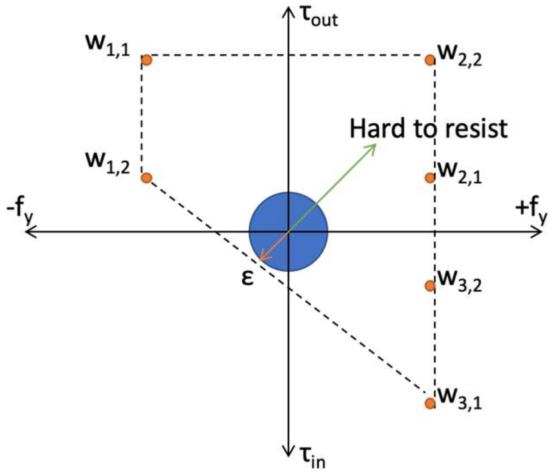
# Principles of Robot Autonomy II

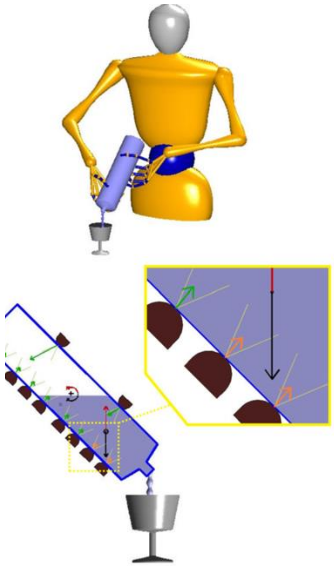
Learning-based Approaches to **Manipulation** & **Interactive Perception**

Jeannette Bohg

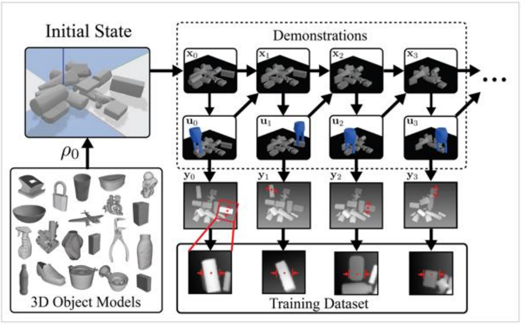
Stanford University

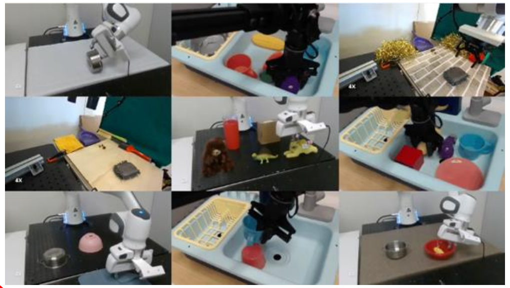
# Learning Outcome for next five Lectures

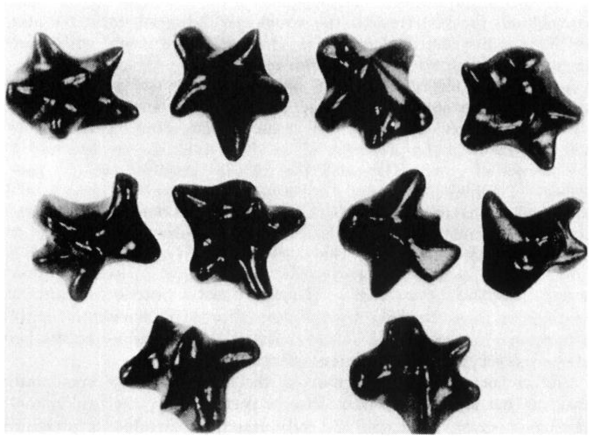


Modeling and Evaluating Grasping and Manipulation



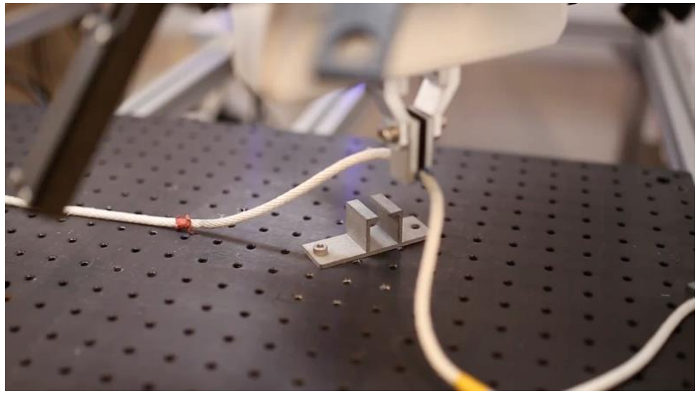



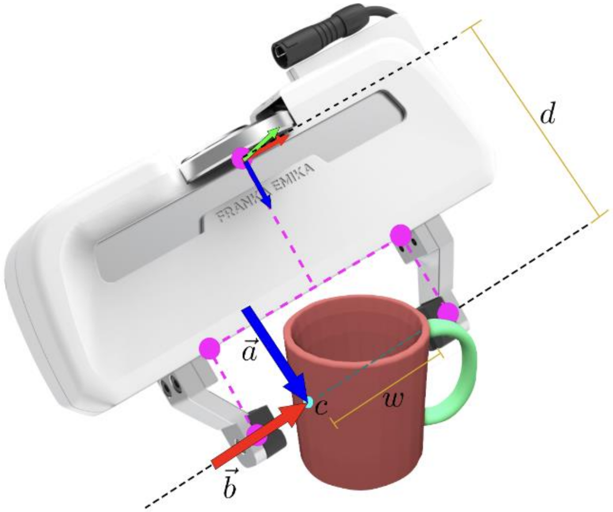


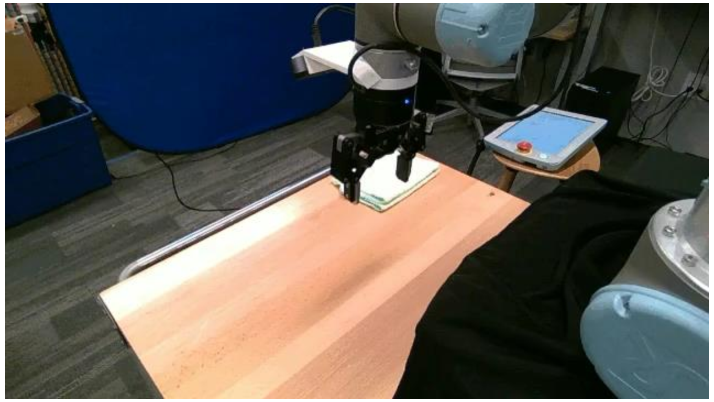
Use Manipulation to Perceive better

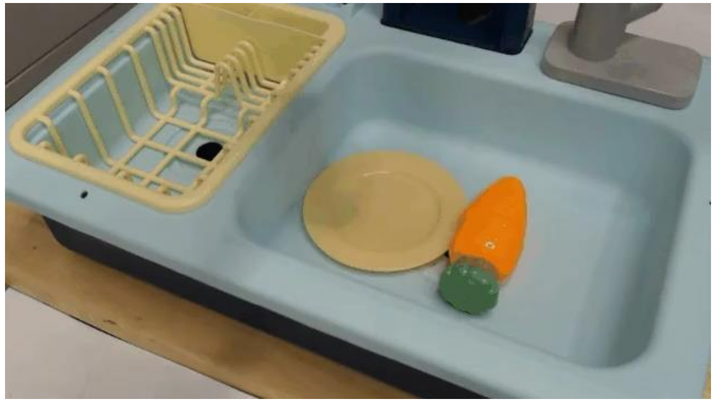
Learning-based Grasping and Manipulation

# Learning-Based Grasping vs Manipulation

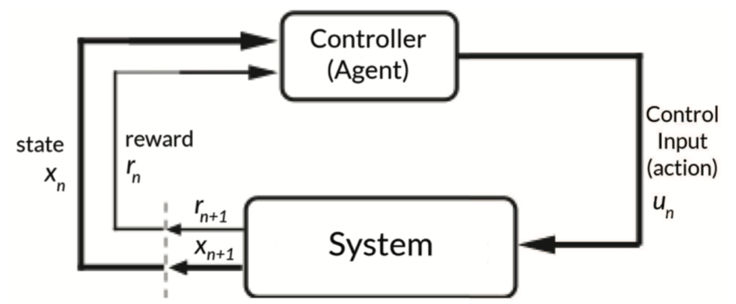




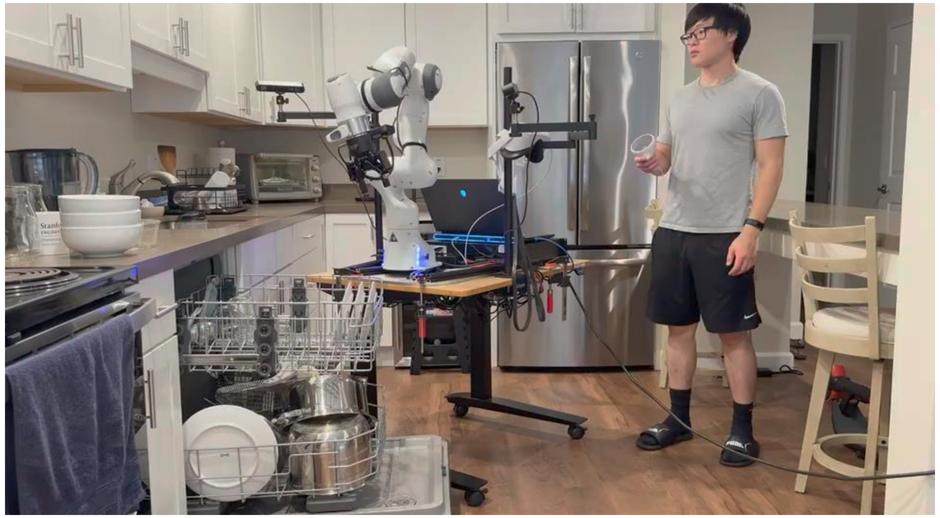



# Main Technical Approaches to Learning Manipulation

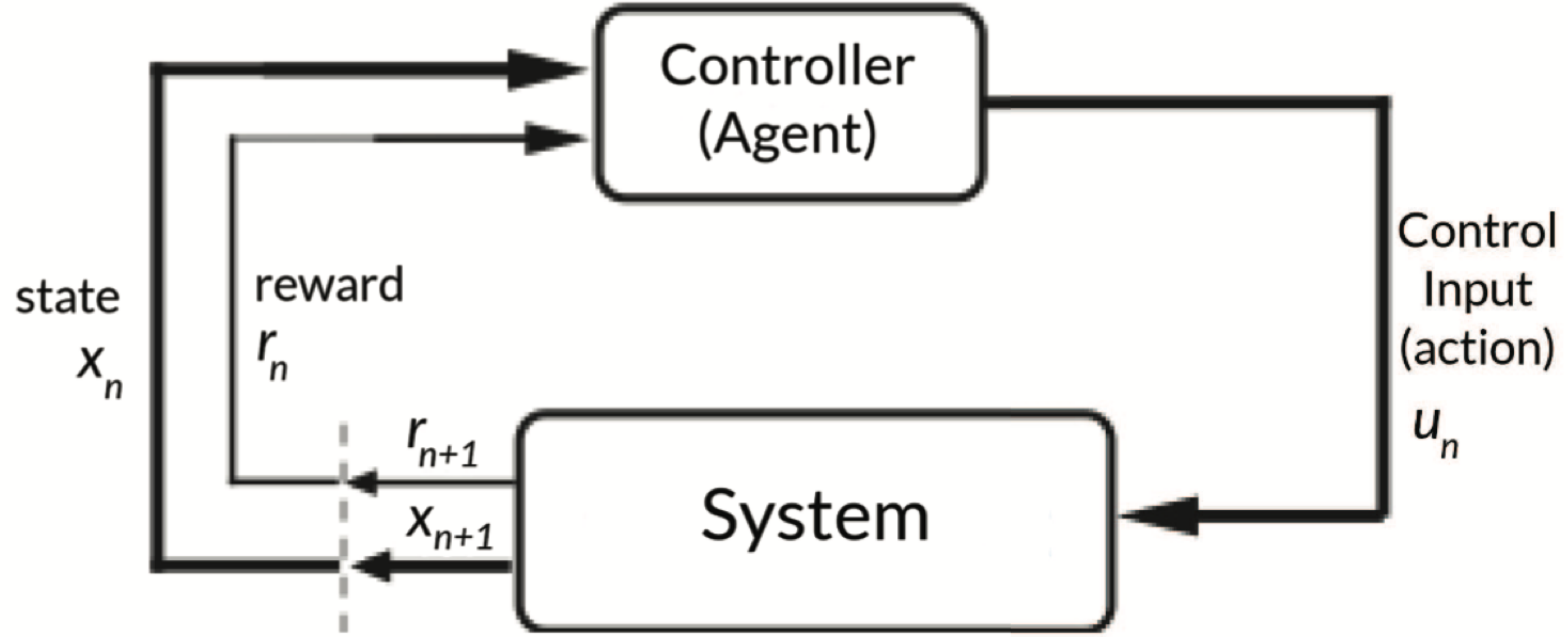



Reinforcement Learning

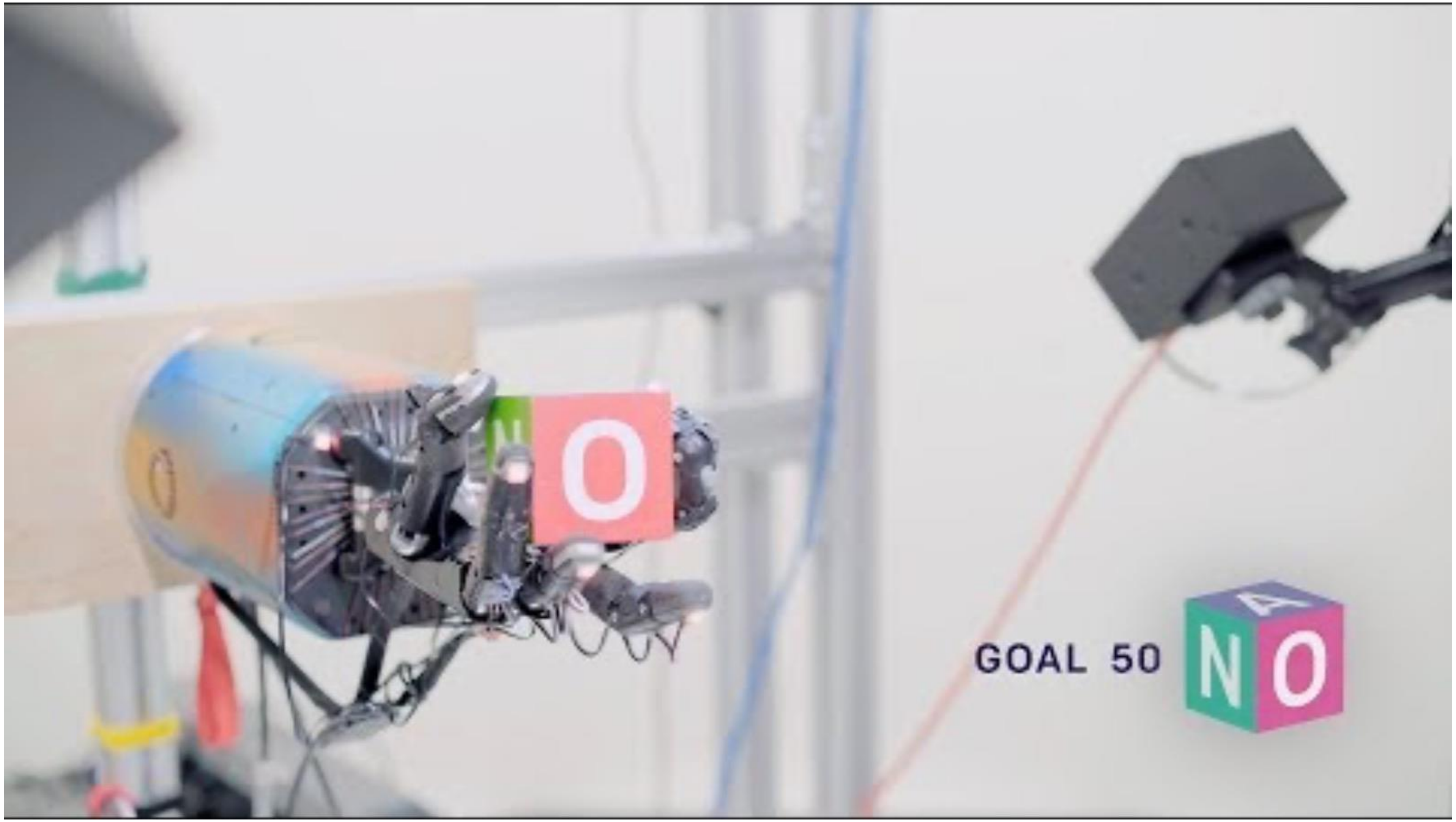


Imitation Learning

# Reinforcement Learning for Manipulation



# Learning Dexterous In-hand Manipulation OpenAI, IJRR, 2020.



# What stuck: Domain Randomization

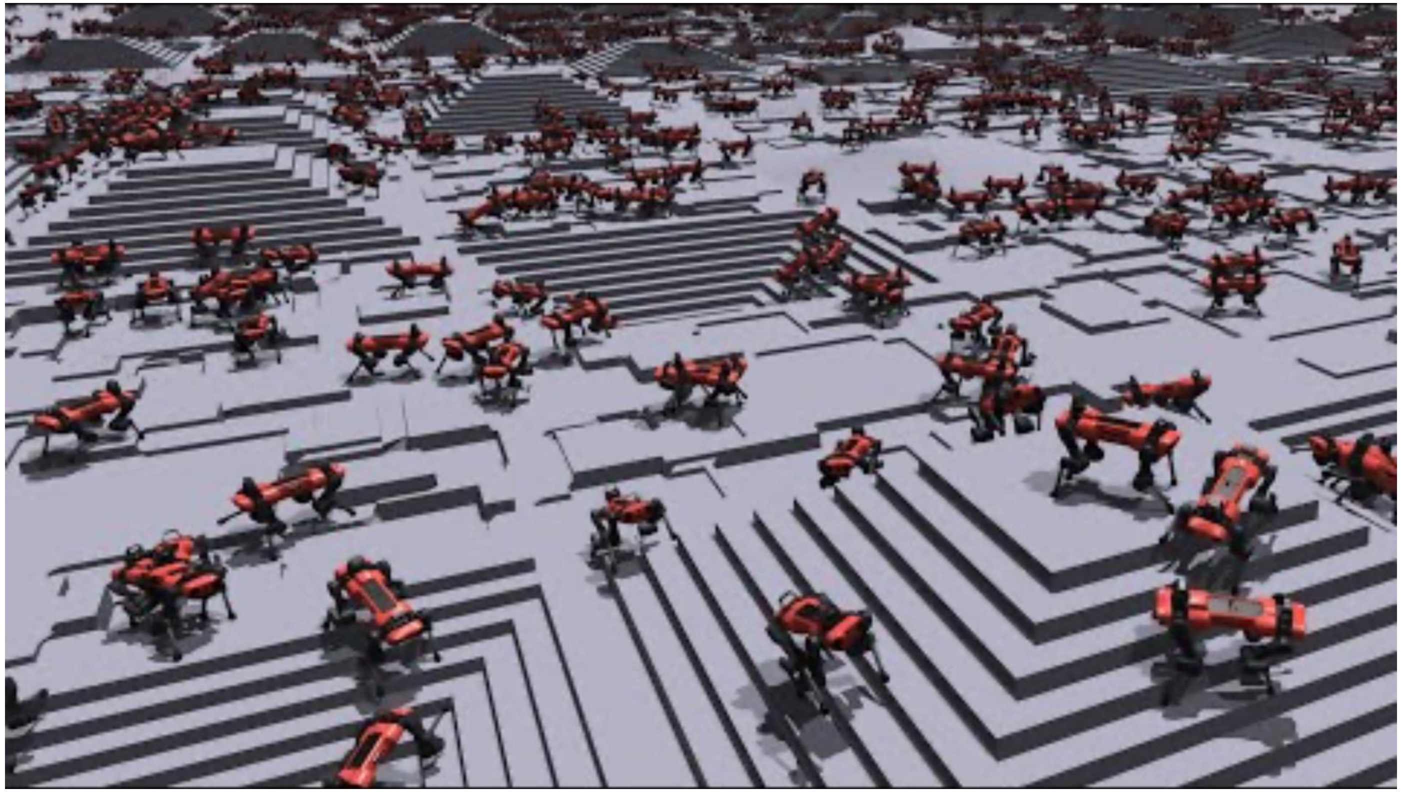
Table 1: Ranges of physics parameter randomizations.

| Parameter | Scaling factor range | Additive term range |
|---|---|---|
| object dimensions | uniform([0.95, 1.05]) | |
| object and robot link masses | uniform([0.5, 1.5]) | |
| surface friction coefficients | uniform([0.7, 1.3]) | |
| robot joint damping coefficients | loguniform([0.3, 3.0]) | |
| actuator force gains (P term) | loguniform([0.75, 1.5]) | |
| joint limits | | $\mathcal{N}(0, 0.15)$ rad |
| gravity vector (each coordinate) | | $\mathcal{N}(0, 0.4)$ $\mathrm{m/s^2}$ |

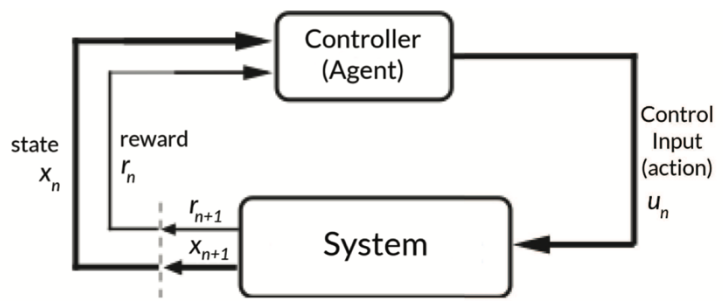
Figure 4: Simulations with different randomized visual appearances. Rows correspond to the renderings from the same camera, and columns correspond to renderings from 3 separate cameras which are simultaneously fed into the neural network.

# The impact on Locomotion

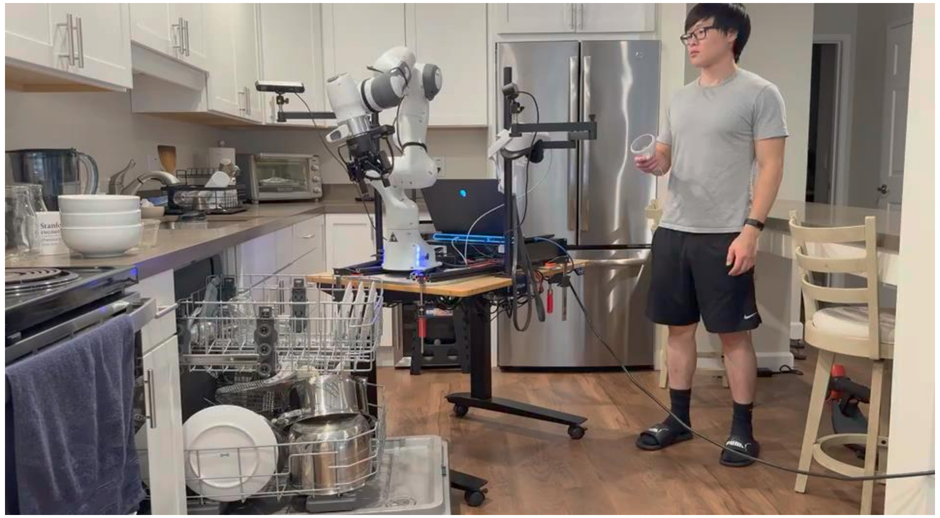
- Successful Sim2Real RL in quadruped and more recently humanoid locomotion
- Learning to walk in minutes using Massively Parallel Deep RL. Rudin et al. CoRL'21



# Main Technical Approaches to Learning Manipulation



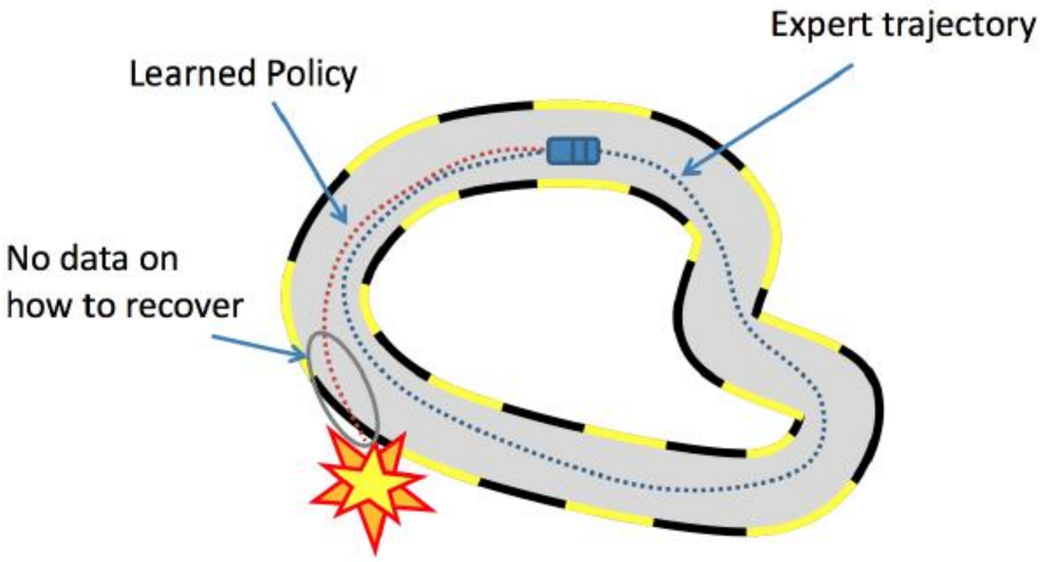

Reinforcement Learning



Imitation Learning

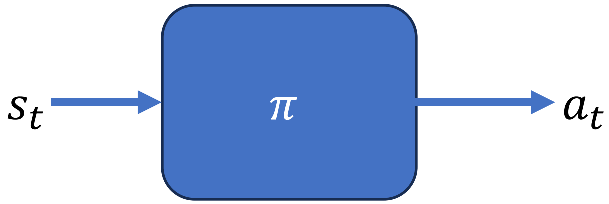
# Imitation learning

- Turning the learning problem into a supervised problem
- Estimate a policy from training examples $(s_0, a_0), (s_1, a_1)$ $(s_2, a_2)$

