Time: 2026-02-19 10:00-11:20am
Location: CSE2 271

CSE 571: AI-Robotics

Deep Perception for Manipulation

Wenlong Huang

Ph.D. Candidate, Stanford University

UNIVERSITY of WASHINGTON

Stanford University

Myth and truth about robot perception

For the following statement, tell us whether you think it is true or false, **and why.**

Statement:

Robot Perception == Computer Vision

3

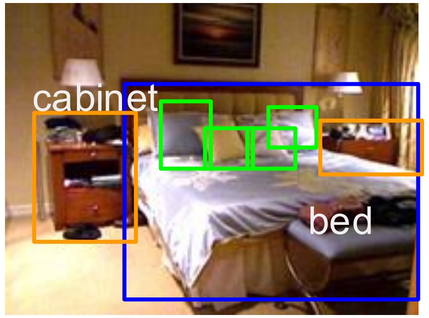
True or false, **and why.**

Slides adapted from EECS227 by Shuran Song

Statement:

Robot Perception == Computer Vision

Not true. Robot need more information: 3D, physical properties …



Object detection

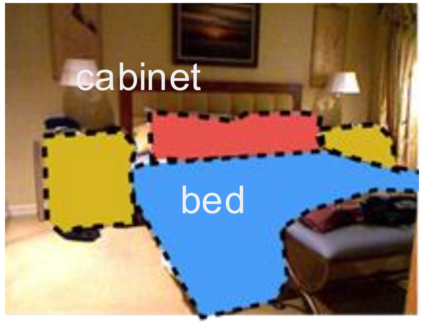
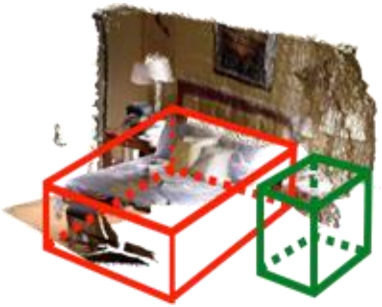


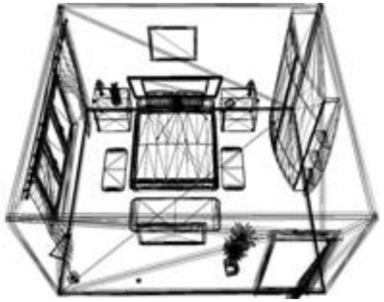
Image segmentation

How far is the cabinet?



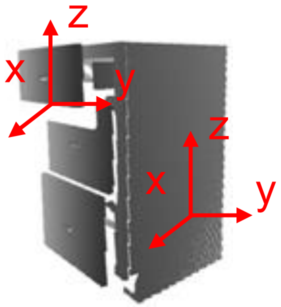
3D object detection

How to get there?



3D reconstruction

How to open the cabinet?



Object articulation

How much force to use

Friction, mass

Slides adapted from EECS227 by Shuran Song

Statement:

Robot perception is harder than computer vision?

True or false, **and why.**

Statement:

Robot perception is harder than computer vision?

Sometimes, yes.

Need to output more information (previous slide)
Sensitive to speed
Sensitive to error

Slides adapted from EECS227 by Shuran Song

Statement:

Robot perception is harder than computer vision?

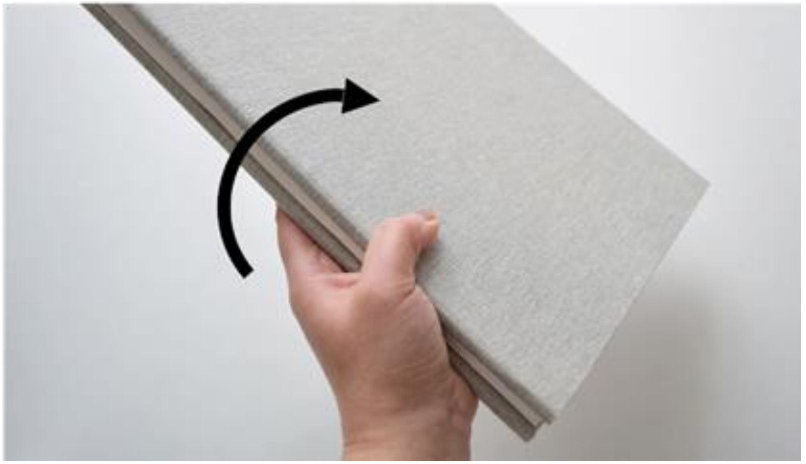
However, sometimes, robot can use action to simplify perception

Slides adapted from EECS227 by Shuran Song

Statement:

Robot perception is harder than computer vision?

Not necessarily. Sometimes, robot can use action to simplify perception



Action	Dipping	Pushing	Flipping
Information	Temperature	Weight	Title
Planing	Swim	Lift up the box	Read the book

Slides adapted from EECS227 by Shuran Song

Statement:

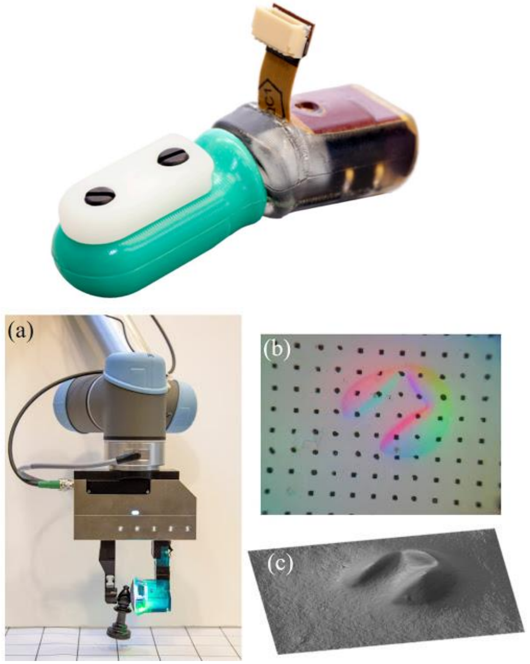
Perception is only about visual data

True or false, **and why.**

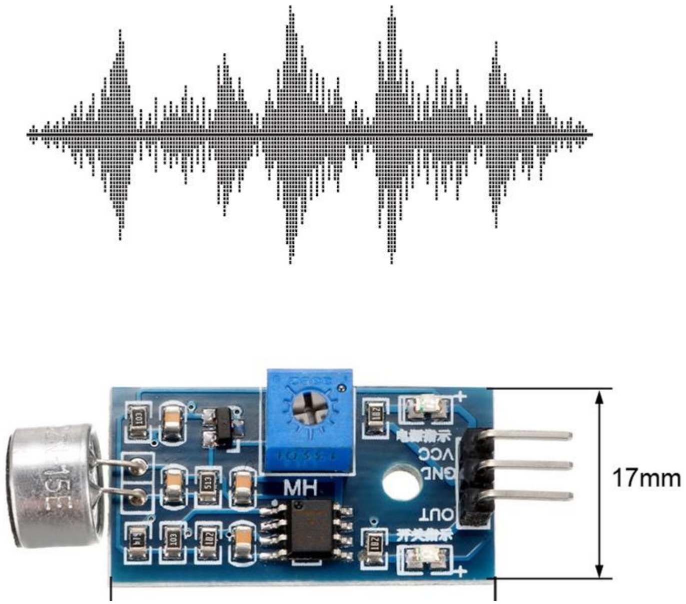
Statement:

Perception is only about visual data

Not true. There are many other sensory modalities a robot can use.



Tactile sensor



Acoustic sensor

Slides adapted from EECS227 by Shuran Song

Why we need different sensors?

Provide complimentary information – sense different signal

Provide redundant information – improve robustness

Help to develop a rich and coherent representation of the word

-- The association between different modality in our experience let us learned a share representation that let us infer other modality when they are absent

Slides adapted from EECS227 by Shuran Song

Statement:

Perception is a module in a robotics system

True or false, **and why.**

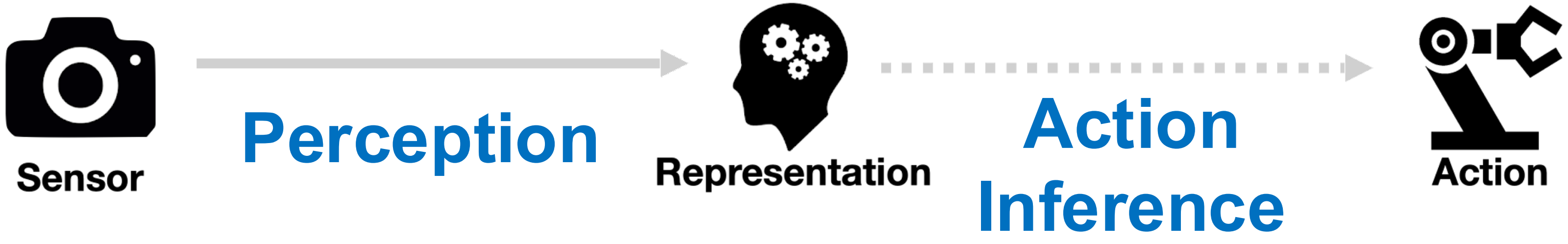
Slides adapted from EECS227 by Shuran Song

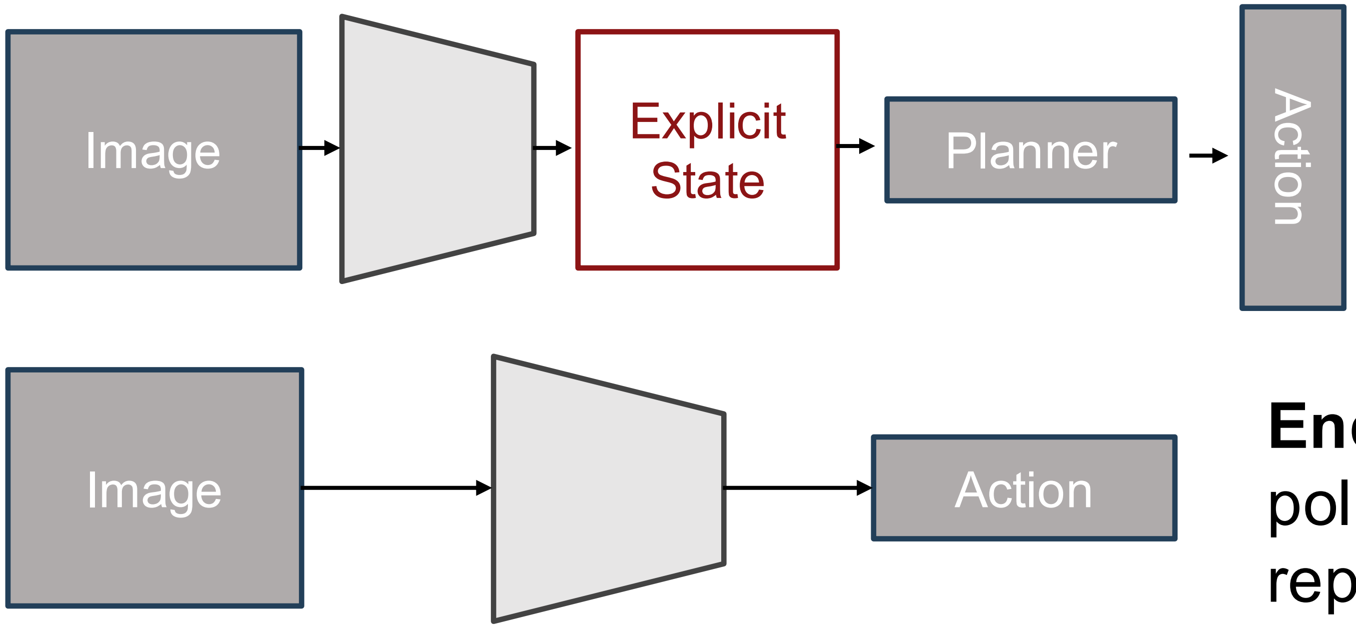
Statement:

Perception is a module in a robotics system

Yes, and No

Role of perception: (different designs)





Modular design: Perception converts sensor input into a representation of the world that could be used for planning

End-to-end design: Visuomotor policy builds an implicit state representation to predict action

Slides adapted from EECS227 by Shuran Song

Statement:

Perception is independent of tasks/objectives

True or false, **and why.**

Statement:

