Learning Keypoint Representations for Robot Manipulation

Yuke Zhu

IROS 2019

Stanford University

NVIDIA

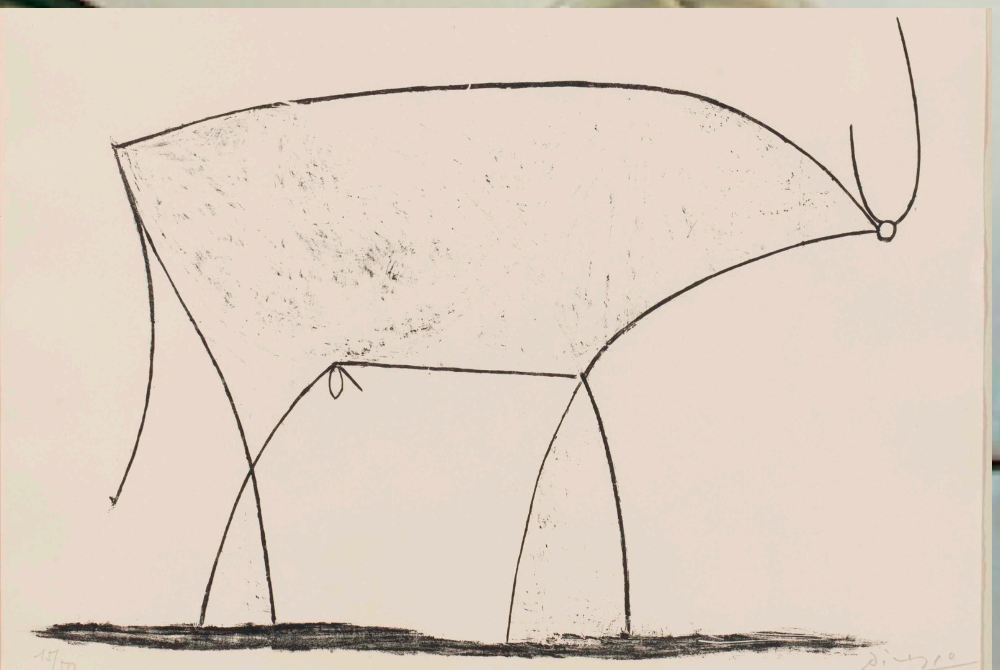
TEXAS
The University of Texas at Austin

"Bull" Pablo Picasso 1945

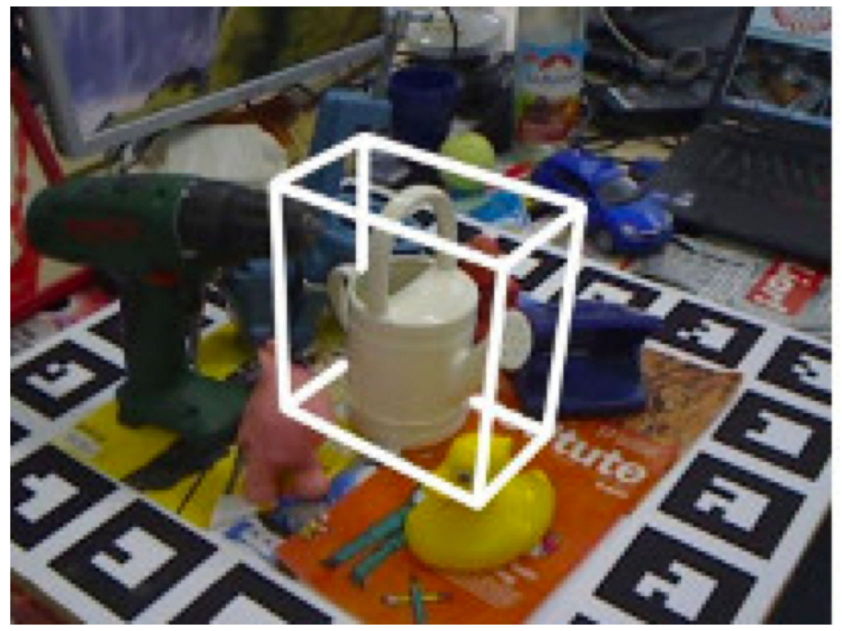

de Paris
Allemagne
Deutschland
Germany
ROUTE
66
KEEP CALM AND CARRY ON SHOPPING
leader of the house
leader of the
Uff da!
CAFE
JAPAN
CALIFORNIA
ZEPHYR

Key to generalization across objects

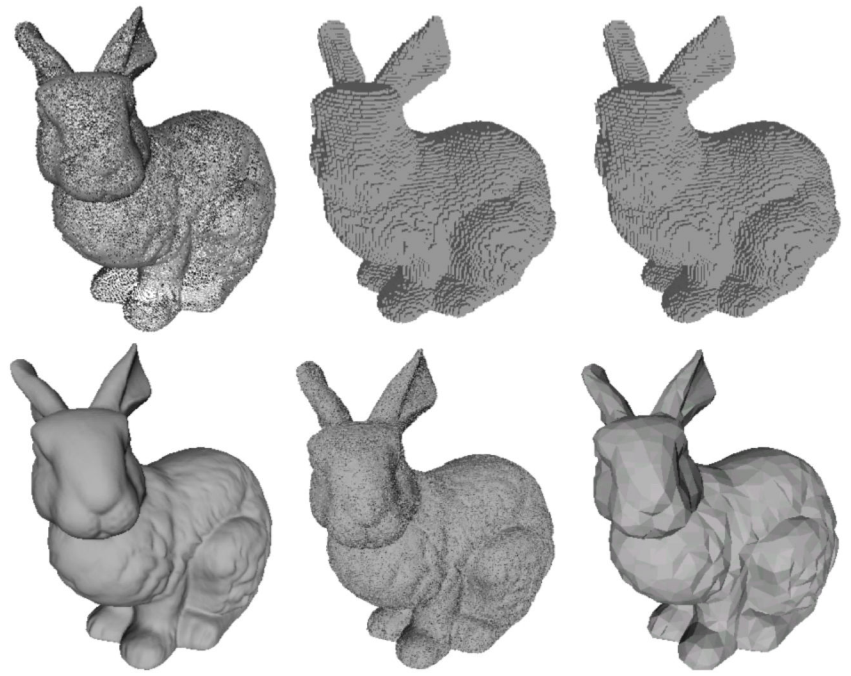
Find an **abstract representation** that can be shared by a family of objects



