Schedule

Start	End	Section	Speaker
13:00	13:45	Overview of Embodied AI	Zhiwei Jia (video)
13:45	14:30	The Basic Frameworks and techniques for Embodied AI	Fanbo Xiang (in person)
14:30	15:15	Design Choices in Embodied AI Environments	Jiayuan Gu (video)
15:15	15:30	Break	
15:30	16:15	Experience and Practices to Debug Simulators	Fanbo Xiang (in person)
16:15	16:35	Real World Robotics and Sim2Real	Rui Chen (video)
16:35	17:00	Embodied AI Tasks in ManiSkill and Visual Learning Challenges	Fanbo Xiang (in person)

Angel Xuan Chang
Simon Fraser University

Rui Chen
Tsinghua University

Jiayuan Gu
UC San Diego

Zhiwei Jia
UC San Diego

Tongzhou Mu
UC San Diego

Yuzhe Qin
UC San Diego

Hao Su
UC San Diego

Xiaolong Wang
UC San Diego

Fanbo Xiang
UC San Diego

UC San Diego

清華大學 Tsinghua University

SFU SIMON FRASER UNIVERSITY

Building and Working in Environments for Embodied AI

CVPR 2022 Tutorial


Angel Xuan Chang
Simon Fraser University


Rui Chen
Tsinghua University


Jiayuan Gu
UC San Diego


Zhiwei Jia
UC San Diego


Tongzhou Mu
UC San Diego


Yuzhe Qin
UC San Diego


Hao Su
UC San Diego


Xiaolong Wang
UC San Diego


Fanbo Xiang
UC San Diego

Schedule

Start	End	Section	Speaker
13:00	13:45	Overview of Embodied AI	Zhiwei Jia (video)
13:45	14:30	The Basic Frameworks and techniques for Embodied AI	Fanbo Xiang (in person)
14:30	15:15	Design Choices in Embodied AI Environments	Jiayuan Gu (video)
15:15	15:30	Break	
15:30	16:15	Experience and Practices to Debug Simulators	Fanbo Xiang (in person)
16:15	16:35	Real World Robotics and Sim2Real	Rui Chen (video)
16:35	17:00	Embodied AI Tasks in ManiSkill and Visual Learning Challenges	Fanbo Xiang (in person)

Angel Xuan Chang
Simon Fraser University

Rui Chen
Tsinghua University

Jiayuan Gu
UC San Diego

Zhiwei Jia
UC San Diego

Tongzhou Mu
UC San Diego

Yuzhe Qin
UC San Diego

Hao Su
UC San Diego

Xiaolong Wang
UC San Diego

Fanbo Xiang
UC San Diego

Overview of Embodied AI: Simulators, Datasets and Tasks

Building and Working in Environments for Embodied AI (part I)

CVPR 2022 Tutorial

UC San Diego

清華大學
Tsinghua University

SFU SIMON FRASER UNIVERSITY

Overview

This part of the tutorial is intended to give an introduction of Embodied AI

- What is Embodied AI and why Embodied AI?
- Why studying Embodied AI in virtual environments?
- What are the key factors of building environments?
- Roadmap for other parts of the tutorial

Outline

- Background
  - Why embodiment in AI and What is Embodied AI?
- What are the Key Factors in the Environments?
  - Simulators
  - Assets
  - Tasks
- Roadmap of the Tutorial

Example: Penalty Kick

You have to learn by doing!

Image source: https://www.thesoccerstore.co.uk/blog/football-goals/best-kids-garden-football-goals/

Embodiment Hypothesis:

"intelligence ***emerges*** in the interaction of an agent with an environment and as a result of sensorimotor activity"

The Development of Embodied Cognition: Six Lessons from Babies

Linda Smith
Psychology Department
Indiana University
Bloomington, IN 47405
smith4@Indiana.edu

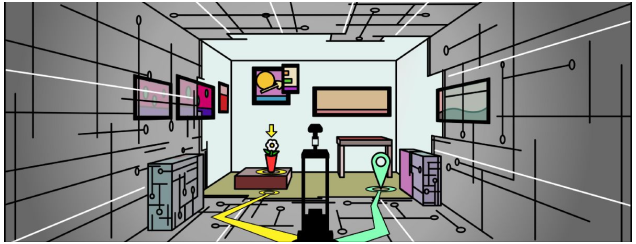
Michael Gasser
Computer Science Department
Indiana University
Bloomington, IN 47405
gasser@Indiana.edu

Abstract The embodiment hypothesis is the idea that intelligence emerges in the interaction of an agent with an environment and as a result of sensorimotor activity. We offer six lessons for *developing* embodied intelligent agents suggested by research in developmental psychology. We argue that starting as a baby grounded in a physical, social, and linguistic world is crucial to the development of the flexible and inventive intelligence that characterizes humankind.