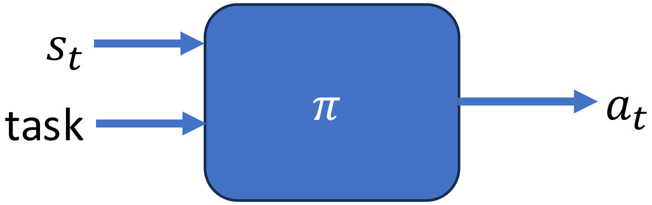



From CS237: Reinforcement Learning

# Two flavors of imitation learning

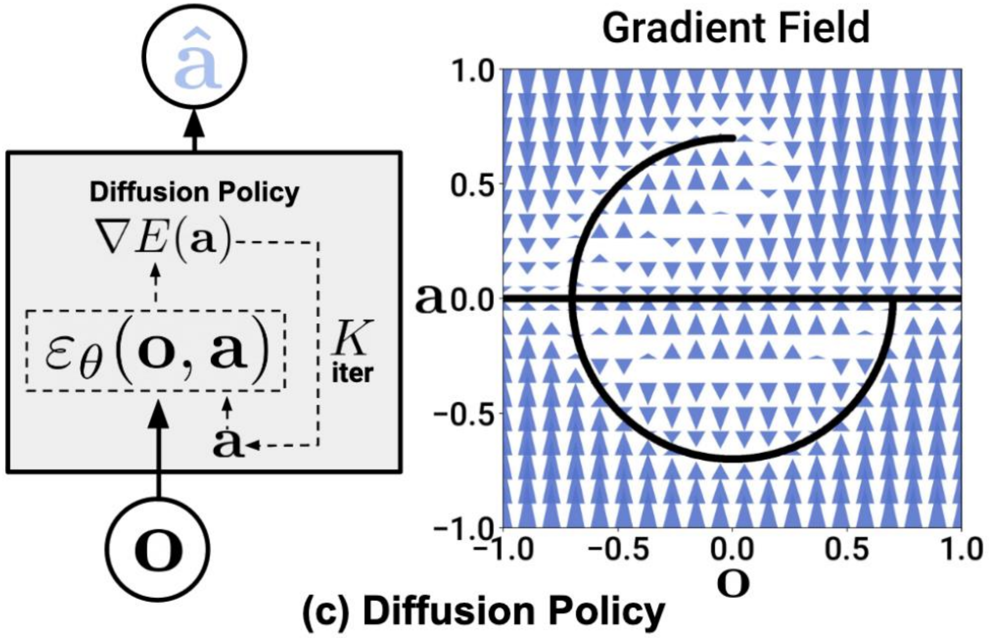



Single task policy

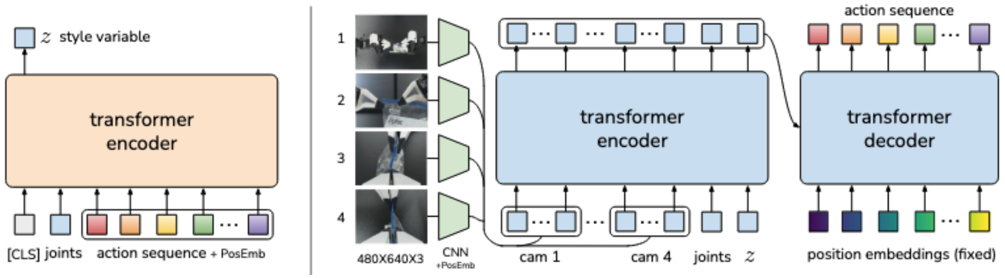



Multi-task (Generalist) policy

# Single-task policies

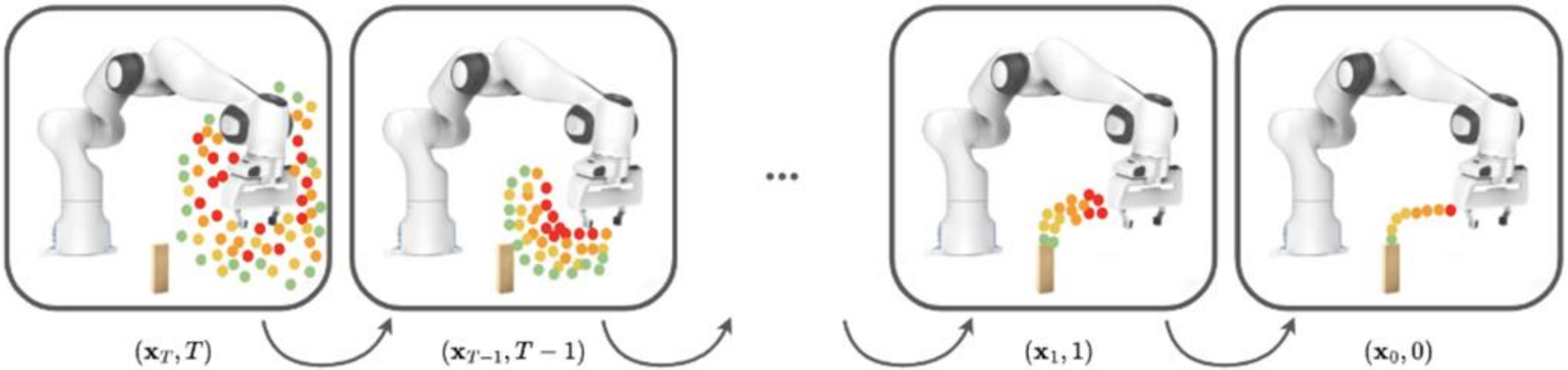



Diffusion Policies



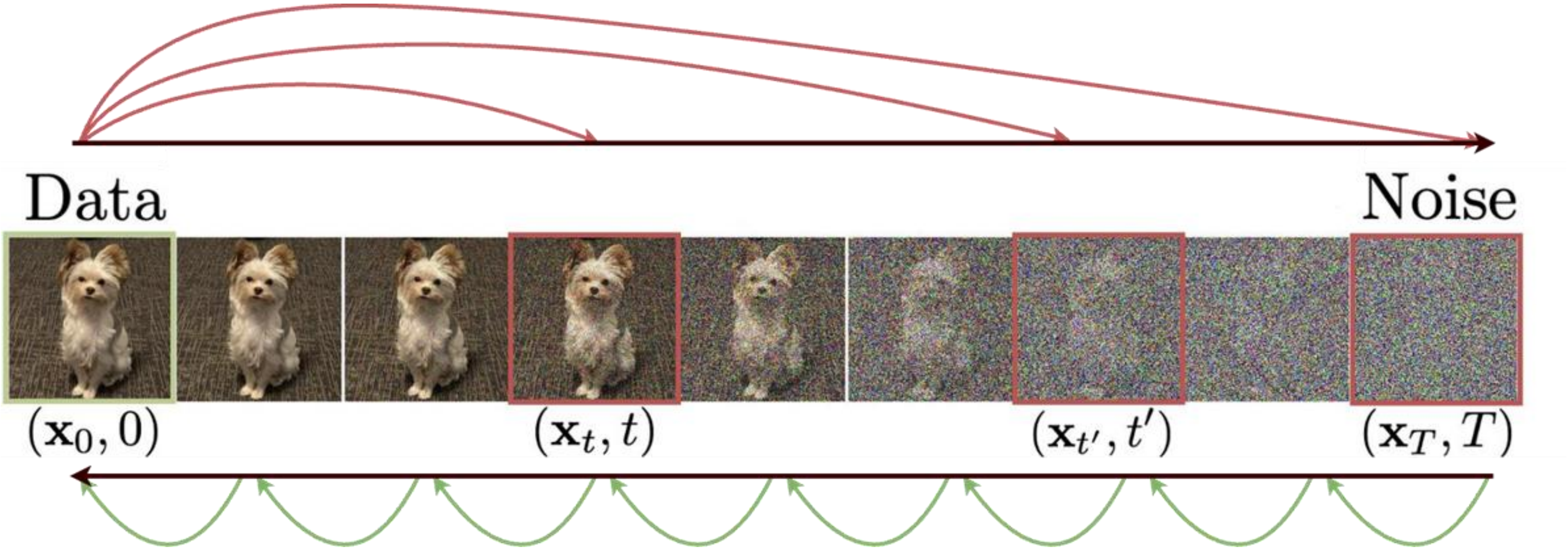

Action Chunking with Transformers

# Diffusion policy



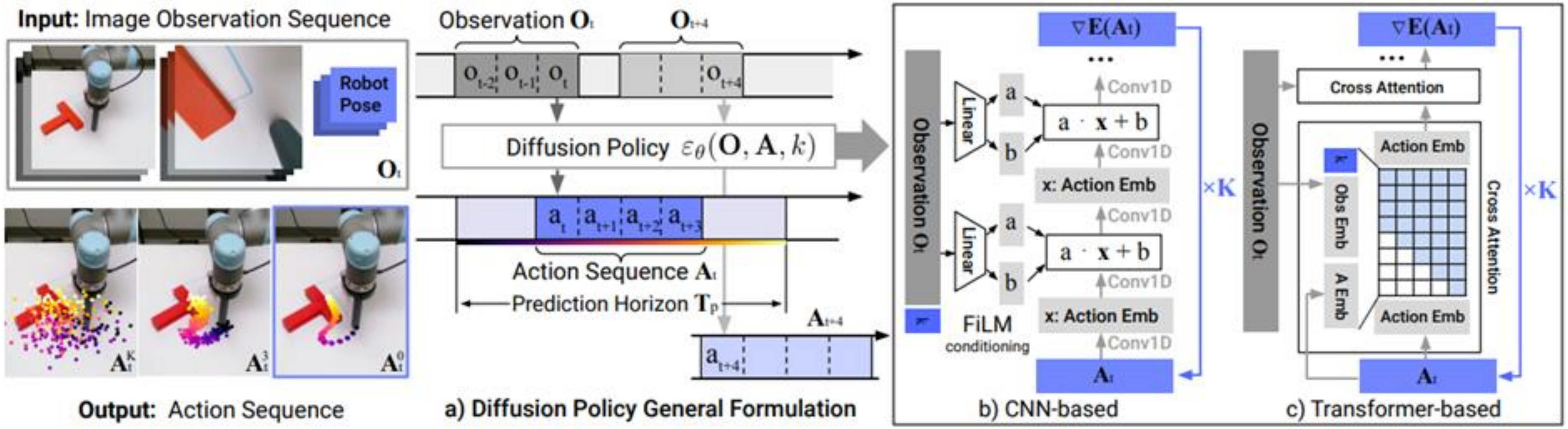

Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. Chi et al. RSS'23

# Idea behind Diffusion from Computer Vision



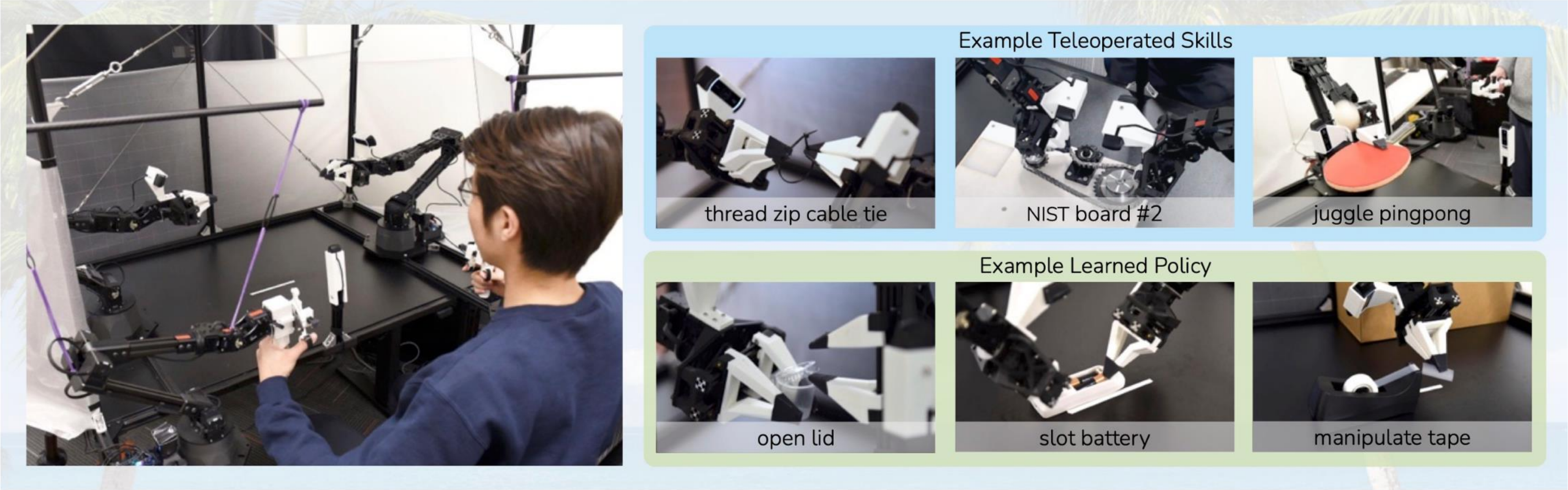
# Diffusion Policy

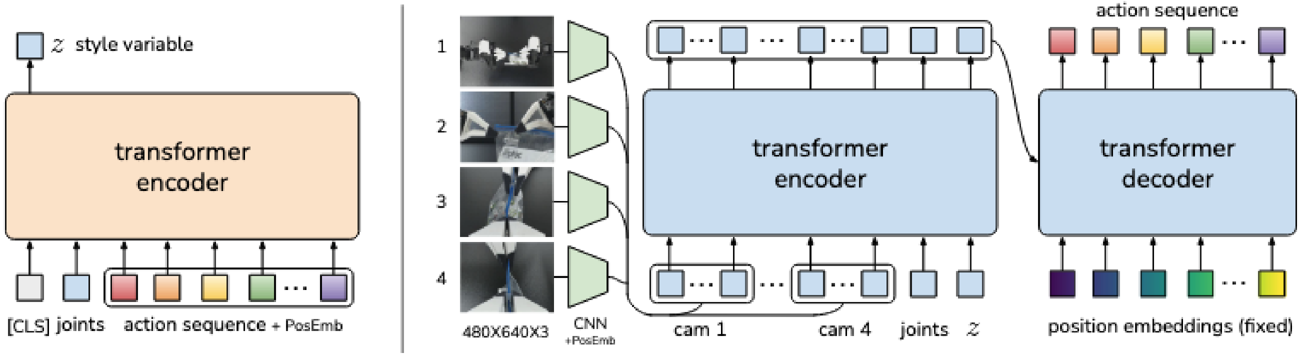
1. Diffuse through actions, condition on observations
2. Chunk observations and actions for smoother control
3. Pro: SOTA success rates, continuous action space, multimodality
4. Con: Inference takes a lot of time b/c of denoising process

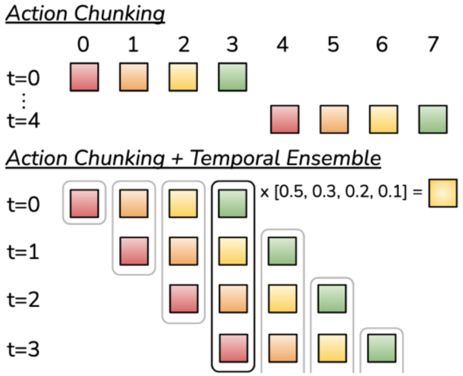



Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. Chi et al. RSS’23

# Action Chunking Transformer (ACT)

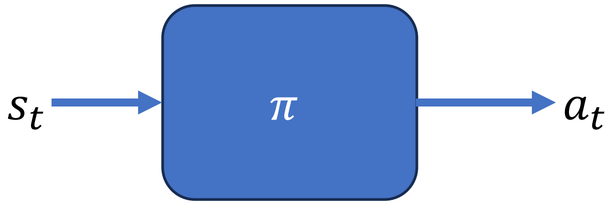


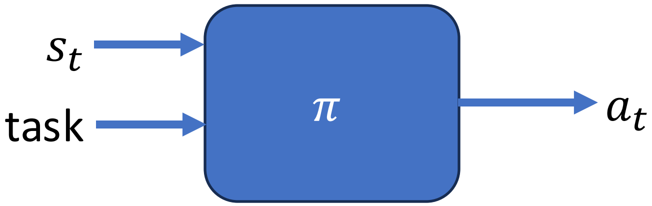




Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware. Zhao et al. RSS'23

# Two flavors of imitation learning

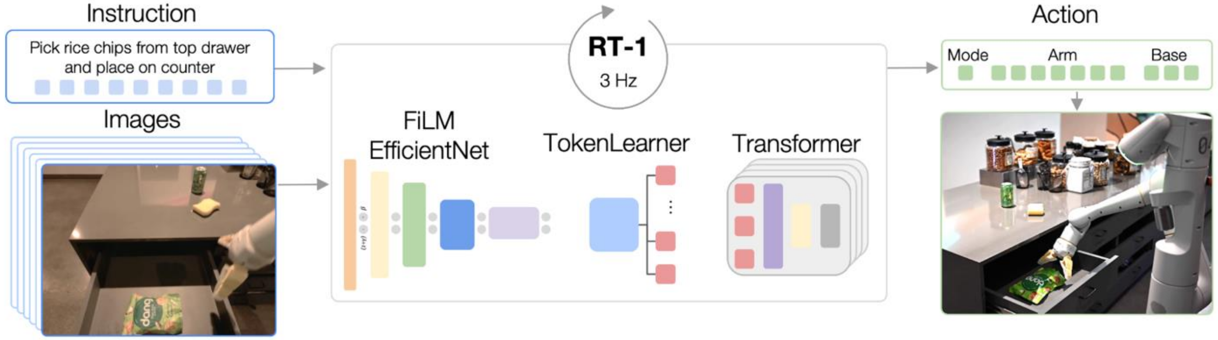



Single task policy

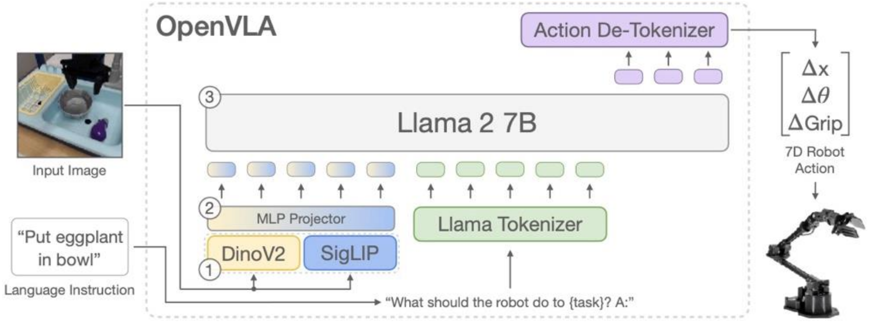



Multi-task (Generalist) policy

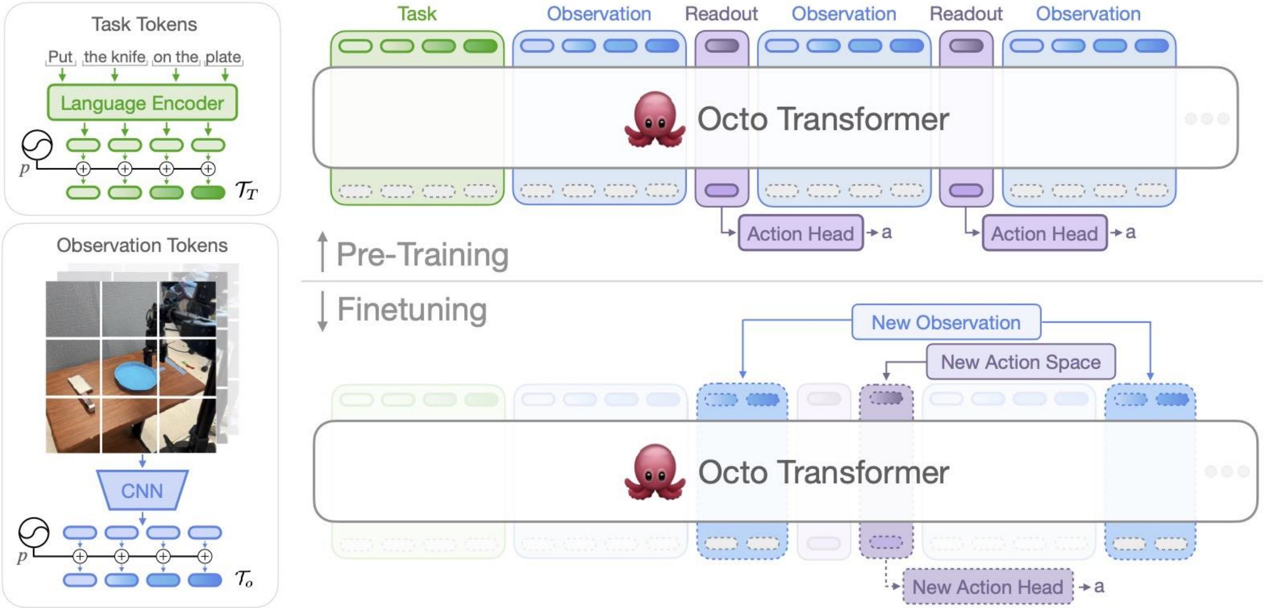
# Multi-task (generalist policies)



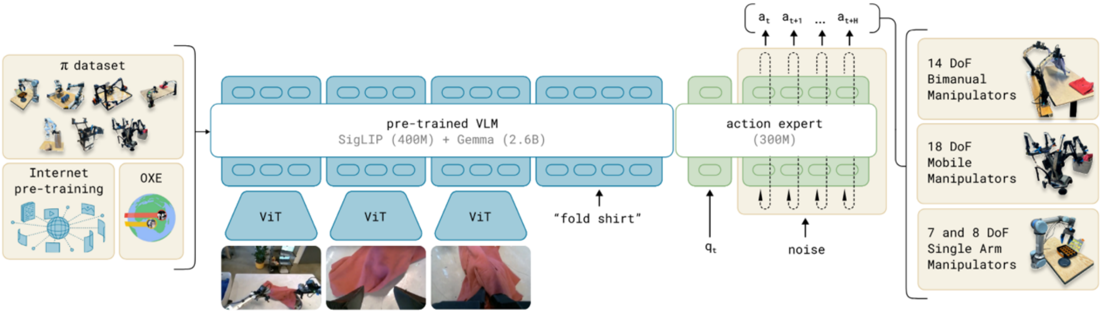

RT-1/2/X




OpenVLA

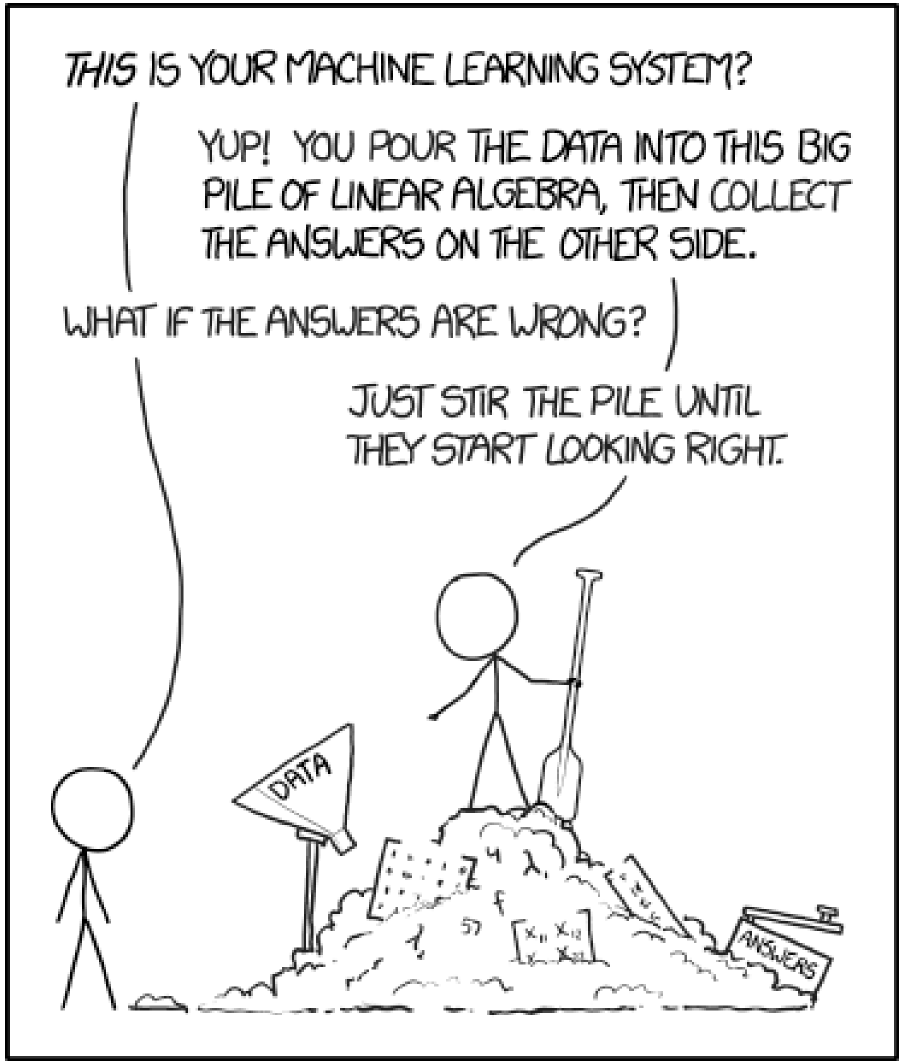



Octo

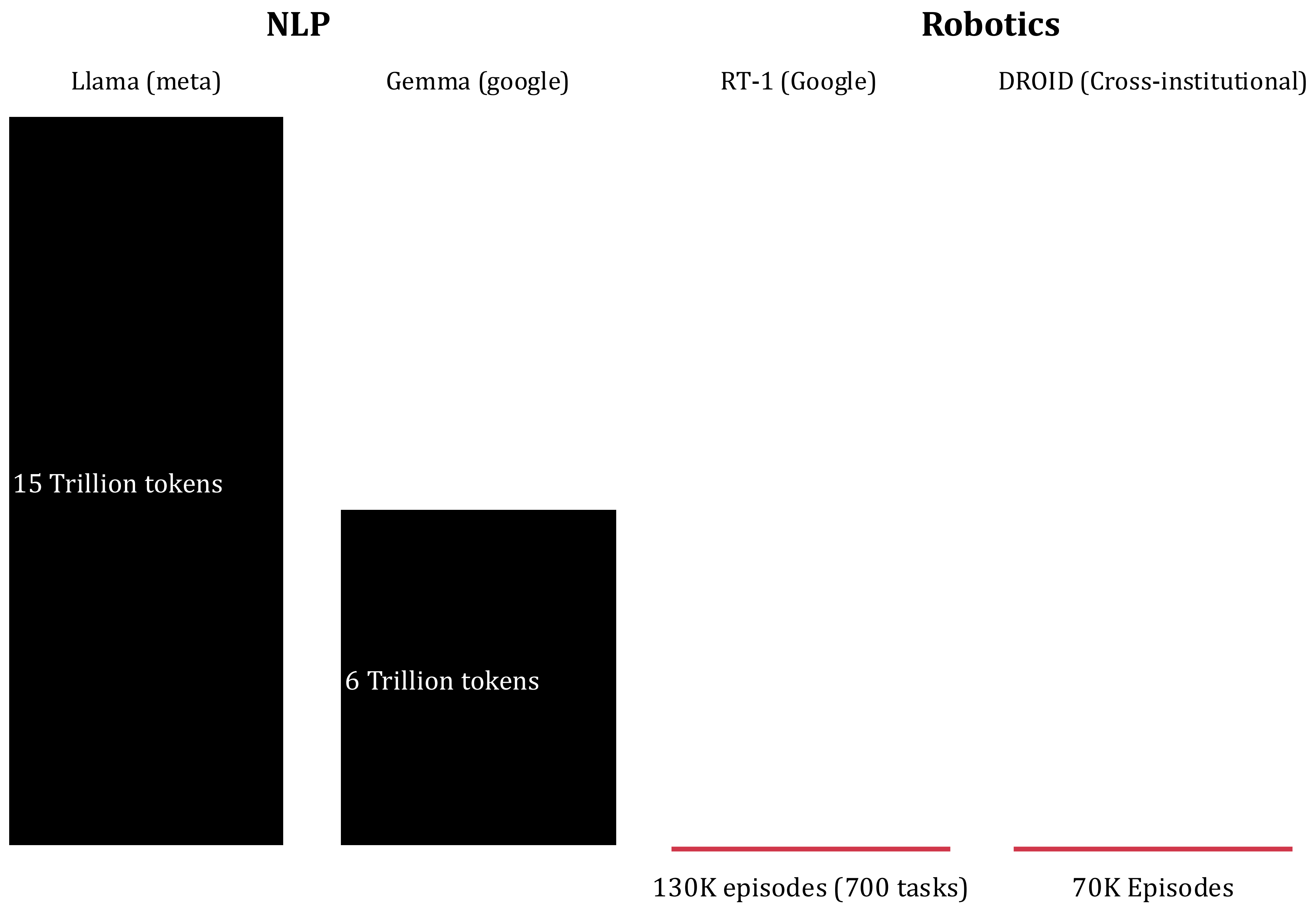


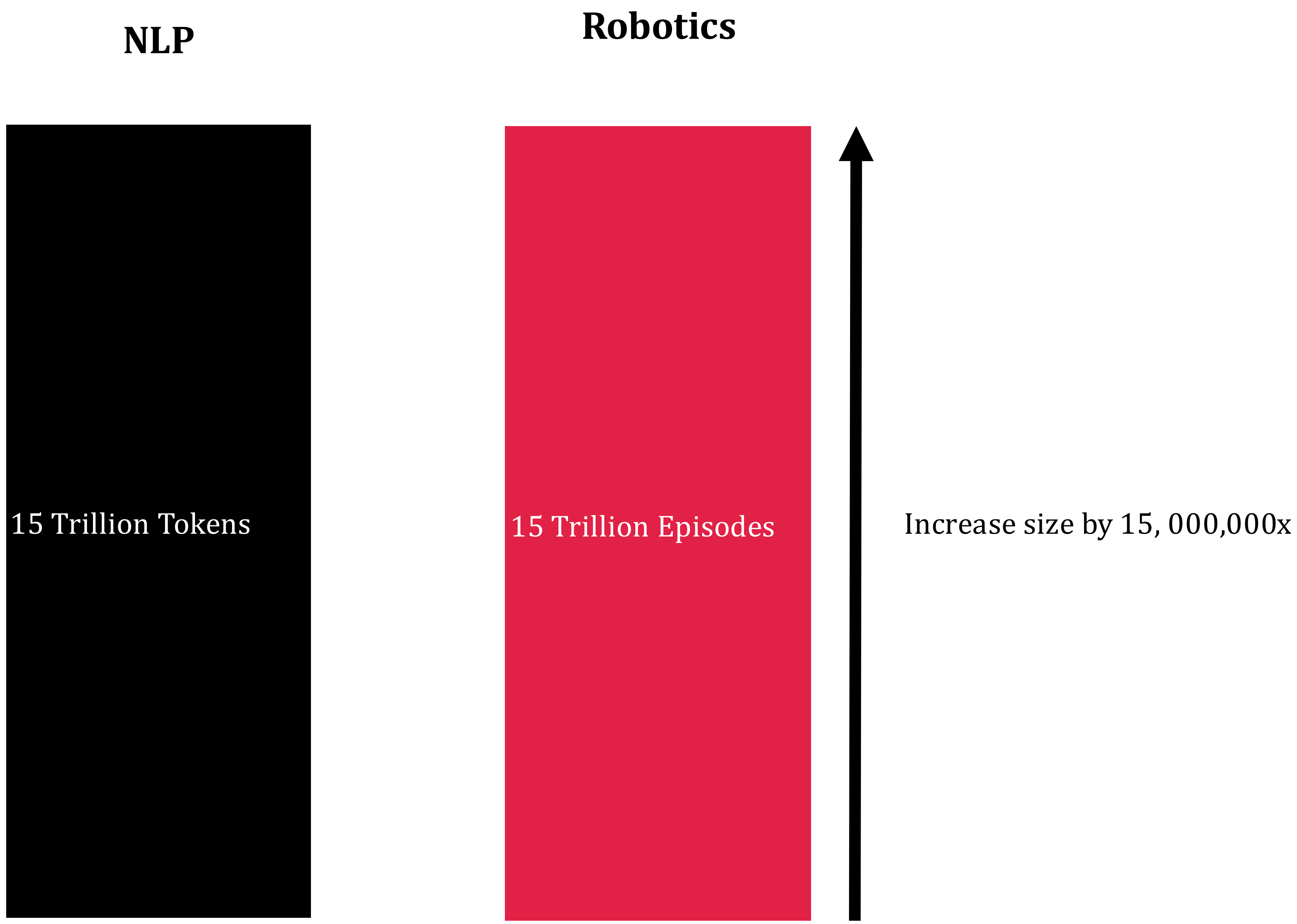

Pi0

# How much data do these policies need?




https://imgs.xkcd.com/comics/machine_learning.png


NLP
Robotics
Llama (meta)
Gemma (google)
RT-1 (Google)
DROID (Cross-institutional)
15 Trillion tokens
6 Trillion tokens
130K episodes (700 tasks)
70K Episodes


NLP
Robotics
15 Trillion Tokens
15 Trillion Episodes
Increase size by 15, 000,000x