6D Object Pose



Full 3D Model



A broad range of **object representations**

lack of details

specific to instance

computational cheap

(relatively) easy to estimate

geometric details

generic to object

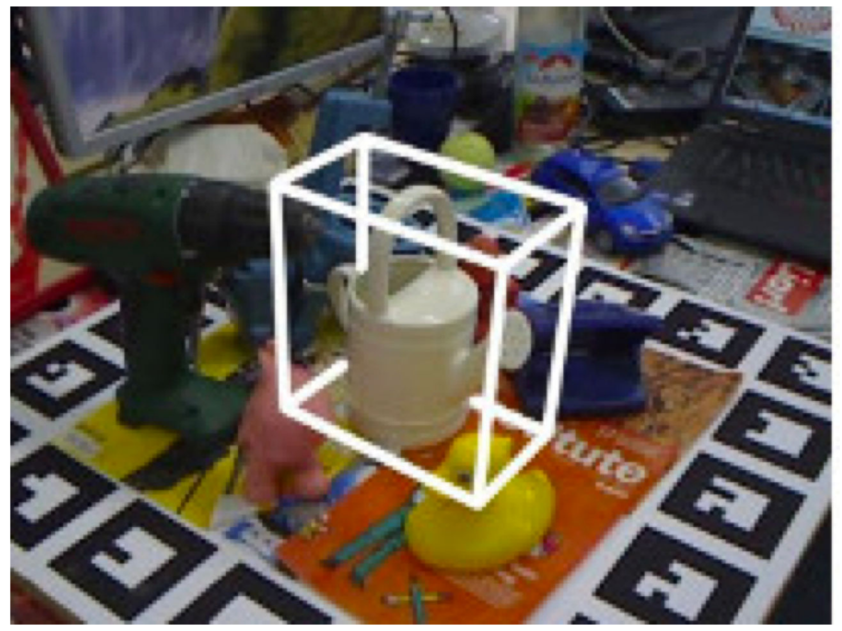
computational expensive

difficult to estimate

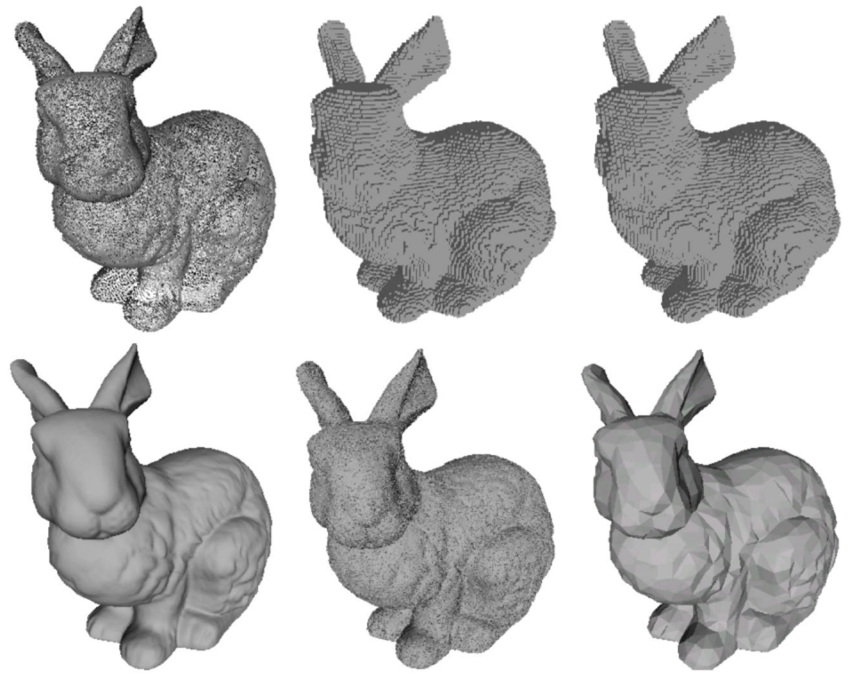
sparse

dense

6D Object Pose



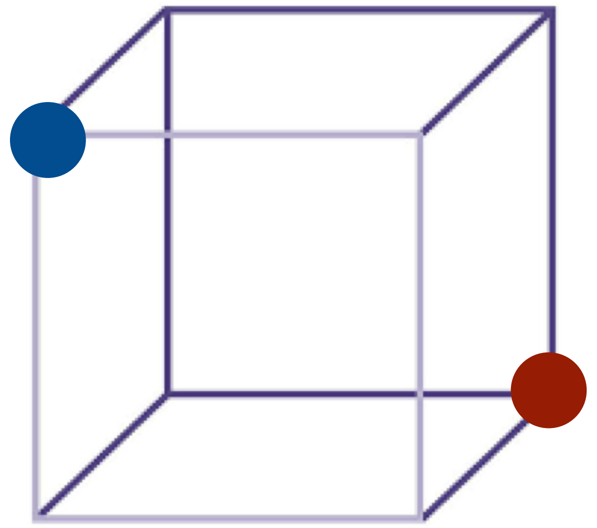
