# Visual Affordance Learning for Robot Manipulation

Yuke Zhu

Toyota Research Institute, August 2021

TEXAS
The University of Texas at Austin

# Traditional Form of Robot Automation


Structured Environments

Fixed Set of Tasks

Pre-programmed Procedures

# General-Purpose Robot Autonomy: Our North Star Goal

Natural Environments

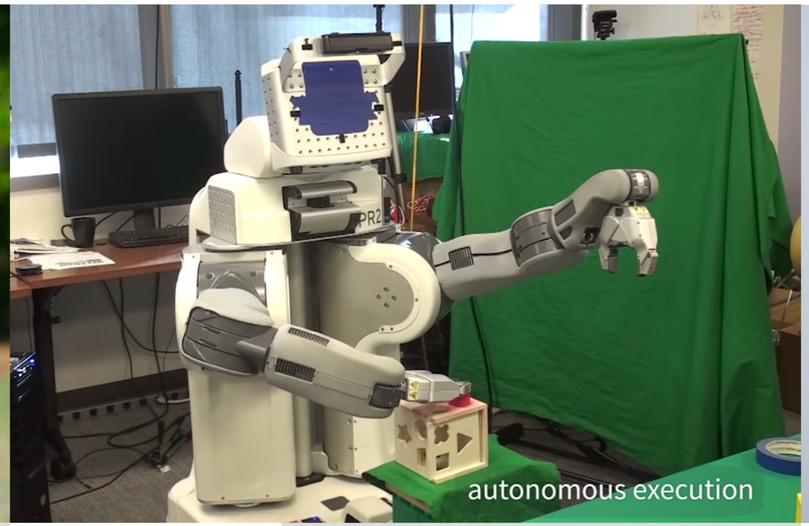
Ever-Changing Tasks

Human Involvement

# The Perception-Action Loop

A key challenge in **robot autonomy** is to close the **perception-action loop**.

[Sa et al. IROS 2014]



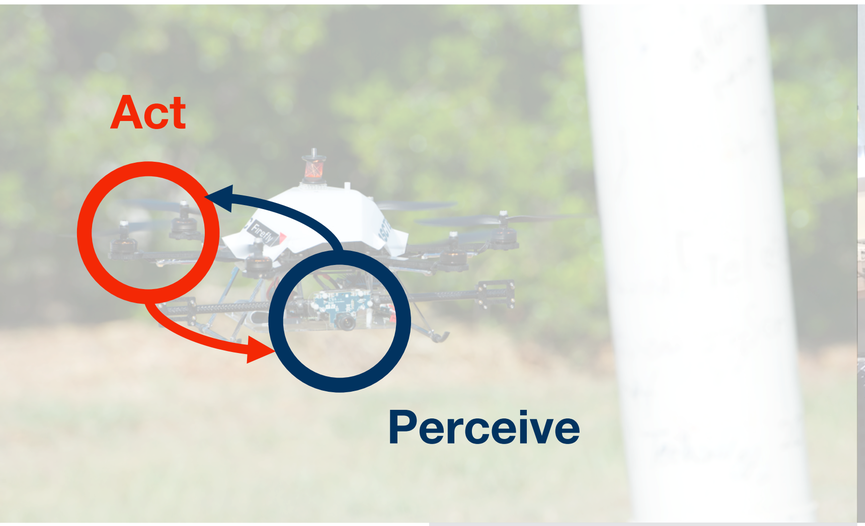

[Levine et al. JMLR 2016]

[Bohg et al. ICRA 2018]

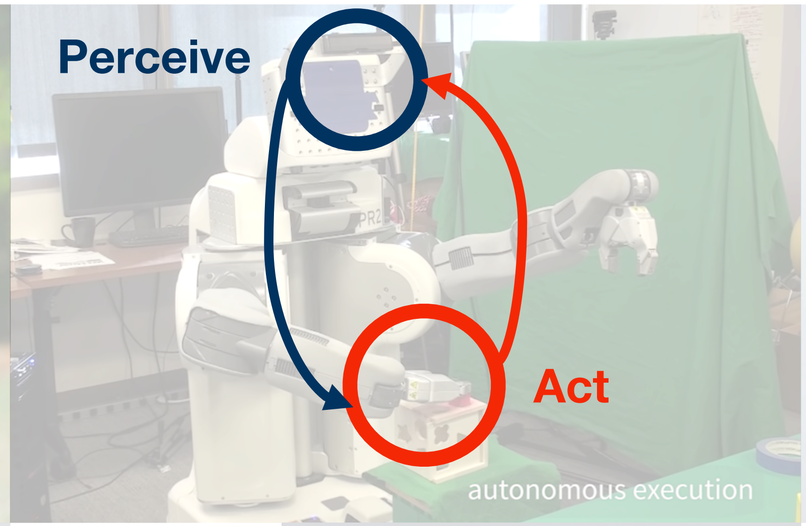
[Gibson 1979; Bajcsy 1988; Ballard 1991; Espiau et al. 1992; Hutchinson et al. 1996; Hamel & Mahony 2002; Kragic & Christensen 2002; Jonschkowski & Brock 2015; Levine et al. 2016; Agrawal et al. 2016; Bojarski et al. 2016; Finn & Levine 2017; Florence et al. 2018]

# The Perception-Action Loop

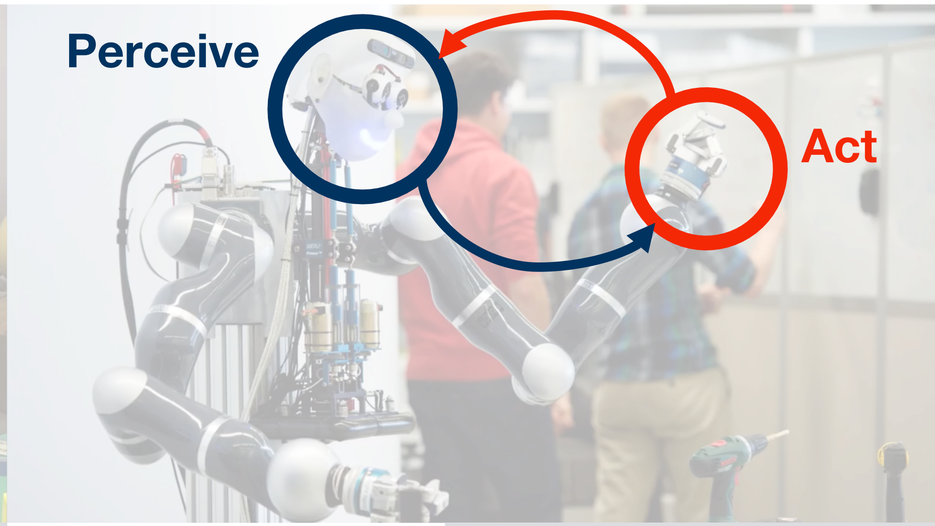
A key challenge in **robot autonomy** is to close the **perception-action loop**.




[Sa et al. IROS 2014]



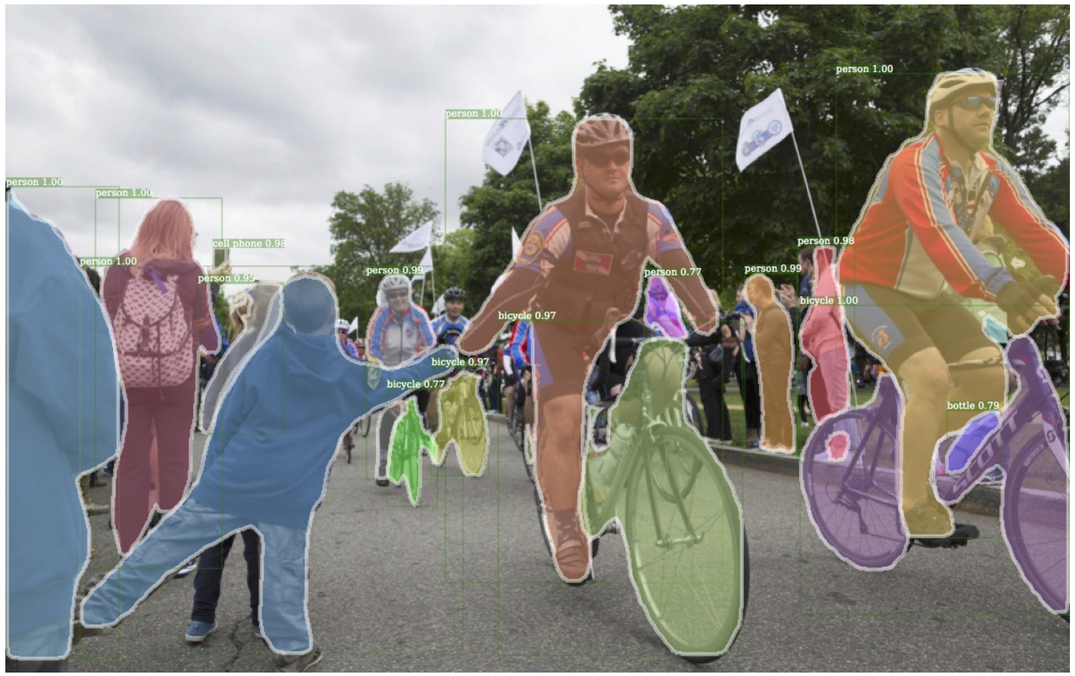

[Levine et al. JMLR 2016]




[Bohg et al. ICRA 2018]

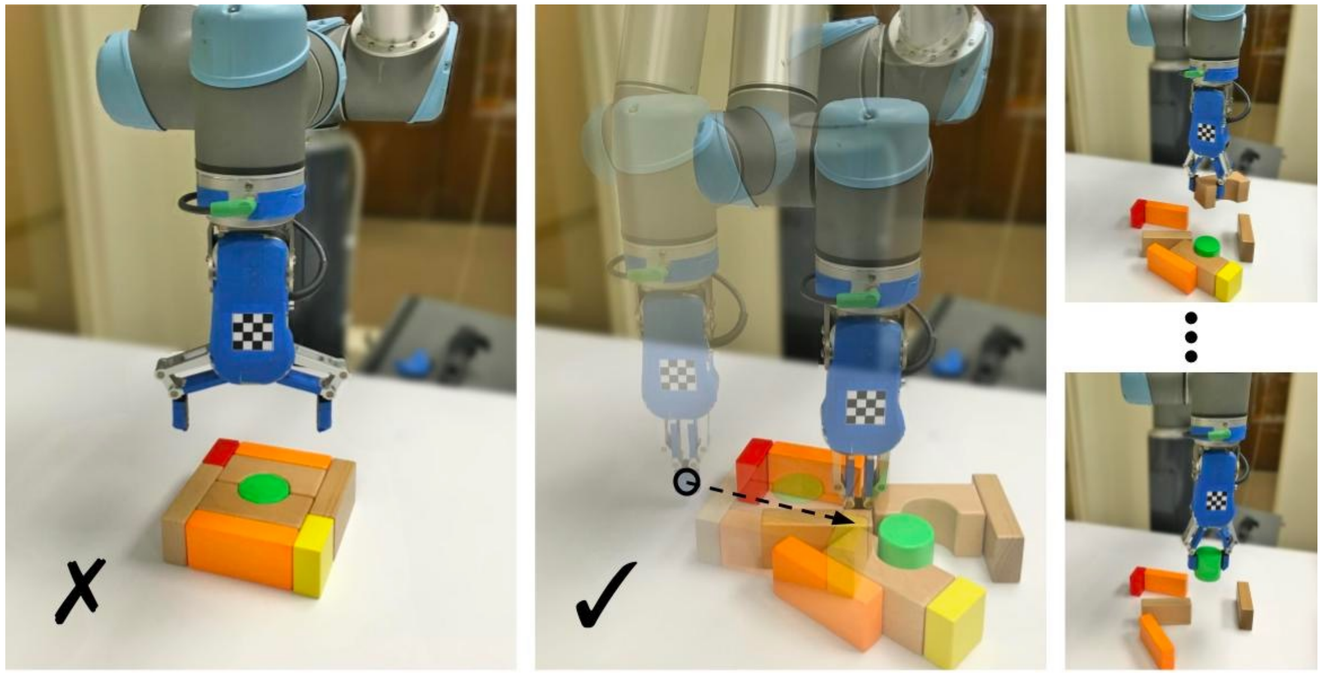
[Gibson 1979; Bajcsy 1988; Ballard 1991; Espiau et al. 1992; Hutchinson et al. 1996; Hamel & Mahony 2002; Kragic & Christensen 2002; Jonschkowski & Brock 2015; Levine et al. 2016; Agrawal et al. 2016; Bojarski et al. 2016; Finn & Levine 2017; Florence et al. 2018]

# The Perception-Action Loop




[Detectron - Facebook AI Research]

Conventional Computer Vision




[Zeng et al., IROS 2018]

Physically-Grounded Robot Perception

# Object Perception

## “What makes a chair a chair?”

**Figure 1** This is a chair as conceptualized by **Ikea**

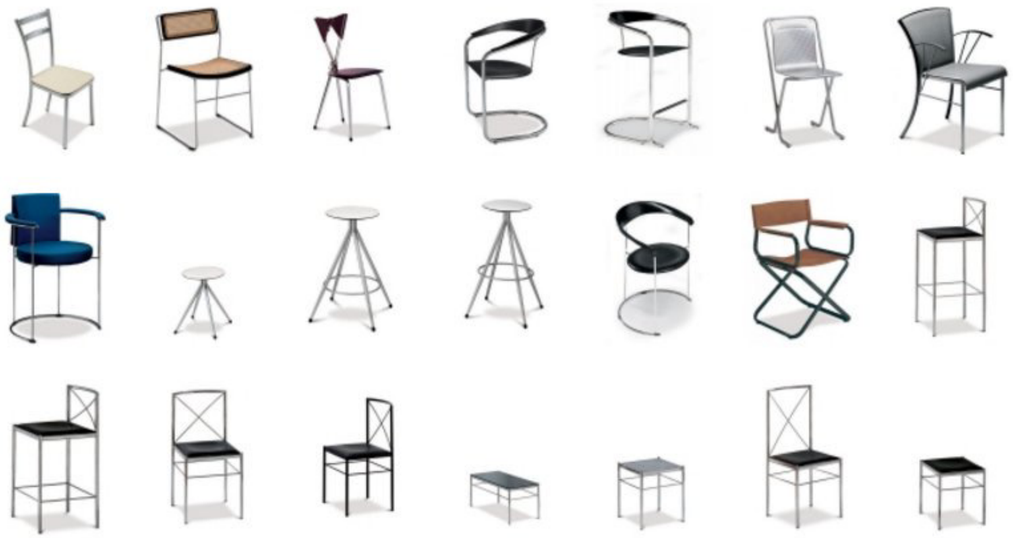
1. A chair has 4 legs... although I can imagine a chair with 3 or even 5
2. A chair usually has arms and a back... although a stool fits in the chair category without either
3. It must be raised above the ground... although things that aren't chairs also do this in a similar way...


By visual attributes

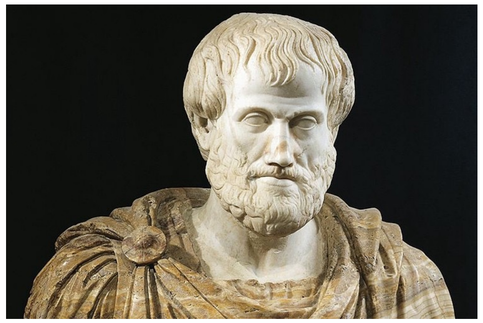
https://www.theuncomfortable.com/

# Object Perception

"What makes a chair a chair?"



"What makes a chair a chair is its ability to be sat on."



Aristotle's *Ontology*

Zhu Xi's *Li* (理)

By functional properties

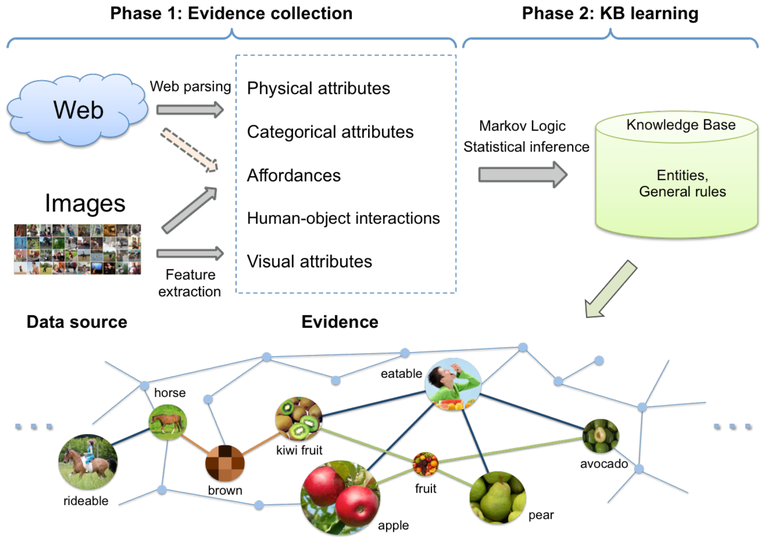
# Object Perception

"What makes a chair a chair?"