Keywords
Development, cognition, language, embodiment, motor control

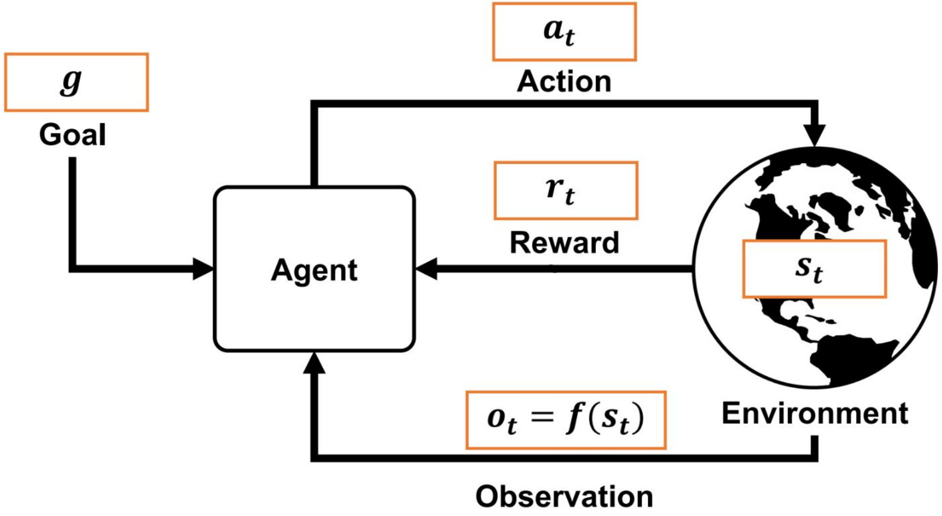
What is Embodied AI?

- Concretely, the study of intelligent agents to solve tasks by
  - **seeing** (usually in an egocentric view)
  - **talking** (via texts or audios)
  - **reasoning** (understand the surroundings and plan)
  - **acting** (through motor controls or high-level actions).
- An interdisciplinary field
  - Embodied AI Workshop @ CVPR 20/21/22



How to Model Agents, Environments & Tasks?

Usually via a Markov Decision Process (MDP)



Slide Credit: Stefan Lee

How to model the intelligent agents, the environment and the tasks?

- Usually via partially-observed Markov decision process

  S, O, A: State, Observation and Action space

  $p(s)$ the observation distribution

  T the dynamics, $\mathcal{R}$ the reward

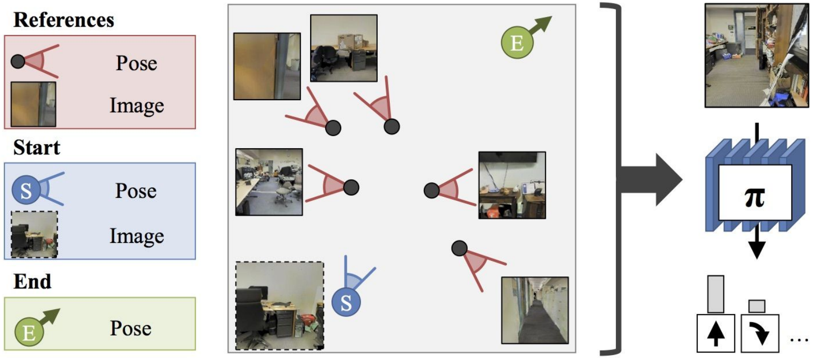
  ρ_0 the initial state distribution

  γ the discount factor

  H the finite horizon

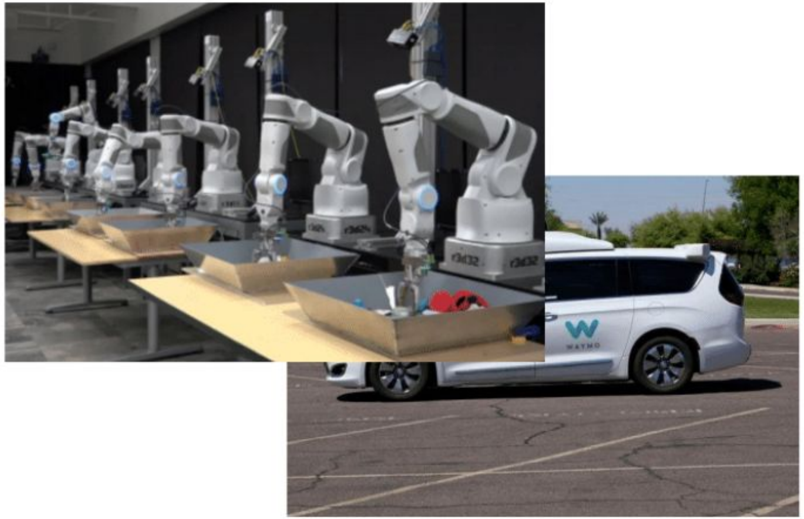
$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\tau\sim\pi} \sum_{t=0}^{H} \gamma^t r_t$$

$$(\mathcal{S}, \mathcal{O}, \mathcal{A}, p(s), T, \mathcal{R}, \rho_0, \gamma, H)$$



Many Embodied AI Work Starts from Virtual Environments. Why?

