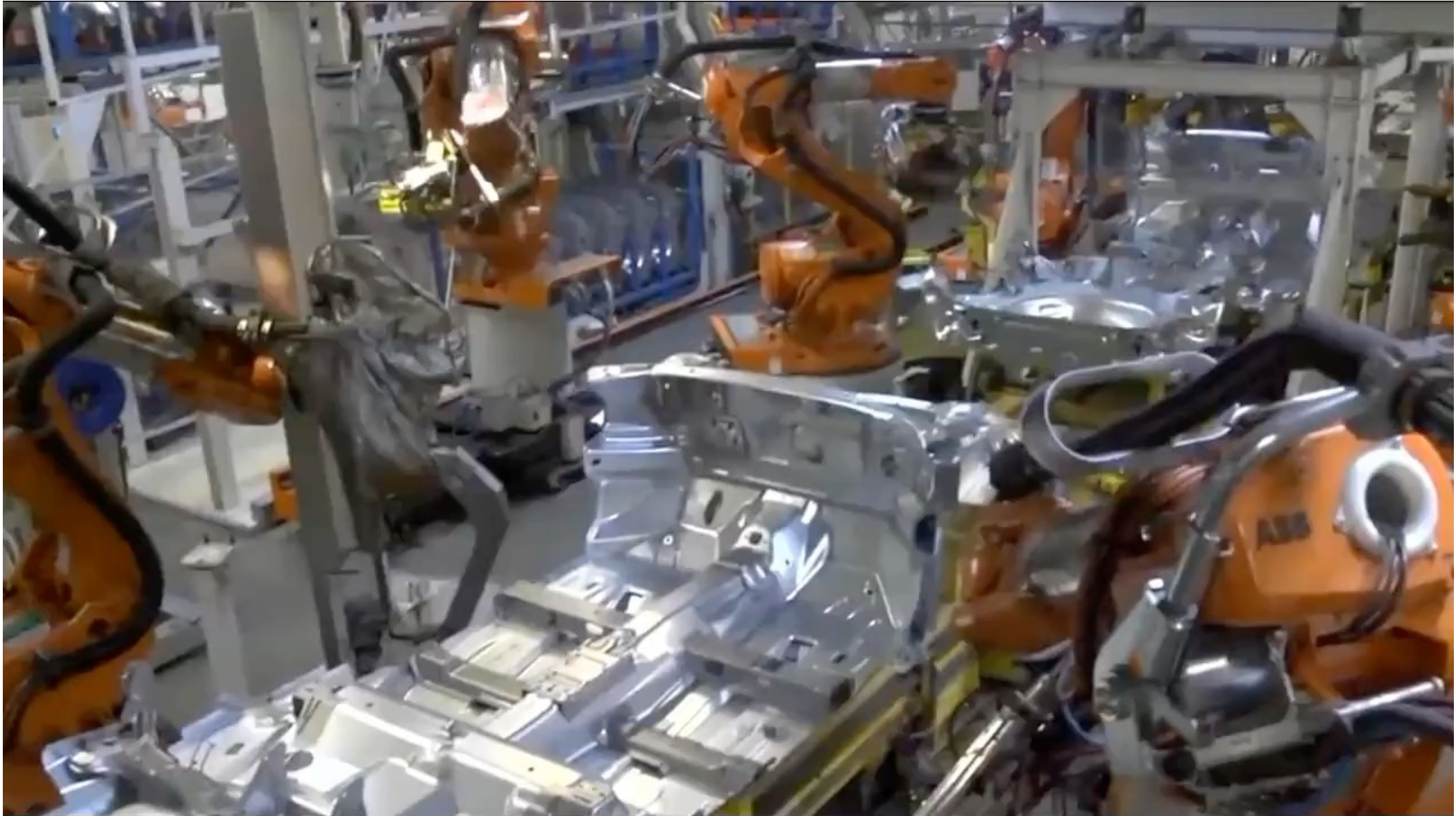


Rethink Robotics smart robot Baxter

Applications of Robotics



Classical Special-Purpose Robots in Car Factory



Classical Special-Purpose Robots in Car Factory

- Predesign and compute the trajectory
- Only replay trajectory when using the arm
- Limitation 1: time-consuming deployment
- Limitation 2: can't flexibly handle multi-tasks

Very different from human intelligence!