Perception is independent of tasks/objectives

No

Objects tell us where to carve the world

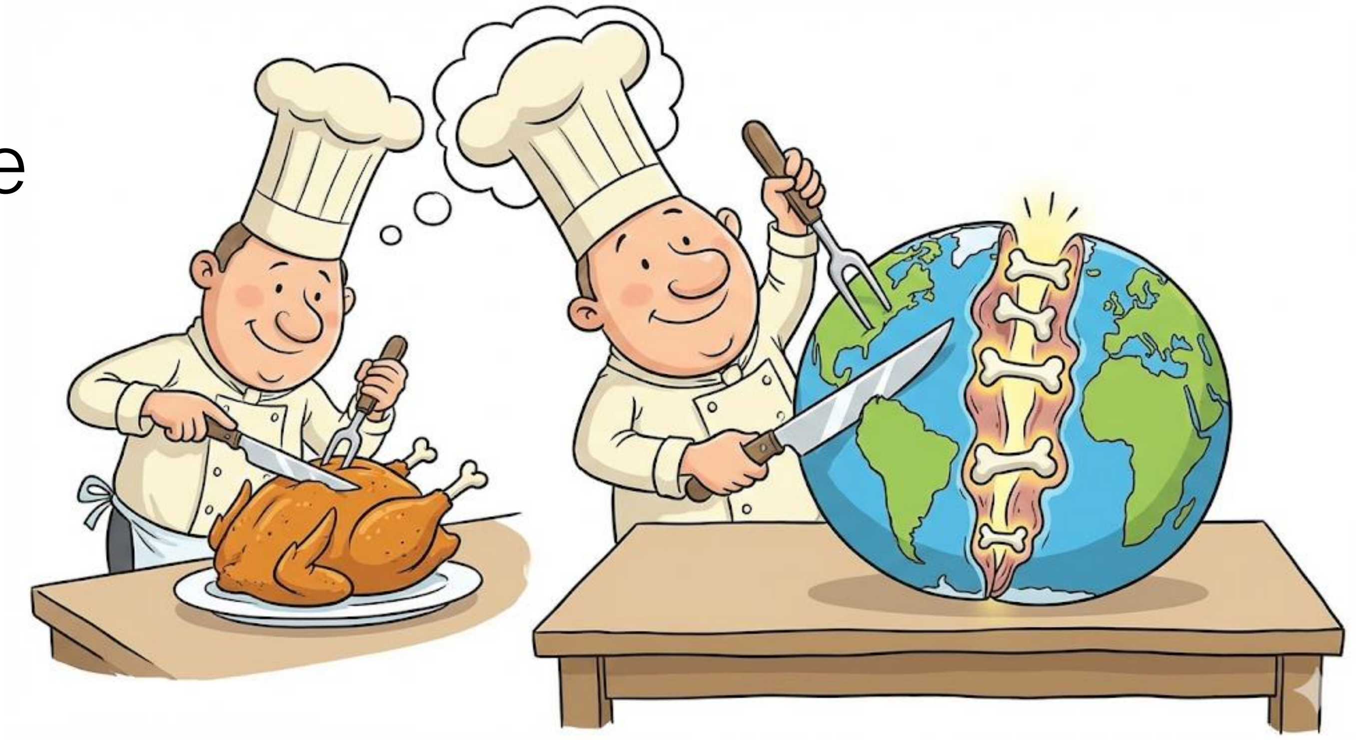


Image credit: Nano Banana Pro

Slides adapted from Leslie Kaelbling

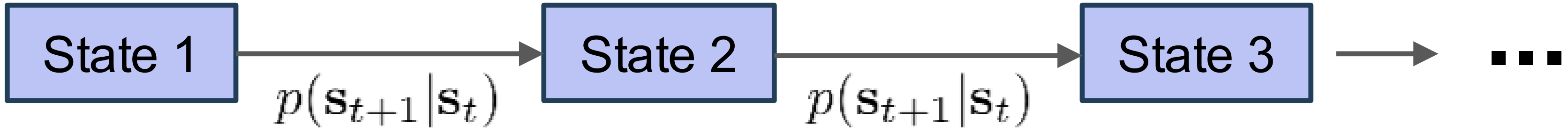
What is an “object” here?

What is an “object” here?

Markov Chain

Andrey Markov

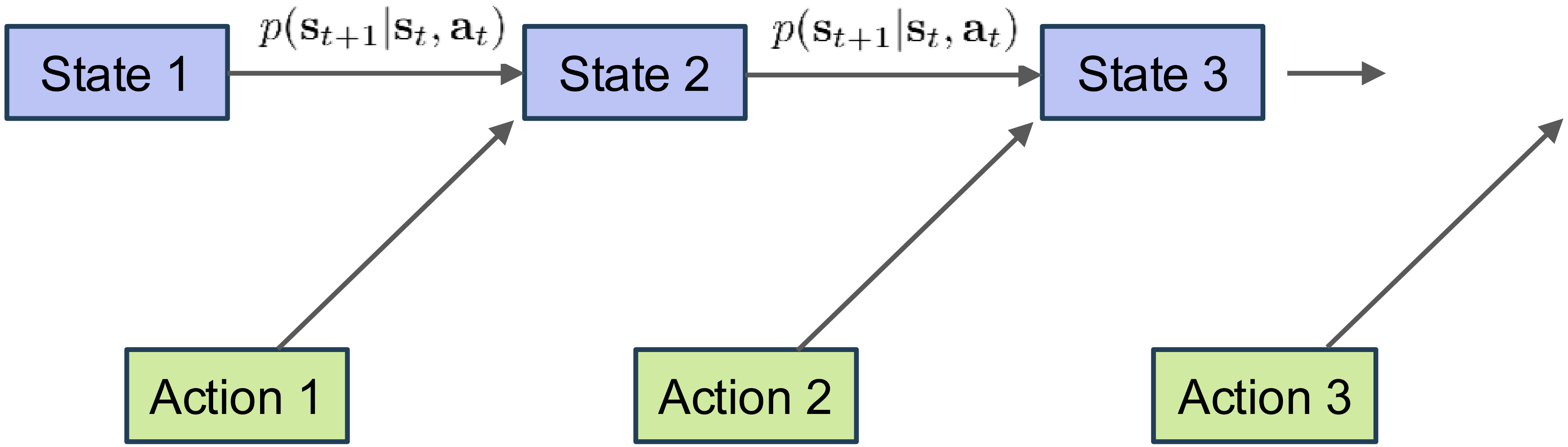
The most basic probabilistic model of a dynamical system
No actions yet, only states



Markov property

$$\mathbf{s}_{t+1} \perp \mathbf{s}_{t-1} | \mathbf{s}_t$$

Markov Decision Process (MDP)

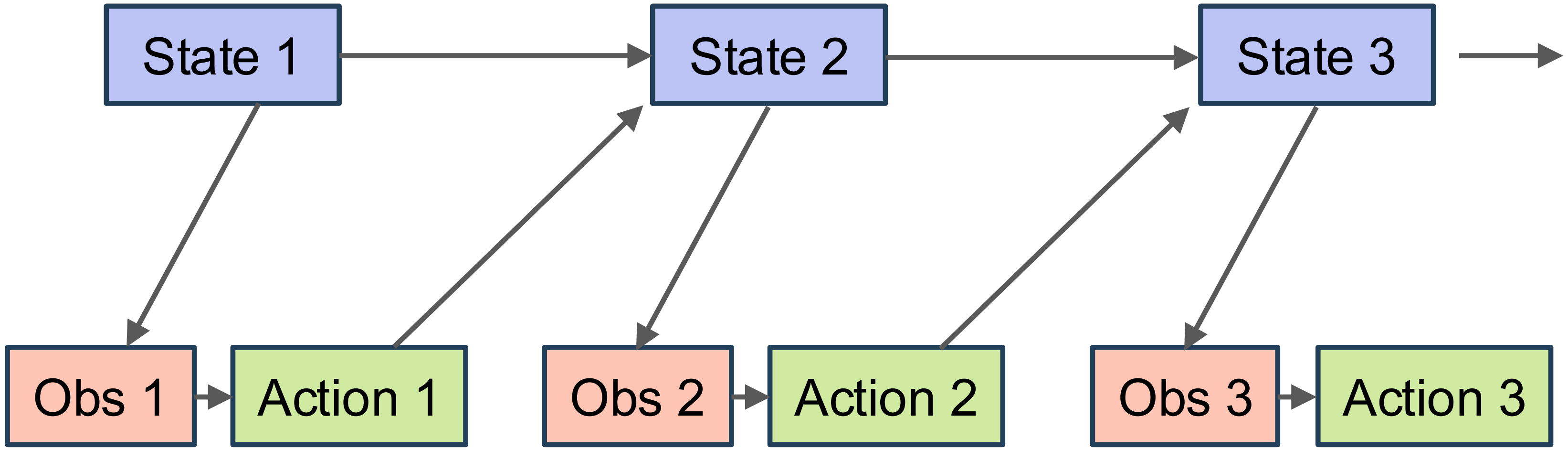


Richard Bellman

Probabilistic model for sequential decision making
Also assumes a given reward function $S \times A \to \mathbb{R}$

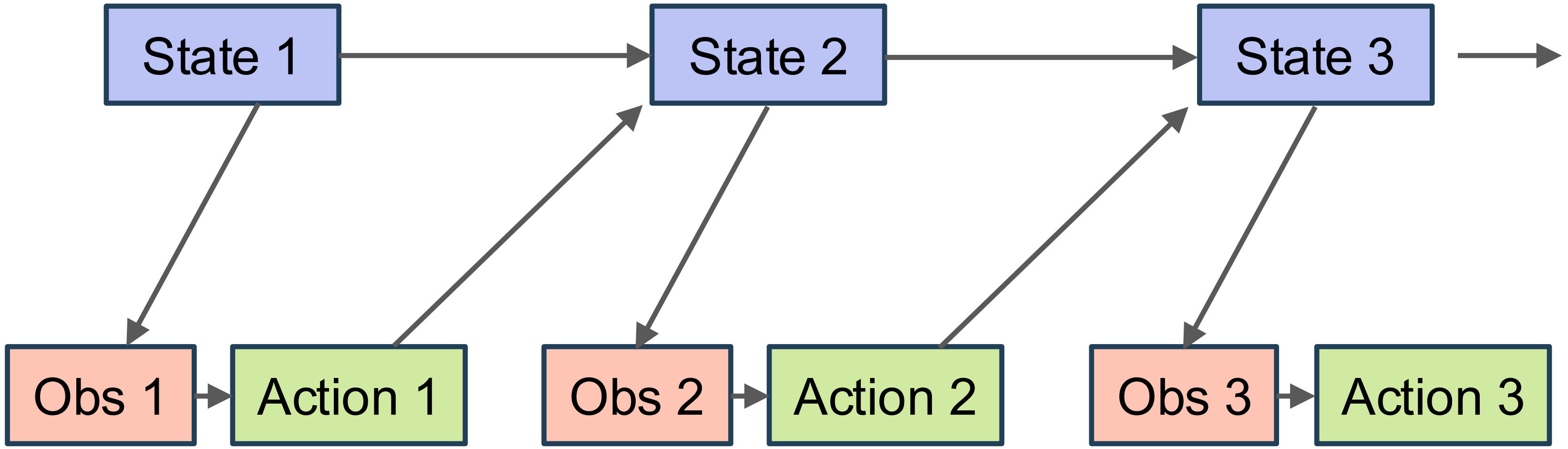
Slides adapted from CS285 by Sergey Levine

Partially Observable Markov Decision Process (POMDP)



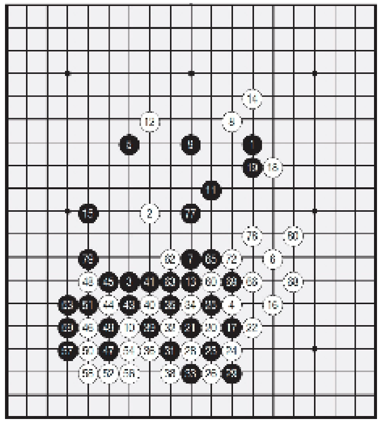
This is the regime real-world robots operate in, where we have access only to observations.

Partially Observable Markov Decision Process (POMDP)



What robot perception does: $\Pr(s_t = s \mid o_{1:t}, a_{1:t-1})$

What is a state? What is a representation?



3^{361} states?

Game of Go
- an exponentially large number of states?
- infeasible to enumerate, memorize, or search

$256^{3x500x500}$?

Images
Image space has exponentially more states than Go.

Examples from MIT - 6.8300/1 Advances in Computer Vision

Slides adapted from CS231a by Jeannette Bohg

What is a state? What is a representation?