- Learning in real world: dangerous and expensive
- Learning in virtual environment: safe and scalable

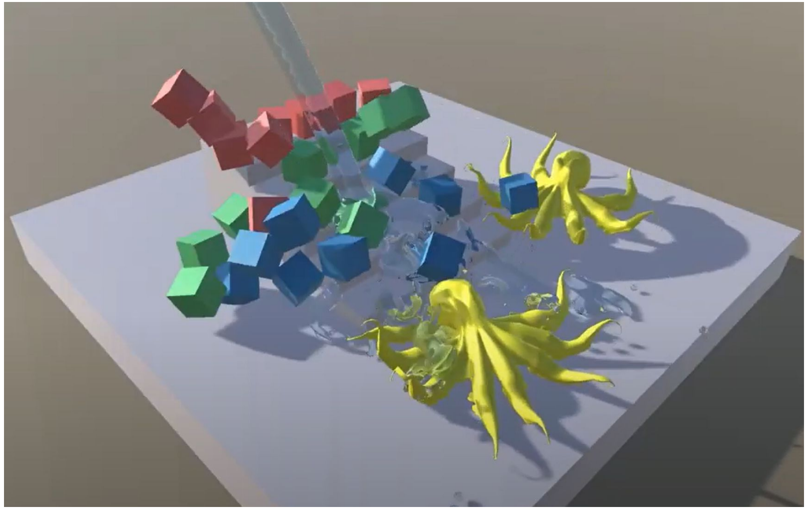


Outline

- Background
- What are the Key Factors in the Environments?
  - Simulators
  - Assets
  - Tasks
- Roadmap of the Tutorial

What is a Simulator?

- Simulator = **Physics Simulation** + **Sensory Signal Rendering**
- Simulation provides the mathematical model of the dynamics
- Rendering provides observations of the robot and its surroundings



Source: Nvidia FleX

What to Consider When Choosing a Simulator?

- Rendering
- Physics
- Speed
- Objects types and properties
- Action modeling
- Human interface

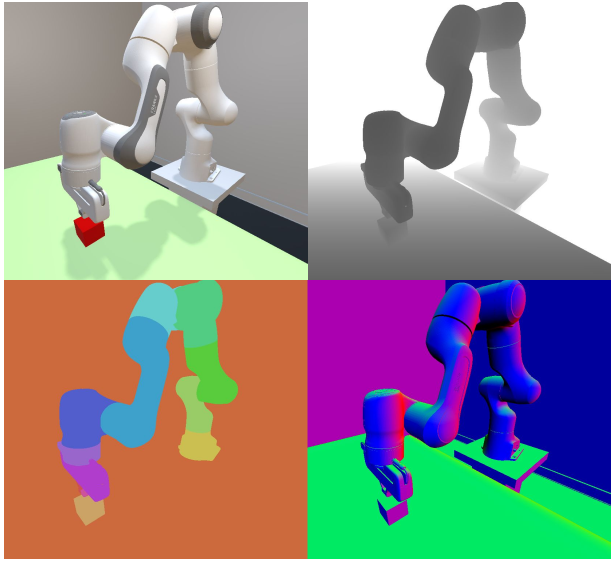
Rendering in Simulators

- What type of sensory signals are supported?

Rendering in Simulators

Common sensory signals

- RGB, depth, surface normal
- Instance/semantic segmentation
- Optical flow, scene flow

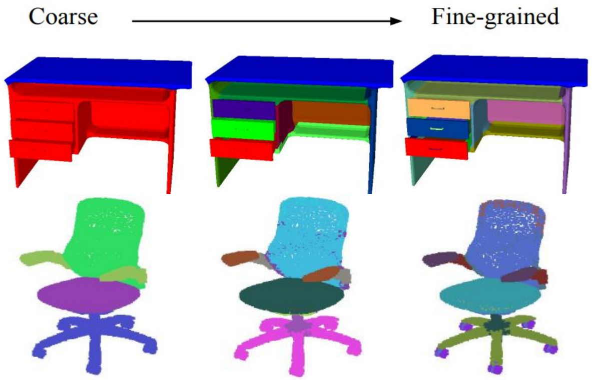


ManiSkill 2022: rgb, depth, semantic segmentation, normal

Rendering in Simulators

Uncommon sensory signals

- Part-level segmentation
  - Help to understand dynamics of articulated objects.
- Acoustic signals
  - Help to understand mass, texture, collision, etc.

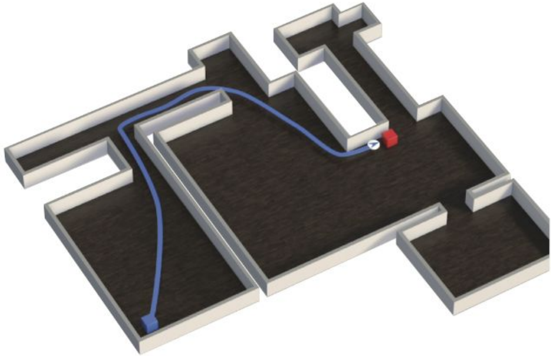


PartNet

ThreeDWorld

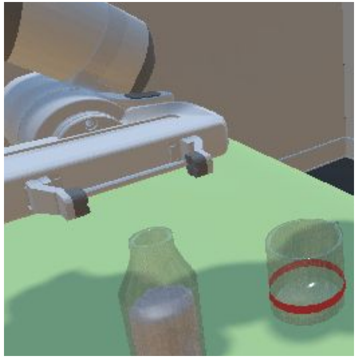
Physics

- Physical vs. non-physical simulation
  - No-physics, rigid-body, articulated body, fluid and soft-body
- Different tasks require different granularity level of physics



Partial Physics
(Figure from DD-PPO, for Visual Navigation)

Full physics
(ManiSkill 2021)



Full physics
(ManiSkill 2022)

Speed of Simulators

- Number of interaction steps per second affects what kind of training approaches are viable for solving tasks in the simulators.
    - **Imitation Learning?**
        - Slow simulators can be used. Only need to evaluate policies.
    - **Model-free RL?**
        - Need very fast simulators. Agents practices many times to learn.
    - **Model-based RL?**
        - Relatively fast simulator can be used. Agents use an internal world model to reduce dependency on interactions.