Full 3D Model



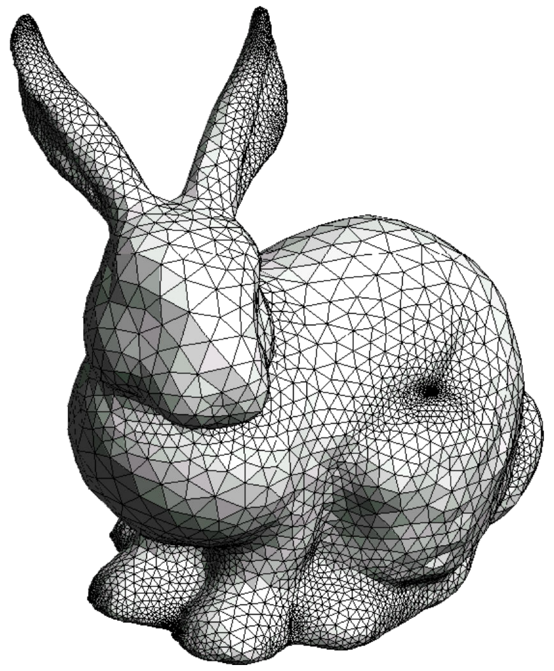
6D pose and **3D model** are

two ends of a spectrum of

point-based representations



2 points

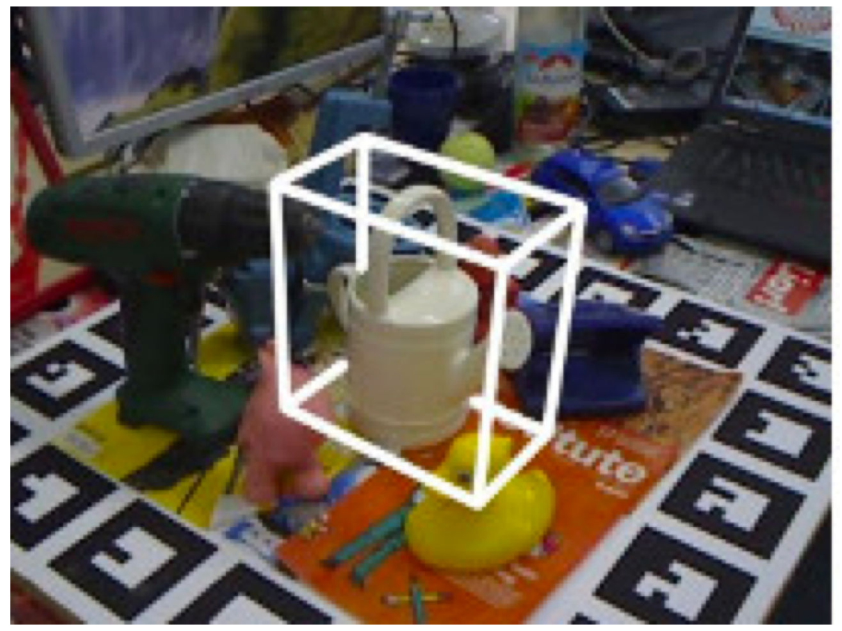


35,947 points (vertices)

sparse

dense

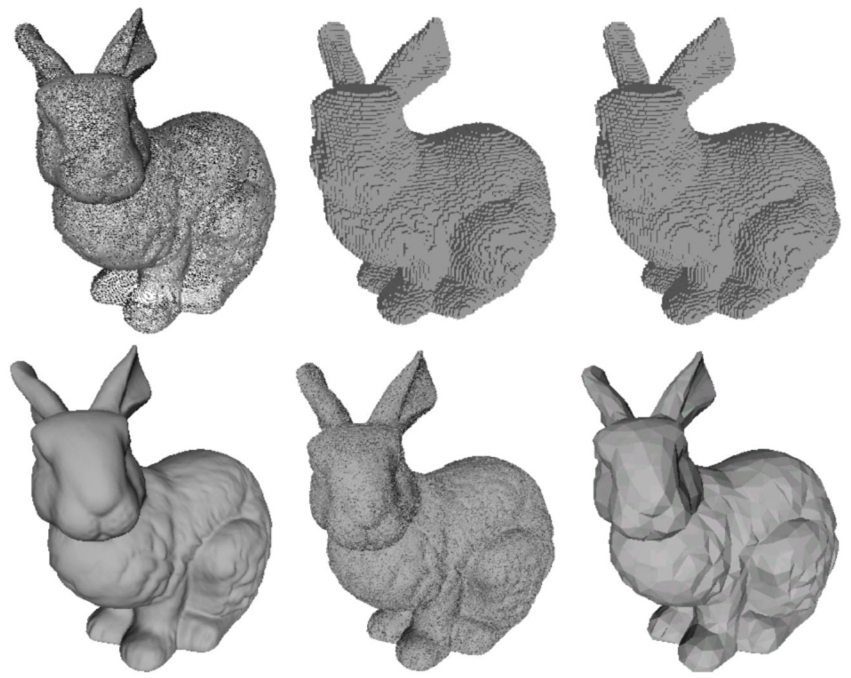
6D Object Pose



Key idea

A handful of discriminative **3D keypoints** as a **compact** and **effective** object representation