DeepMind Control Suite. Tassa et al. 2018

Representations for Autonomous Driving

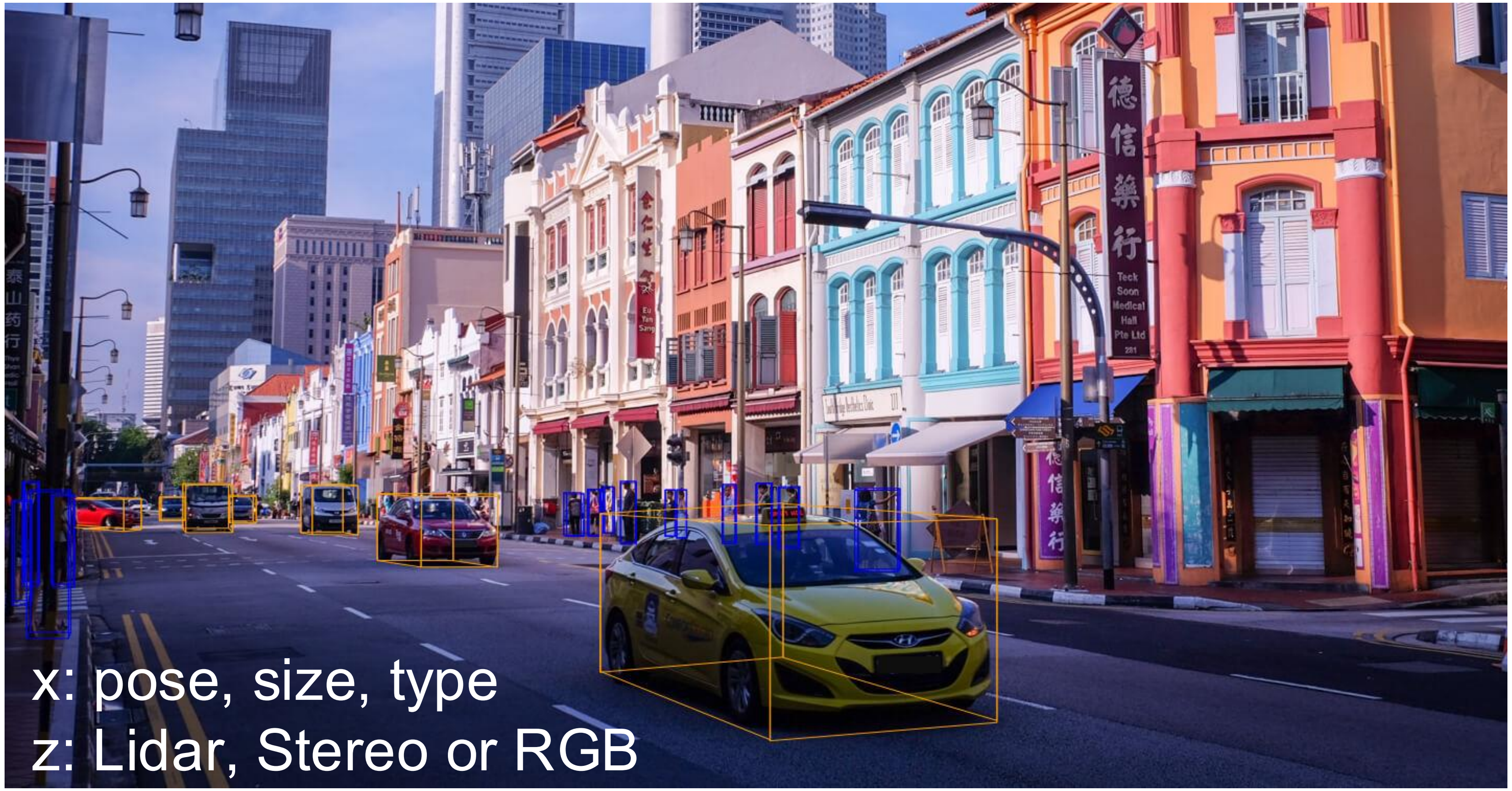
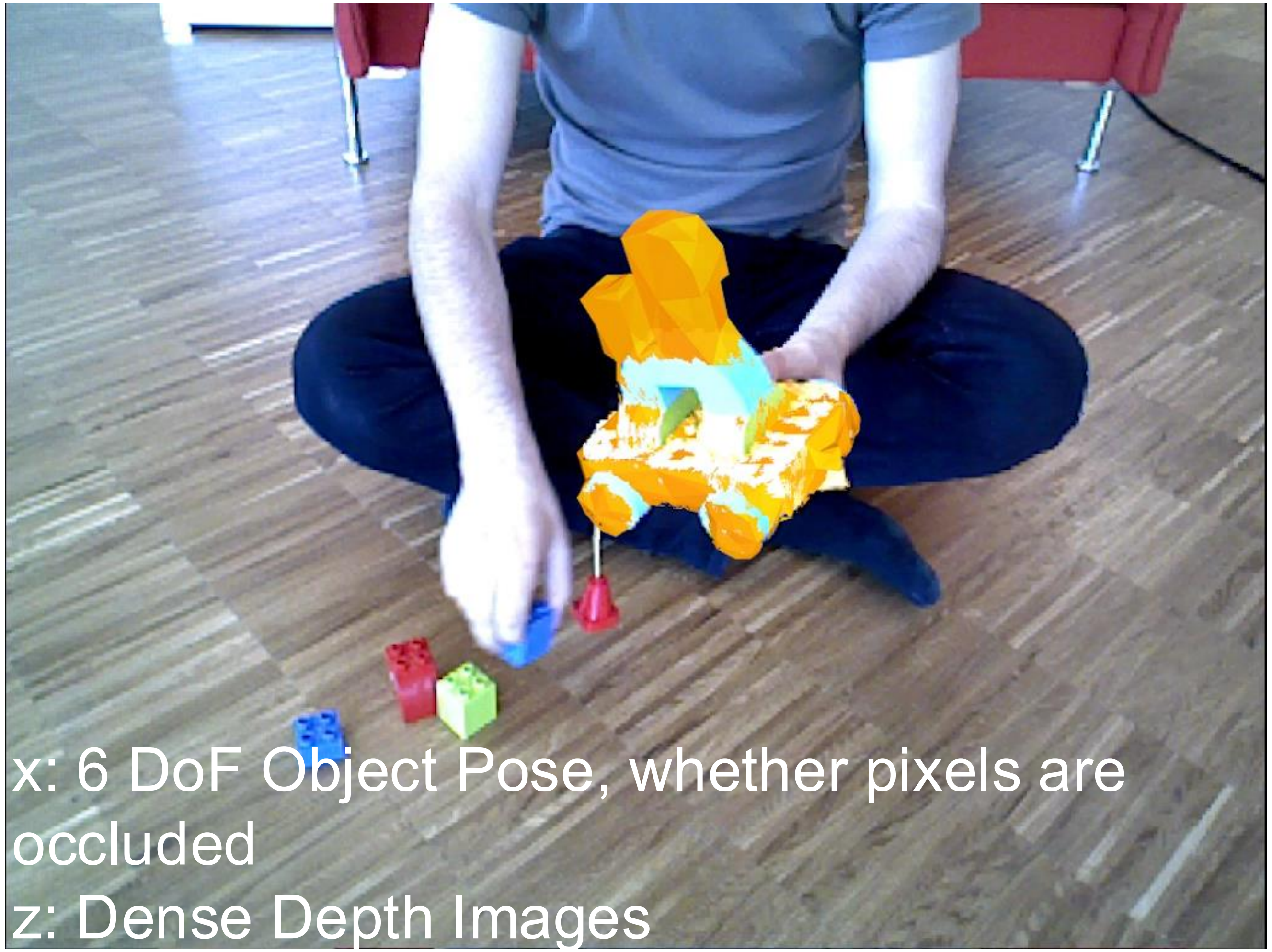


Image adapted from NuScenes by Motional. nuscenes.org

Representations for Manipulation



Manuel Wühtrich et al. "Probabilistic Object Tracking using a Depth Camera", IROS 2013

What is a state? What is a representation?

Pose?

Slides adapted from EECS227 by Shuran Song

Meaning in English

"the way that someone or something is shown or described:"

"a sign, picture, model, etc. of something"

- Cambridge Dictionary

Representations in Cognitive Science

Symbolic View

"[…] a hypothetical internal cognitive symbol that represents external reality" (Morgan '14)

"[…] a formal system for making explicit certain entities or types of information […]" (Marr '10)

"[…] intermediaries between the observing subject and the objects, processes or other entities observed in the external world. These intermediaries […] represent to the mind the objects of that world." (Wikipedia - Mental Representations - Representationalism)

Embodied View

"… actions are directly triggered by stimuli in the environment without the need for internal representations" (Gibson '66, Zech '19 on Embodied Cognition)

"… actions are represented by their anticipated effect, that is, action representations essentially entail a mental model of a needed future environmental state" (Jeannerod '06, Zech '19)

Representations in Machine Learning

“Features”, “A good representation is also one that is useful as input to a supervised predictor.” (Bengio ’14)

“create a representation of the data to provide the model with a useful vantage point into the data's key qualities. […] to train a model, you must choose the set of features that best represent the data.” (Google Crash Course of ML Concepts)

“ […] world models, internal models of how the world works.”; “(1) estimate missing information about the state of the world not provided by perception, (2) predict plausible future states of the world.” (YLC ’22)

Slides adapted from CS231a by Jeannette Bohg

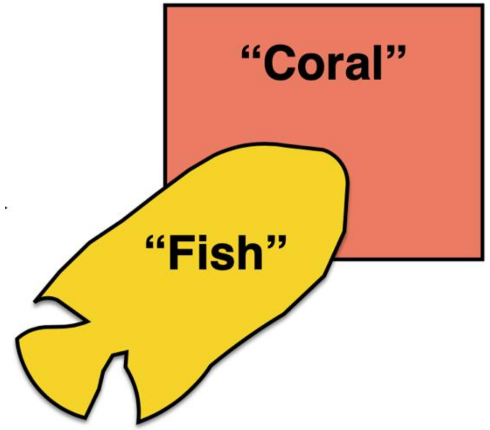
Requirements for Good Representations

"Solving a problem simply means representing it so as to make the solution transparent."

Herbert A. Simon, Sciences of the Artificial

Requirements for Good Representations

- Compact (minimal)
- Explanatory (sufficient)
- Disentangled (independent factors)
- Hierarchical (feature reuse)
- Generalizes over many tasks



[See "Representation Learning", Bengio 2013, for more commentary]

Case Studies of State Representations in Manipulation

- Simulator State (and Real2Sim2Real)
- Object-Centric Descriptors
- Affordances
- Keypoints
- Latent
- Hierarchical/Hybrid

Simulator State

DeepMind Control Suite. Tassa et al. 2018

Simulator State

BEHAVIOR-1K

URDF

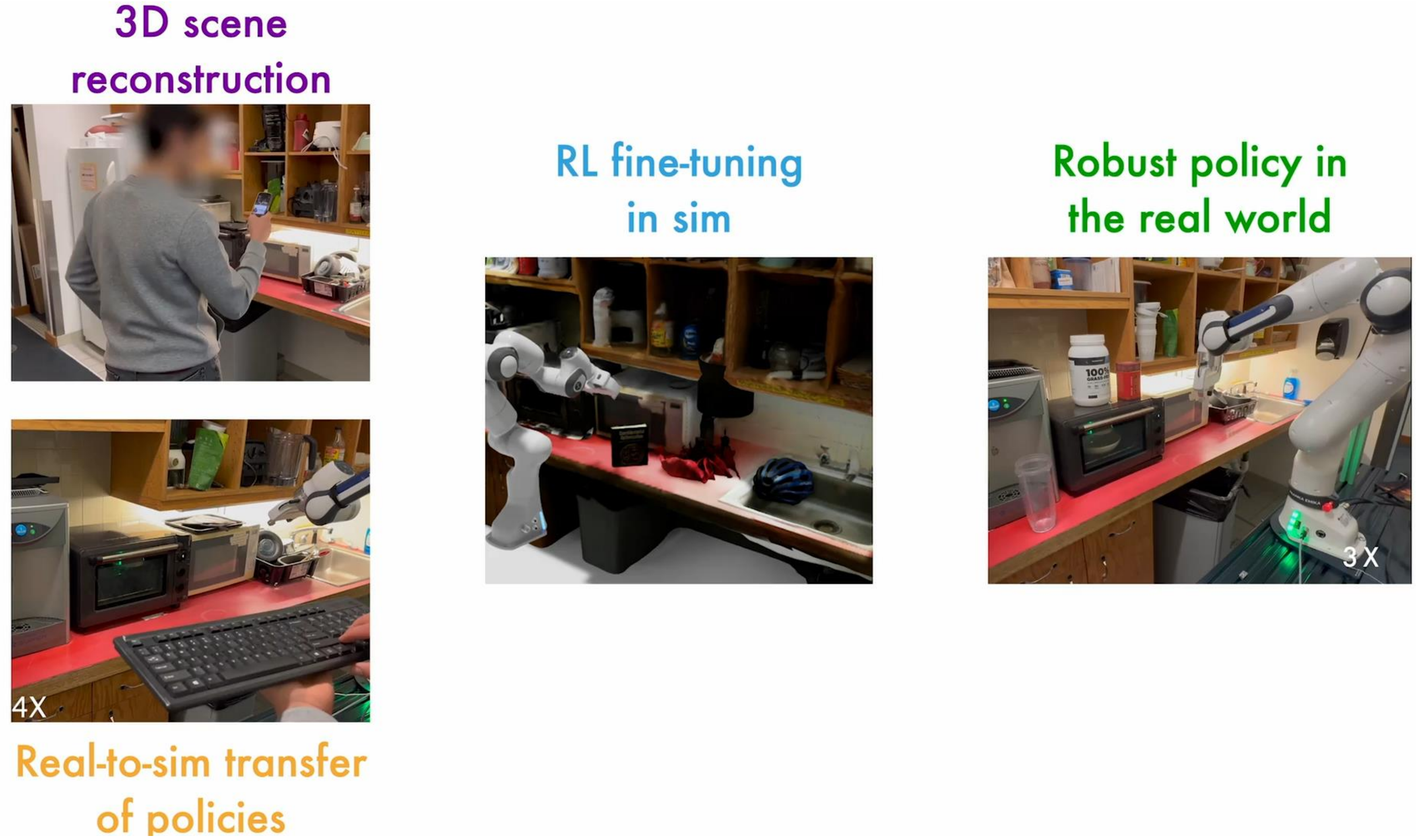
What:
Unified Robot Description Format Kinematic and basic physics description of a robot

How:

- XML format
- Tags: link or joint
- Kinematic tree structure
- Order in the file does not matter

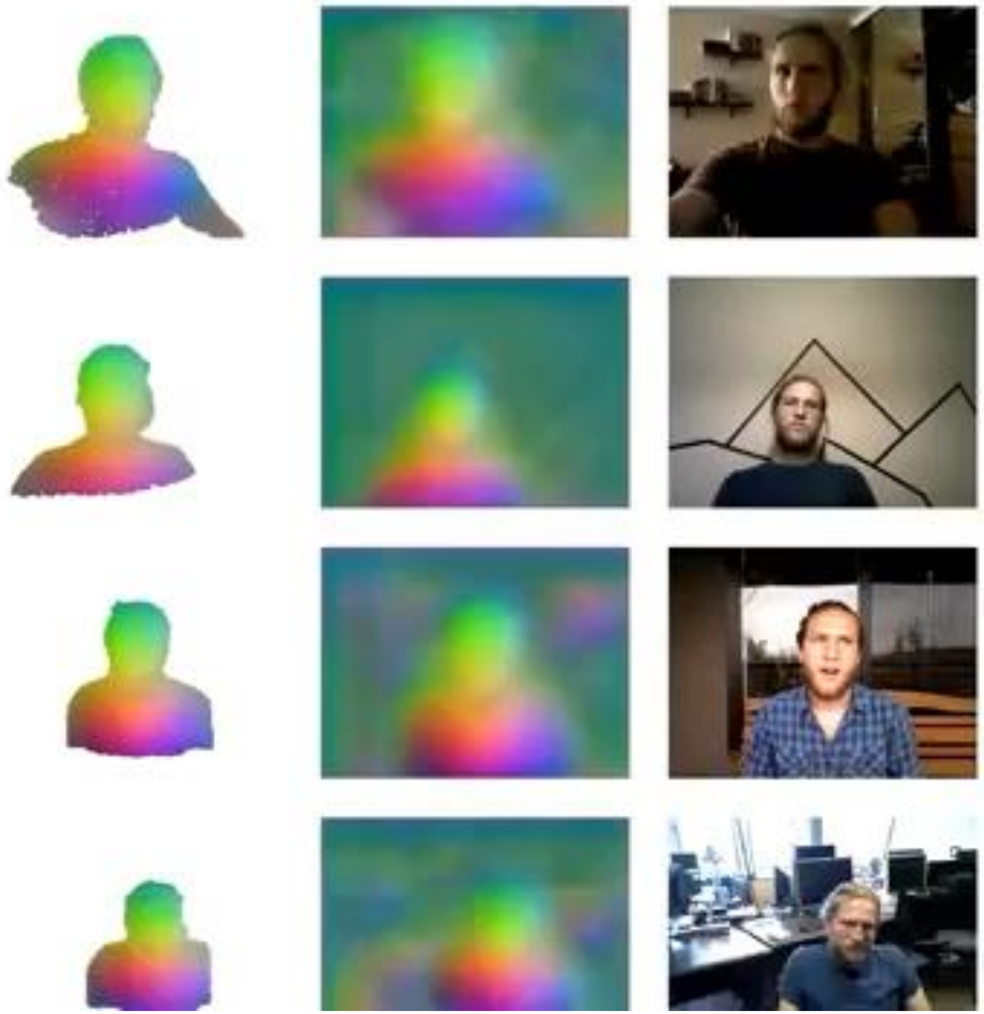
Slides adapted from EECS227 by Shuran Song

