Objects, Skills, and the Quest for Compositional Robot Autonomy

Yuke Zhu

February 18, 2022

TEXAS Robotics

TEXAS
The University of Texas at Austin

Two Worlds of Building Robot Autonomy

James Webb Space Telescope

robotic systems engineering

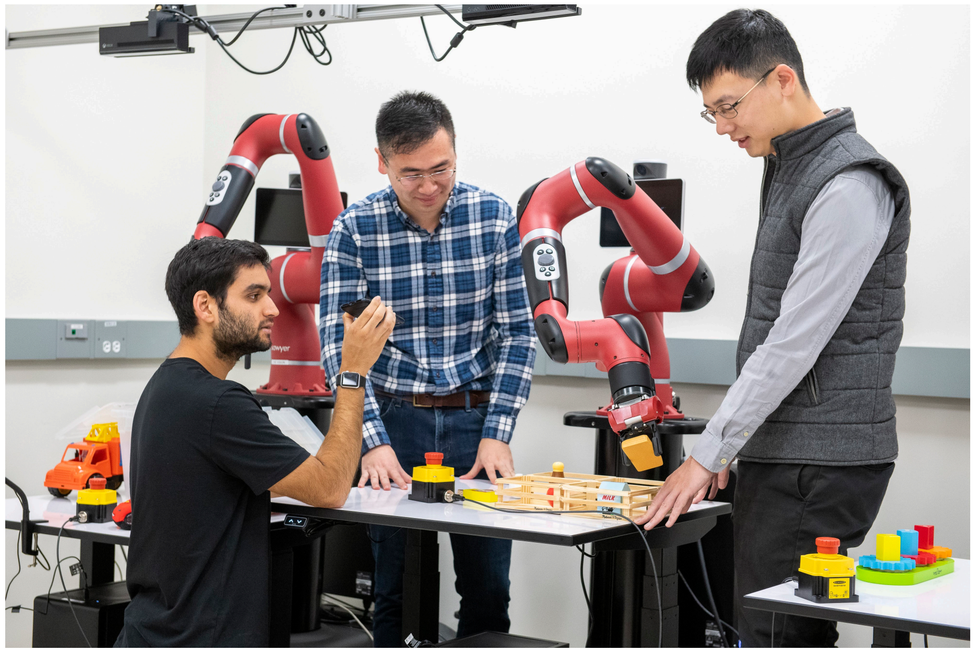
- special-purpose robotic systems
- high engineering cost: over 9,000,000,000 dollars
- decades of teamwork: 20+ years, originally planned for 2013, launched in the final week of 2021
- extremely complex yet reliable system: over 300 are “single points of failure”

"The James Webb Space Telescope — making 300 points of failure reliable" Robert Barron

Two Worlds of Building Robot Autonomy

- general-purpose robotic systems
- intractable to be manually engineered
- months of research effort: small team of graduate students and researchers
- one-off and unreliable system: publish a paper with ~70% success rate and move on to the next one

General-Purpose Collaborative Robots



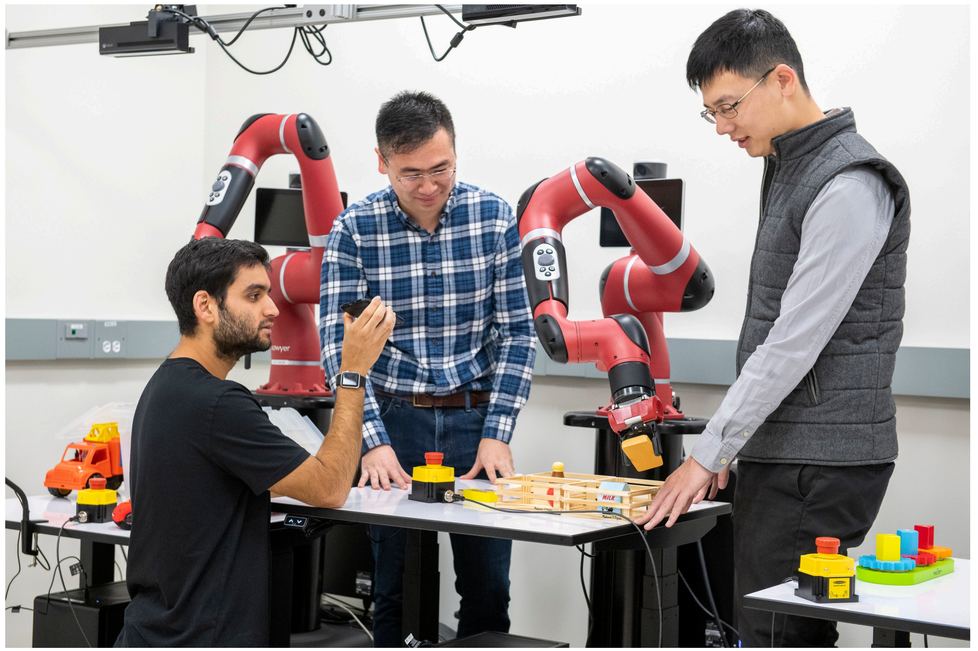
robot learning research

Two Worlds of Building Robot Autonomy

James Webb Space Telescope

robotic systems engineering

General-Purpose Collaborative Robots



robot learning research

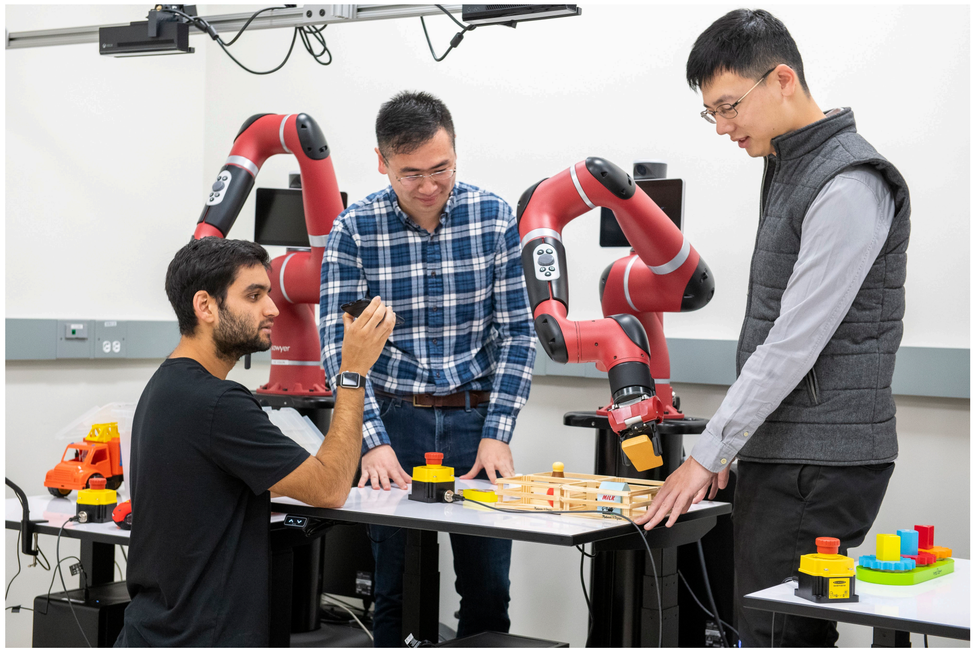
Key to **deployable robot autonomy** in the wild!

James Webb Space Telescope

robotic systems engineering

+

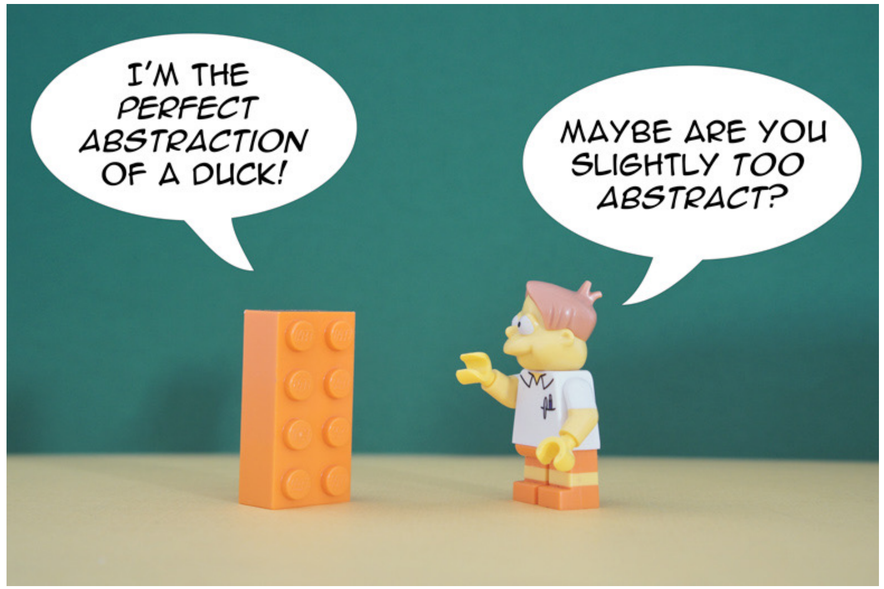
General-Purpose Collaborative Robots



robot learning research

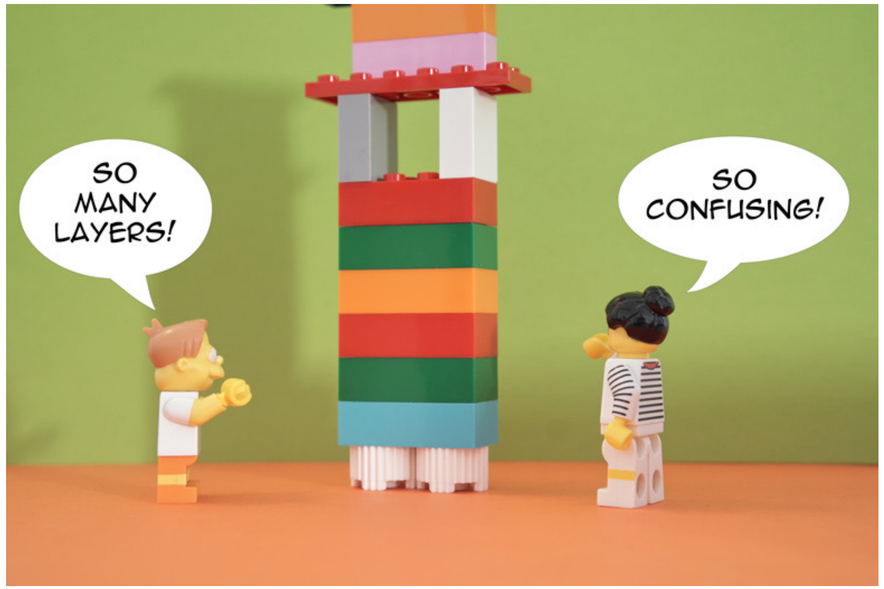
Lessons from Systems Engineering Principles

Abstraction



"division of labor"

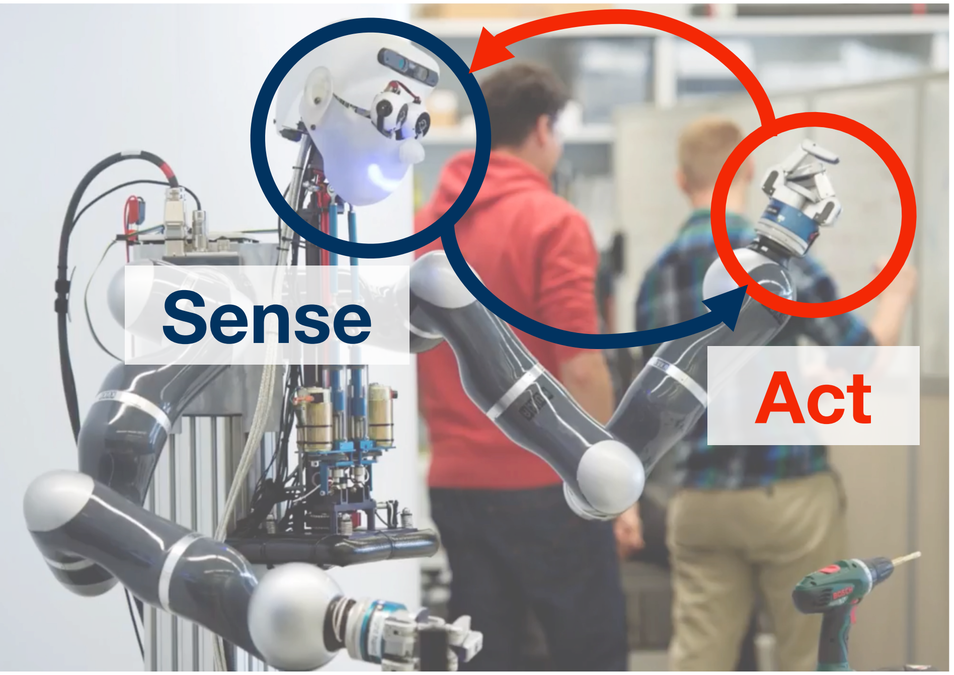
Composition



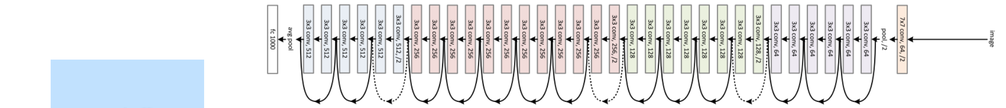
"harmony of labor"

"Types, abstraction, and parametric polymorphism." John Reynolds, 1983
"What Are Abstractions in Software Engineering with Examples." Matthieu Cneude, 2019
"Between abstraction and composition..." Jonathan Sterling, 2021

Abstraction and Composition in Robot Autonomy



π :



$\mathcal{O}$

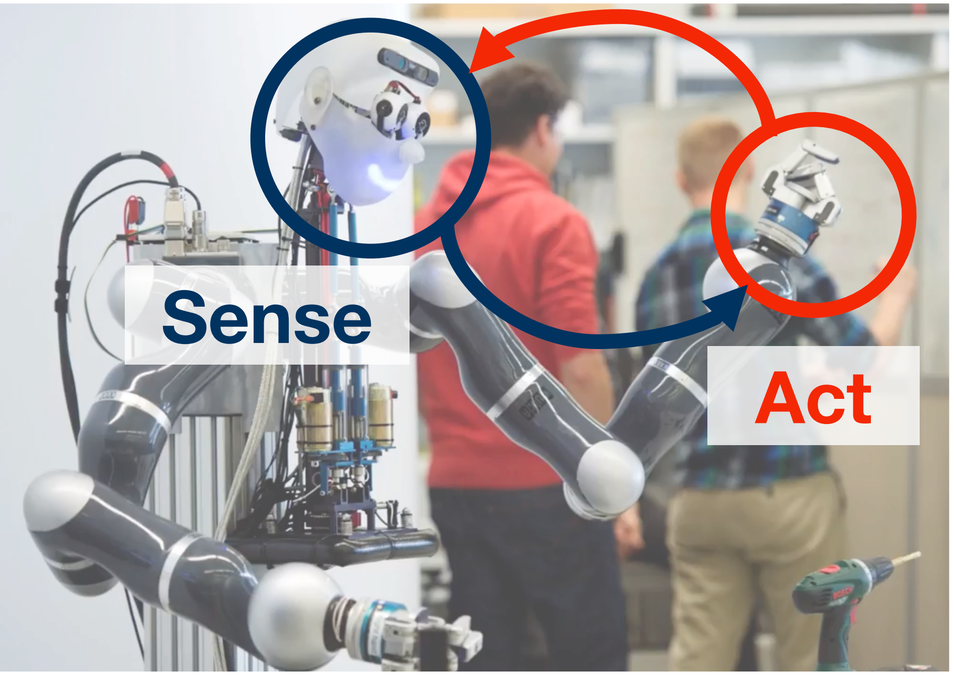
sensory data

$\mathcal{A}$

motor actions

the “pixels to torques” approach

Abstraction and Composition in Robot Autonomy



$$\pi : \mathcal{O} \rightarrow \phi \rightleftharpoons \psi \leftarrow \mathcal{A}$$

state abstraction

action abstraction

sensory data

motor actions

compositional robot autonomy stack

Abstraction and Composition in Robot Autonomy

[Source: Detectron2, FAIR]

composition of objects

$\pi : \mathcal{O} \rightarrow \phi \rightleftharpoons \psi \leftarrow \mathcal{A}$

state abstraction

action abstraction

sensory data

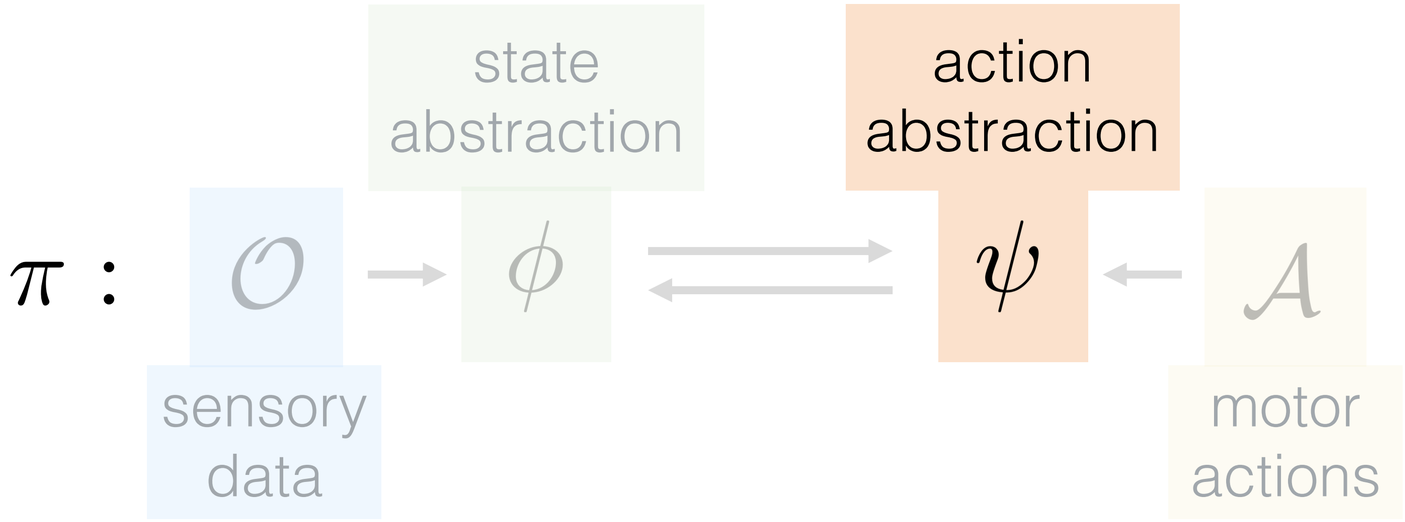
motor actions

compositional robot autonomy stack

Abstraction and Composition in Robot Autonomy

[Source: Daniel M. Wolpert]

composition of skills



compositional robot autonomy stack

Abstraction and Composition in Robot Autonomy

Neural Task Programming (NTP): Hierarchical Policy Learning as Neural Program Induction

Objects: color cubes

Skills: pick and place

→ Strong generalization!