Applications of Robotics



Medical Robotics

Applications of Robotics



Applications of Robotics



Warehouse Robotics

Applications of Robotics

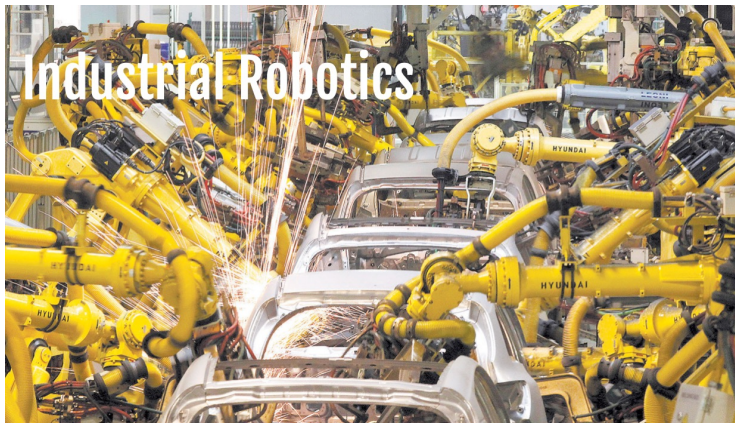


Applications of Robotics



Past, Now and Future of Robotics

Past & Existing



Industrial robots

Now & Happening



Autonomous cars

Future & Our Dream



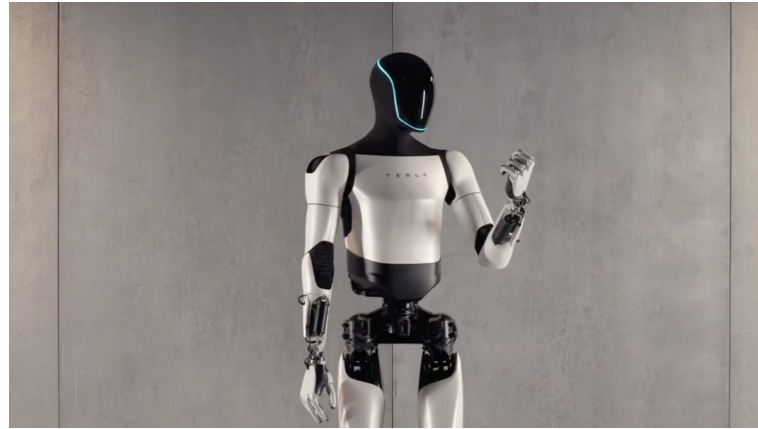
**Generalist Robots
— Humanoids**

Generalist Robots

OpenAI + Figure



Tesla Optimus

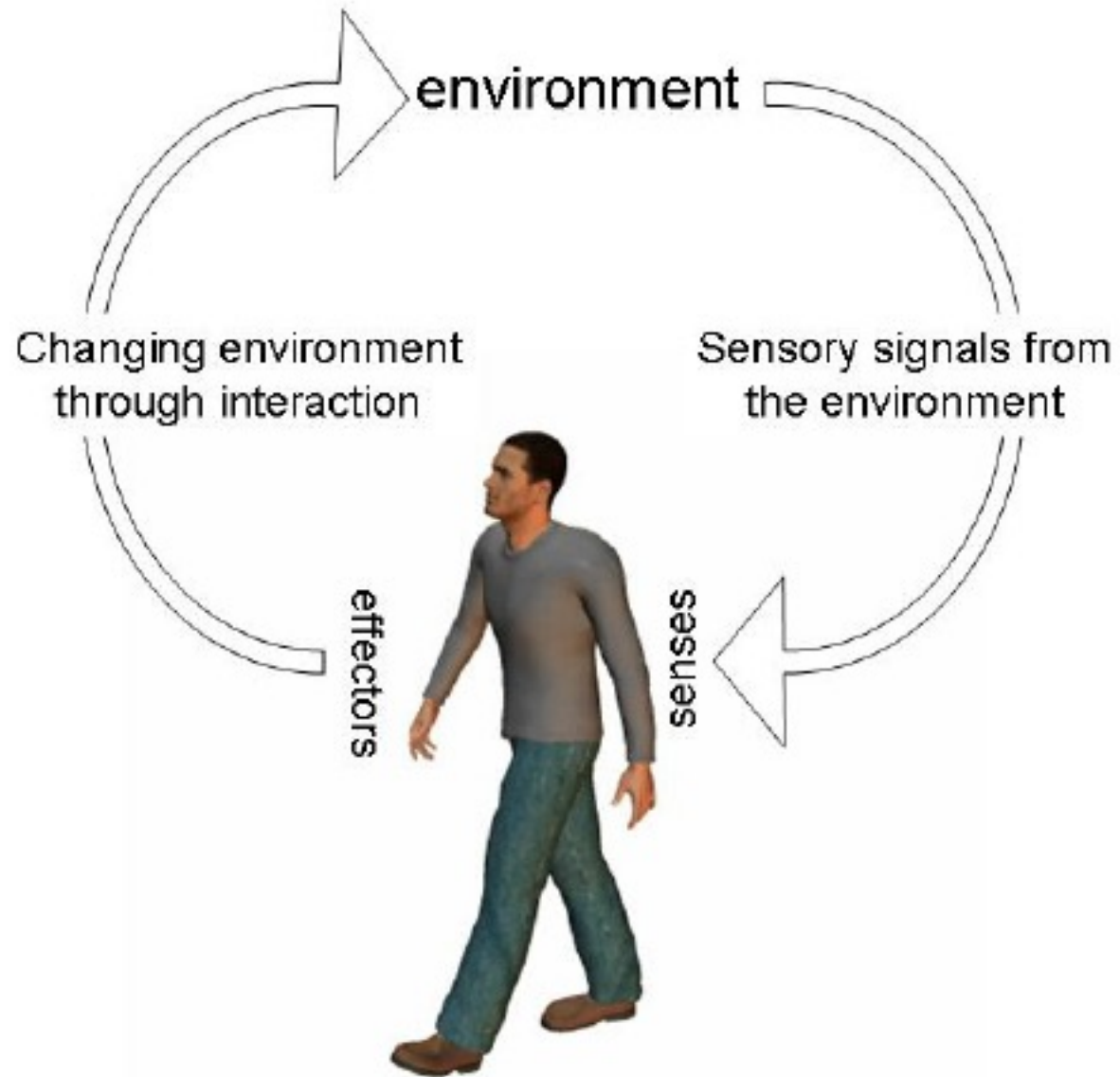


Galbot G1



- Task generalists
- Environment generalists

Perception-Action Loop



Feature 1: High Degree-of-Freedom Body



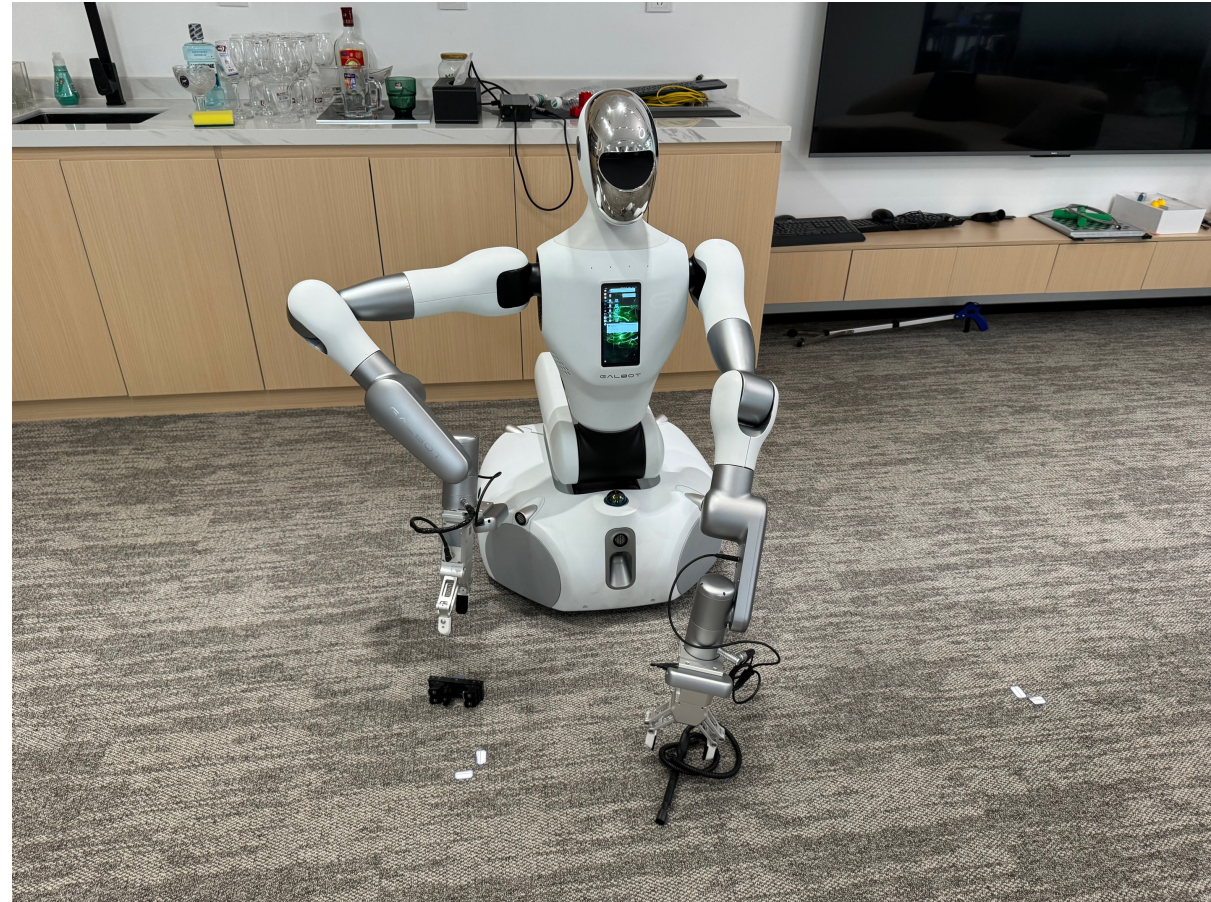
Feature 1: High Degree-of-Freedom Body



Wheel-based Humanoid Robots: Galbot G1

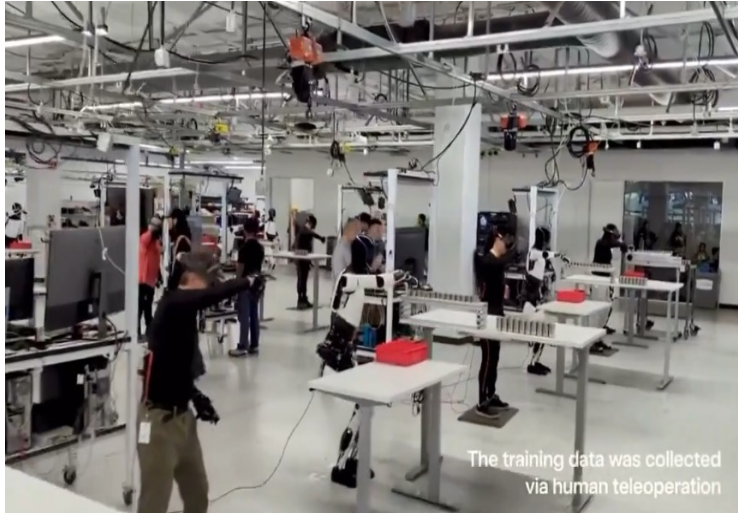


Stand up to reach 2.4 m high



Lower down to pick up objects
on the floor

Feature 2: Data-driven Physical Intelligence

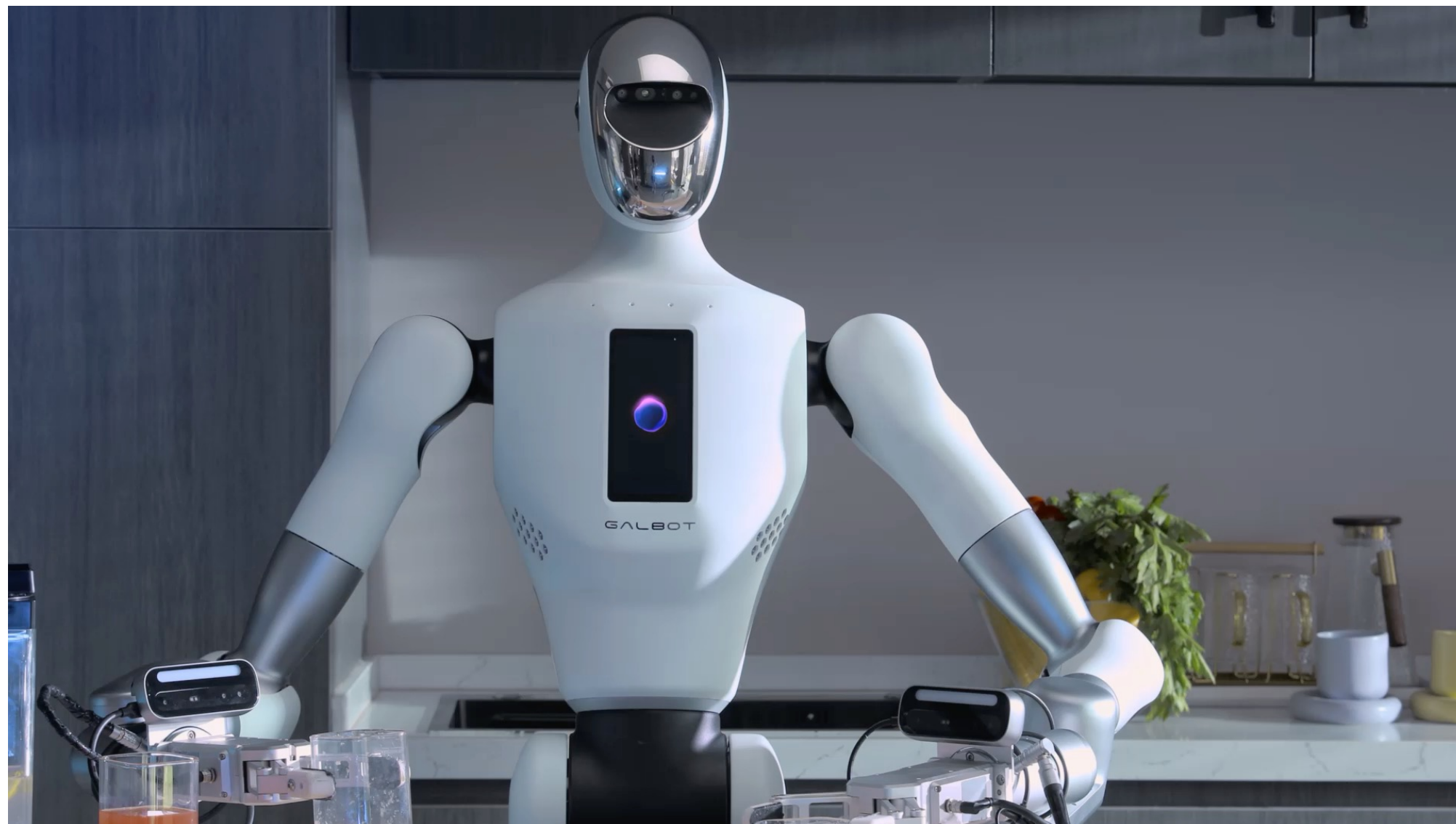


Tesla Data Factory

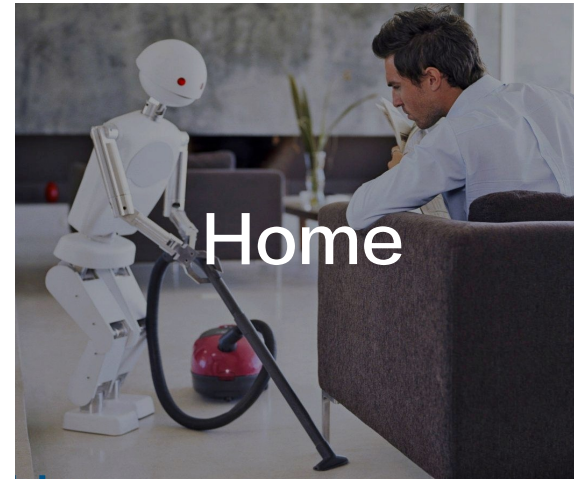
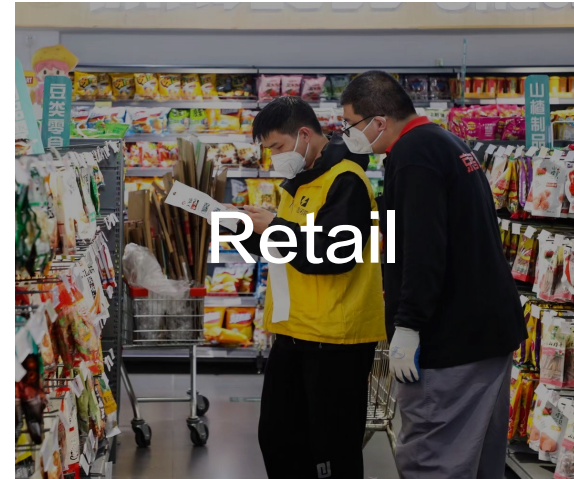


Battery Sorting

Galbot G1



Cross-Industry Applications of Generalist Robots



Revolutionary significance: cross-Industry applications, address future challenges from labor shortages and senior care

Car Factory: Sunroof Glass Loading



机器人装载天窗玻璃
Robot loading the sunroof glass

Car Factory: Box Transferring



Retail: Pick Up



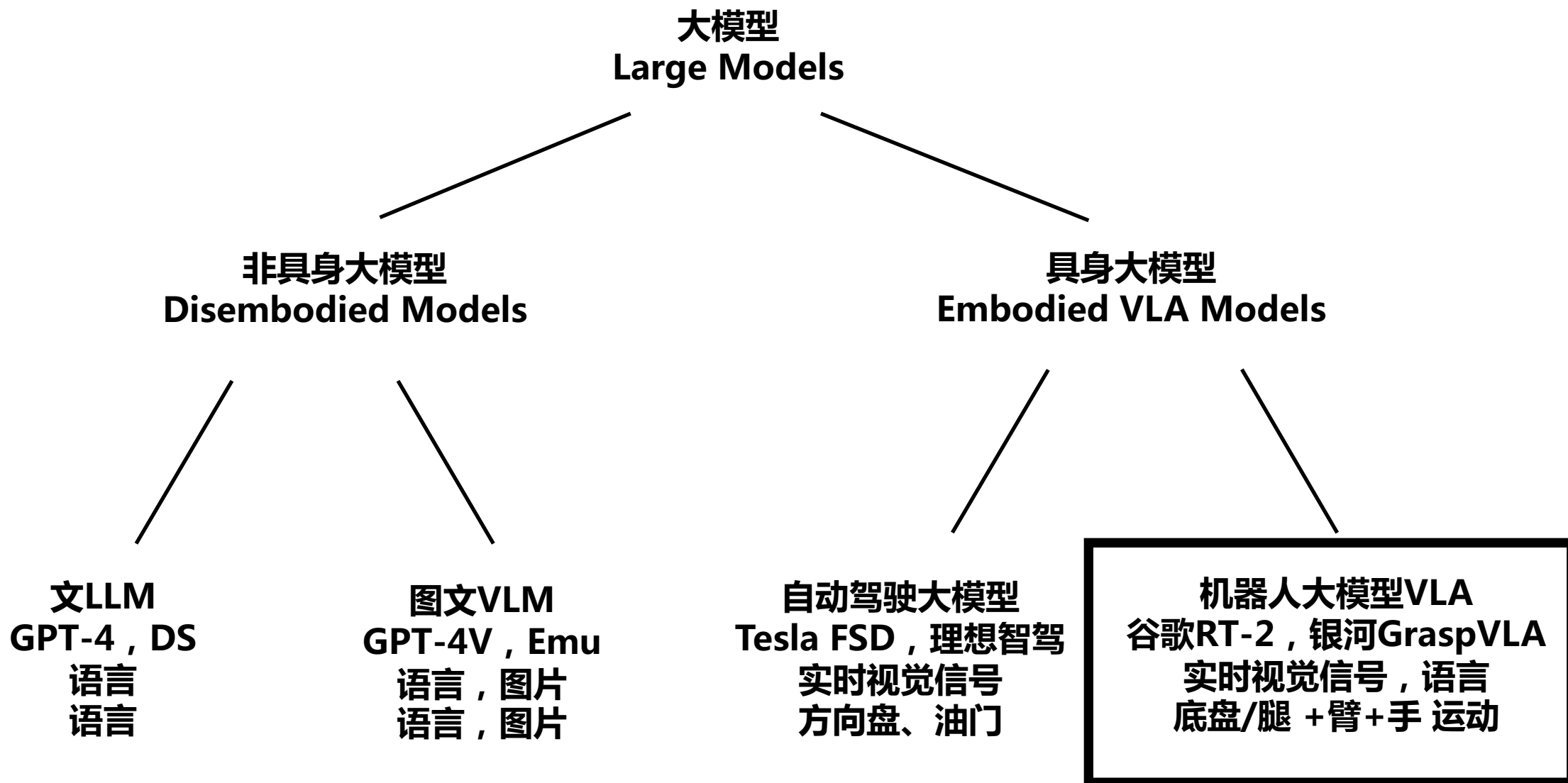
Retail: Restocking



Hotel Reception

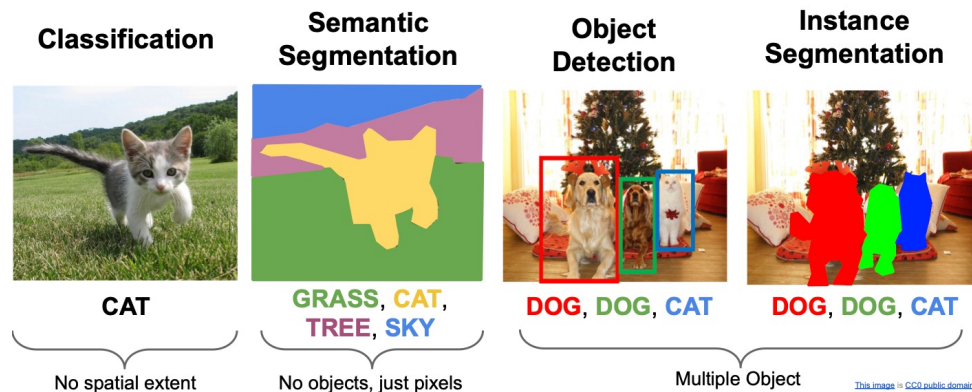


Large Model Collection



What Computer Vision Can Do Well Now?