Bathroom

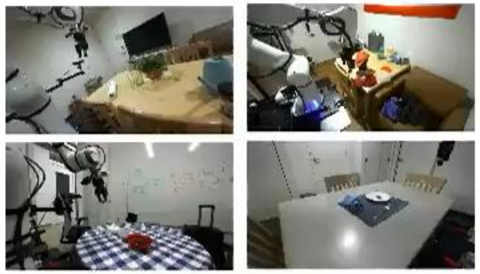
# DROID

## Distributed Robot Interaction Dataset

- 76k Episodes
- 564 Scenes
- 52 Buildings
- 13 Institutions
- 86 Tasks / Verbs

Kitchen

Dining Room

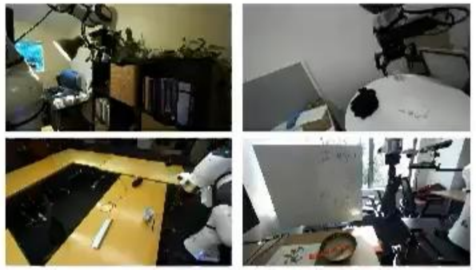


Bedroom

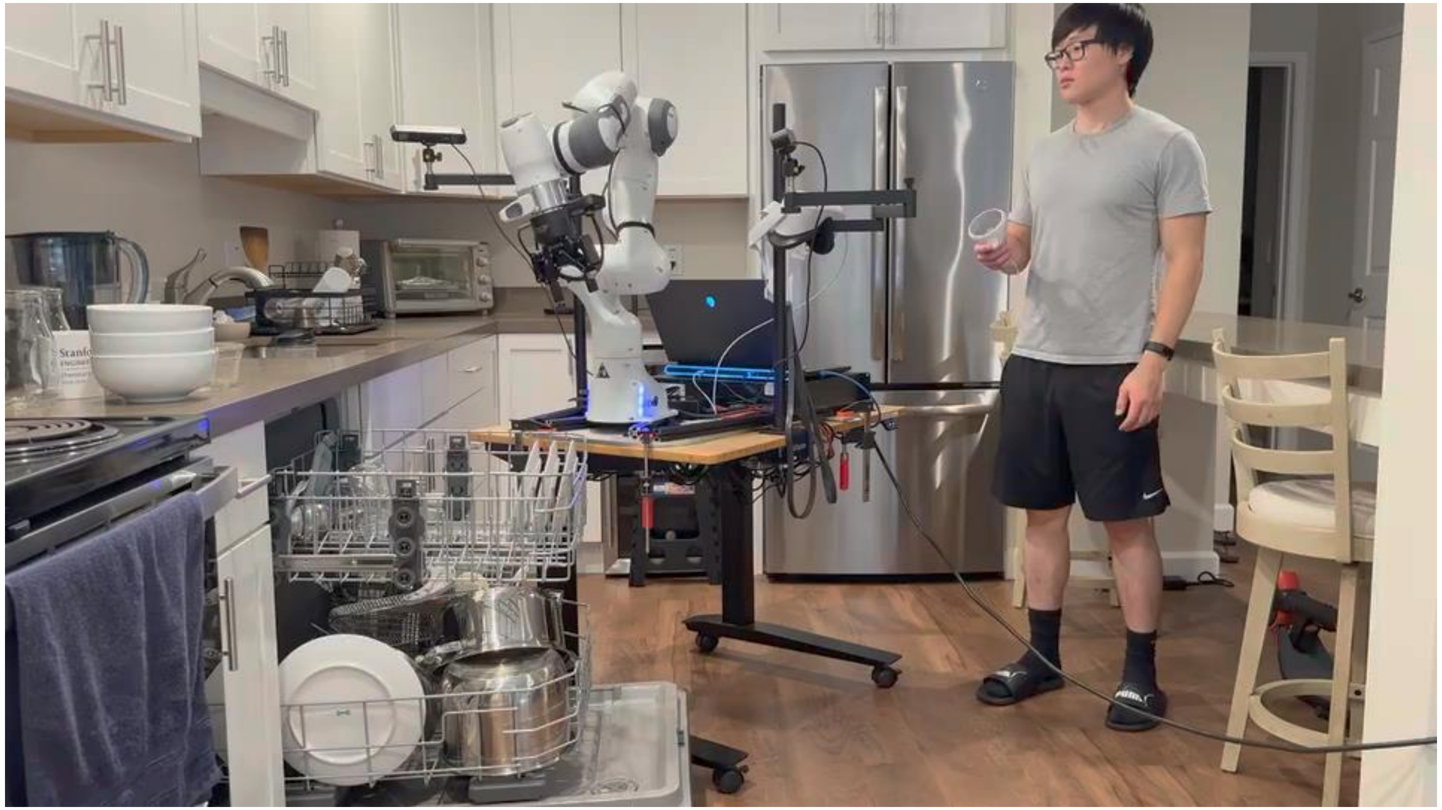
Laboratory

Laundry Room

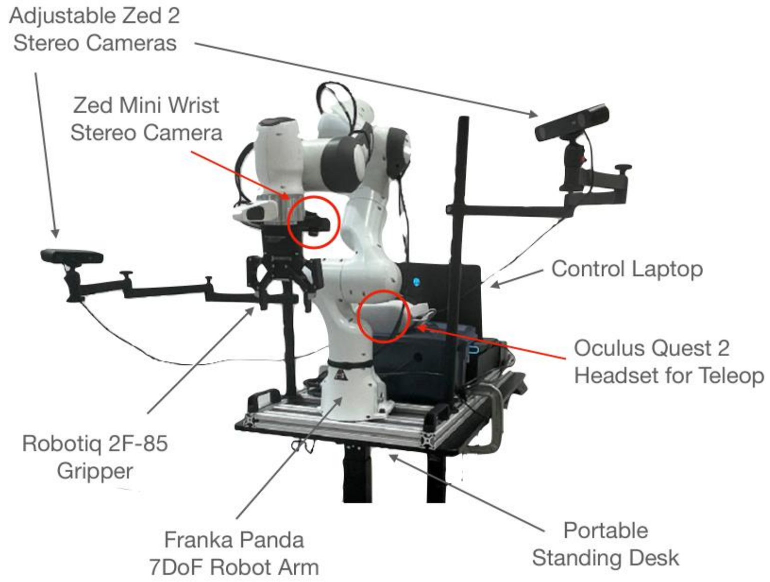
Office



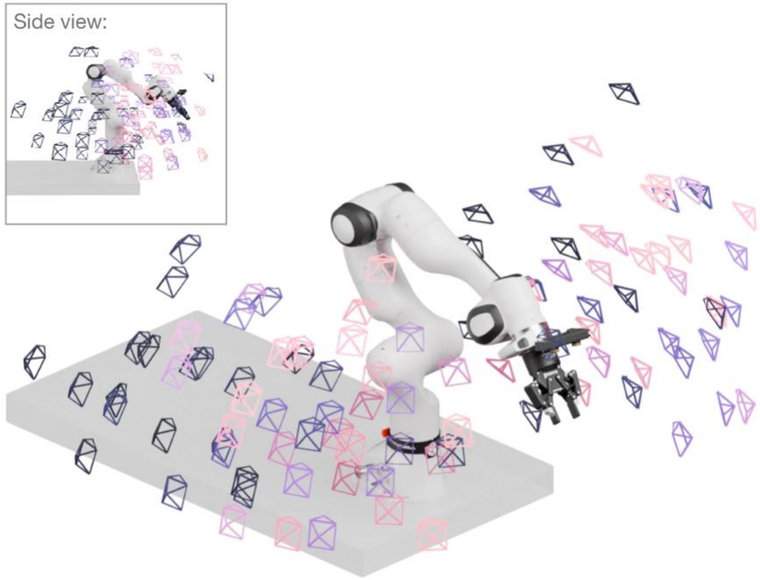
# Data through Teleoperation in the Wild

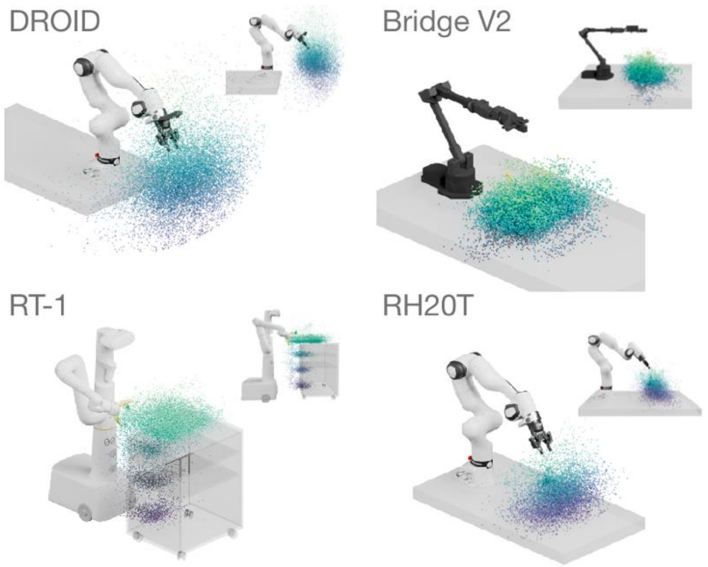


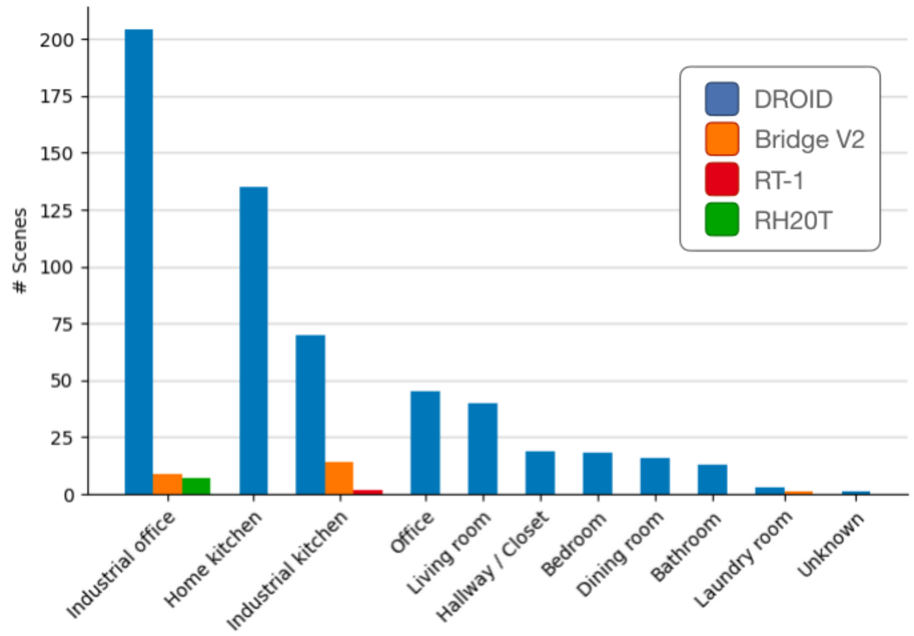
DROID: A Large-Scale In-the-Wild Robot Manipulation Dataset. Khazatsky, Pertsch et al. ‘24

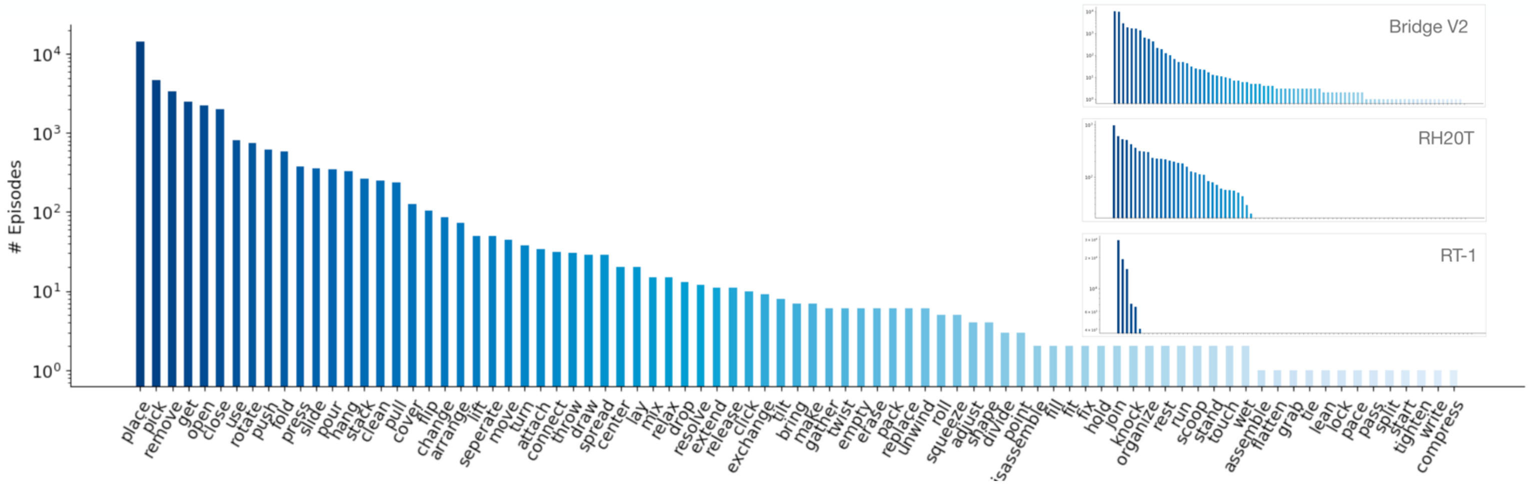


# Dataset of Unprecedented Diversity



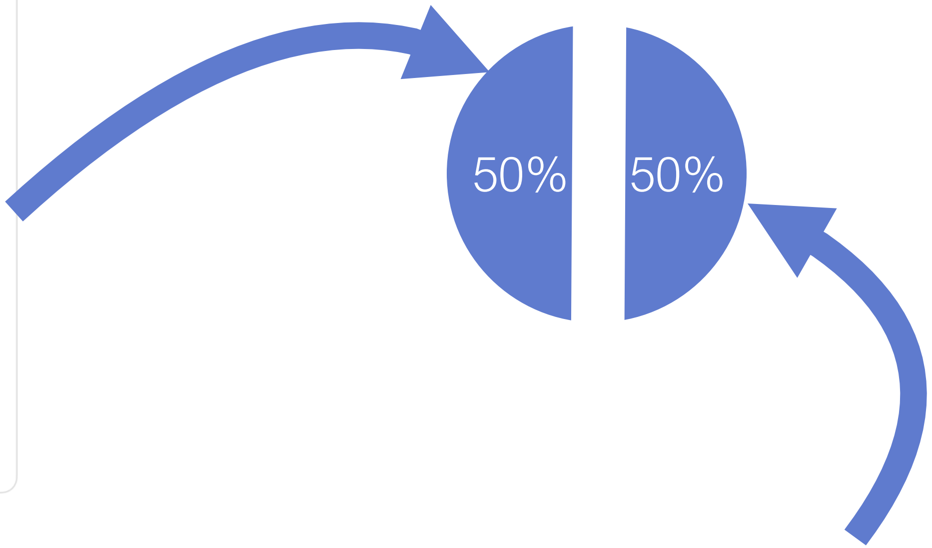
# How do you learn with this data?

## DROID

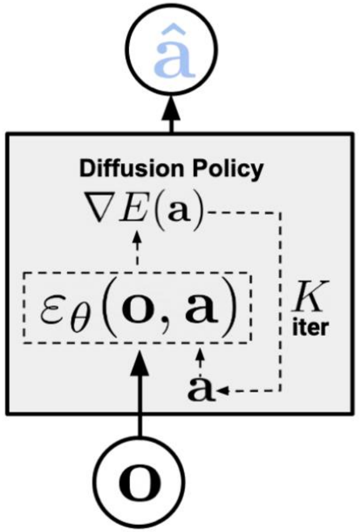
**D**istributed **R**obot **I**nteraction **D**ataset

- 🤖 **76k** Episodes
- 🖼️ **564** Scenes
- 🏠 **52** Buildings
- 🏫 **13** Institutions
- 💪 **86** Tasks / Verbs

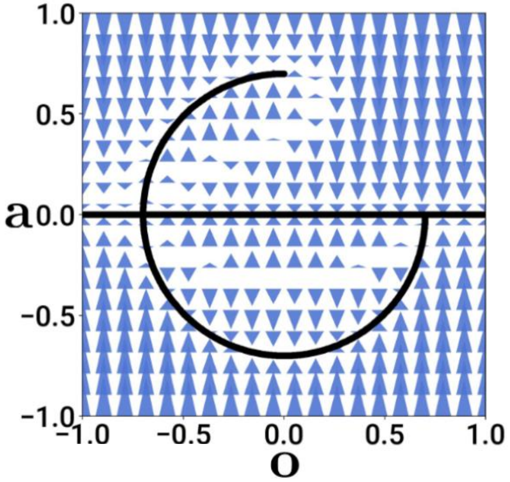
**Training Batch** → **Diffusion Policy Co-Training**




**Small Data per Testscenario**




Gradient Field




Close Waffle Maker | Place Chips on Plate | Put Apple in Pot | Toasting | Clean up Desk | Cook Lentils

Lab Evaluation | Office Evaluation | Household Evaluation

## NLP

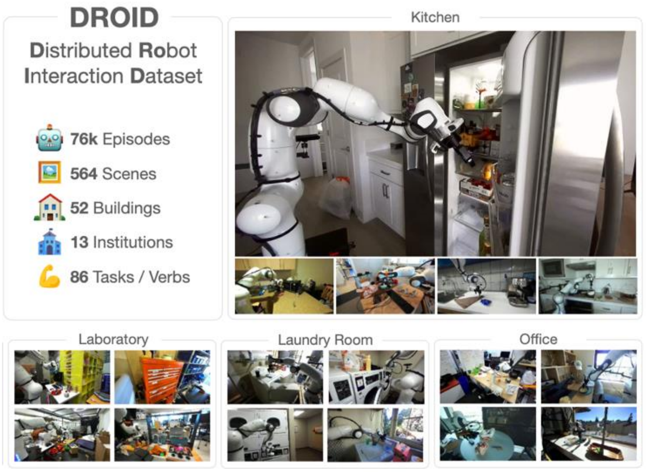
15 Trillion Tokens

## Robotics

15 Trillion Episodes

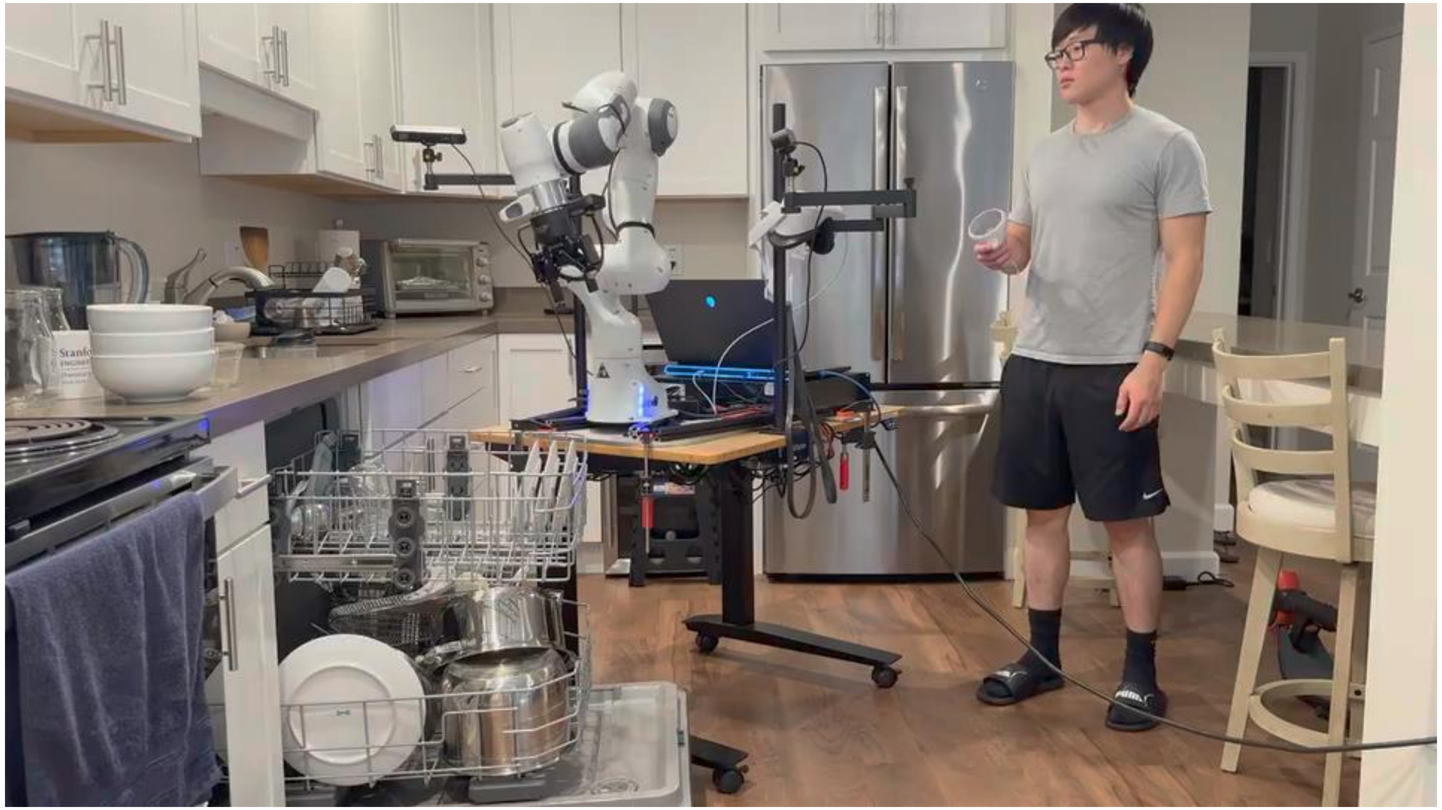
Increase size 15,000,000x

Impossible on real hardwarex




DROID: A Large-Scale In-the-Wild Robot Manipulation Dataset. Khazatsky, Pertsch et al. ‘24

# How much data can we really collect like this?



DROID: A Large-Scale In-the-Wild Robot Manipulation Dataset. Khazatsky, Pertsch et al. ‘24

## NLP

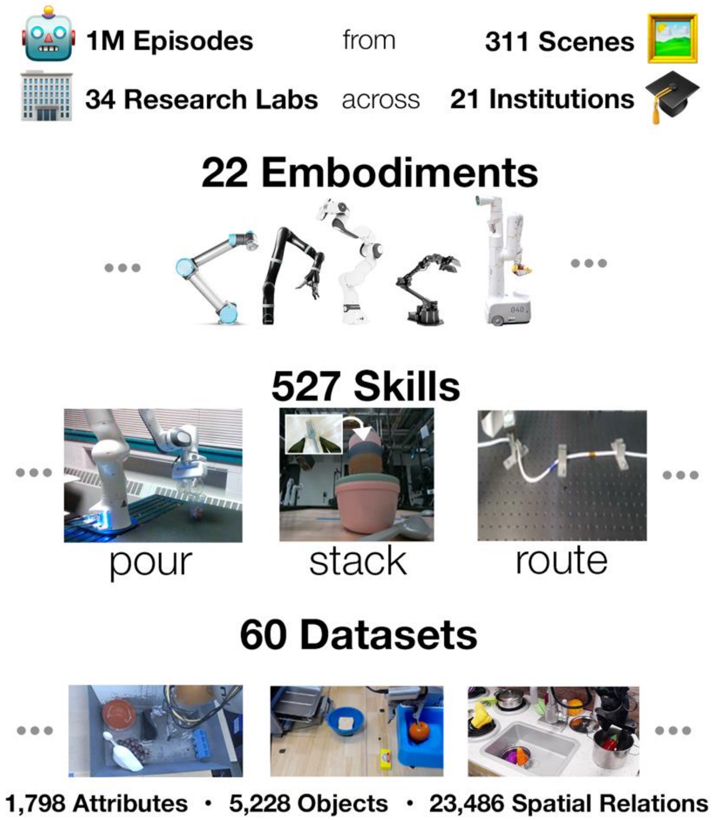
## Robotics

15 Trillion Tokens

15 Trillion Episodes

Increase size 15,000,000x

Impossible on real hardware




Open-X Embodiment: Robotic Learning Datasets and RT-X Models. ICRA. 2024. Best Paper Award.

## NLP

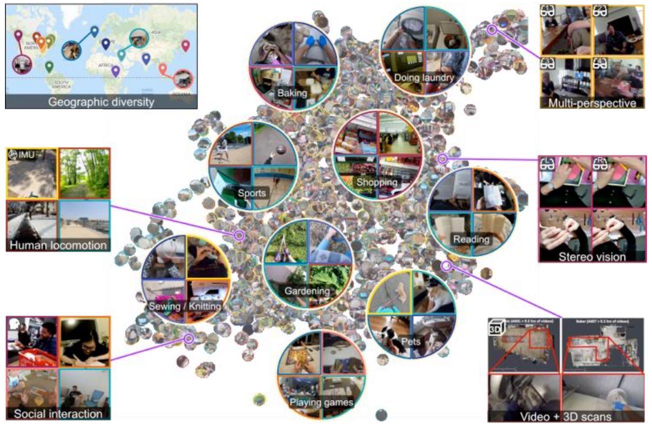
## Robotics

15 Trillion Tokens

15 Trillion Episodes

Increase size 15,000,000x

Impossible on real hardware




Ego4D

Epic Kitchens

## NLP

## Robotics

15 Trillion Tokens

15 Trillion Episodes

Increase size 15,000,000x

Impossible on real hardware

Universal Manipulation Interface: In-The-Wild Teaching without In-The-Wild Robots. Chi, Chu et al. arXiv. 2024.

## NLP

15 Trillion Tokens

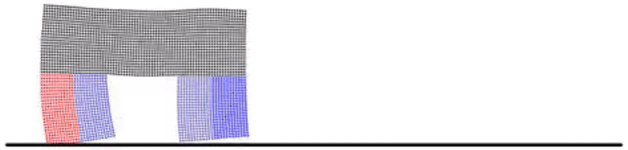
## Robotics

15 Trillion Episodes

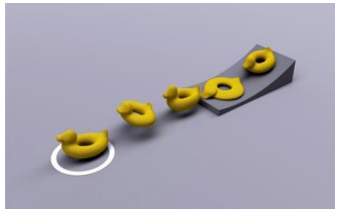
Increase size 15,000,000x

Impossible on real hardware

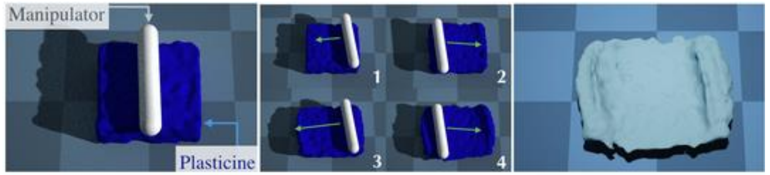
Warp [NVIDIA ’22]



DiffTaichi [Hu et al. ’20]



DiffPD [Du et al. ’22]




Initial State | Manipulation | Target Plasticine Shape

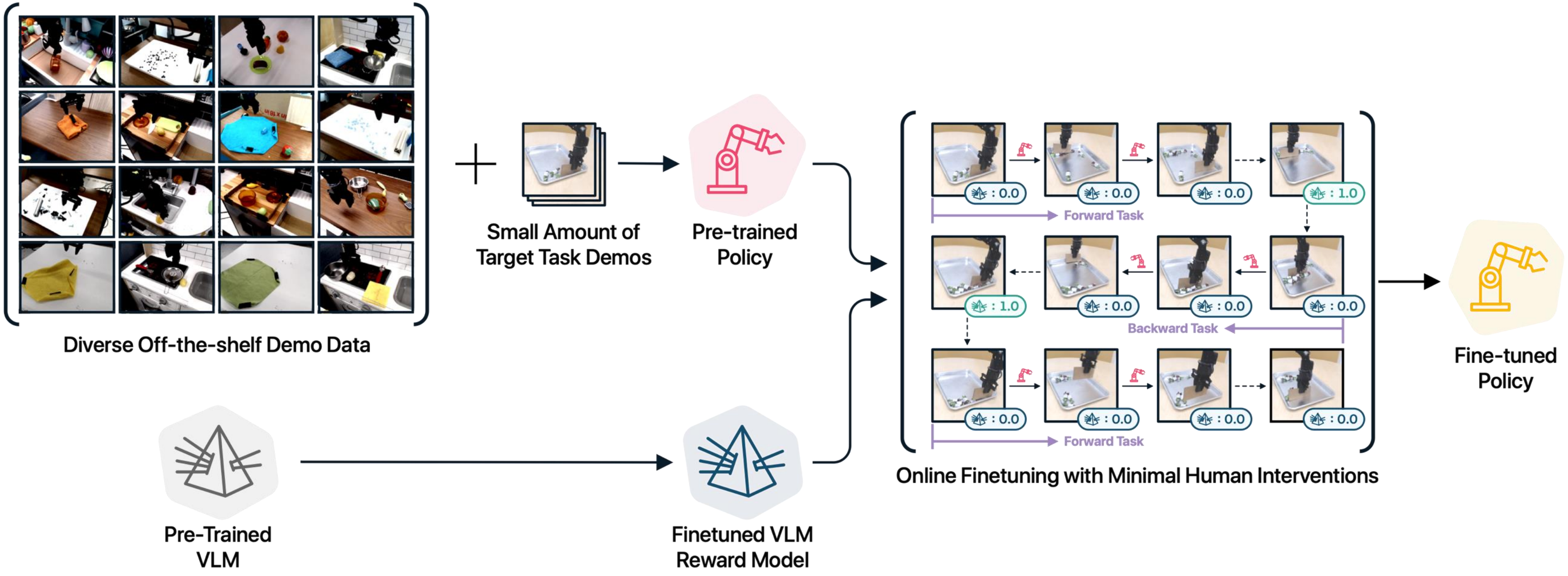
PlasticineLab [Huang et al. ’21]

# Beyond data scaling, what else do we need?

- Planning by sequencing short skills
- More than just RGB data -> depth, tactile, …
- Better and more efficient algorithms
- Online adaptation and exploration