"Neural Task Programming: Learning to Generalize Across Hierarchical Tasks." Xu et al. ICRA 2018

"Neural Task Graphs: Generalizing to Unseen Tasks from a Single Video Demonstration." Huang et al. CVPR 2019

Abstraction and Composition in Robot Autonomy

Neural Task Programming (NTP): Hierarchical Policy Learning as Neural Program Induction

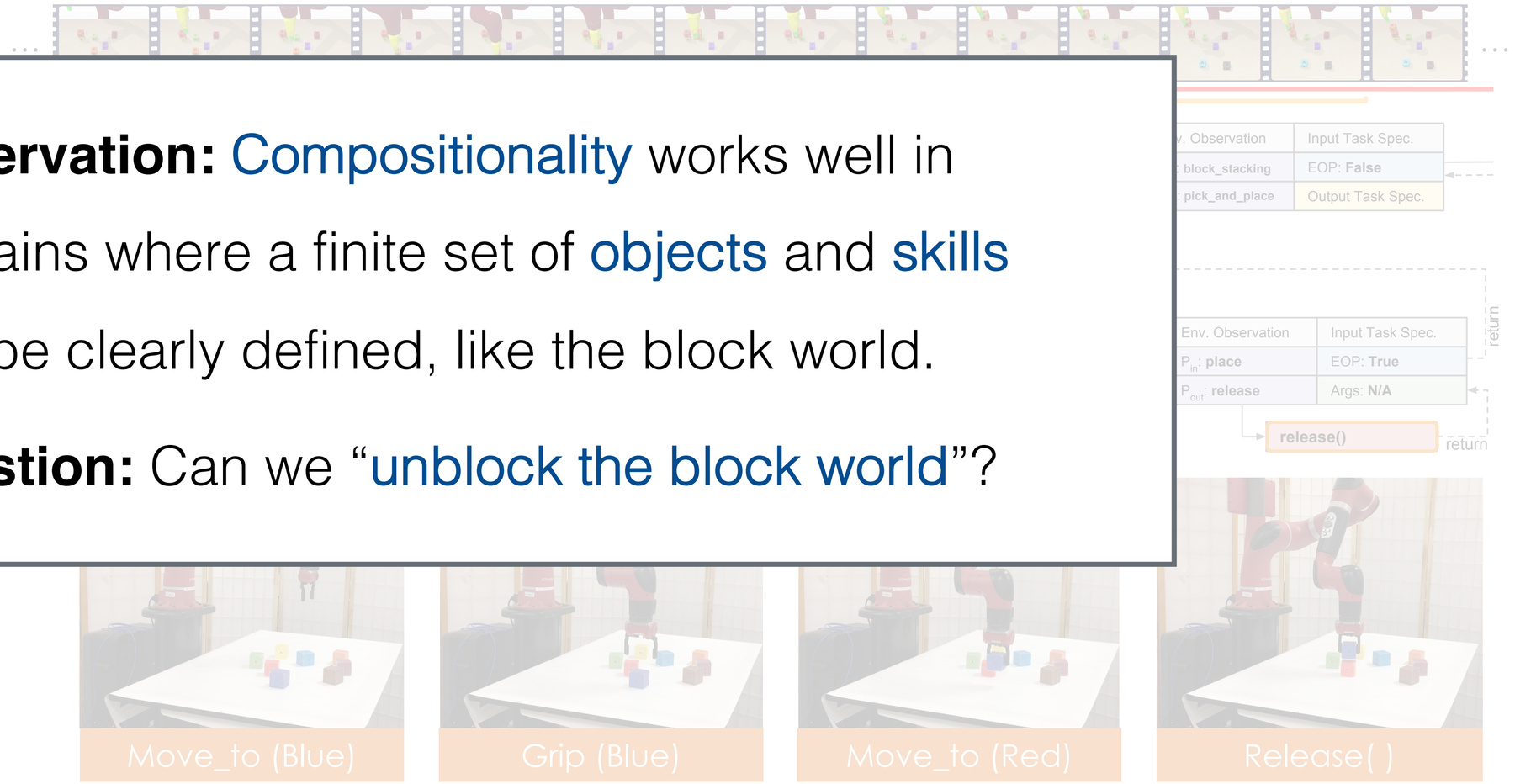
Objects: color b

Skills: pick and

→ Strong generalization

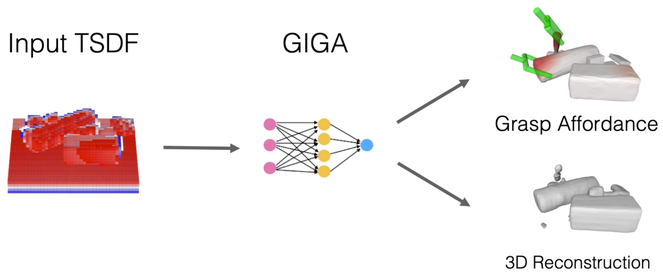
Observation: Compositionality works well in domains where a finite set of objects and skills can be clearly defined, like the block world.

Question: Can we "unblock the block world"?

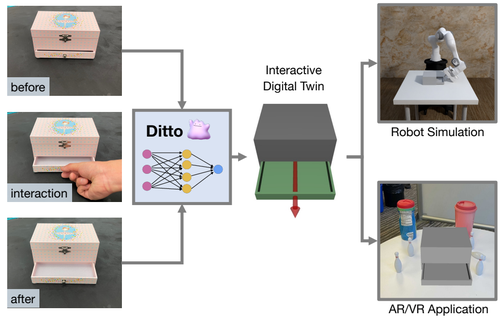


Compositional Robot Autonomy with Objects and Skills

Learning object abstractions from embodied interactions

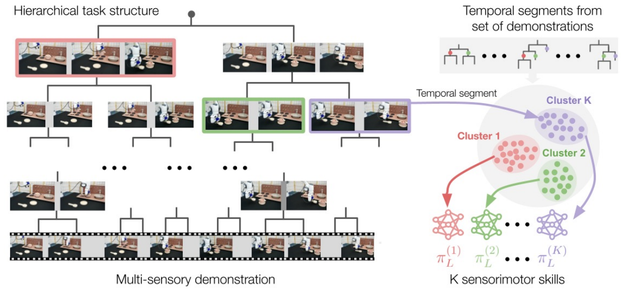


GIGA [Jiang et al., RSS 2021]

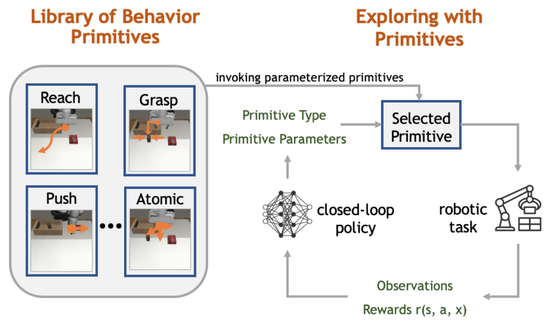


Ditto [Jiang et al., arXiv 2022]

Compositional task modeling with skill abstractions



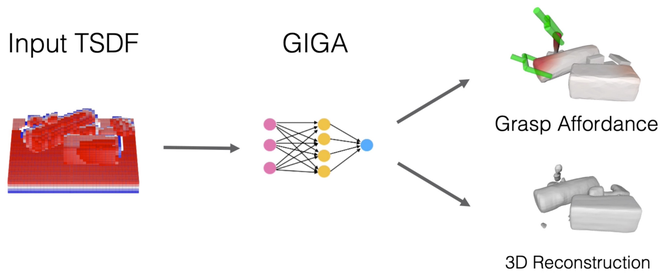
BUDS [Zhu et al., RA-L 2022]



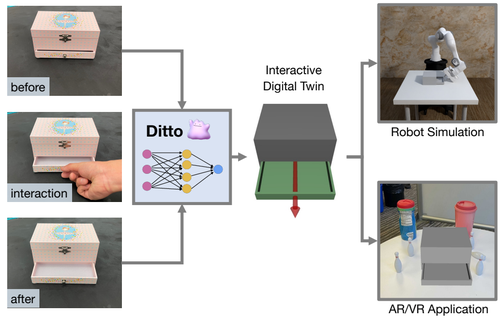
MAPLE [Nasiriany et al., ICRA 2022]

Compositional Robot Autonomy with Objects and Skills

Learning object abstractions from embodied interactions

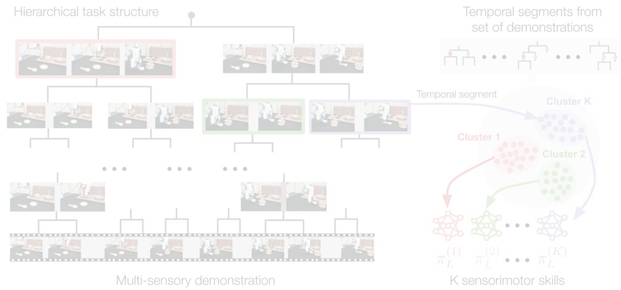


GIGA [Jiang et al., RSS 2021]

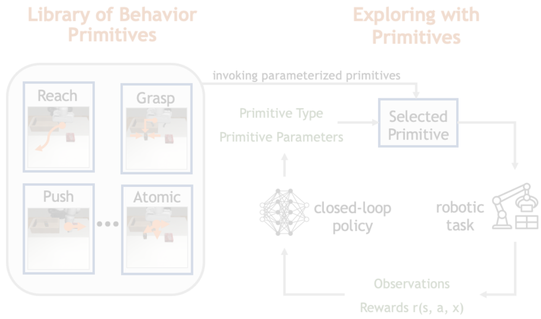


Ditto [Jiang et al., arXiv 2022]

Compositional task modeling with skill abstractions

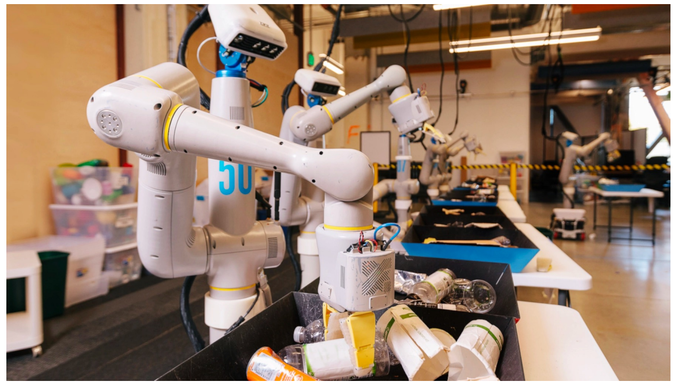


BUDS [Zhu et al., RA-L 2022]



MAPLE [Nasiriany et al., ICRA 2022]

Object Representations for Robotic Grasping



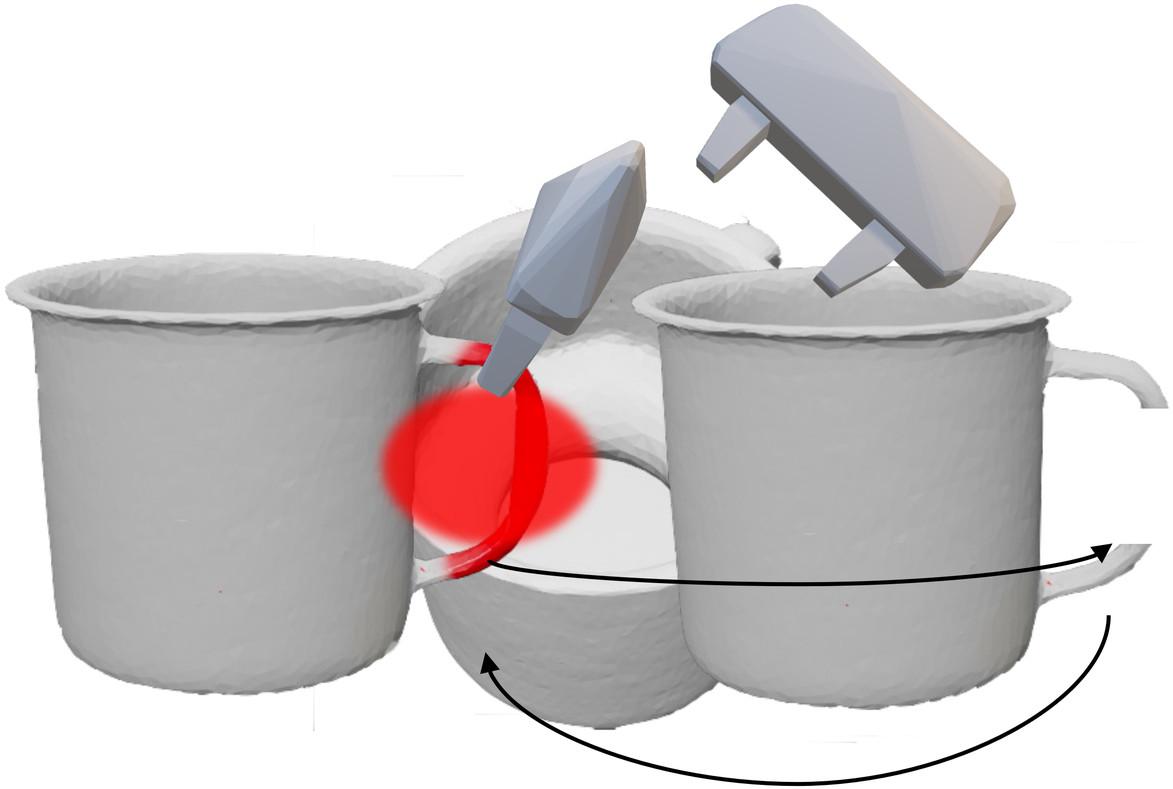
- **Input:** Partial point cloud of workspace
- **Output:** 6-DoF grasp pose (3D position and orientation)

Affordance and geometry reasoning are connected.

Likelihood of grasp success
and grasp parameters

Affordance

Predict affordance of
reconstructed part

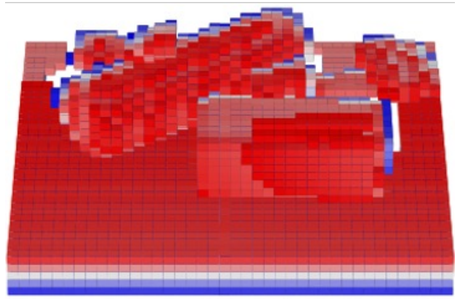


Geometry

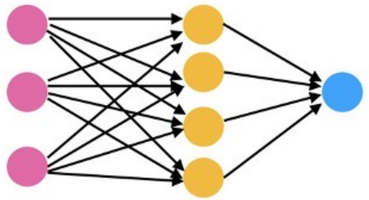
Reconstruct
graspable region

GIGA: Grasping via Implicit Geometry and Affordance

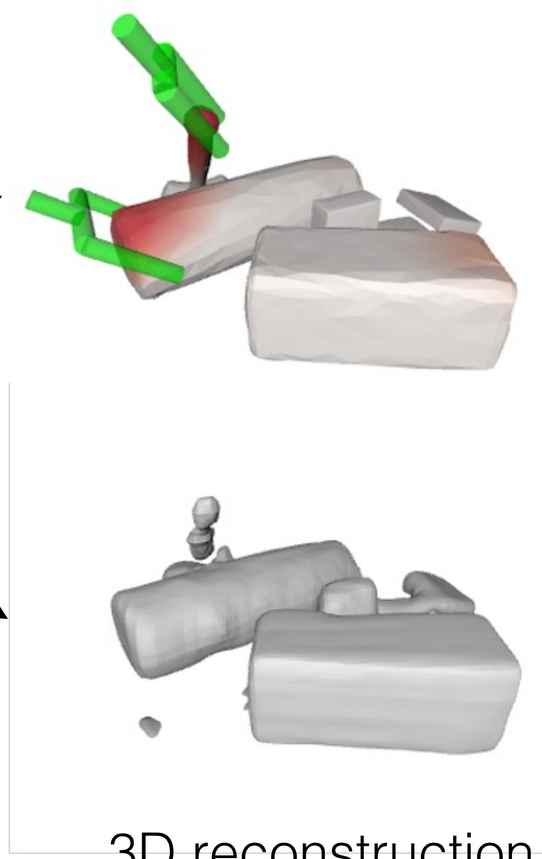
Input TSDF



GIGA



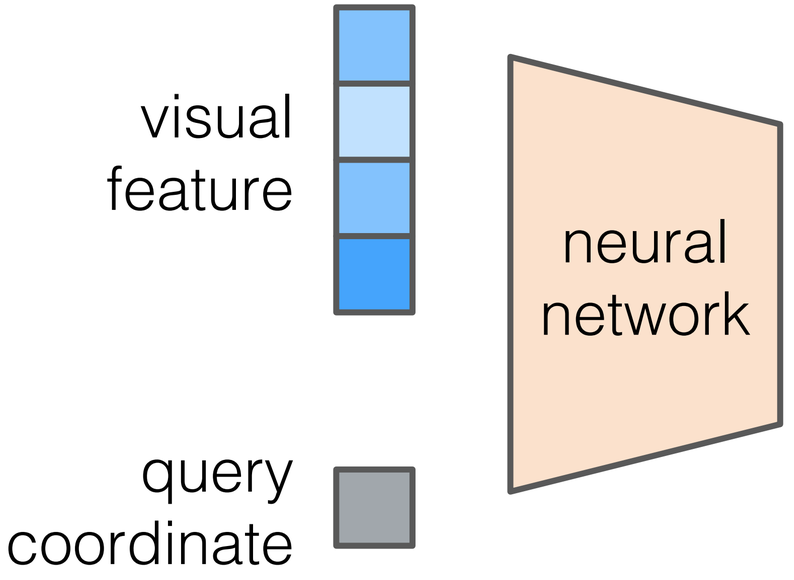
Grasp affordance

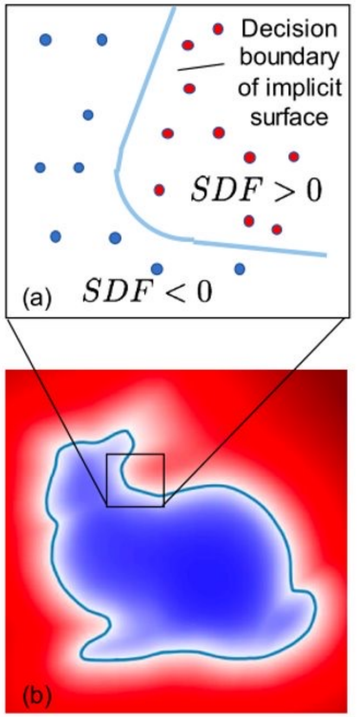


3D reconstruction

Neural Fields for 3D Scenes

- Neural networks that map from coordinate to quantities (SDF, occupancy)
- Smooth, continuous encoding of 3D scenes
- Fully differentiable models that can be trained with rich supervisions