- Pattern recognition (classification, detection, segmentation ...) in images, videos, and text on datasets typically curated from the internet.



Internet AI

Vision-Language Foundation Model

MM-REACT Bot



AI: 1. This is an image.

AI: 1. This is an image of a kitchen with white cabinets and appliances.

2. There is a glass pitcher with liquid in it on the countertop.

3. There is a black and silver microwave on the countertop.

4. There is a white mug with a cartoon design on it on the countertop.

5. There is a pink container with a handle on the countertop.

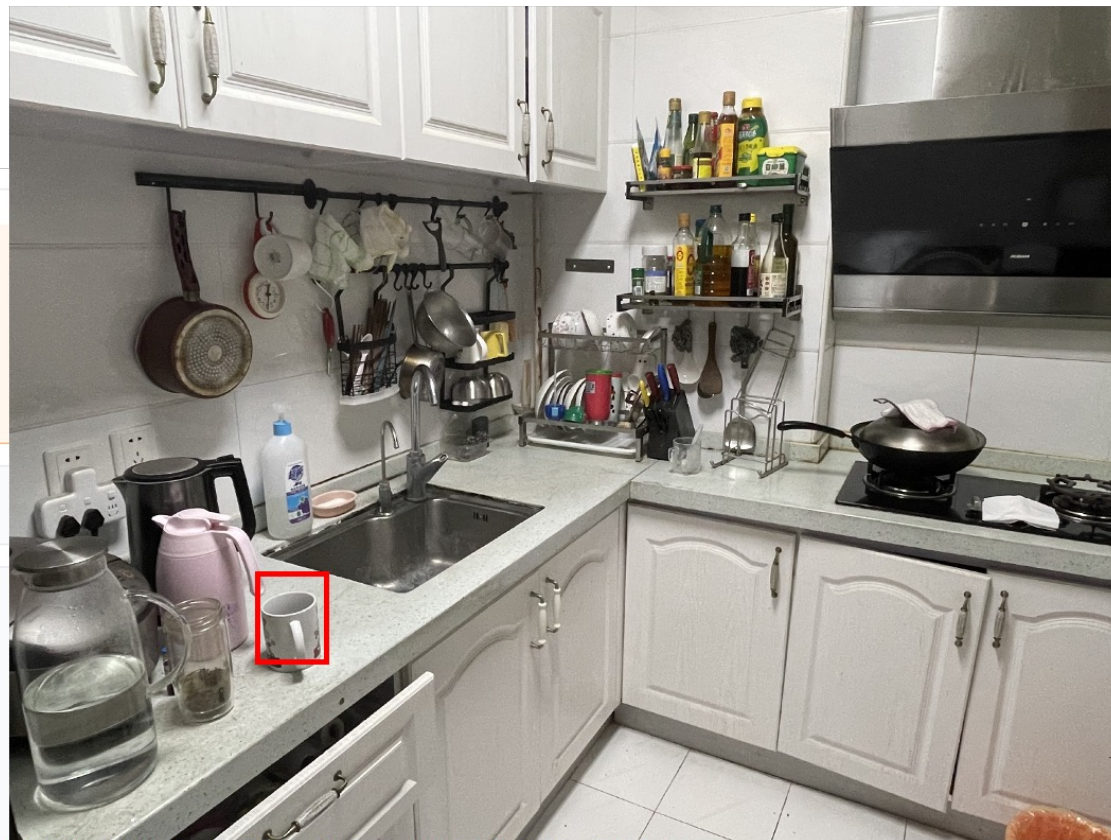
6. There is a white cabinet with a handle on the wall.

7. There is a pan with a white cloth on top on the stove.

8. There are several object tags detected in this image including indoor, countertop, cabinetry, home appliance, sink, cabinet, wall, kitchen stove, gas stove, major appliance, kitchen sink, cooktop, oven, tap, cupboard, plumbing fixture, small appliance, microwave oven, drawer, kitchen appliance, cookware and bakeware, exhaust hood, bottle, kitchen, stove, appliance, room, and pot.

To summarize, this is an image of a kitchen with white cabinets and appliances. There is a glass pitcher, a black and silver microwave, a white mug with a cartoon design, and a pink container on the countertop. There is a white cabinet on the wall and a pan with a white cloth on top on the stove. The image is likely taken indoors in a home or apartment.

Upload a pic and ask!



What Computer Vision Can't Do Well?

- Complete tasks through **interaction** with their environment (enable **action** in an **environment**)
- Bridge the gap between perception and action (**learning by doing**).

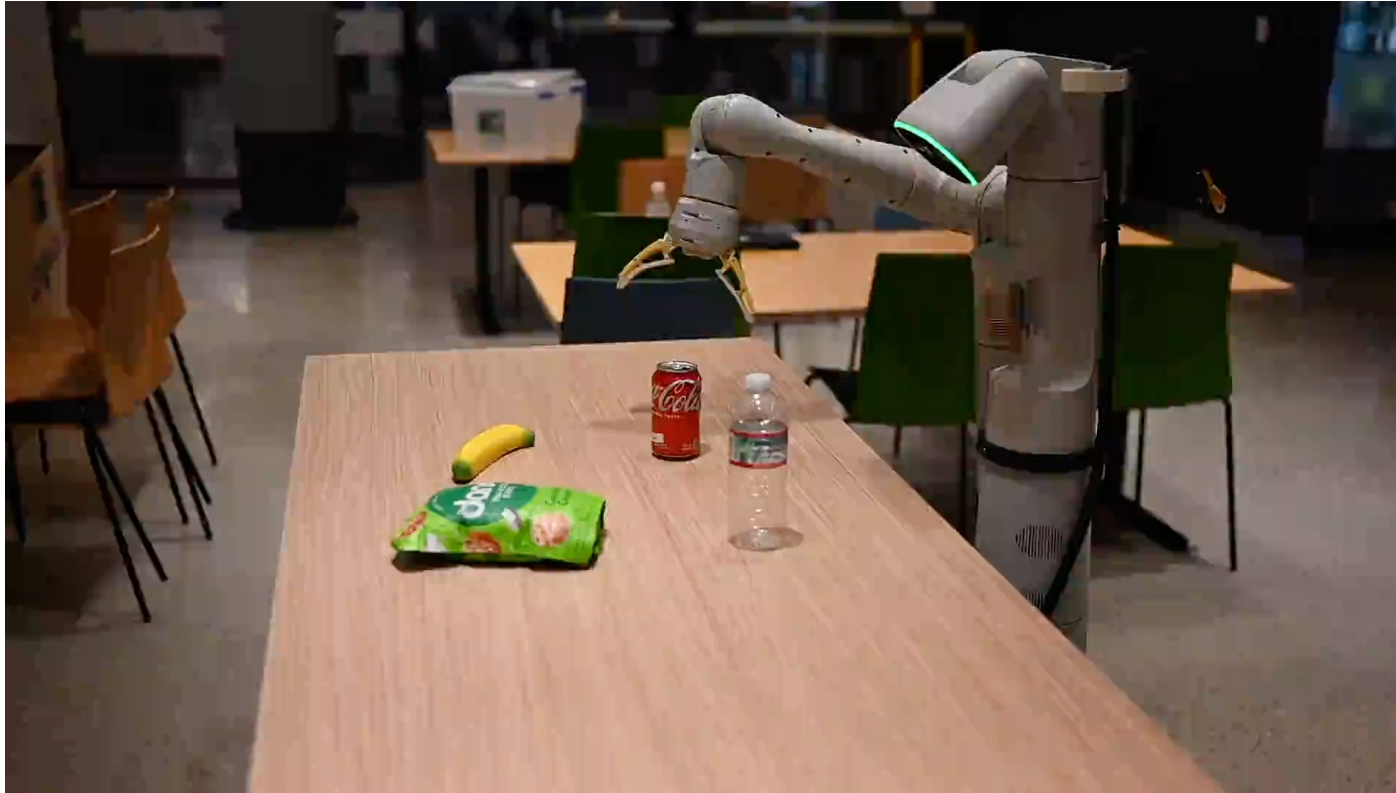


A closed loop
of perception,
cognition and
action !

Embodied AI

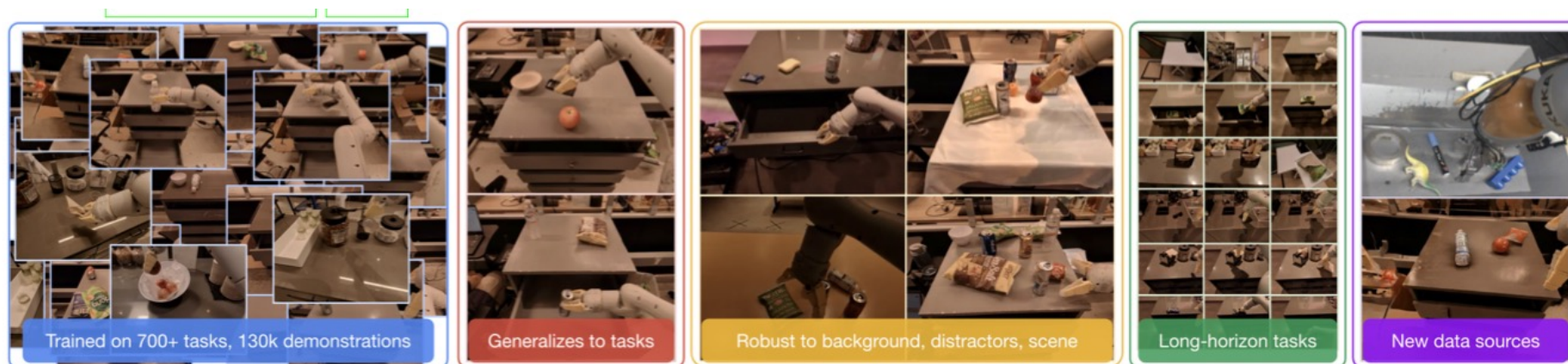
Embodied Large Model: RT-2

Input : language, vision, other proprioceptive / sensor signals
Output : language or action



Google proposed RT-2, a VLA model, to address open-instruction manipulation.

Problem: Data Hungry and Limited Scene Generalization



(b) RT-1's large-scale, real-world training (130k demonstrations) and evaluation (3000 real-world trials) show impressive generalization, robustness, and ability to learn from diverse data.

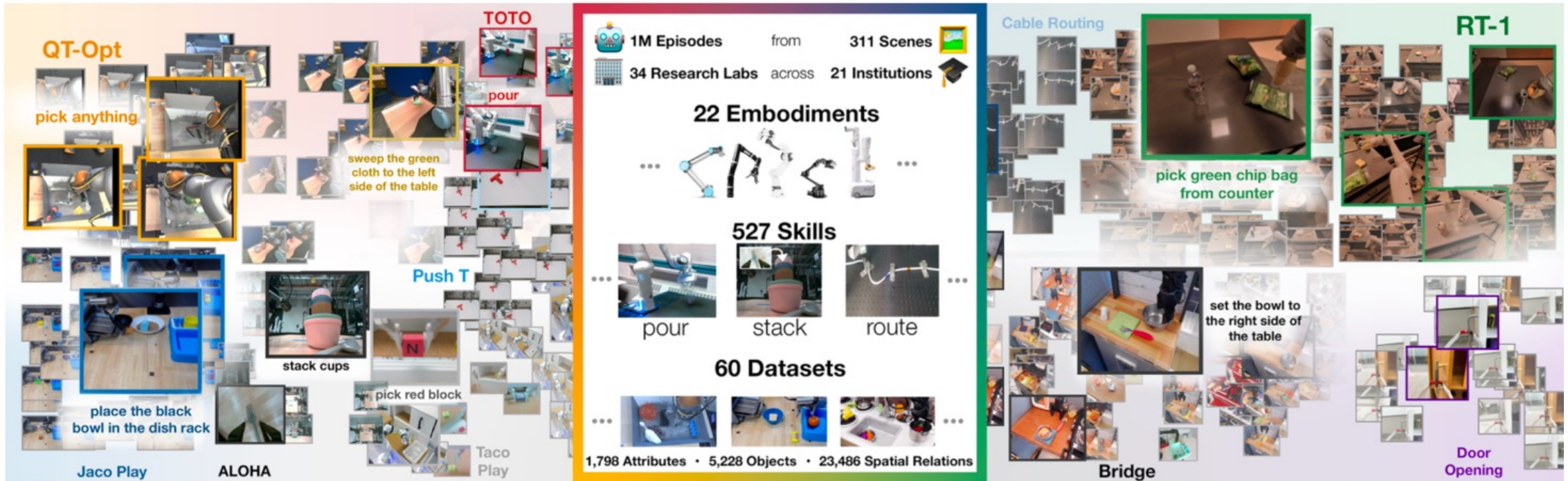
Figure 1: A high-level overview of RT-1's architecture, dataset, and evaluation.

- **Data collection and evaluations:** Noah Brown, Justice Carbajal, Joseph Dabis, Tomas Jackson, Utsav Malla, Deeksha Manjunath, Jodily Peralta, Emily Perez, Jornell Quiambao, Grecia Salazar, Kevin Sayed, Jaspiar Singh, Clayton Tan, Huong Tran, Steve Vega, and Brianna Zitkovich.