**Affordances** are *possibilities for actions* that the environment *affords* to the agent. (Gibson, 1977)

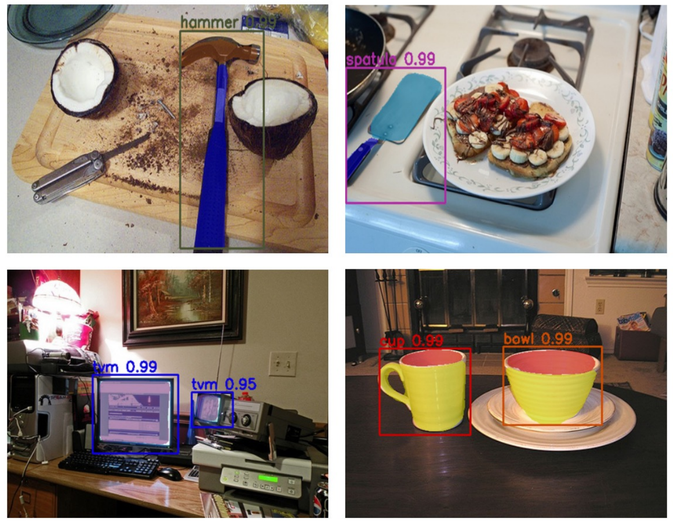
[Gibson 1979; Kirklik 1993; Zaff 1995; Stoytchev 2015; Amant 1999; Bousmalis 2018; Detry 2011; Zhu et al. 2017; Mahler 2017; Nagarajan 2020; Ugur 2007; Dang 2020; Fang 2018; Song 2010; Zeng 2018; Abel 2014; Abel 2015; Cruz 2016; Khetarpal 2020; Ardon 2020; Mandikal 2020]

# Affordance Learning in Robotics



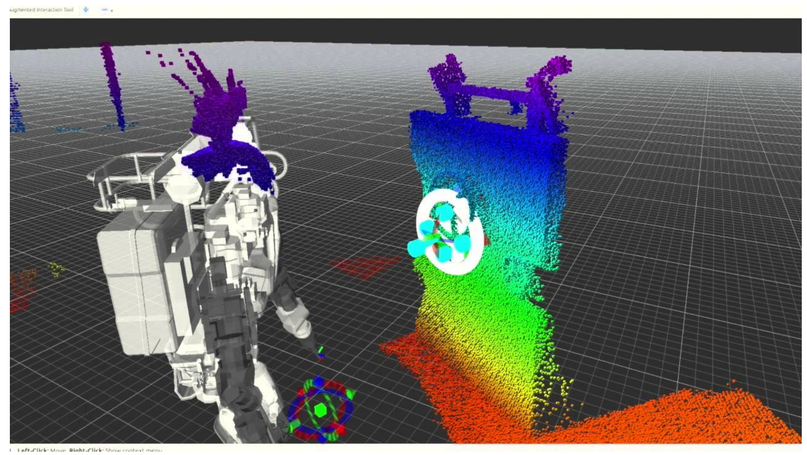

Semantic Affordance

[Zhu et al. 2014, Varadarajan & Vincze 2012]




Detection & Segmentation

[Katz et al. 2013; Do et al. 2018]



Affordance Template

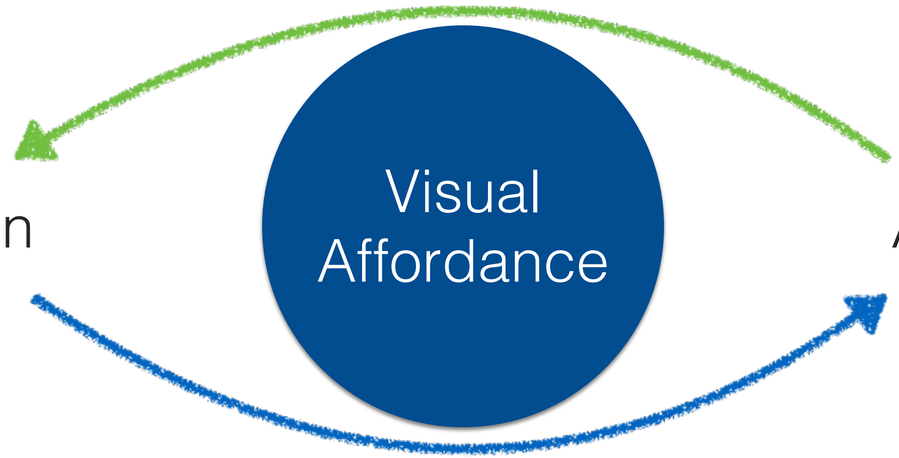
[Hart et al. 2015; Pohl et al. 2020]

Prior methods focus on learning from human supervision and building staged pipelines.

[Gibson 1979; Kirklik 1993; Zaff 1995; Stoytchev 2015; Amant 1999; Bousmalis 2018; Detry 2011; Zhu et al. 2017; Mahler 2017; Nagarajan 2020; Ugur 2007; Dang 2020; Fang 2018; Song 2010; Zeng 2018; Abel 2014; Abel 2015; Cruz 2016; Khetarpal 2020; Ardon 2020; Mandikal 2020]

# Visual Affordance Learning

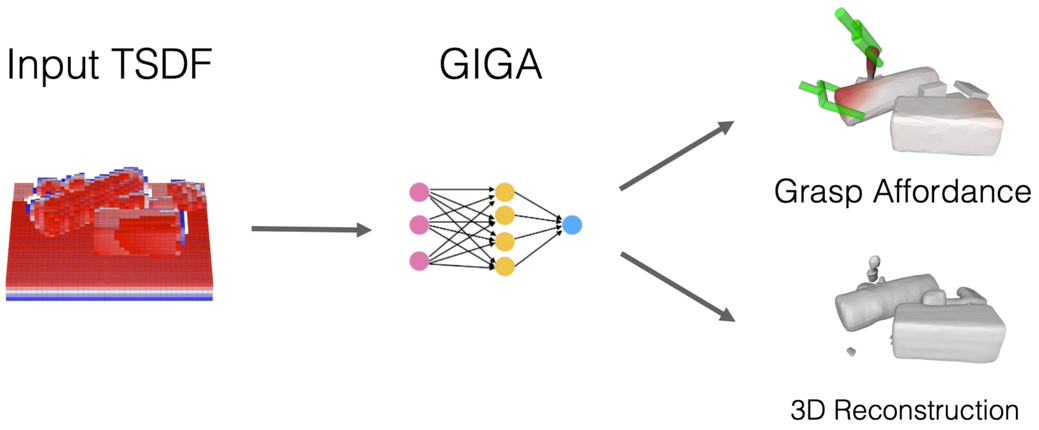
Bridging **visual perception** and **robot action** through **visual affordance**




# Visual Affordance Learning

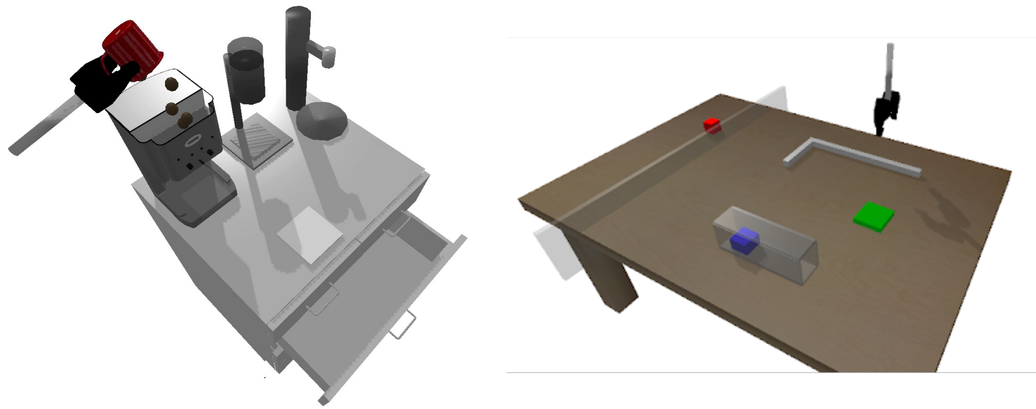
Bridging **visual perception** and **robot action** through **visual affordance**

## Learning grasp affordance with implicit representations




["GIGA" Jiang et al. RSS 2021]

## Planning long-horizon manipulation tasks with skill affordances

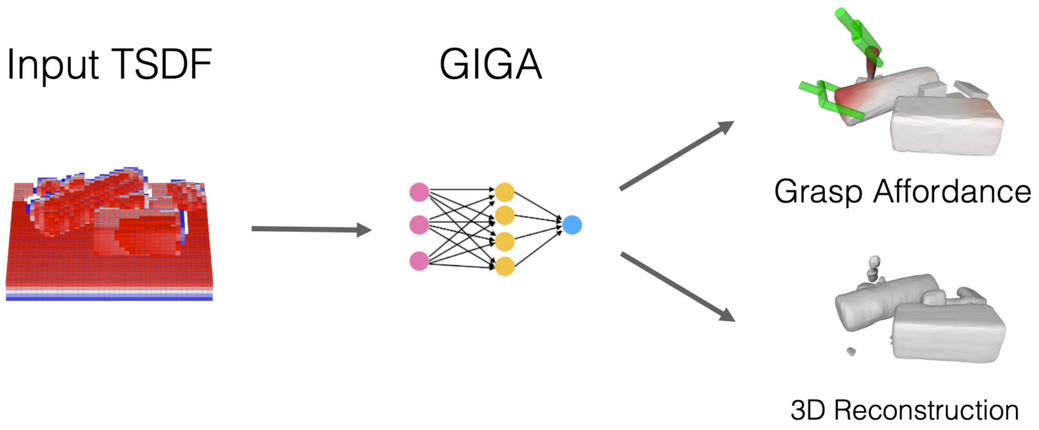


["Deep Affordance Foresight" Xu et al. ICRA21]

# Visual Affordance Learning

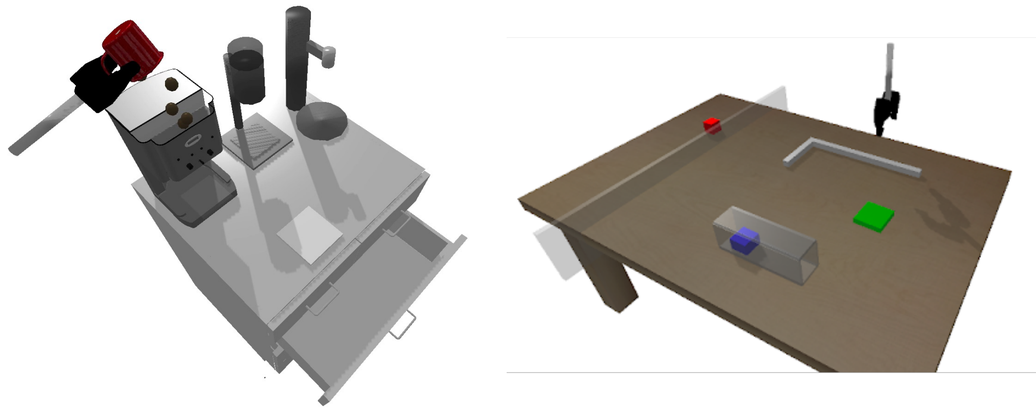
Bridging **visual perception** and **robot action** through **visual affordance**

## Learning grasp affordance with implicit representations




["GIGA" Jiang et al. RSS 2021]

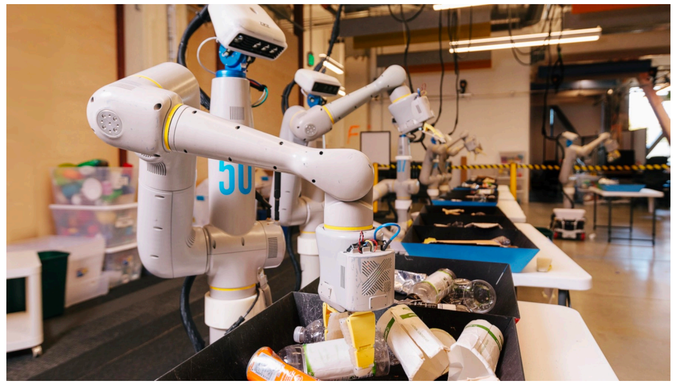
## Planning long-horizon manipulation tasks with skill affordances



["Deep Affordance Foresight" Xu et al. ICRA21]

# 6-DoF Robotic Grasping in Clutter

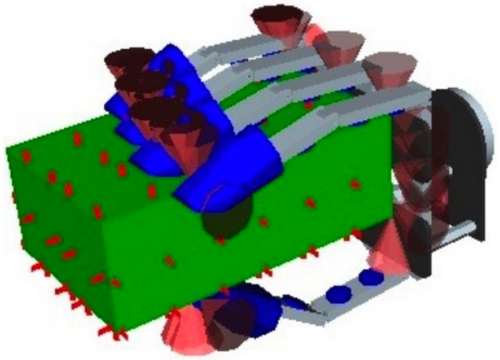




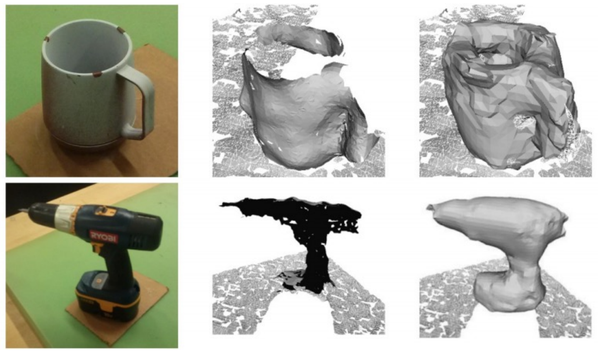
- Important modules in robot manipulation
  - Bin Picking
  - Part Assembly
  - Logistics

- **Input:** Partial point cloud of workspace
- **Output:** 6-DoF grasp pose (3D position and orientation)

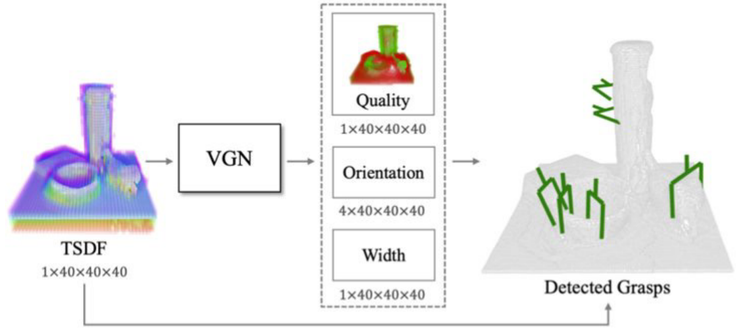
# How to Predict Grasps



[Miller et al. 2003, Goldfeder et al. 2007,
Hübner et al. 2008, Diankov et al. 2008]



[Bohg et al. 2011, Varley et al. 2017,
Lundell et al. 2019]




[Mahler et al. 2017, Morrison et al. 2018,
Liang et al. 2019, Breyer et al. 2020]

## Geometry Analysis

- Analytical solution
- Require full 3D model

## Reconstruction → Grasp

- Working with visual input
- Information bottleneck

## End-to-End Deep Learning

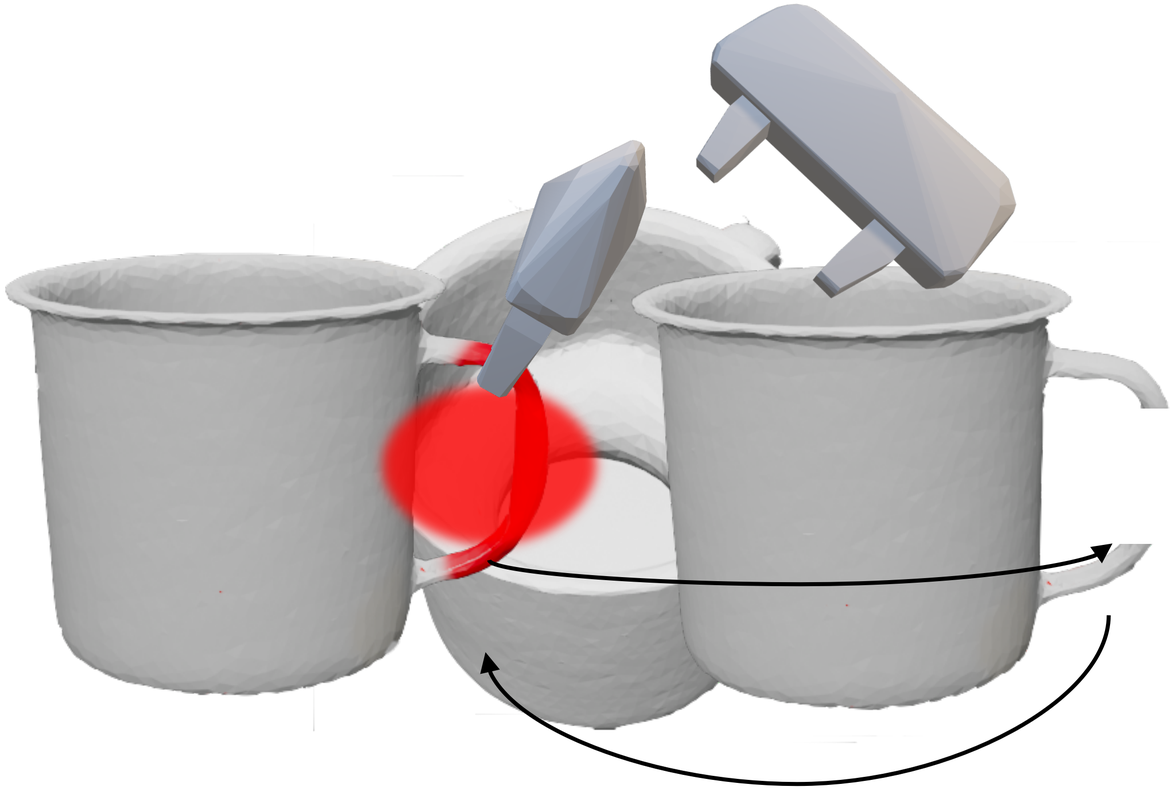
- High grasp performance
- Limited generalization