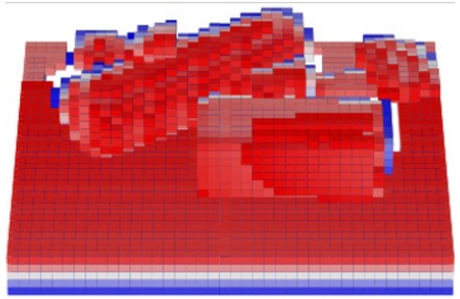




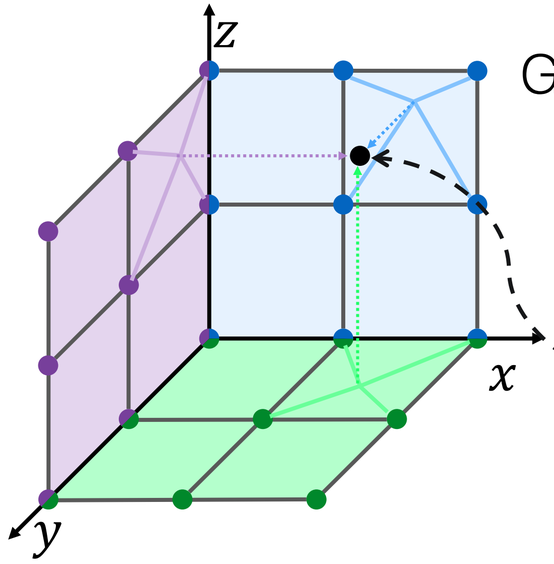
DeepSDF, Park et al. CVPR 2019

GIGA: Grasping via Implicit Geometry and Affordance

Input TSDF



GIGA Model



Structured feature grids

Grasp center (x, y, z)

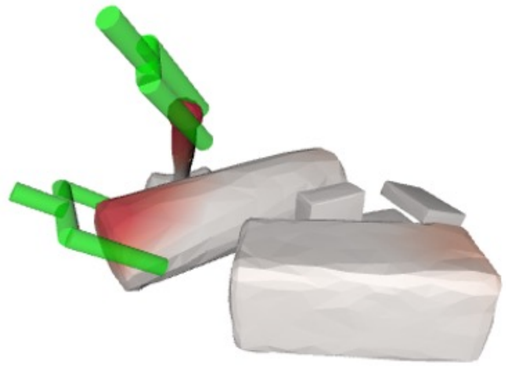
(x', y', z') Query point

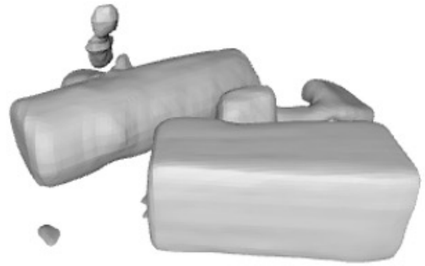
local feature

Affordance Decoder

Geometry Decoder

Grasp affordance



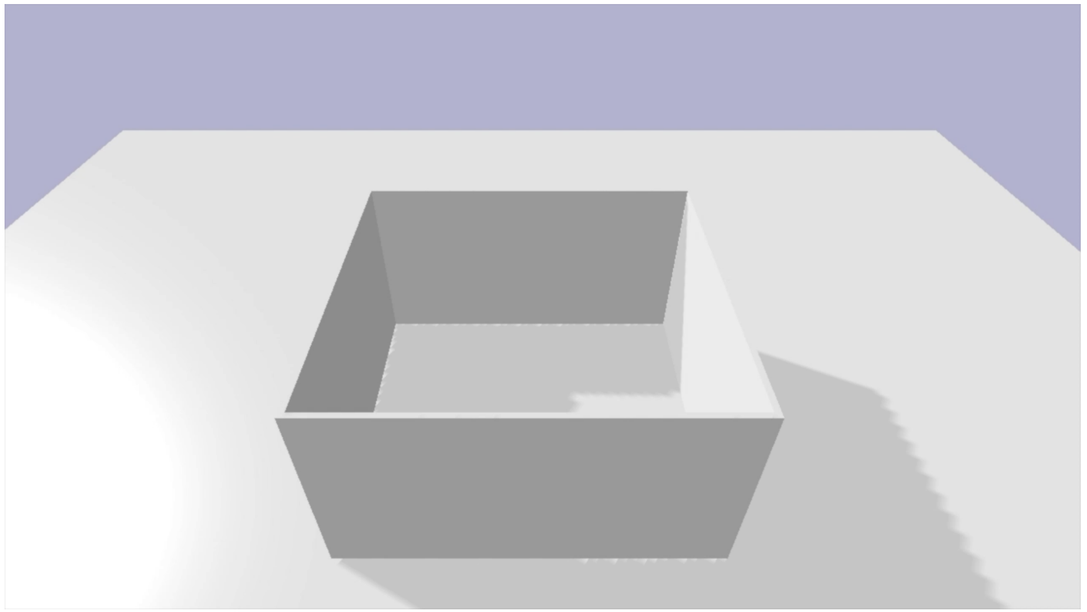


3D reconstruction

Self-Supervised Data Collection

- Affordance labels from grasp trials in simulation.
- 3D geometry labels from ground-truth object meshes.

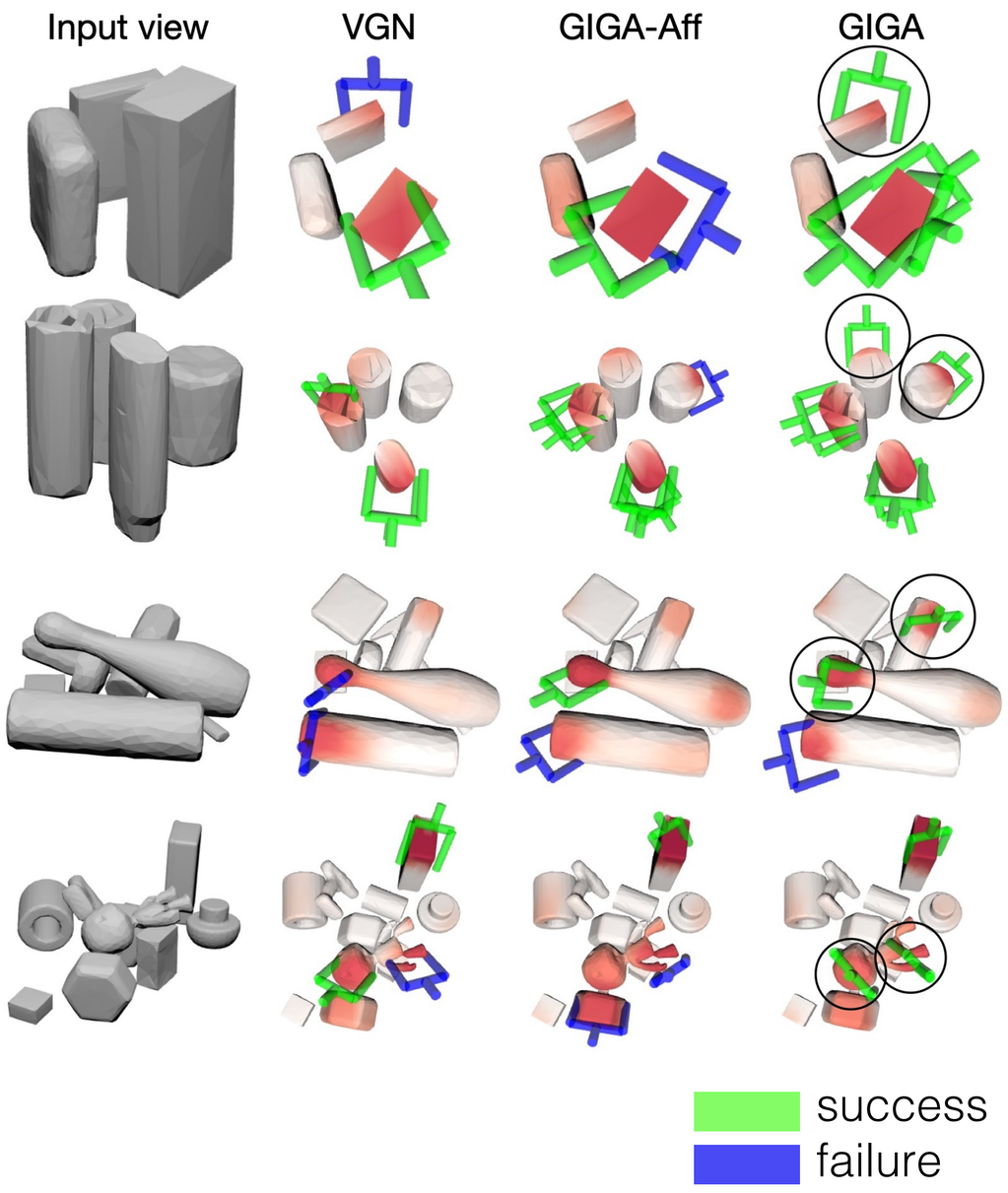
Packed scene



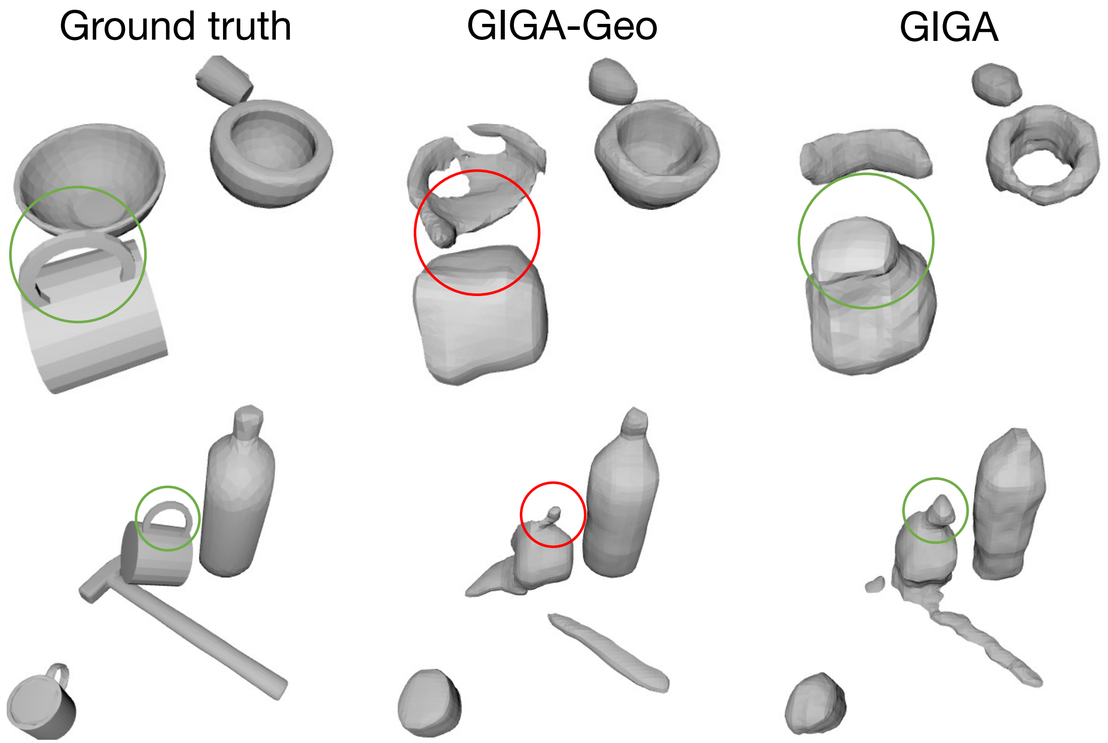
Pile scene

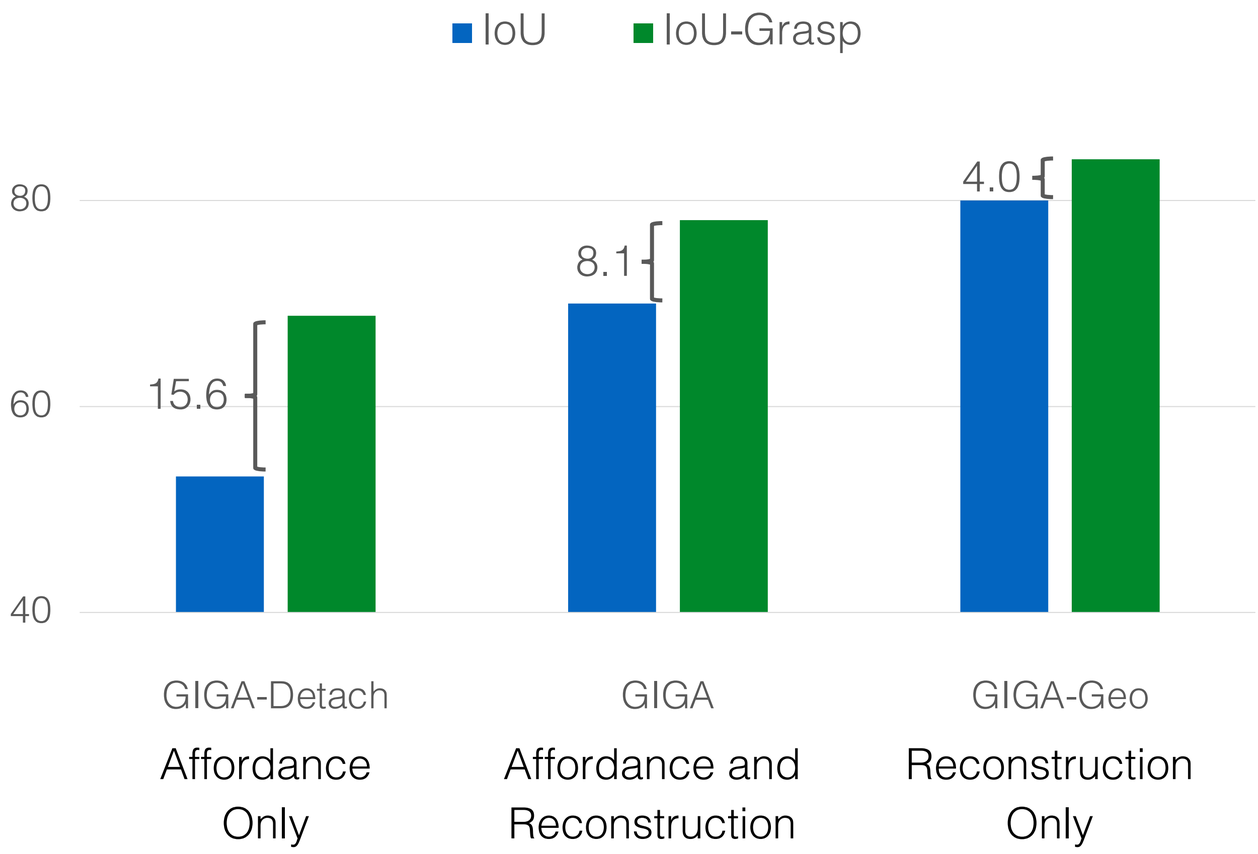

Packed: 30x

Geometry Learning Facilitates Occluded Grasps.



Reconstruction Focuses on Graspable Parts





GIGA: Learning Object Representations

- Task-dependent: optimized for downstream grasping tasks
- Multimodal: captures synergies between affordance and geometry for robotic grasping.
- Structured: combines 3D voxel grids with neural field methods
- Self-supervised: trained with self-supervised interactions

Low Affordance

High Affordance

Zhenyu Jiang

"Synergies Between Affordance and Geometry: 6-DoF Grasp Detection via Implicit Representations." Jiang, Zhu, Svetlik, Fang, and Z.

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

Can we build object models from interaction?

Interaction creates novel sensory stimuli for object learning.

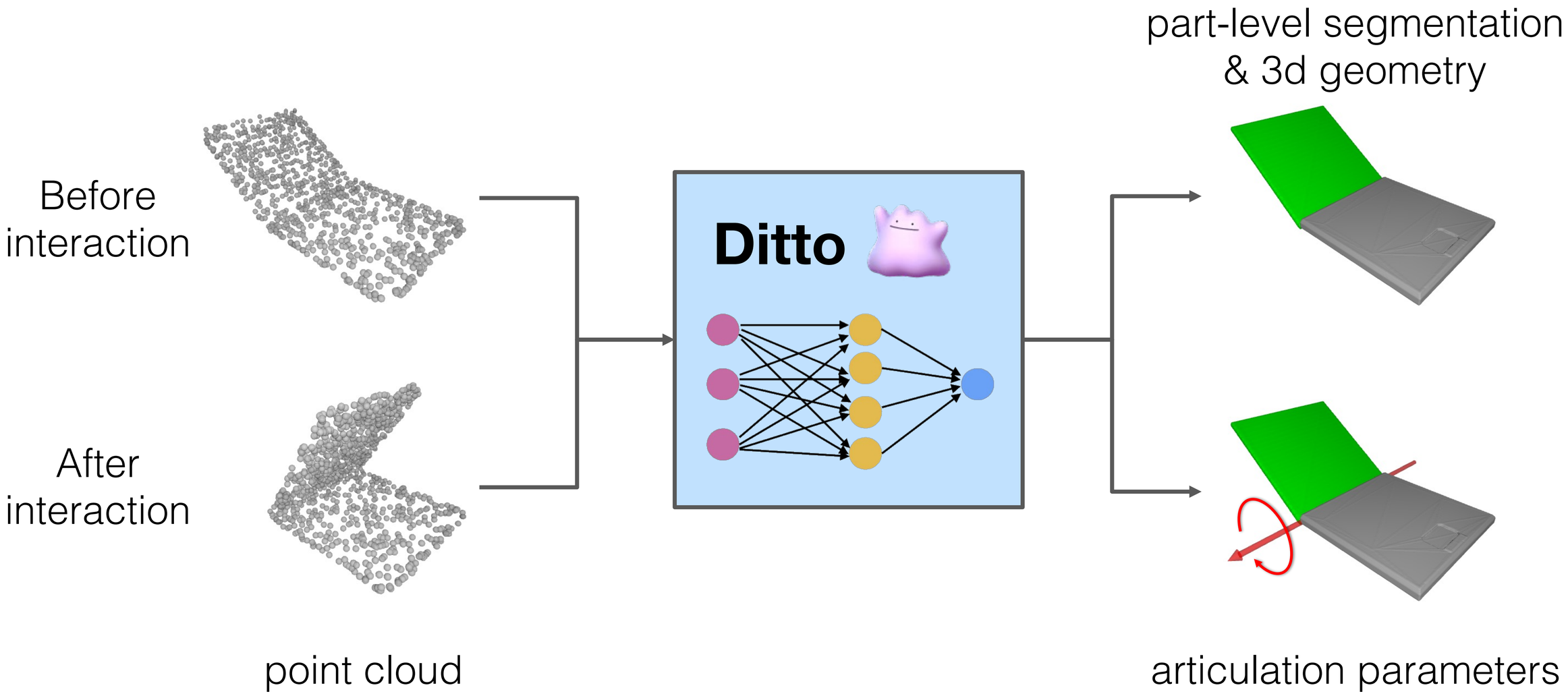
Can we build object models from interaction?