Demonstrations are costly (130K demo, 17 months, 16 people, 13 robots)

Real Data Collection

Open X-Embodiment: Robotic Learning Datasets and RT-X Models

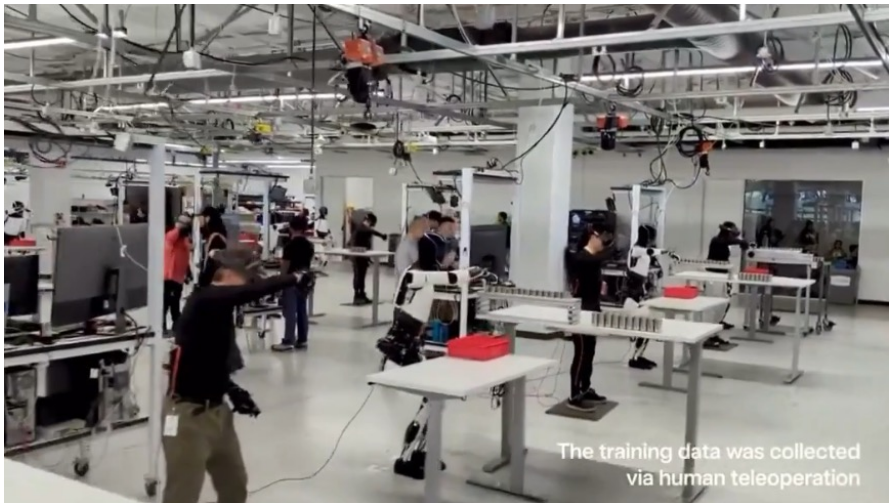


ICRA 2024 Best Conference Paper

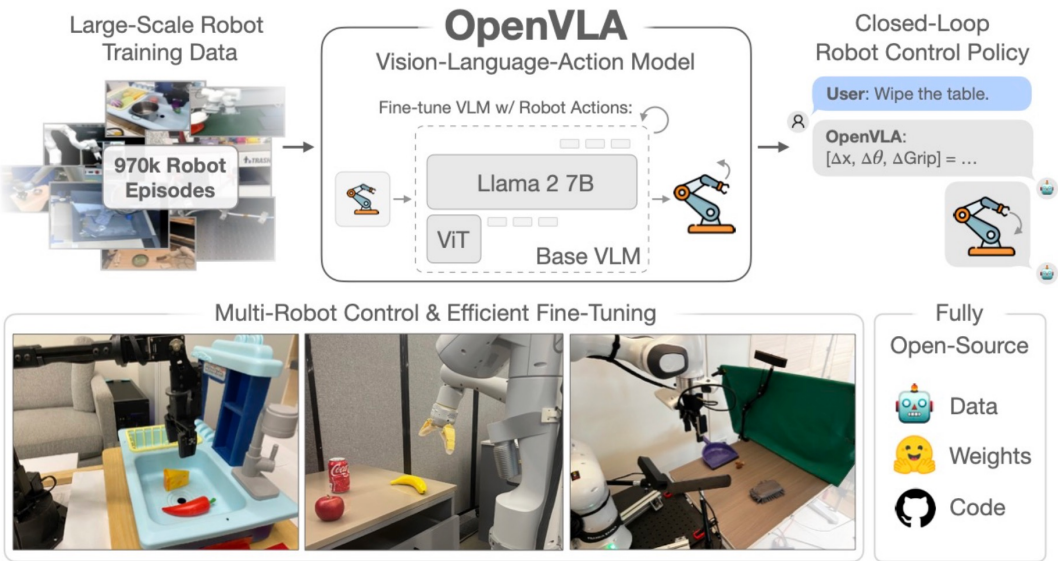
1M+ trajectories, mainly covering pick-and-place, articulated object manipulation, etc.

具身数据难题：真实数据采集贵，难以Scale Up到大模型所需的量

Tesla雇佣40人团队进行人形机器人遥操作产生数据



Stanford Aloha工作进行遥操作采集



大模型可能需要上亿条数据才能泛化！

Large-Scale Synthetic Data



GraspVLA

GALBOT
银河通用



BAAI
百度研究院



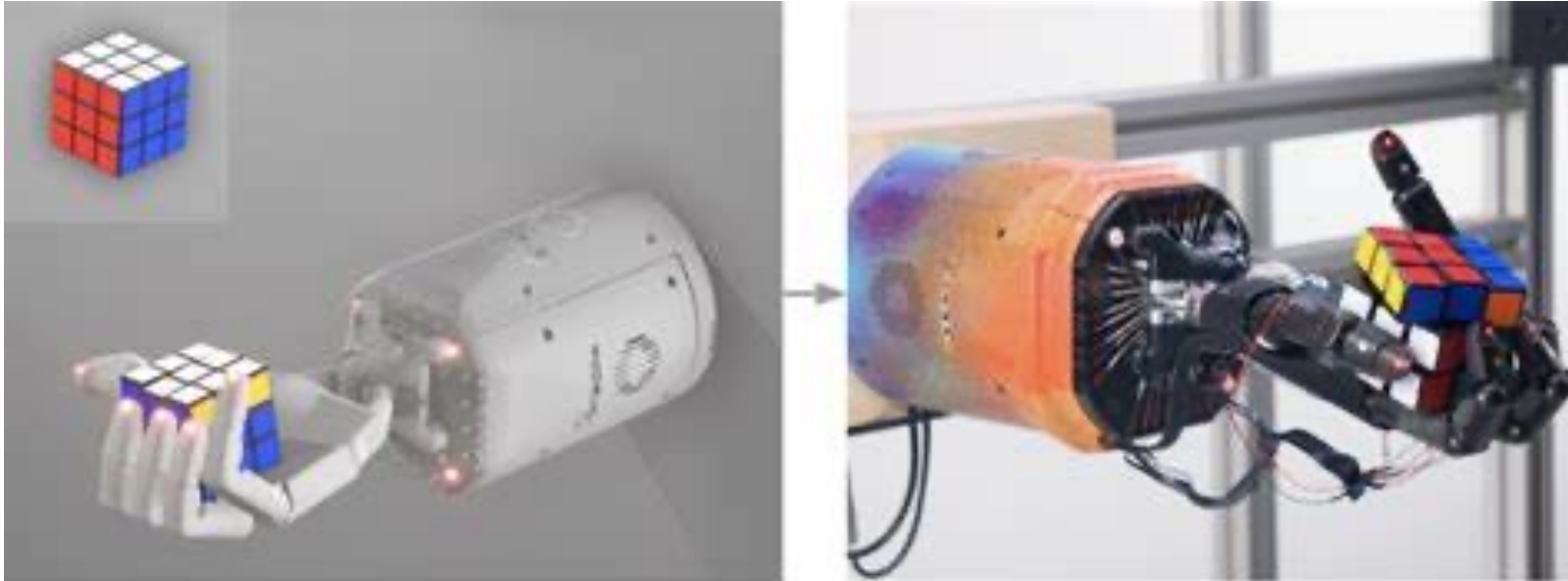
Learning Interaction in Simulation Environments



Simulation and Photorealistic Rendering:

- Annotation-free
- Time efficient
- Transferable to real world

Where to learn? In Sim or Real?



- For problems that is already done in real, learning in simulation requires to generalize to real.
- For problems that is still far to be done in real, simulation environment is a good place for exploration.

Embodied AI

Classic AI



Classical:
“intelligence as
computation”



Problems?

What remains?

- AI still heavily biased toward representation and computation.
- vs. natural (also human) intelligence:
 - Embodied
 - emergent from sensory-motor and interaction processes



Embodied Agents



Embodiment

- “intelligence requires a body”
- Interplay / task distribution
 - Brain
 - Body (morphology – shape, materials, ...)
 - Environment
- Principal of ecological balance
 - match in complexity of sensory, motor, and neural system

Classical AI vs. Embodied AI

- Classical AI
 - Thinking, reasoning, abstract problem solving
- Embodied AI
 - Movement, physical interaction with the real world

“Why do plants not have brains? The answer is actually quite simple: they don’t have to move.”

Lewis Wolpert, UCL

Disembodied and Embodied AI

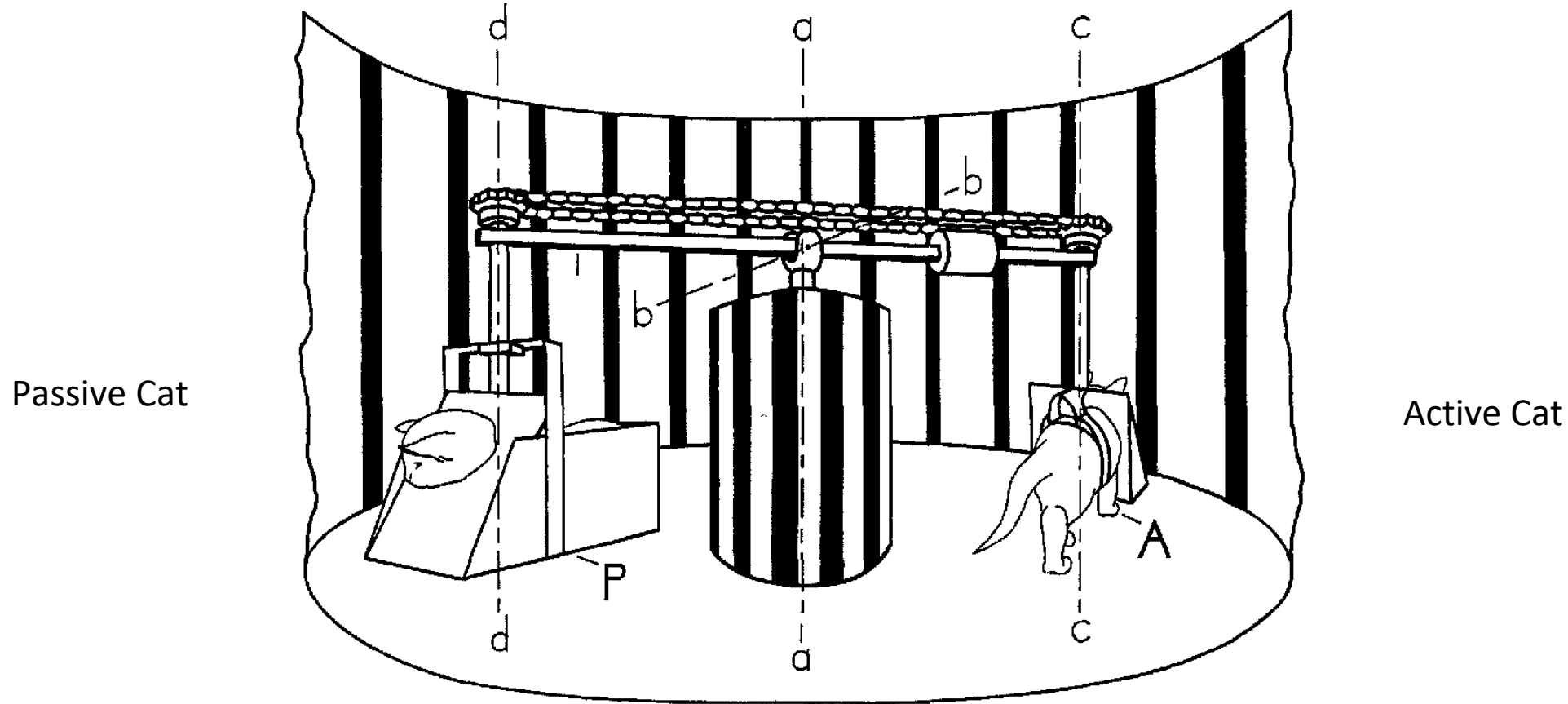


We may hope that machines will eventually compete with men in all purely intellectual fields. But which are the best ones to start with? Even this is a difficult decision. Many people think that a very abstract activity, like the playing of chess, would be best. It can also be maintained that it is best to provide the machine with the best sense organs that money can buy, and then teach it to understand and speak English. This process could follow the normal teaching of a child. Things would be pointed out and named, etc. Again I do not know what the right answer is, but I think both approaches should be tried.

—Alan Turing

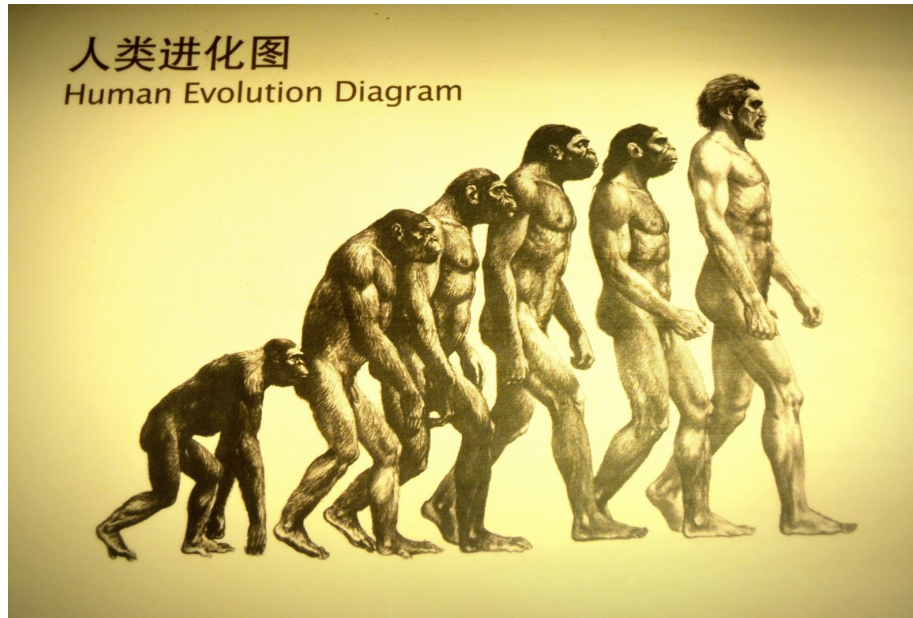
Alan Turing. *Computing Machinery and Intelligence*.