# Affordance and geometry reasoning are not isolated

Likelihood of grasp success
and grasp parameters

## Affordance

Predict affordance of
reconstructed part

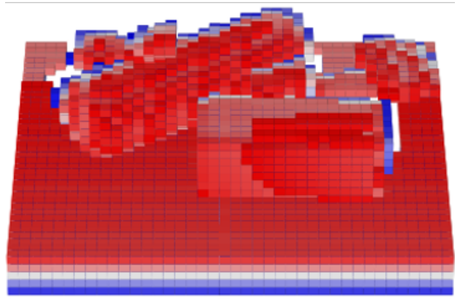


## Geometry

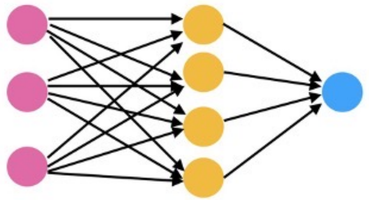
Reconstruct
graspable region

# Overview

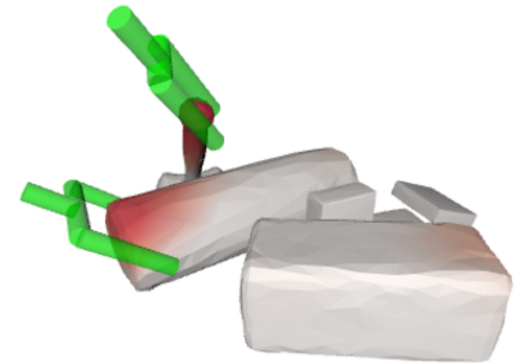
Input TSDF

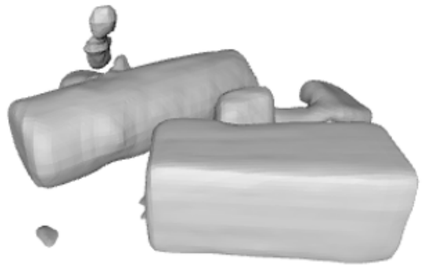


GIGA



Grasp affordance

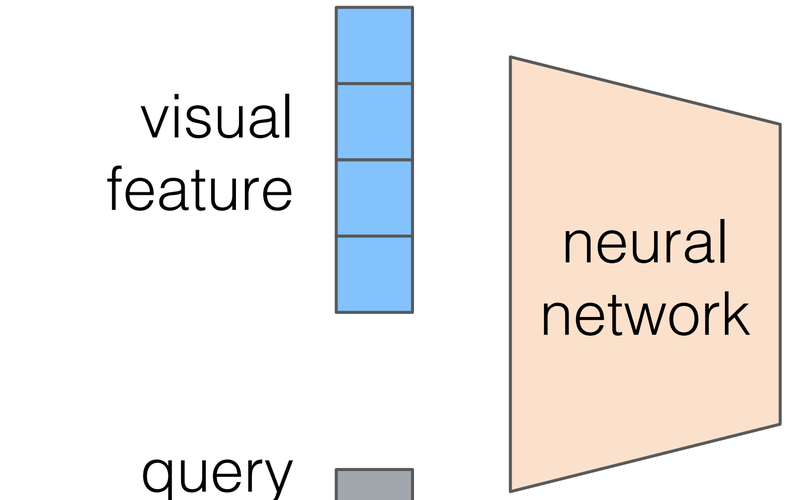


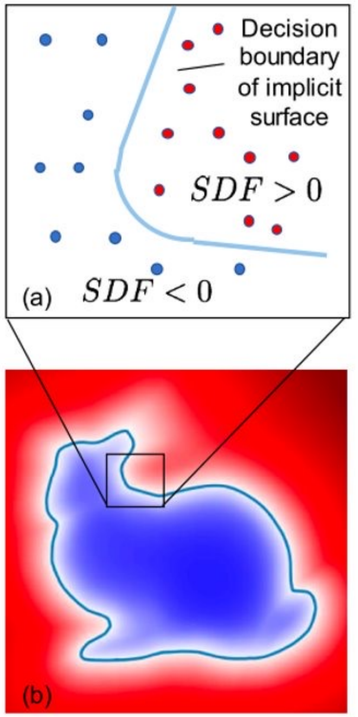


3D reconstruction

# Implicit Neural Representations for 3D Shapes

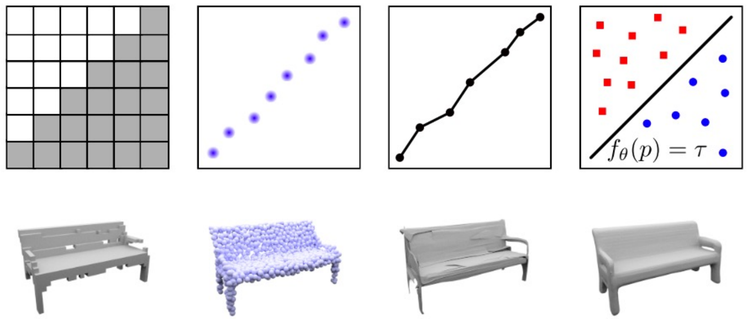
- Functions that map from coordinate to quantities (SDF, occupancy).
- Functions are parametrized with neural networks.
- Shape bound is defined by level set of the parametrized functions.



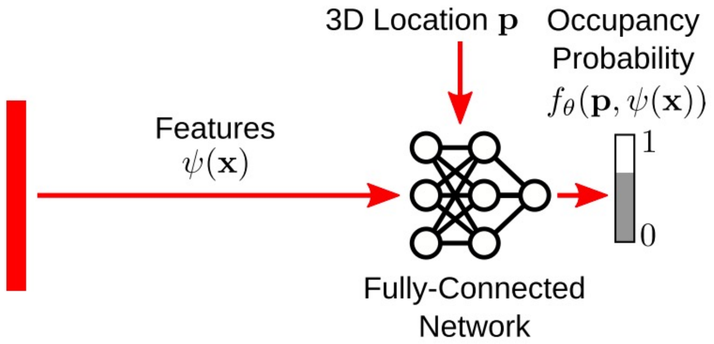

DeepSDF, Park et al. CVPR 2019

# Implicit Neural Representations: Advantages



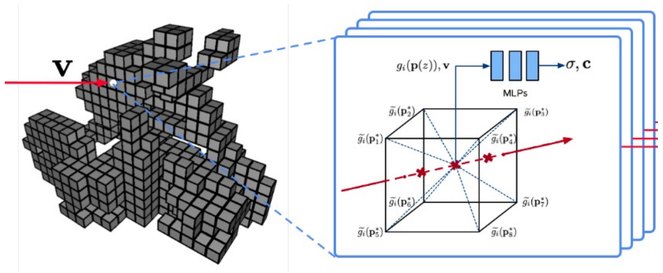

OccNets, Mescheder et al. CVPR 2019

- Continuous & memory-efficient




OccNets, Mescheder et al. CVPR 2019

- End-to-end differentiable

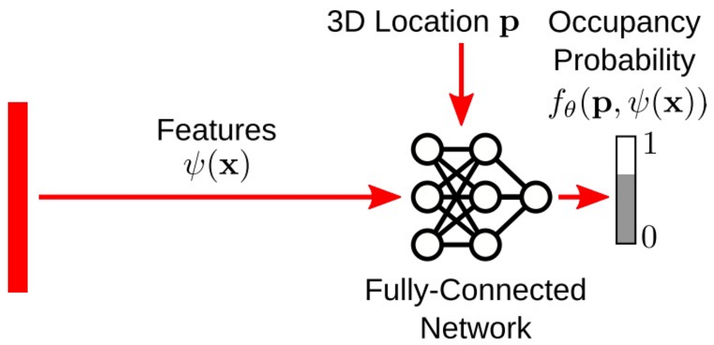



NSVF, Liu et al. NeurIPS 2020

- Adaptively allocating representation resources

# Structured Implicit Functions

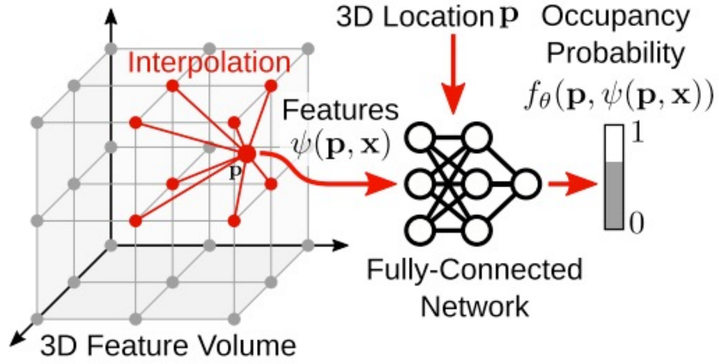
Global implicit function




OccNets, Mescheder et al. CVPR 2019

- Single global feature
- Implicit function conditioned on global feature
- Overly smooth reconstruction

Structured implicit function

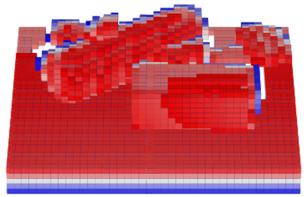



ConvONets, Peng et al. ECCV 2020

- Structured feature grid
- Local features linearly sampled from the feature grid
- Fine-grained local details

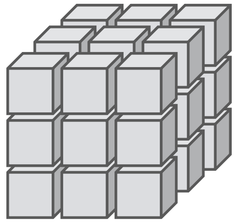
# GIGA Architecture

Input TSDF



3D Conv

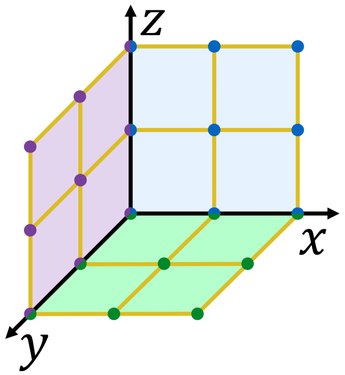
3D feature grid



Projection

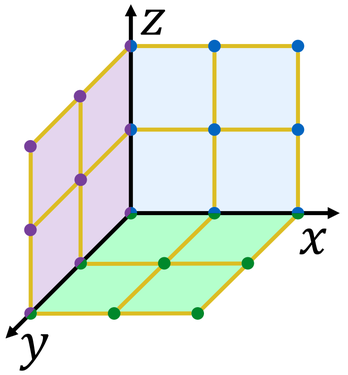
Aggregation

Projected 2D features

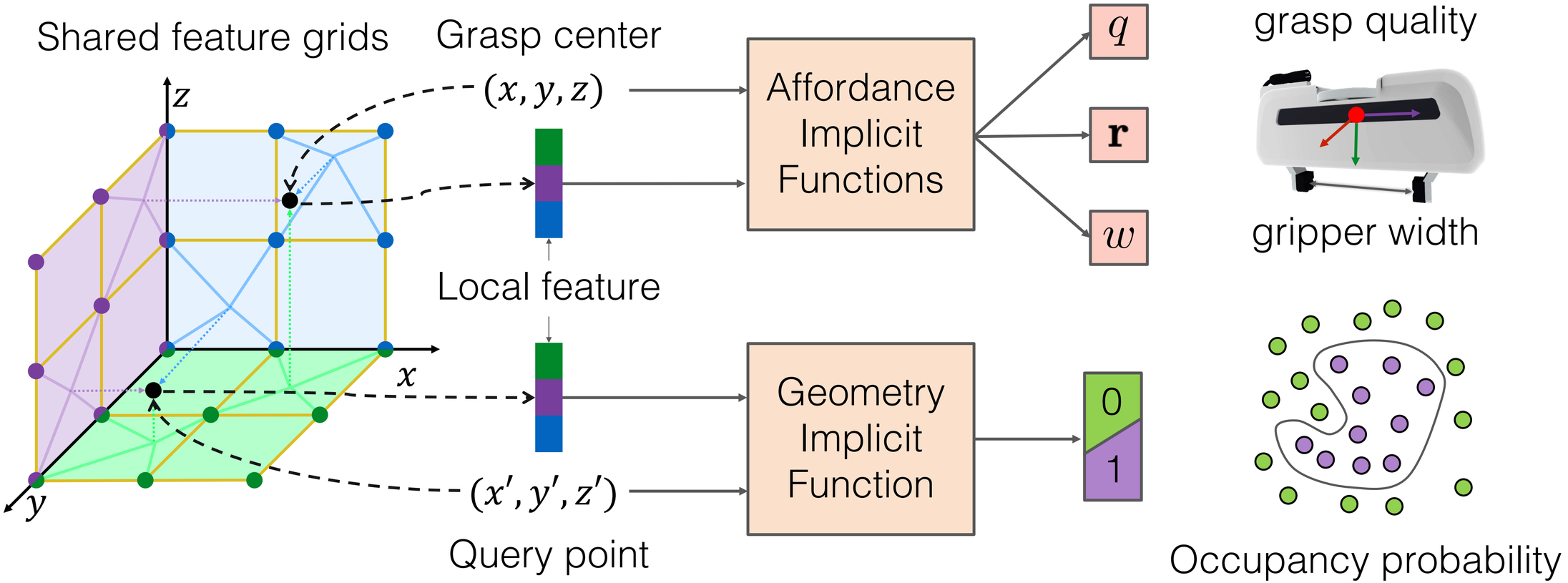



2D U-Nets

Structured feature grids

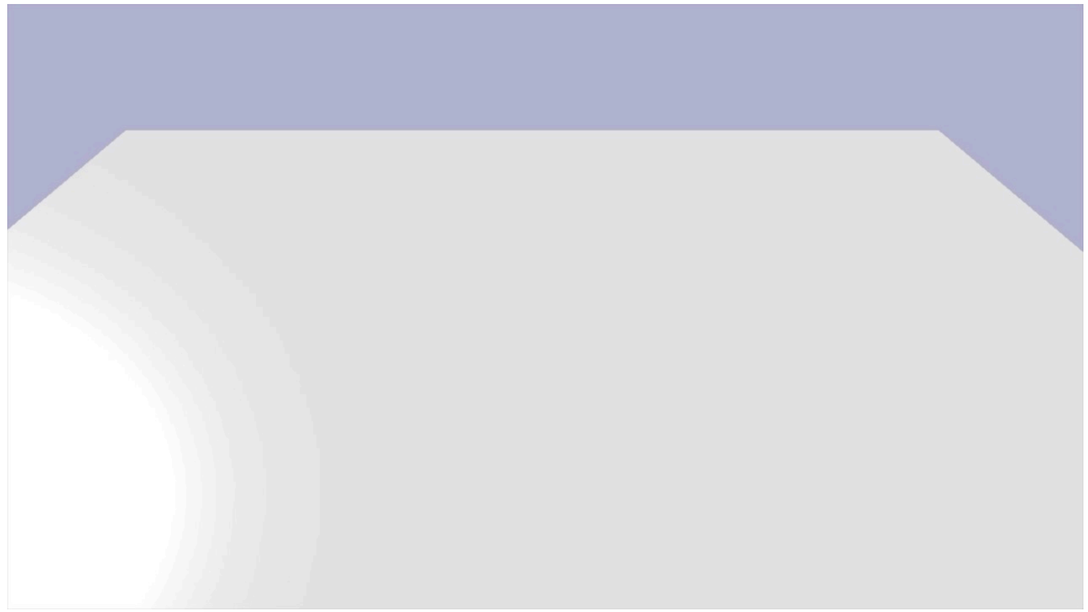


# GIGA Architecture

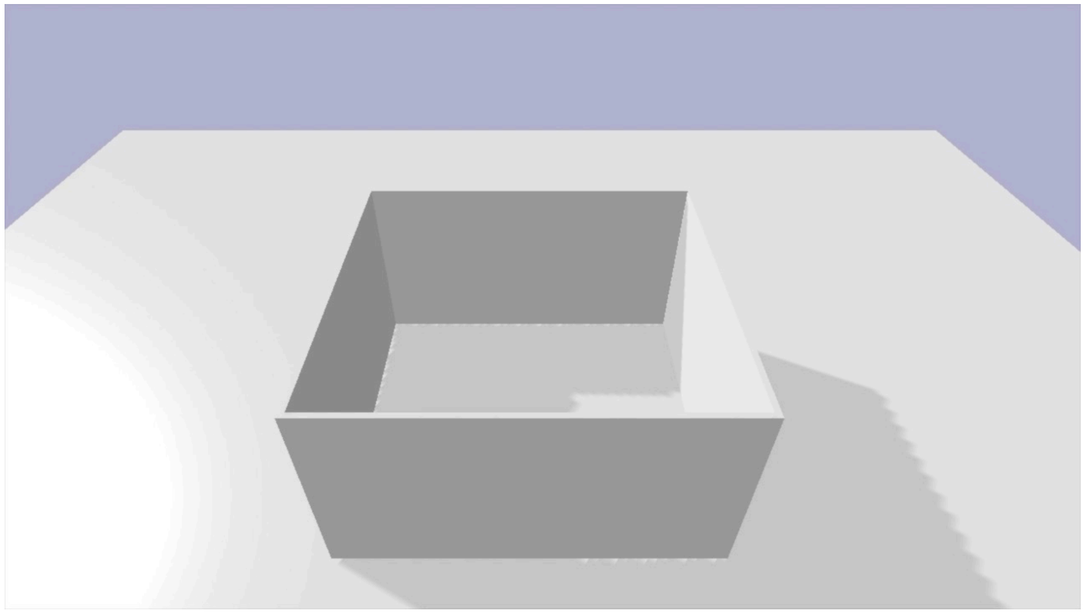


# Self-Supervised Data Collection

- Grasp affordance labels from physical trials in simulation.
- 3D geometry labels from ground-truth object meshes.