digital twin
creation

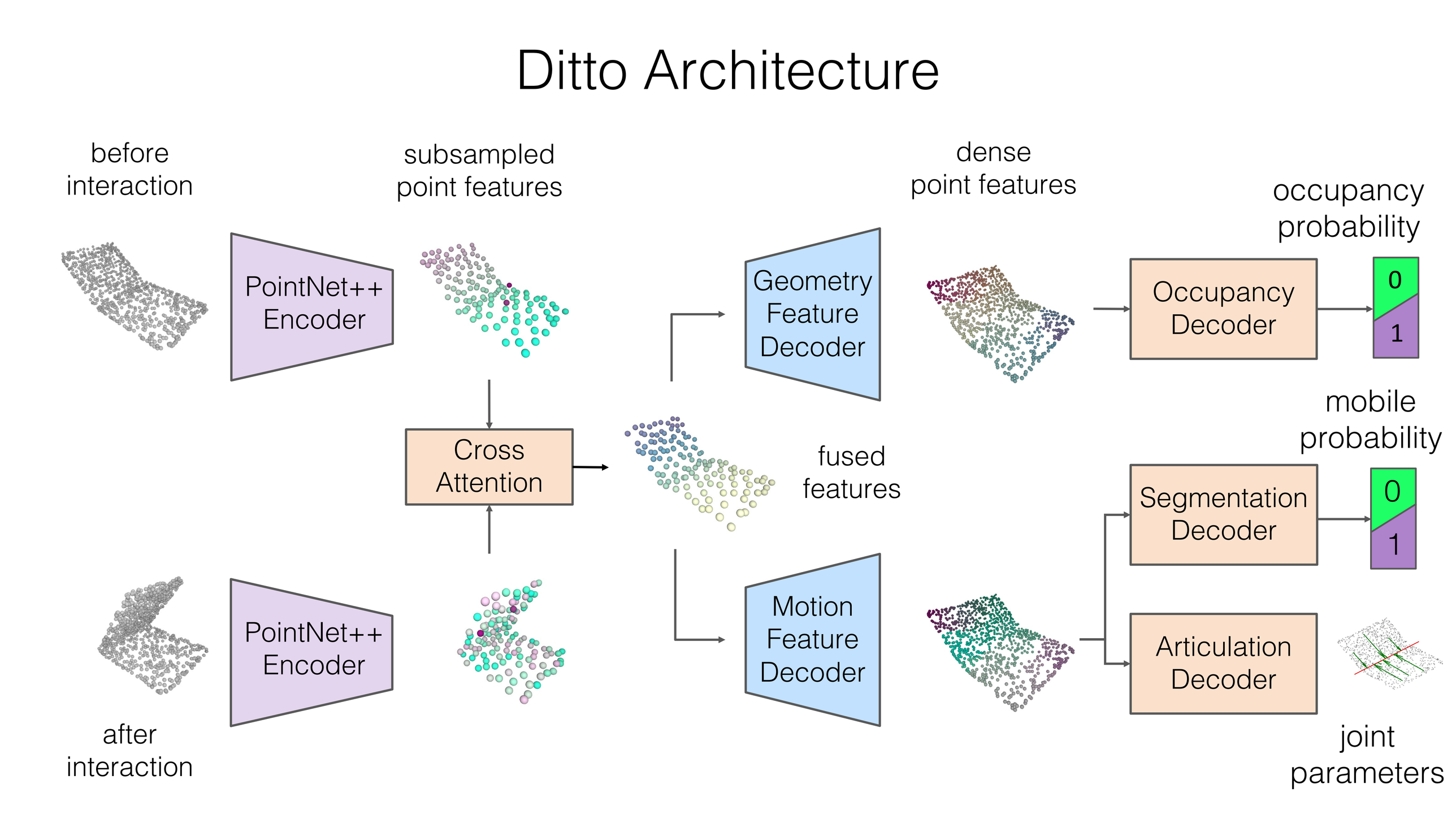
physical object
in universe

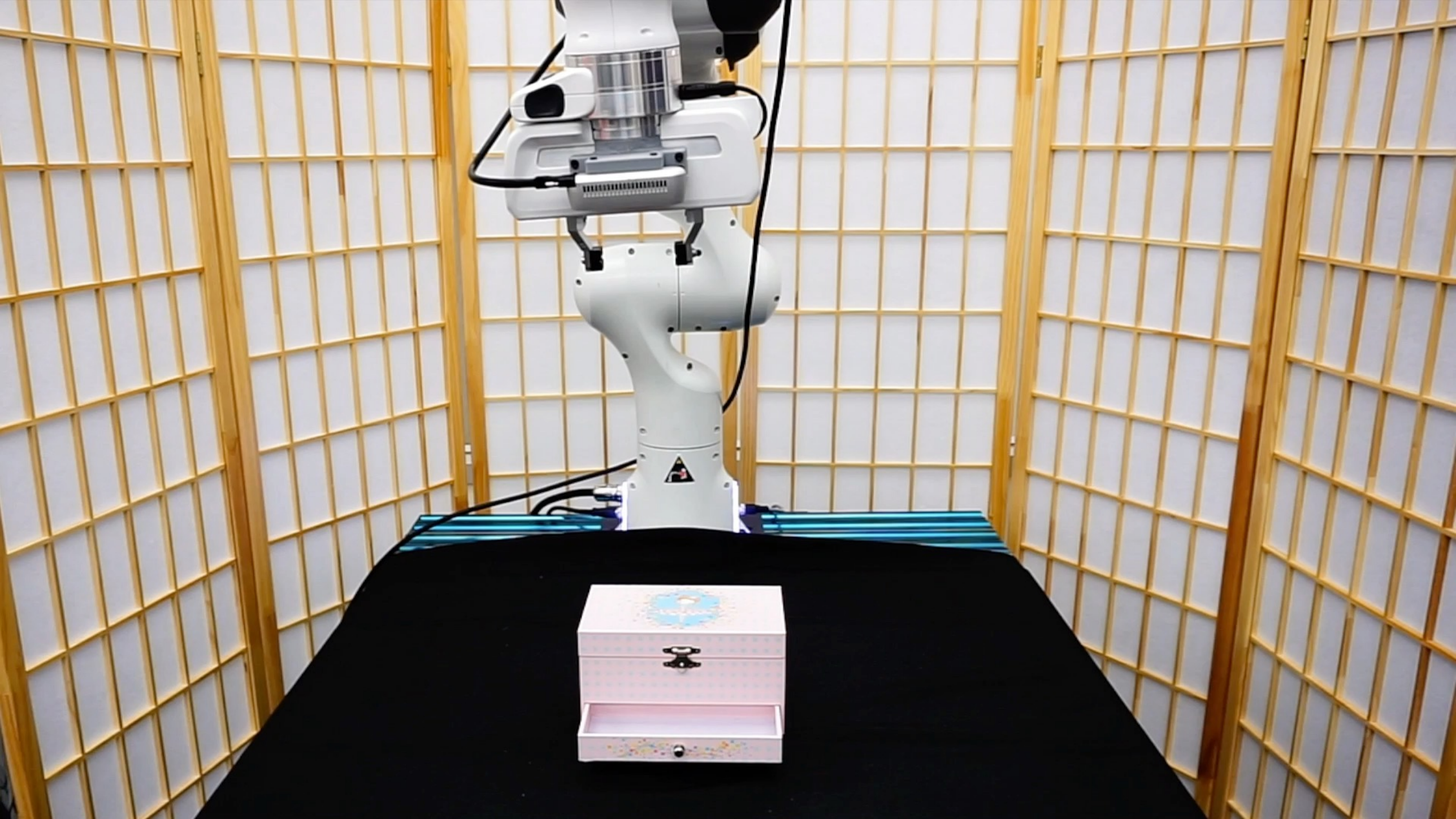
virtual object
in “metaverse”

Ditto: Digital Twin of Articulated Objects

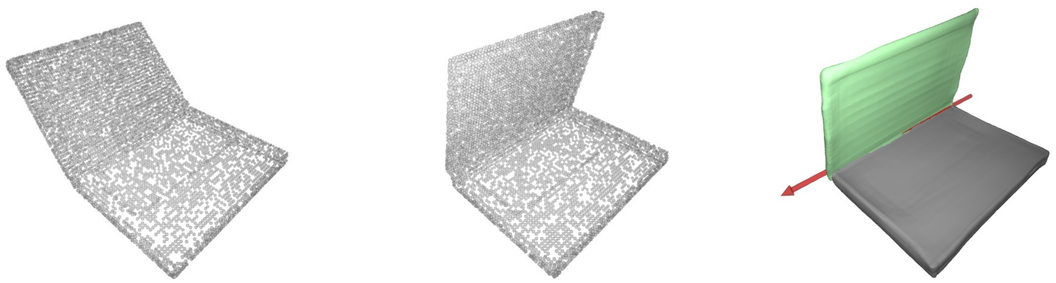


Ditto Architecture

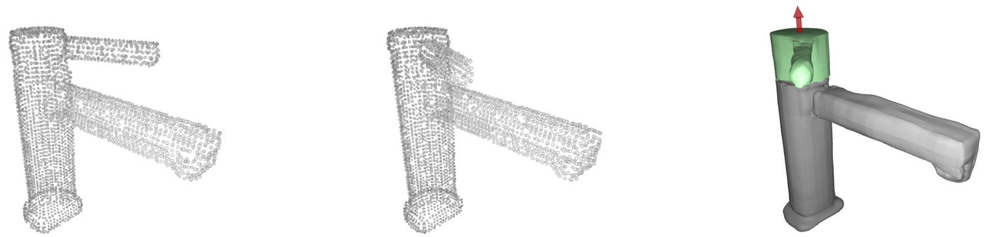




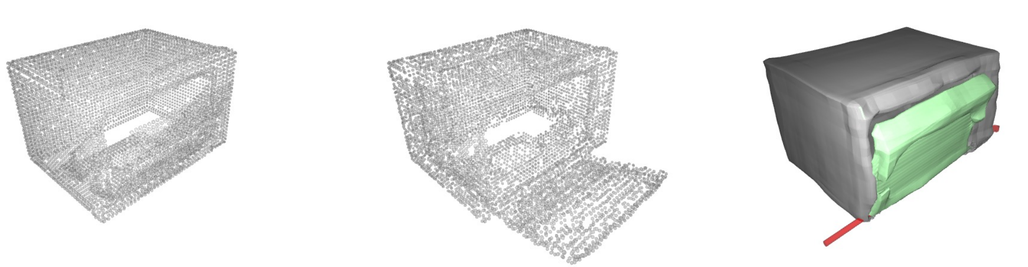
Recreated Digital Twins of Articulated Objects



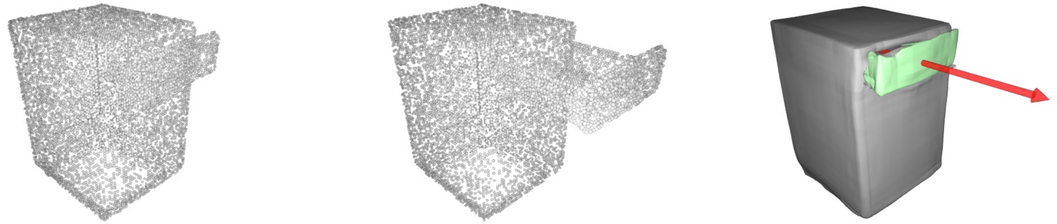
laptop



faucet



oven

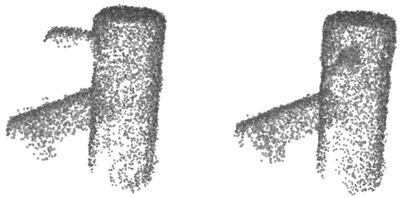


drawer

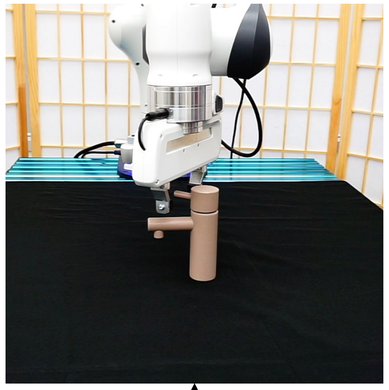
From Real World to Simulation and Back

physical world

Real-World
Observations



Sensing

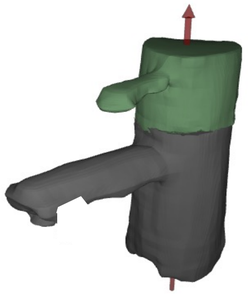


Real-World
Manipulation

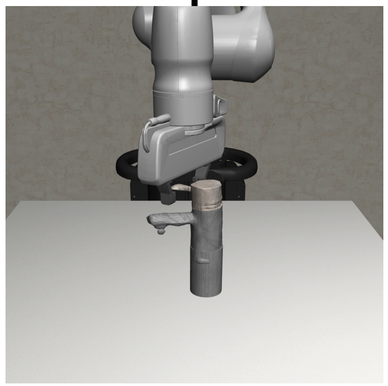
Recreation

Transfer

Digital
Twin



Simulation



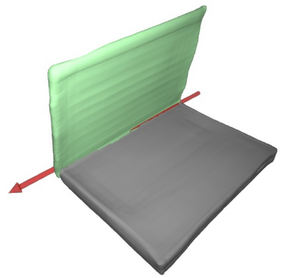
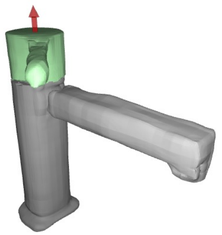
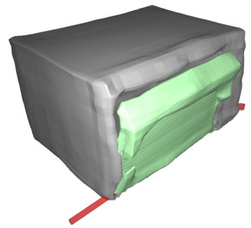
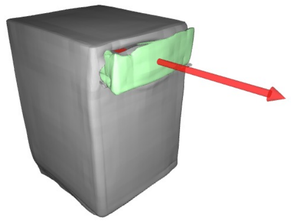
Physics
Engine

virtual world

https://robosuite.ai

Ditto: Building Interactive Object Models

- Embodied interactions: Emit useful sensory information for understanding an object
- Structured implicit neural representations: Jointly encode geometry and articulation
- Digital twins: Bridge simulation and the real world

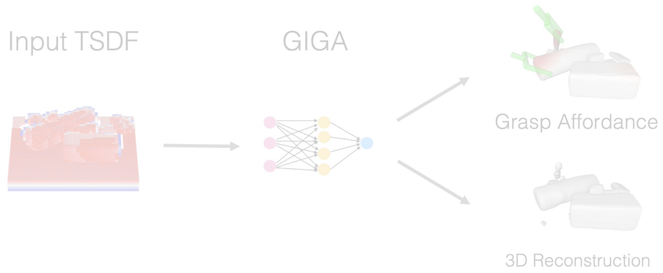






Zhenyu Jiang

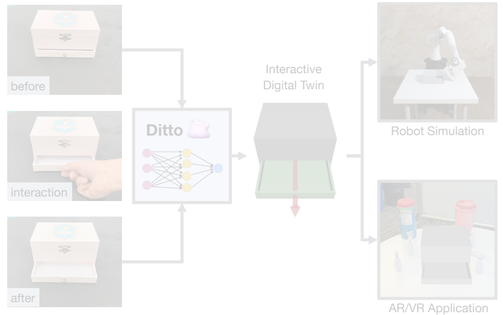
"Building Digital Twins of Articulated Objects from Interaction." Jiang, Hsu, and Z.
Code and models are available at https://ut-austin-rpl.github.io/Ditto

Compositional Robot Autonomy with Objects and Skills

Learning object abstractions from embodied interactions

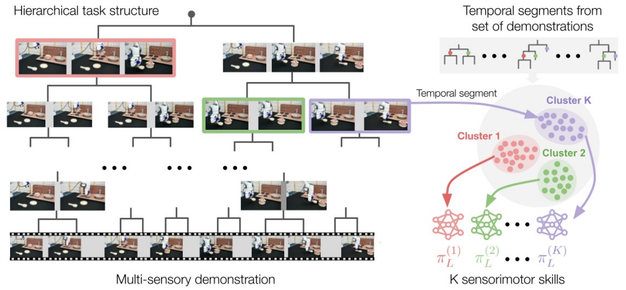


GIGA [Jiang et al., RSS 2021]

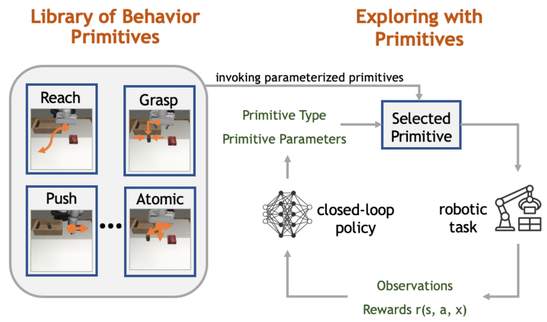


Ditto [Jiang et al., arXiv 2022]

Compositional task modeling with skill abstractions

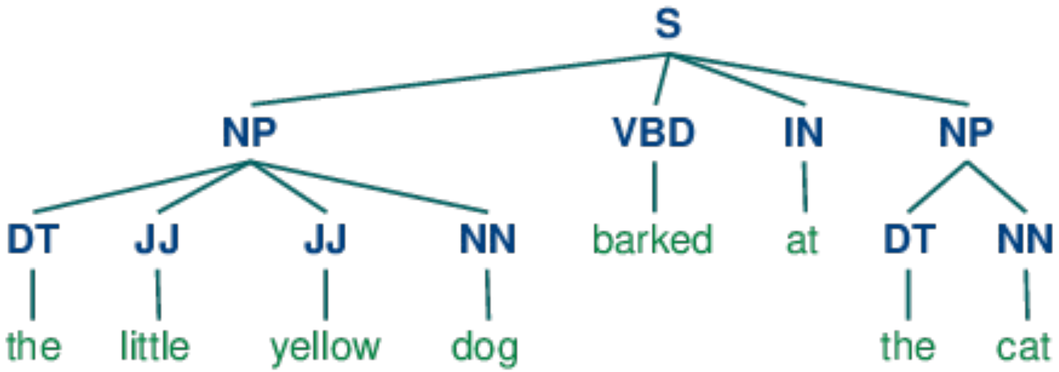


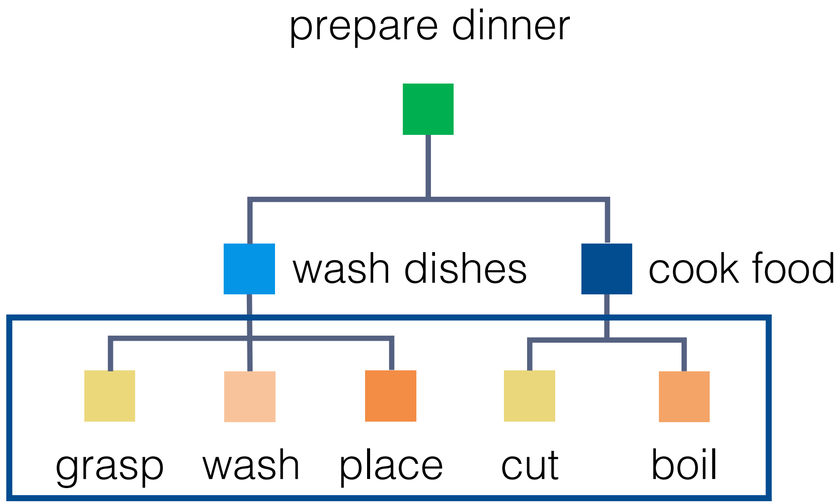
BUDS [Zhu et al., RA-L 2022]



MAPLE [Nasiriany et al., ICRA 2022]