A Brief History of Embodied AI

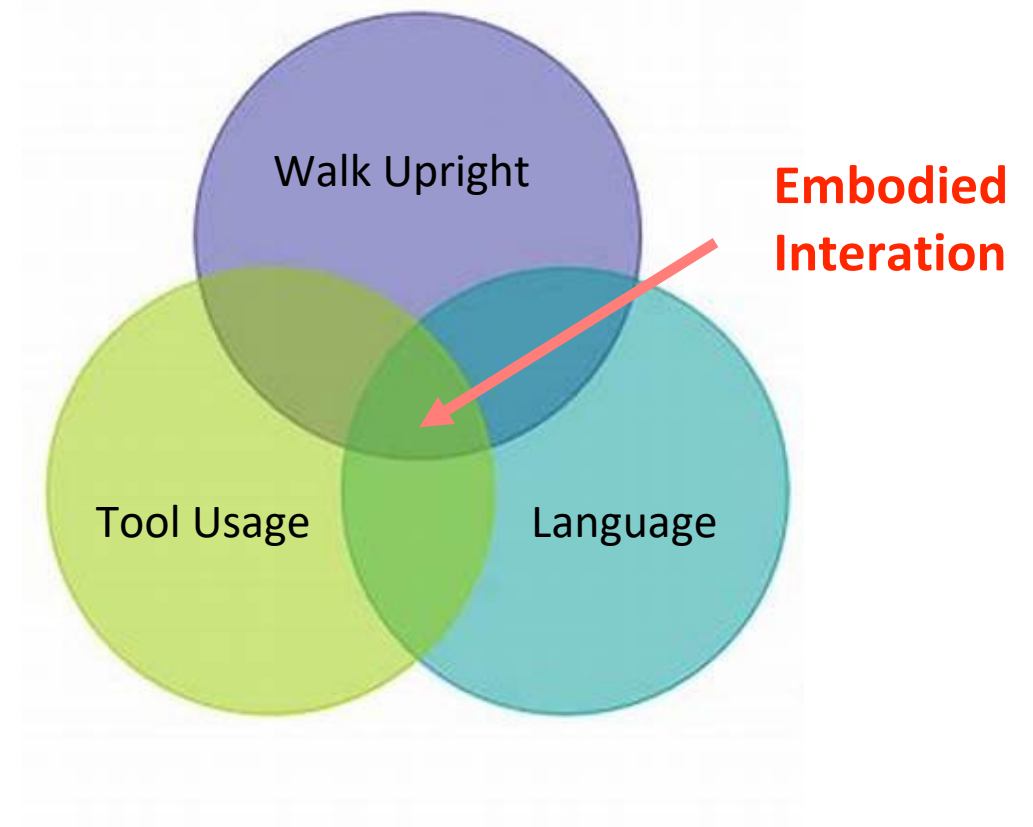


Held, R., & Hein, A. (1963). Movement-produced stimulation in the development of visually guided behavior. *Journal of comparative and physiological psychology*, 56(5), 872.

Evolution of Human Intelligence



The keys to evolution of human intelligence:



Visual Motor Coordination/Integration

- Visual motor control is the ability to coordinate visual information with motor output, where the eyes provide sensory feedback to adjust body motion.
- It is crucial for coordinating the hands, legs, and the rest of the body's movements with what the eyes perceive.

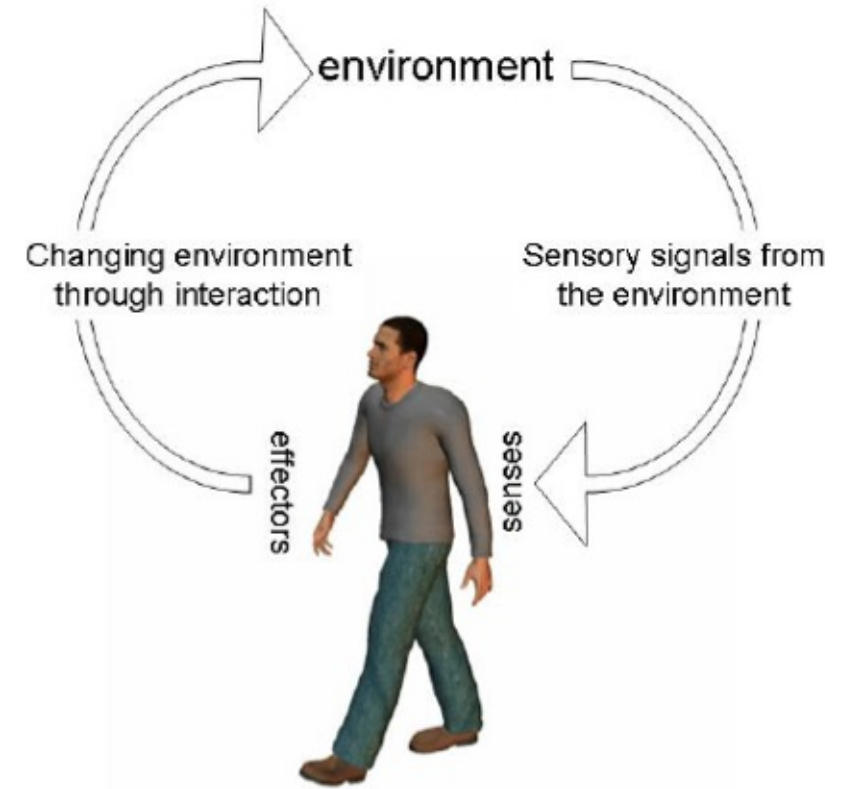
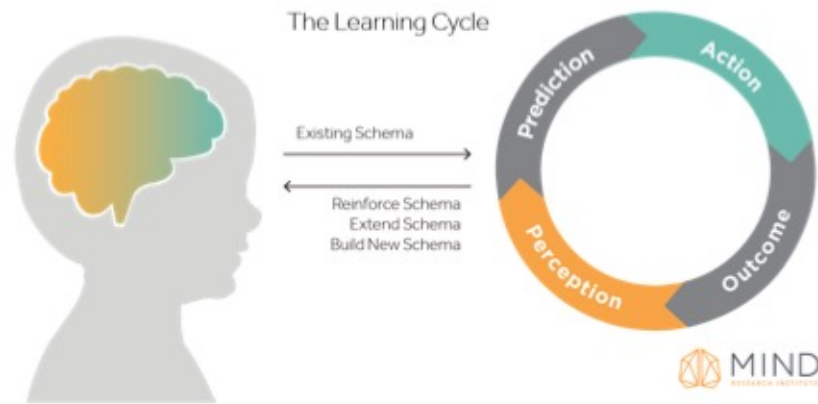


Examples of Visuomotor coordination: Eye-Hand Coordination



How Humans Learn: Perception-Action Loop

- Perceive, forms hypotheses, and then take action to examine.
- Our brain makes sense of the world around us by creating and testing hypotheses about the way the world works.



Perception-Action Loop

Embodied AI and Embodied Vision

- 👁️ **See:** perceive their environment through vision or other senses.
- 🗣️ **Talk:** hold a natural language dialog grounded in their environment.
- 👂 **Listen:** understand and react to audio input anywhere in a scene.
- 📍 **Act:** navigate and interact with their environment to accomplish goals.
- 🧠 **Reason:** consider and plan for the long-term consequences of their actions.

Embodied AI is the field for solving AI problems for virtual robots that can move, see, speak, and interact in the virtual world and with other virtual robots — these simulated robot solutions are then transferred to real world robots.

--- Luis Bermudez, Overview of Embodied Artificial Intelligence



Home robots

Course Logistics

Objective: A Pioneering Course on Embodied AI

- A frontier course on Embodied AI
- Covering basics in robotics and deep learning based vision and robotic system, from a modern perspective of Embodied AI
- To lay a solid foundation for conducting research in Embodied AI

Logistics

- Instructor
 - He Wang (hewang@pku.edu.cn)
- TAs:
 - Haoran Liu (jiangran.lyu@stu.pku.edu.cn)
 - Hao Shen (dotandung@stu.pku.edu.cn)
- Class Time & Location
 - Monday 3:10PM - 6:00PM
 - In-person: Room 207, Teaching Building 3

Prerequisite

- Courses:
 - Deep Learning
 - Computer Vision
 - University-level Mechanics
- Coding:
 - Proficiency in Python
 - Familiar with PyTorch

Course Website

- Website accessible to everyone:
- https://pku-epic.github.io/Intro2EAI_2025/
 - Syllabus
 - Publishing Assignments
- Internal course website in <https://course.pku.edu.cn/>
 - Slides download
 - Assignment submission
 - Grades
 - Discussion board



Syllabus

EVENT	DATE	TOPIC	CONTENTS
Lecture	02/17/2025 15:10 Monday	Overview	Overview of Embodied AI



Resources

- No required textbooks.
- Online Courses for references:
 - On Deep Learning and Computer Vision:
 - Ian Goodfellow and Yoshua Bengio and Aaron Courville. *Deep Learning*. MIT Press, 2016.
 - [Stanford CS 231N](#)
 - On Reinforcement Learning:
 - [Berkeley CS 285](#)
 - On Robotics:
 - [UCSD CSE291](#)

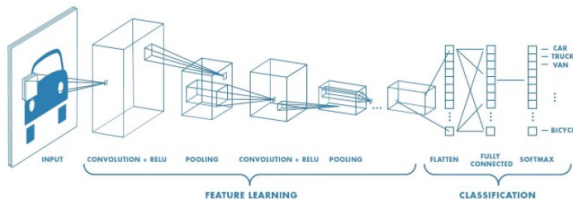
The most effective way to learn and check things

- Just google it!
- Search in wikipedia!

tds Published in Towards Data Science · Follow

 Sumit Saha
Dec 16, 2018 · 7 min read

A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way



Artificial Intelligence has been witnessing a monumental growth in bridging the gap between the capabilities of humans and machines. Researchers and enthusiasts alike, work on numerous aspects of the field to make amazing things happen. One of many such areas is the domain of Computer Vision.

The agenda for this field is to enable machines to view the world as humans do, perceive it in a similar manner and even use the knowledge for a multitude of tasks such as Image & Video recognition, Image Analysis & Classification, Media Recreation, Recommendation Systems, Natural Language Processing, etc. The advancements in Computer Vision with Deep Learning has been constructed and perfected with time, primarily over one particular algorithm — a **Convolutional Neural Network**.

WIKIPEDIA
The Free Encyclopedia

Not logged in · Talk · Contributions · Create account · Log in

Article · Talk · Read · Edit · View history · Search Wikipedia

Generative adversarial network

From Wikipedia, the free encyclopedia

Not to be confused with Adversarial machine learning.

A **generative adversarial network (GAN)** is a class of machine learning frameworks designed by Ian Goodfellow and his colleagues in June 2014.^[1] Two neural networks contest with each other in a game (in the form of a zero-sum game, where one agent's gain is another agent's loss).

Given a training set, this technique learns to generate new data with the same statistics as the training set. For example, a GAN trained on photographs can generate new photographs that look at least superficially authentic to human observers, having many realistic characteristics. Though originally proposed as a form of *generative model* for unsupervised learning, GANs have also proved useful for semi-supervised learning,^[2] fully supervised learning,^[3] and reinforcement learning.^[4]

The core idea of a GAN is based on the "indirect" training through the discriminator, another neural network that is able to tell how much an input is "realistic", which itself is also being updated dynamically.^[5] This basically means that the generator is not trained to minimize the distance to a specific image, but rather to fool the discriminator. This enables the model to learn in an unsupervised manner.

Contents [hide]

1 Method

2 Applications

2.1 Fashion, art and advertising

2.2 Science

2.3 Video games

2.4 Concerns about malicious applications

2.5 Transfer learning

2.6 Miscellaneous applications

3 History

4 Classification

4.1 Bidirectional GAN

5 References

6 External links

Part of a series on

Machine learning and data mining

Problems

Supervised learning

(classification · regression)

Clustering

Dimensionality reduction

Structured prediction

Anomaly detection

Artificial neural network

Autoencoder · Cognitive computing · Deep learning · DeepDream · Multilayer perceptron · RNN (LSTM · GRU · ESN) · Restricted Boltzmann machine · GAN · SOM · Convolutional neural network (U-Net) · Transformer (Vision) · Spiking neural network · Memristor · Electrochemical RAM (ECRAM)

Reinforcement learning

Theory

Machine-learning venues

Related articles

V · T · E

Courseworks and Grading Policy

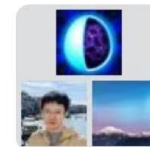
- 3 assignments: in total 40%
- 1 real-world lab: 20%
- Class participation: up to 5% bonus
- Look up class schedule (https://pku-epic.github.io/Intro2EAI_2025/schedule/) for release and due dates.