Stanford University

# RoboFuME: Making Robot Fine-tuning Easy by Improving Autonomy



# Principles of Robot Autonomy II

Learning-based Approaches to **Manipulation** & **Interactive Perception**

Jeannette Bohg

Stanford University


FIRE CHIEF


# Interactive Perception

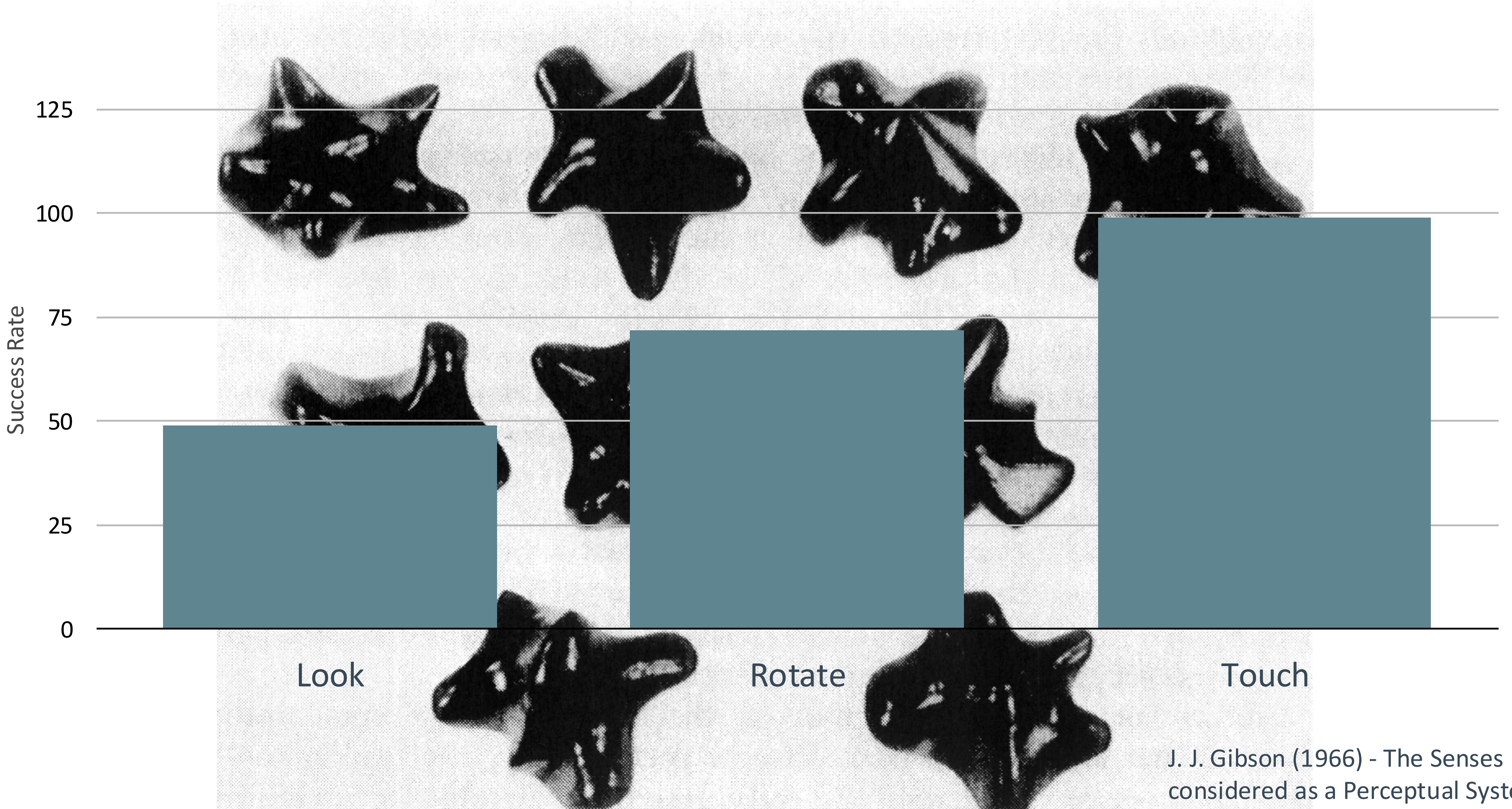
$$S \times A \times t$$

Sensory Data Actions Time

Control

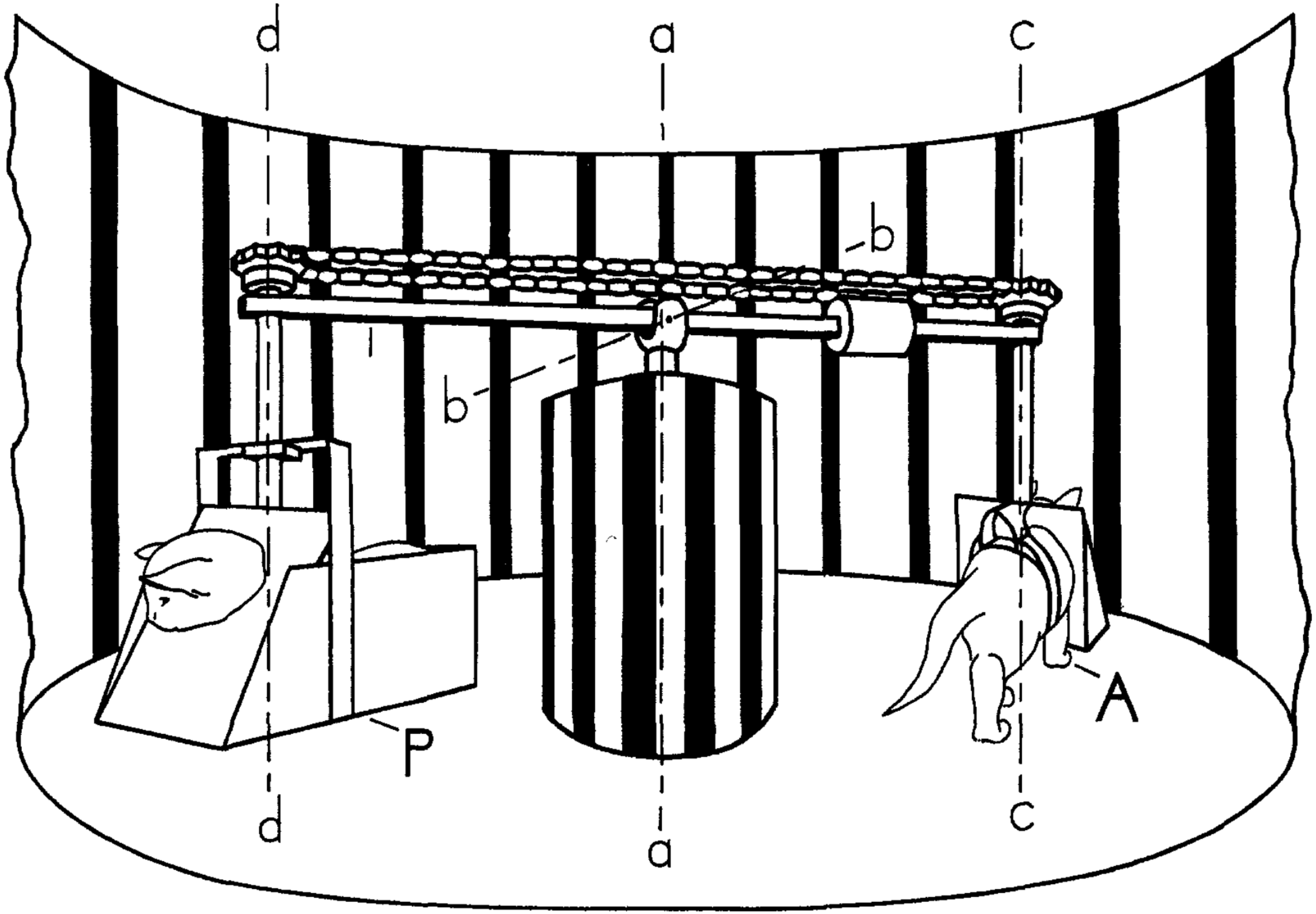
Interactive Perception - Leveraging Action in Perception and Perception in Action.
Bohg, Hausman, Sanakaran, Brock, Kragic, Schaal and Sukhatme. TRO '17.

# Exploiting Multi-Modality



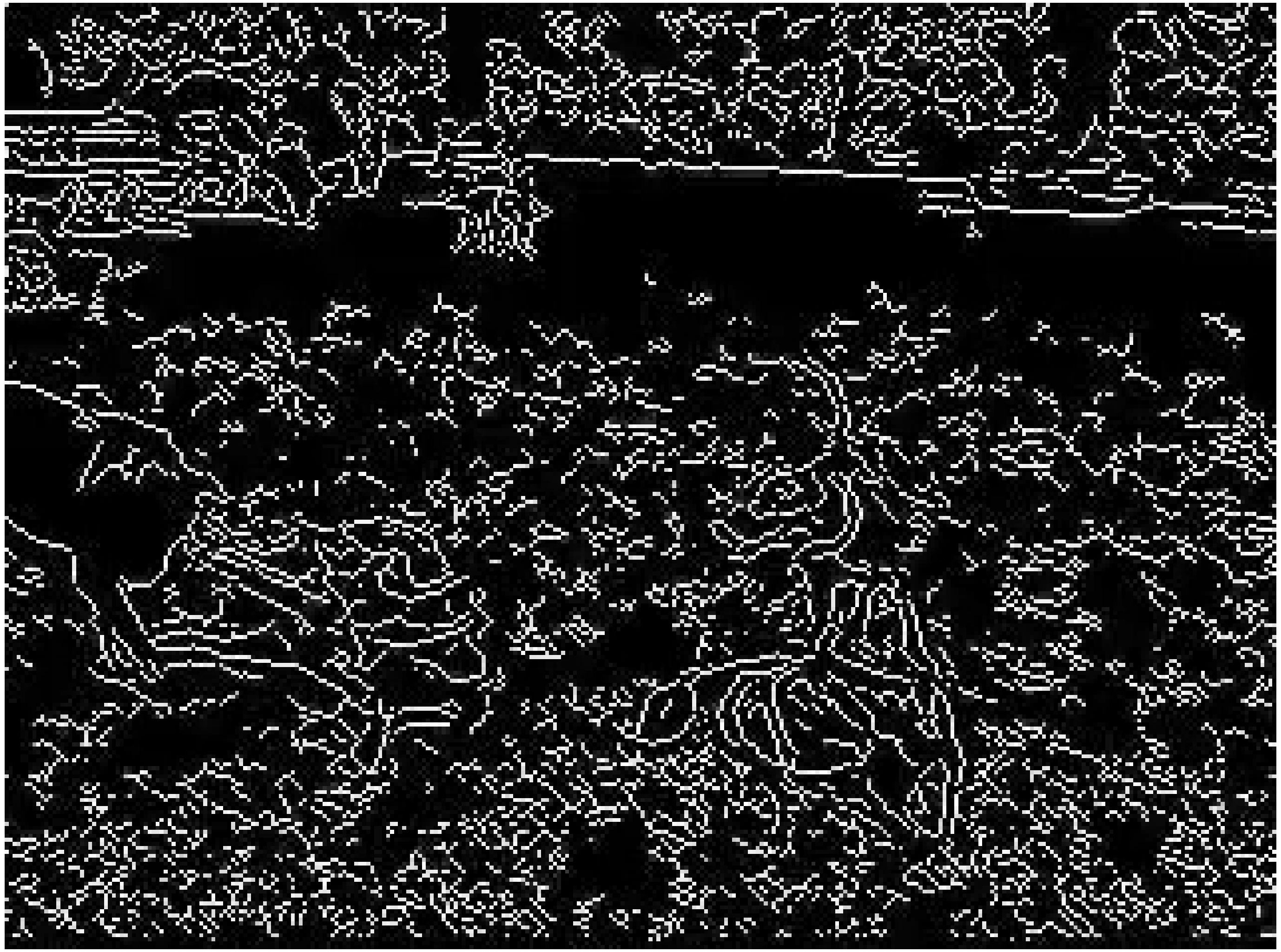

J. J. Gibson (1966) - The Senses considered as a Perceptual System.

# Concurrency of Motion and Sensing




Held and Hein (1963). Movement-Produced Stimulation in the Development of Visually-Guided Behaviour

# Accumulation over Time



Thanks to Octavia Camps at Northeastern University, Boston

# Interactive Perception

$$\boxed{S \times A} \times t$$

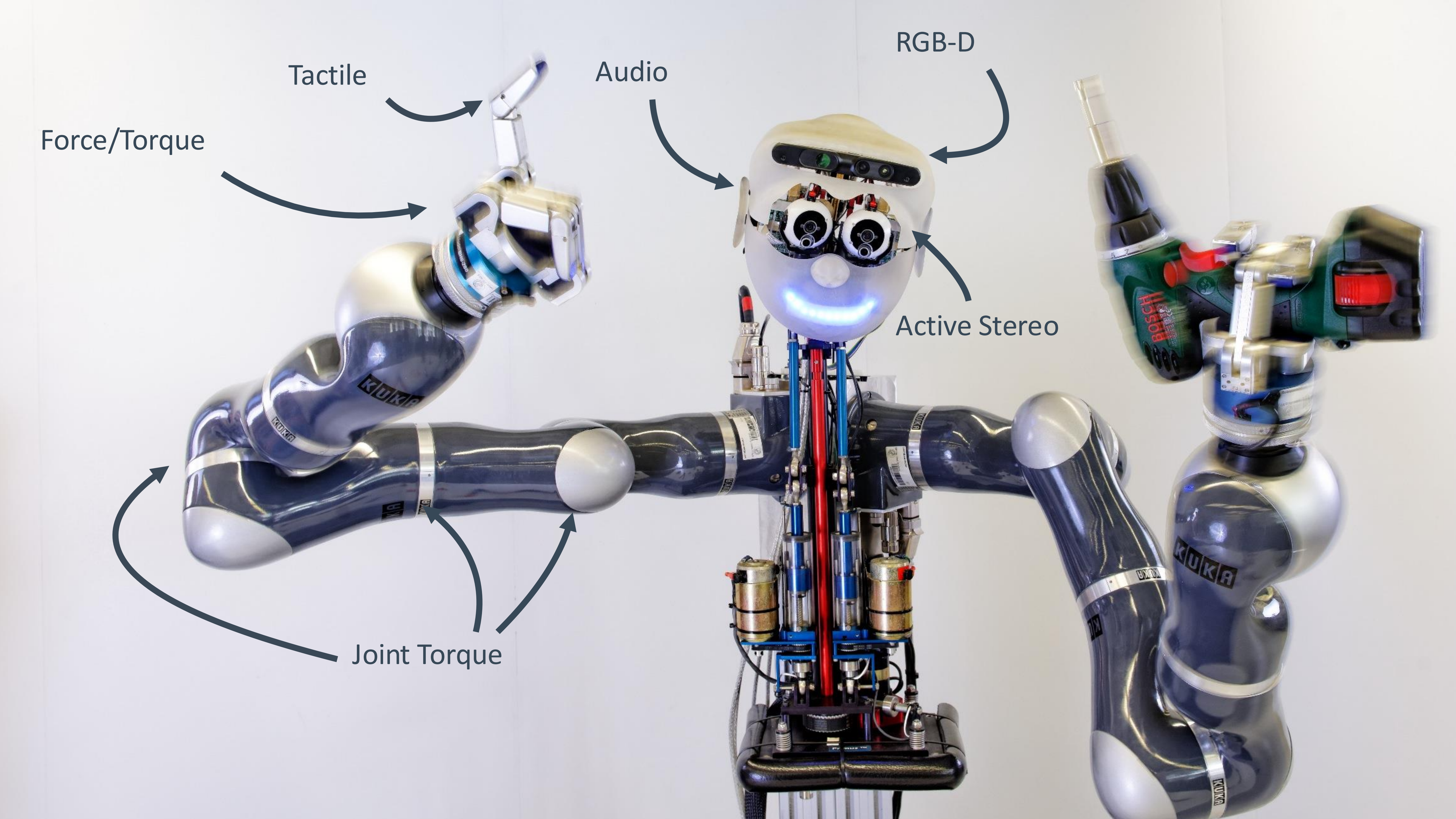
Sensory Data

Actions

Time

Selfsupervised Learning of a
Multimodal Representation

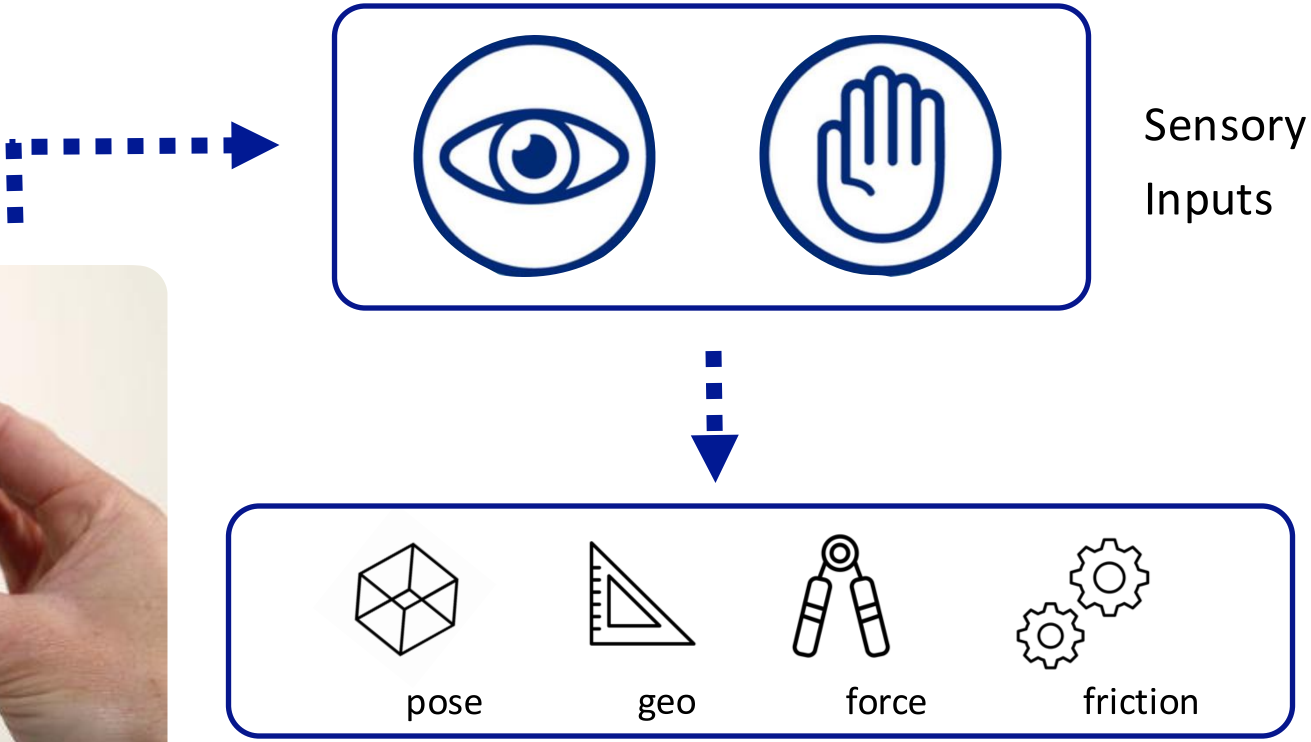
Interactive Perception - Leveraging Action in Perception and Perception in Action.
Bohg, Hausman, Sanakaran, Brock, Kragic, Schaal and Sukhatme. TRO '17.



Tactile
Audio
RGB-D
Force/Torque
Active Stereo
Joint Torque


Jenga
Jenga
Jenga
Jenga

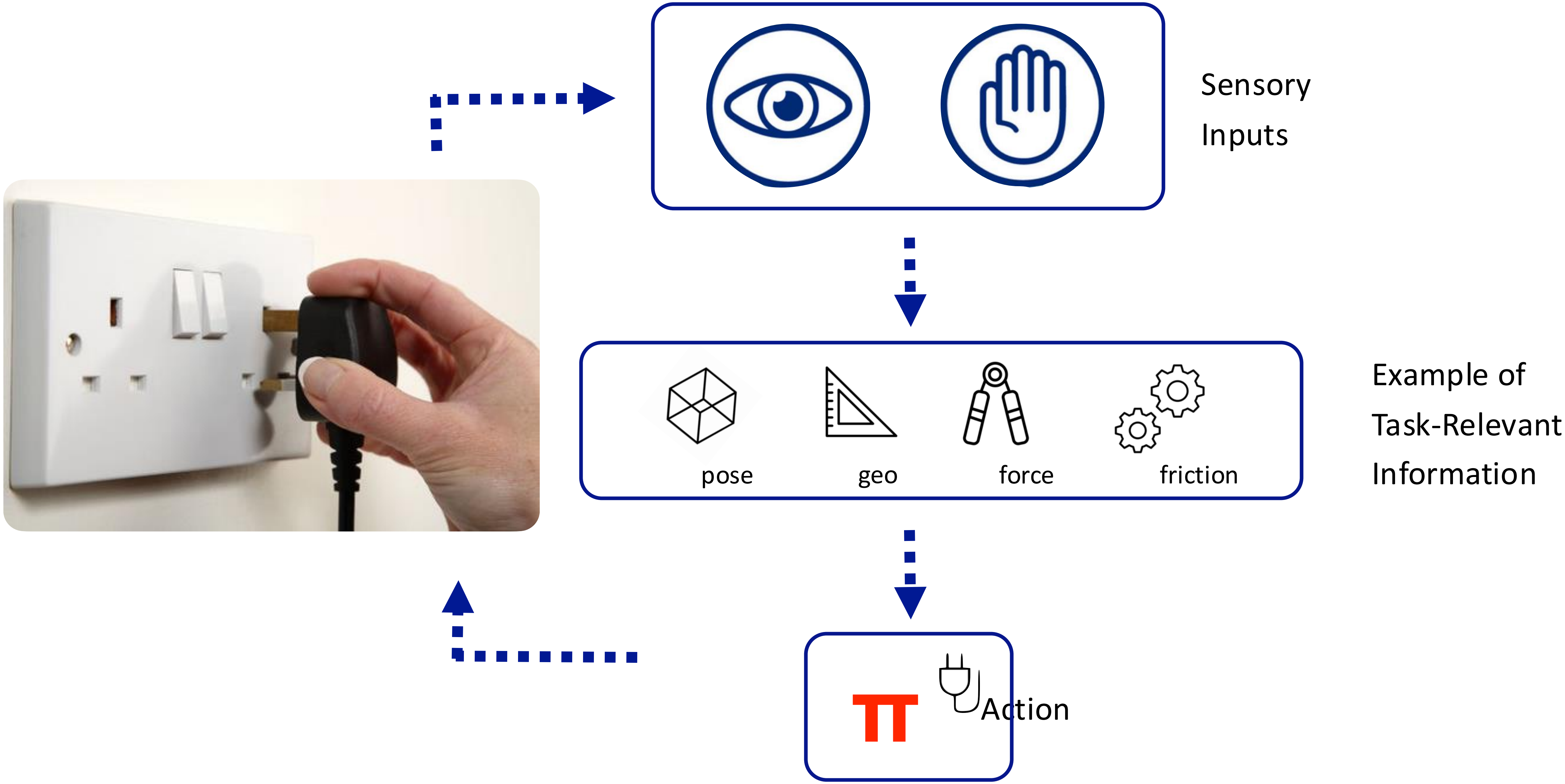
# Human multimodal sensor-motor coordination



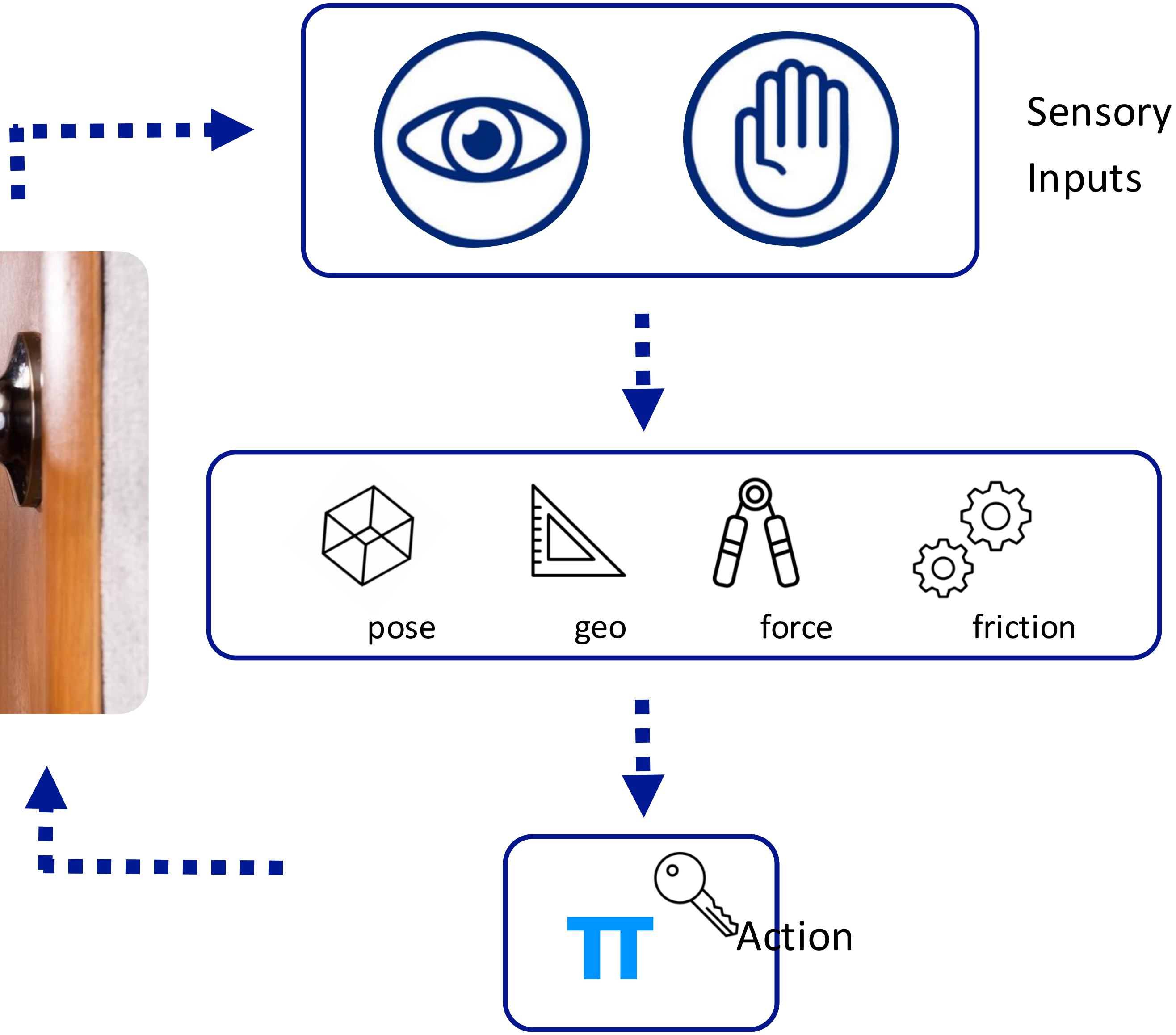

Sensory Inputs

Example of Task-Relevant Information

# Human multimodal sensor-motor coordination



# Human multimodal sensor-motor coordination



# Generalizable representations for multimodal inputs

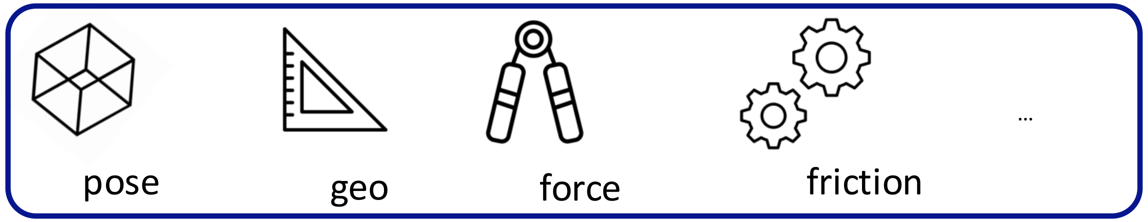
# Generalizable representations for multimodal inputs