Full 3D Shape



lack of details

specific to instance

computational cheap

(relatively) easy to estimate

3D keypoints

geometric details

generic to object

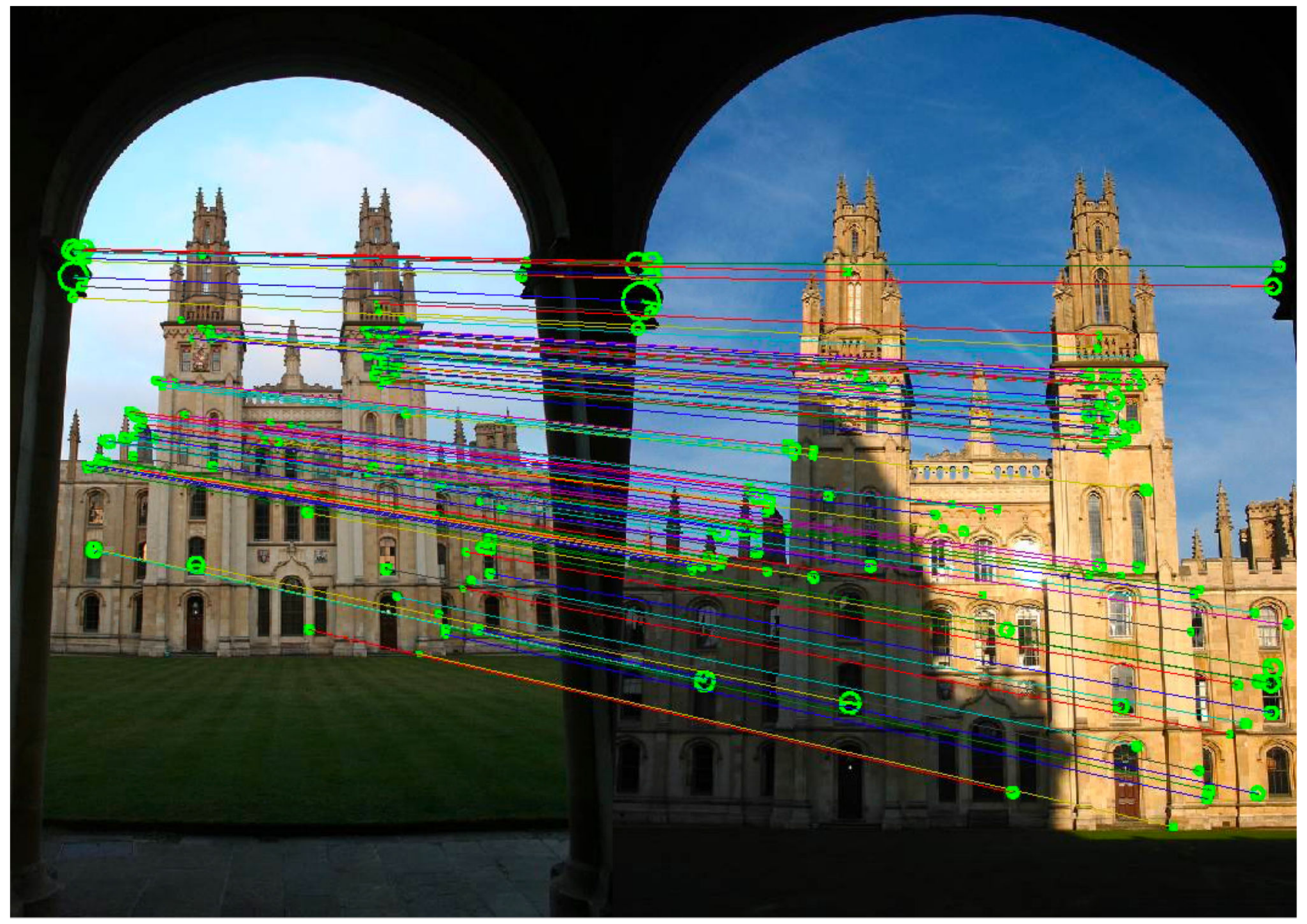
computational expensive