Assignments

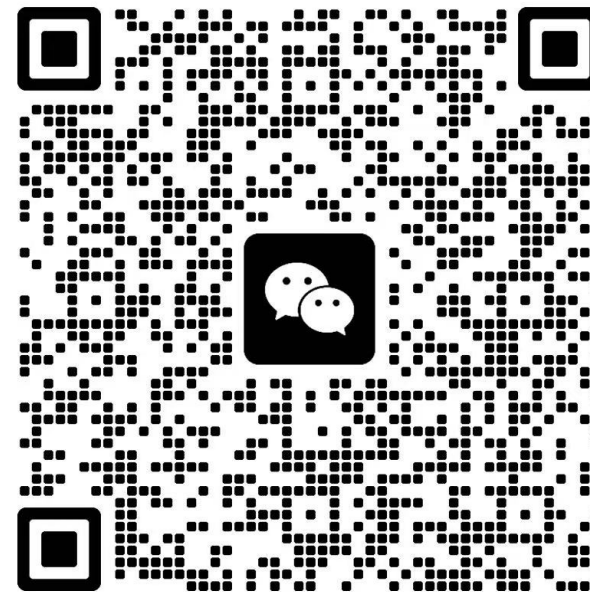
- Each assignment will give you two weeks to finish.
- Late policy for assignments
 - If 1 day (0 - 24 hours) past the deadline, 10% off
 - If 2 day (24 - 48 hours) past the deadline, 20% off
 - Zero credit if more than 2 days.

WeChat Group

- WeChat group:
 - Use for notifications and announcements



群聊: 25 春 EAI 导论



该二维码7天内(2月24日前)有效, 重新进入将更新

Course Organization

A Typical Syllabus of Introduction to Robotics

- Robotics is a branch of engineering that involves the conception, design, manufacture and operation of robots.

What is missing?

Introduction to Robotics

CS223A

Stanford School of Engineering

Topics include

- Spatial descriptions
- Kinematics
- Jacobians
- Statics
- Dynamics
- Motion planning and control
- Trajectory generation

Intelligence! Intelligence! Intelligence!

Course Organization: Intro to Robotics

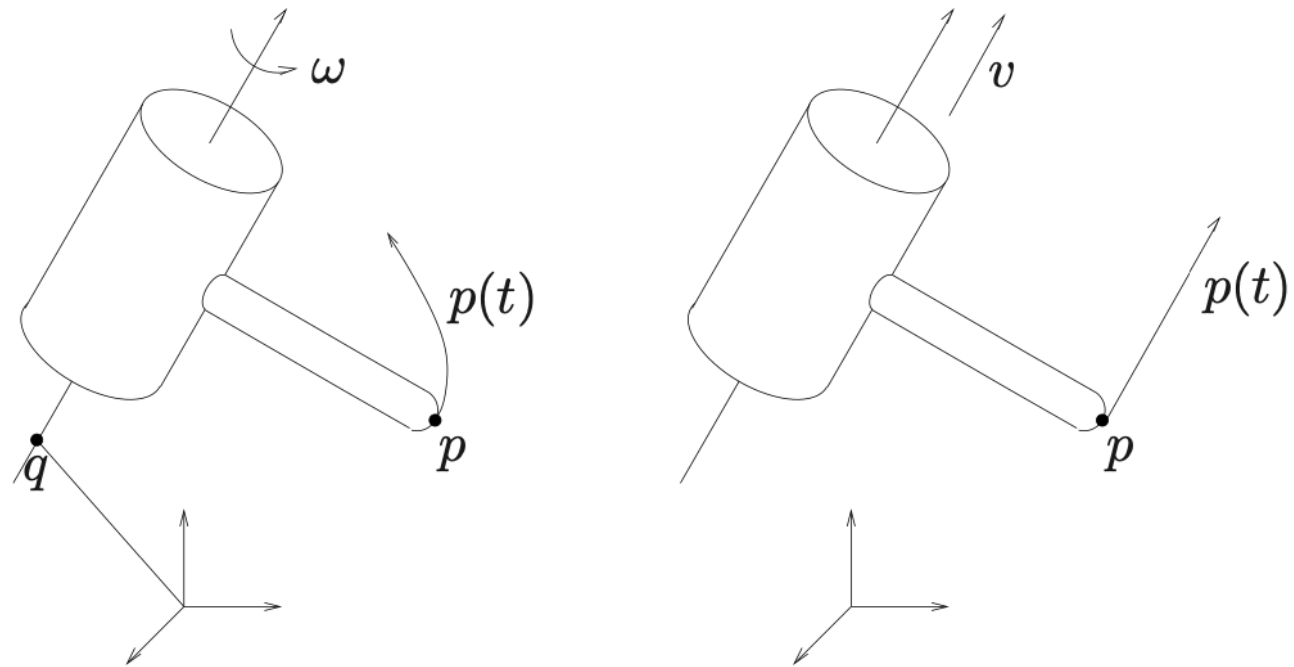
Lecture	02/24/2025 15:10 Monday	Basics of Robotics I	Kinematics of Rigid Bodies and Articulated Objects
Lecture	03/03/2025 15:10 Monday	Basics of Robotics II	Motion Planning
Lecture	03/10/2025 15:10 Monday	Basics of Robotics III	Control Algorithms (PD, PID)

(Tentative schedule)

For the complete and up-to-date schedule, see https://pku-epic.github.io/Intro2EAI_2025/schedule/.