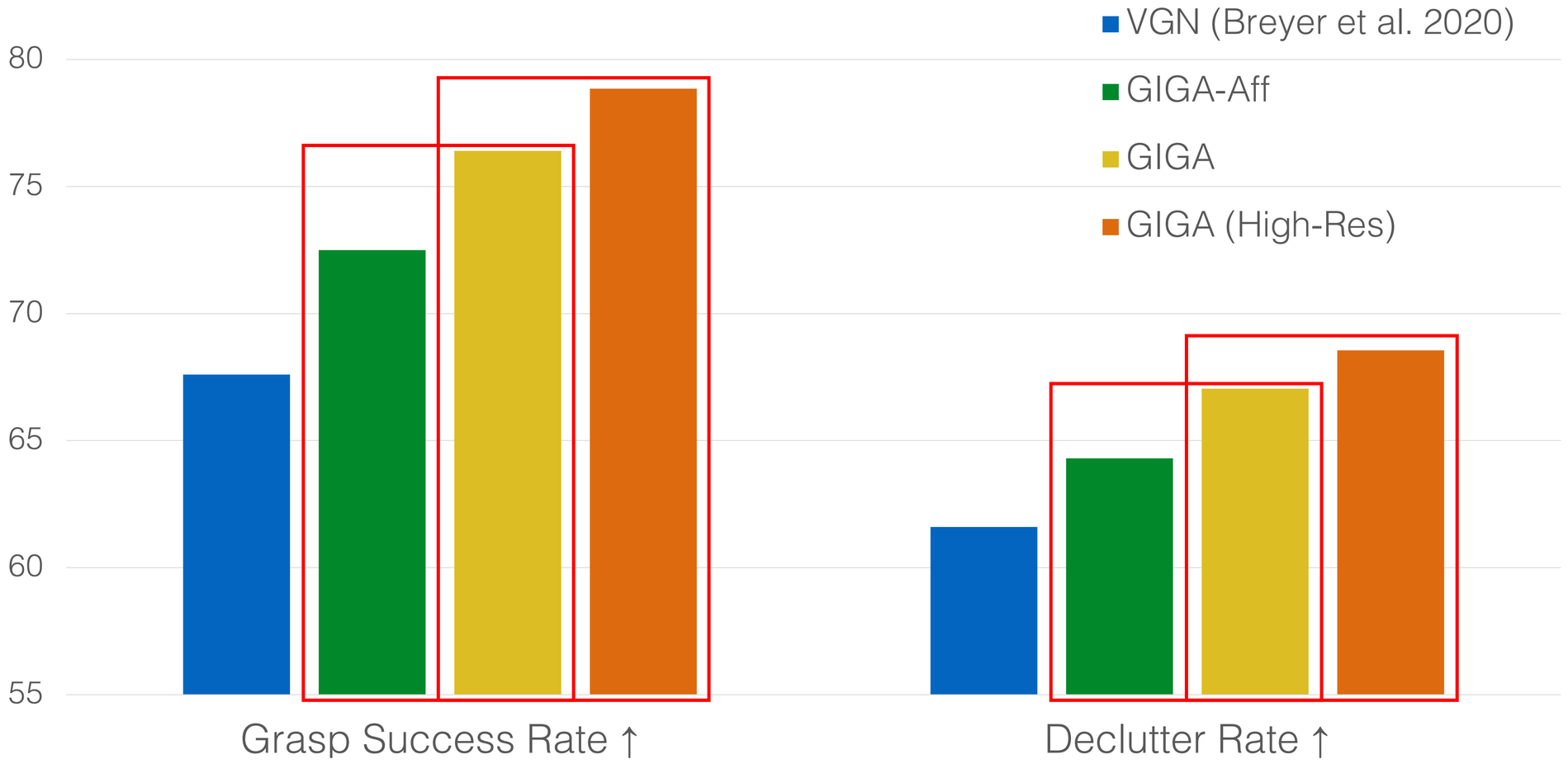
Packed scene



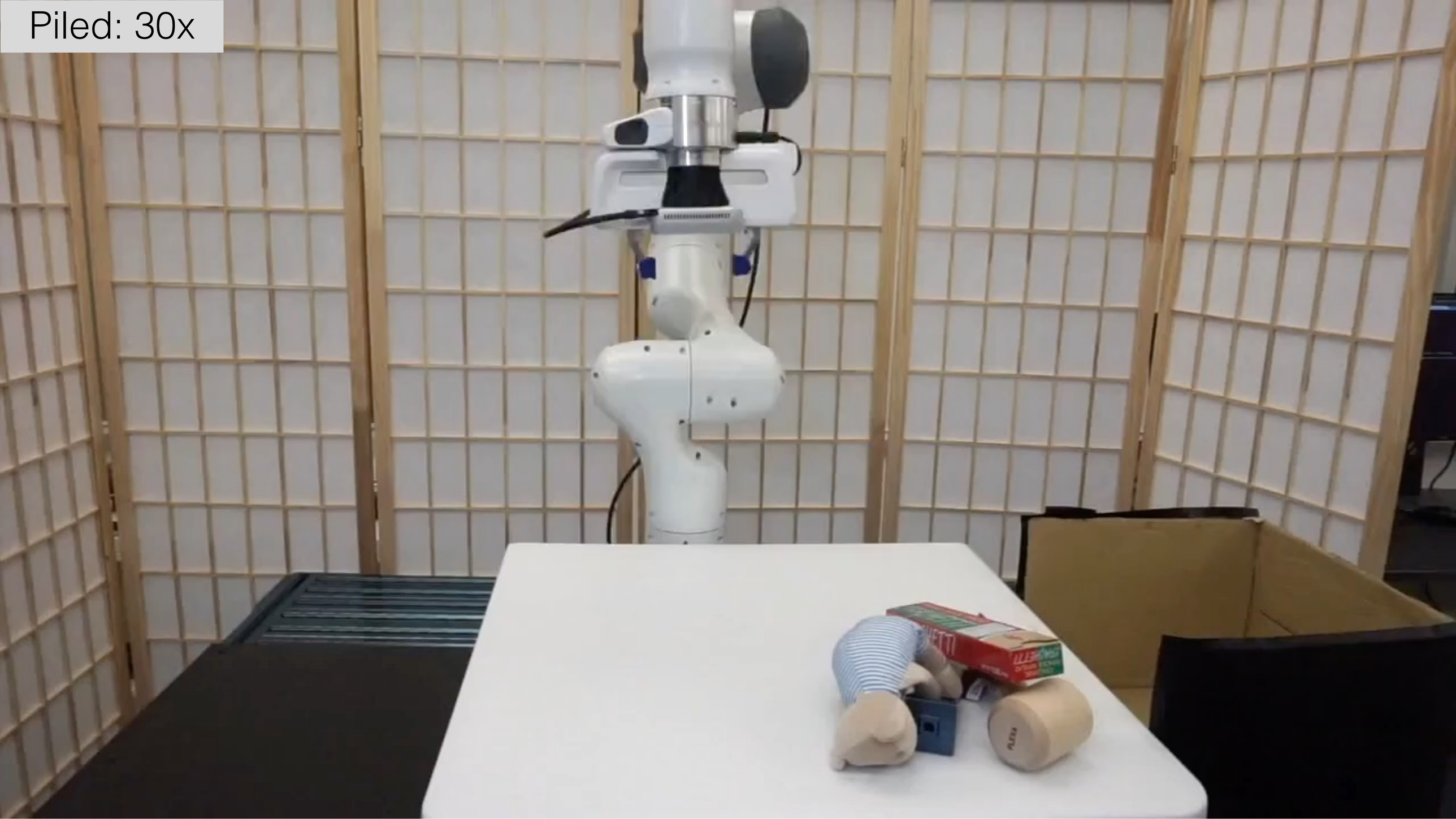
Pile scene

# Quantitative Comparison

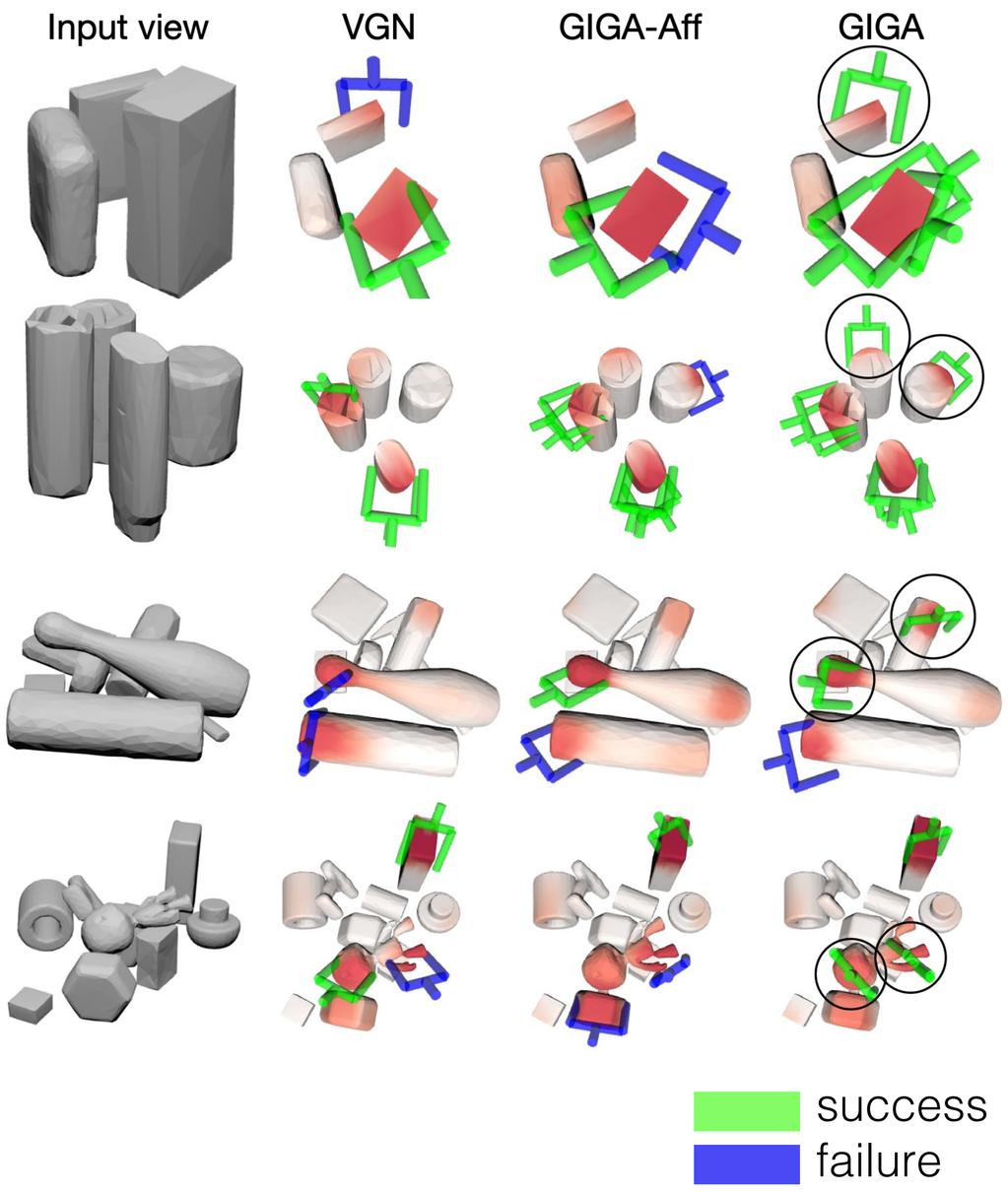
- Geometry learning facilitates affordance learning
- Continuity of implicit function enables higher performance




Packed: 30x

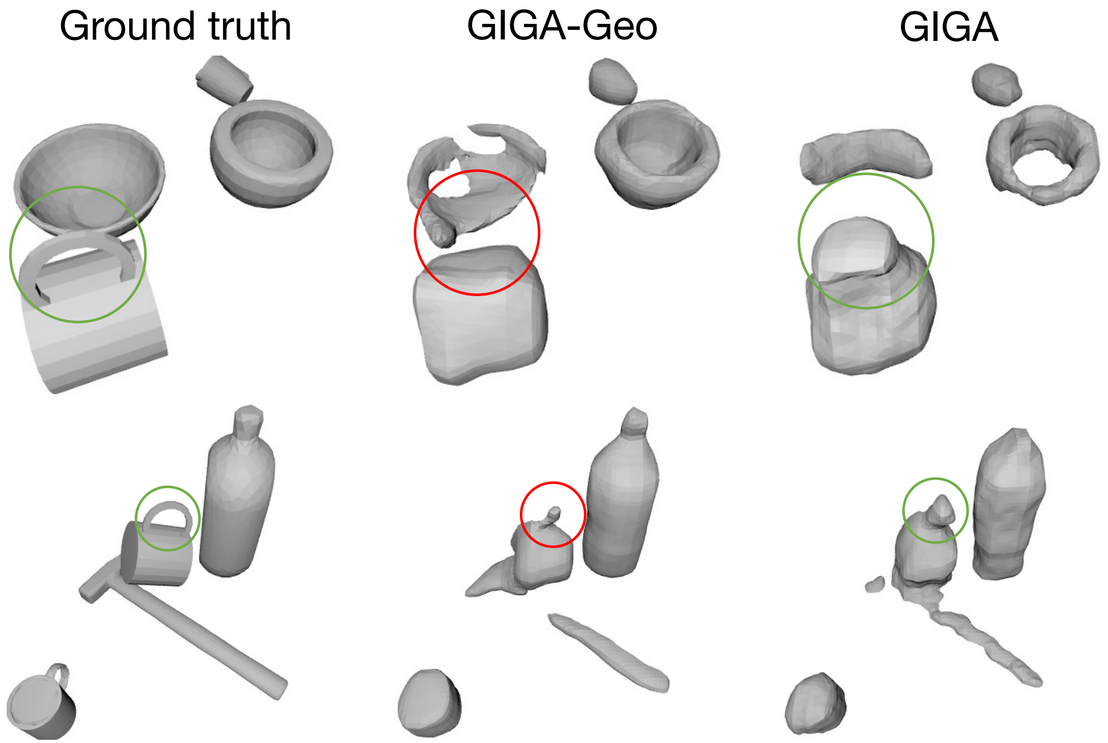

Piled: 30x

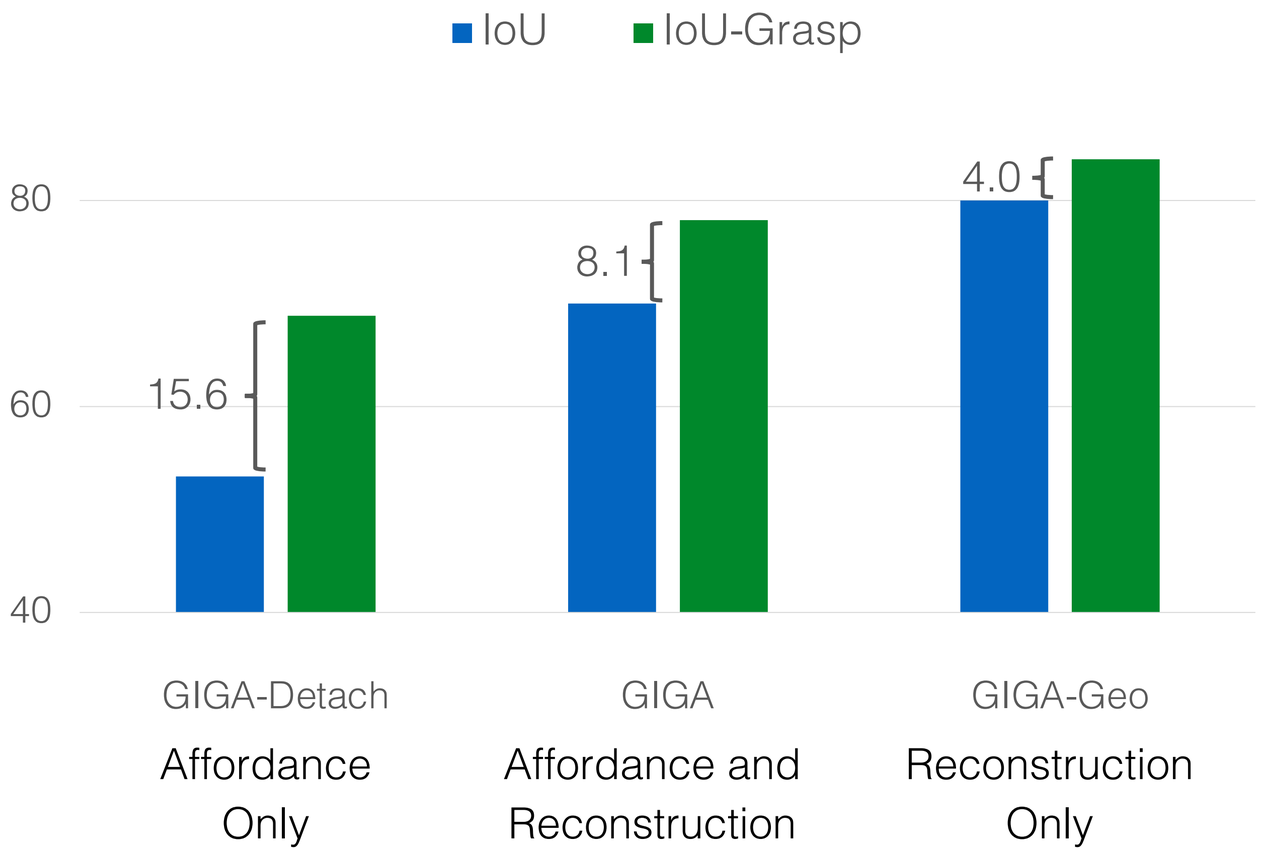
# Geometry Learning Facilitates Occluded Grasps





# Reconstruction Focuses on Graspable Parts



# GIGA: Summary

- Synergies between **affordance** and **geometry**

  Better grasp prediction, especially in occluded regions

  3D reconstruction focuses on action-relevant parts

- Structured **implicit neural representation**

  Continuous and compact representation for both affordance and geometry

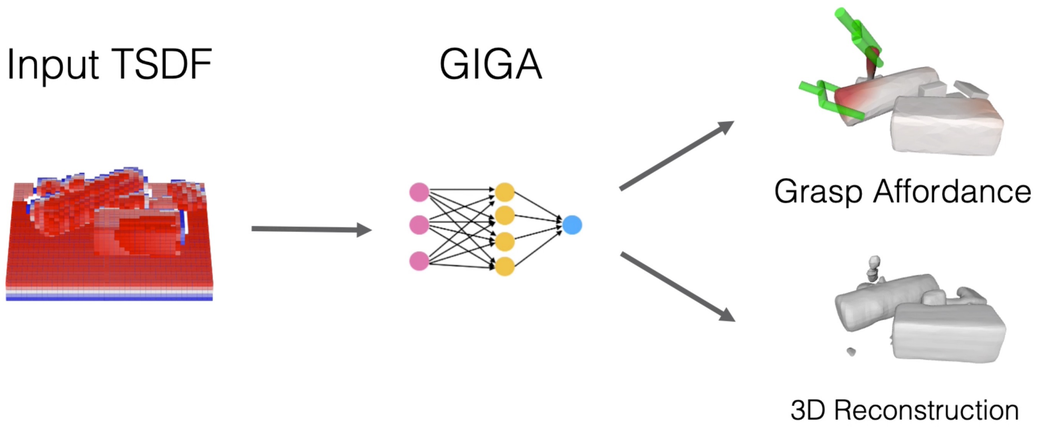
  Combine voxel grids with neural implicit functions

Code and models are available at https://sites.google.com/view/rpl-giga2021

# Visual Affordance Learning

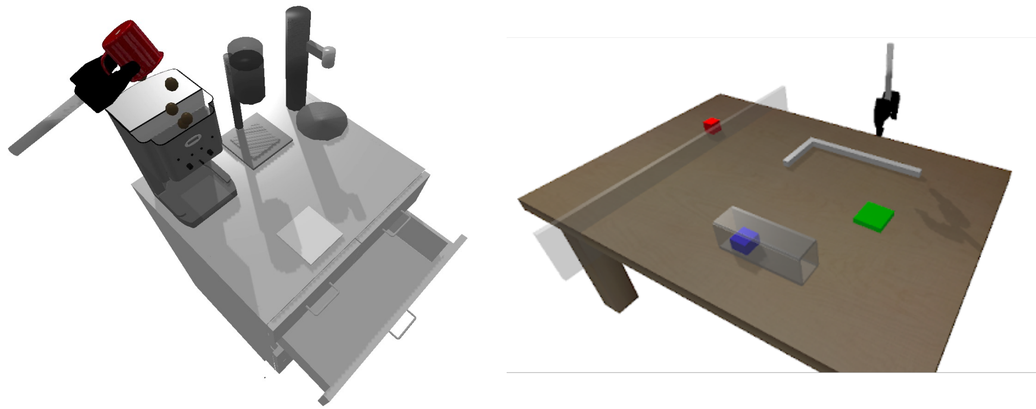
Bridging **visual perception** and **robot action** through **visual affordance**