Real2Sim2Real with Simulator State



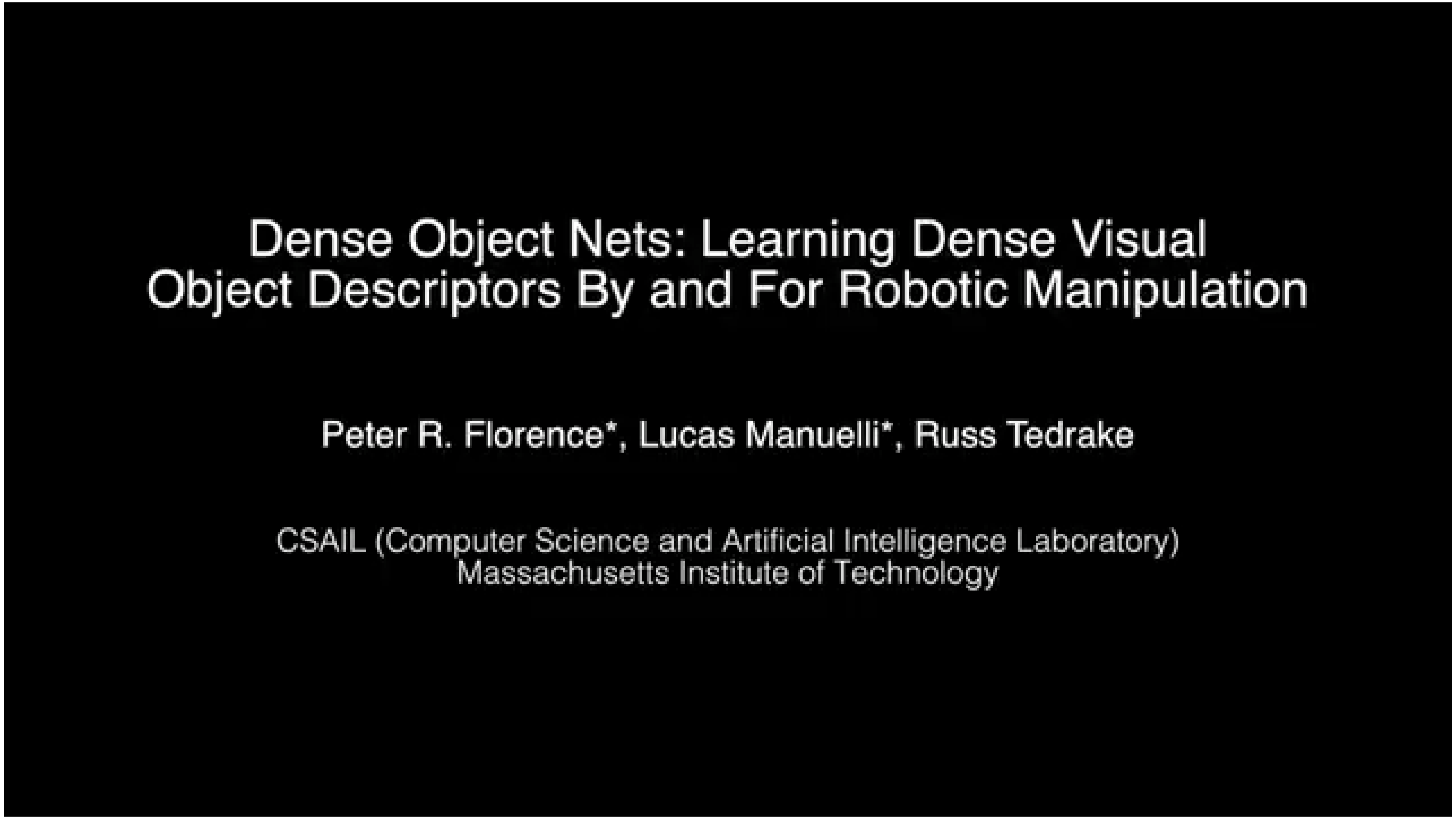
Torne et al., RSS 2024

Object-Centric Descriptors



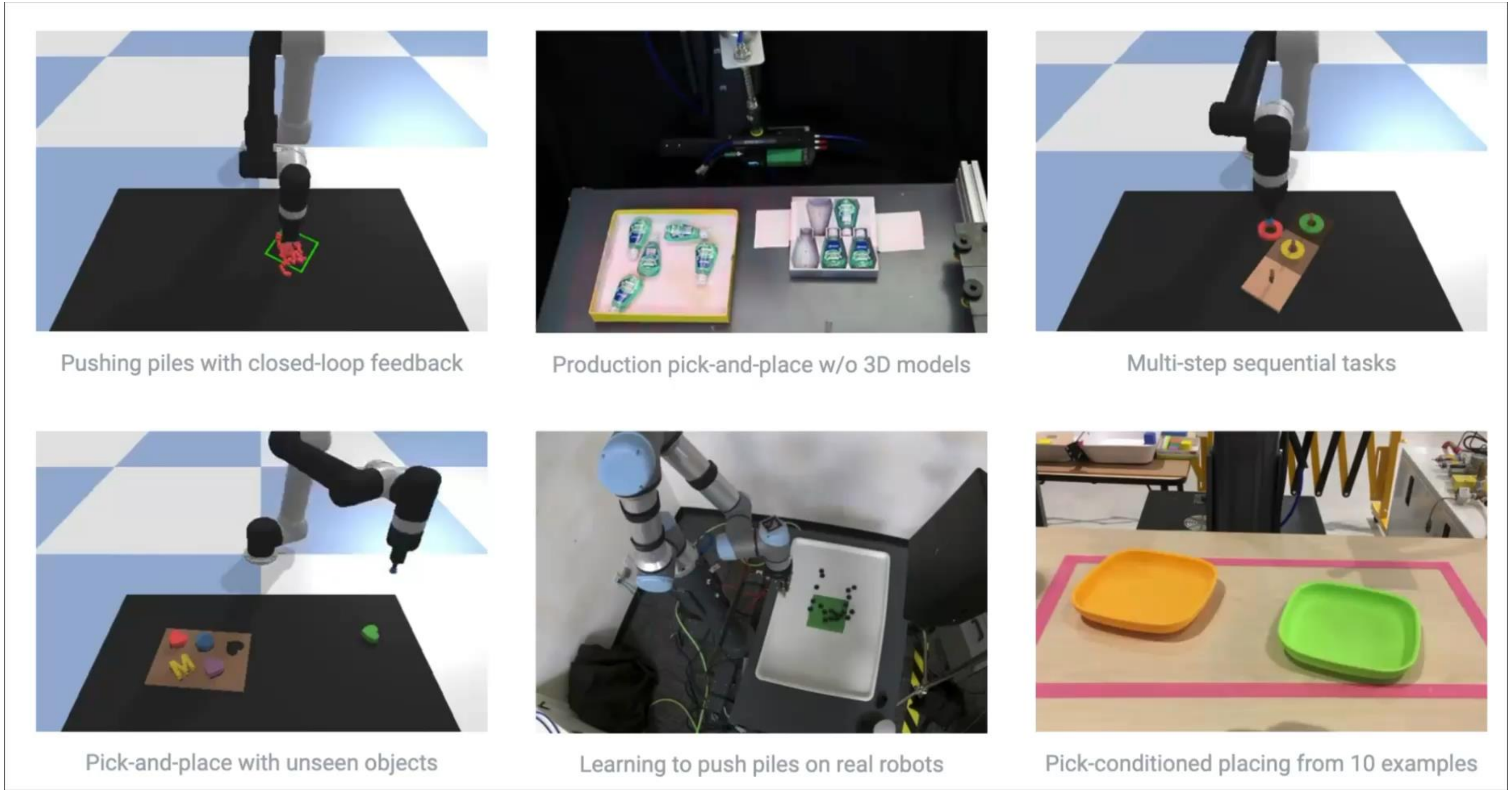
Schmidt et al., ICRA 2017

Object-Centric Descriptors



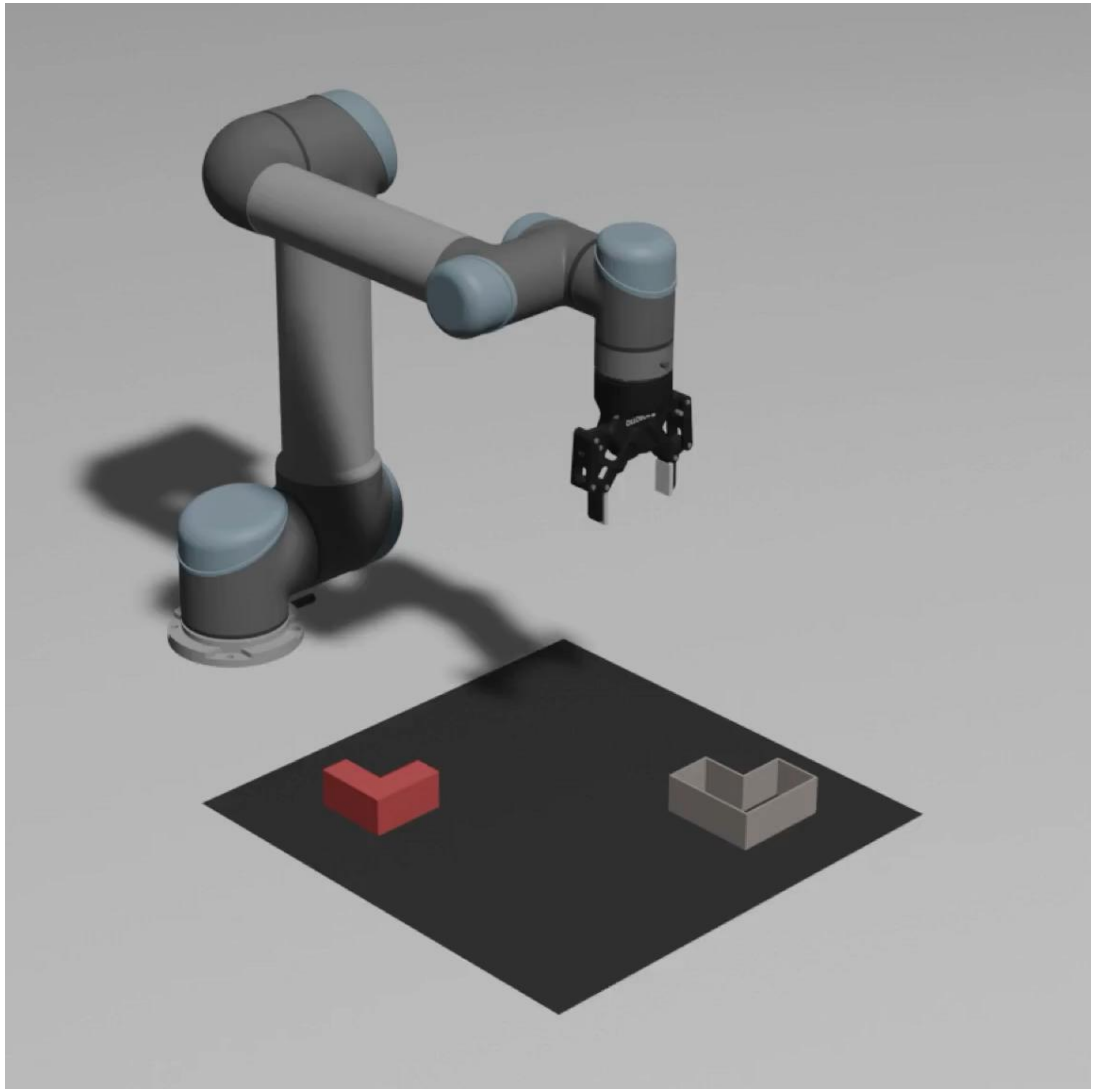
Florence et al., RA-L 2020 (Best Paper Award)

Affordances



Zeng et al., CoRL 2020 (Best Paper Presentation Award Finalist)

Affordances



Zeng et al., CoRL 2020 (Best Paper Presentation Award Finalist)