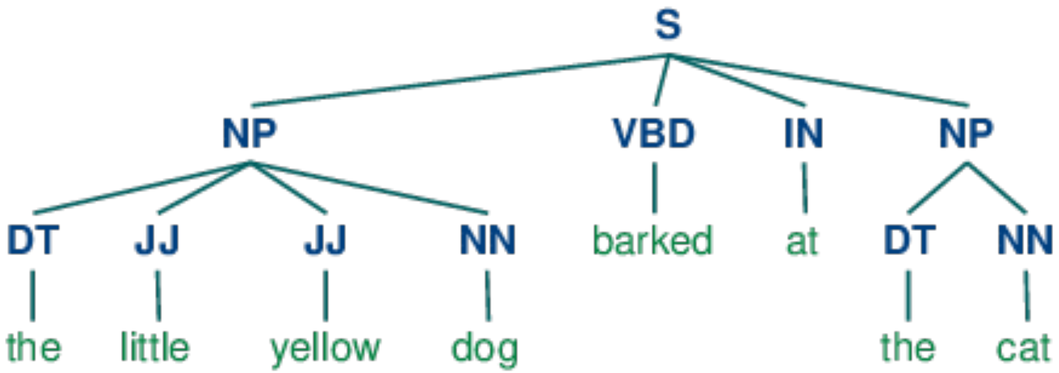
Compositionality in Language and Behaviors

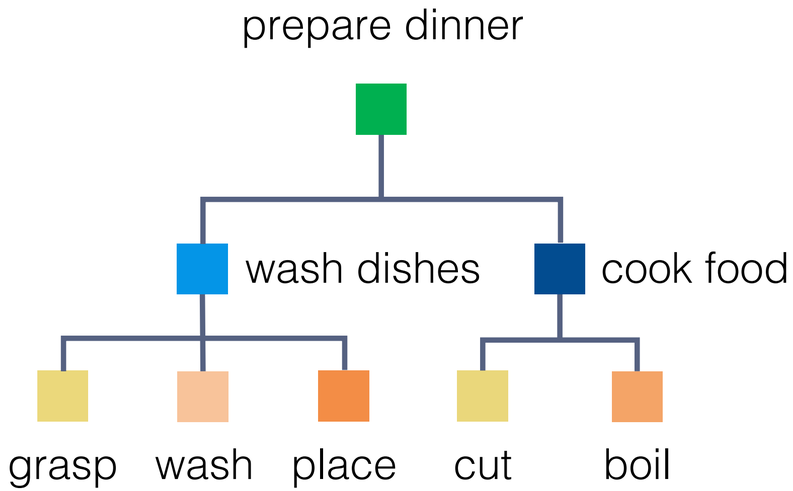




sensorimotor skills: perceptually grounded, temporally extended behaviors

Compositionality in Language and Behaviors



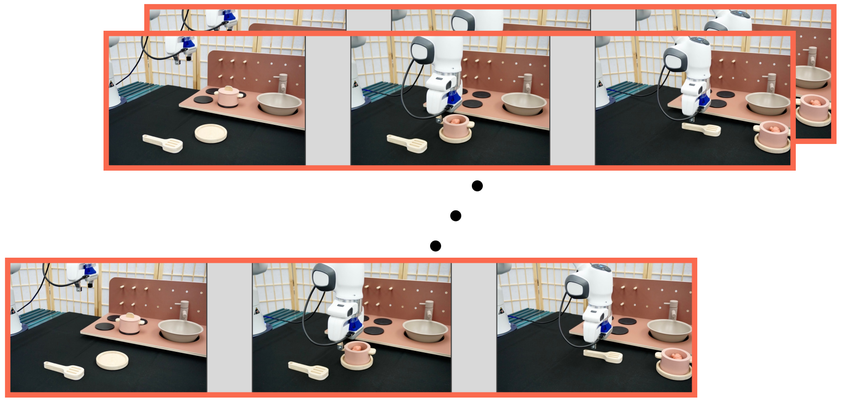


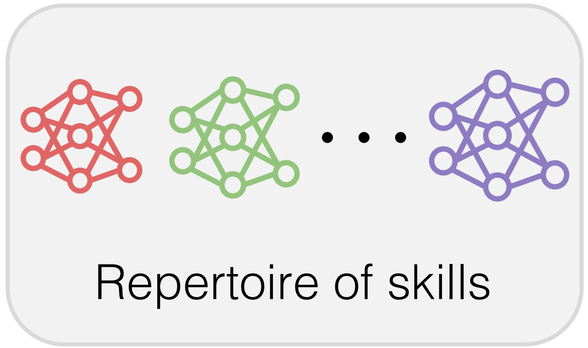
The Context Principle: words have meaning only as constituents of (hence, presumably, only in virtue of their use in) sentences

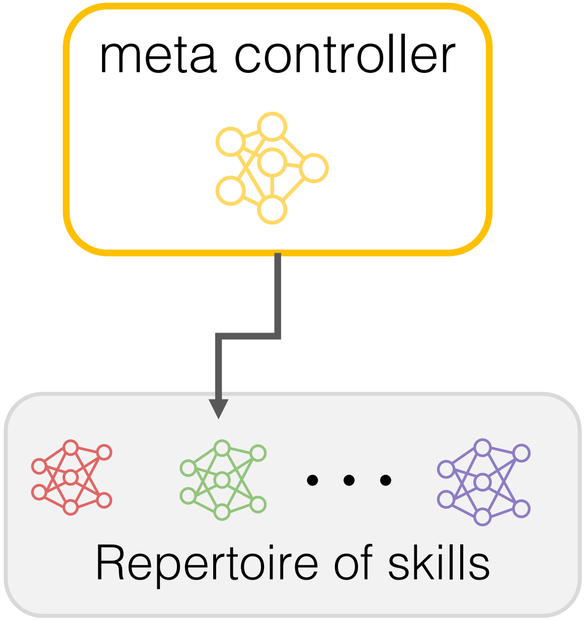
The Compositionality Principle: the meaning of the whole sentence is a function of the meanings of its parts

"How to Stop Worrying About Compositionality", Aurelie Herbelot

BUDS: Bottom-Up Discovery of sensorimotor Skills







Unstructured demonstrations

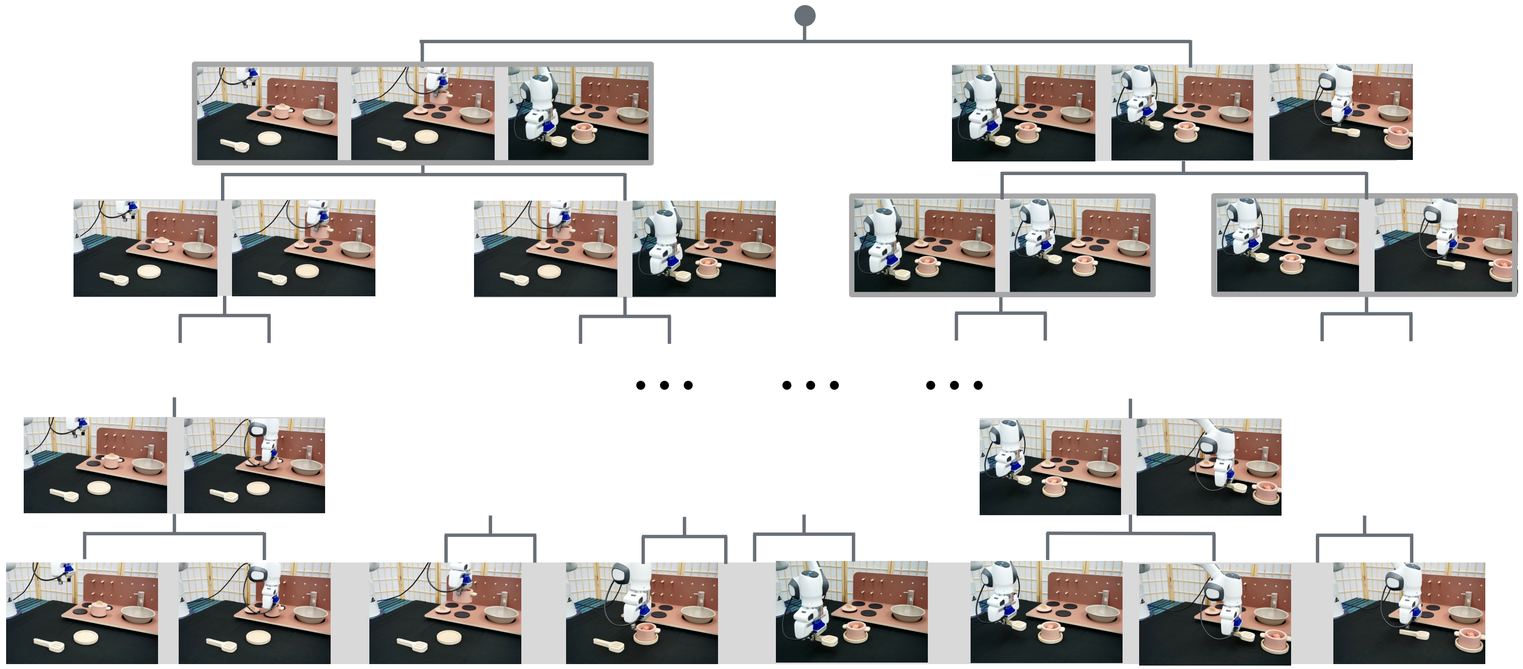
Sensorimotor skills

Composition of skills

- No temporal annotations
- Operating on raw sensory data

BUDS: Bottom-Up Discovery of sensorimotor Skills

Hierarchical task structure

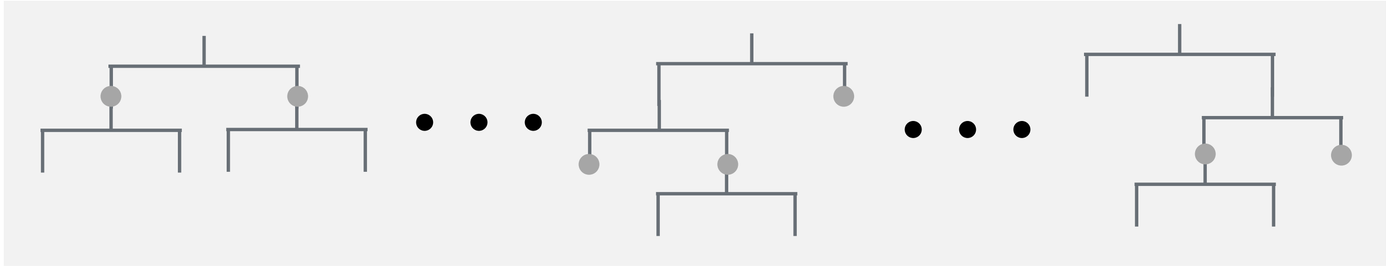


Unstructured demonstration

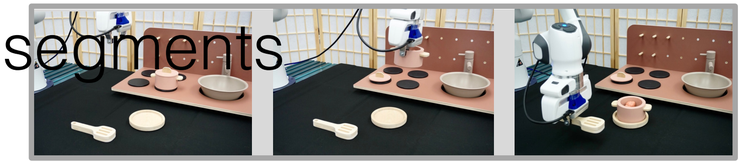
Bottom-up approach

- Building hierarchical task structures
- Agglomerative clustering on temporal segments w/ multi-sensory cues
- Identify reusable skills from recurring temporal patterns

Temporal segments from set of demonstrations

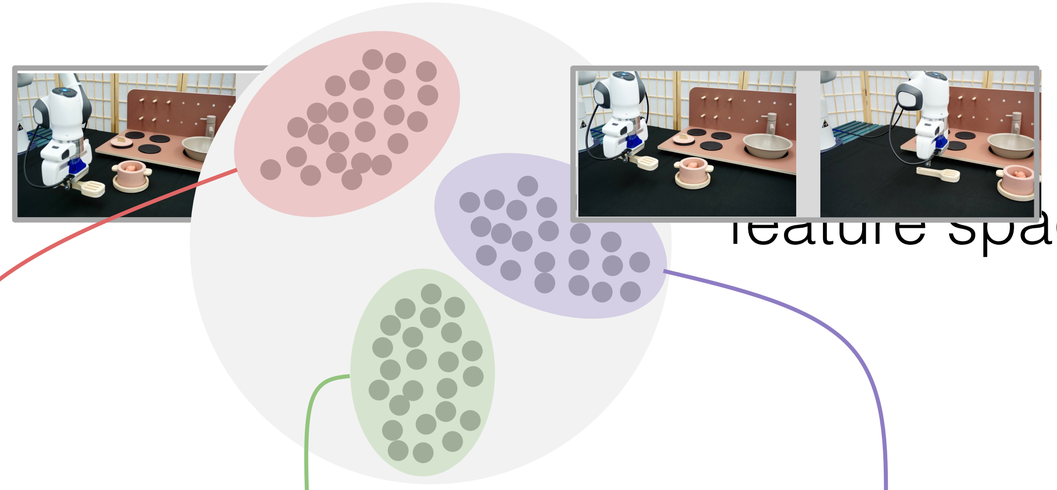


Temporal segments



Compose Skills to Solve Tasks

Cluster segments into skills



feature space

image input

meta controller

Learn each skill

Skill Index

Subgoal

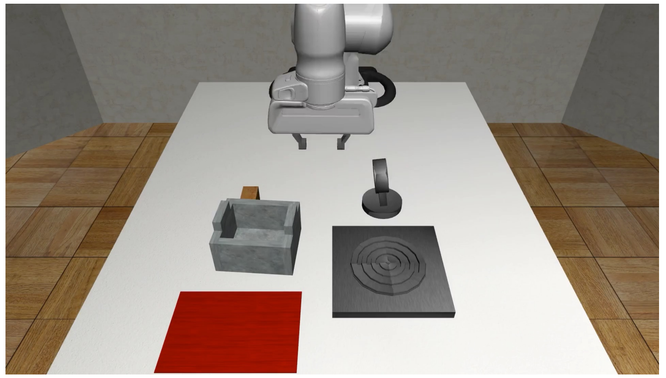
Compute feature for each temporal segment

$\pi_L^{(1)}$ $\pi_L^{(2)}$ . . . $\pi_L^{(K)}$

K sensorimotor skills

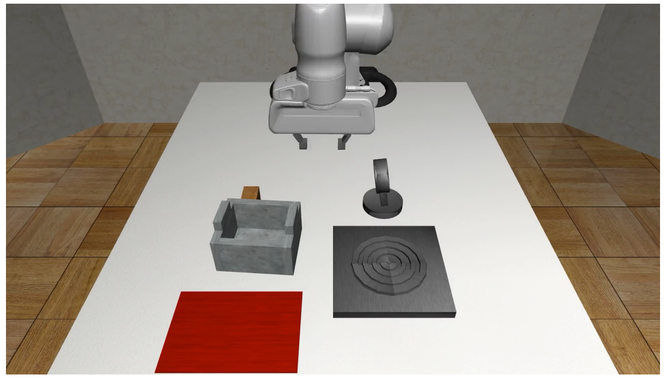
Learned skills improve task success.

Behavioral Cloning
(Zhang et al. 2018)



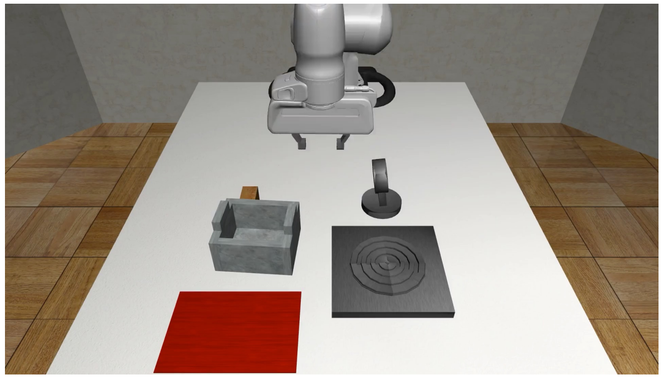
24.4% success

CHAMP
(Niekum et al. 2015)



23.4% success

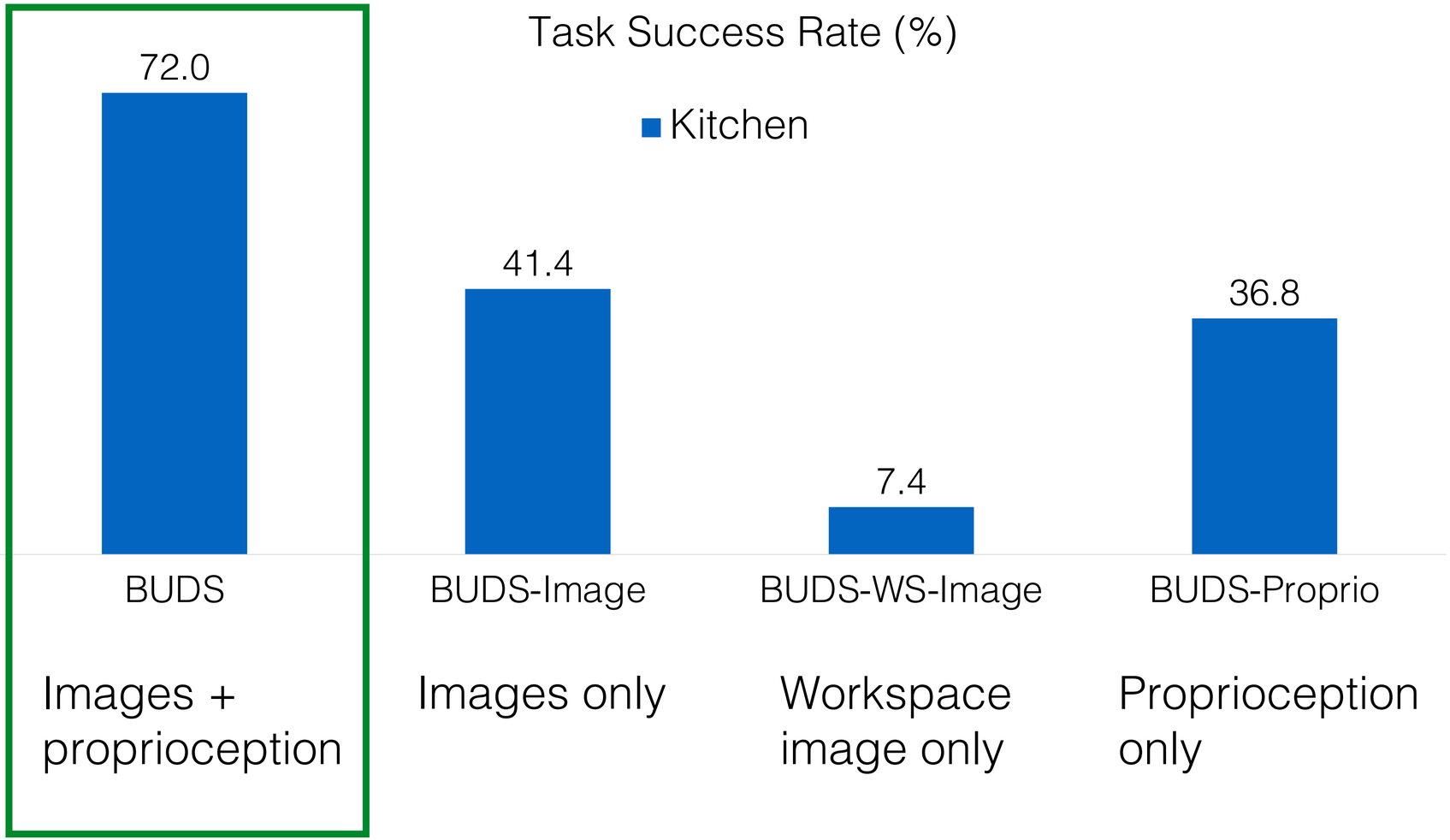
BUDS
(Ours)