difficult to acquire

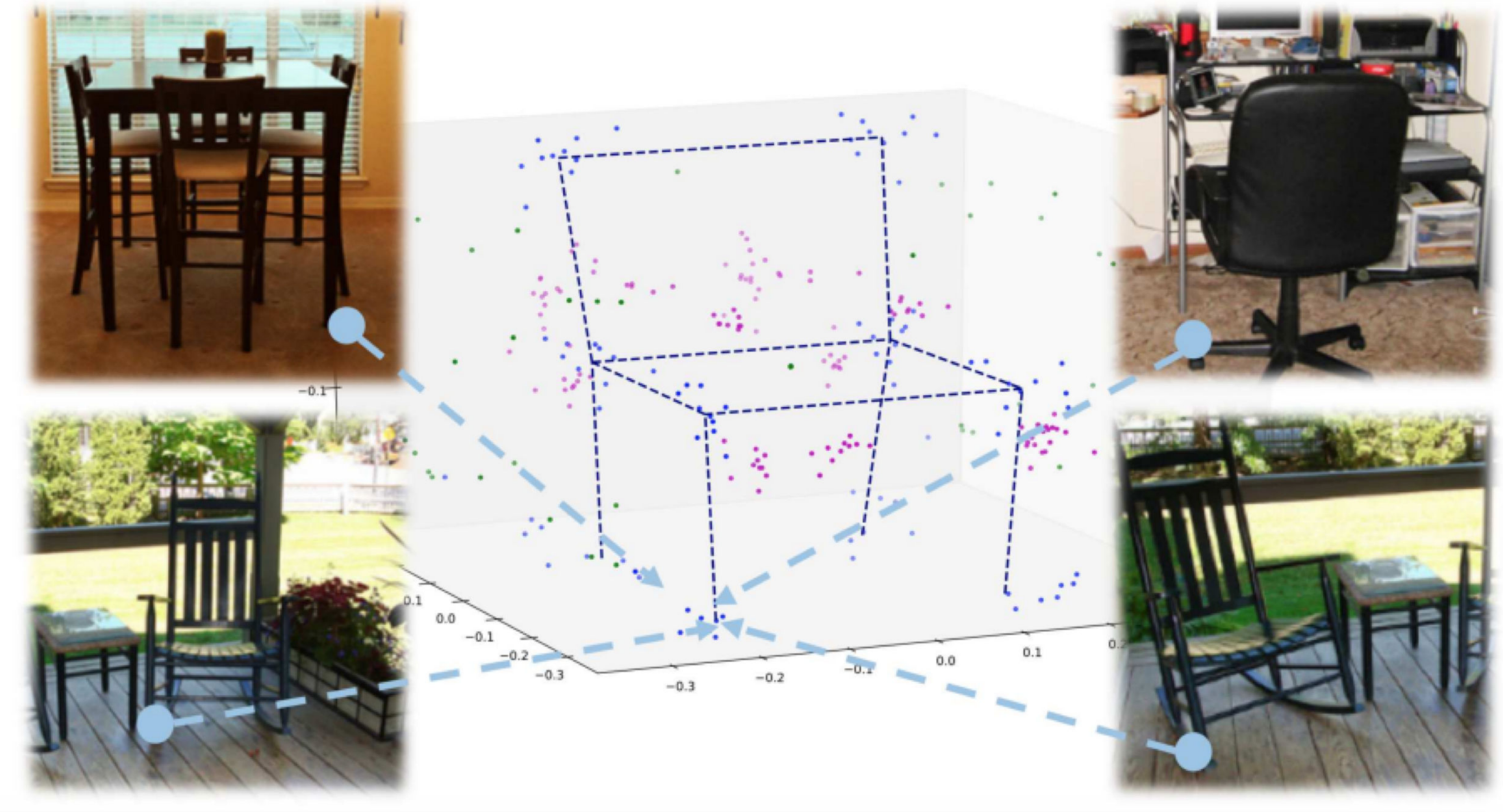
sparse

dense

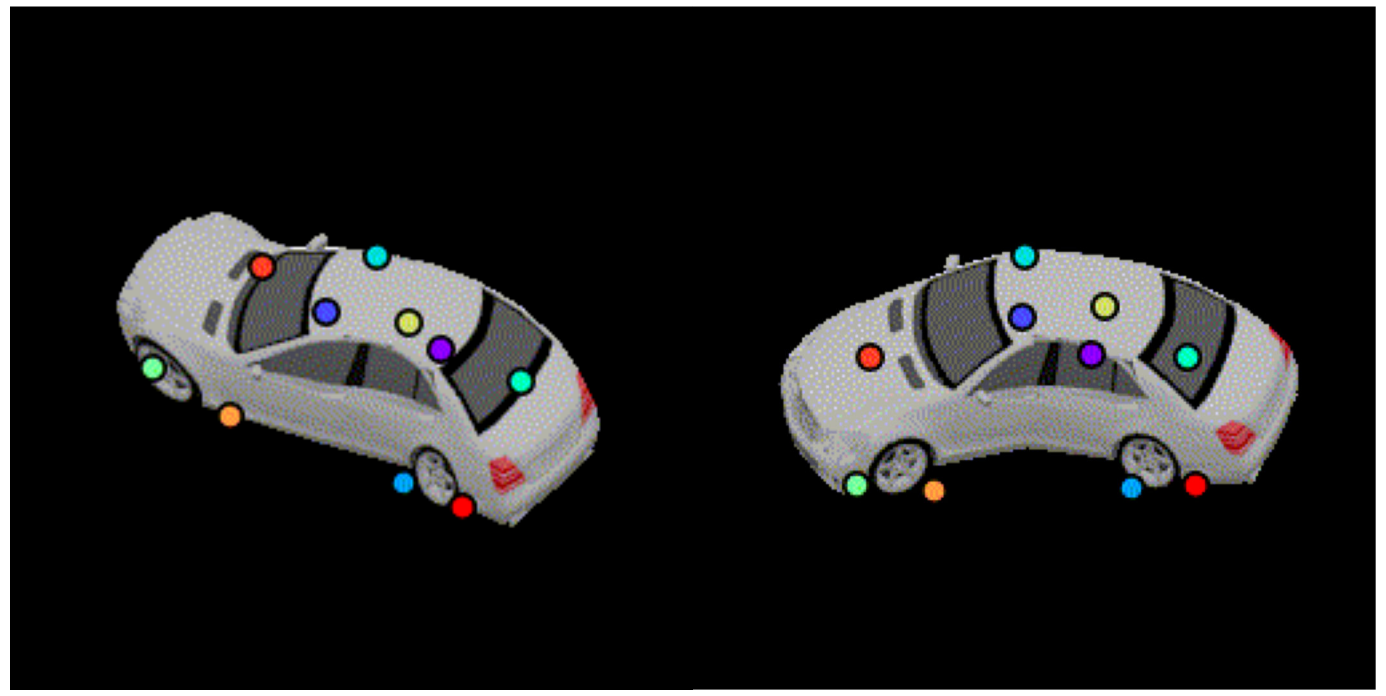
Keypoint Representations



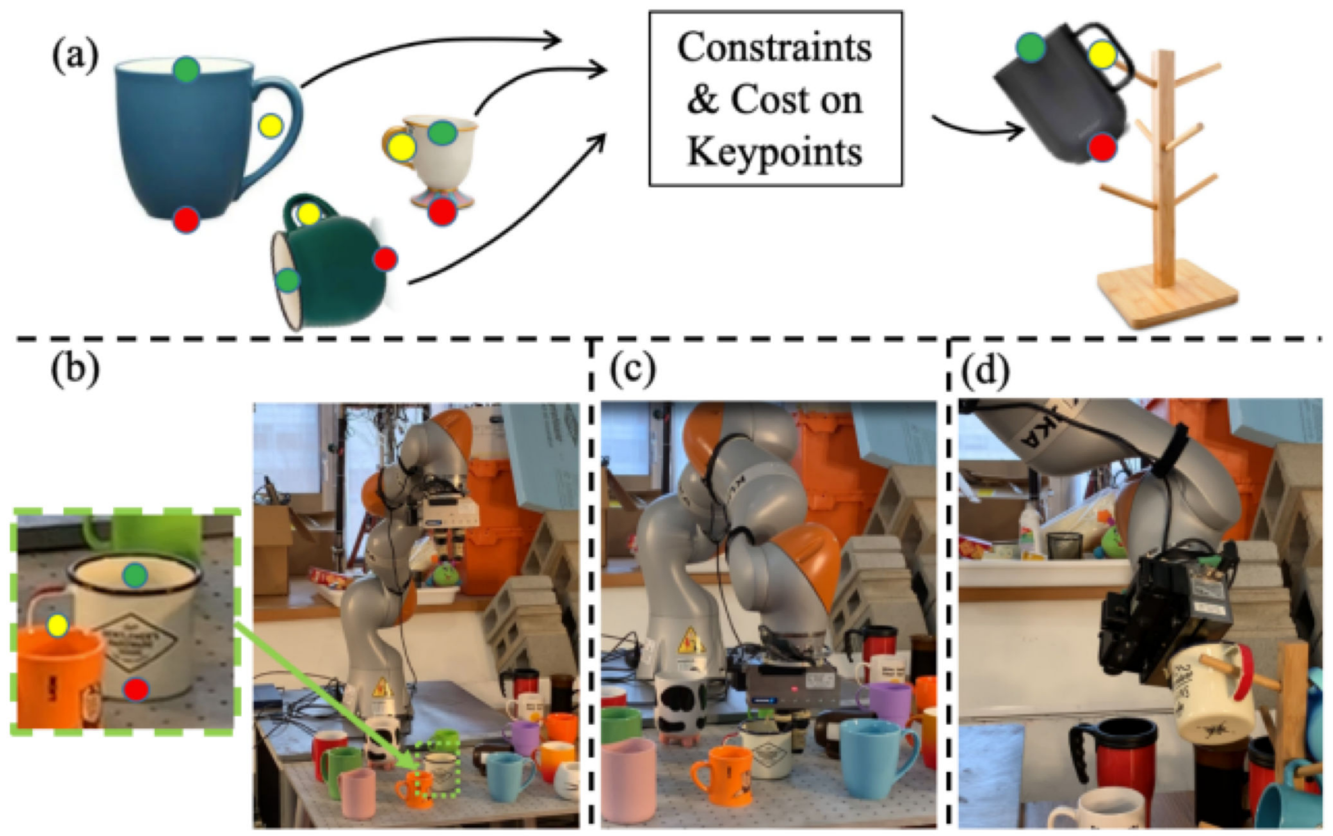
Compact and discriminative
[SIFT, Lowe 2004]



Robust towards object variations
[Zhou et al. 2018]



Handles occlusions and deformations
[KeypointNet, Suwajanakorn et al. 2018]



Informative for robot control
[kPAM, Manuelli et al. 2019]