$$\pi(f(o_1, o_2, o_3 \ldots o_n)) = a$$

multimodal sensory inputs

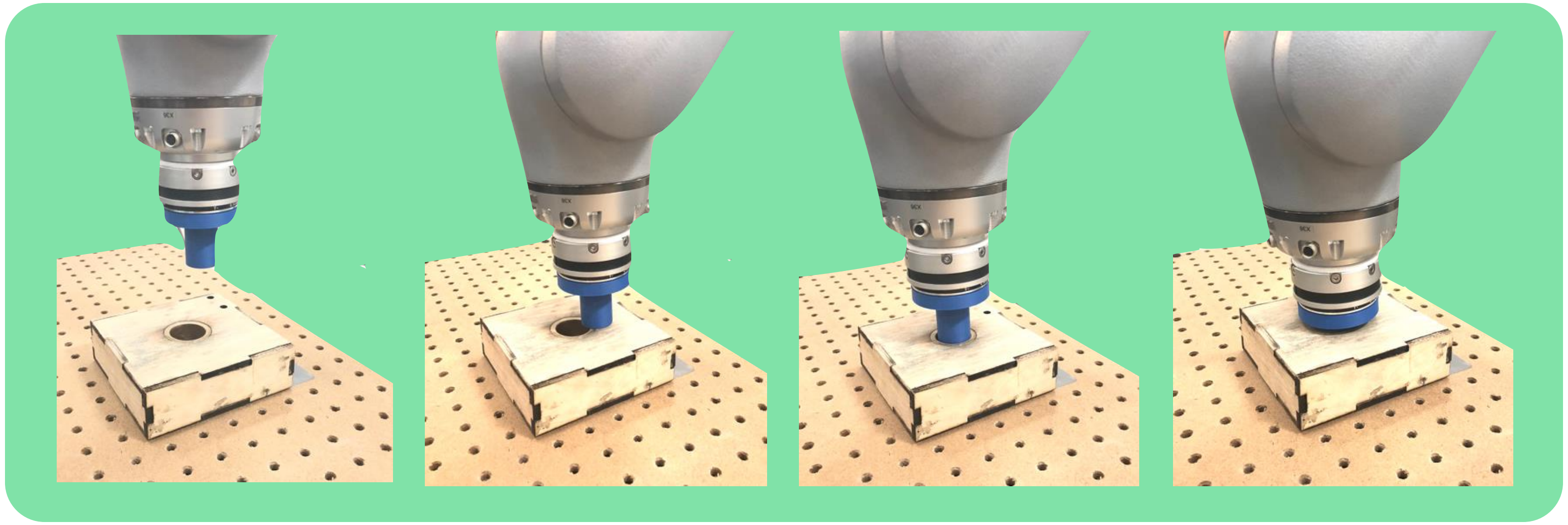


Learn representation

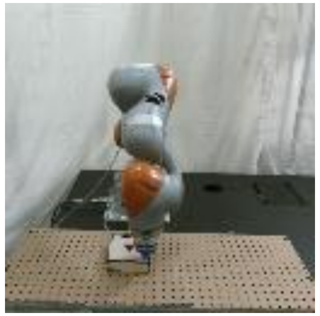



learn policy
for new task instances

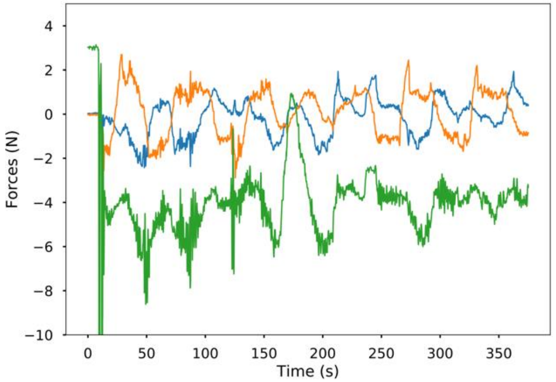
# Experimental setup



Multimodal sensory inputs



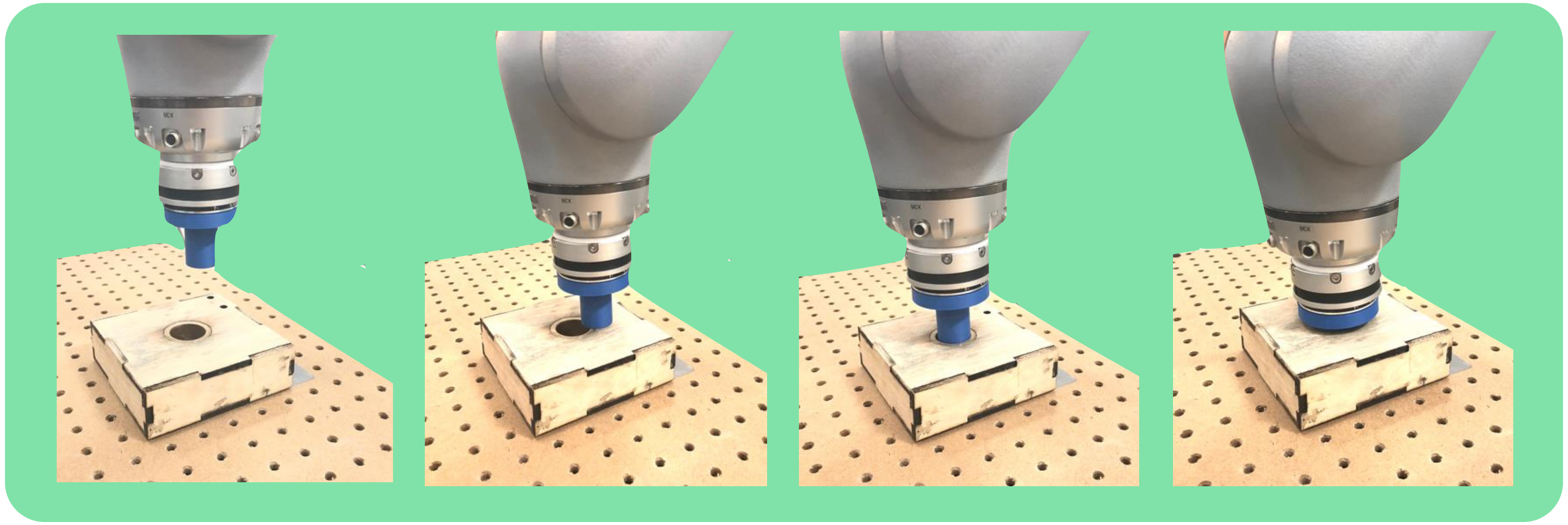
RGB




force/torque

robot states

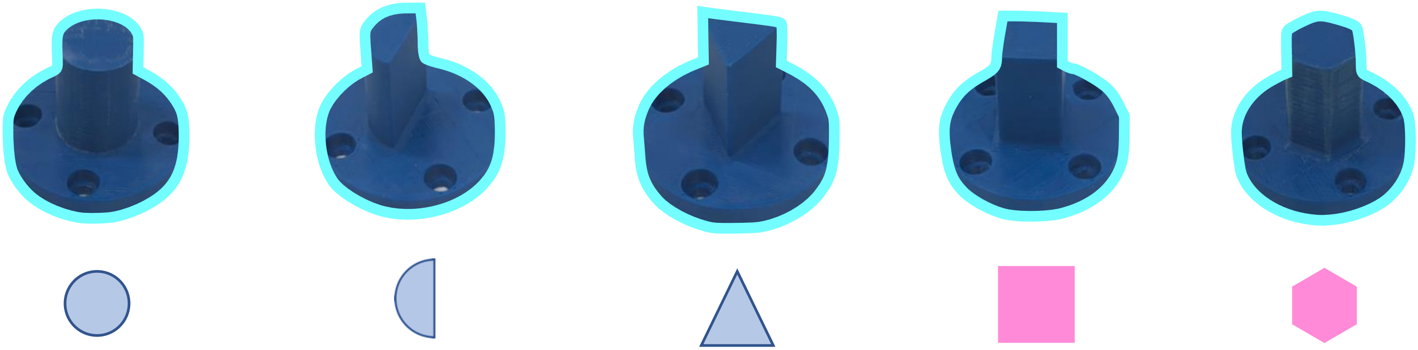
# Experimental setup



Training

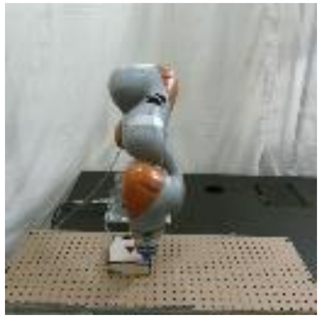
Testing

Peg geometry

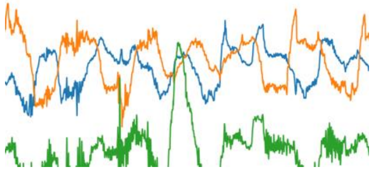


# Learning generalizable representation

Inputs

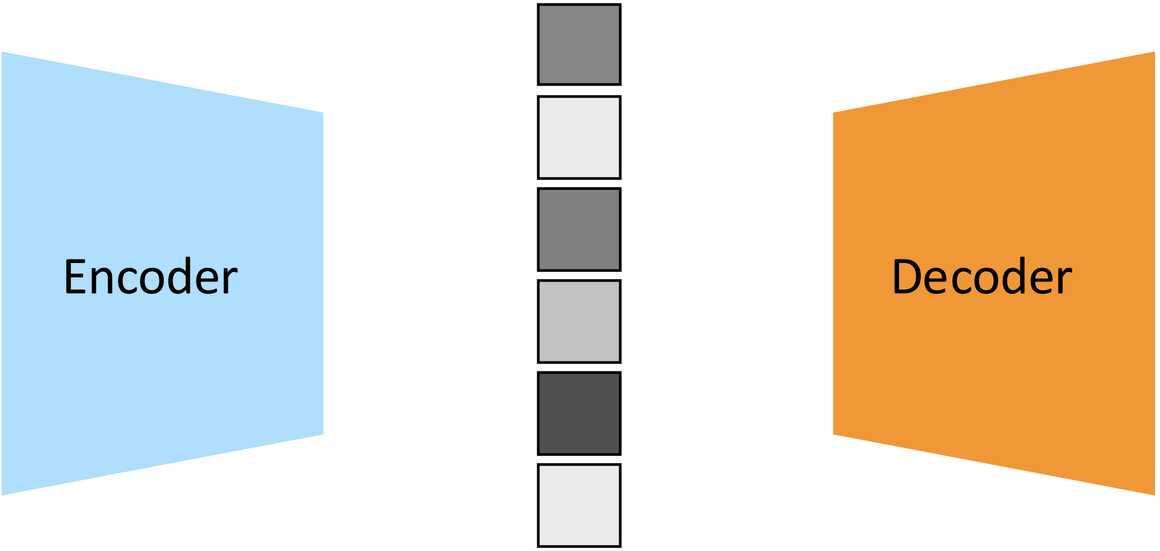


RGB image



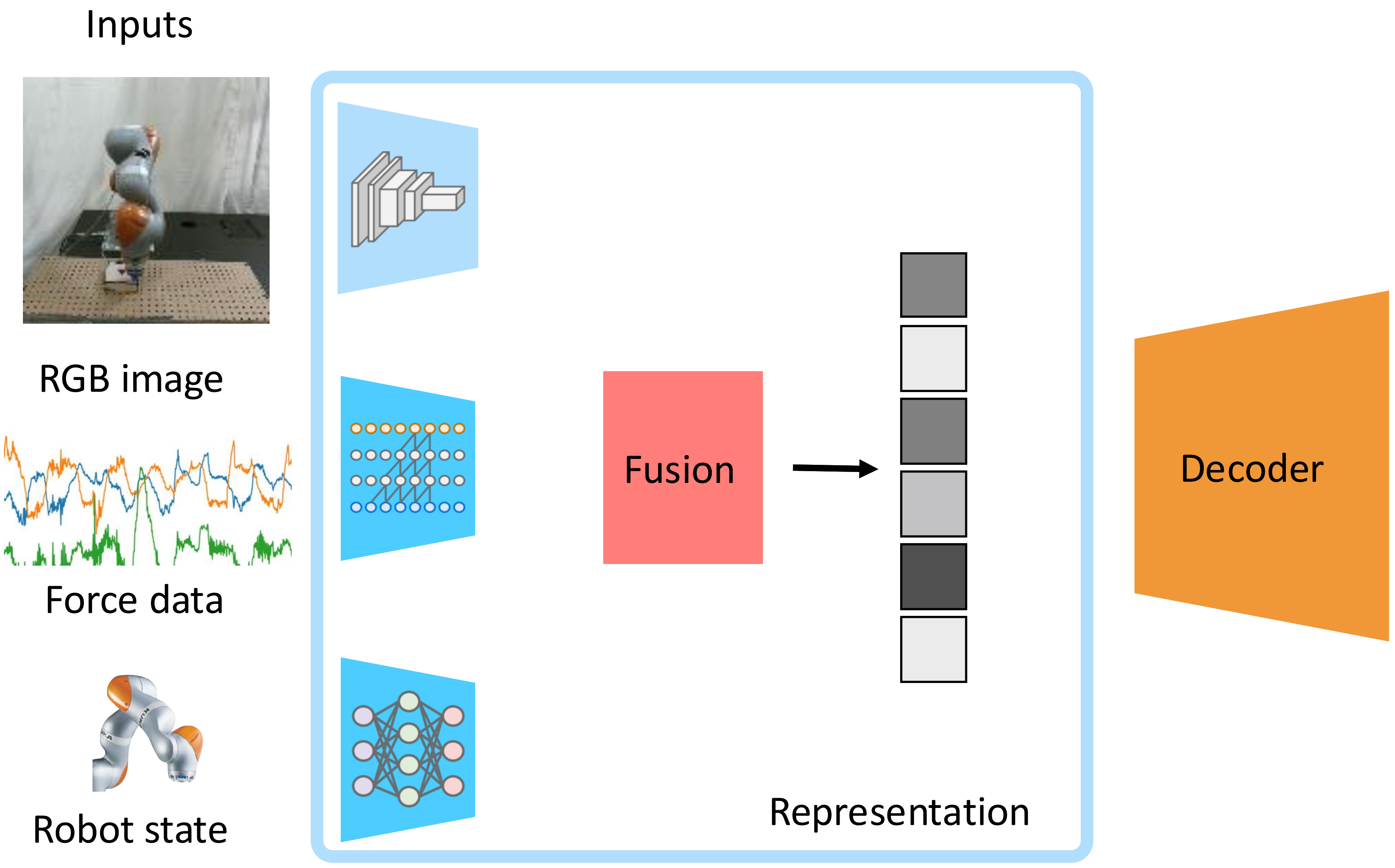
Force data

Robot state

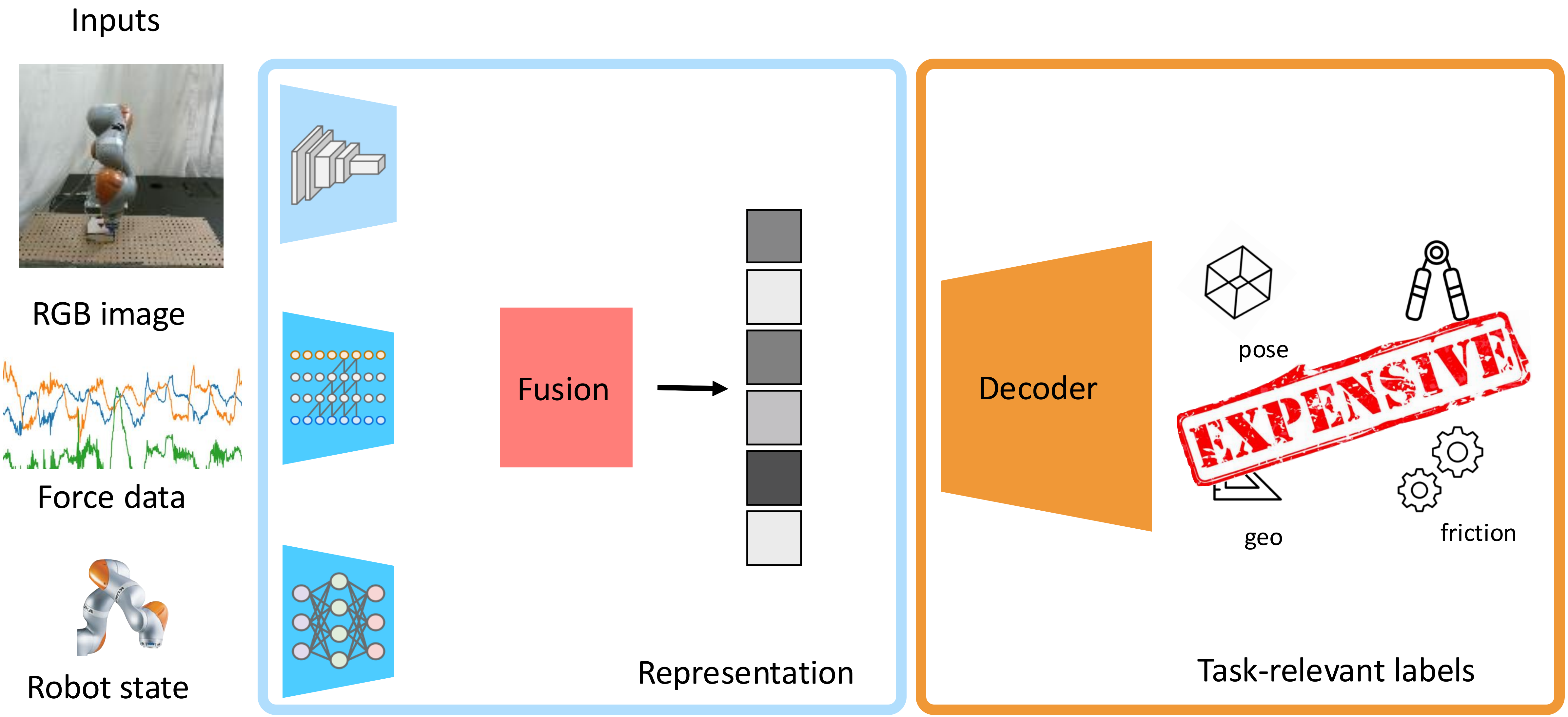



Representation

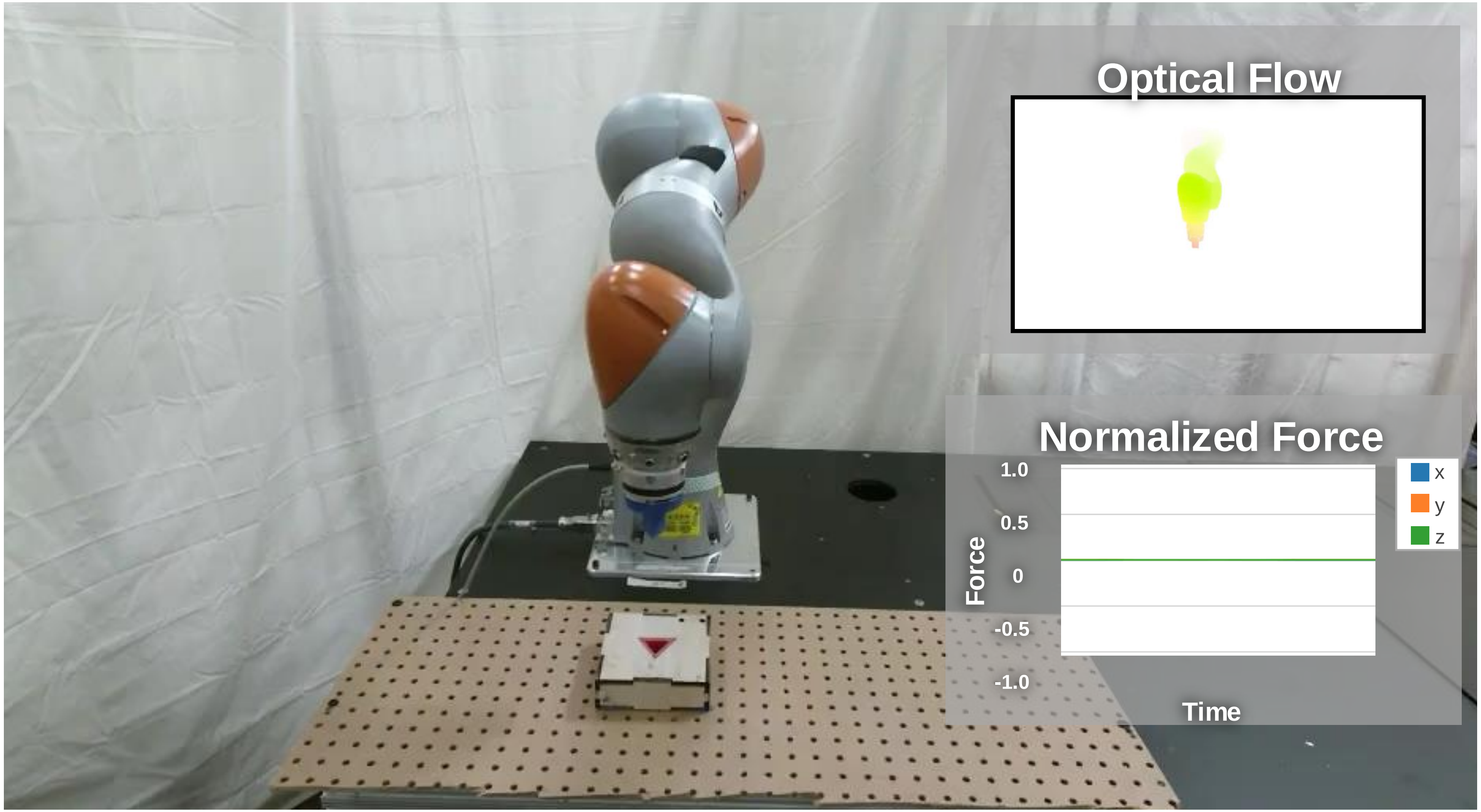
# Learning generalizable representation



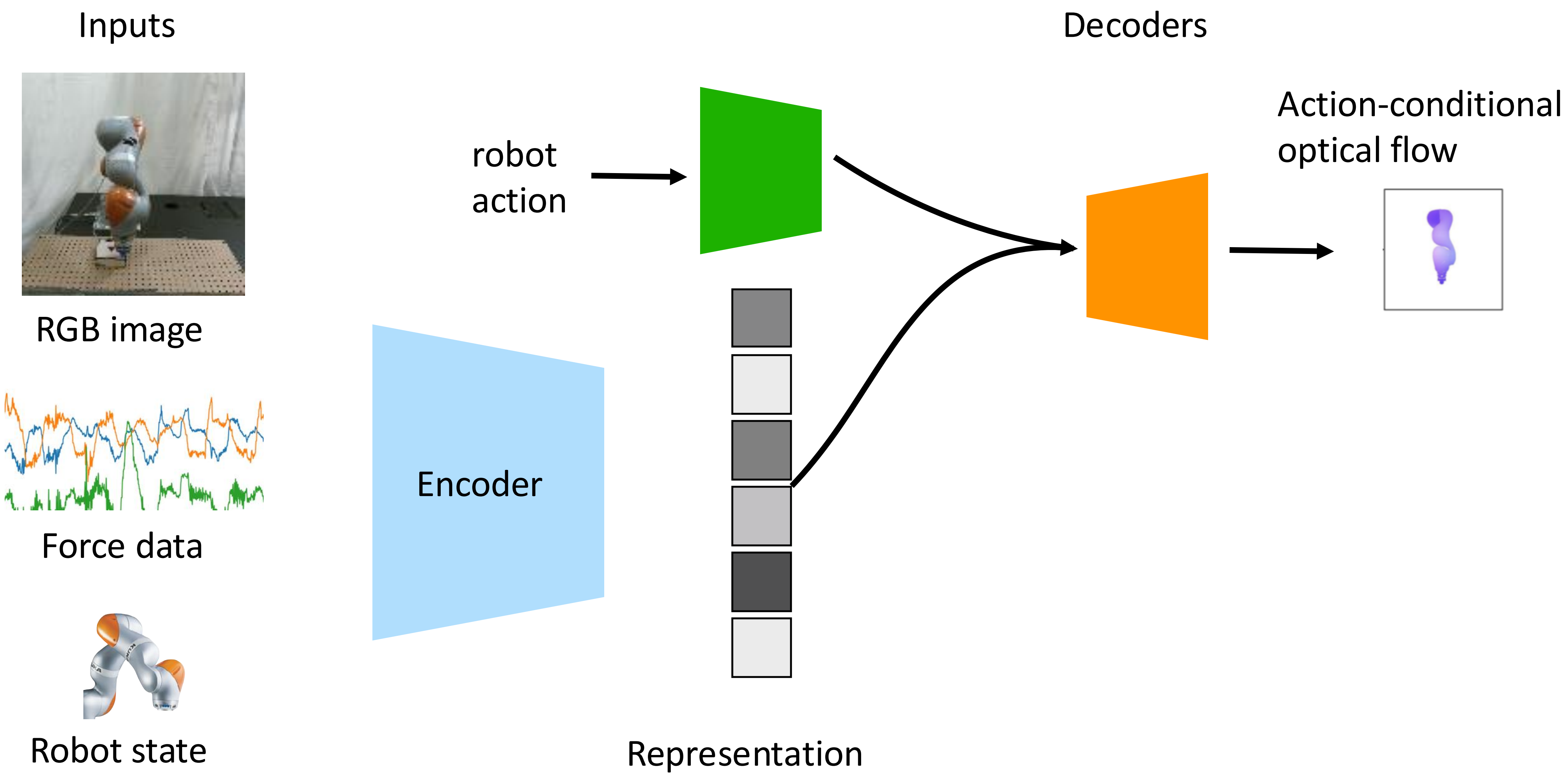
# Learning generalizable representation



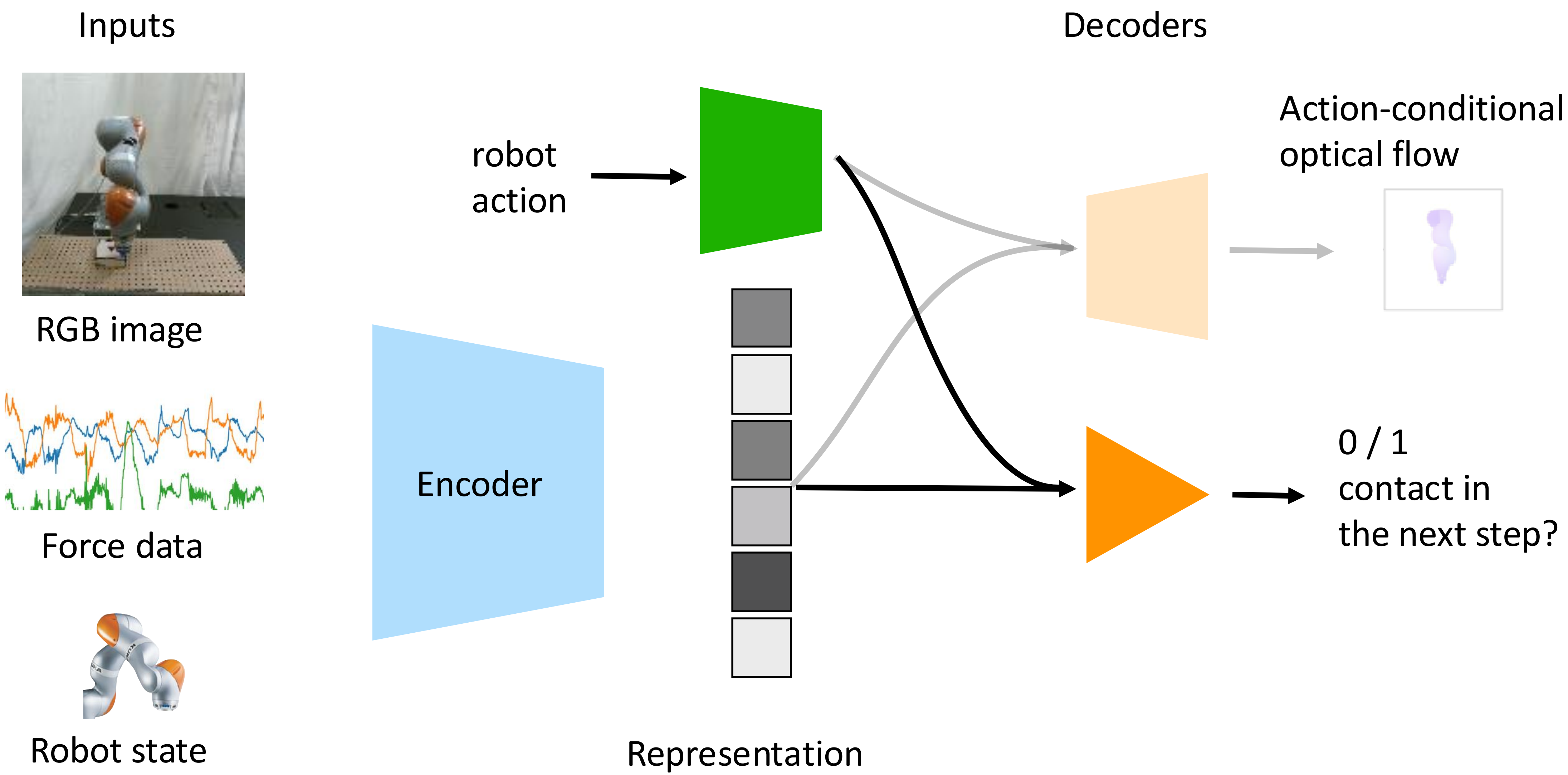
# Collect labeled data with self-supervision



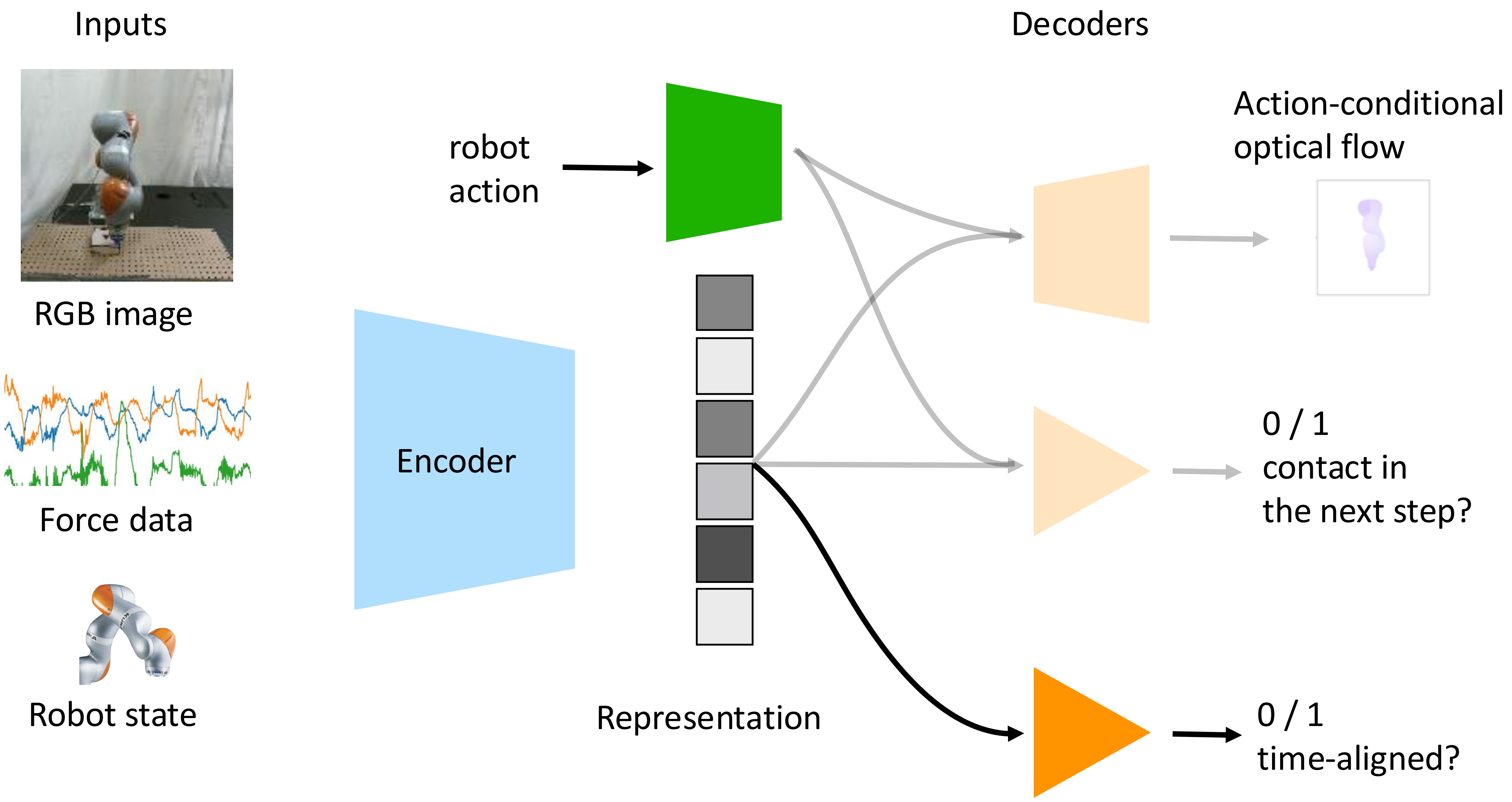
# Dynamics prediction from self-supervision



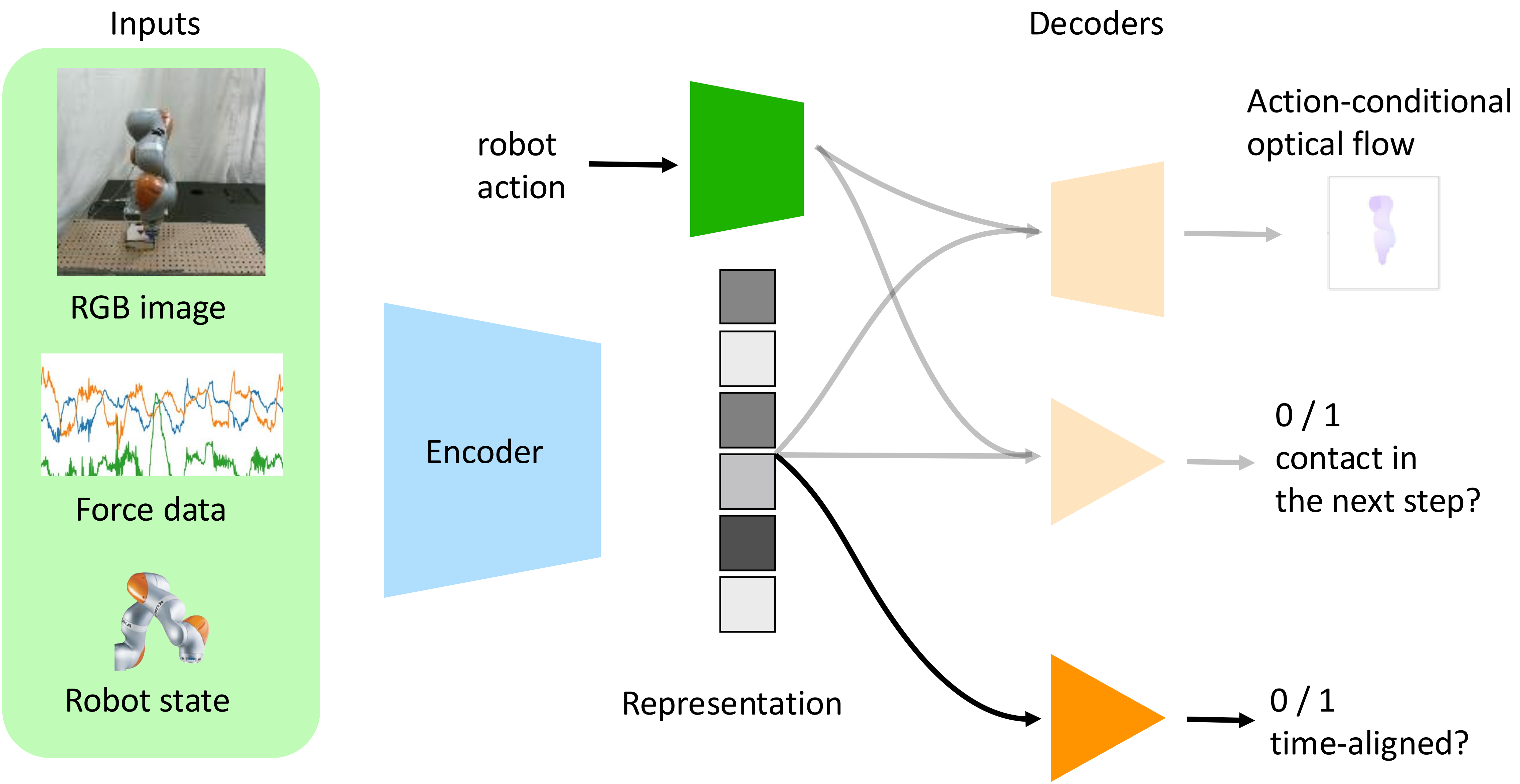
# Dynamics prediction from self-supervision