Speed of Simulators

Deciding factors of speed

- Speed of underlying engines
- Environment complexity
    - Geometry complexity
    - Interaction complexity
    - Rendering complexity

Objects Types and Properties

By kinematic structure:

- Rigid-body objects
- Articulated objects
- Soft-body objects

ManiSkill 2021

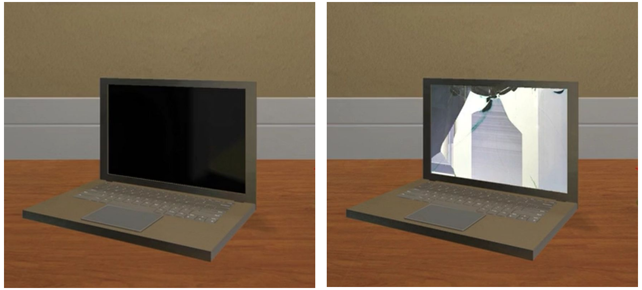
ManiSkill 2022

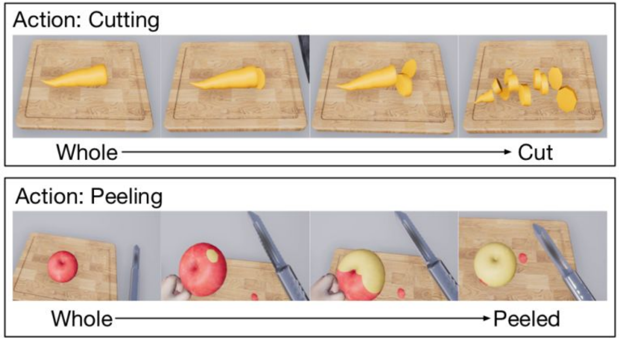
Objects Types and Properties

Other actionable properties

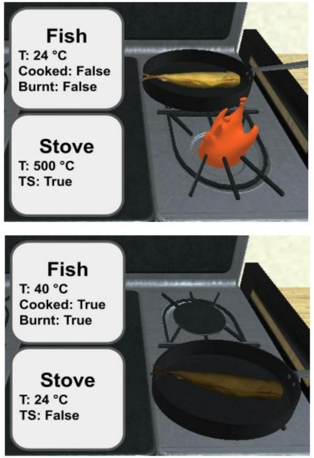
- AI2-THOR: cooled, broken, sliced, etc.
- iGibson 2.0: heated, cooked, etc.
- VRKitchen: cut, peeled, juiced, etc.

…


“Broken” from AI2-THOR



“Peeling” and “Cutting” from VRKitchen



“Cooked” and “heated” from iGibson 2.0

Action Modeling

What actions are supported? It depends on

- Types of **physics simulation** (introduced before)
- Types and properties of **objects** (introduced before)
- Types of **robots** (introduced next)

Actions can be classified as

- Low-level, high-level, somewhere in between, etc.

Action Modeling

Robot models to use

For example:

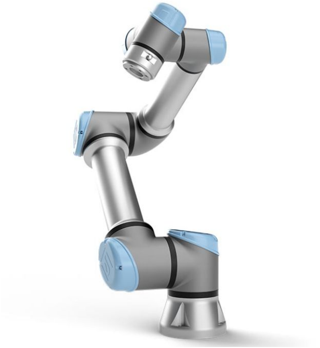
- Fetch
- Franka
- Kuka
- UR
- More at https://robots.ros.org/

Common concern: versatility vs. realisticity

Fetch

Franka

Kuka



UR

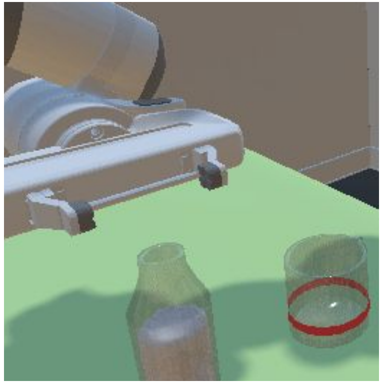
Action Modeling

Low-level vs. High-level Actions

- **Low-level** actions (e.g., motor controls)
  - Necessary for actual robot deployment to the real world
- **High-level** actions (e.g., [Action] [TargetObjA] [TargetObjB])
  - Good for long-horizon tasks, skill chaining, task planning, etc.
- Somewhere **in between**

High-level (iTHOR)

In between (ManipulaTHOR)

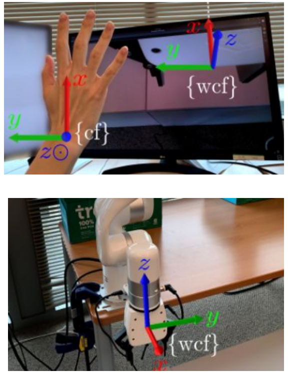


Low-level (ManiSkill 2022 in July)