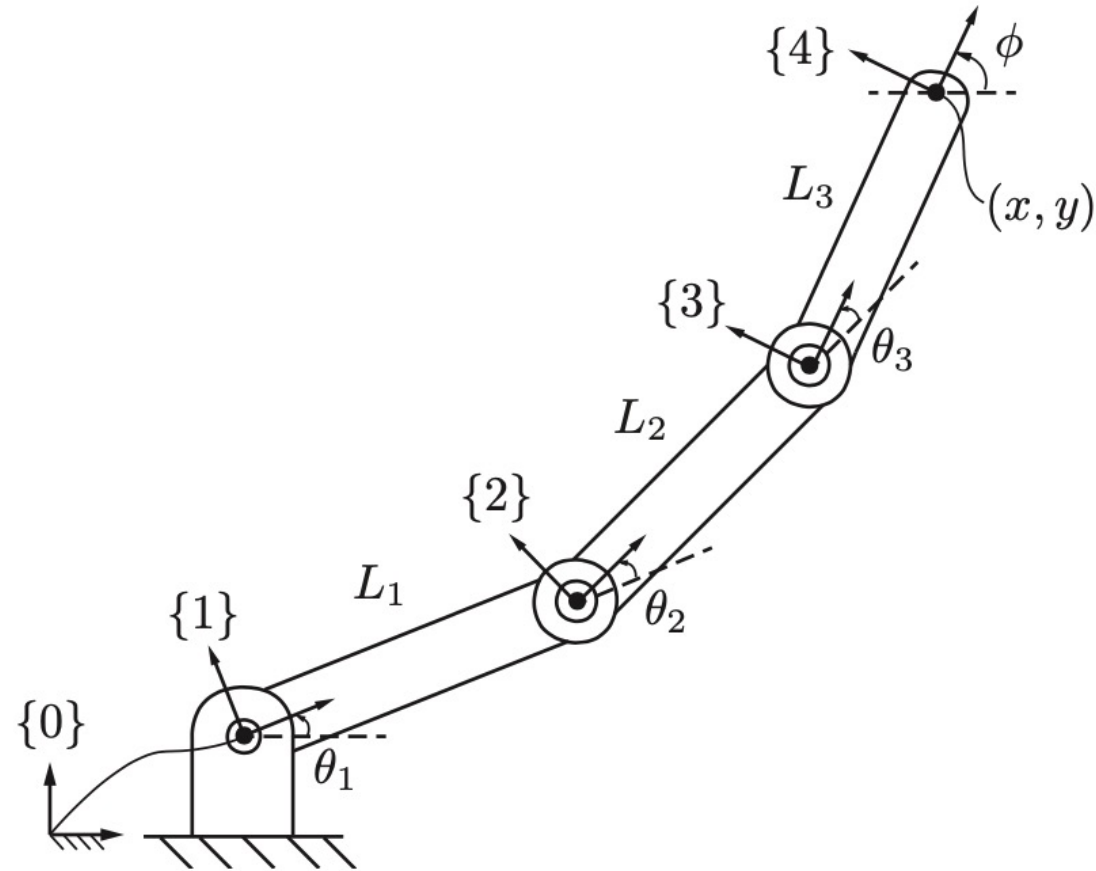
Topics Covered

- Modeling Robots by Rigid-Body Geometry



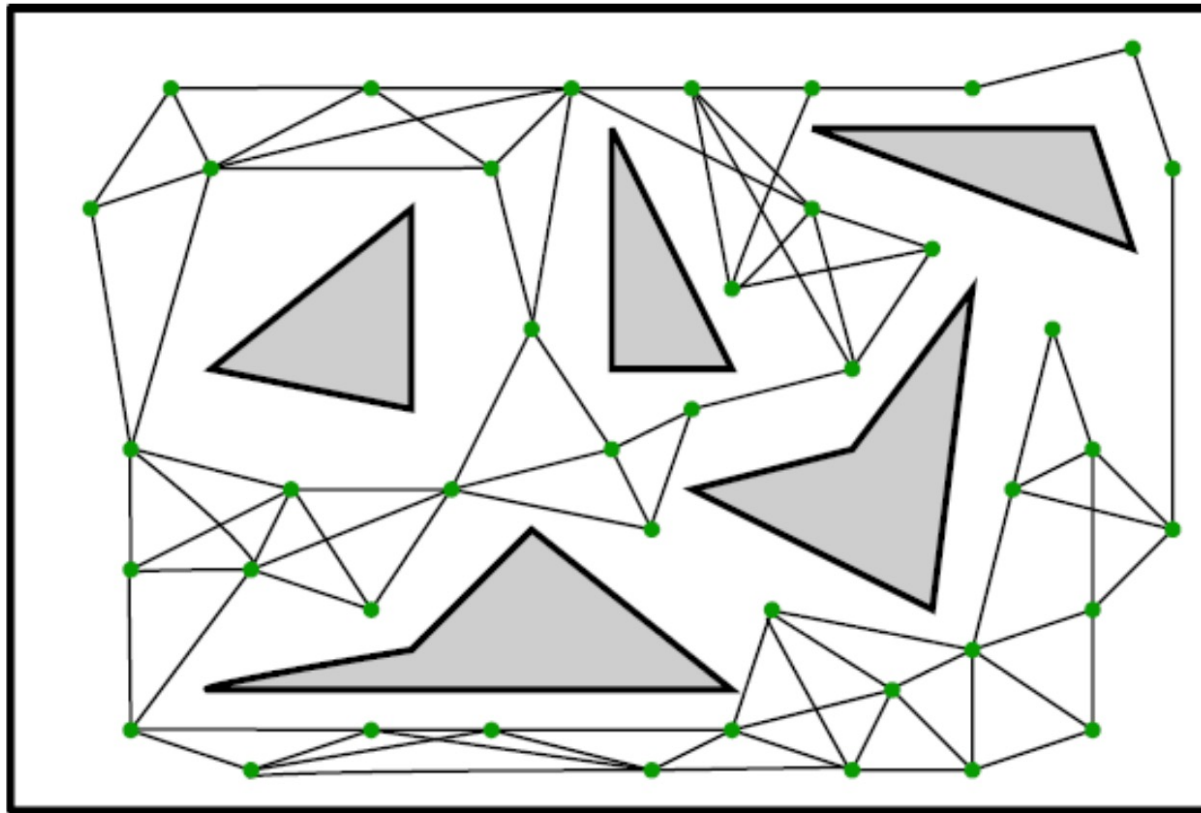
Topics Covered

- Forward and Inverse Kinematics of Robots



Topics Covered

- Classical Planning and Control



Course Organization: Robot 3D Vision

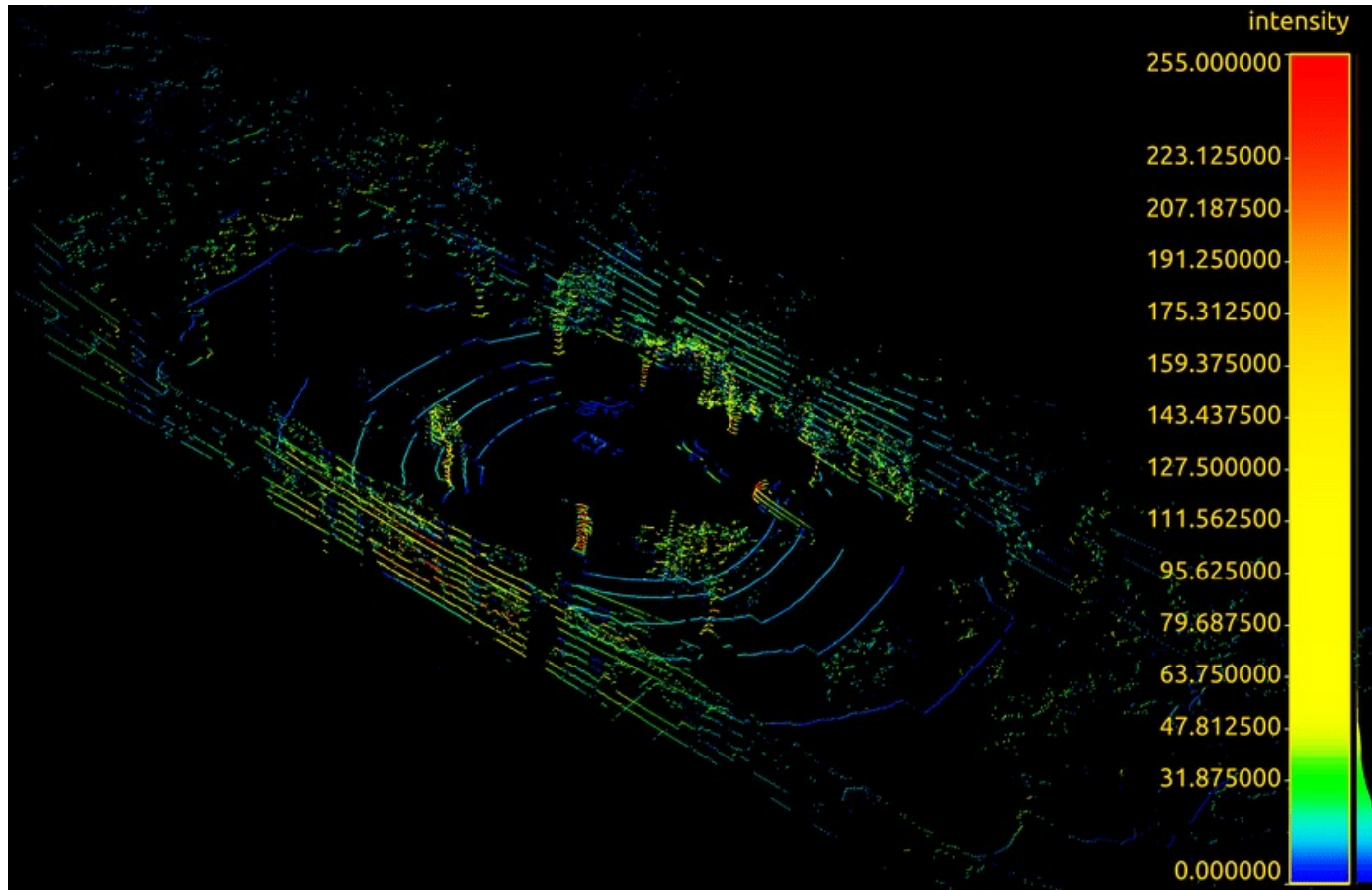
Lecture	03/17/2025 15:10 Monday	Vision and Grasping I	Scene Perception Based on 3D Vision
Lecture	03/24/2025 15:10 Monday	Vision and Grasping II	6D Object Pose Estimation for Known Object Grasping, and Grasp Pose Estimation for Generic Objects

(Tentative schedule)

For the complete and up-to-date schedule, see https://pku-epic.github.io/Intro2EAI_2025/schedule/.

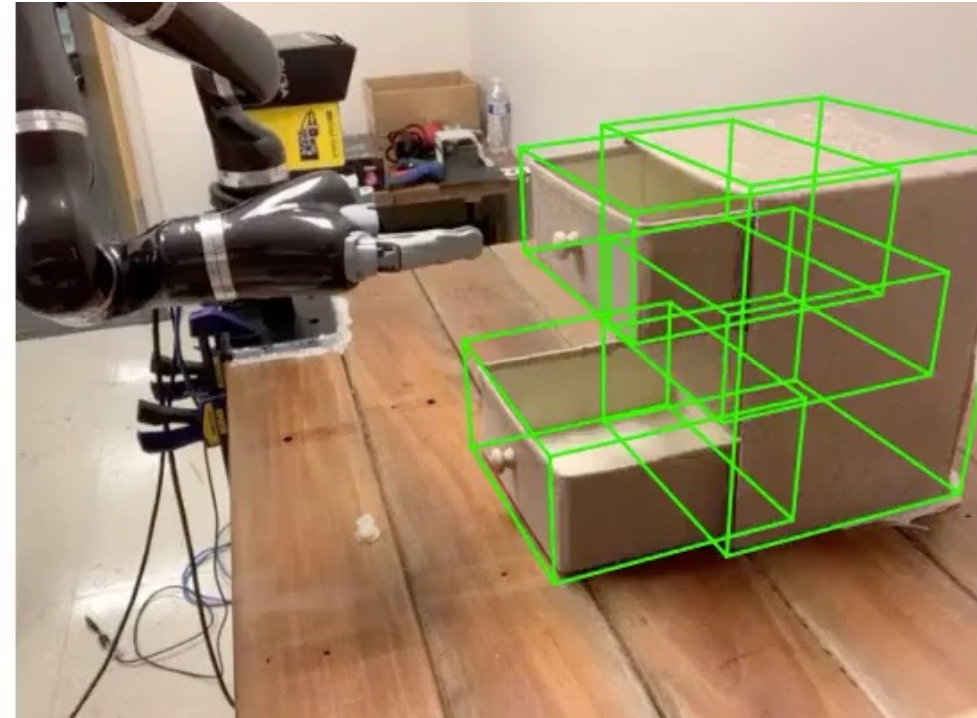
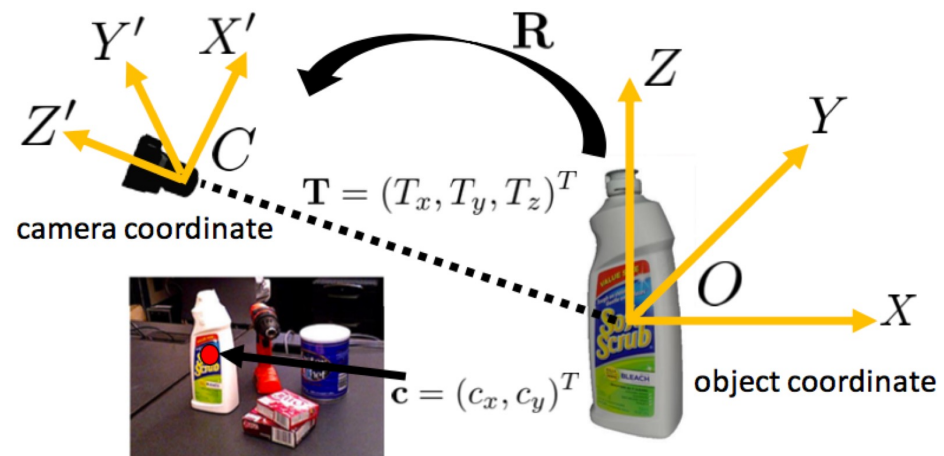
Topics Covered