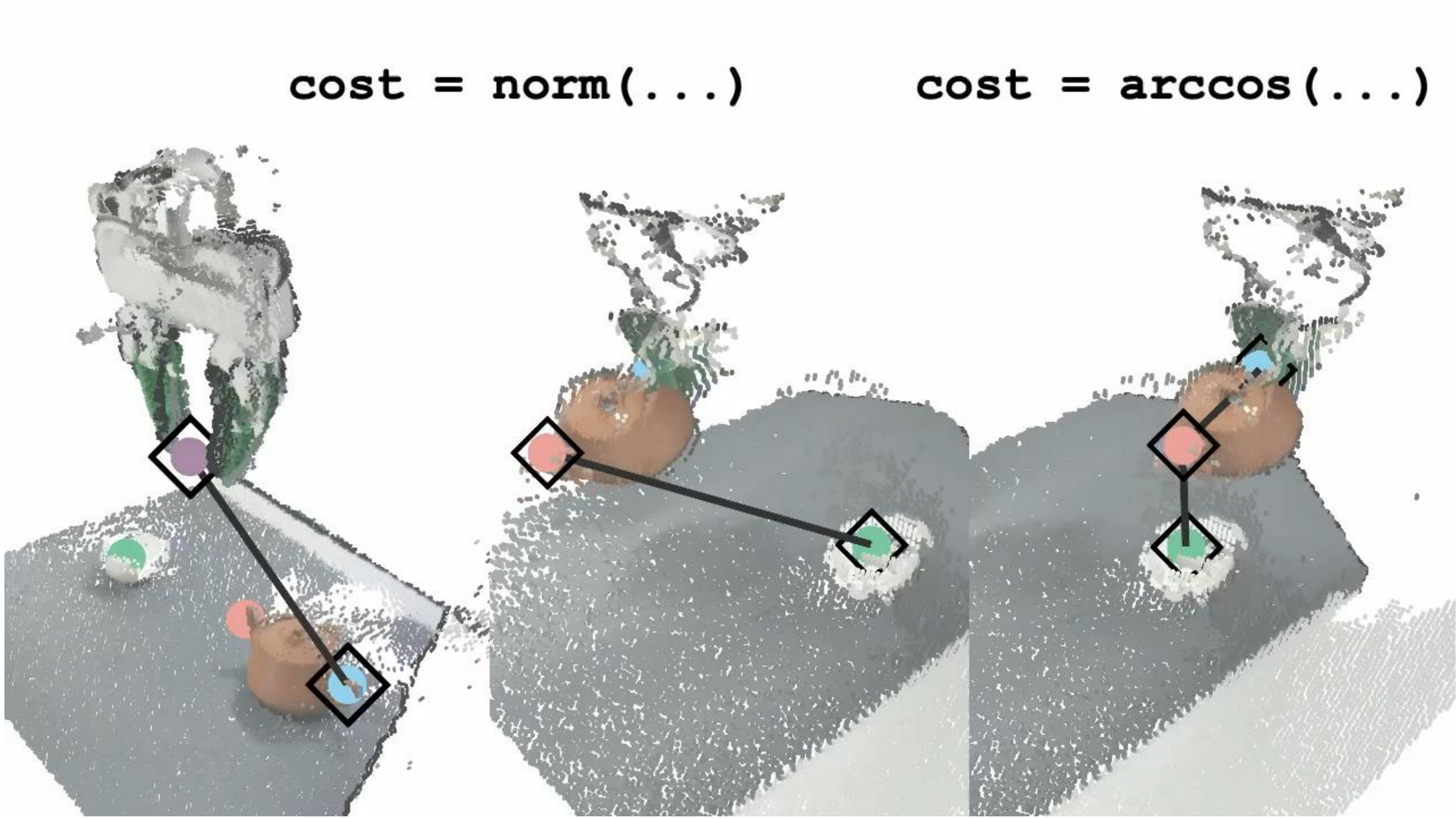
Affordances

Tang et al., ICRA 2025 (Best Paper Award Finalist)

Keypoints

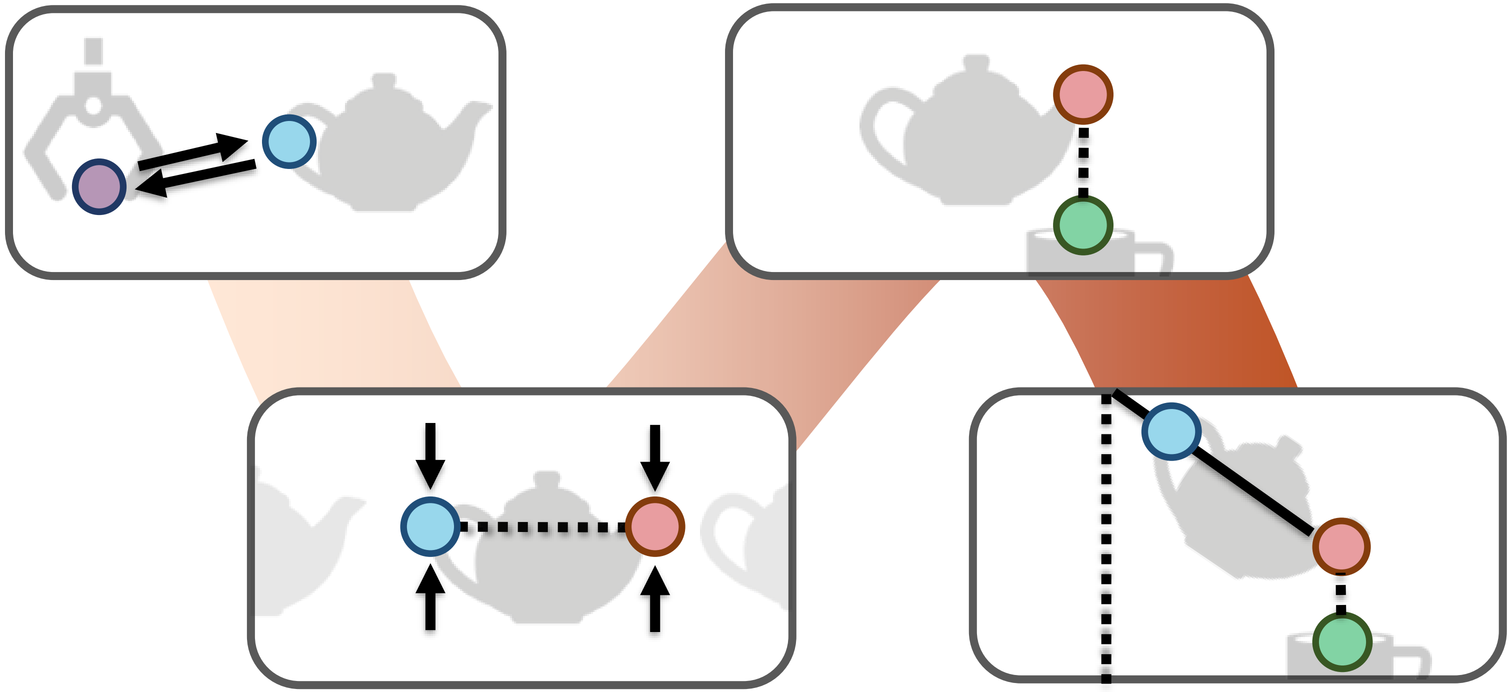
Manuelli et al., ISRR 2019

Keypoints



Huang et al., CoRL 2024 (Best Paper Award at CoRL LEAP)

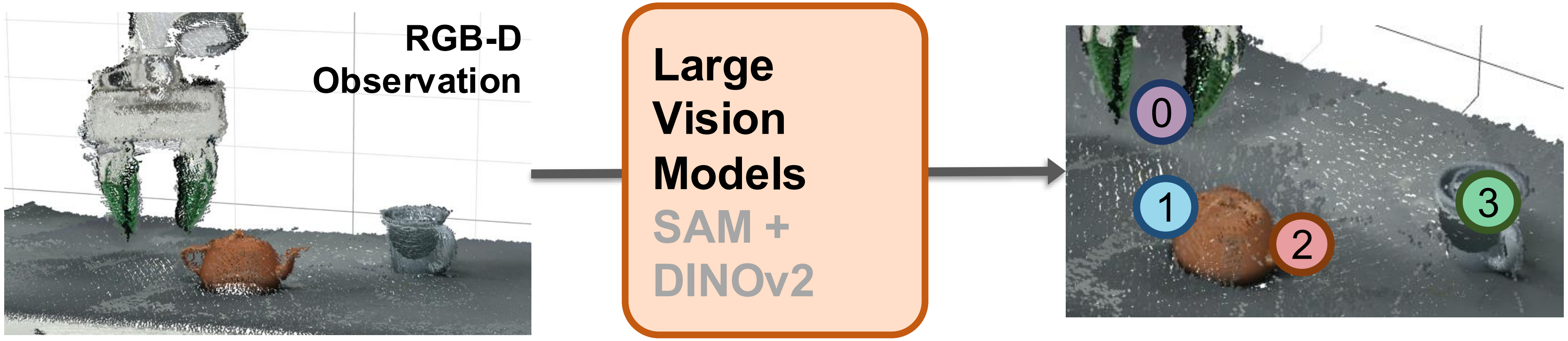
Keypoints



Huang et al., CoRL 2024 (Best Paper Award at CoRL LEAP)

Keypoints

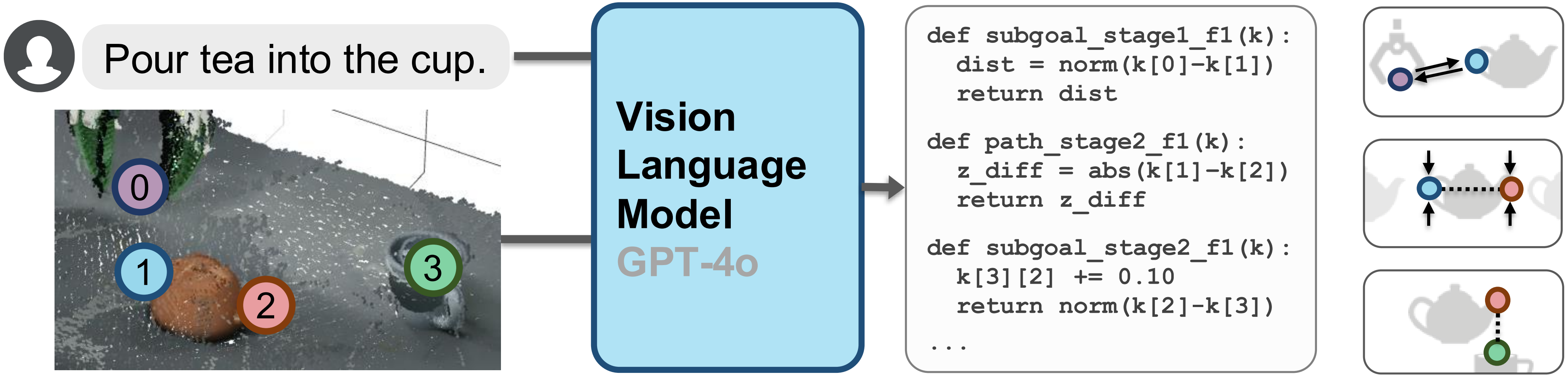
- **Step 1**: Obtain a set of semantically meaningful keypoints in the scene



Huang et al., CoRL 2024 (Best Paper Award at CoRL LEAP)

Keypoints

- Step 1: Obtain a set of semantically meaningful keypoints in the scene
- **Step 2**: Visually prompt VLM to write keypoint-based constraint code.



Huang et al., CoRL 2024 (Best Paper Award at CoRL LEAP)

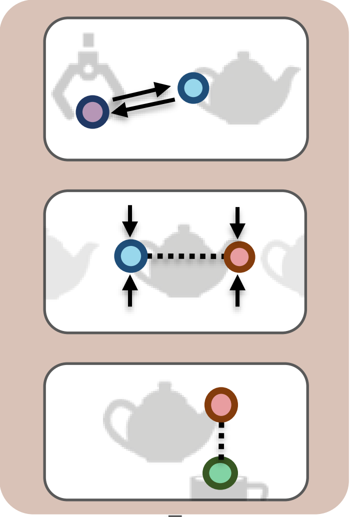
Keypoints

- **Step 1**: Obtain a set of semantically meaningful keypoints in the scene
- **Step 2**: Visually prompt VLM to write keypoint-based constraint code.
- **Step 3**: Perform constrained optimization to obtain robot actions.

Pour tea into the cup.

Vision Language Model
GPT-4o

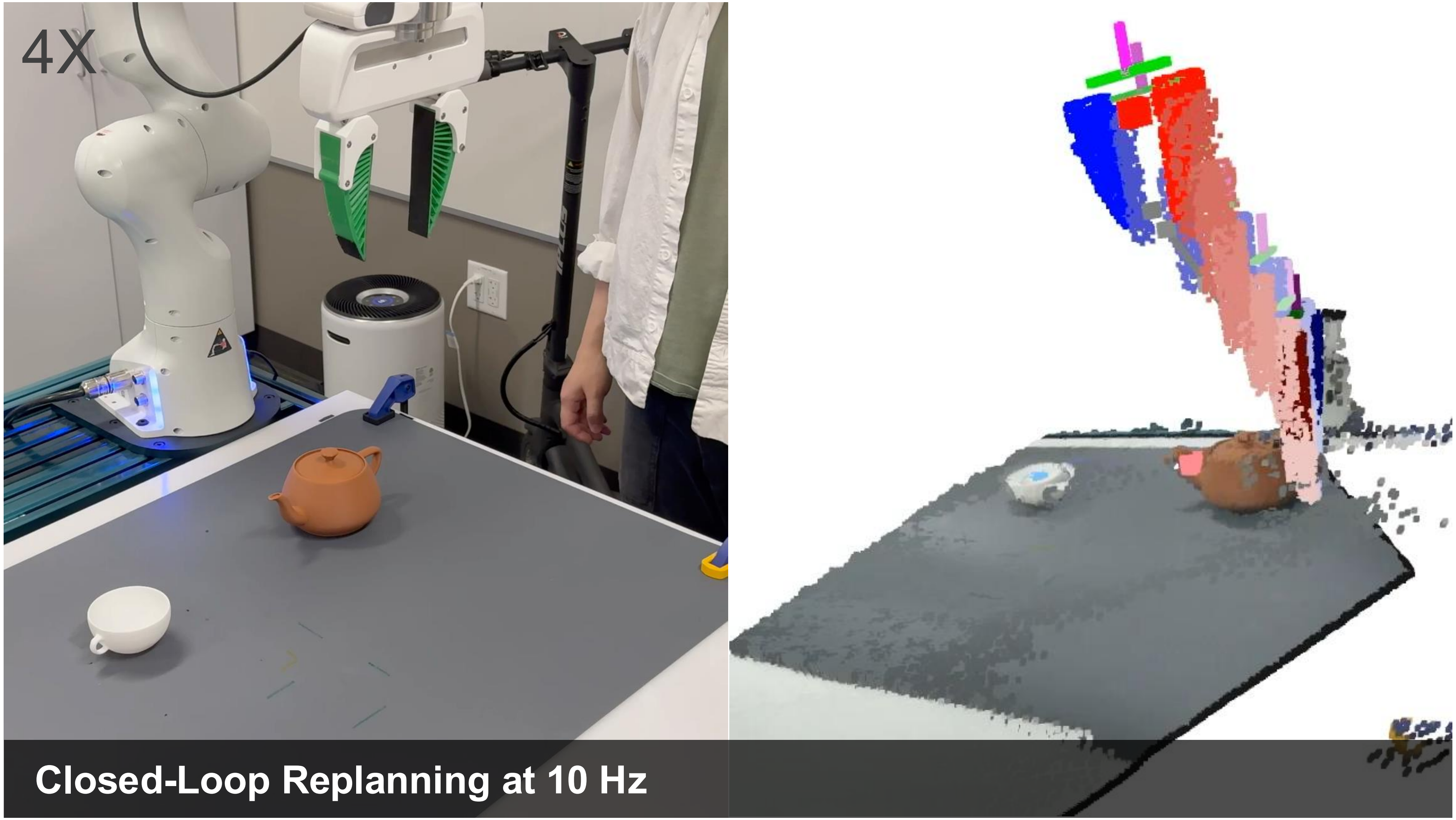
```
def subgoal_stage1_f1(k):
  dist = norm(k[0]-k[1])
  return dist

def path_stage2_f1(k):
  z_diff = abs(k[1]-k[2])
  return z_diff

def subgoal_stage2_f1(k):
  k[3][2] += 0.10
  return norm(k[2]-k[3])
...
```