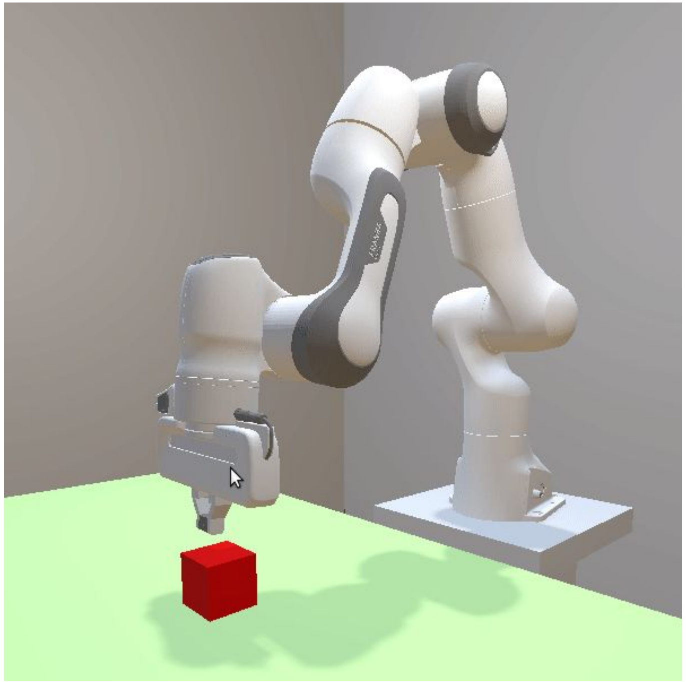
Human Interface

- Mouse & keyboard
- Virtual reality (VR)
- Vision-based Teleoperation

VRKitchen



“Single RGB-D Camera Teleoperation for General Robotic Manipulation”



SAPIEN

Outline

- Background
    - What is Embodied AI and why Embodied AI?
- What are the Key Factors in the Environments?
    - Simulators
    - Assets
    - Tasks
- Roadmap of the Tutorial

What are Assets?

In simulation, we load data structures stored as files to

- **specify each object** by its properties;
- **define a scene** by the arrangement of objects;
- **represent demonstrations** with trajectories or human instructions;

  …

We call these data structures as assets.

Object Asset

Geometry and kinematic structure

- Mesh, revolute vs. prismatic joint, etc.

Optical material properties

- Reflection model, texture, etc.

Dynamical material properties

- Friction, mass properties (density, inertia), elasticity & plasticity, etc.

Other properties

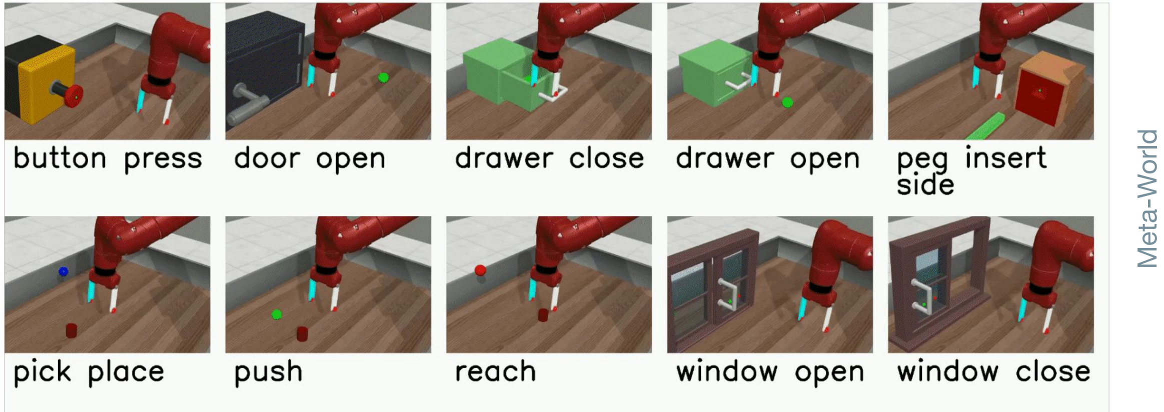
- Acoustics, thermodynamics, etc.

Object Asset Example 1: Grasping Assets

- YCB
- EGAD!

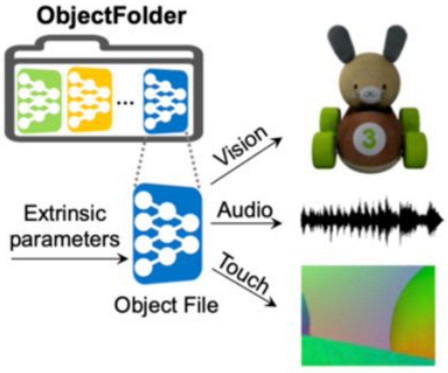
Object Asset Example 2: General Manipulation Skill Assets

- PartNet-Mobility
- DoorGym
- Objects from iTHOR
- Meta-World



Object Asset Example 3: Multisensory Object Assets

- ObjectFolder
- ThreeDWorld



ObjectFolder

ThreeDWorld

Scene Assets

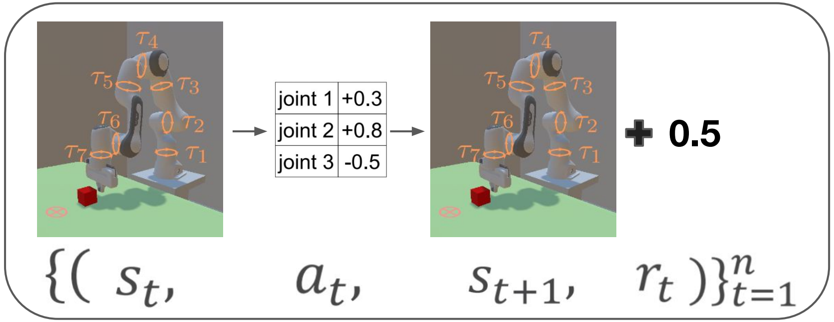
- Static
  - E.g, Habitat-Matterport 3D Dataset
- Interactable
  - E.g., iTHOR scenes