72.0% success

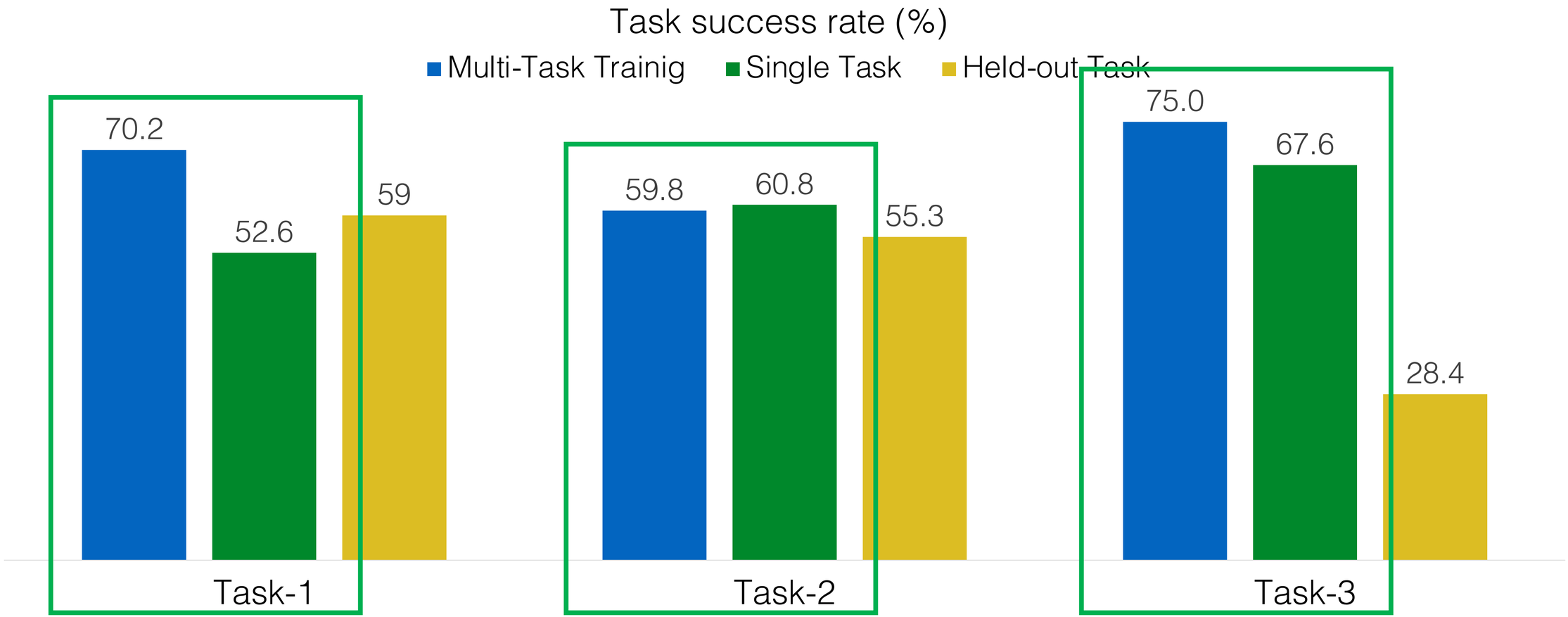
- Trained on 30min demonstrations for each task

Multi-sensory cues improve skill segmentation.



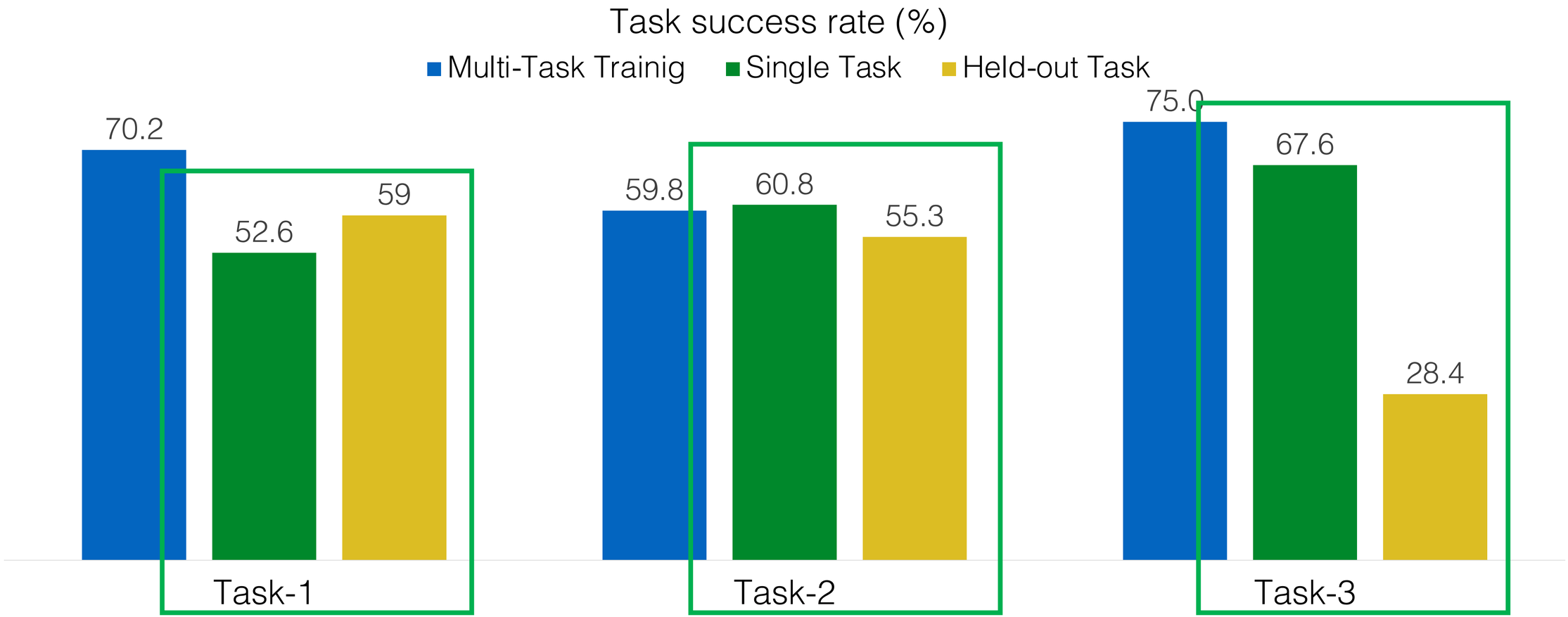
"Event Structure in Perception and Conception." Zacks and Tversky, 2001

Multi-task learning improves quality of skills.



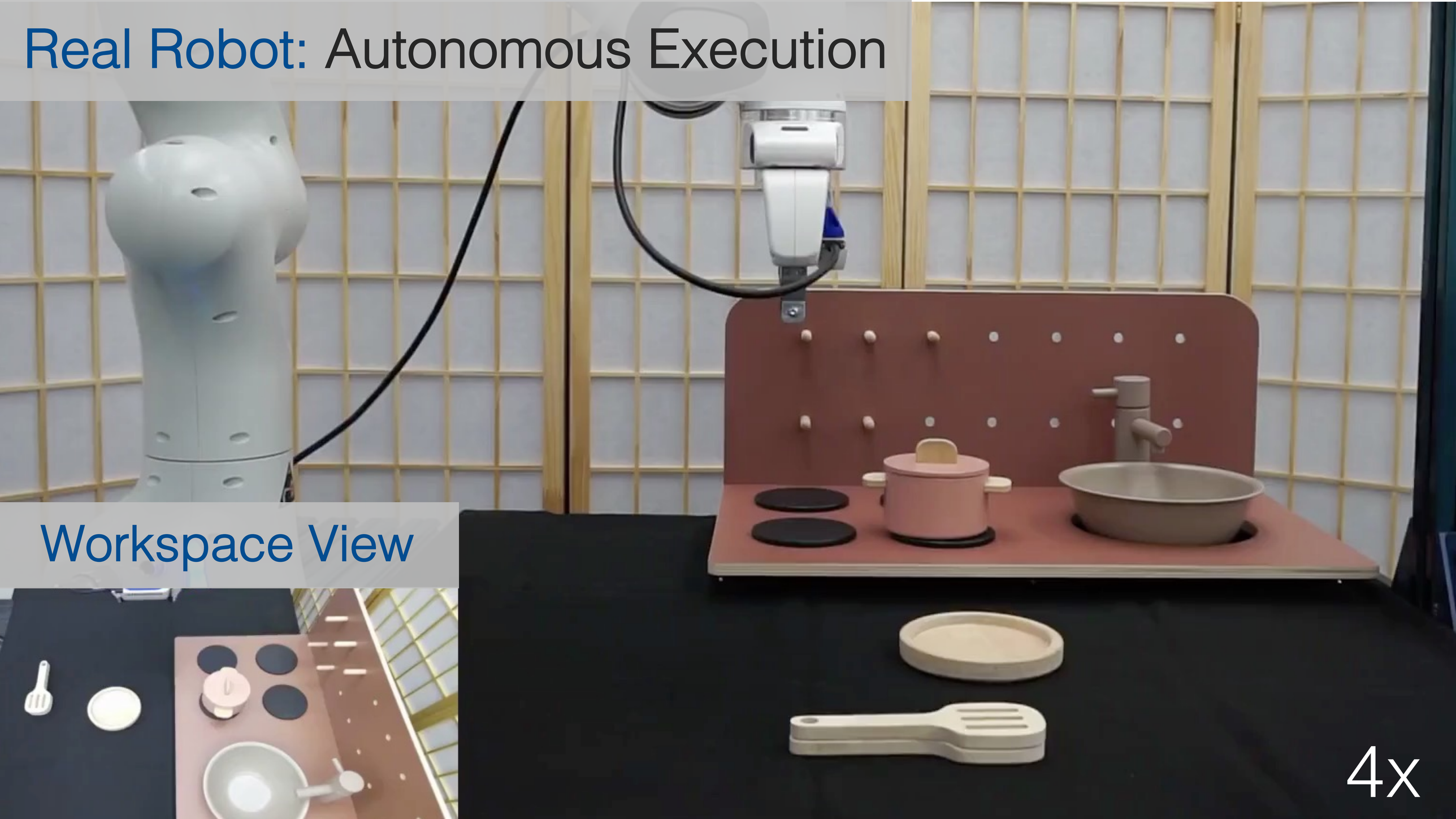
Multi-task learning in Kitchen leads to an 8% increase in success rate.

Discovered skills are reusable in novel tasks.



Through reusing the skills while re-training the meta-controller.

Real Robot: Autonomous Execution

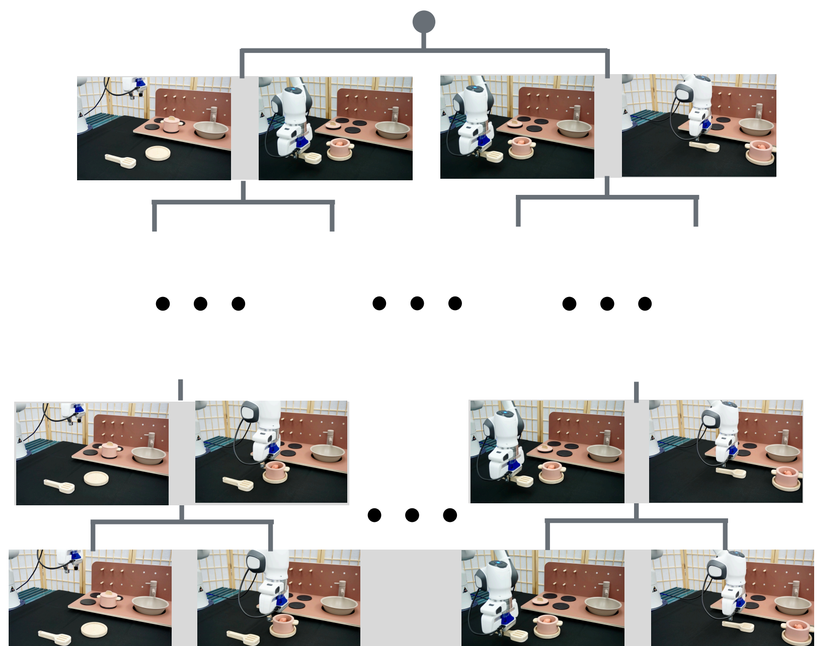


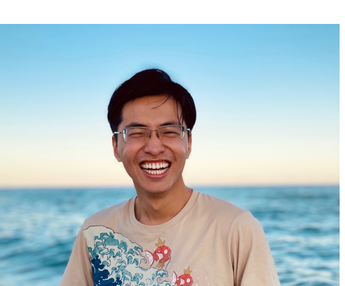
Workspace View

4x

BUDS: Bottom-Up Sensorimotor Skill Discovery

- Discover reusable sensorimotor skills from unstructured task demonstrations
- Multi-sensory, multi-task learning improves skill quality
- Sample-efficient learning of complex behaviors using 30min of demonstrations



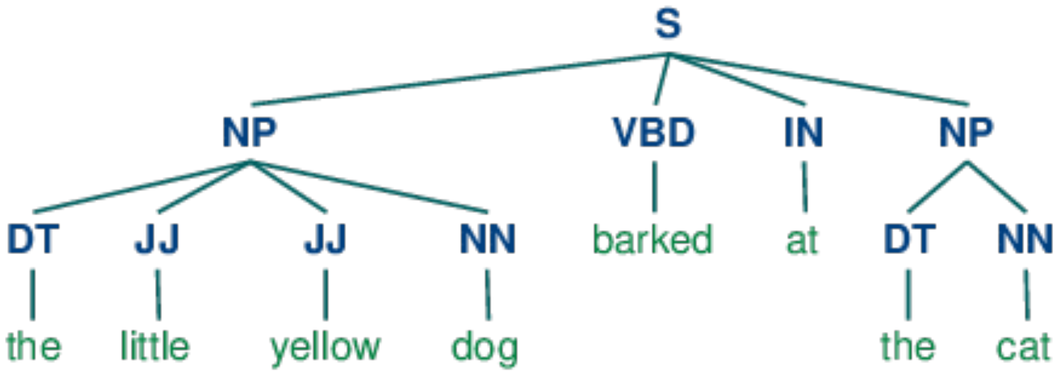


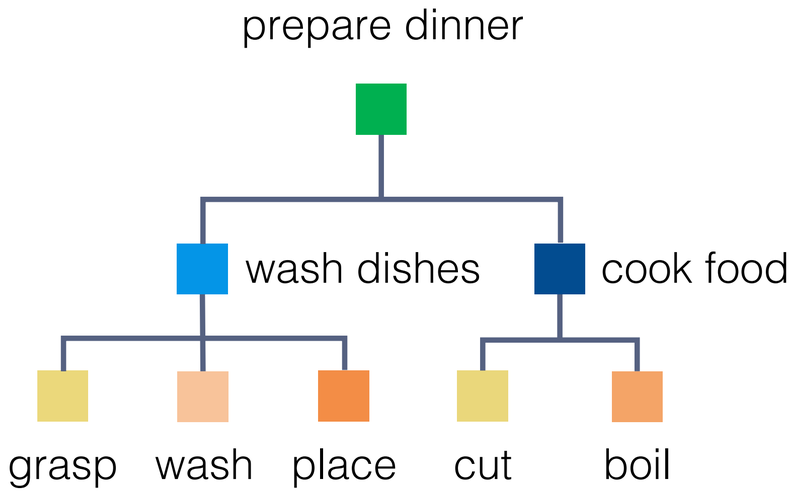
Yifeng Zhu

"Bottom-Up Skill Discovery from Unsegmented Demonstrations for Long-Horizon Robot Manipulation." Zhu, Stone, and Z.

Code and models are available at https://github.com/UT-Austin-RPL/BUDS

Compositionality in Language and Behaviors

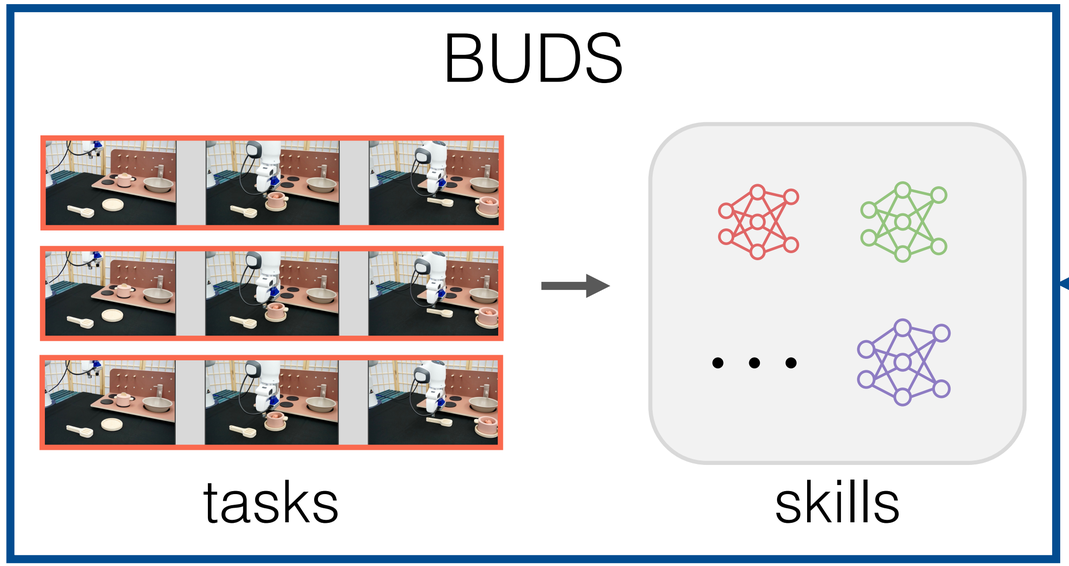




The Context Principle: words have meaning only as constituents of (hence, presumably, only in virtue of their use in) sentences

The Compositionality Principle: the meaning of the whole sentence is a function of the meanings of its parts