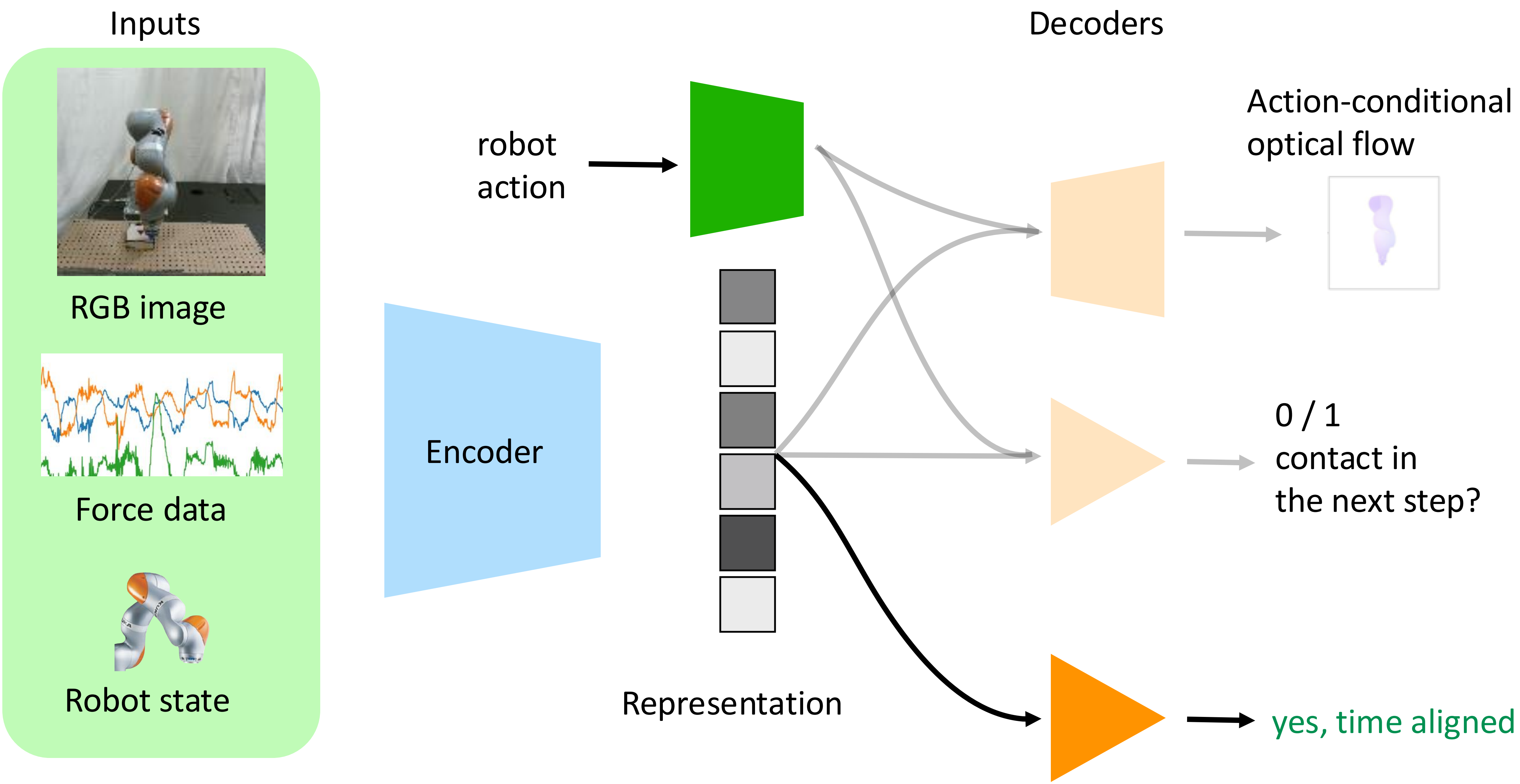
# Concurrency prediction from self-supervision



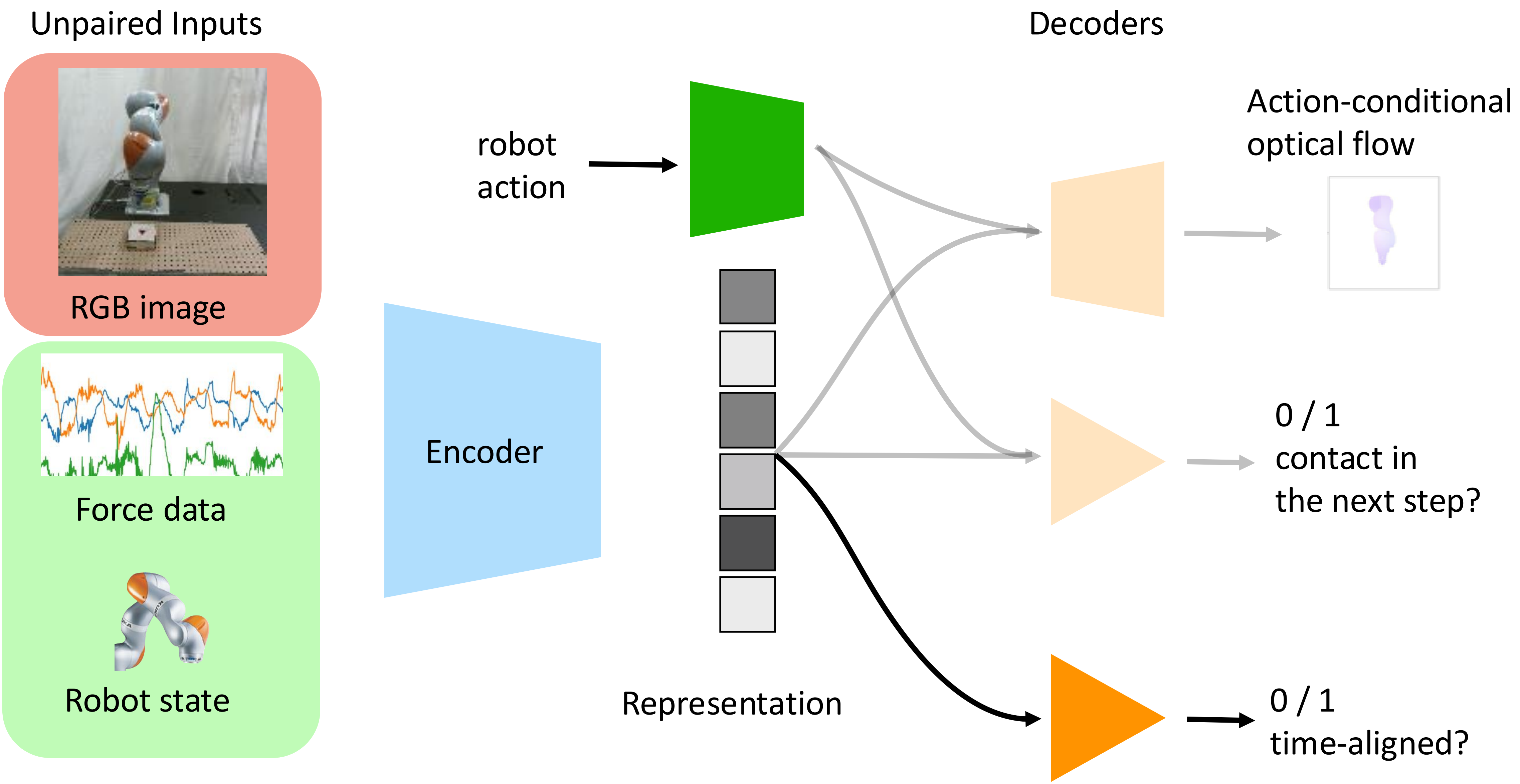
# Concurrency prediction from self-supervision



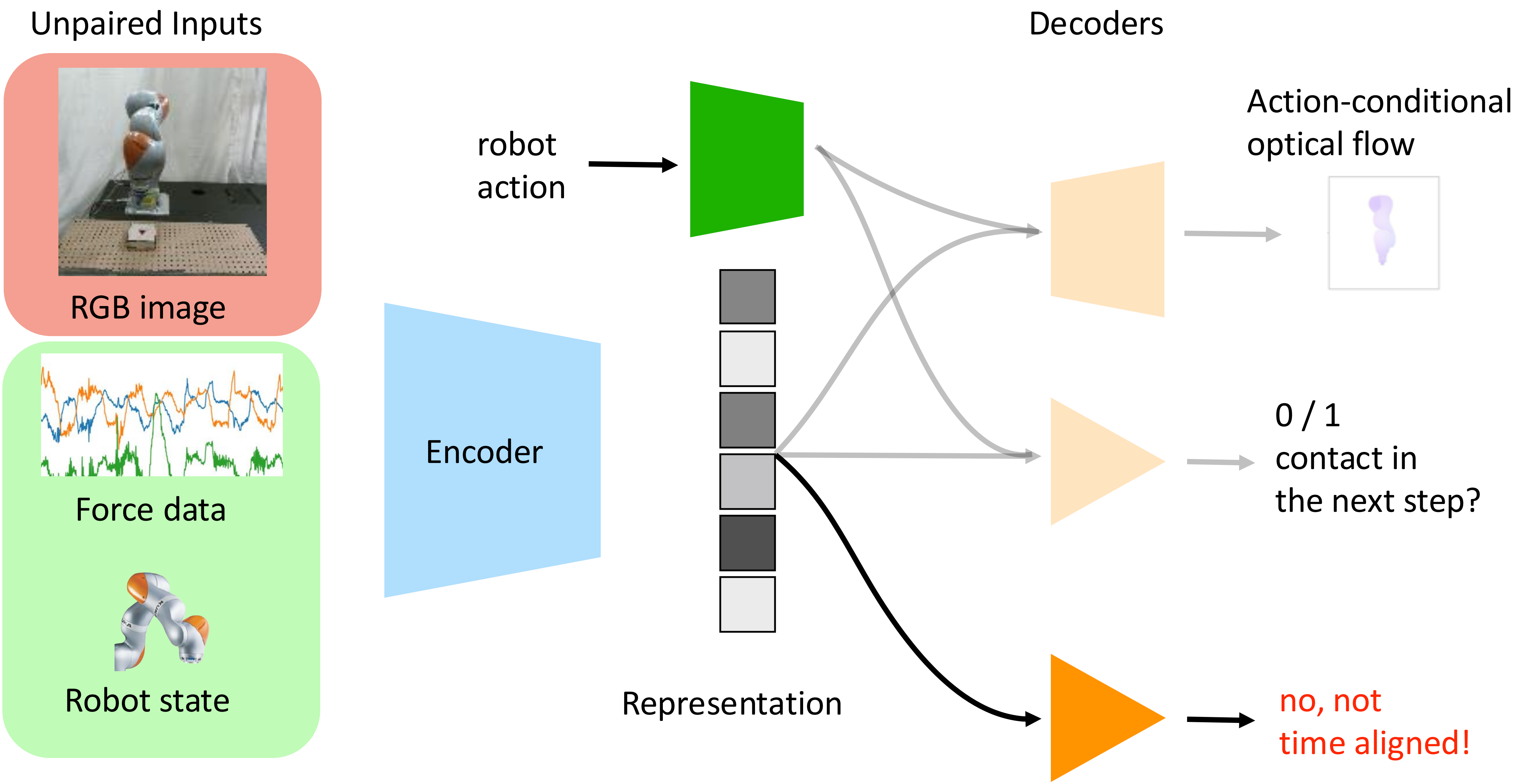
# Concurrency prediction from self-supervision



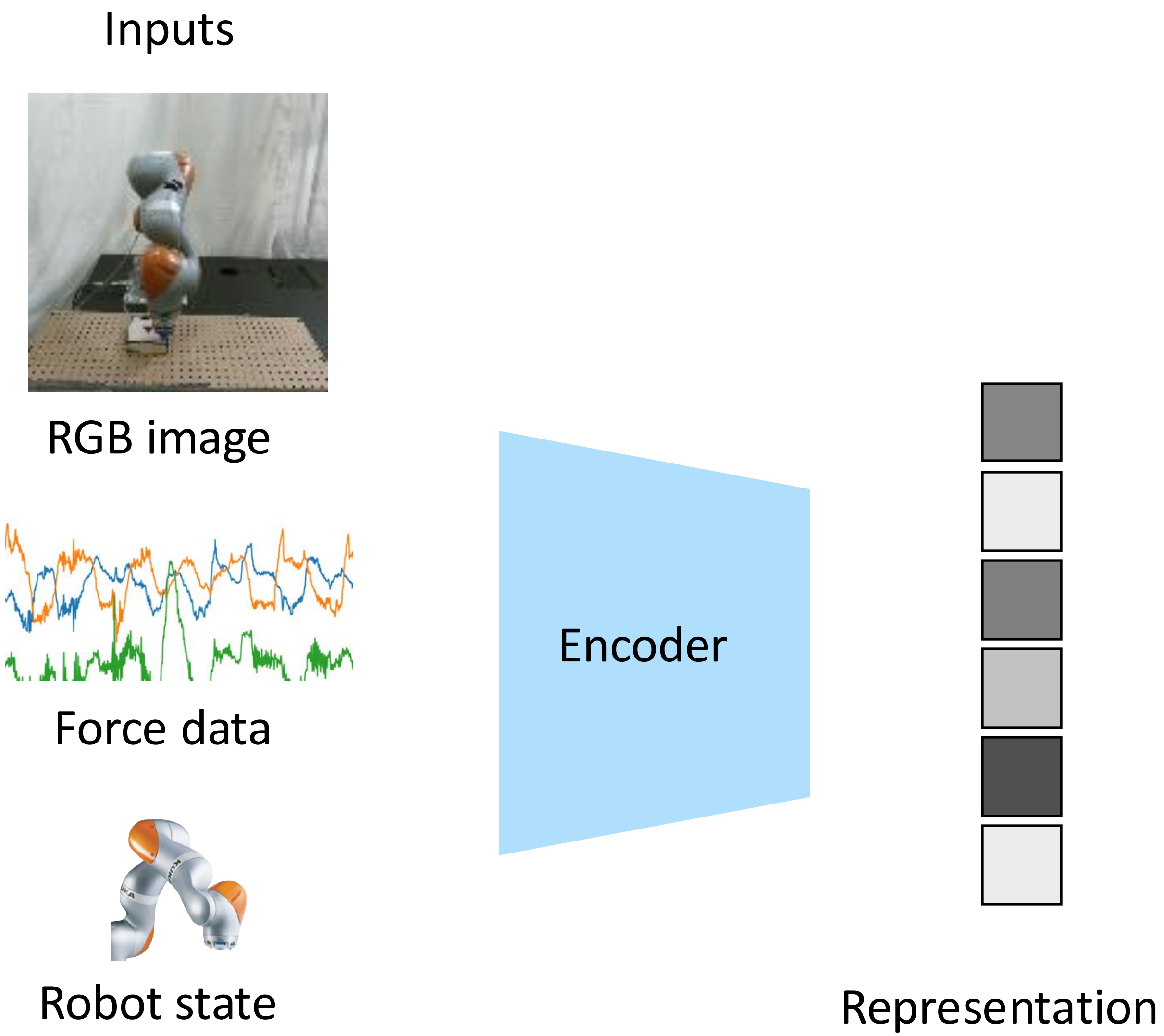
# Concurrency prediction from self-supervision



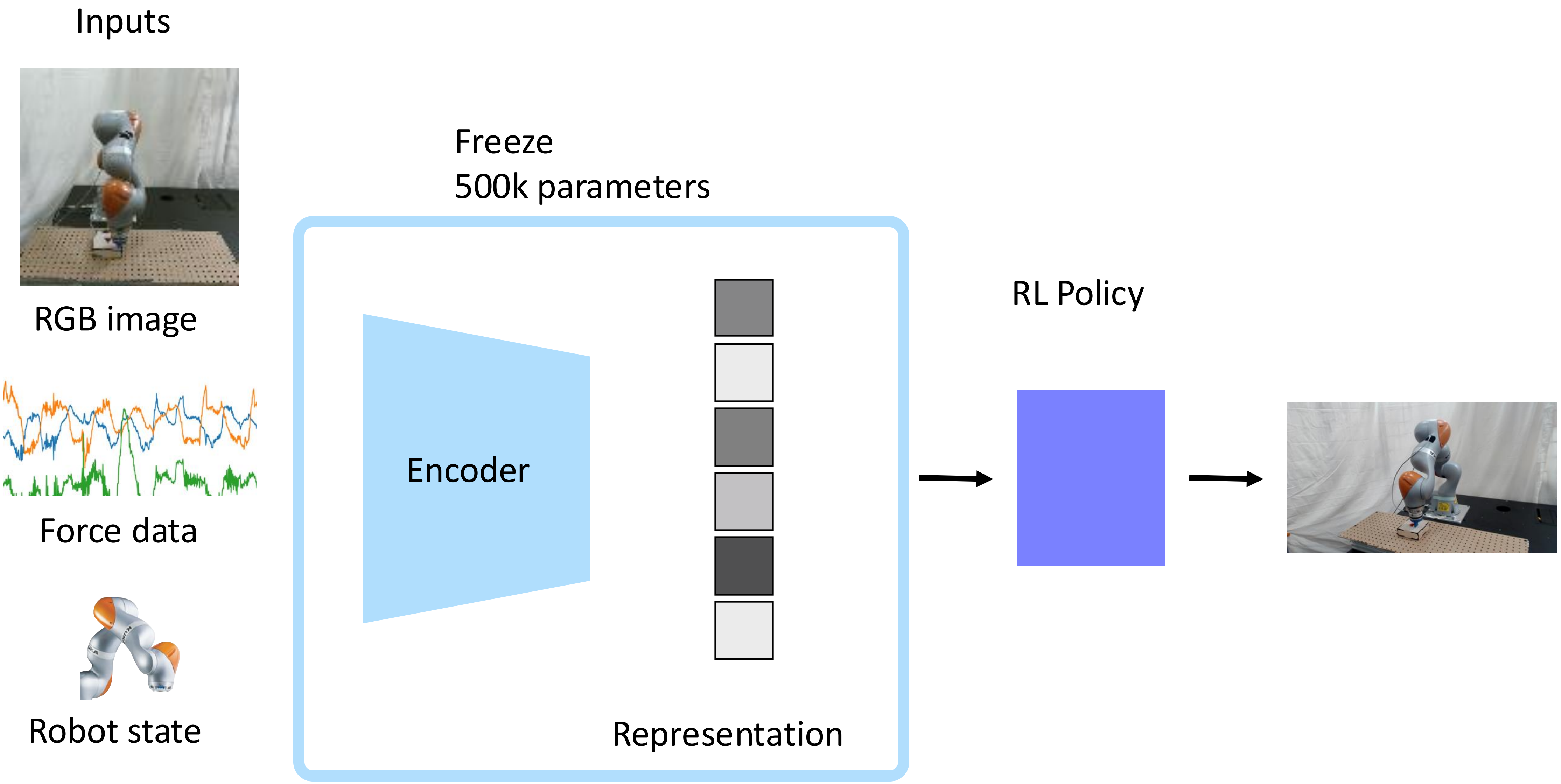
# Concurrency prediction from self-supervision



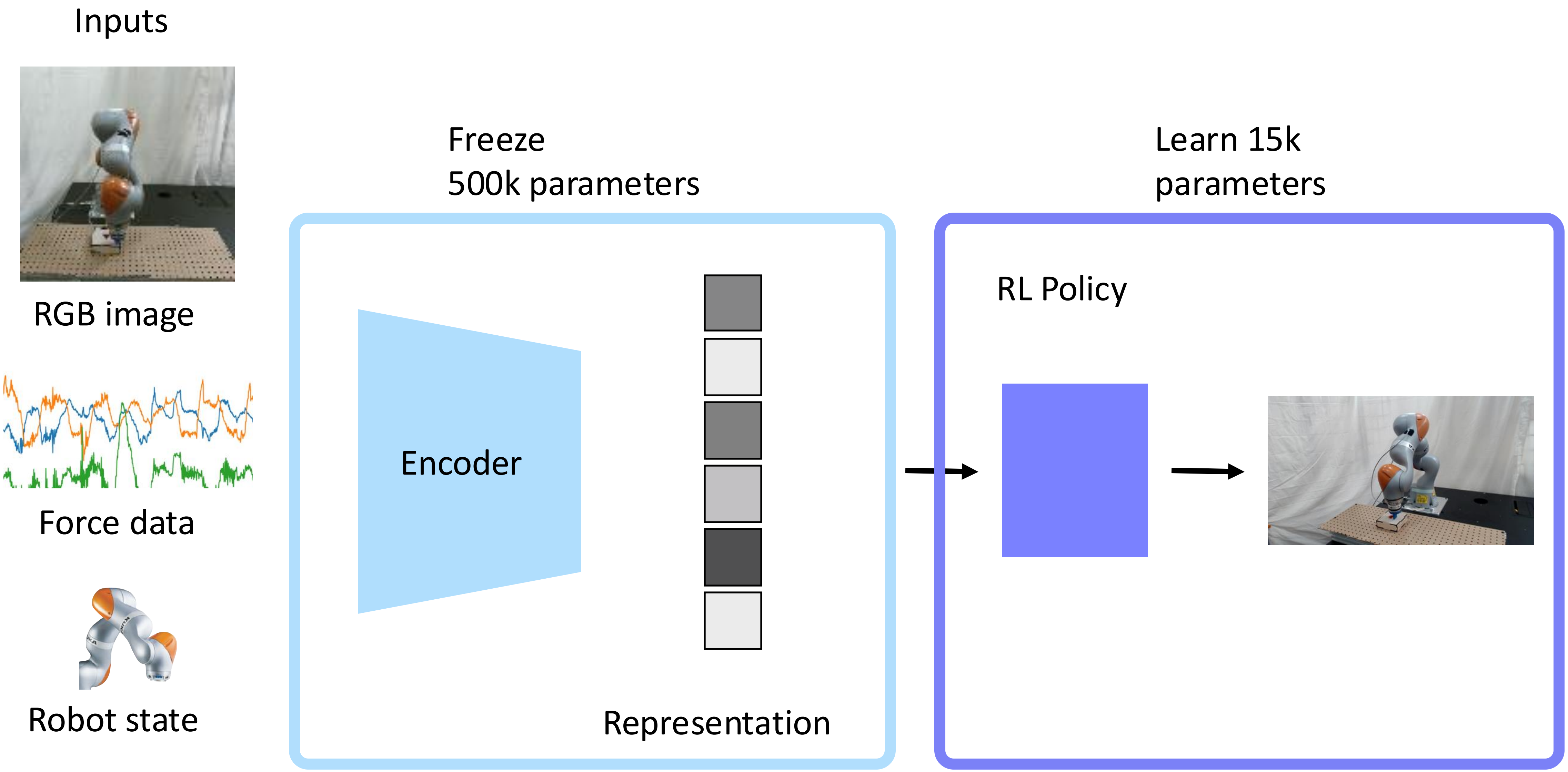
# Learning sample efficient policies



# How to efficiently learn a policy?



# How to efficiently learn a policy?



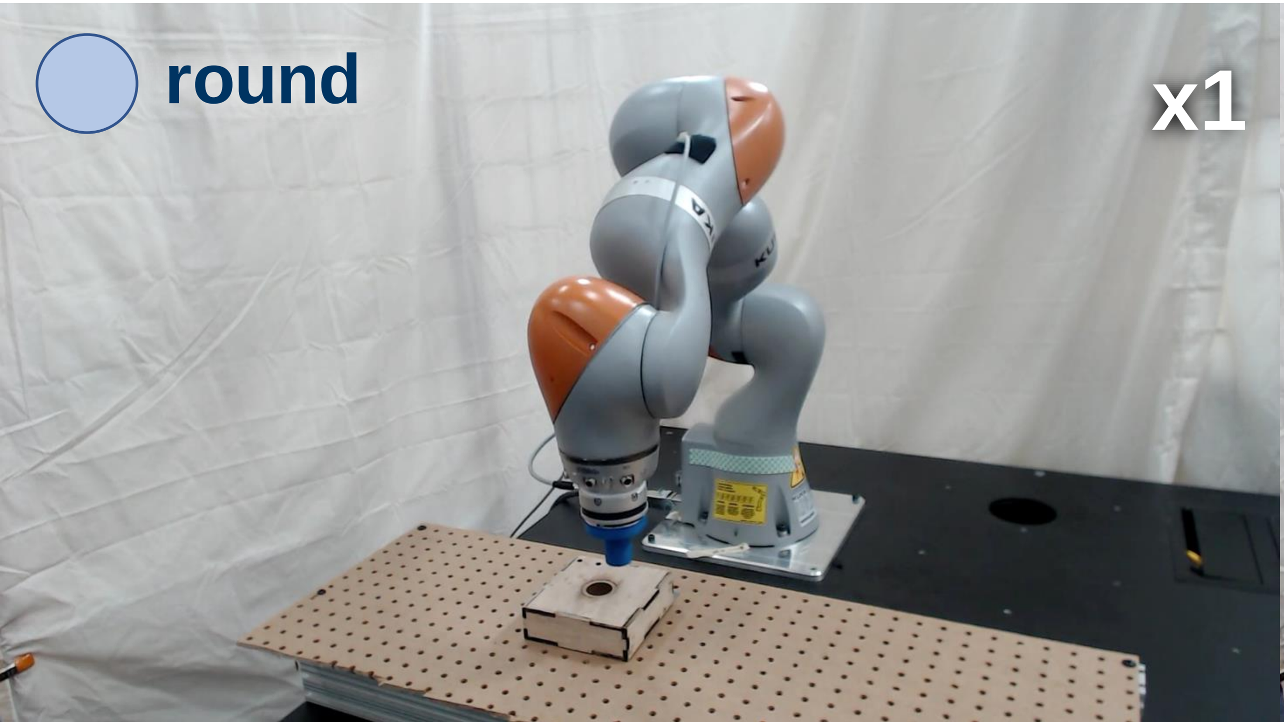
# We evaluate our representation with policy learning

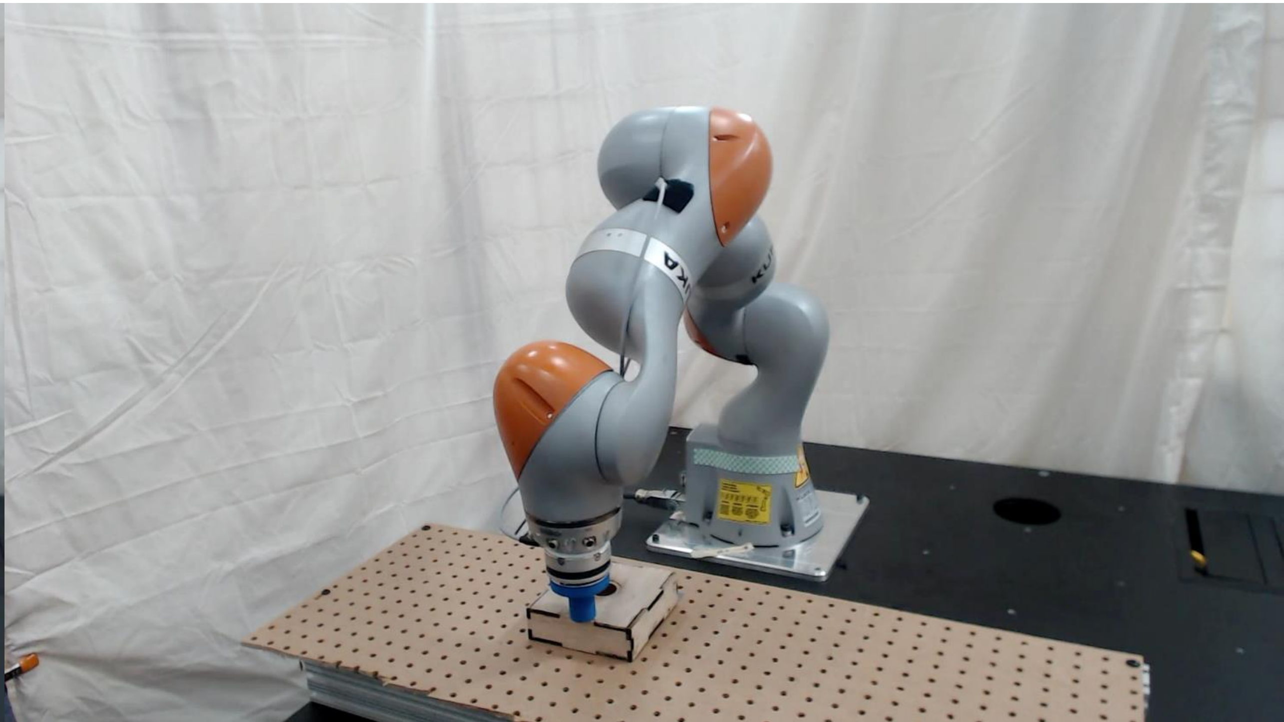
Episode 0

0% success rate

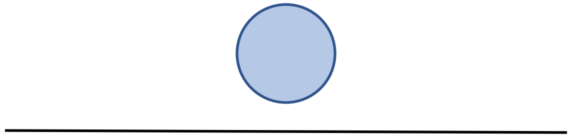
Episode 100

21% success rate



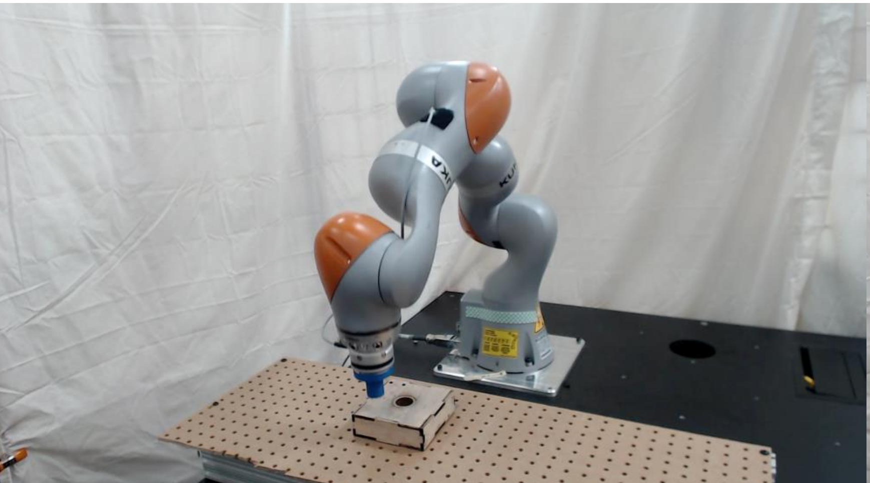



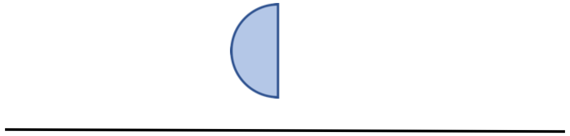
# We efficiently learn policies in 5 hours