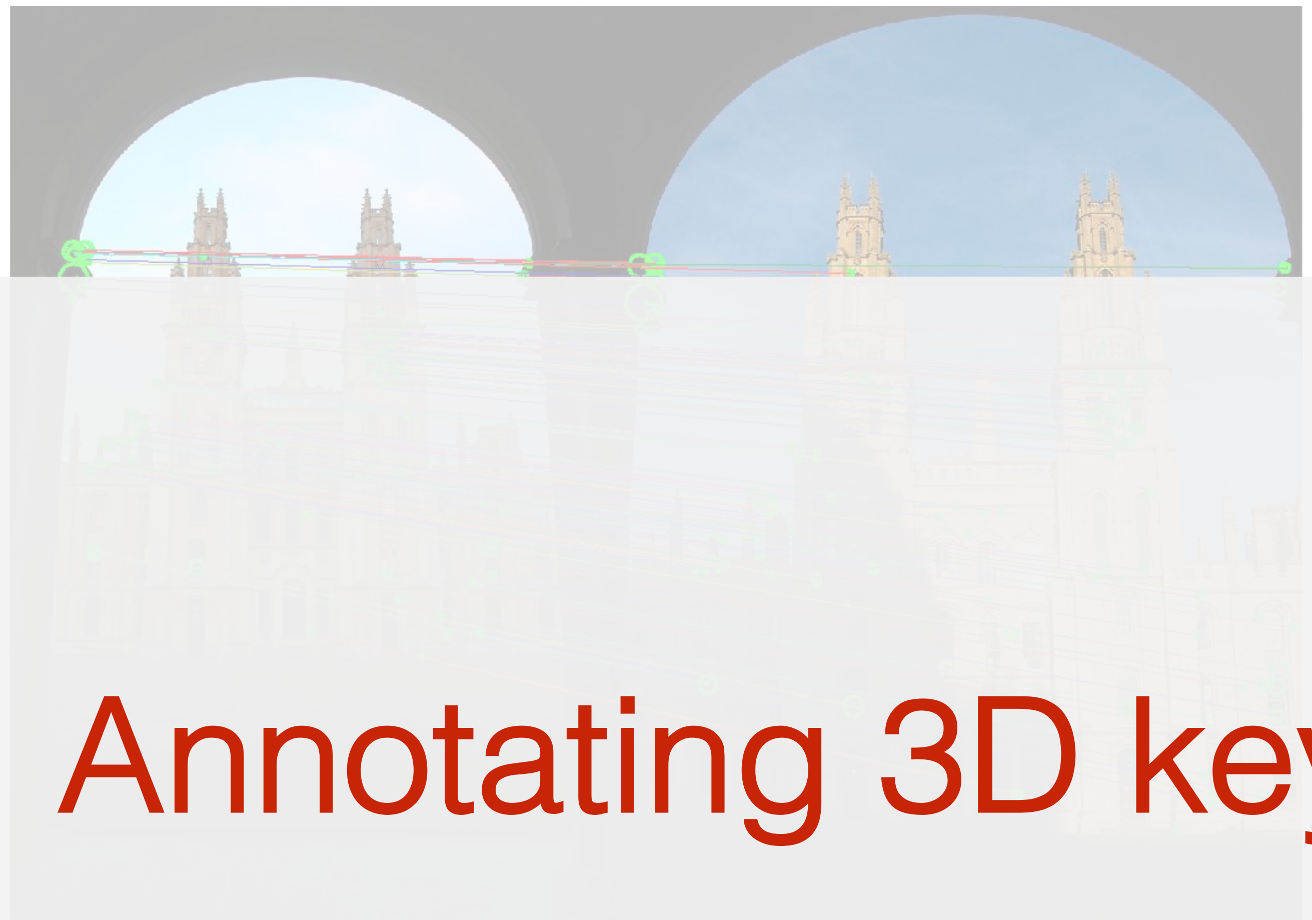
Keypoint Representations

Problem:

Annotating 3D keypoints is tedious and ambiguous.

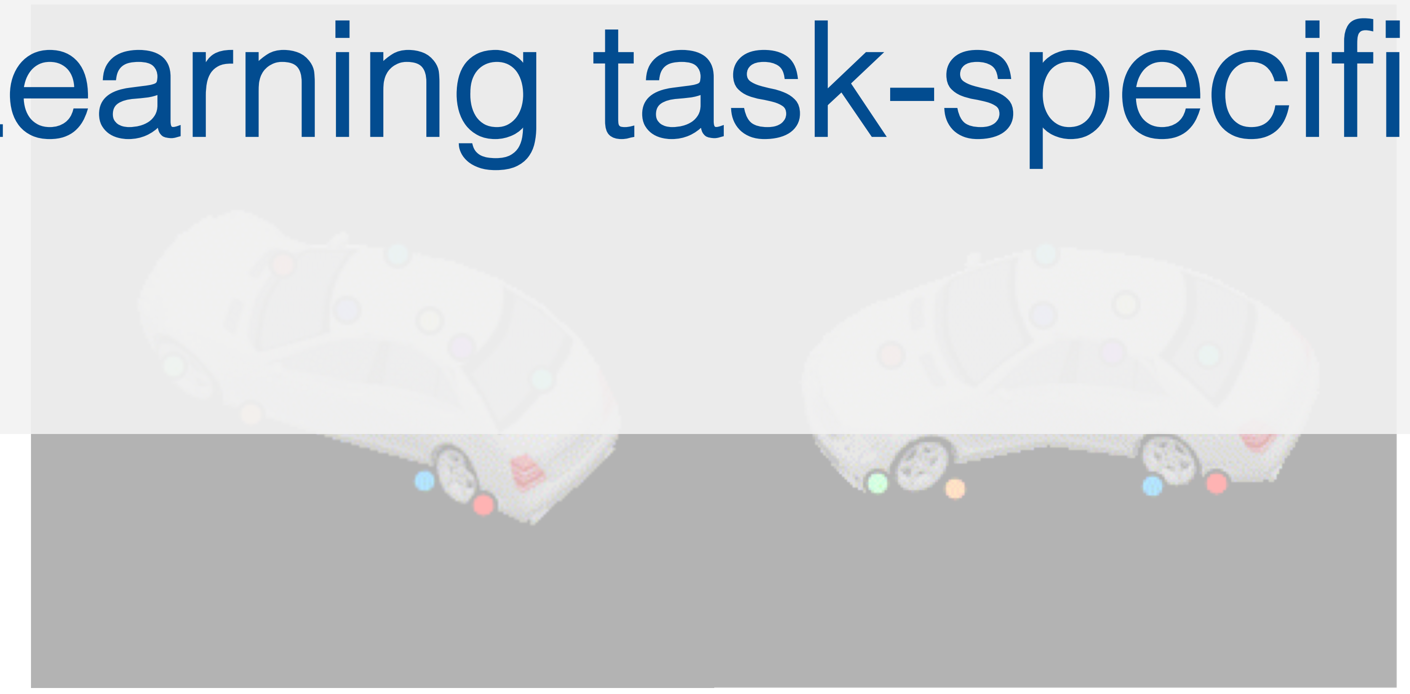
Our Solution:

Learning task-specific keypoints without manual labeling.