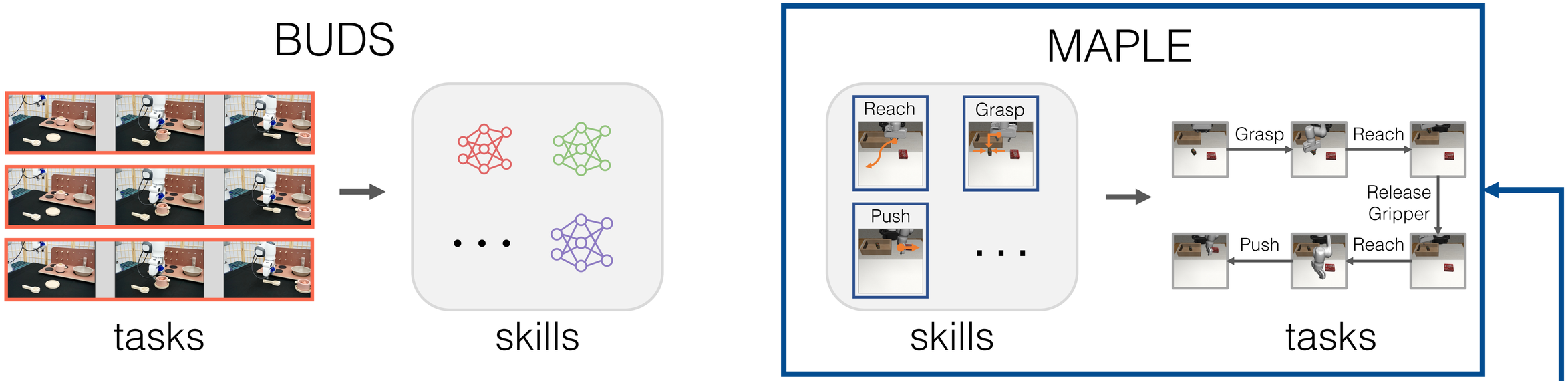
Compositionality in Language and Behaviors



The Context Principle: words have meaning only as constituents of (hence, presumably, only in virtue of their use in) sentences

The Compositionality Principle: the meaning of the whole sentence is a function of the meanings of its parts

Compositionality in Language and Behaviors

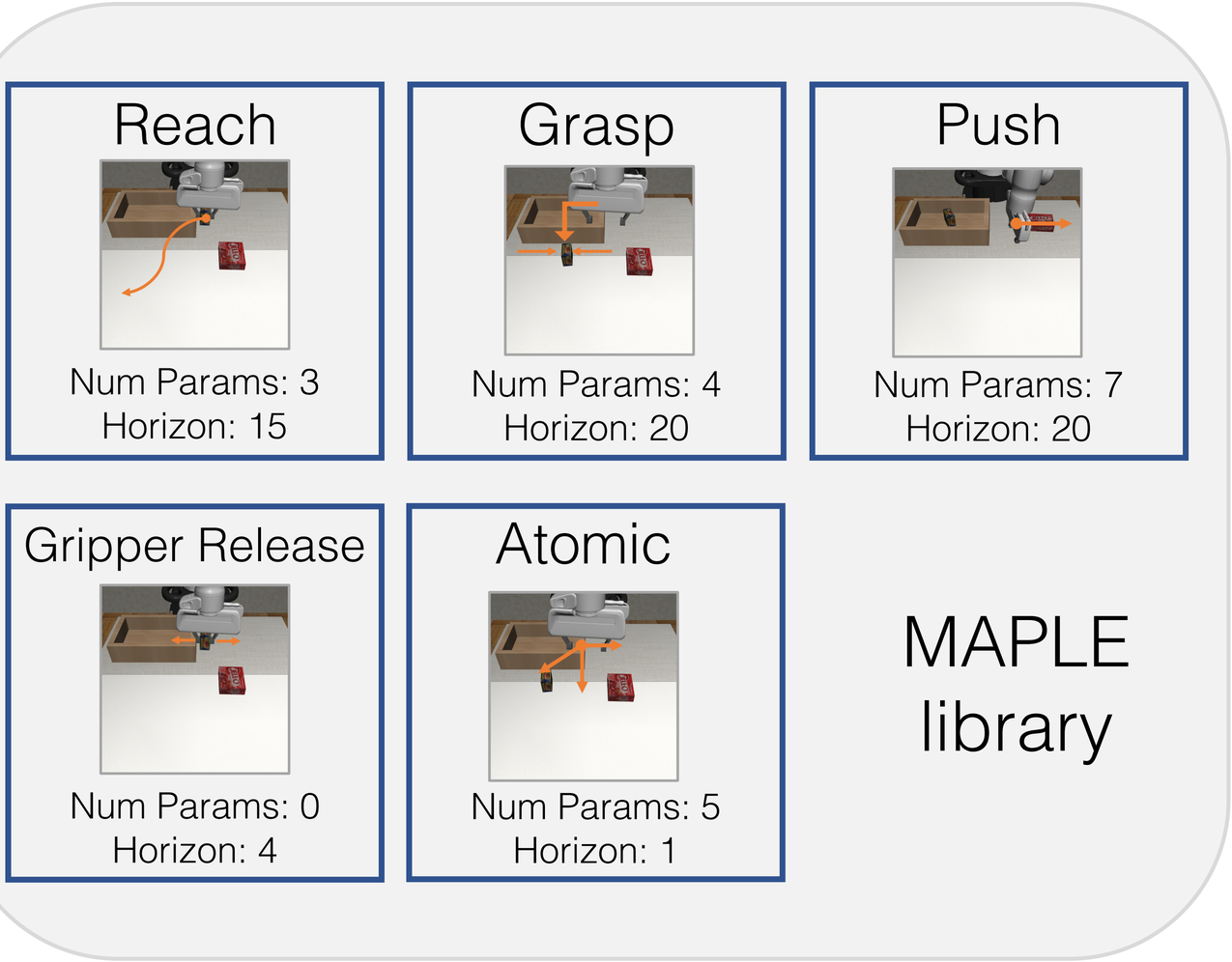


The Context Principle: words have meaning only as constituents of (hence, presumably, only in virtue of their use in) sentences

The Compositionality Principle: the meaning of the whole sentence is a function of the meanings of its parts

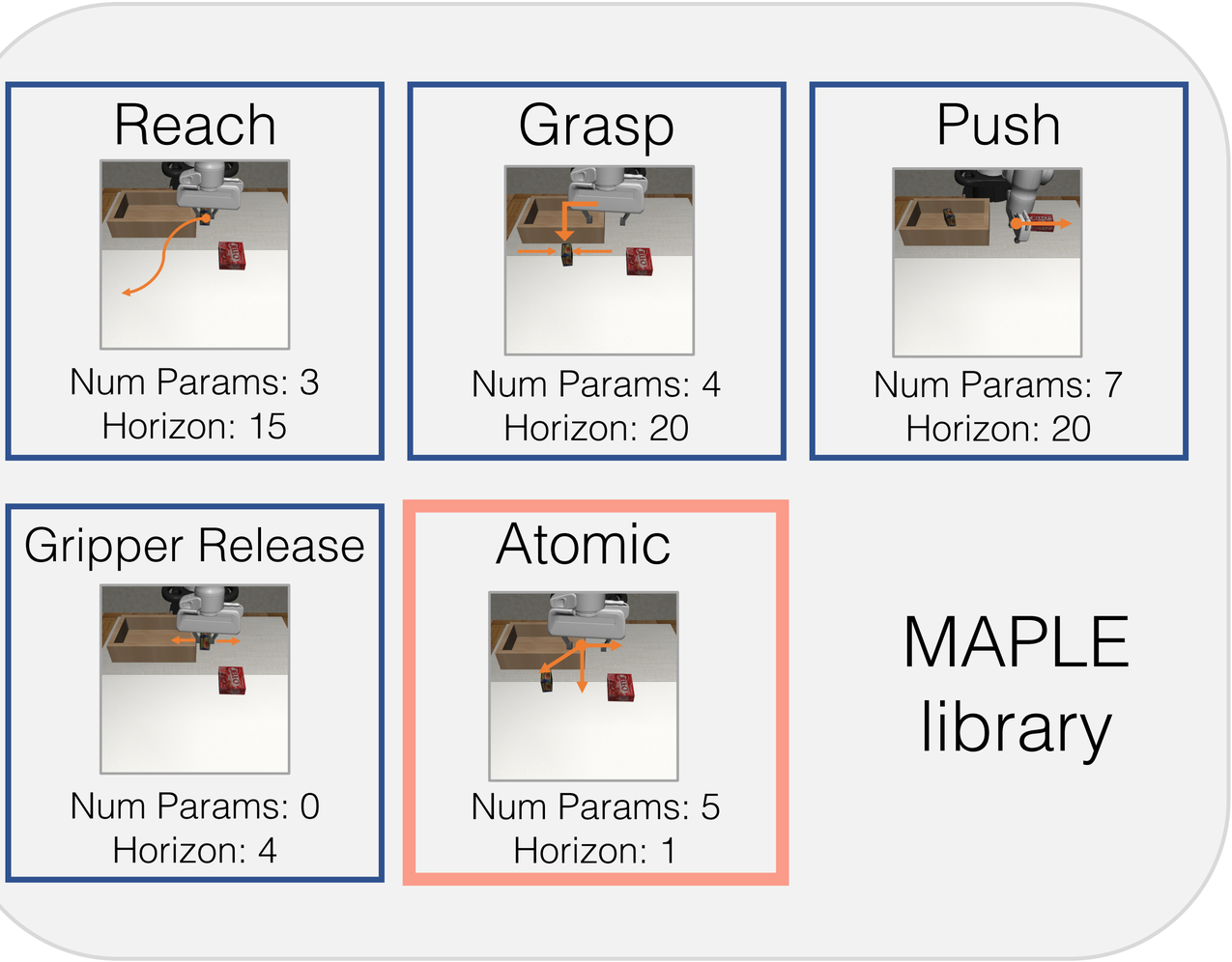
MAPLE: Manipulation Primitive-augmented Reinforcement Learning

- Heterogenous library of primitives
- Primitives are parameterized with inputs. For example, Grasp(x, y, z, ψ): reach location (x, y, z) at angle ψ

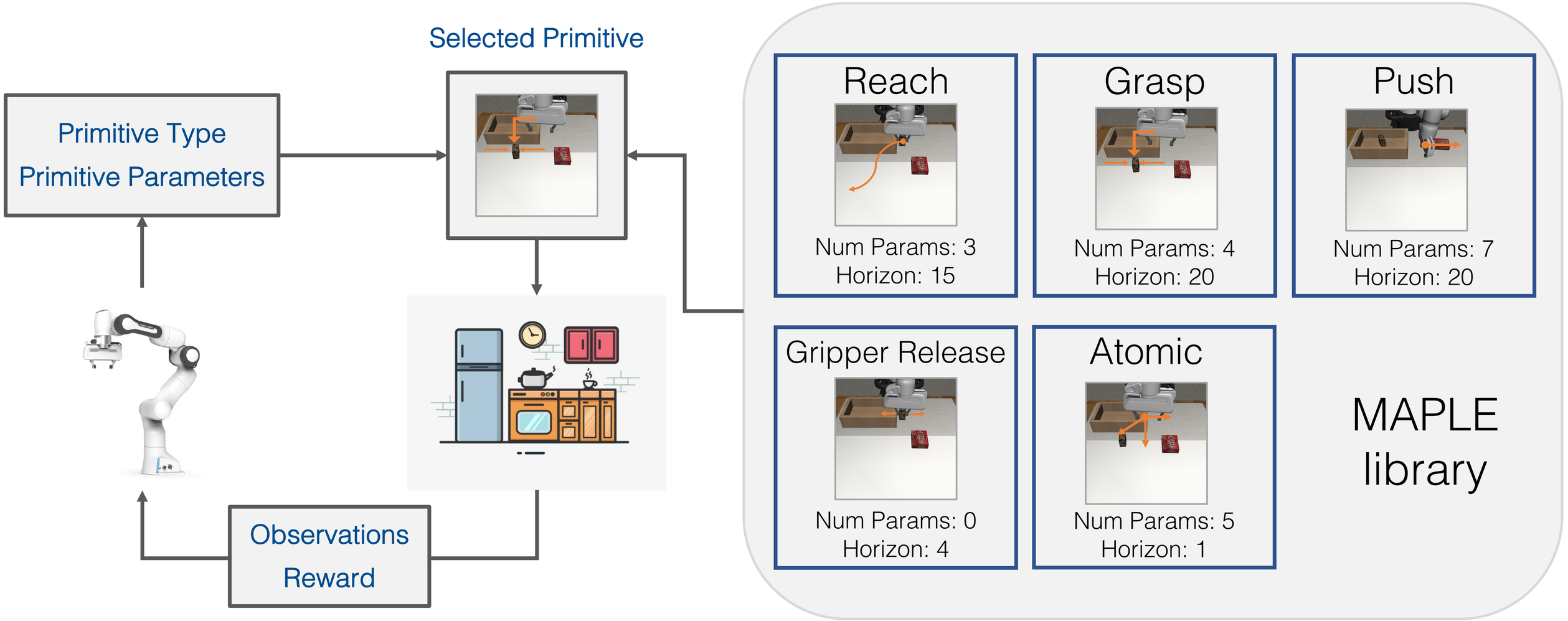


MAPLE: Manipulation Primitive-augmented Reinforcement Learning

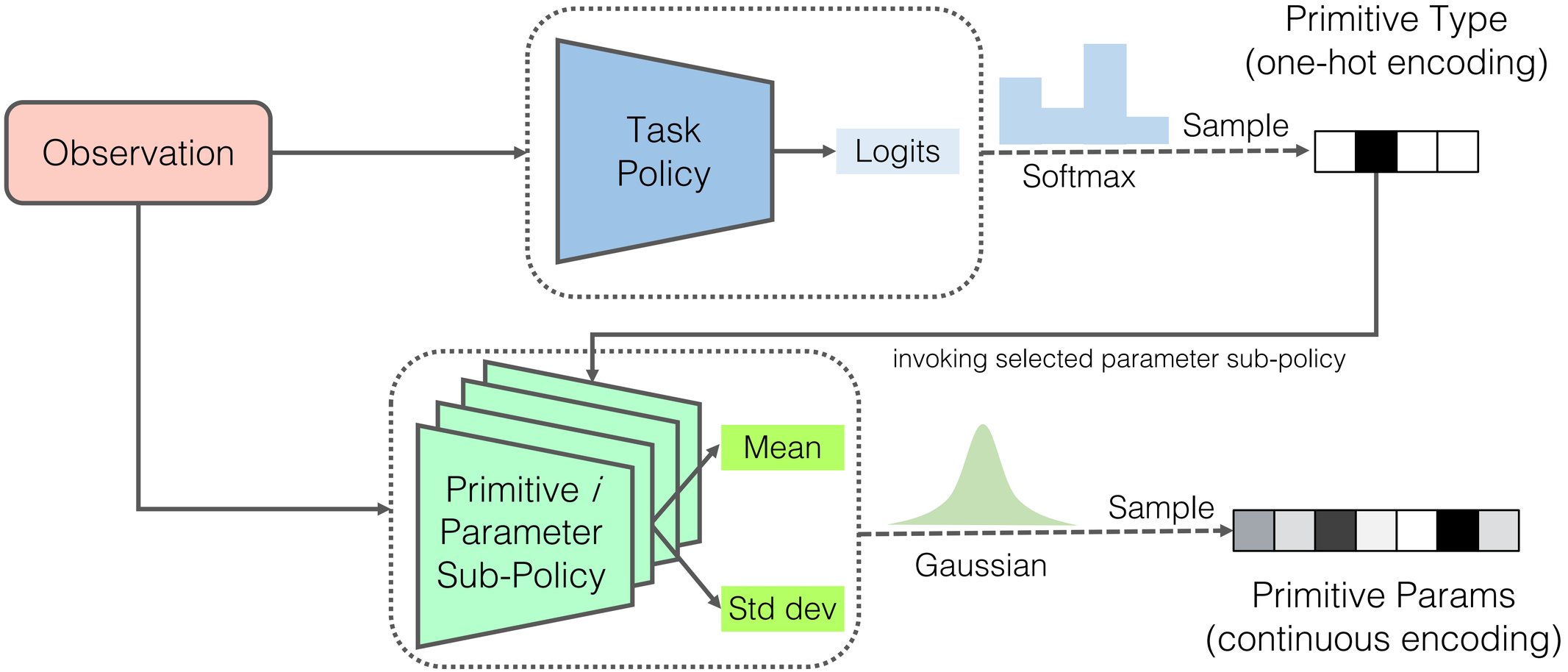
- Heterogenous library of primitives
- Primitives are parameterized with inputs. For example, Grasp(x, y, z, ψ): reach location (x, y, z) at angle ψ
- Atomic primitive dedicated to low-level motor actions



MAPLE: Manipulation Primitive-augmented Reinforcement Learning

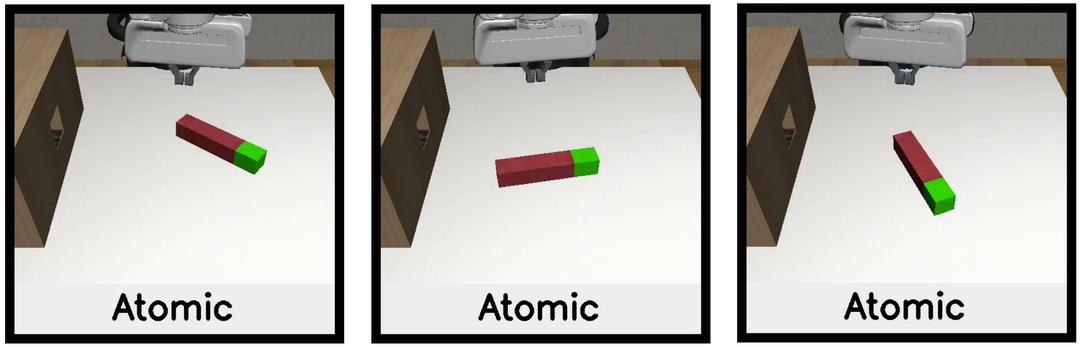


MAPLE: Manipulation Primitive-augmented Reinforcement Learning

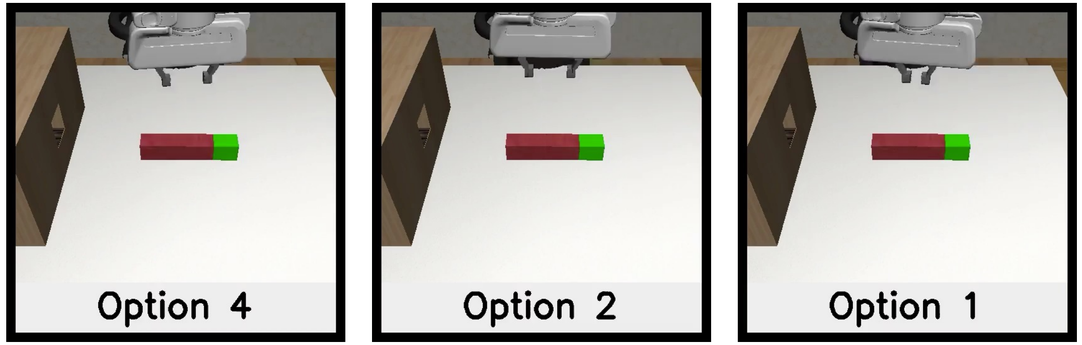


Evaluation: Peg Insertion

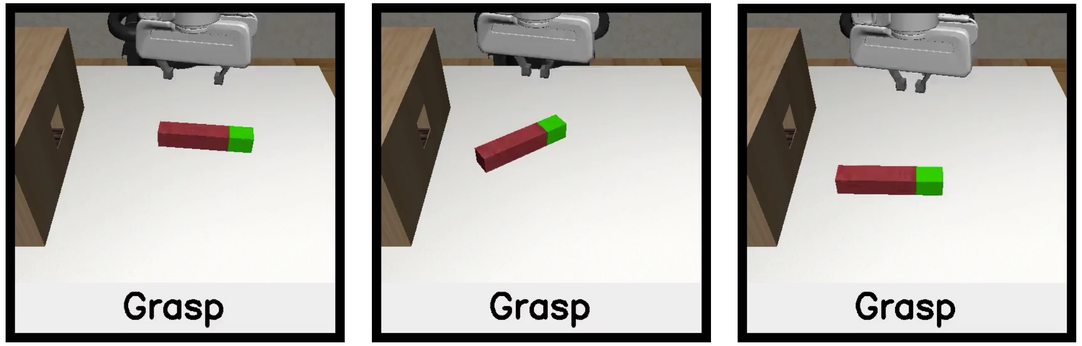
Atomic (0% Success)