Episode 300

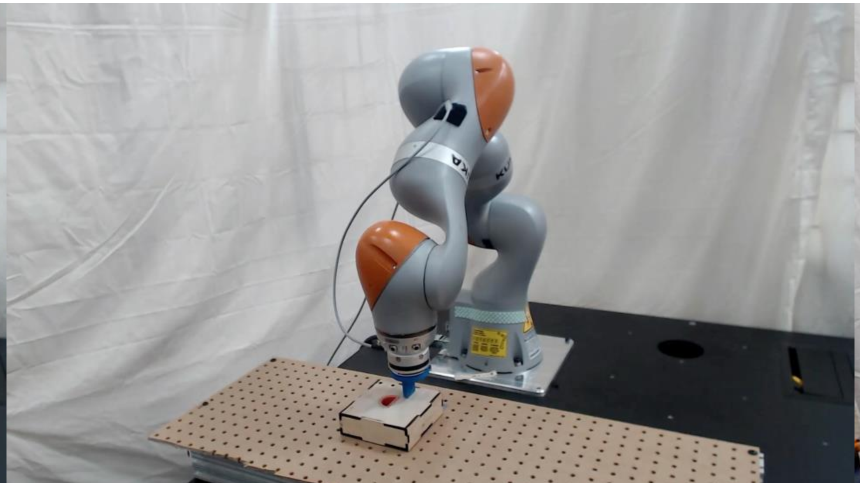
73% success rate

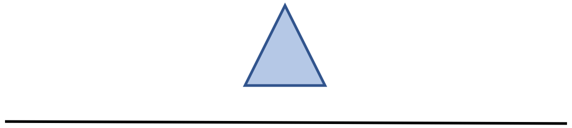




Episode 300

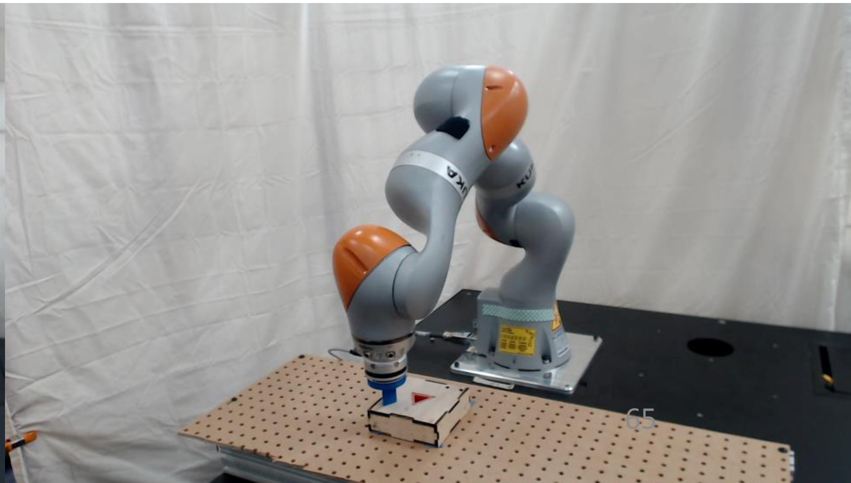
71% success rate





Episode 300

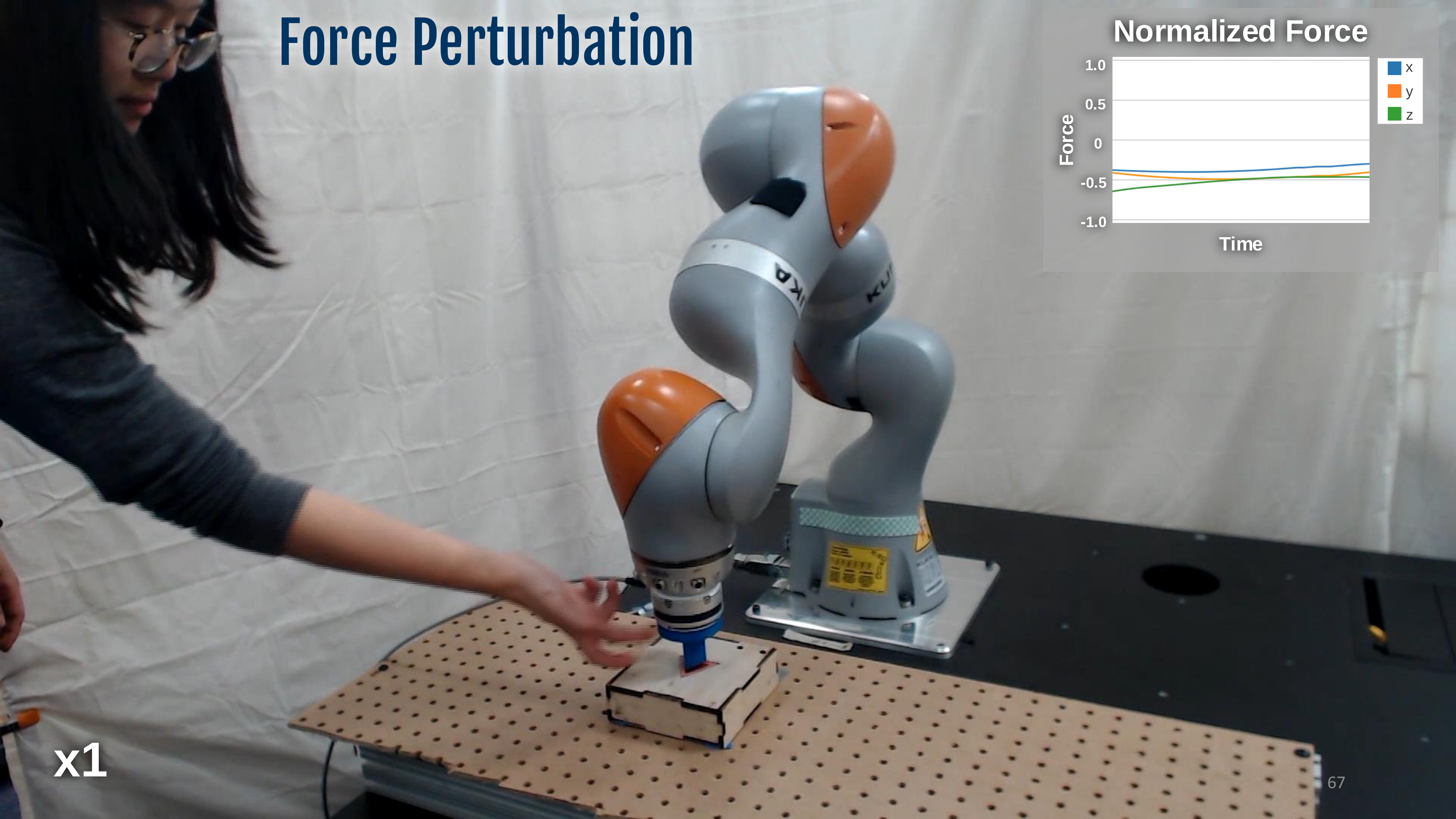
92% success rate



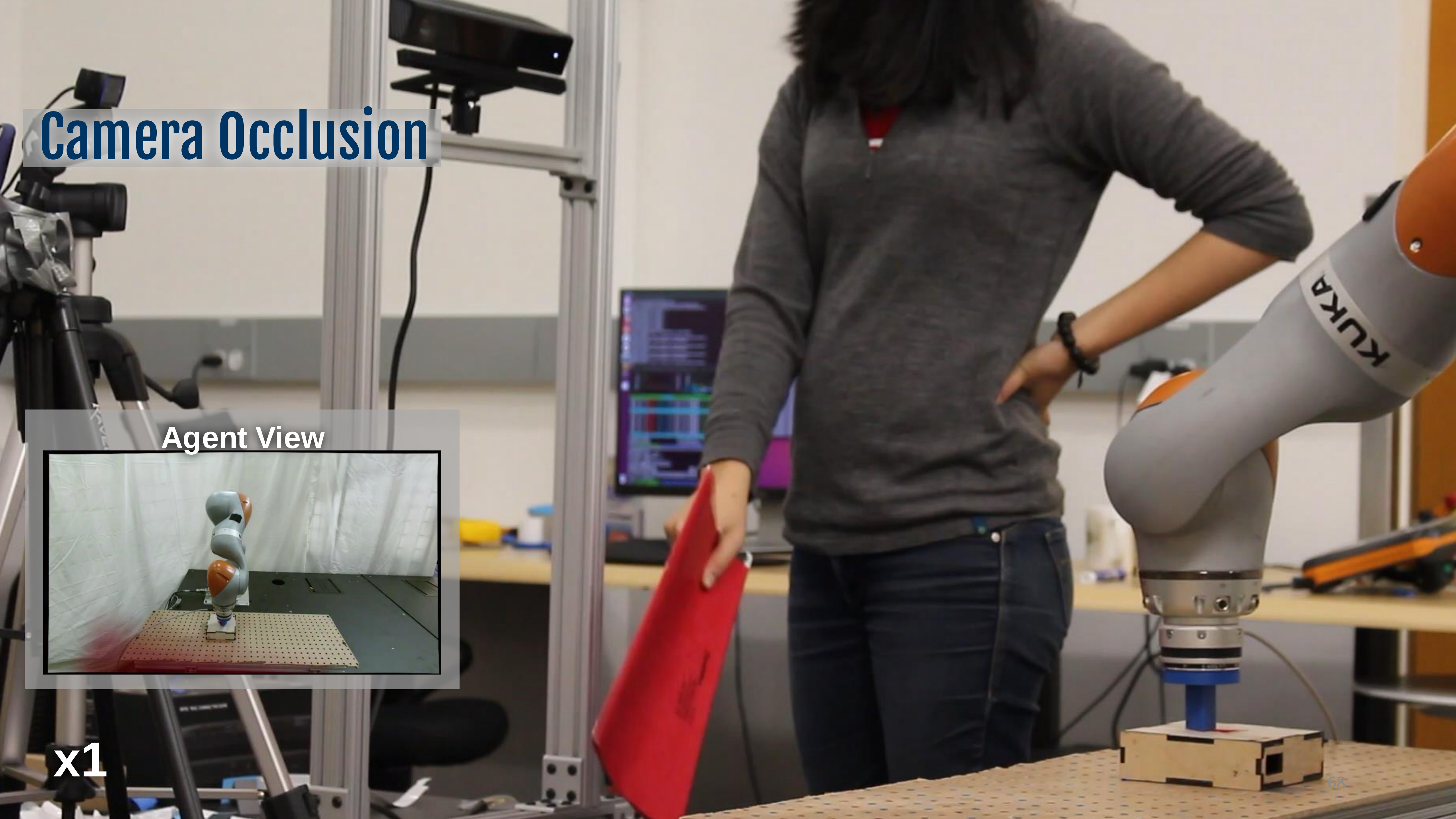
# Our multimodal policy is robust against sensor noise

1. Force Perturbation
2. Camera Occlusion

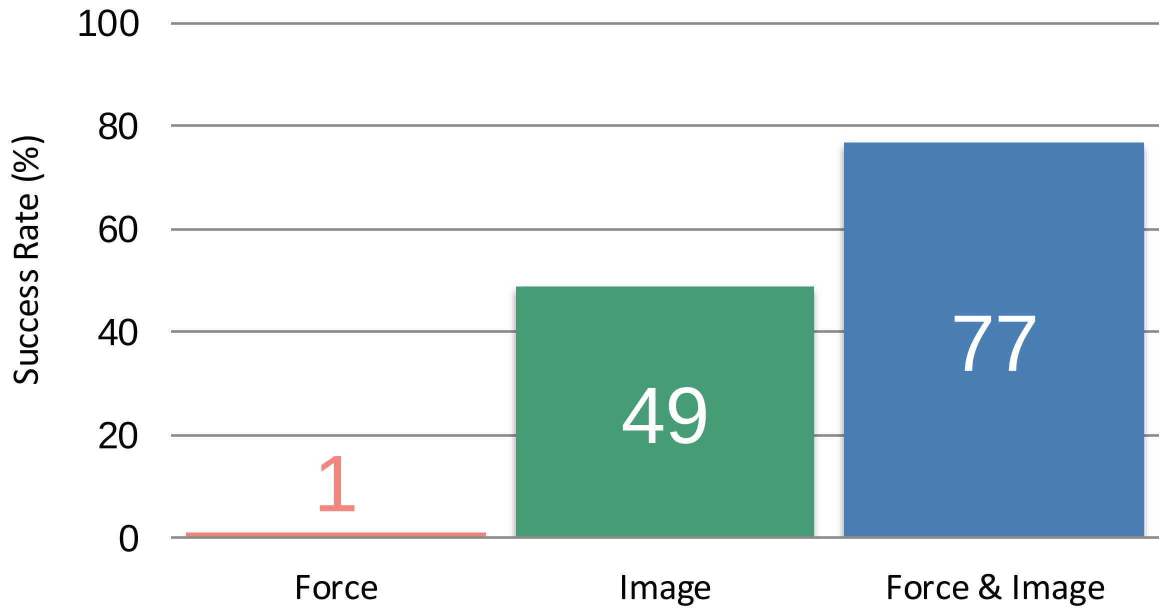
# Force Perturbation



# Camera Occlusion



# How is each modality used?




Force Only: Can’t find box

Image Only: Struggles with peg alignment

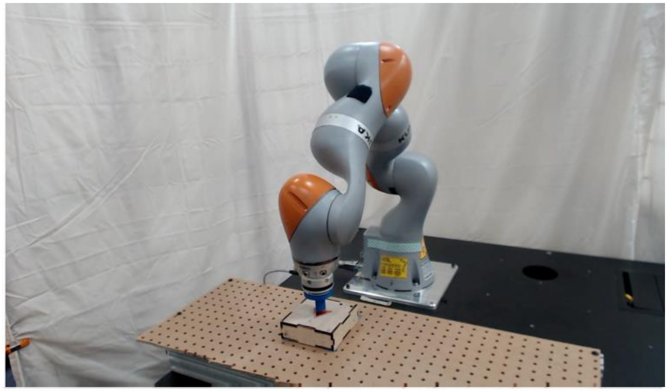
Force & Image: Can learn full task completion

Simulation Results
(Randomized box location)

# Does our representation generalize to new geometries?

# Does our representation generalize to new geometries?

92% Success Rate



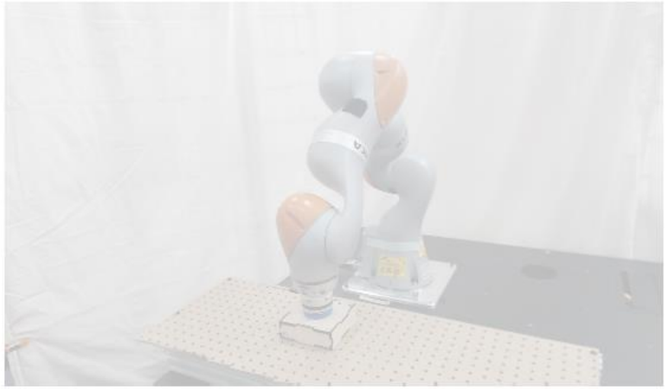
Tested on

Representation

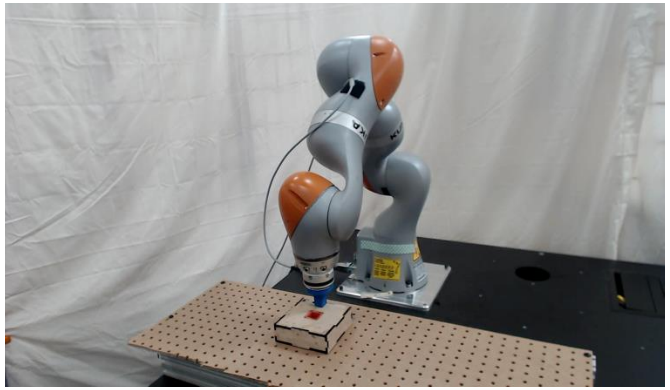
Policy

# Does our representation generalize to new policies?

92% Success Rate



62% Success Rate



Policy does not transfer

Tested on

Representation

Policy

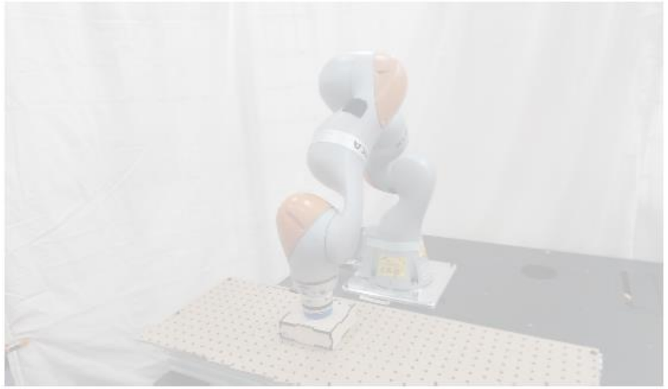
Tested on

Representation

Policy

# Does our representation generalize?

92% Success Rate

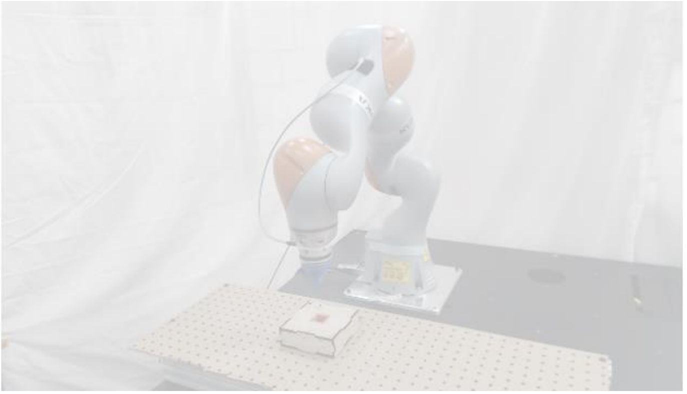


Tested on

Representation

Policy

62% Success Rate



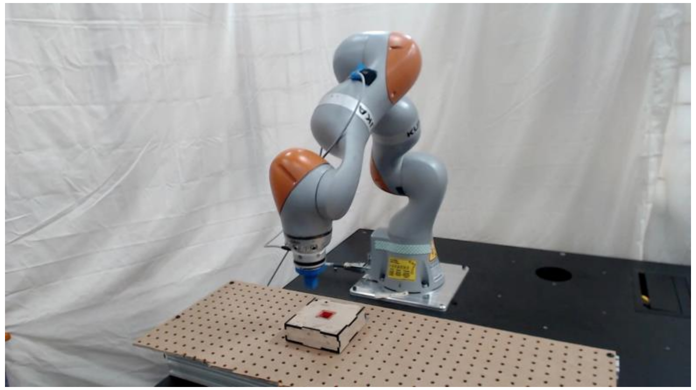
Policy does not transfer

Tested on

Representation

Policy

92% Success Rate



Representation transfers

Tested on

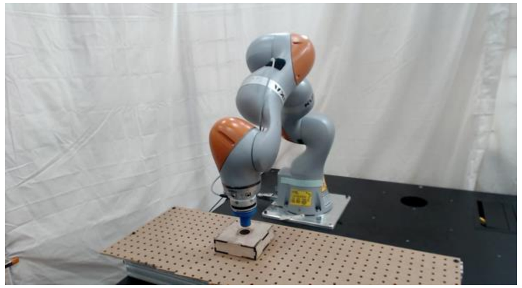
Representation

Policy

# Overview of method

## Self-supervised data collection

$o_{RGB}, o_{force}, o_{robot}$

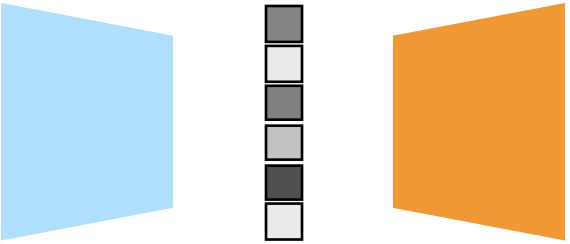


100k data points

90 minutes

## Representation learning

$f(o_{RGB}, o_{force}, o_{robot})$

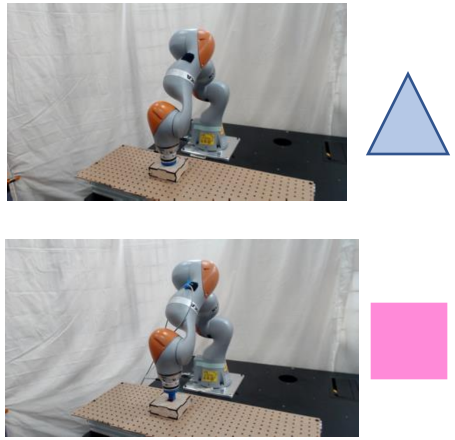


20 epochs on GPU

24 hours

## Policy learning

$\pi(f(\cdot)) = a$



Deep RL

5 hours

Lee et al. Making Sense of Vision and Touch: Self-Supervised Learning of Multimodal Representations for Contact-Rich Tasks. ICRA'19. *Best Paper Award.*

# Lessons Learned

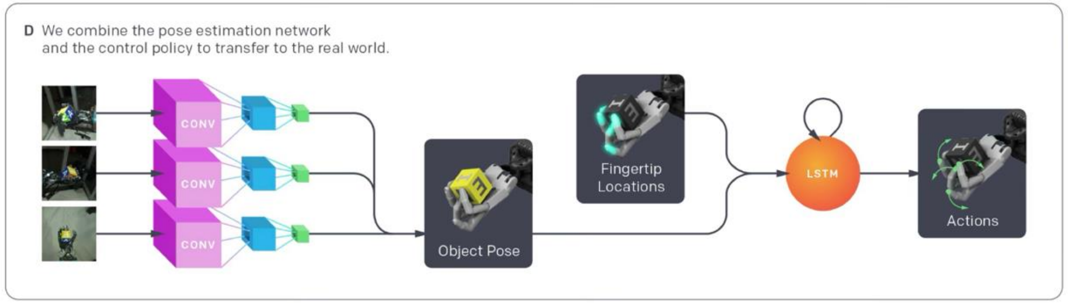
1. Self-supervision gives us rich learning objectives

2. Representation that captures concurrency and dynamics can generalize across task instances

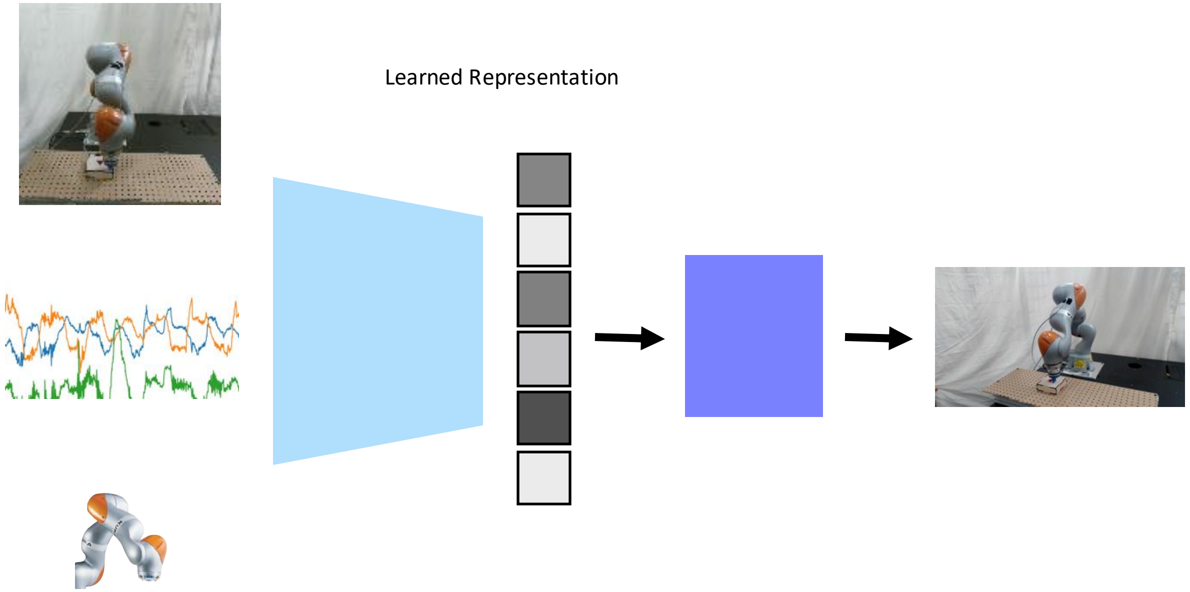
3. Our experiments show that multimodal representation leads to learning efficiency and policy robustness

# State Representation - Physically Meaningful or Learned?

Interpretable Representation




Learned Representation



# Interactive Perception

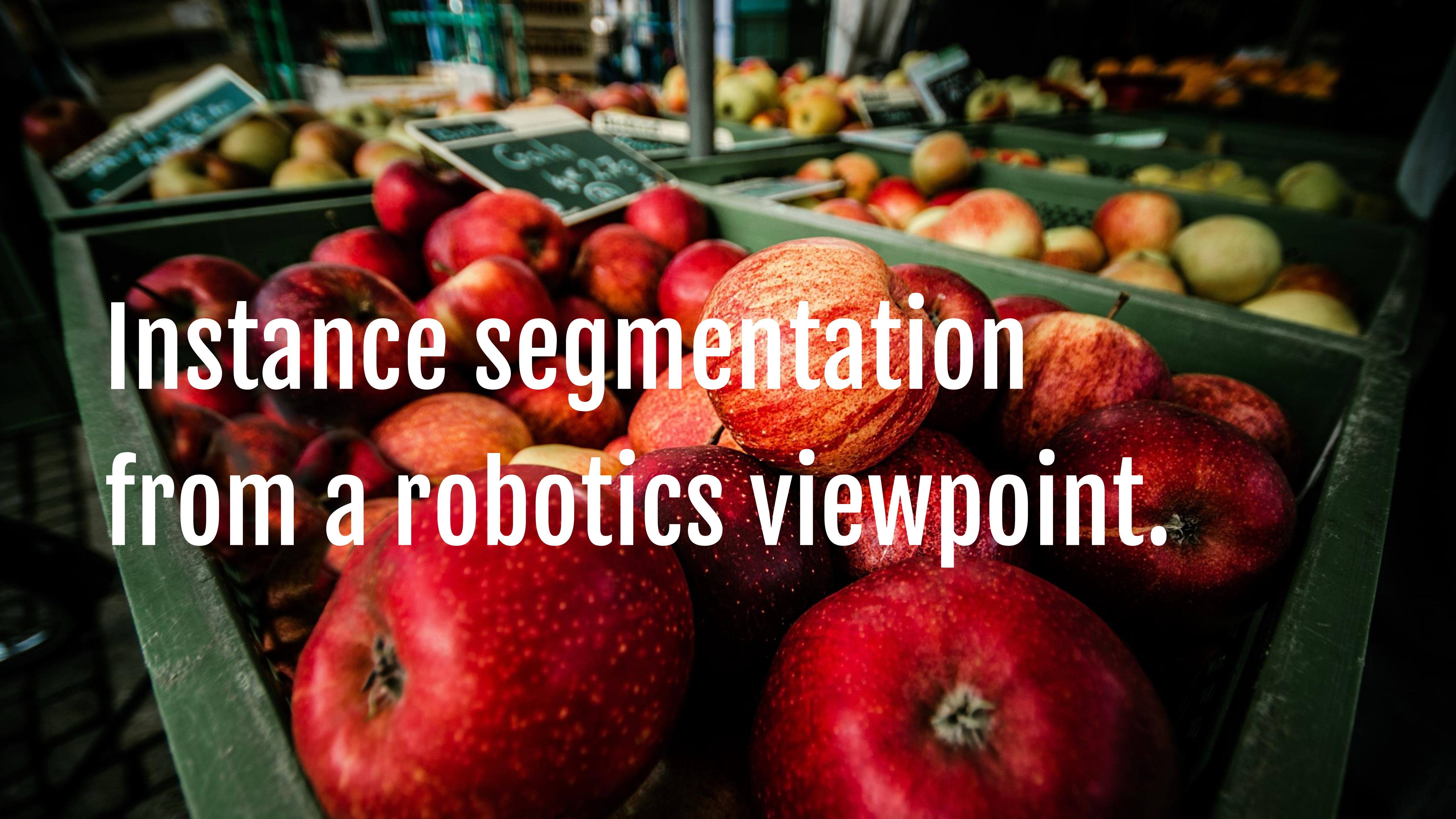
$$S \times A \times t$$

Sensory Data Actions Time

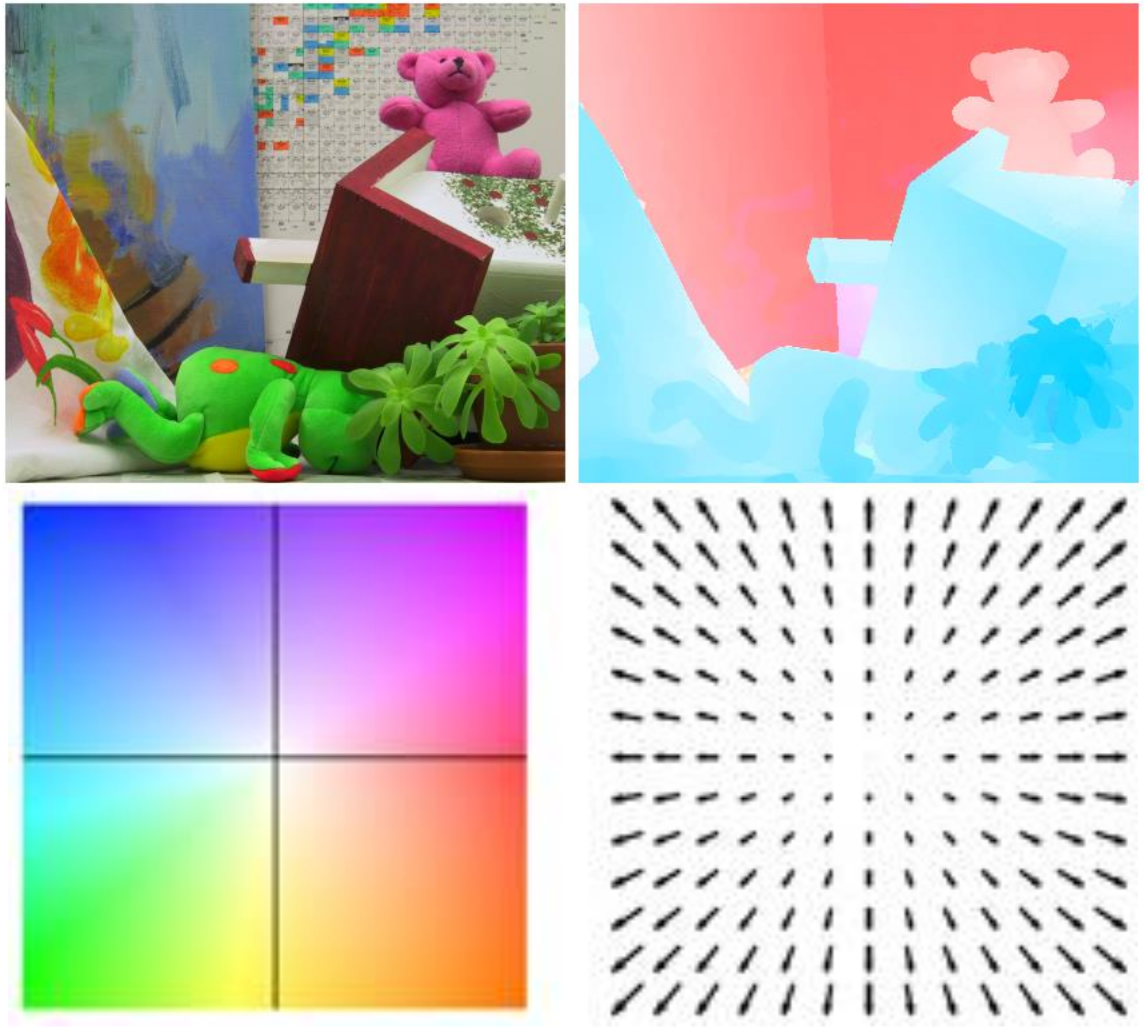
Exploiting RGB, Depth and Motion for Instance Segmentation