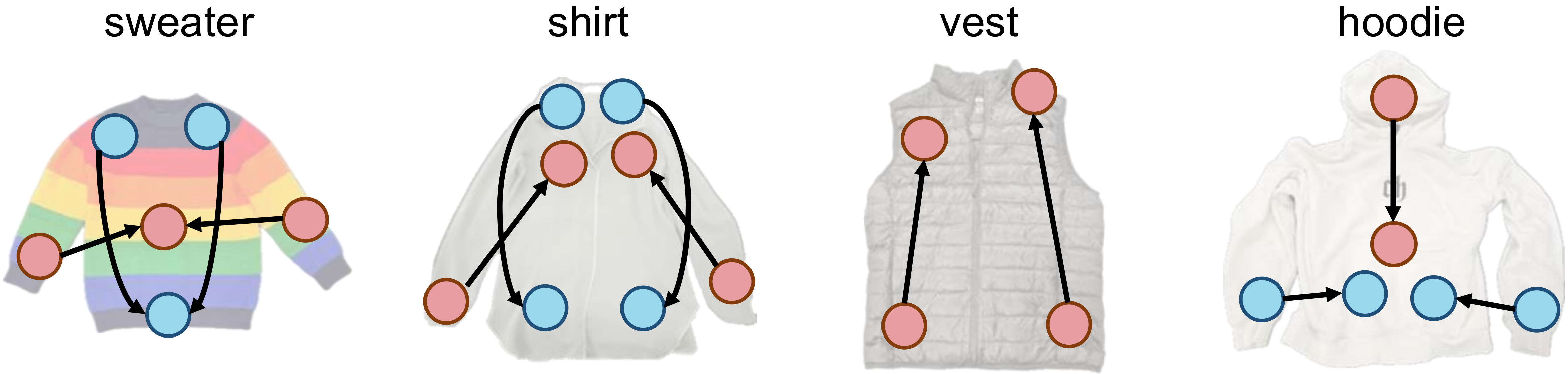
Constrained Optimization

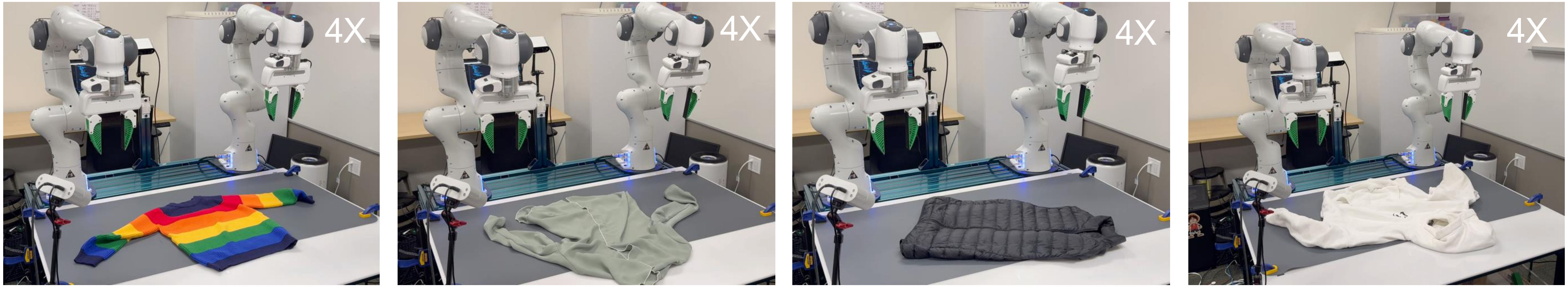
End-Effector Actions

Huang et al., CoRL 2024 (Best Paper Award at CoRL LEAP)



Huang et al., CoRL 2024 (Best Paper Award at CoRL LEAP)





Huang et al., CoRL 2024 (Best Paper Award at CoRL LEAP)

Hierarchical/Hybrid State Representation

Text Instruction

Image Obs

Vision Language Model

High-Level
State Representation
(Implicit)

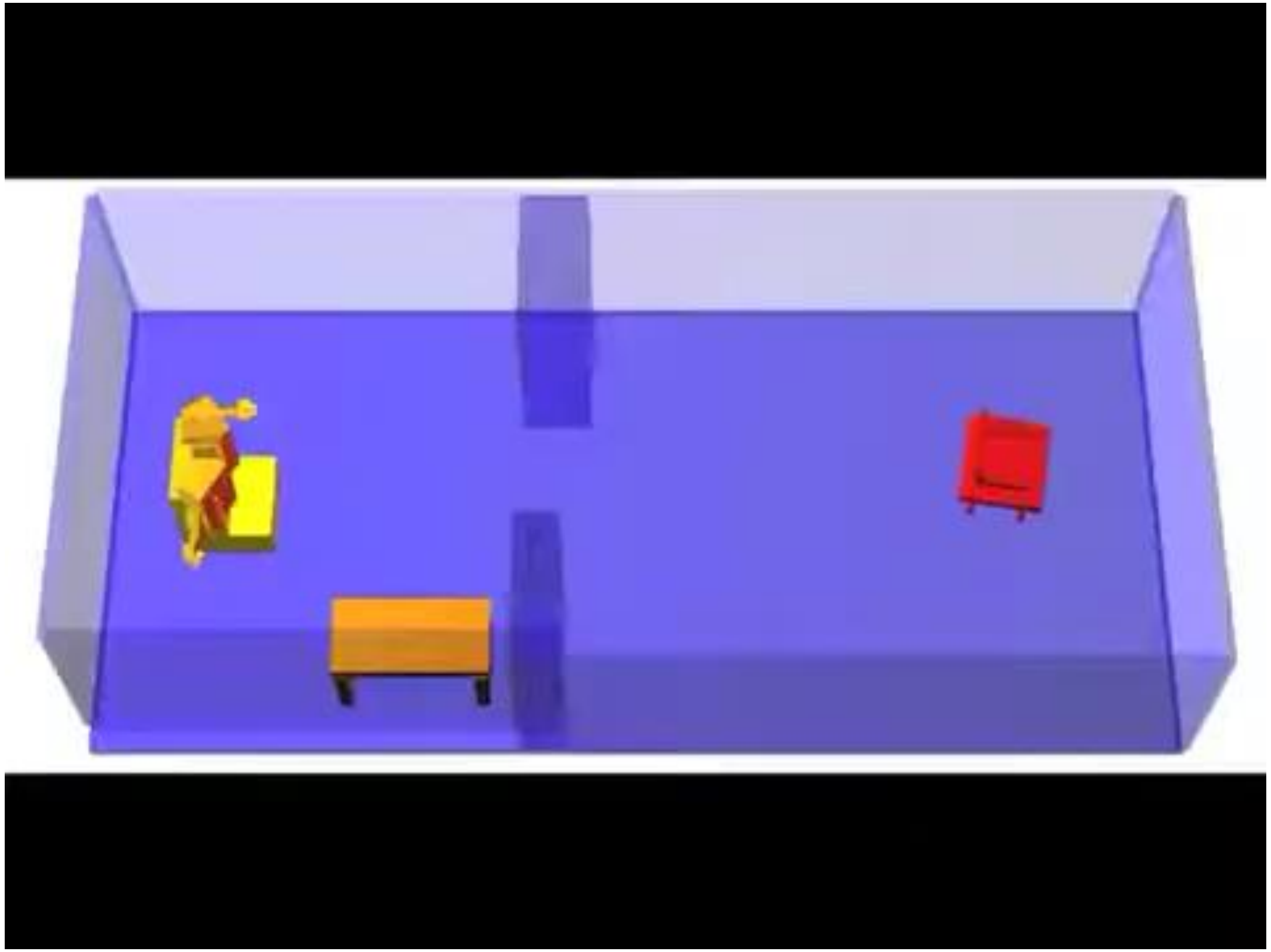
Low-Level State Representation
(Keypoints + Point Cloud)

Hierarchical/Hybrid State Representation

Task and Motion Planning

Discrete state: predicates (e.g., “grasped” or “not grasped”)

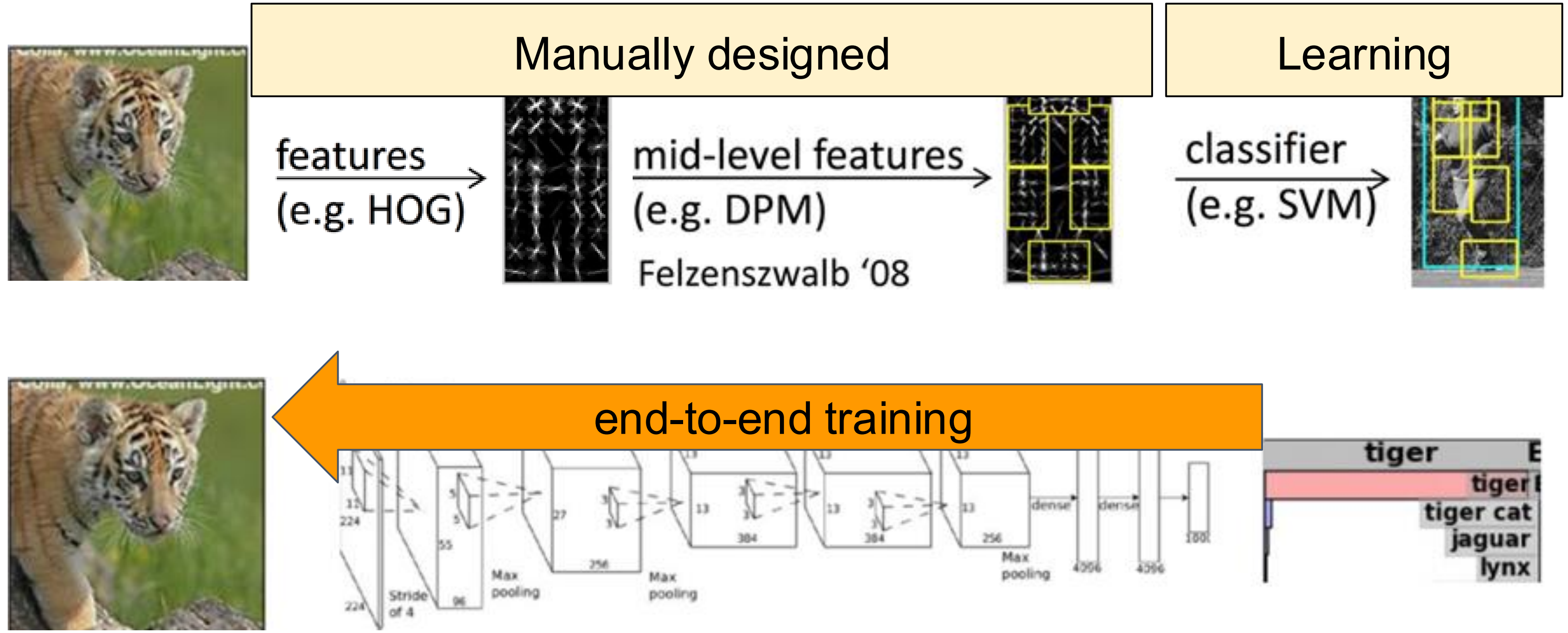
Continuous state: collision geometry of the robot and the scene



Levihn et al., IROS 2013

Latent State Representation

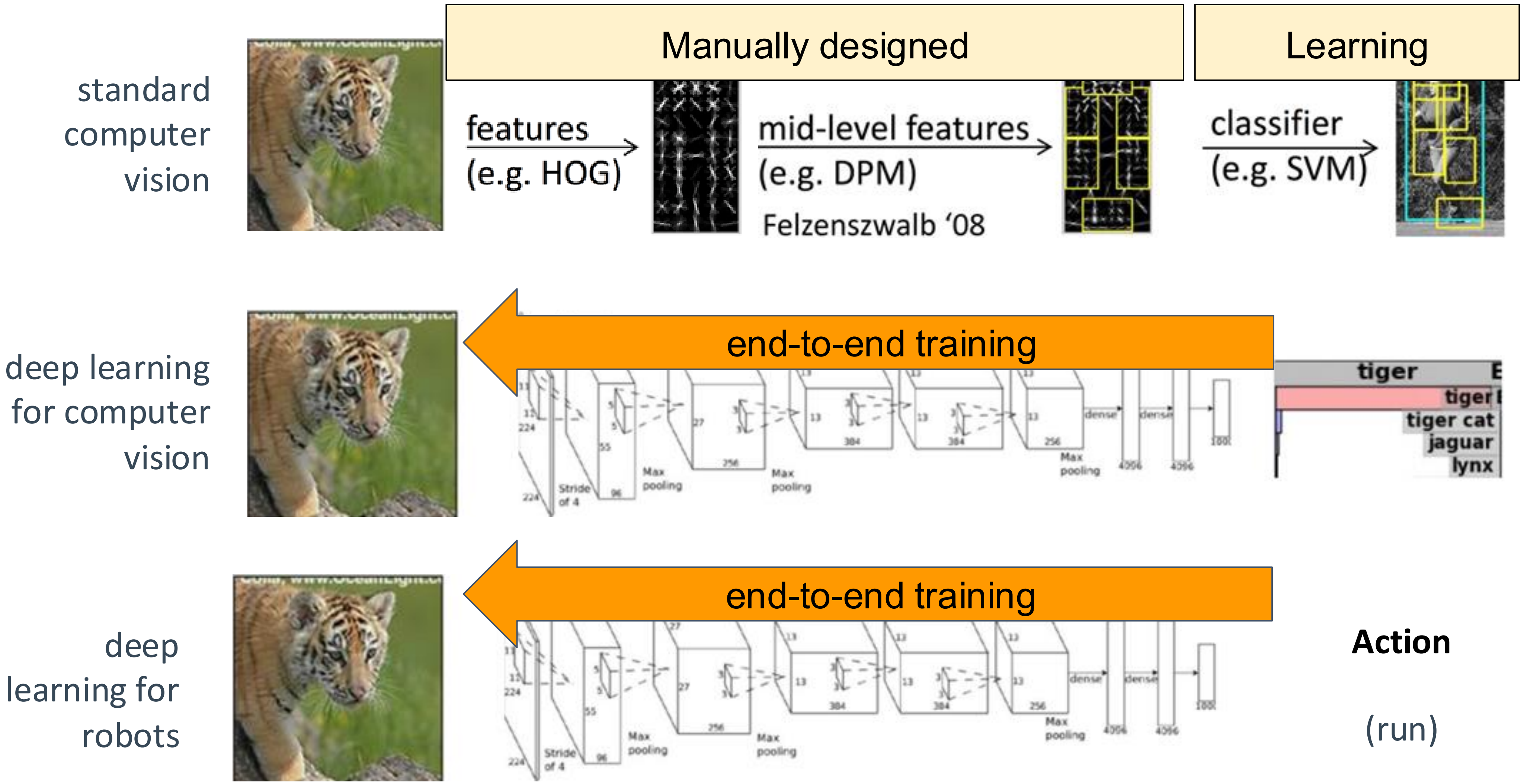
Why Latent State Representation?