## Learning affordance landscape with implicit representations



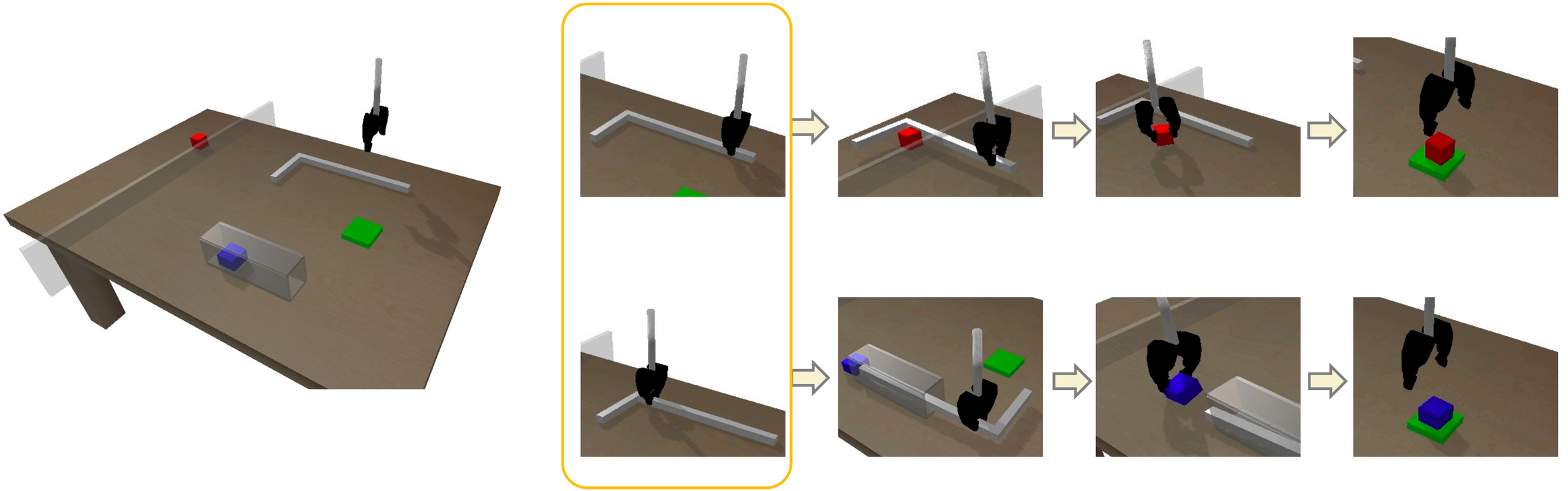

["GIGA" Jiang et al. RSS 2021]

## Planning long-horizon manipulation tasks with skill affordances

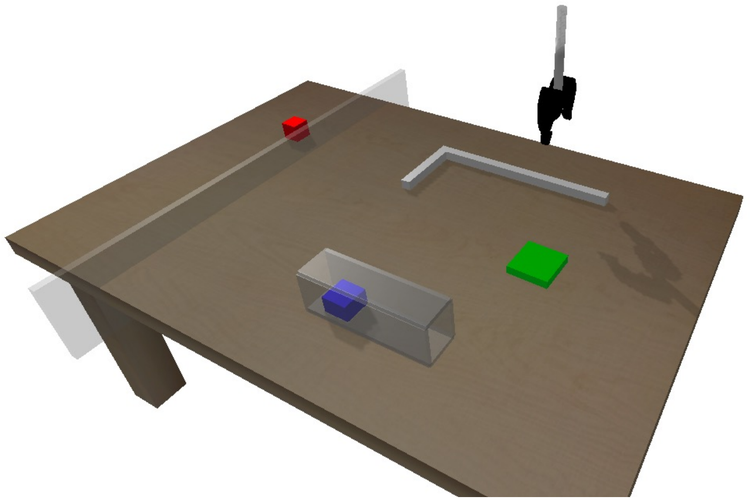


["Deep Affordance Foresight" Xu et al. ICRA21]

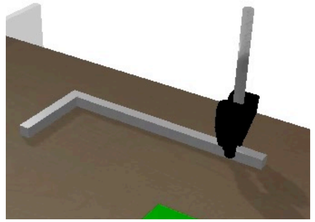
# Classical notion of **affordance** is **not** suitable for planning



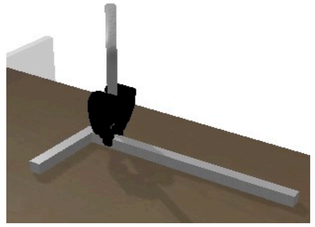
# Classical notion of **affordance** is **not** suitable for planning

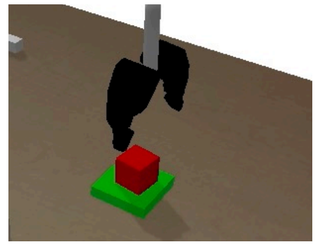


feasible? **YES**



feasible? **YES**

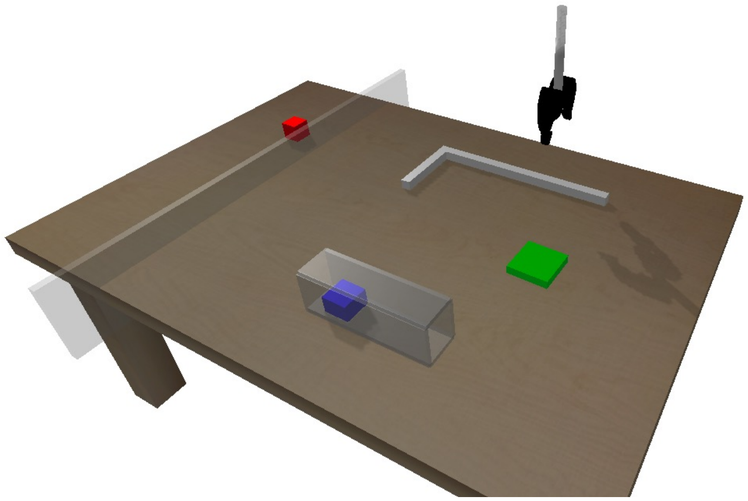




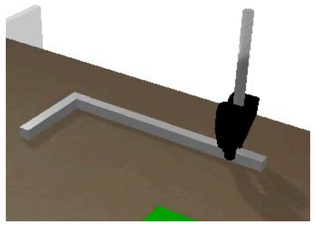
**Classical affordance**: whether an action is *feasible*

No way to choose actions with respect to a long-horizon task goal

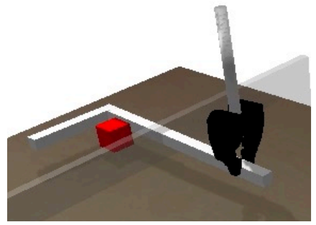
# Classical notion of **affordance** is **not** suitable for planning



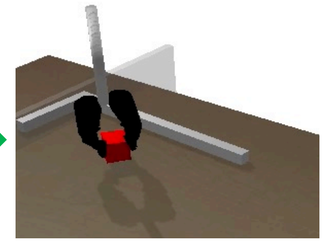
enables hooking red cube? **YES**

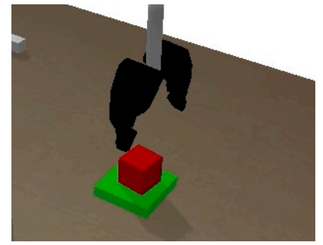


enables grasping red cube? **YES**

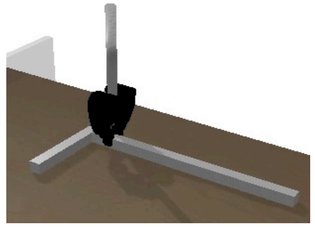


enables placing on target? **YES**





enables fetching red cube? **NO**



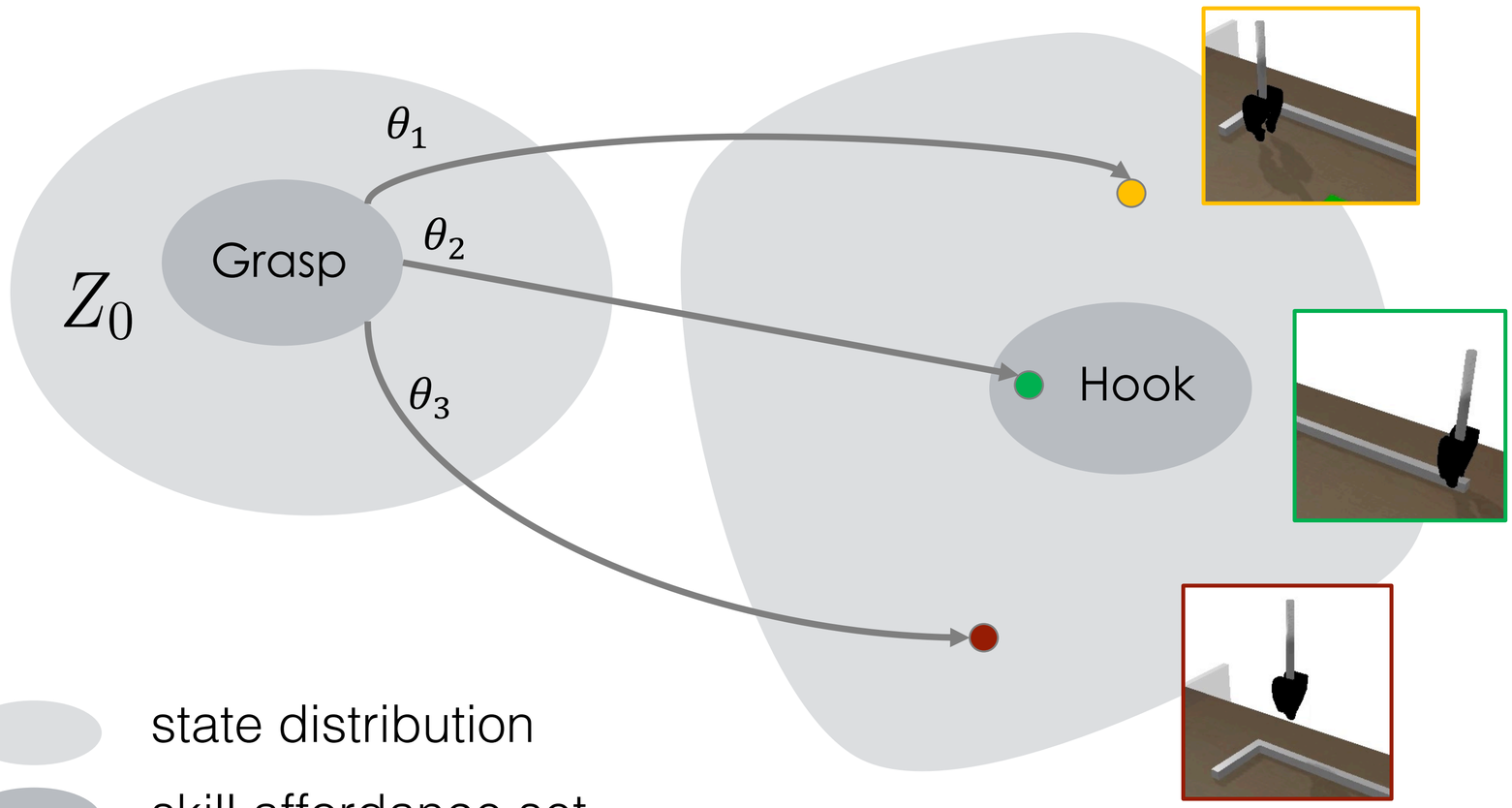
**Classical affordance**: whether an action is *feasible*

**Our new affordance**: will an action **make future actions feasible**?

Choose actions that enable the subsequent steps in the task plan

# Skill Affordances

**Parameterized skill plan** grasp($\theta$) ⟶ hook($\theta$) ⟶ grasp($\theta$) ⟶ $\cdots$




Affordance function:

$$\mathcal{A}_{\pi,\theta}(s)$$

Whether state $s$ is in the affordance set of $(\pi, \theta)$

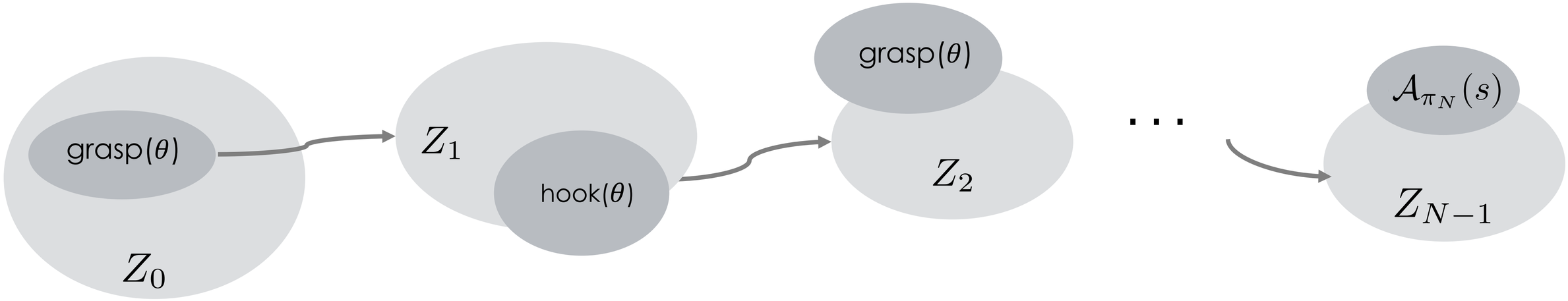
Xu et al. “Deep Affordance Foresight” ICRA 2021

# Verifying that a plan is executable with affordances

**Parameterized skill plan** grasp($\theta$) ⟶ hook($\theta$) ⟶ grasp($\theta$) ⟶ ···

How likely is this plan executable by the robot?

The plan is executable if every skill in the plan is afforded.




state distribution

skill affordance set

transition function

Probability that a plan is executable:

$$C_{plan}(\{(\pi_i, \theta_i)\}_{i=1}^{N}) = \sum_{s \in S} Z_{N-1}(s) \mathcal{A}_{\pi_N, \theta_N}(s)$$

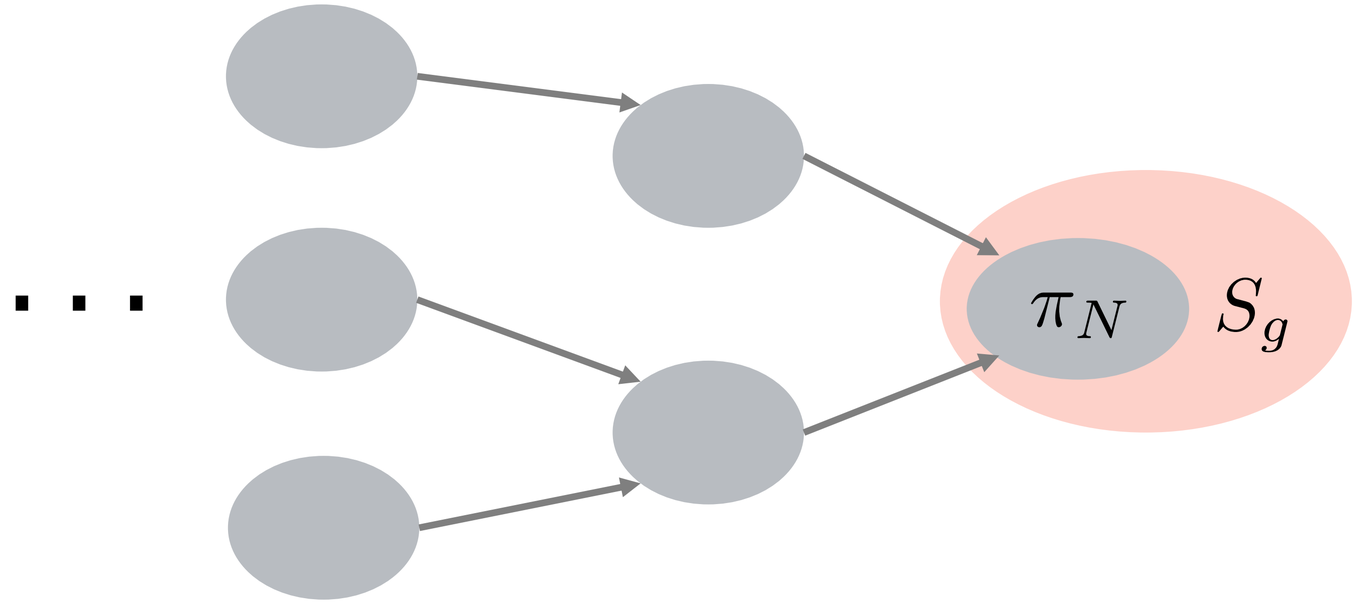
state distribution

affordance likelihood

# Planning towards a Goal

We define a goal $g \in G$ as a condition function that checks whether a state satisfies the goal. We denote the set of states that satisfies $g$ as $S_g$ .

Any plan that ends with a skill whose affordance set is a *subset* of $S_g$ are goal-directed plans for $g$ .