Selfsupervised Learning of a Multimodal Representation

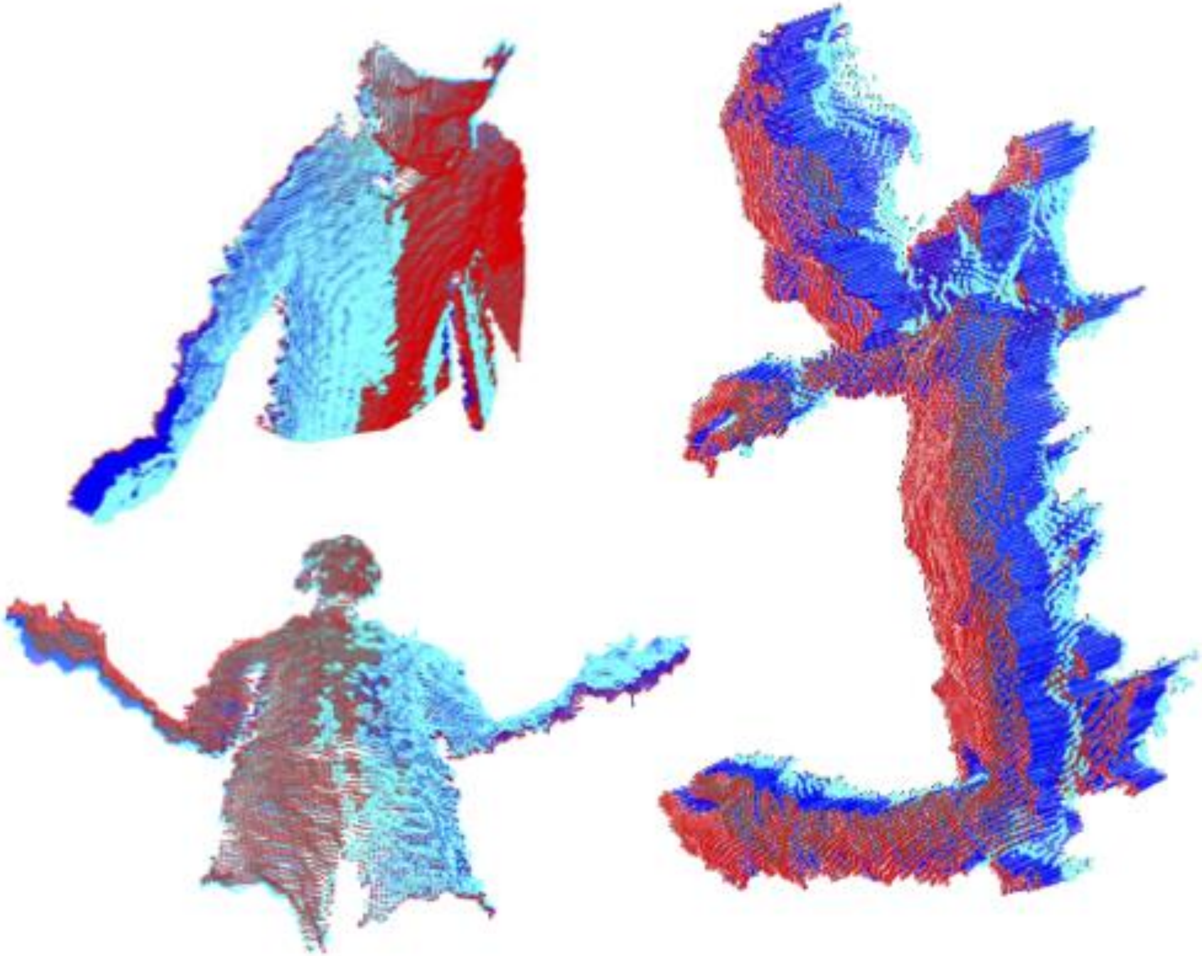
Interactive Perception - Leveraging Action in Perception and Perception in Action.
Bohg, Hausman, Sanakaran, Brock, Kragic, Schaal and Sukhatme. TRO ’17.


Instance segmentation
from a robotics viewpoint.

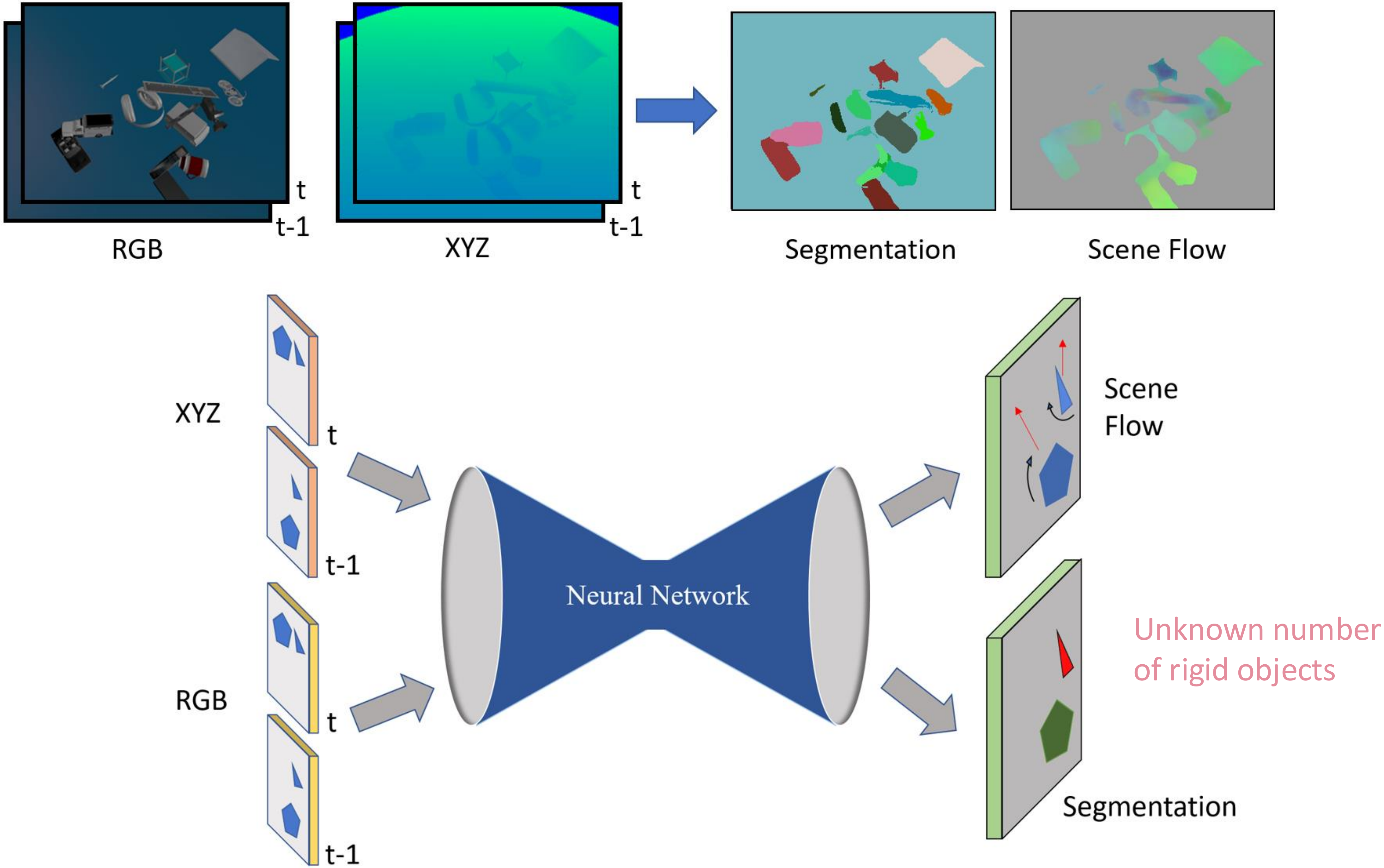
# Optical Flow - Scene Flow



A Database and Evaluation Methodology for Optical Flow. Baker et al. IJCV. 2011

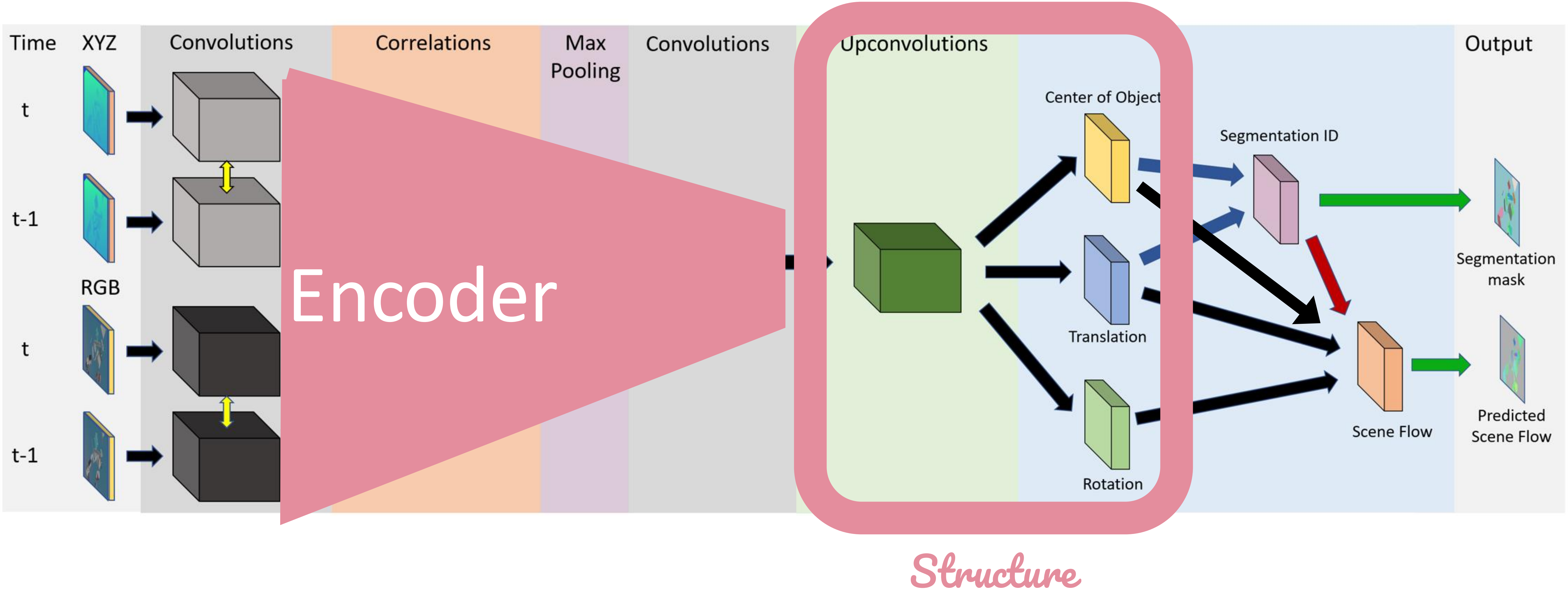


A Primal-Dual Framework for Real-Time Dense RGB-D Scene Flow. Jaimez at al. ICRA, 2015.



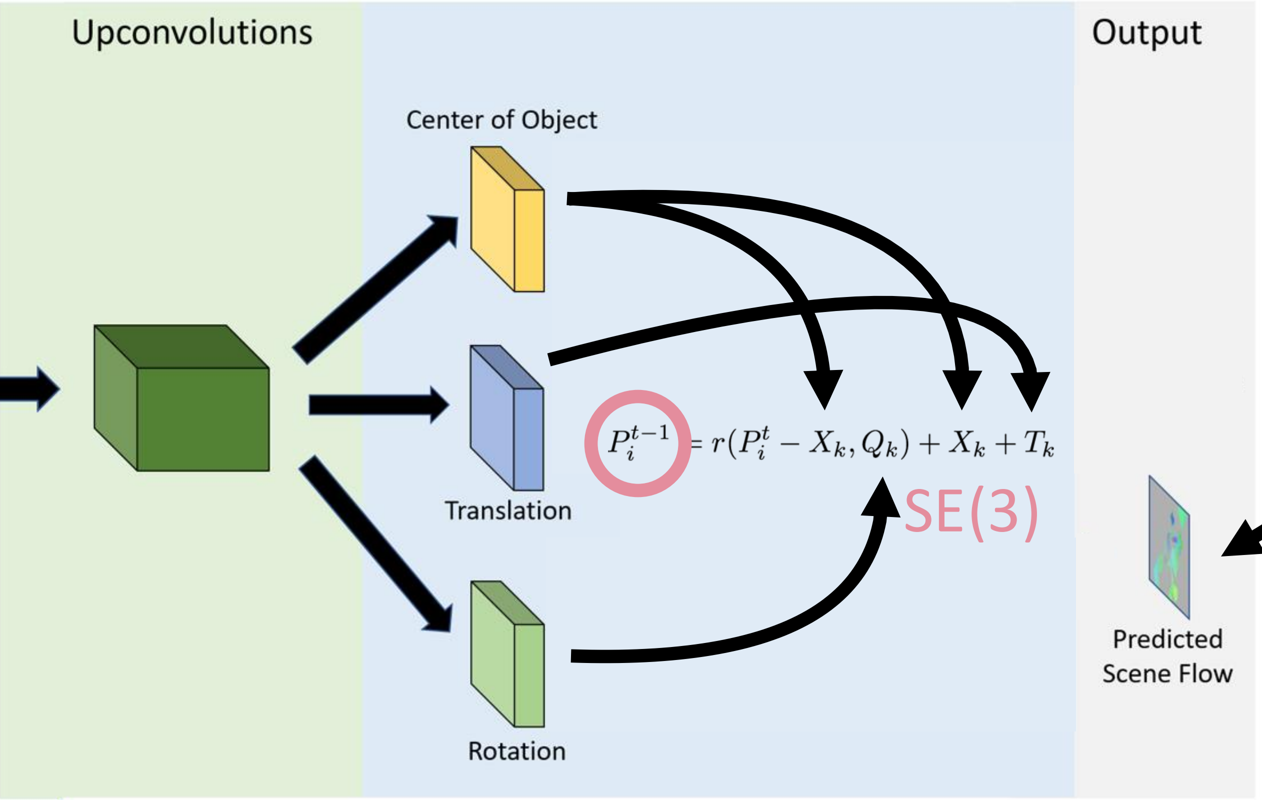

Motion-based Object Segmentation based on Dense RGB-D Scene Flow. Shao et al. Submitted to RAL + IROS. 2018. Pre-print on arXiv.

# Pixel-wise Prediction




Motion-based Object Segmentation based on Dense RGB-D Scene Flow. Shao et al. Submitted to RAL + IROS. 2018. Pre-print on arXiv.

# Computing Scene Flow



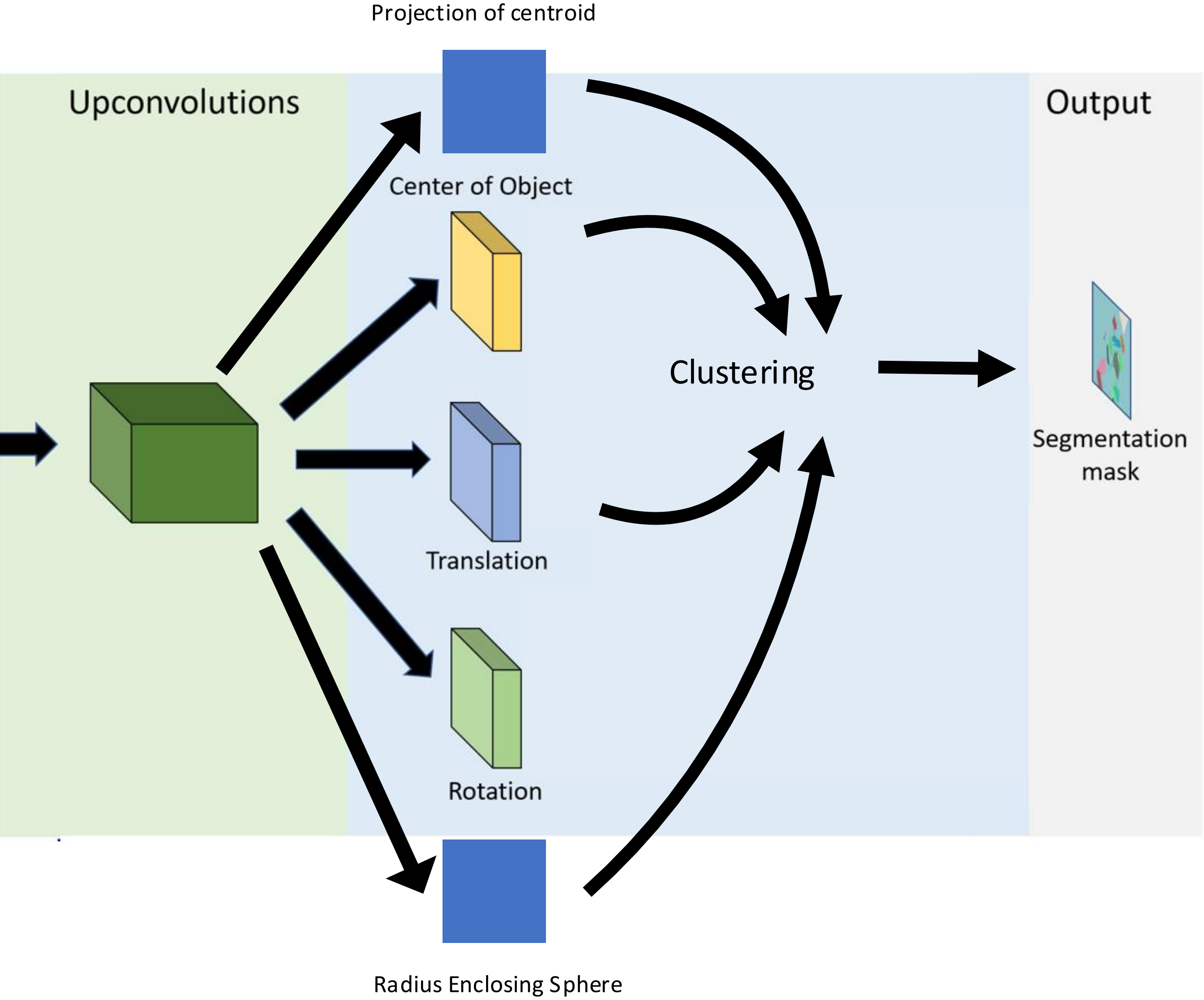

Scene Flow = Displacement of one 3D point $P_i^t$ over time

$$S_i = P_i^{t-1} - P_i^t$$

?

Motion-based Object Segmentation based on Dense RGB-D Scene Flow. Shao et al. Submitted to RAL + IROS. 2018. Pre-print on arXiv.

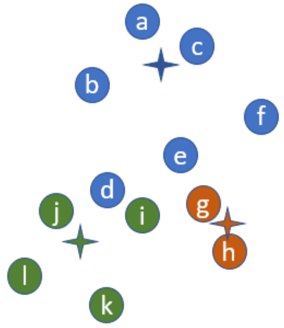
# Motion-based Object Segmentation




Pixels that move rigidly belong to one object.

Trajectory feature $\xi_k = [X_k, X_k + T_k]$

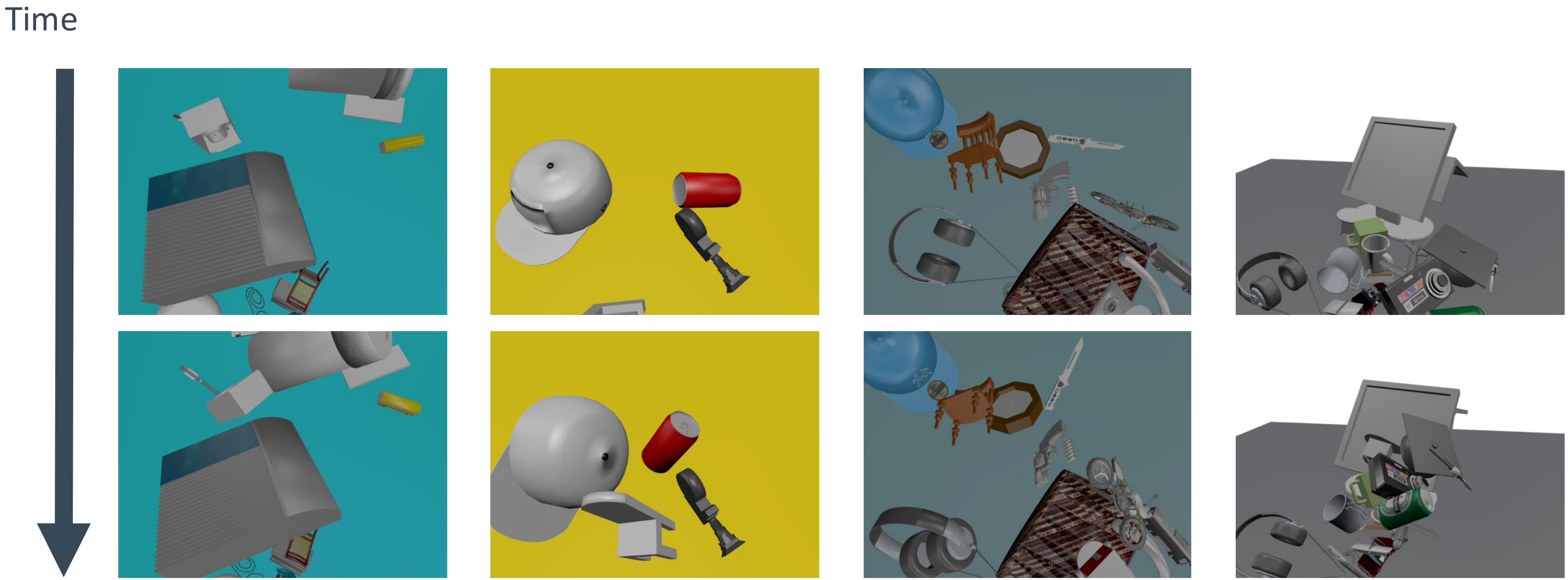
Clustering




3D World Space at Frame t

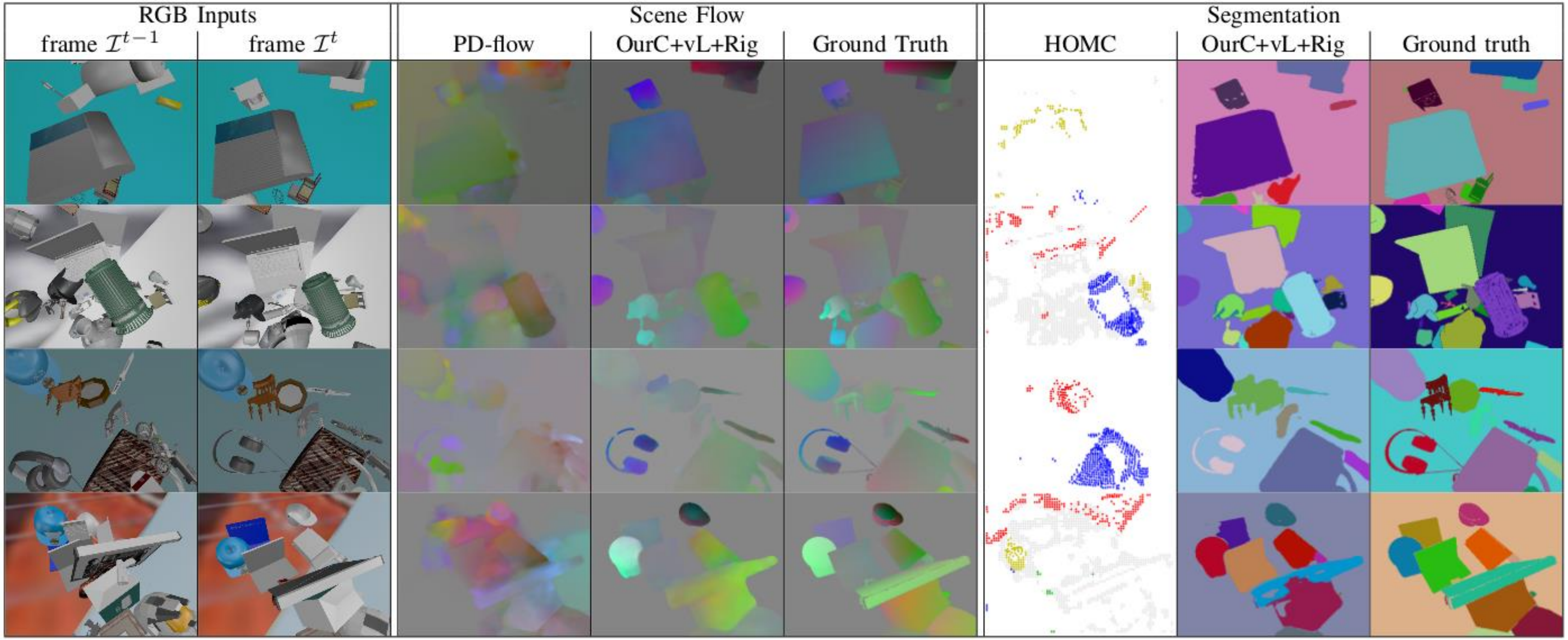
Motion-based Object Segmentation based on Dense RGB-D Scene Flow. Shao et al. Submitted to RAL + IROS. 2018. Pre-print on arXiv.

# Data Set of Frame Pairs



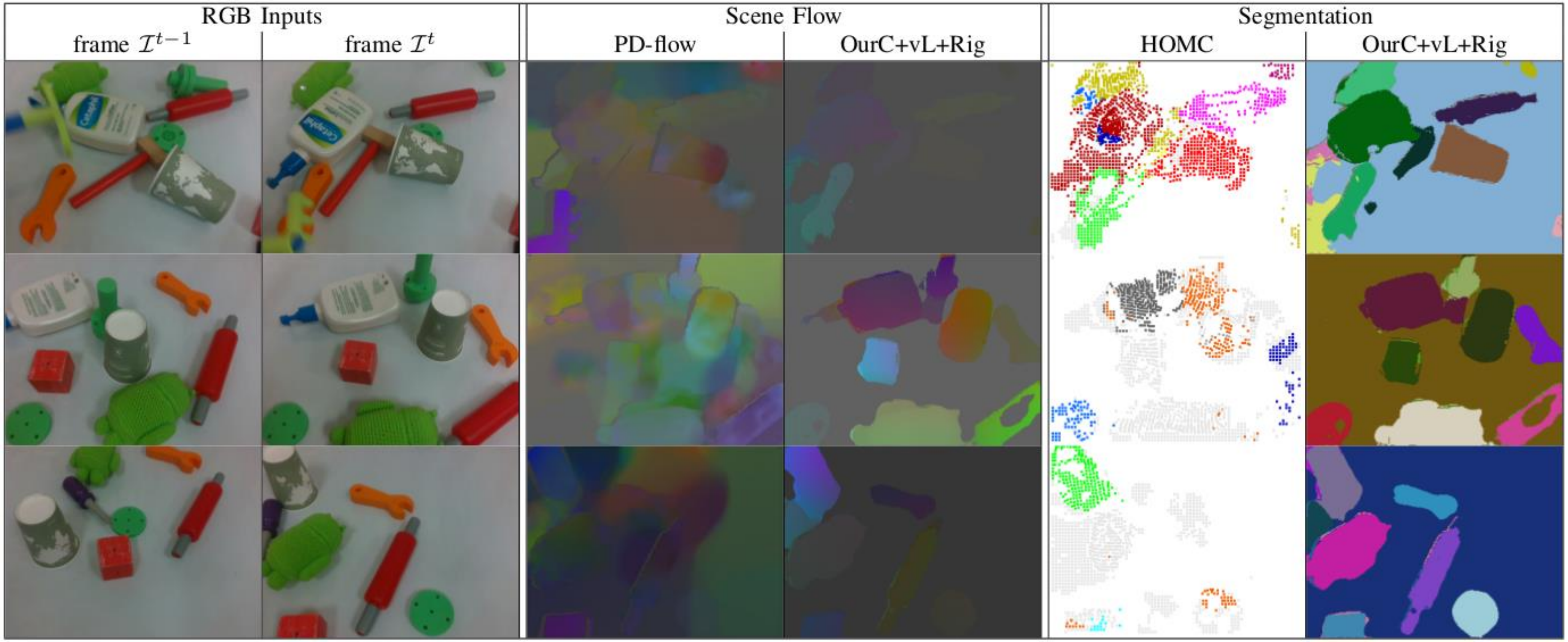

Motion-based Object Segmentation based on Dense RGB-D Scene Flow. Shao et al. Submitted to RAL + IROS. 2018. Pre-print on arXiv.

# Qualitative Results - Synthetic




Motion-based Object Segmentation based on Dense RGB-D Scene Flow. Shao et al. Submitted to RAL + IROS. 2018. Pre-print on arXiv.

# Qualitative Results - Real




Motion-based Object Segmentation based on Dense RGB-D Scene Flow. Shao et al. Submitted to RAL + IROS. 2018. Pre-print on arXiv.

# Interactive Perception

$$\boxed{S} \times A \times \boxed{t}$$

Sensory Data    Actions    Time

Exploiting RGB, Depth and Motion for Instance Segmentation

Interactive Perception - Leveraging Action in Perception and Perception in Action.
Bohg, Hausman, Sanakaran, Brock, Kragic, Schaal and Sukhatme. TRO ’17.

# Interactive Perception

$$\boxed{S \times A \times t}$$

Sensory Data Actions Time

Homework 3 - Problem 3 Learning Intuitive Physics

Interactive Perception - Leveraging Action in Perception and Perception in Action.
Bohg, Hausman, Sanakaran, Brock, Kragic, Schaal and Sukhatme. TRO ’17.

# Conclusions

- Interaction generates a rich sensory signal that eases perception.
- Action-conditional, multi-modal representation helps contact-rich manipulation and generalizes.

# Next time

Wednesday – Imitation Learning I – Dorsa Sadigh