iTHOR

Habitat-Matterport 3D

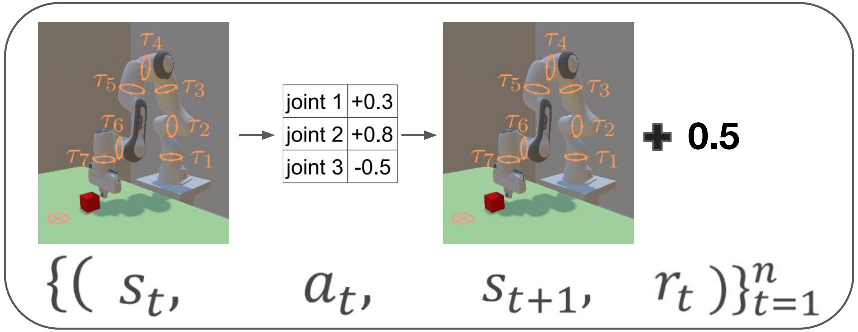
Demonstration Assets

State (ManiSkill)



Video (HOI4D)

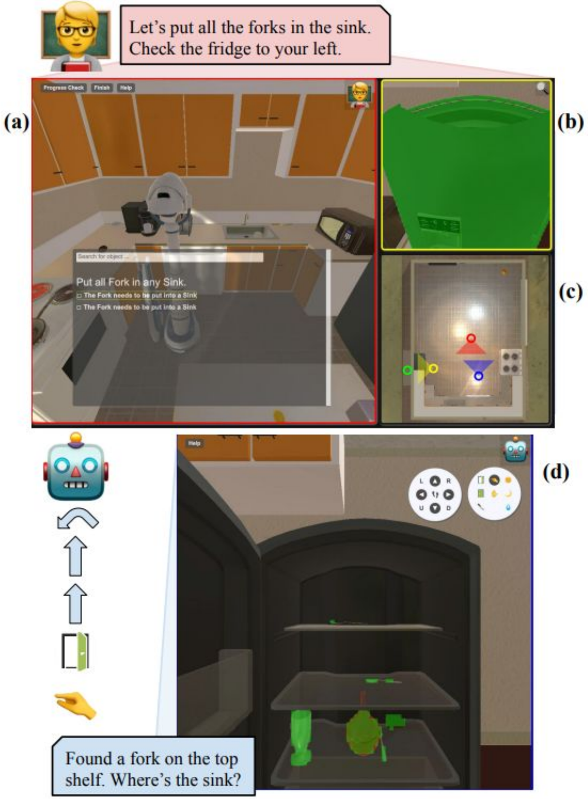
Language (ALFRED)



τ_4
τ_5
τ_3
τ_6
τ_2
τ_7
τ_1
joint 1 +0.3
joint 2 +0.8
joint 3 -0.5
+ 0.5
$\{(s_t, \quad a_t, \quad s_{t+1}, \quad r_t)\}_{t=1}^{n}$

Language-oriented Datasets & Assets

- Language also as outputs
  - Examples:
    - Embodied Question Answering
    - Vision-and-Dialogues Navigation
      - VLN + dialogues
    - Task-driven Embodied Agents that Chat
      - ALFRED + dialogues
  - Cons:
    - Even harder to train online as dialogue generation is an open question.



Outline

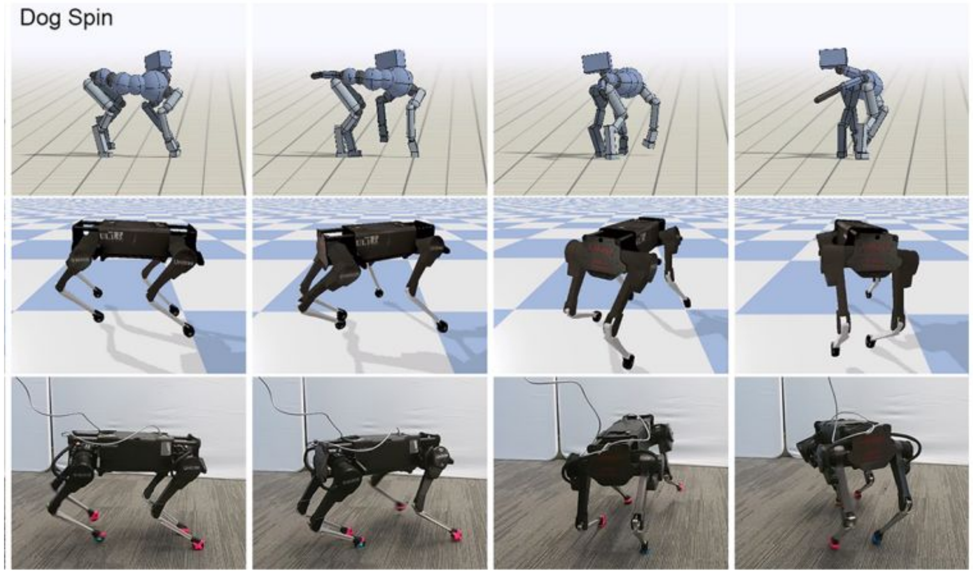
- Background
  - What is Embodied AI and why Embodied AI?
- What are the Key Factors in the Environments?
  - Simulators
  - Assets
  - Tasks
- Roadmap of the Tutorial