Benefit:

- No manual feature engineering
- Optimize the features for the target task (diff. task requires diff. representations)

Latent State Representation in Policies



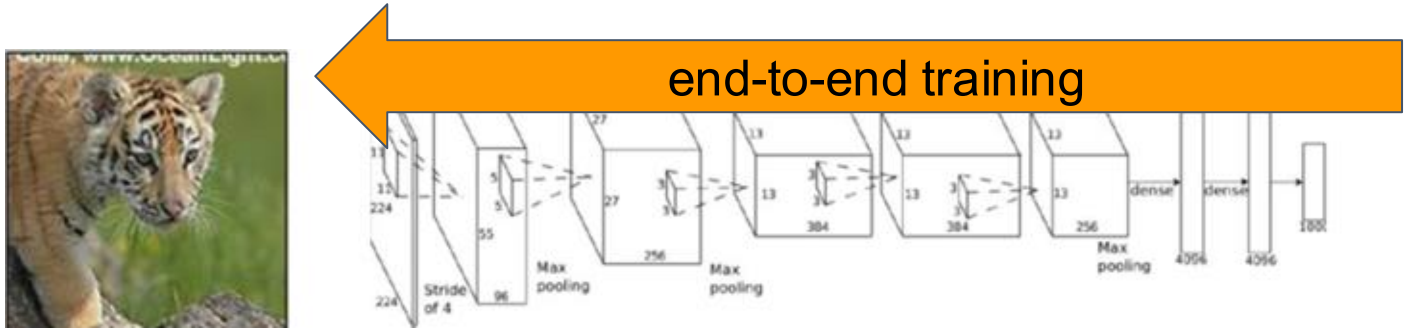
Slides adapted from EECS227 by Shuran Song

Latent State Representation in Policies

Benefit (hopefully …):

- Less manual engineering on the intermediate representation between perception and planning (Object pose? Grasp pose? Object shape?)
- Optimize representation for the target task (diff. task requires diff. representation)
- More robust to perception error

deep learning for robots

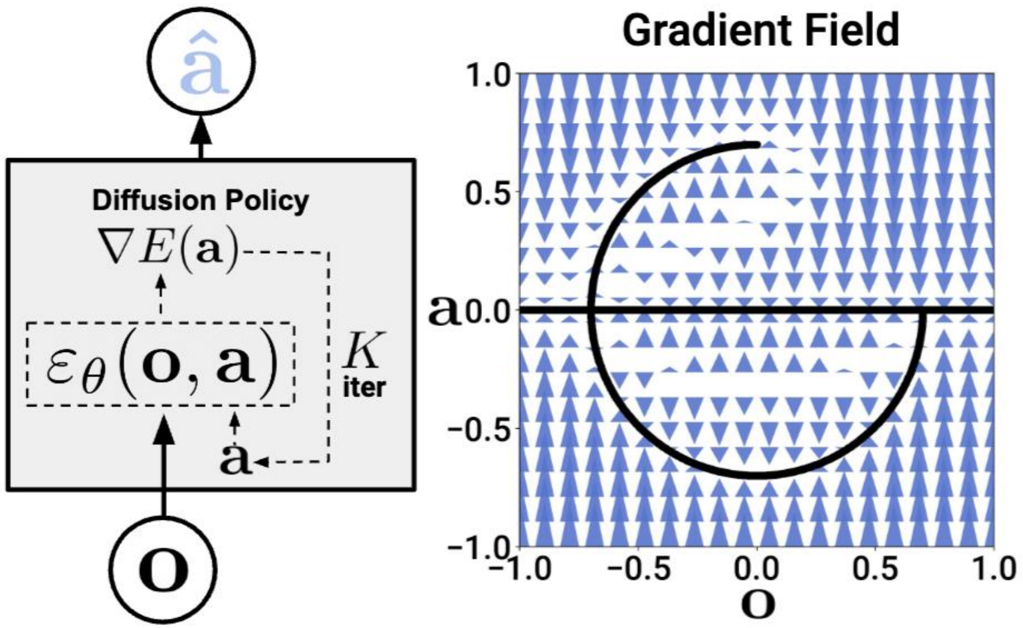


Action

(run)

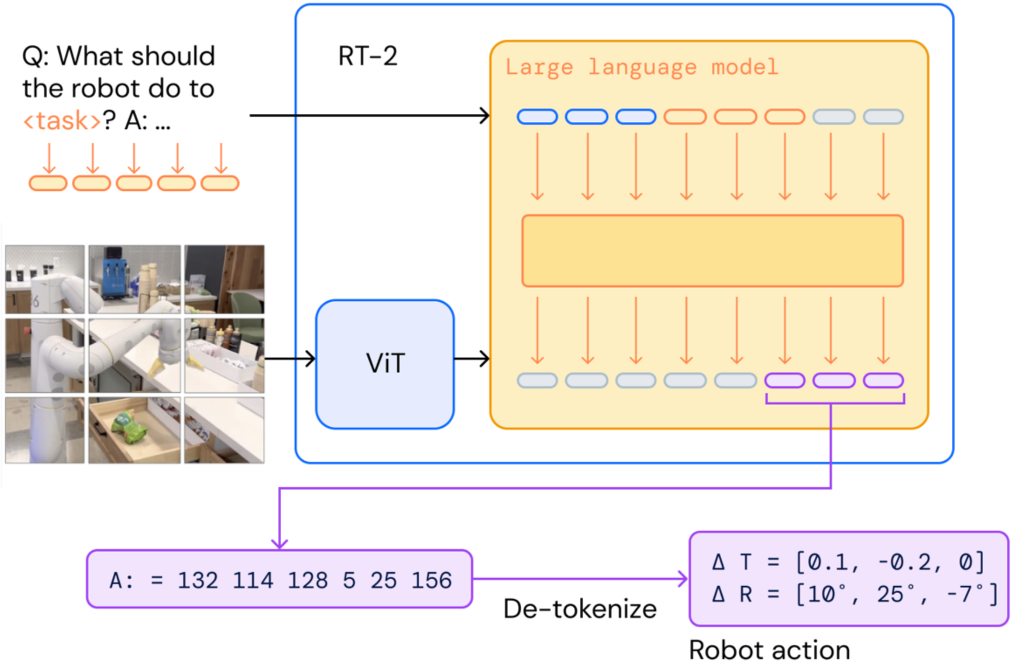
Slides adapted from EECS227 by Shuran Song

Latent State Representation in Policies



Diffusion Policy

[Chi et al., RSS 2023]



VLA Models

[Brohan et al., CoRL 2023]

Latent State Representation in Policies

$\pi_{0.5}$

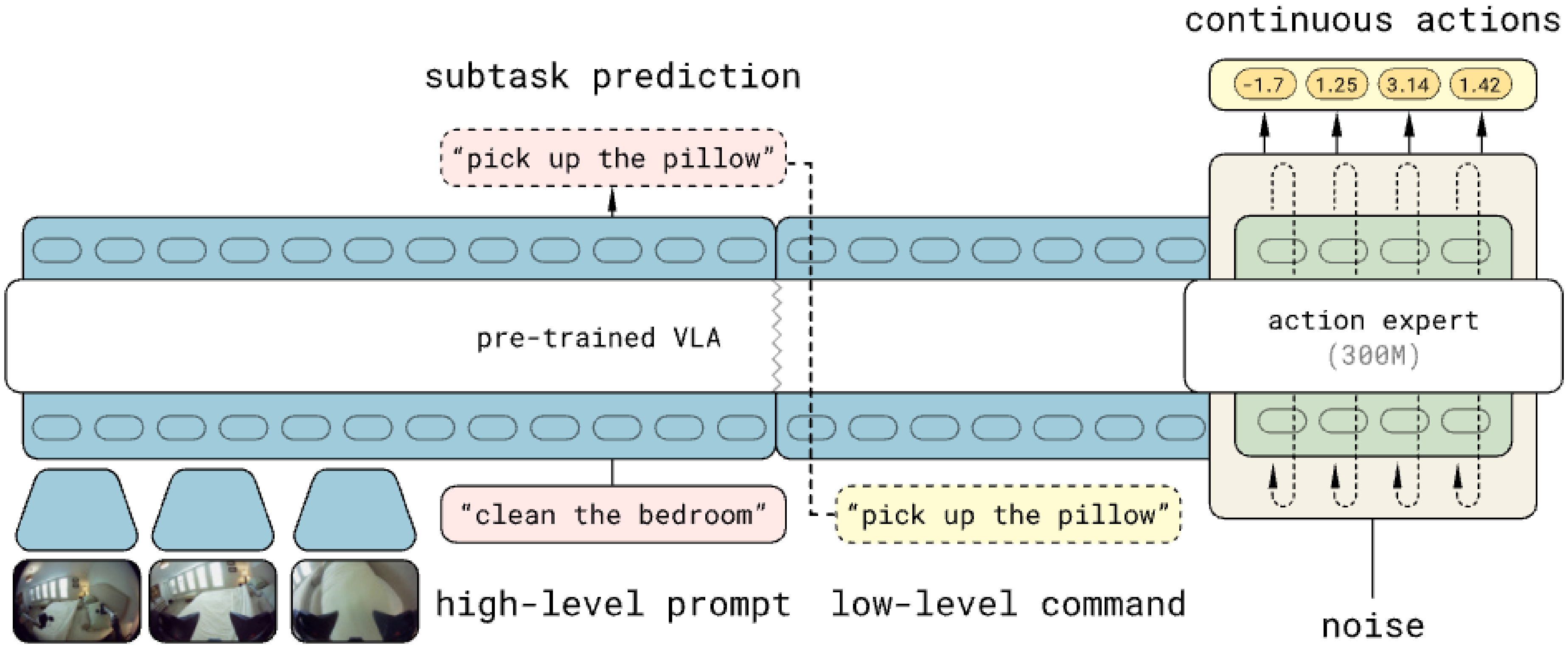
[Physical Intelligence, 2025]

Research Frontier

- Perceptual representations for end-to-end policies
- Multimodal perception (Perception beyond vision)
  - Touch, Audio, etc
- Interactive and active perception
  - “We see in order to move; we move in order to see.” – William Gibson
- Perception for world modeling
  - Beyond “perceive to act”, also answer the “what if” queries based on actions

Perceptual Representations for End-to-End Policies

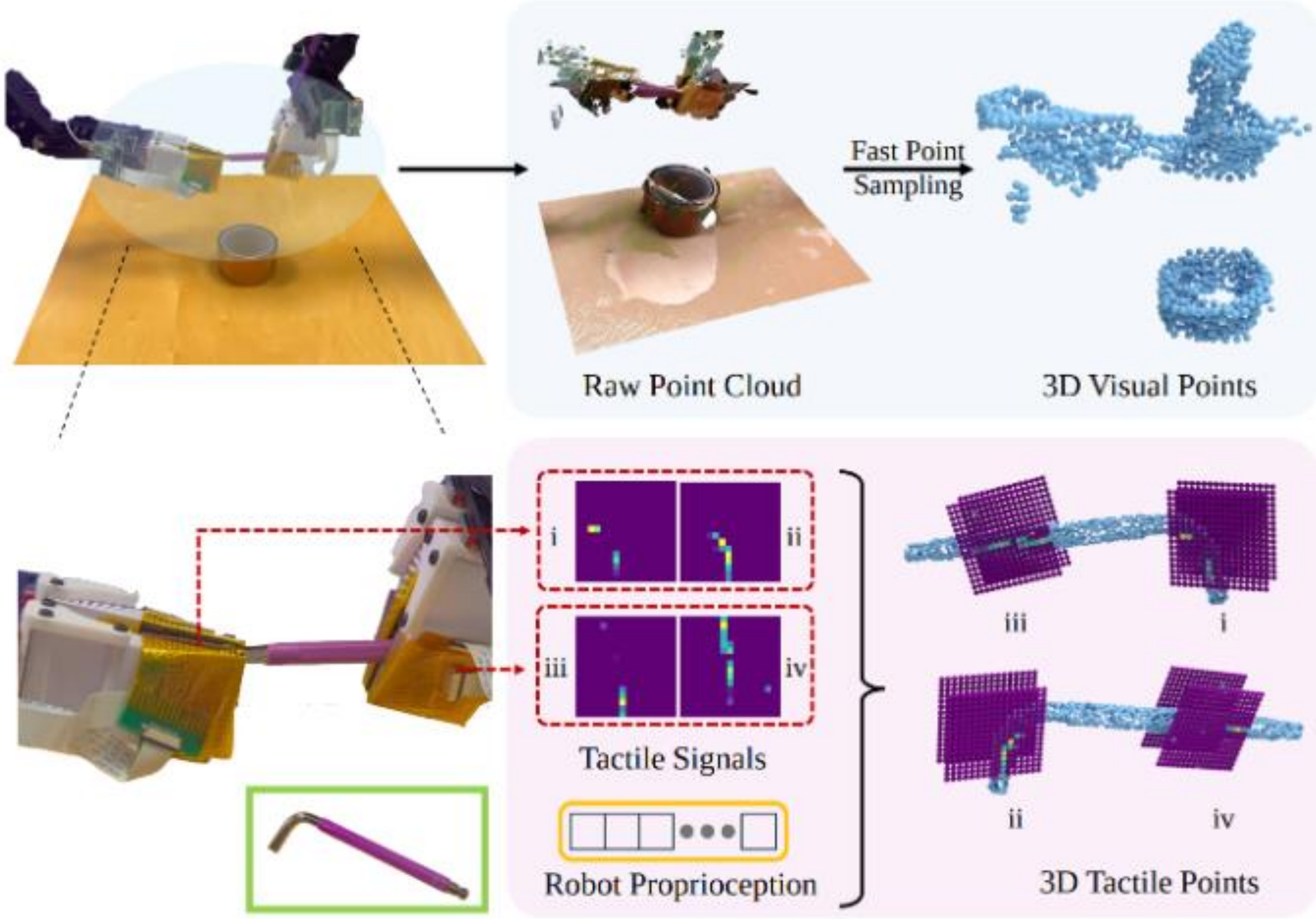
- What should be the pre-trained representation?
- What properties should it have?
- Is vision-language encoder the best choice?