We denote all plans directed at goal $g$ as $\mathcal{P}_g$

# Planning with Affordance

Probability that a plan is executable:

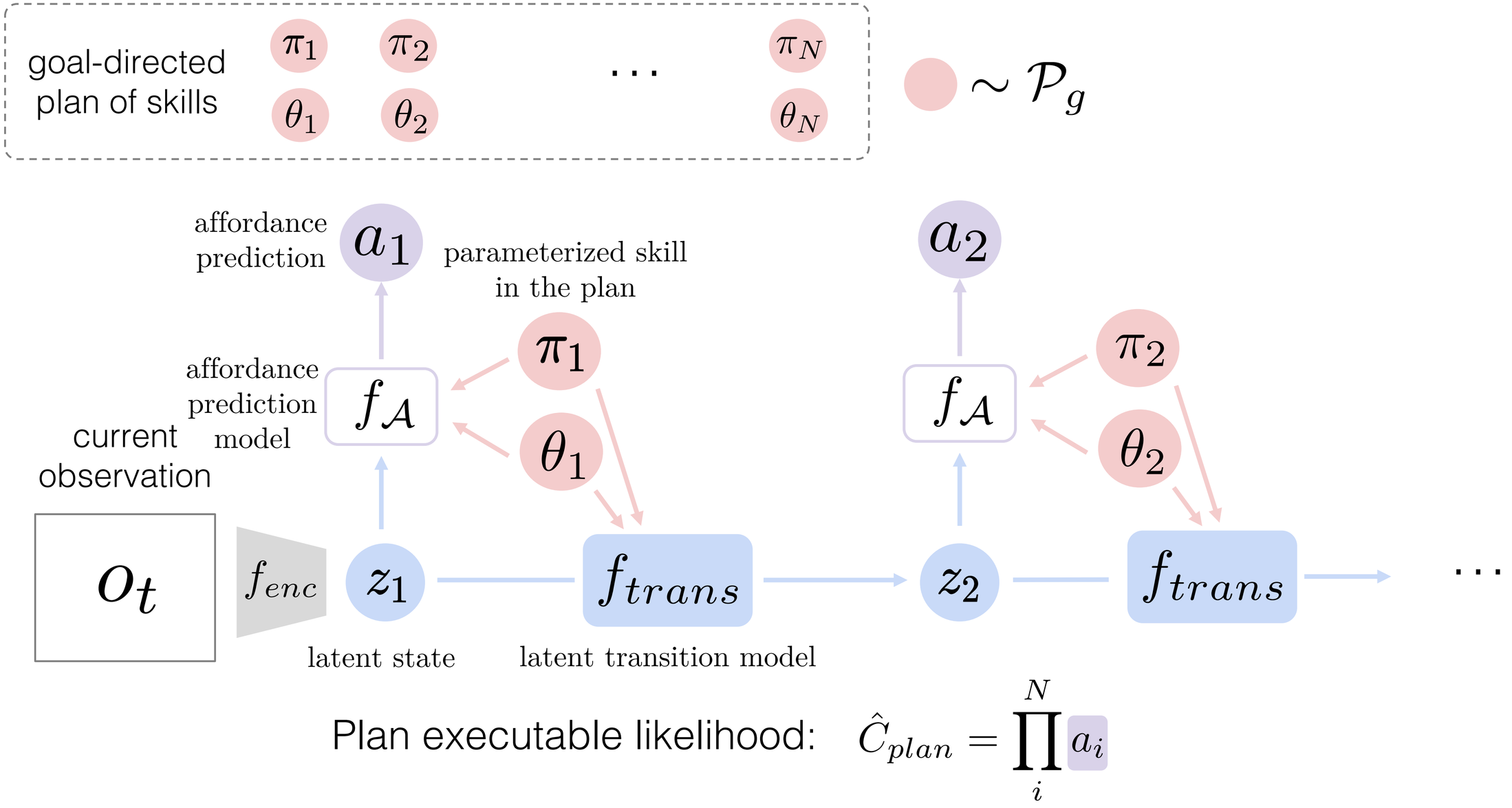
$$C_{plan}(\{(\pi_i, \theta_i)\}_{i=1}^{N}) = \sum_{s \in S} Z_{N-1}(s) \mathcal{A}_{\pi_N, \theta_N}(s)$$

Search for goal-directed plans that are most likely executable:

$$\arg\max_{p \in \mathcal{P}_g} C_{plan}(p)$$

How do we plan in a partially-observed environment with unknown dynamics?

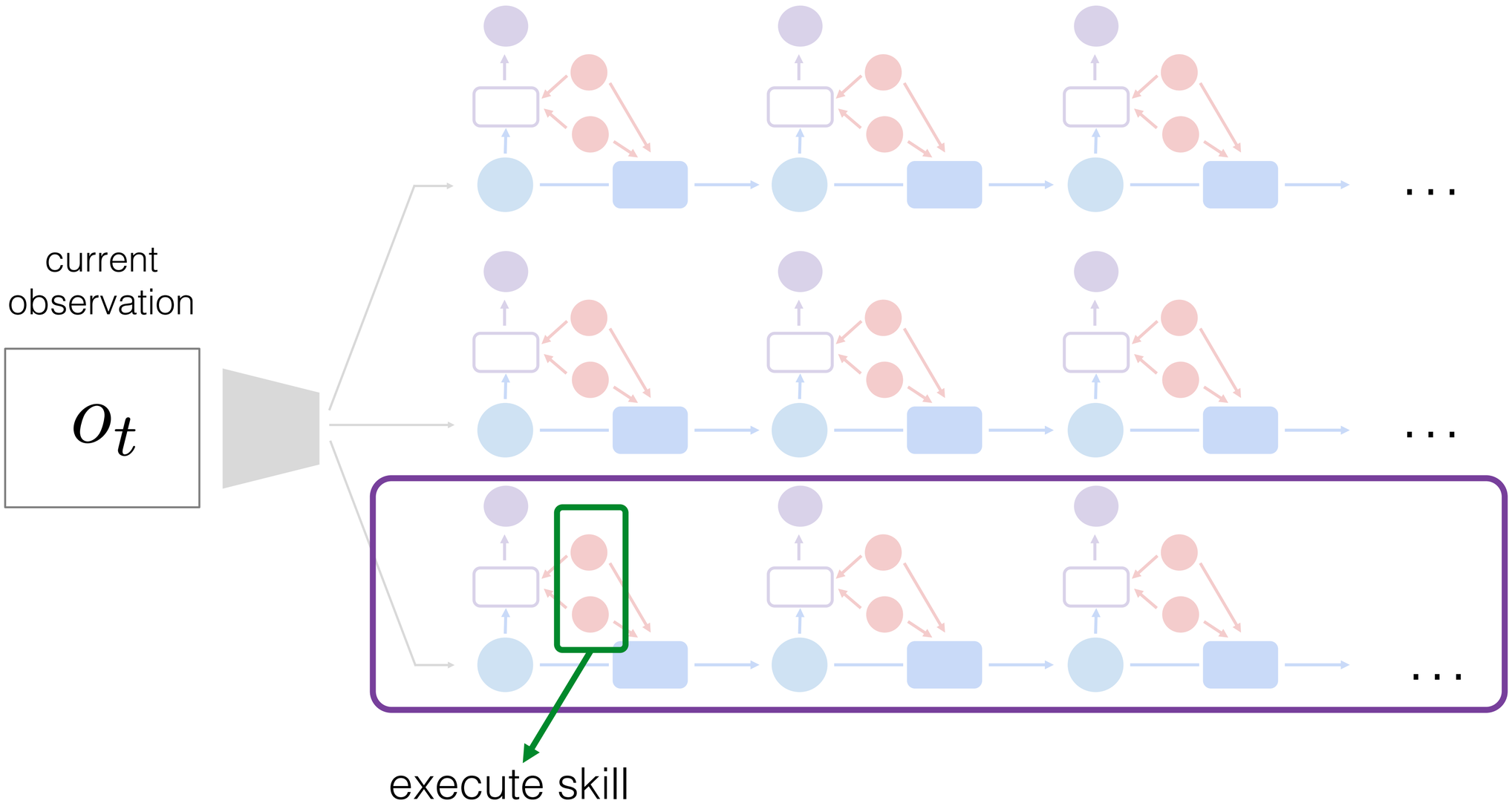
# Learning to Plan with Deep Affordance Foresight (DAF)



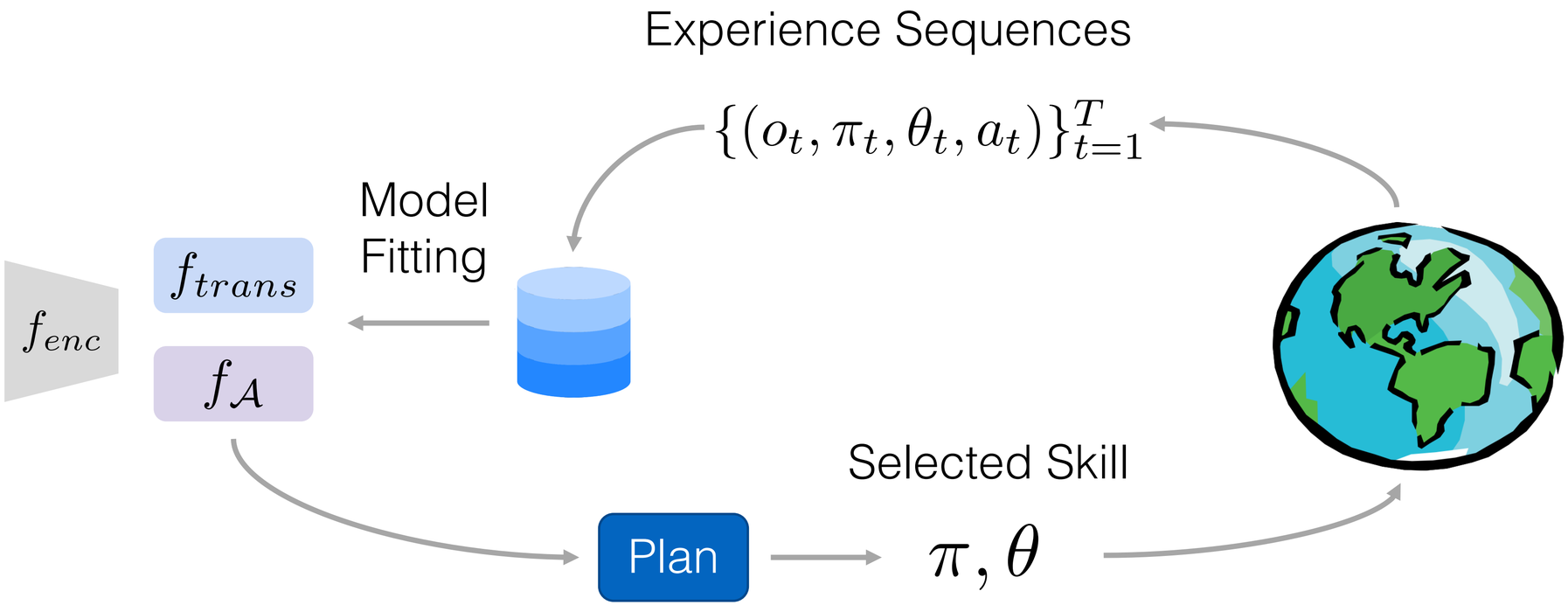

Plan executable likelihood: $\hat{C}_{plan} = \prod_i^N a_i$

Xu et al. “Deep Affordance Foresight” ICRA 2021

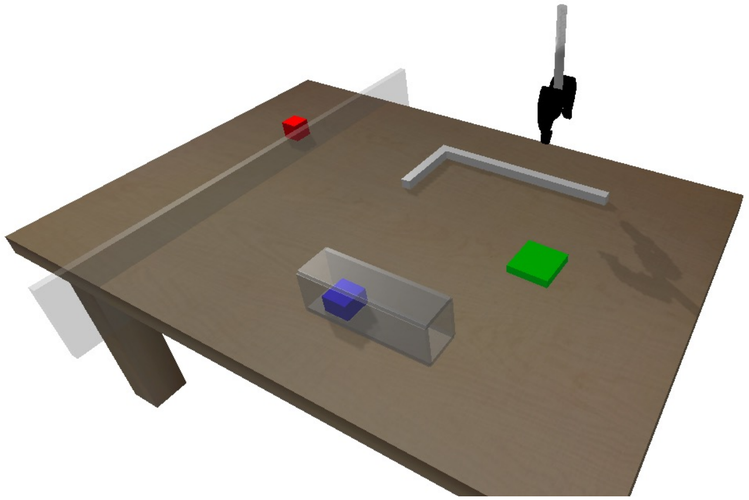
# DAF with Model-Predictive Control (MPC)