Example Tasks in Embodied AI

- Locomotion
- Visual navigation
- Object manipulation
- Rearrangement

Example of Locomotion - Legged Robot Control

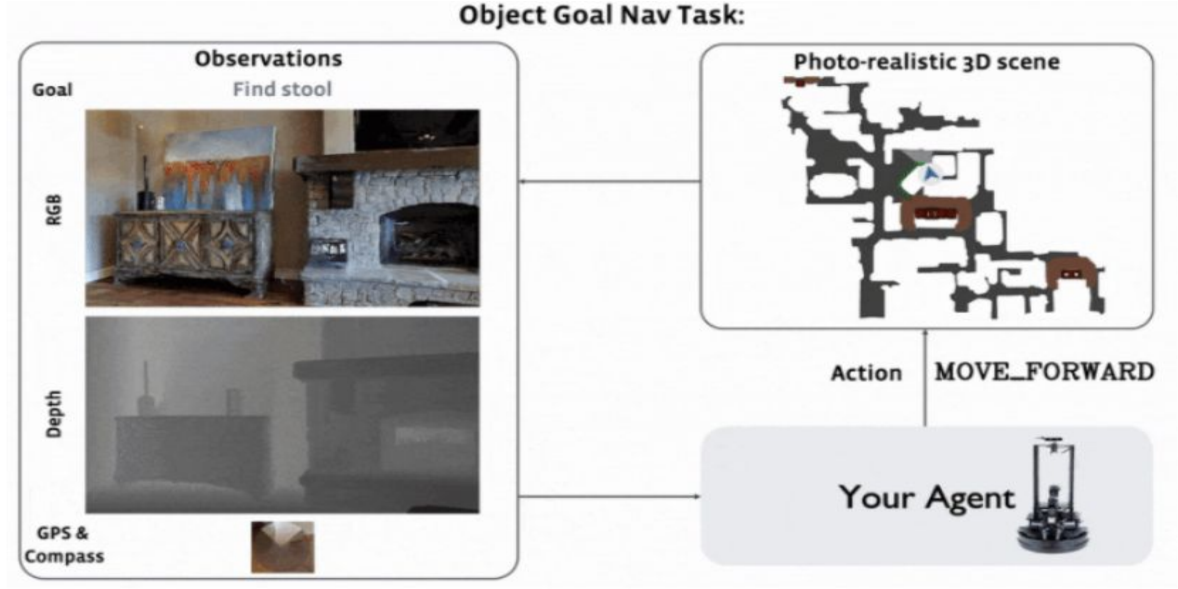
- Control a robot dog to perform a series of actions



"Learning Agile Robotic Locomotion Skills by Imitating Animals"

Example of Visual Navigation (VN) - Object Goal Navigation

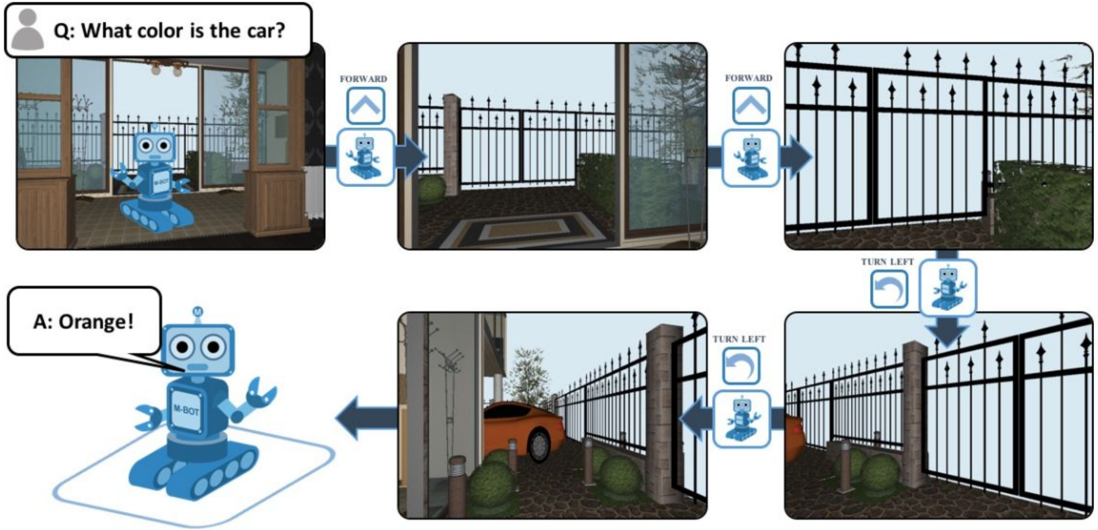
- Specify an object category and ask the agent to find it