pi-0.5, Physical Intelligence

Multimodal Perception

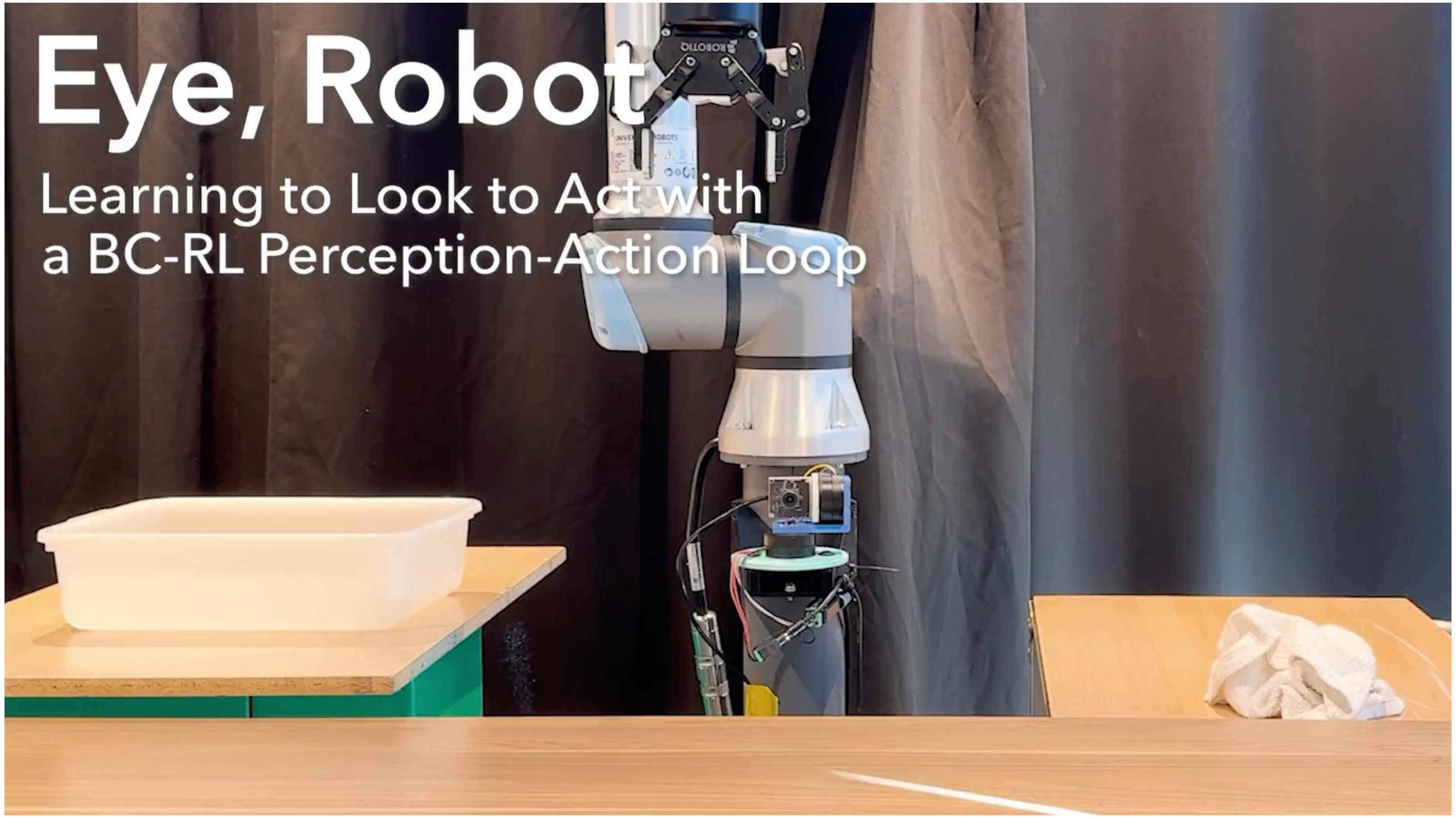
- What is the role of modalities other than vision?
- How to effectively perform sensor fusion?



3D-ViTac. Huang et al., CoRL 2024

Interactive and Active Perception

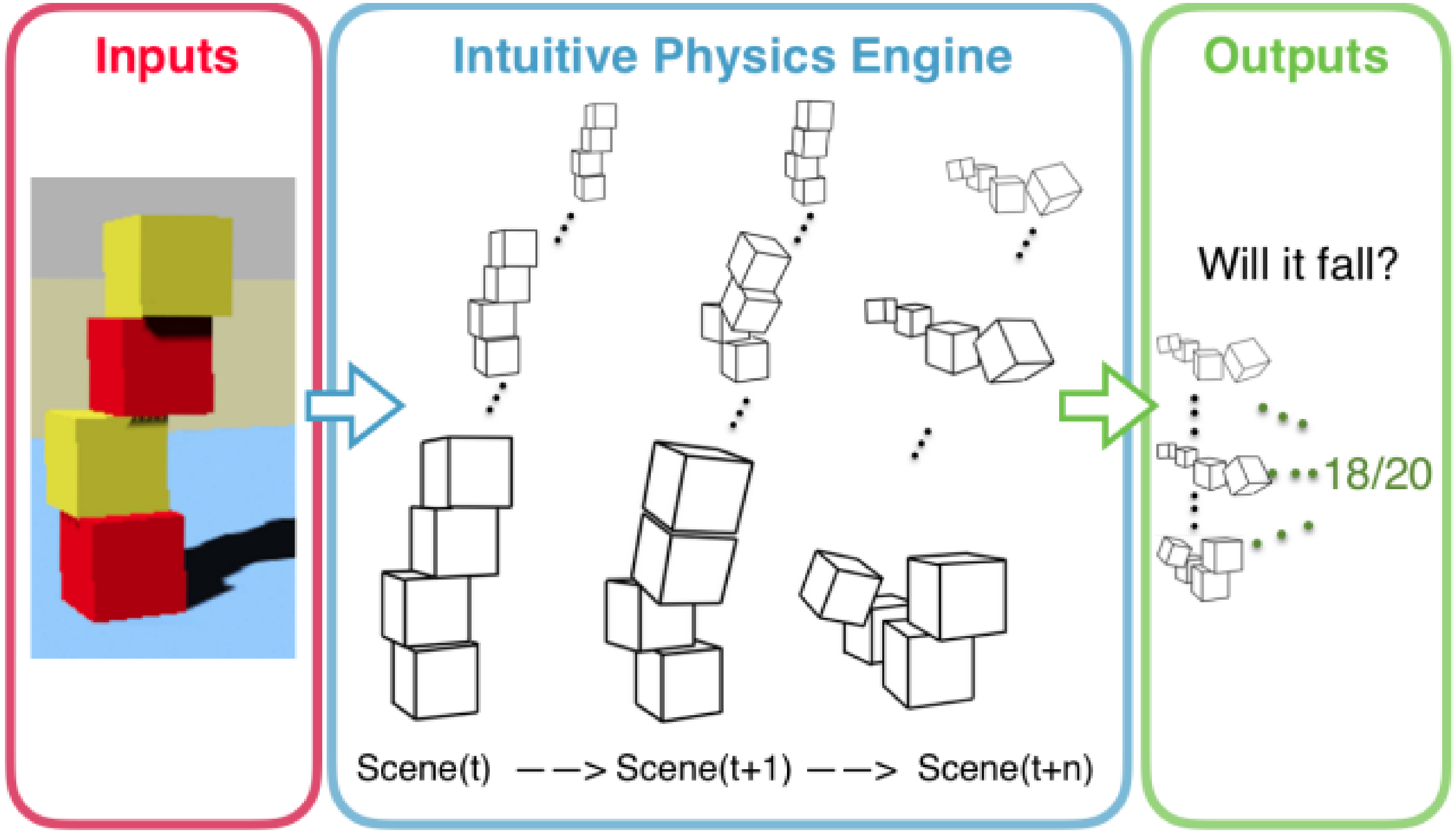
- Humans perceive actively to act with eye gaze, but robots often have fixed cameras.
- How should robots perceive to act and act to perceive?



Eye, Robot. Kerr et al., CoRL 2025

Perception for World Modeling

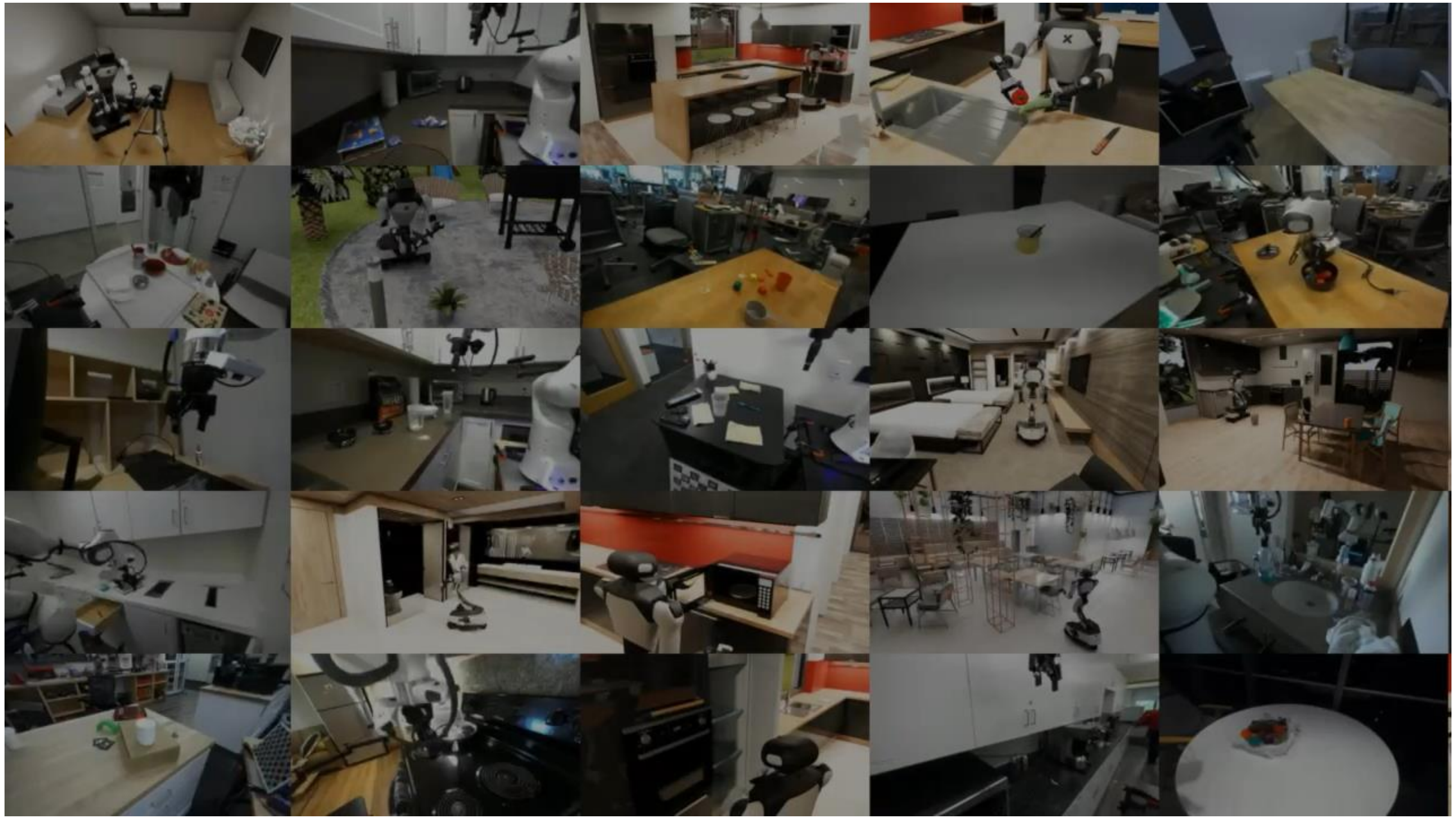
- Humans possess “intuitive physics” commonsense that supports counterfactual reasoning of “what if” queries
- Critical for purposeful manipulation (System 2)
- How can robots achieve the same?



Zhang et al., CogSci 2016

Perception for World Modeling

- Humans possess “intuitive physics” commonsense that supports counterfactual reasoning of “what if” queries
- Critical for purposeful manipulation (System 2)
- How can robots achieve the same?



Huang et al., 2026