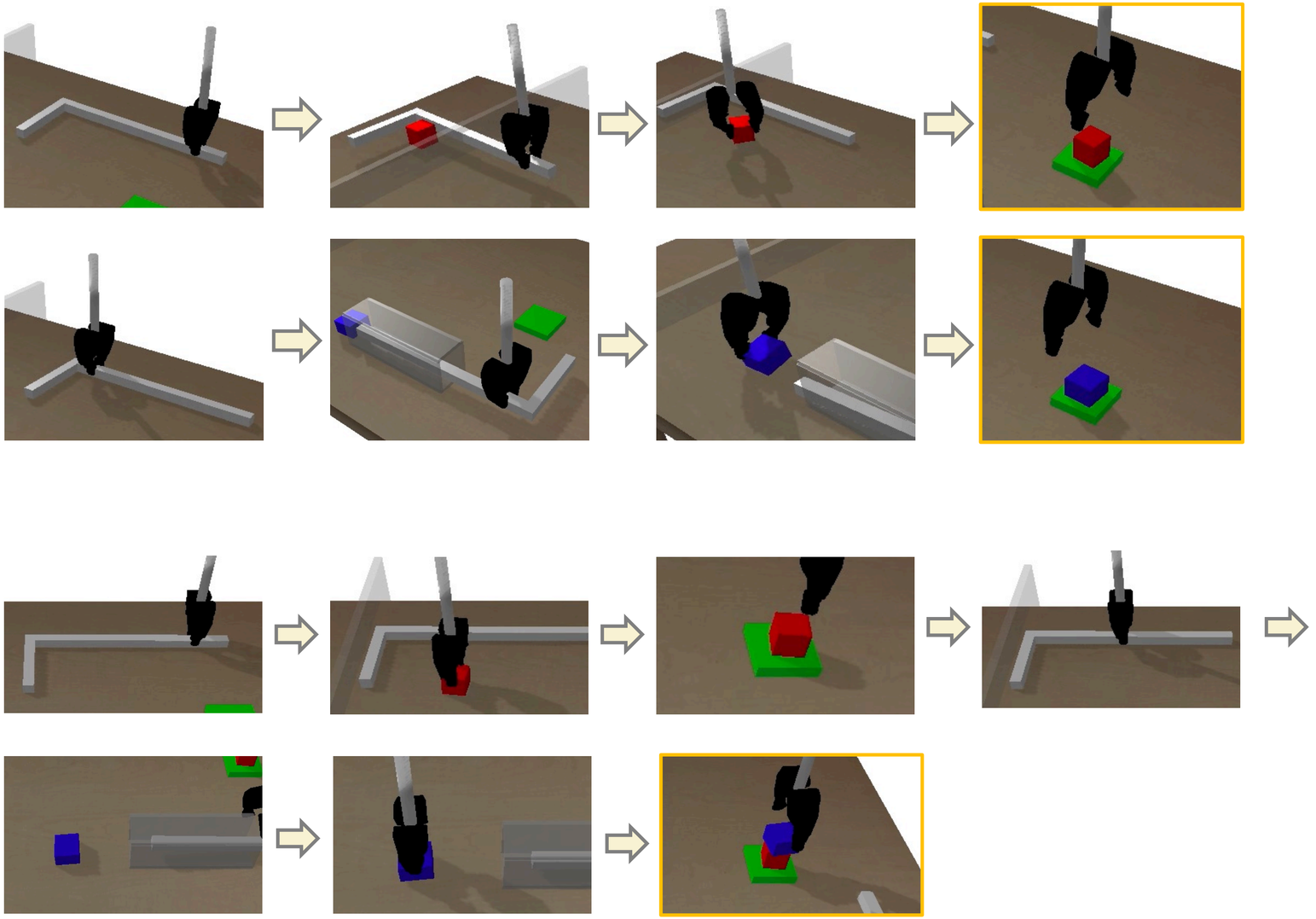


# Learning Through Exploration

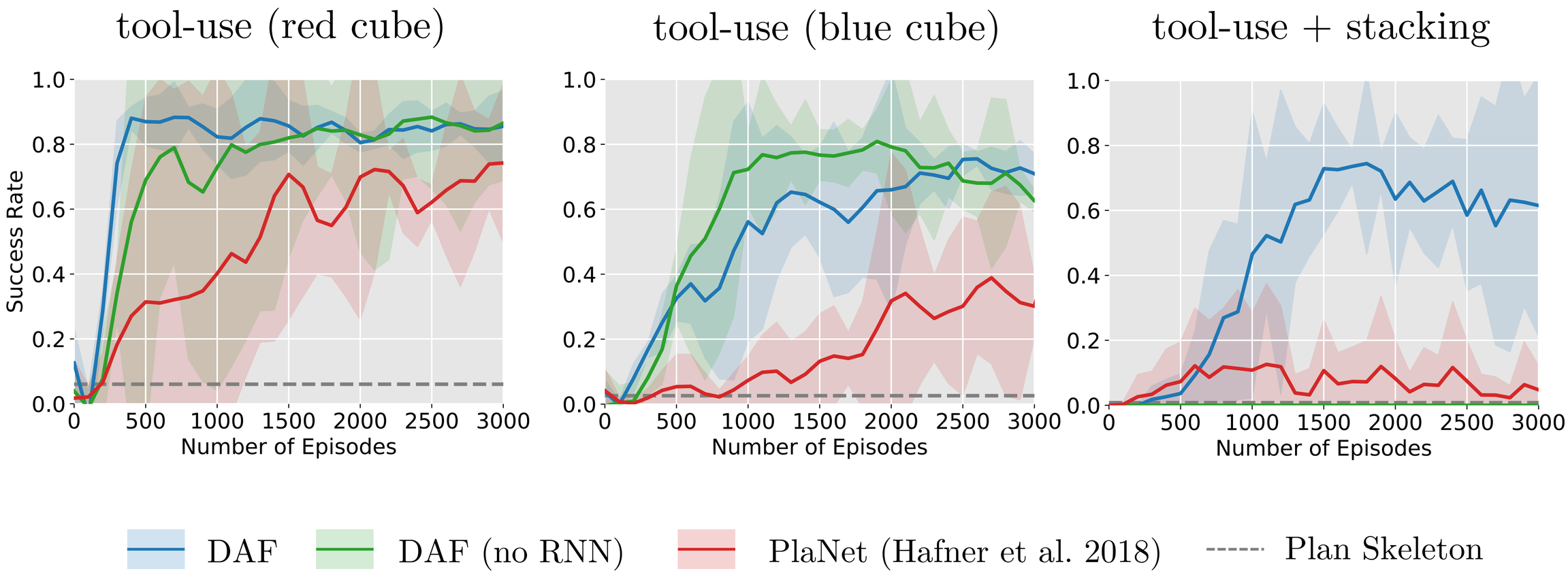


# Experiments: Tool-Use + Stacking

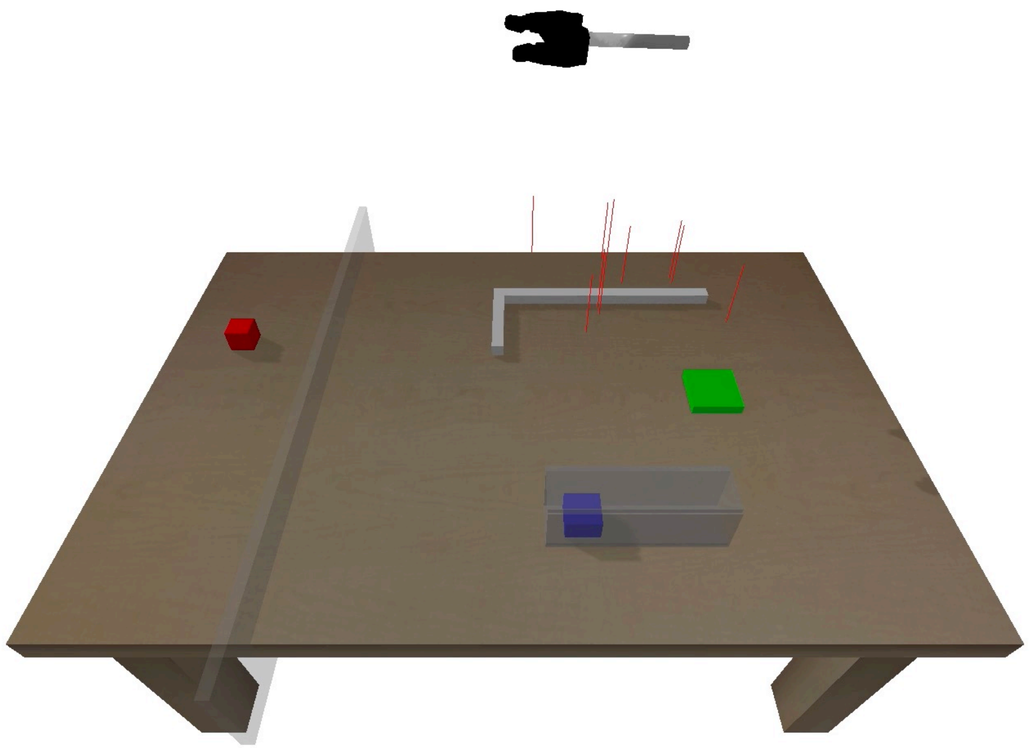




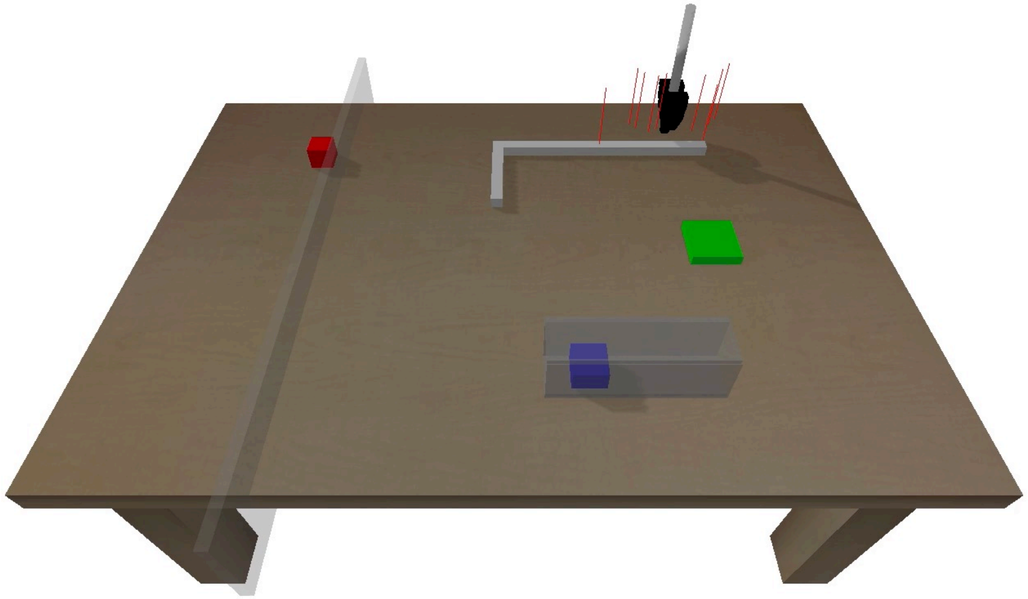
# Results: Tool-Use



# Sample Rollouts: Tool-Use + Stacking

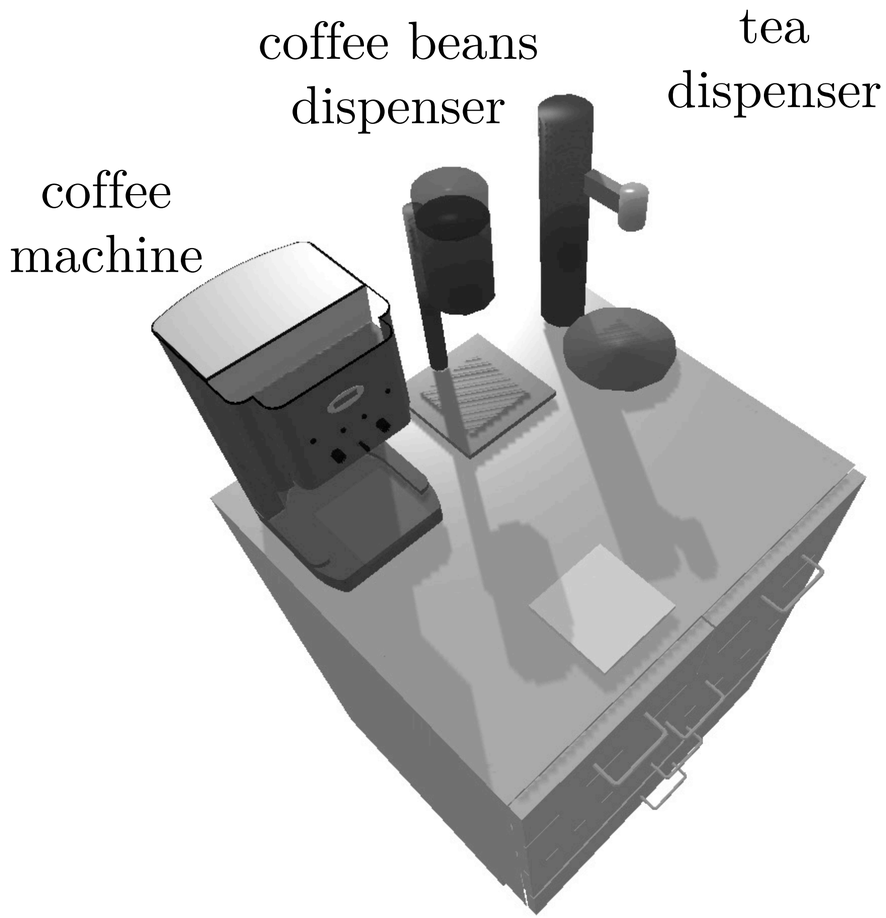


PlaNet

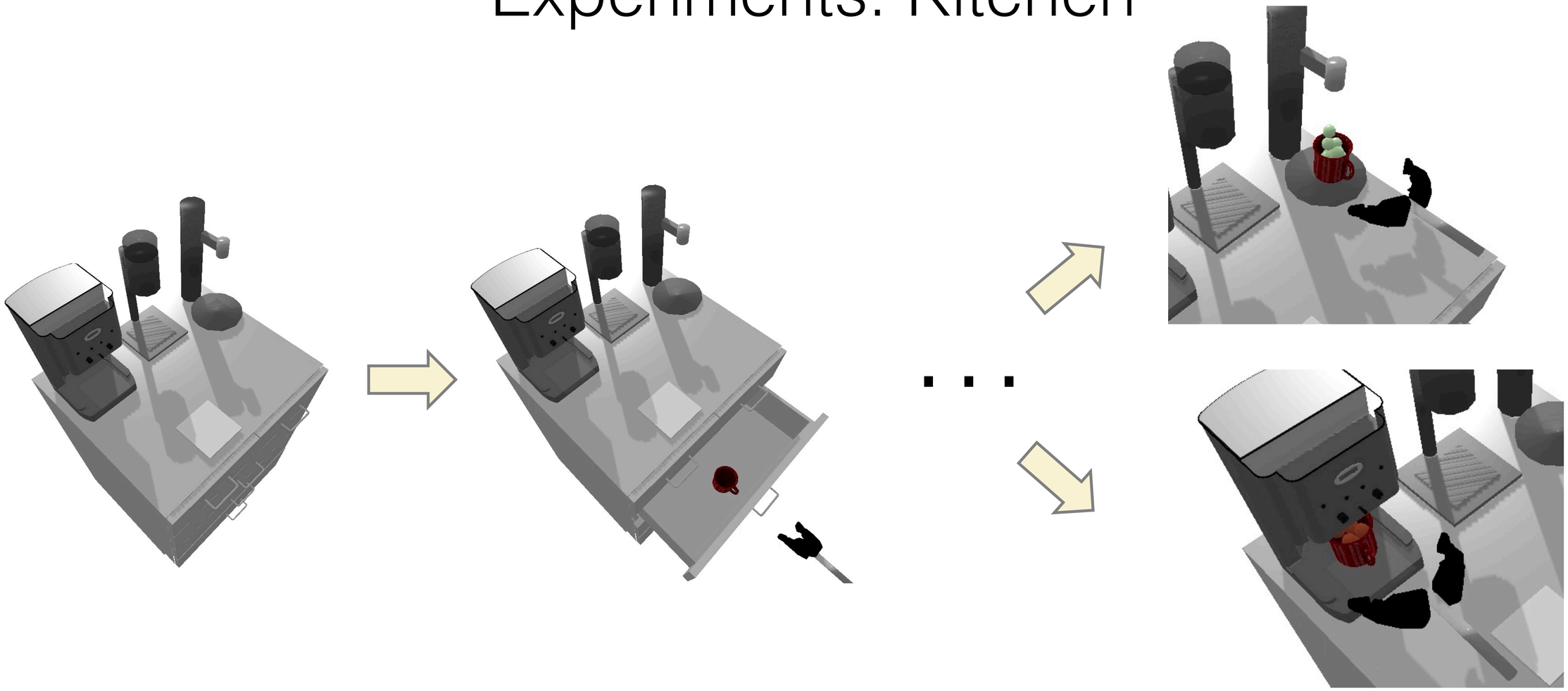


DAF (Ours)

# Experiments: Kitchen



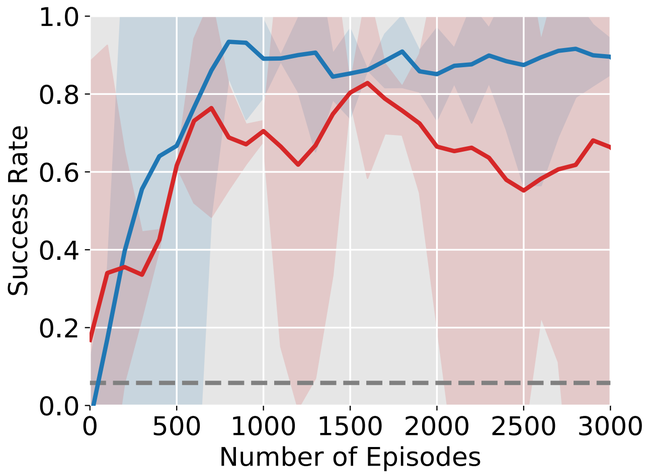
# Experiments: Kitchen



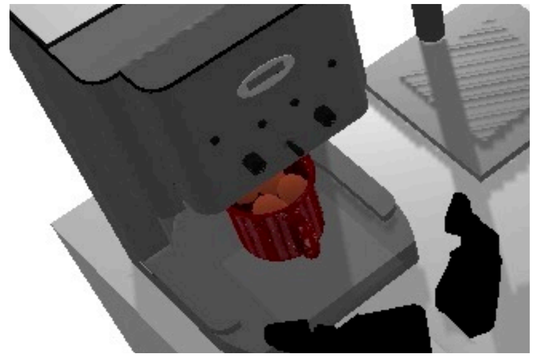
- Affordances for fetching the mug is shared among the two tasks.
- Our method should be able to learn the task much faster & better

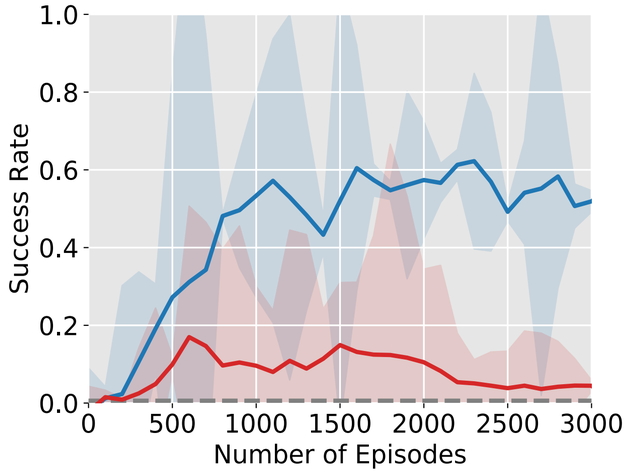
# Results: Kitchen

get tea



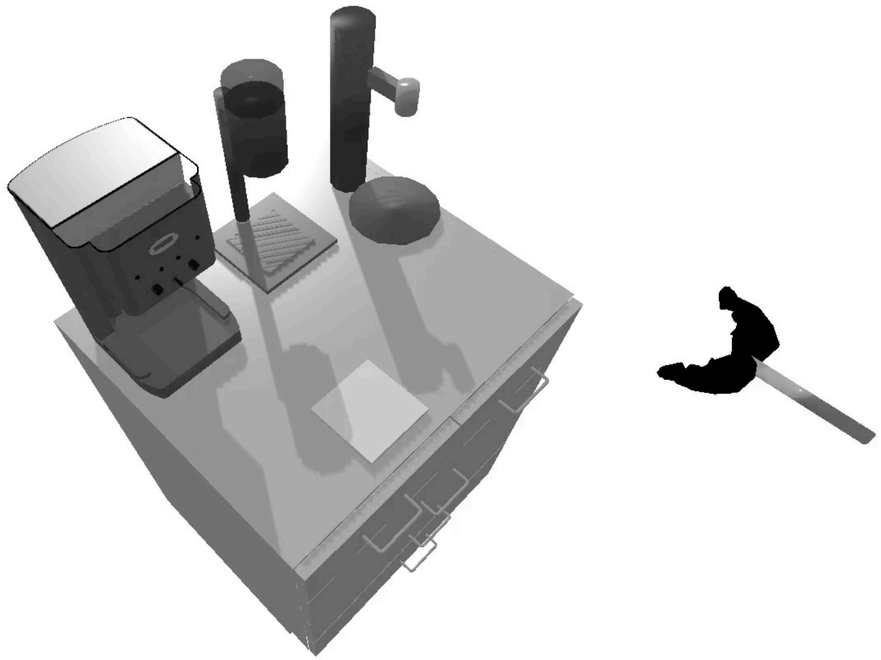

get coffee



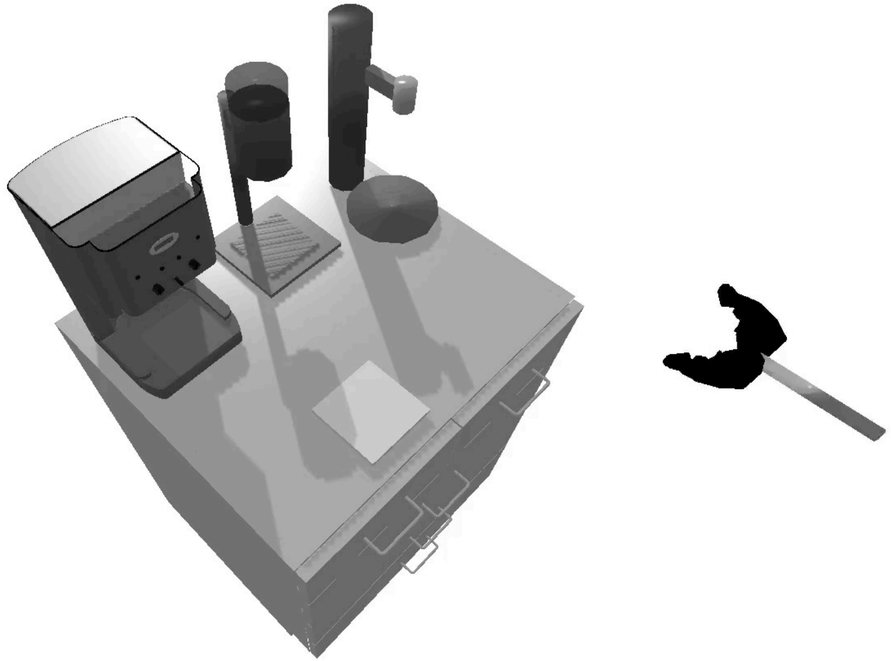



DAF (image)  PlaNet (image)  plan skeleton

# Rollouts: Kitchen (Coffee)

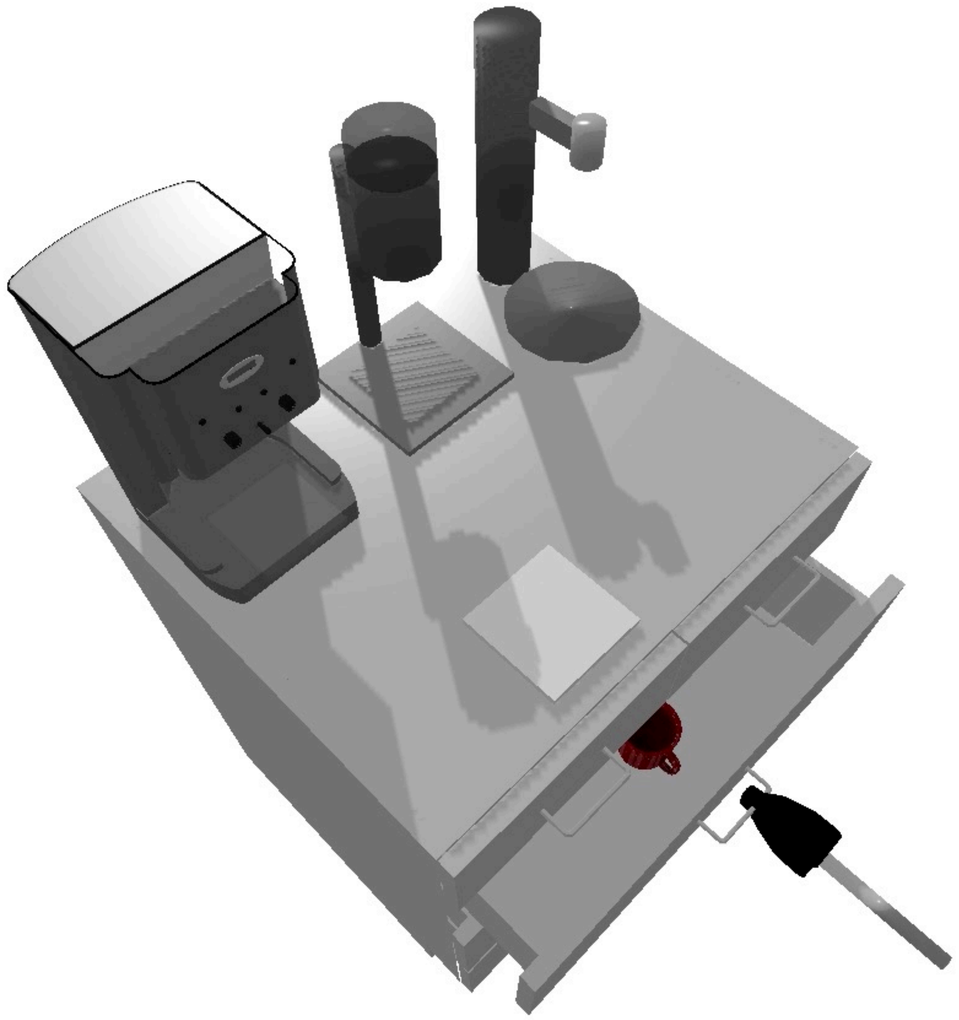


PlaNet

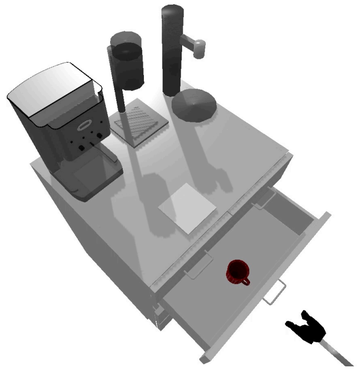


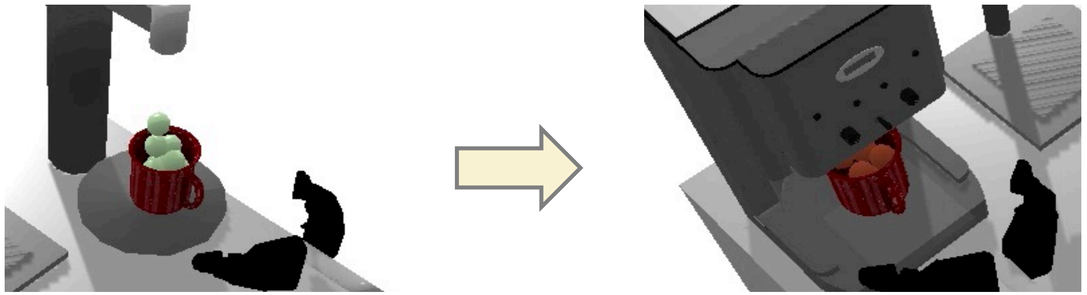
DAF (Ours)