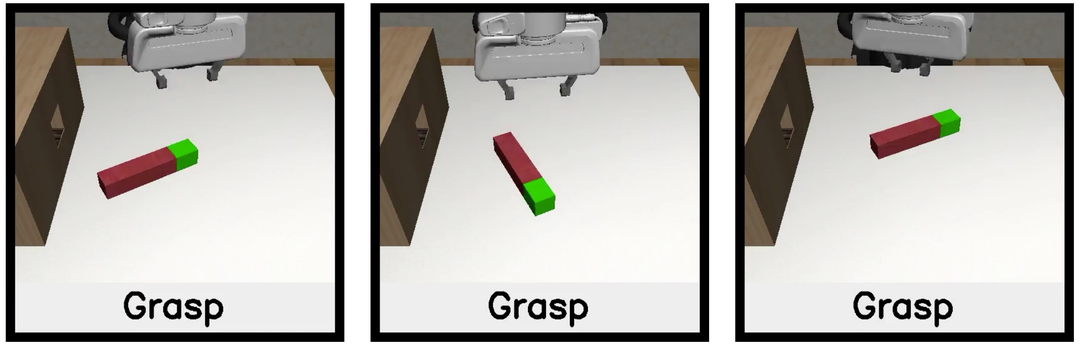
DAC (16% Success)



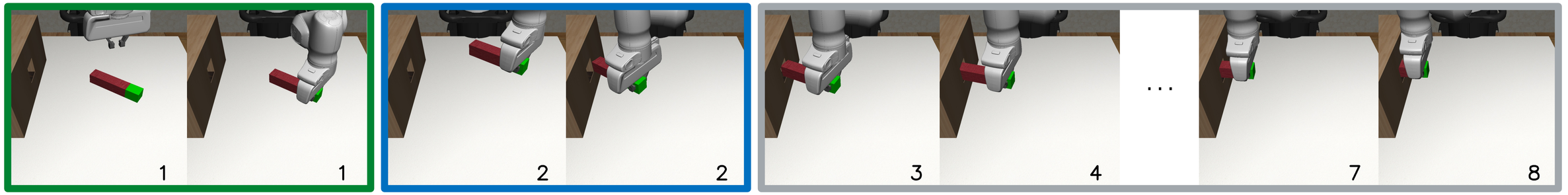
MAPLE (Non-Atomic) (0% Success)



MAPLE (ours) (100% Success)



Evaluation: Compositionality



Primitive Types

Grasp Reach

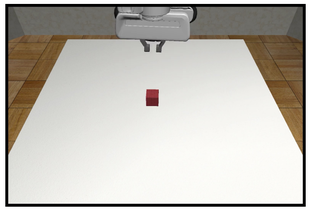
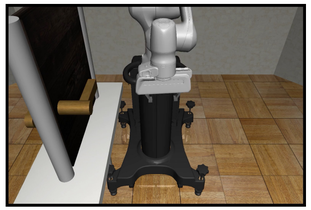
Push Atomic

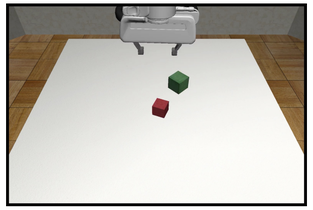
Gripper Release

Task Sketch

scaffold of behavior primitives

Evaluation: Compositionality



Primitive Types

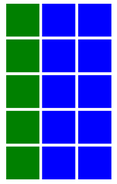
Grasp Reach

Push Atomic

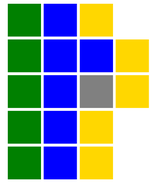
Gripper Release

Lift
$(score = 1.0)$

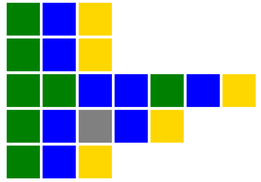
Door
$(score = 1.0)$



Pick and Place
$(score = 0.87)$



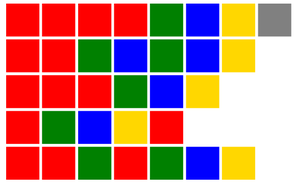
Stack
$(score = 0.96)$



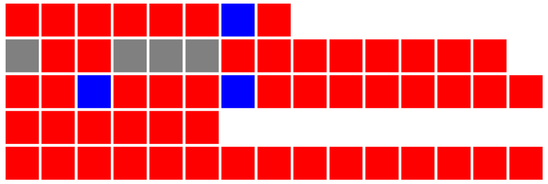
more compositional

Nut Assembly
$(score = 0.87)$

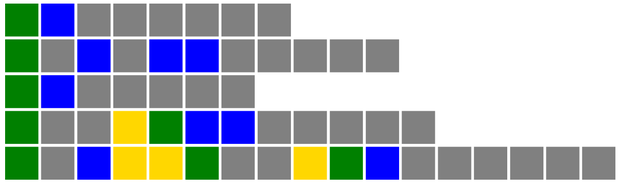
Cleanup
$(score = 0.72)$



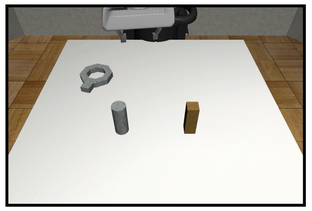
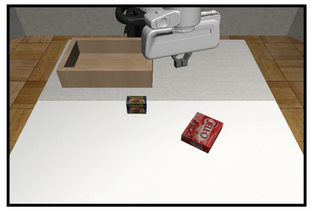
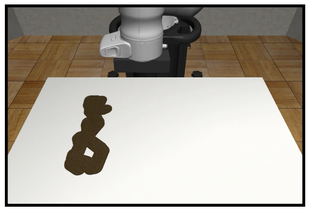
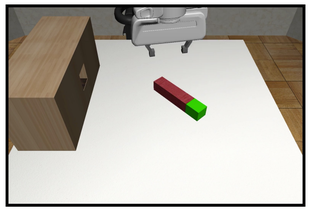
Wipe
$(score = 0.67)$



Peg Insertion
$(score = 0.24)$



less compositional

From Digital Twin Training to Real Robot Deployment

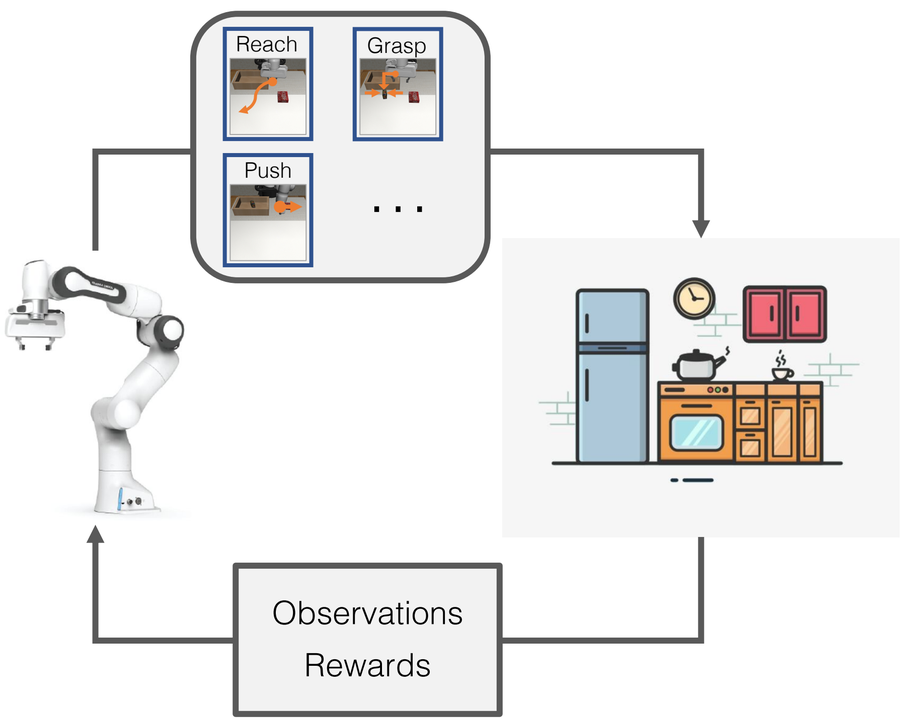
MAPLE Policy Trained in Digital Twin

Zero-Shot Policy Transfer to Real Robot

“Fast Uncertainty Quantification for Deep Object Pose Estimation.” Shi, Zhu, Tremblay, Birchfield, Ramos, Anandkumar, and Z.

MAPLE: Primitive-Augmented Reinforcement Learning

- Heterogeneous behavior primitives to scaffold long-horizon manipulation tasks
- Hierarchical policy to invoke behavior primitives as modular APIs
- Emergent compositional structures from primitive composition



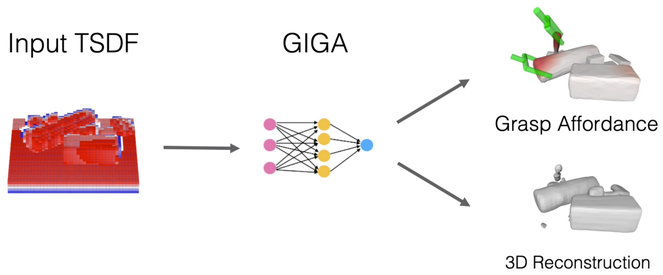
Soroush Nasiriany

"Augmenting Reinforcement Learning with Behavior Primitives for Diverse Manipulation Tasks." Nasiriany, Liu, and Z.

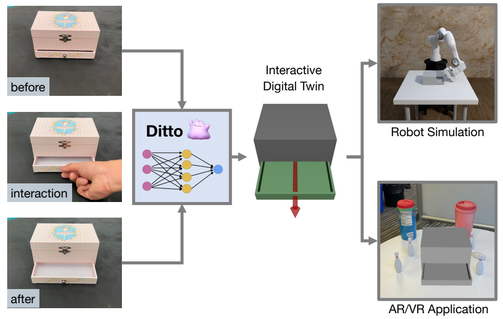
Code and models are available at https://github.com/UT-Austin-RPL/maple

Compositional Robot Autonomy with Objects and Skills

Learning object abstractions from embodied interactions

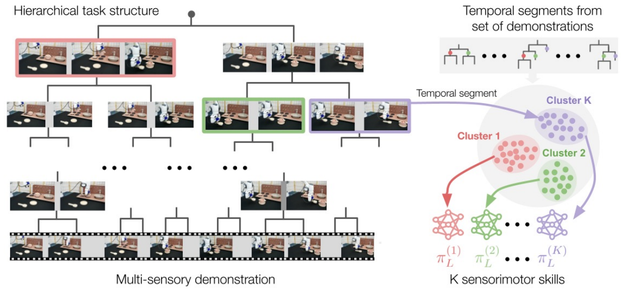


GIGA [Jiang et al., RSS 2021]

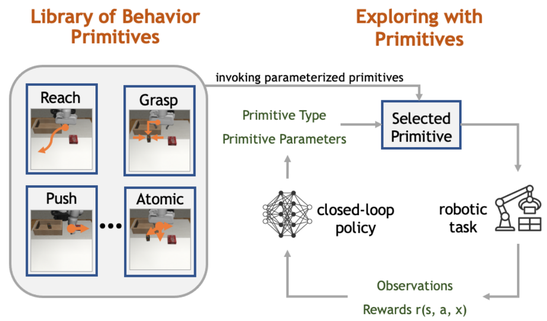


Ditto [Jiang et al., arXiv 2022]

Compositional task modeling with skill abstractions



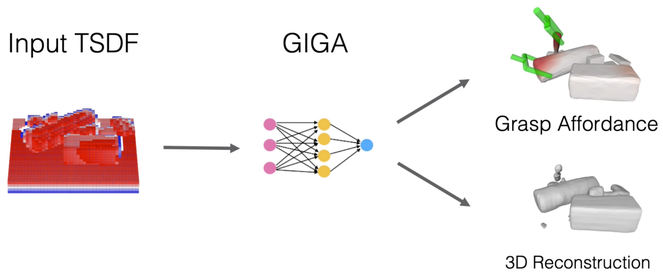
BUDS [Zhu et al., RA-L 2022]



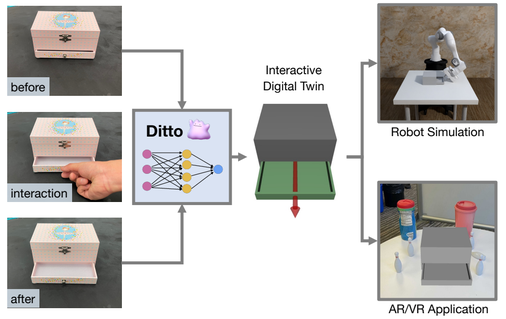
MAPLE [Nasiriany et al., ICRA 2022]

Compositional Robot Autonomy with Objects and Skills

Learning object abstractions from embodied interactions



GIGA [Jiang et al., RSS 2021]



Ditto [Jiang et al., arXiv 2022]

What is an Object?

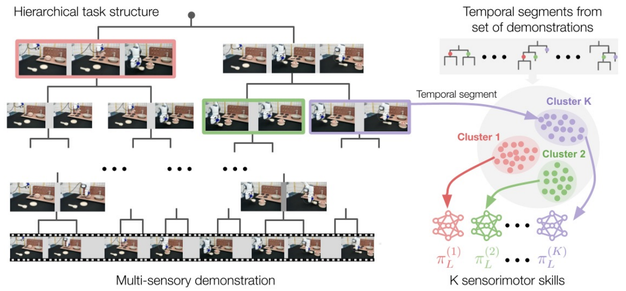
- Context-, task- dependent
- Multi-modal, multi-faceted
- Learned from embodied interaction

Compositional Robot Autonomy with Objects and Skills

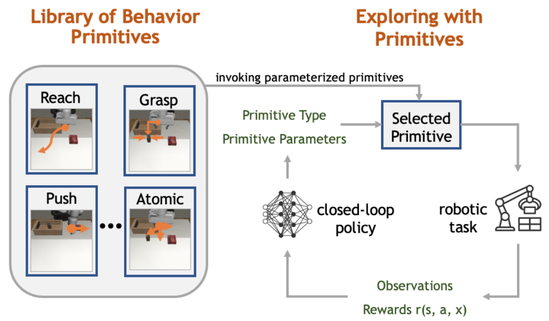
What is a Skill?

- Context vs Compositionality principle
- Versatility and reusability
- Temporal abstraction and perceptual grounding

Compositional task modeling with skill abstractions

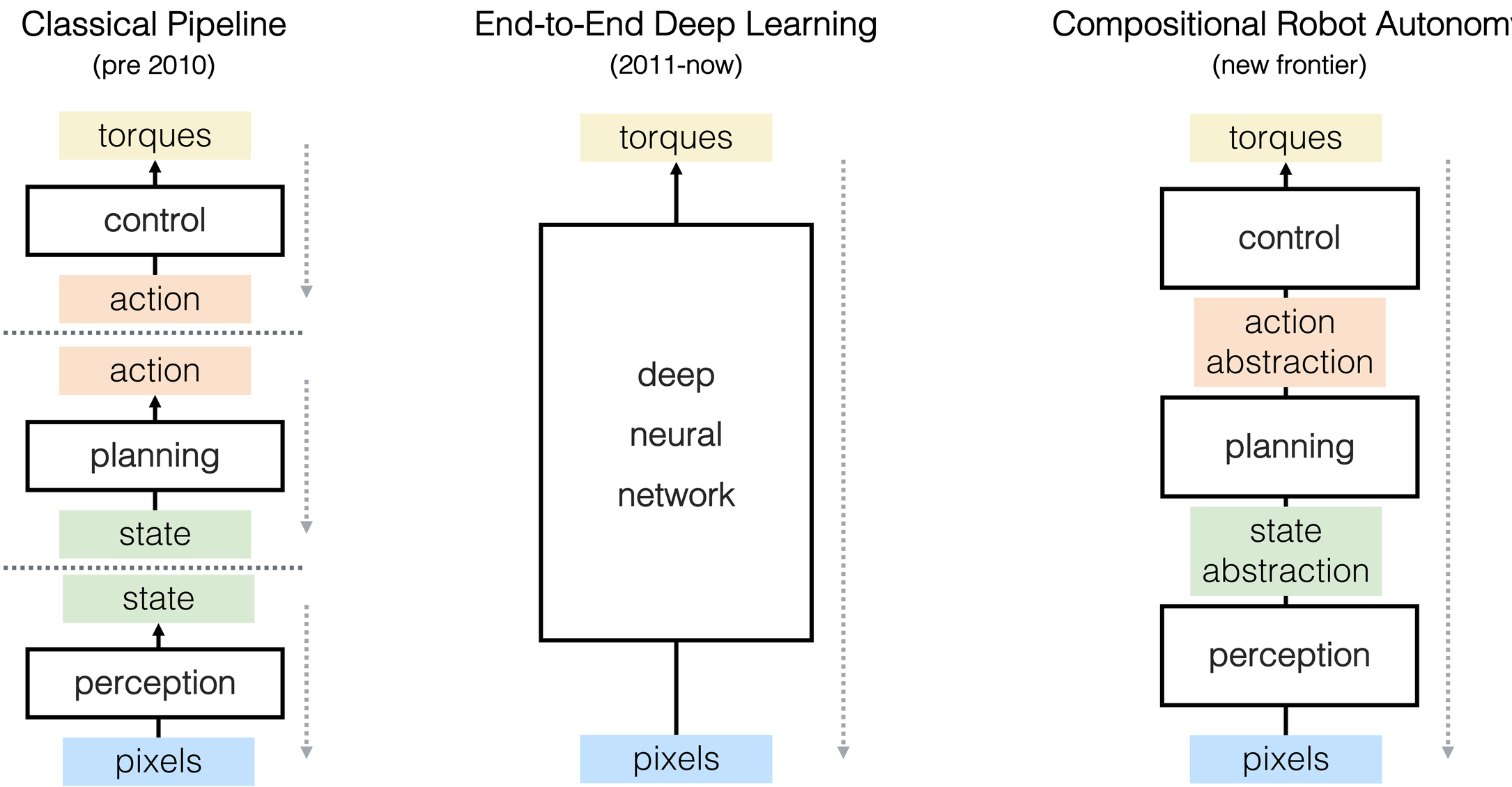


BUDS [Zhu et al., RA-L 2022]



MAPLE [Nasiriany et al., ICRA 2022]

AI Architectures of Robot Autonomy



Acknowledgement

UT Robot Perception & Learning Lab

UT Robot Perception and Learning (RPL) Lab

https://rpl.cs.utexas.edu/