- Depth and Point Cloud Perception



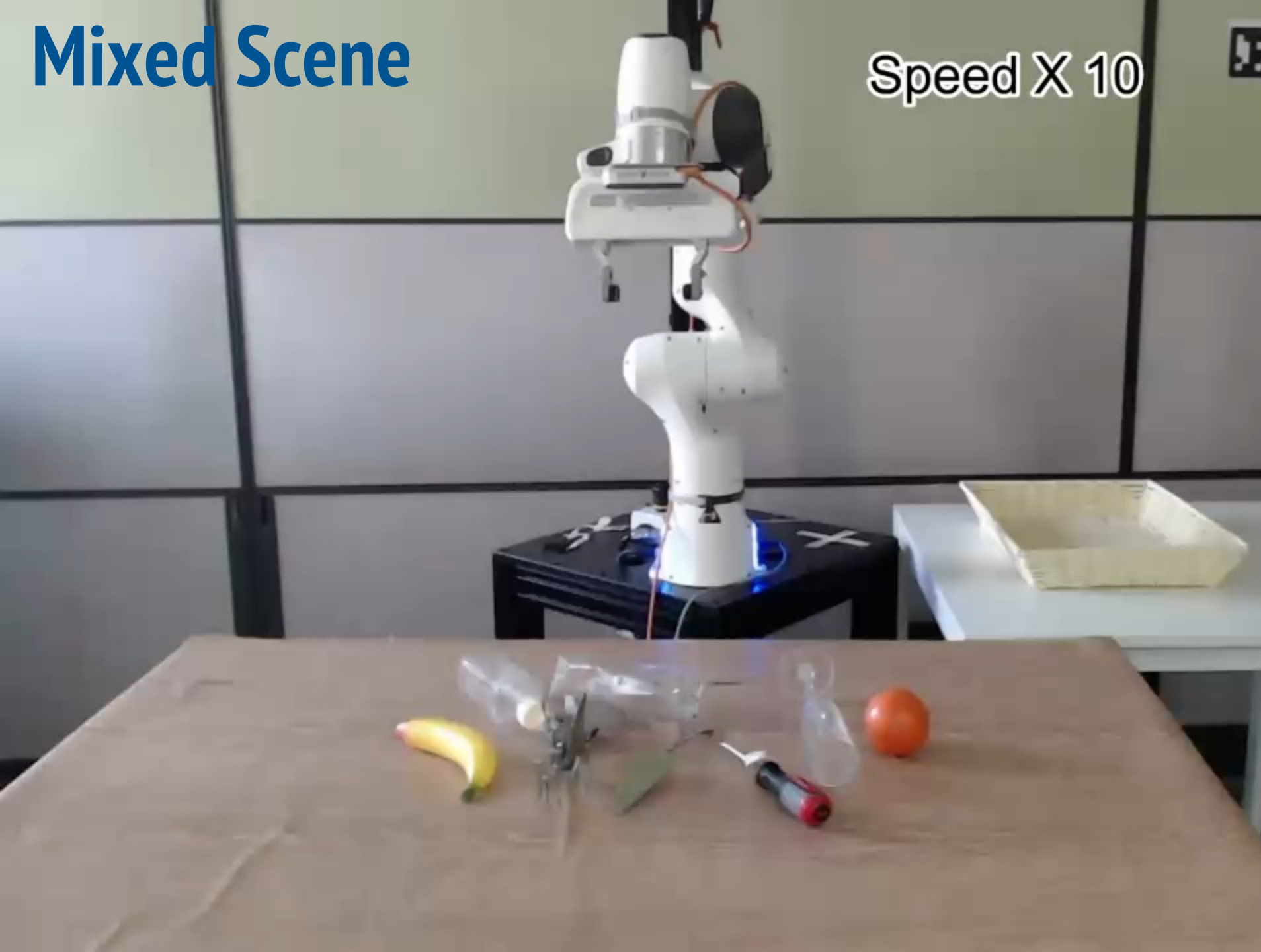
Topics Covered

- Object Pose Estimation
- Generalizable Object Pose Tracking



Mixed Scene

Speed X 10

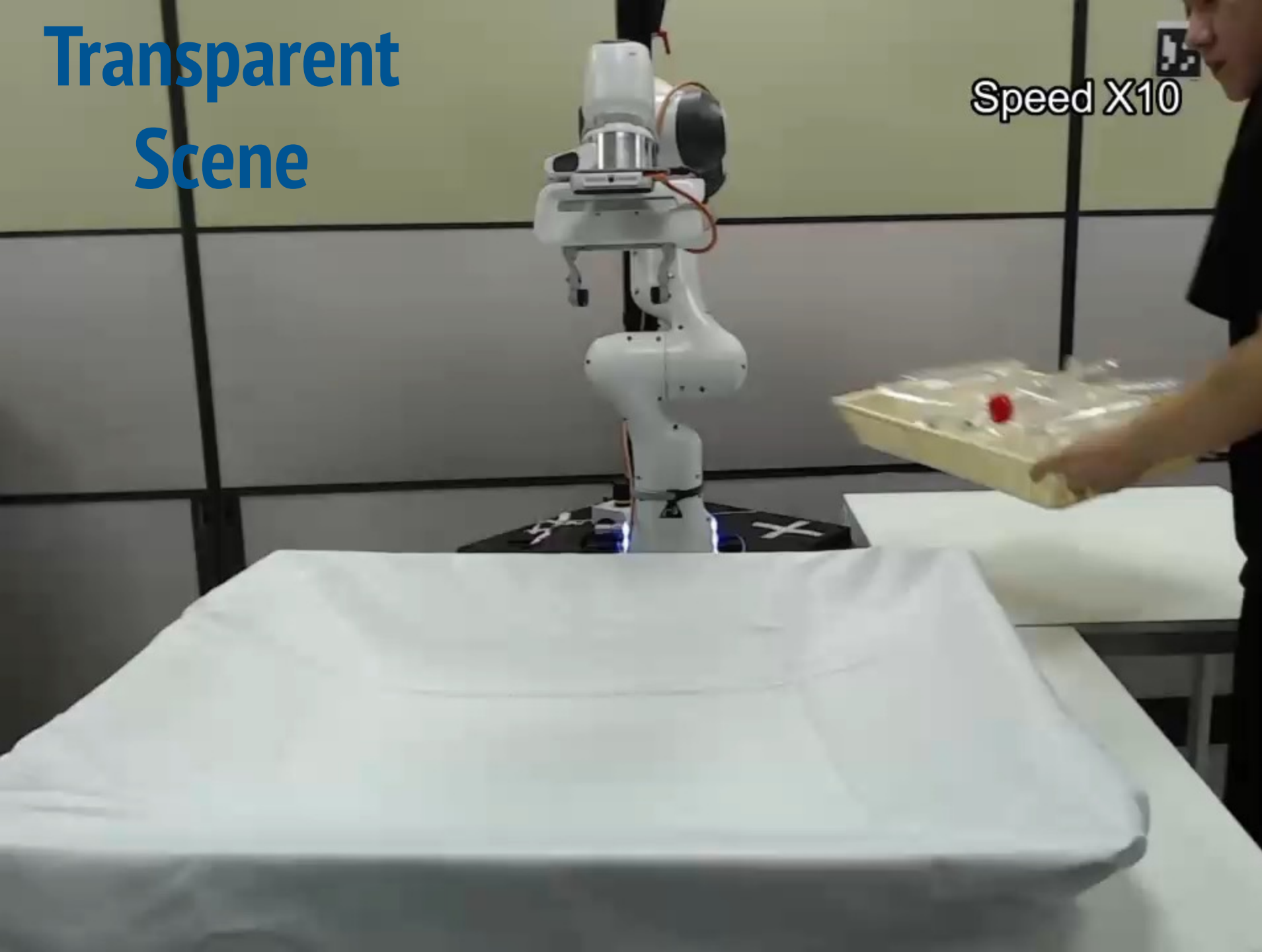


Transparent Scene

Speed X10

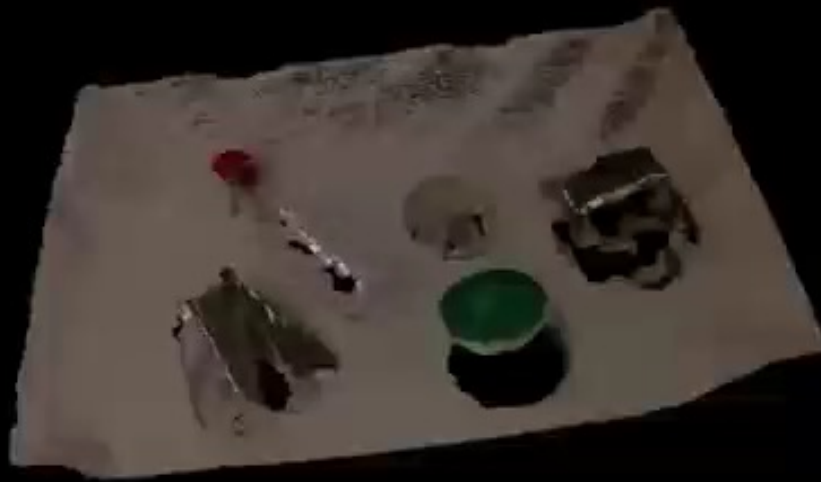
Grasp Poses

RViz Visualization



Galbot Mobile Grasping

Raw Depth



D3RoMA



* Captured by RealSense D415

Dexterous Grasp Diffusion Model



our model can grasp transparent and specular objects despite being trained solely on perfect depth point clouds,



Course Organization: Policy Learning

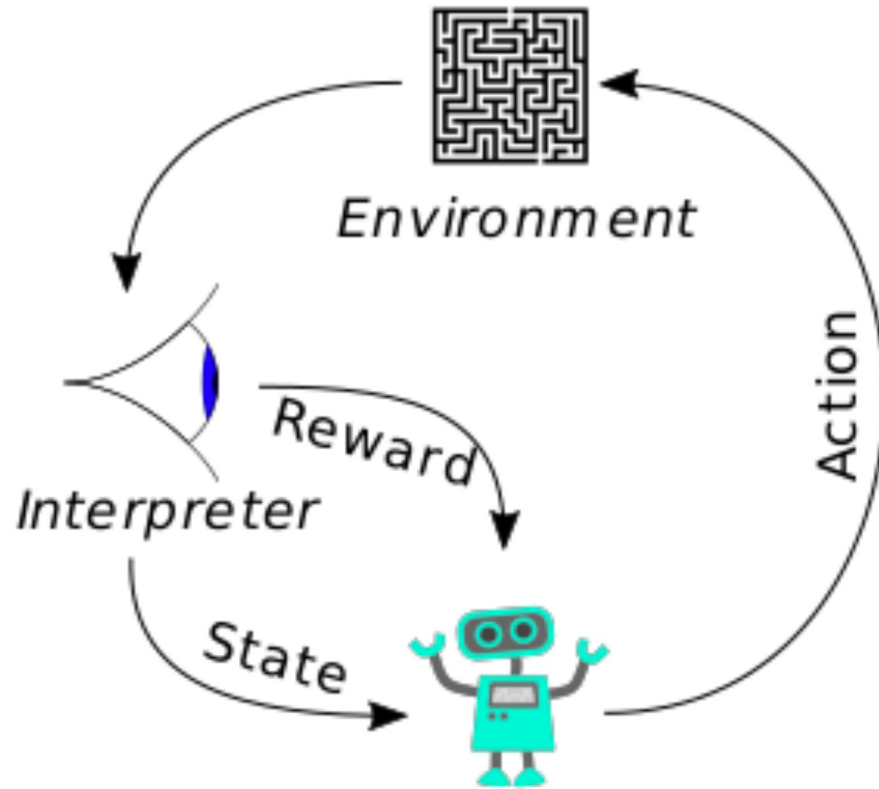
Lecture	03/31/2025 15:10 Monday	Policy Learning I	Reinforcement Learning
Lecture	04/07/2025 15:10 Monday	Policy Learning II	Imitation Learning
Lecture	04/14/2025 15:10 Monday	Policy Learning III	Diffusion Policy
Lecture	04/21/2025 15:10 Monday	Policy Learning IV	Midterm Examination

(Tentative schedule)

For the complete and up-to-date schedule, see https://pku-epic.github.io/Intro2EAI_2025/schedule/.

Topics Covered

- Model-free (on-policy, off-policy) RL



Playing Atari with Deep Reinforcement Learning

Volodymyr Mnih, Koray Kavukcuoglu, David Silver

{vlad, koray, david}@openai.com

We present the first deep reinforcement learning model that can learn to play a wide range of Atari games from raw pixels. Our model is a convolutional neural network that takes raw frames as input and outputs a distribution over actions. We apply this model to a variety of Atari games and find that it outperforms a human expert on 29 out of 49 games.

Trust Region Policy Optimization

John Schulman, Sergey Levine, Philipp Moritz, Michael Jordan, Pieter Abbeel
University of California, Berkeley

JOSCHU@EECS.BERKELEY.EDU

Abs

We describe an iterative policy optimization algorithm, with guaranteed convergence. By making several theoretically-justified practical algorithmic choices, we achieve a practical algorithm, called Trust Region Policy Optimization (TRPO).

Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor

Tuomas Haarnoja¹, Aurick Zhou¹, Pieter Abbeel¹, Sergey Levine¹

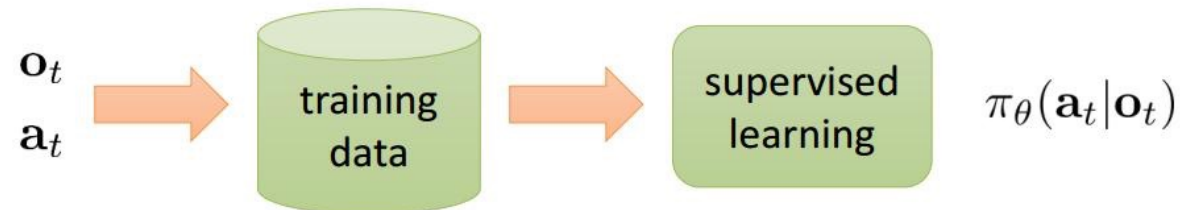
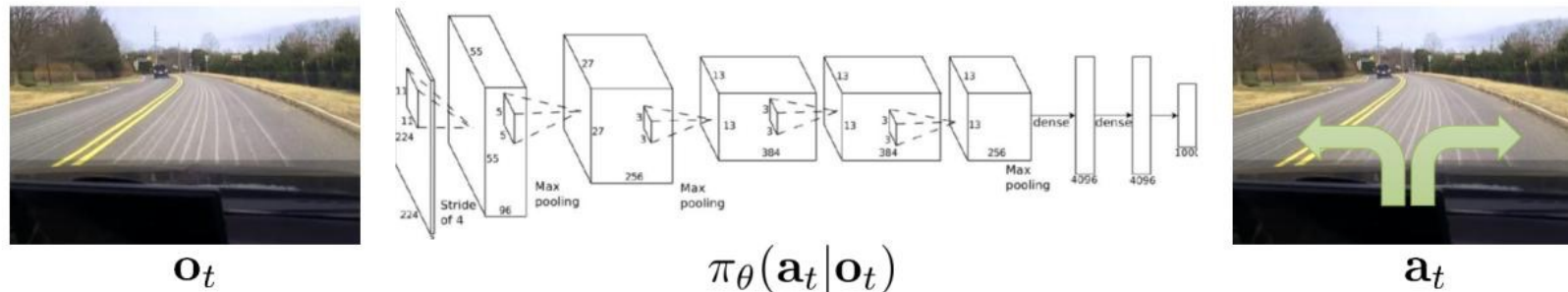
Abstract

Model-free deep reinforcement learning (RL) algorithms have been demonstrated on a range of challenging decision making and control tasks. However, these methods typically suffer from two major challenges: very high sample complexity and brittle convergence properties, which necessitate meticulous hyperparameter tuning. Both of these challenges severely limit the applicability

of these methods in real-world domains has been hampered by two major challenges. First, model-free deep RL methods are notoriously expensive in terms of their sample complexity. Even relatively simple tasks can require millions of steps of data collection, and complex behaviors with high-dimensional observations might need substantially more. Second, these methods are often brittle with respect to their hyperparameters: learning rates, exploration constants, and other settings must be set carefully for different problem settings to achieve good results. Both of these challenges

Topics Covered

- Imitation Learning



behavioral cloning

Course Organization: Robotic Tasks

Lecture	04/28/2025 15:10 Monday	Policy Learning V	Legged Robot Locomotion
Lecture	05/12/2025 15:10 Monday	Policy Learning VI	Robotic Gripper Manipulation and Dexterous Hand Manipulation

(Tentative schedule)

For the complete and up-to-date schedule, see https://pku-epic.github.io/Intro2EAI_2025/schedule/.

Topics Covered

- Object Grasping

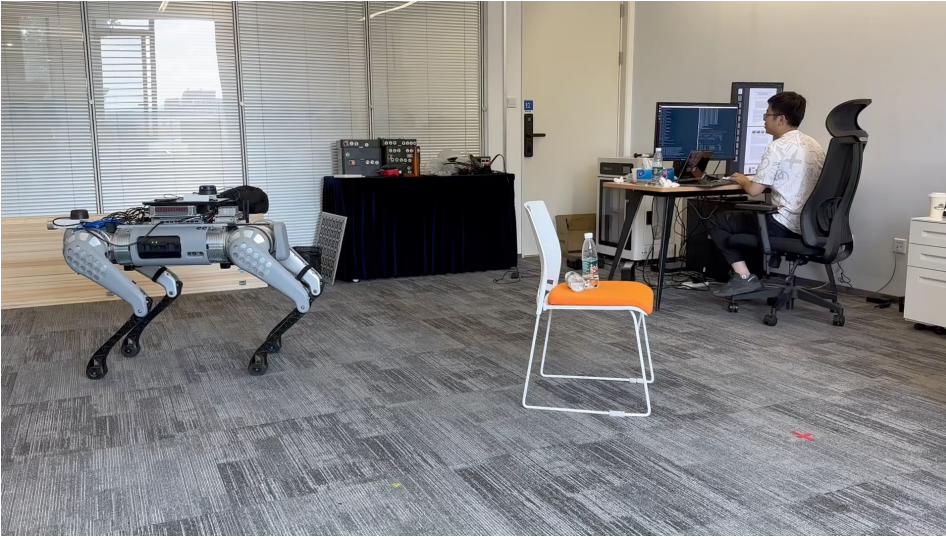


- Articulated Object Manipulation

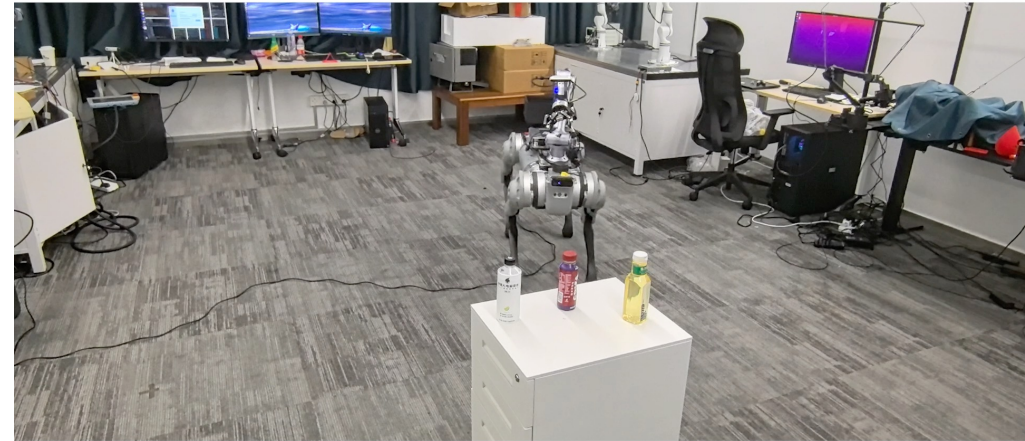


Topics Covered

- Navigation and Mobile Manipulation



Real-world Experiments Speed up x3



Topics Covered

QuadWBG

Topic Covered

Lecture	05/19/2025 15:10 Monday	Multimodal LLM in EAI I	Pretrained Large Models (CLIP) and GPT-4v/4o and other Multimodal LLMs
Lecture	05/26/2025 15:10 Monday	Multimodal LLM in EAI II	Modular General Manipulation System and End-to-end Vision-Language-Action (VLA) Models



Introduction to Embodied AI

Next week: Lecture 2,
The Basics of Robotics