# DAF is reactive and can recover from mistakes



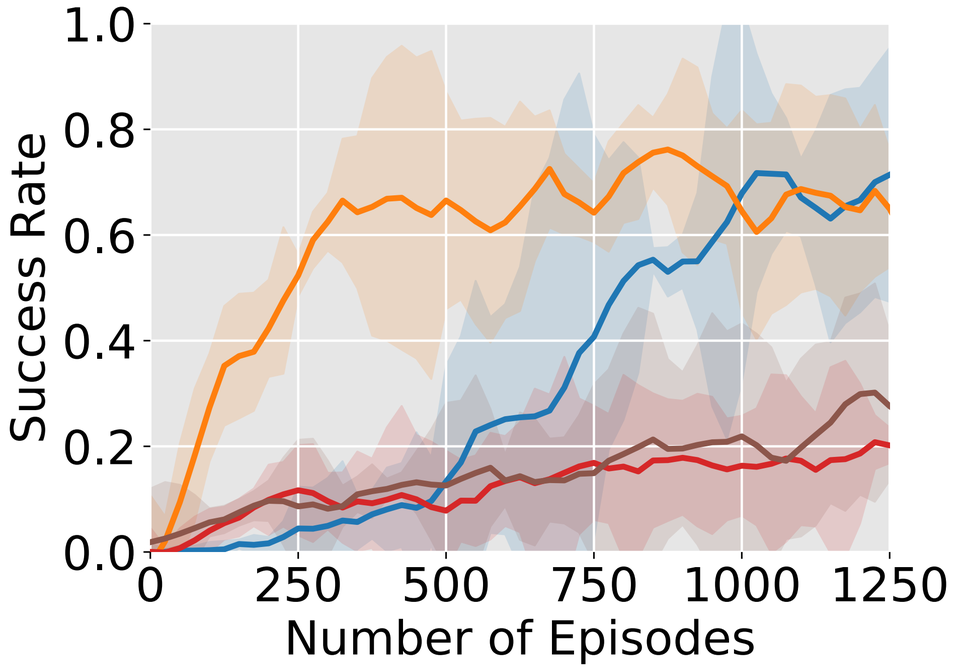
# Transfer learned affordances





base task: get tea

transfer task: get coffee



# Deep Affordance Foresight: Summary

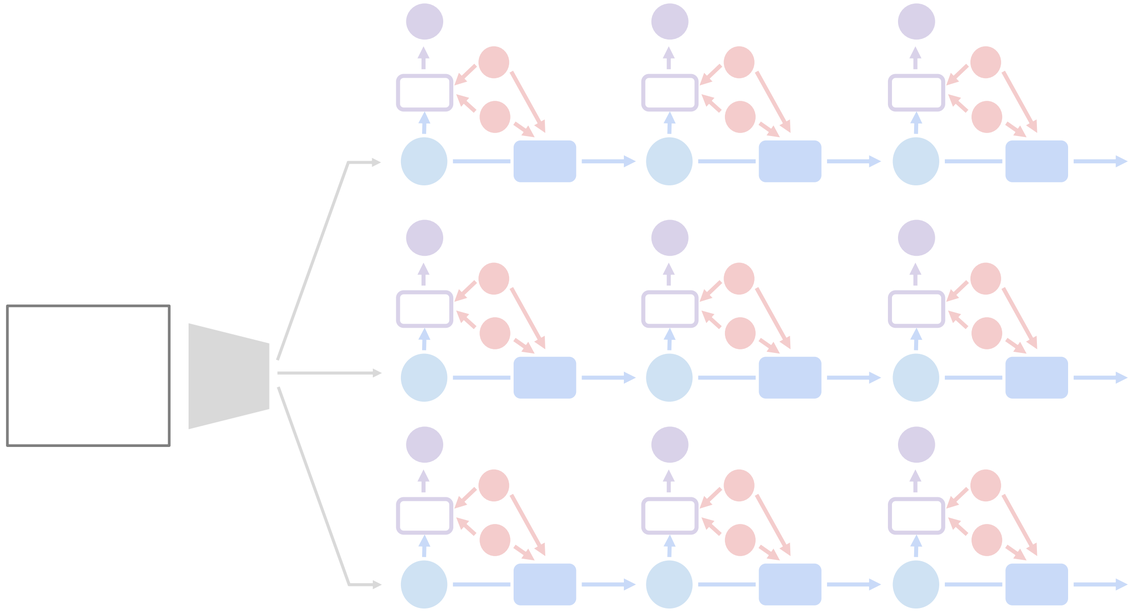
- New notion of **skill affordance**

  action abstraction that supports long-horizon planning

  estimated from visual observations

- Latent-space **planning** algorithm

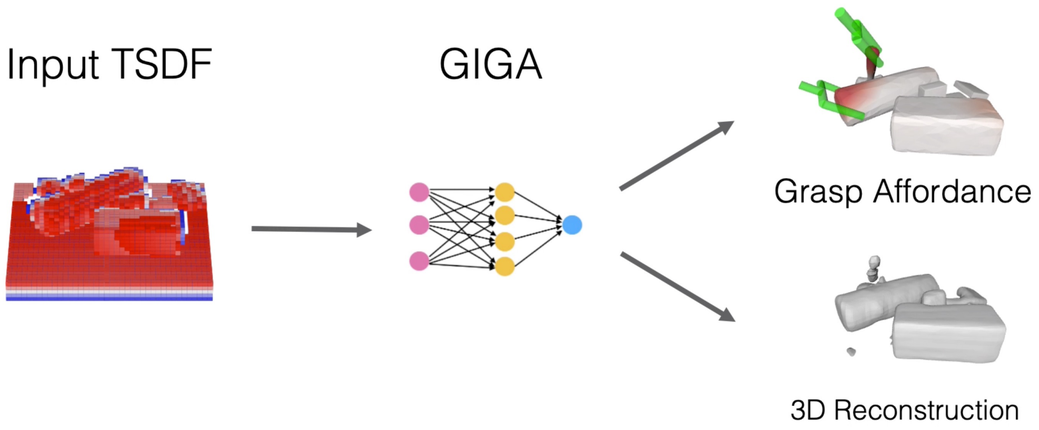
  DAF can plan through **raw perception** and **complex non-rigid dynamics**



# Visual Affordance Learning

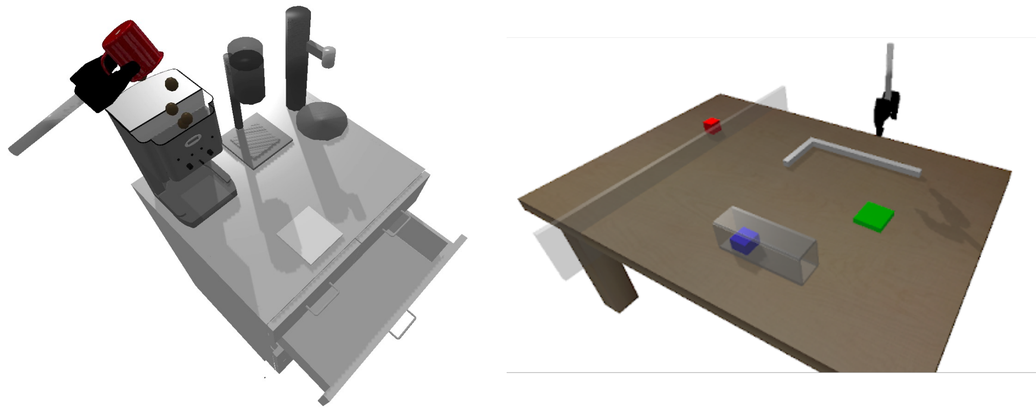
Bridging **visual perception** and **robot action** through **visual affordance**

## Learning grasp affordance with implicit representations




["GIGA" Jiang et al. RSS 2021]

## Planning long-horizon manipulation tasks with skill affordances



["Deep Affordance Foresight" Xu et al. ICRA21]

# Visual Affordance Learning

Bridging **visual perception** and **robot action** through **visual affordance**

## human-centered

(human environments are shaped by affordance)

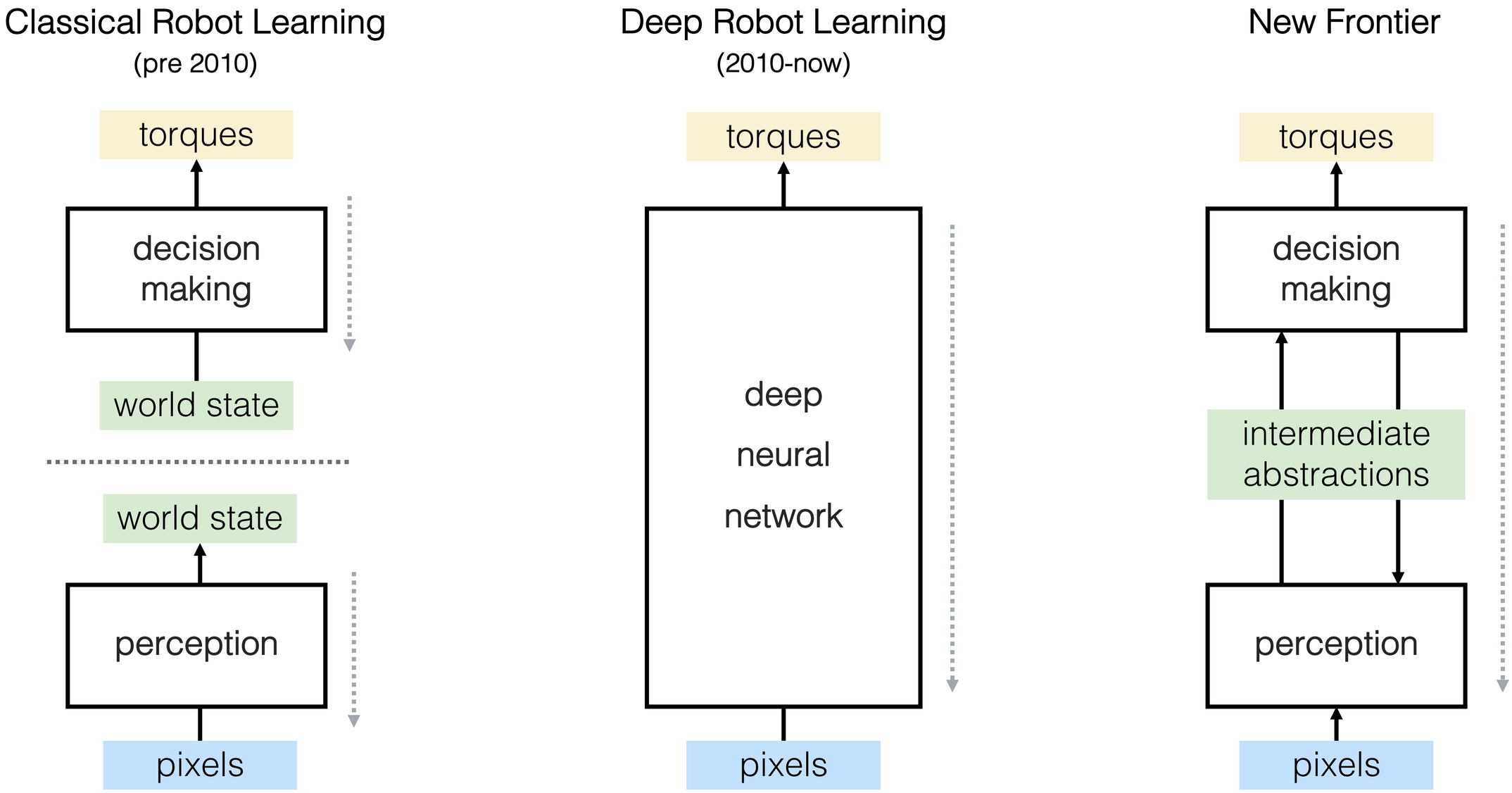

## physically-grounded

(informative representation for robot actions)

## task-agnostic

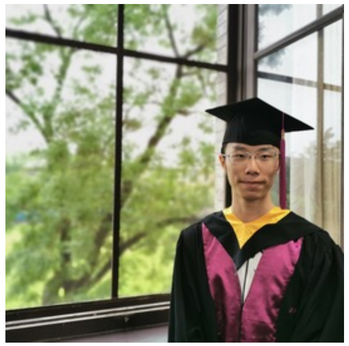
(inherent properties that facilitate knowledge transfer)

# Architectures of Robot Manipulation



# Acknowledgement


## UT Robot Perception and Learning Lab https://rpl.cs.utexas.edu/

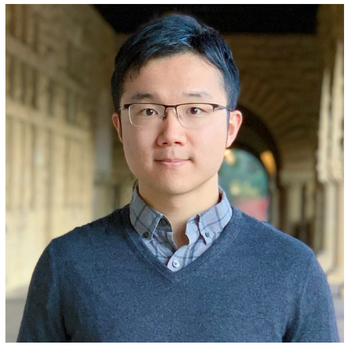

Zhenyu Jiang

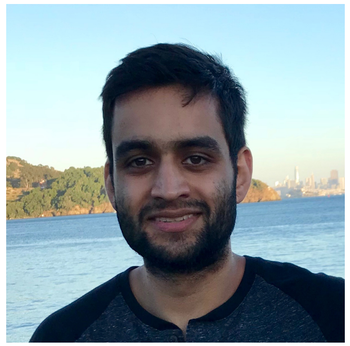

Yifeng Zhu


Soroush Nasiriany


Maxwell Svetlik


Danfei Xu


Kuan Fang


Ajay Mandlekar


Roberto Martín-Martín

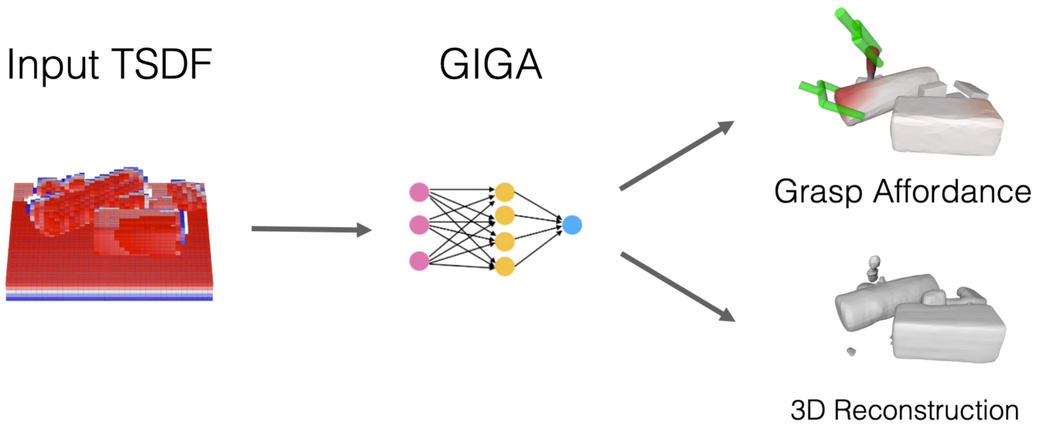

Fei-Fei Li


Silvio Savarese

# Visual Affordance Learning

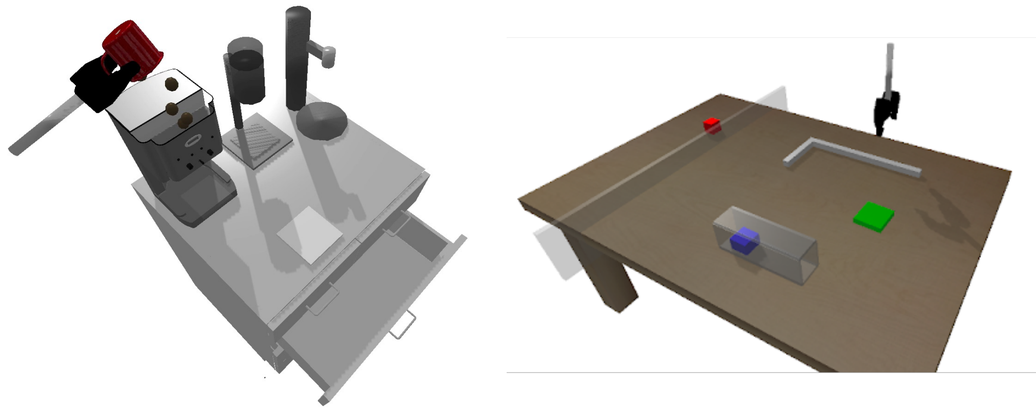
Bridging **visual perception** and **robot action** through **visual affordance**

## Learning affordance landscape with implicit representations




["GIGA" Jiang et al. RSS 2021]

## Planning long-horizon manipulation tasks with skill affordances



["Deep Affordance Foresight" Xu et al. ICRA21]