From Habitat 1.0

Example of VN with Language - Embodied Question Answering

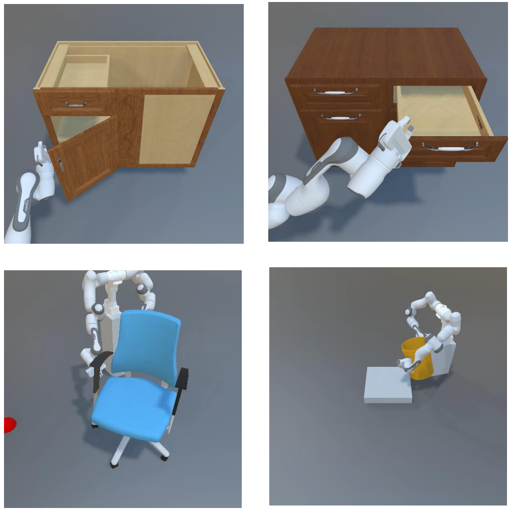
- Ask an agent to answer a question which requires it to navigate in the scene



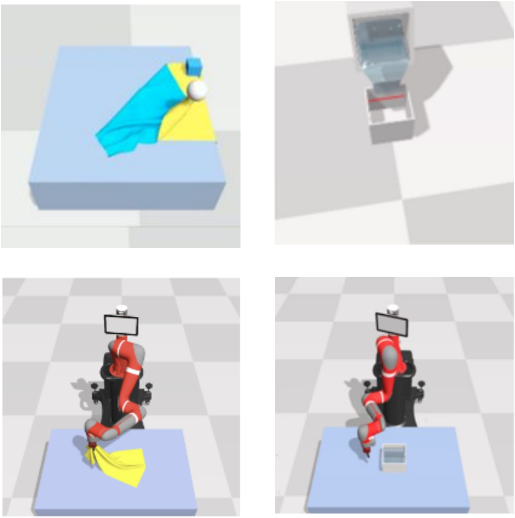
From https://embodiedqa.org/

Examples of Object Manipulation - ManiSkill and SoftGym

- Rigid/articulated object manipulation example - ManiSkill
- Soft-body manipulation example - SoftGym



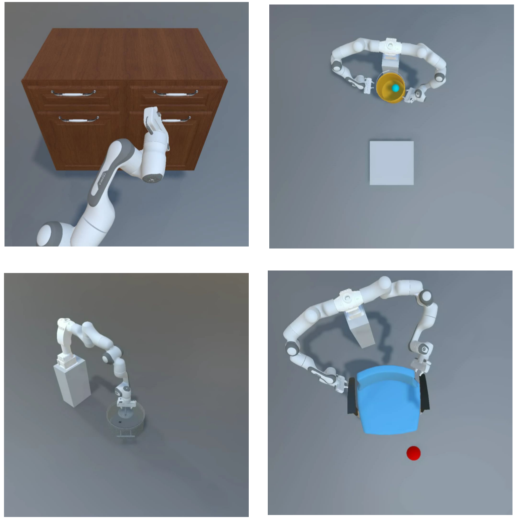
ManiSkill 2021
(Note: ManiSkill 2022 will add softbody simulation tasks)



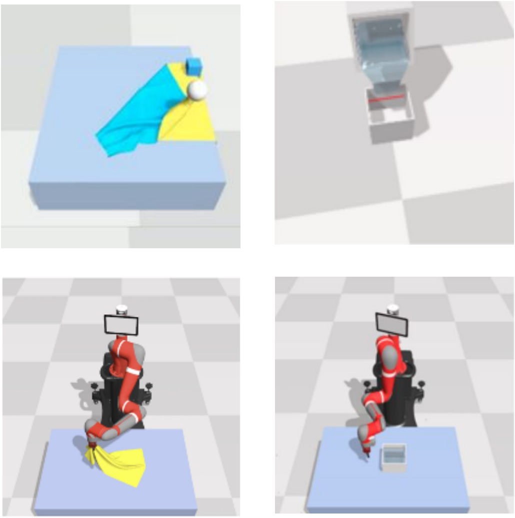
SoftGym

Examples of Object Manipulation - ManiSkill and SoftGym

- Rigid/articulated object manipulation example - ManiSkill
- Soft-body manipulation example - SoftGym



ManiSkill



SoftGym

Example of Object Manipulation with Language - ALFRED

- Ask the agent to perform object manipulations by high-level or step-by-step language instructions

Example of Rearrangement - AI2-THOR Rearrangement Challenge

- Ask the agent to bring poses of the objects to a specified configuration

Summary

- Background
  - Why embodiment in AI and What is Embodied AI?
- What are the Key Factors in the Environments?
  - Simulators
  - Assets
  - Tasks
- Roadmap of the Tutorial

Summary

- Background
  - Why embodiment in AI and What is Embodied AI?
- What are the Key Factors in the Environments?
  - Simulators
  - Assets
  - Tasks
- Roadmap of the Tutorial

- **Part II: The Basic Frameworks and Techniques for Embodied AI**
  - Problem formulation, basics to RL/planning/control/simulation, environment construction example
- **Part III: Design Choices in Modern Embodied AI Environments**
  - Design factors, case studies by popular embodied AI frameworks
- **Part IV: Experiences and Practices to Debug Simulators**
  - Common issues, simulation debugging, environment optimization
- **Part V: Real Robot and Sim-to-Real**
  - Causes of domain gaps, techniques and tips to address sim-to-real gap
- **Part VI: Embodied AI Tasks in ManiSkill and Visual Learning Challenges**
  - Summary of ManiSkill 2021, preview of ManiSkill 2022