Compact and discriminative
[SIFT, Lowe 2004]

Robust towards object variations
[Zhou et al. 2018]

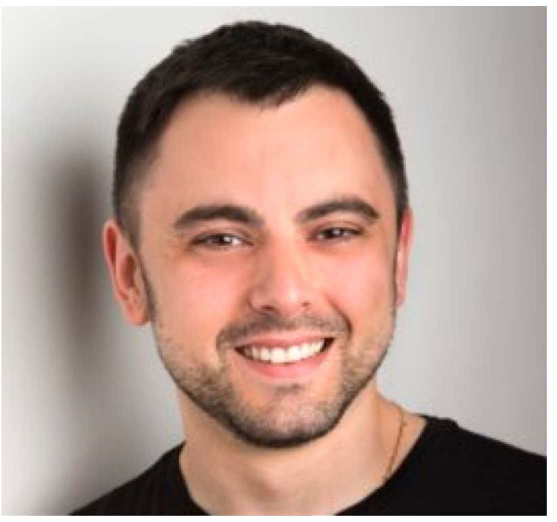
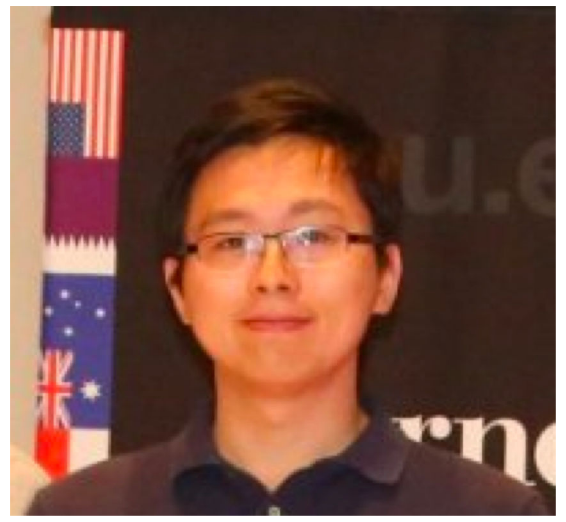


Handles occlusions and deformations
[KeypointNet, Suwajanakorn et al. 2018]

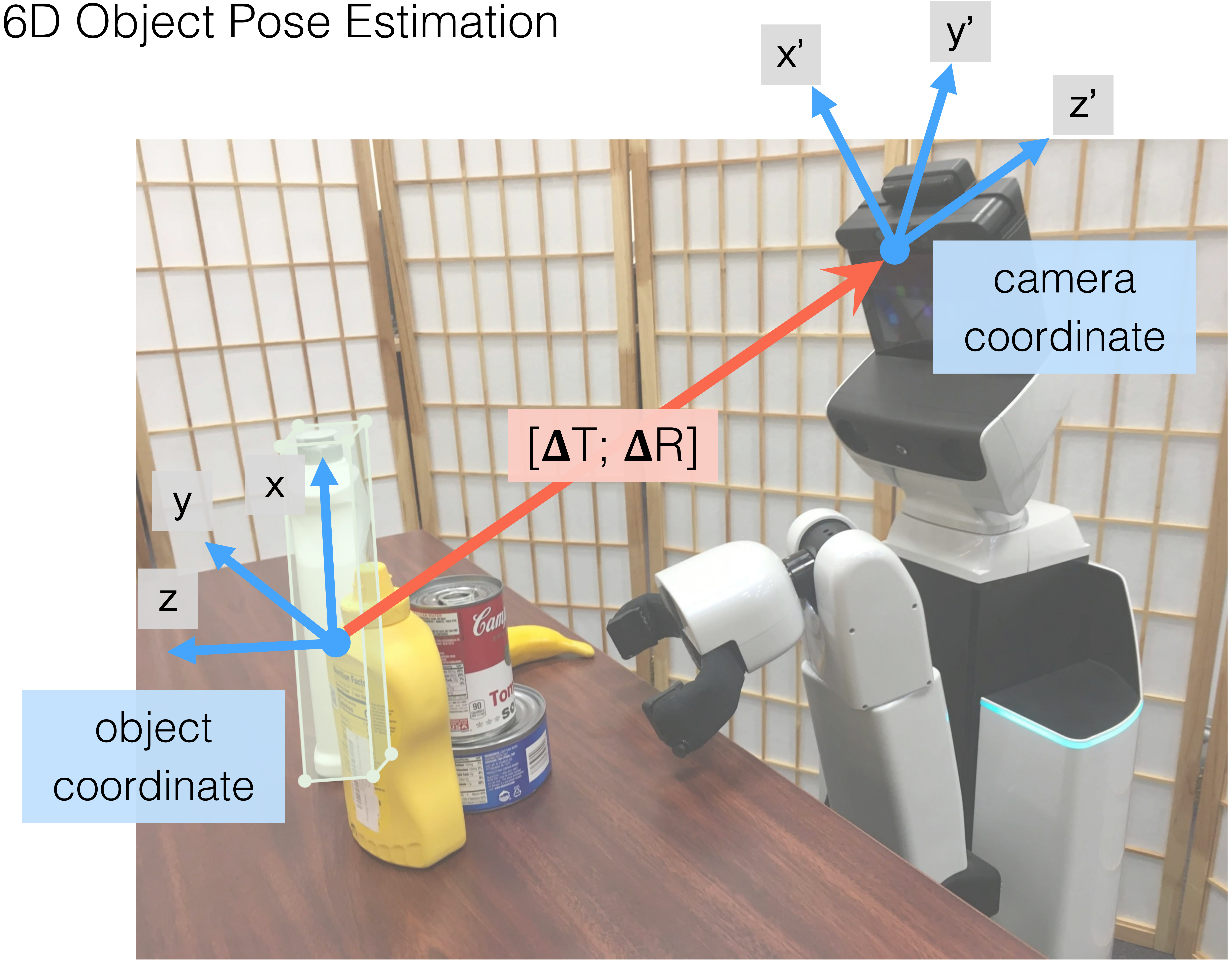
Informative for robot control
[kPAM, Manuelli et al. 2019]

6-Pack: Category-level 6D Pose Tracker with Anchor-Based Keypoints

Chen Wang, Roberto Martín-Martín, Danfei Xu, Jun Lv,
Cewu Lu, Li Fei-Fei, Silvio Savarese, **Yuke Zhu**

6D Object Pose Estimation



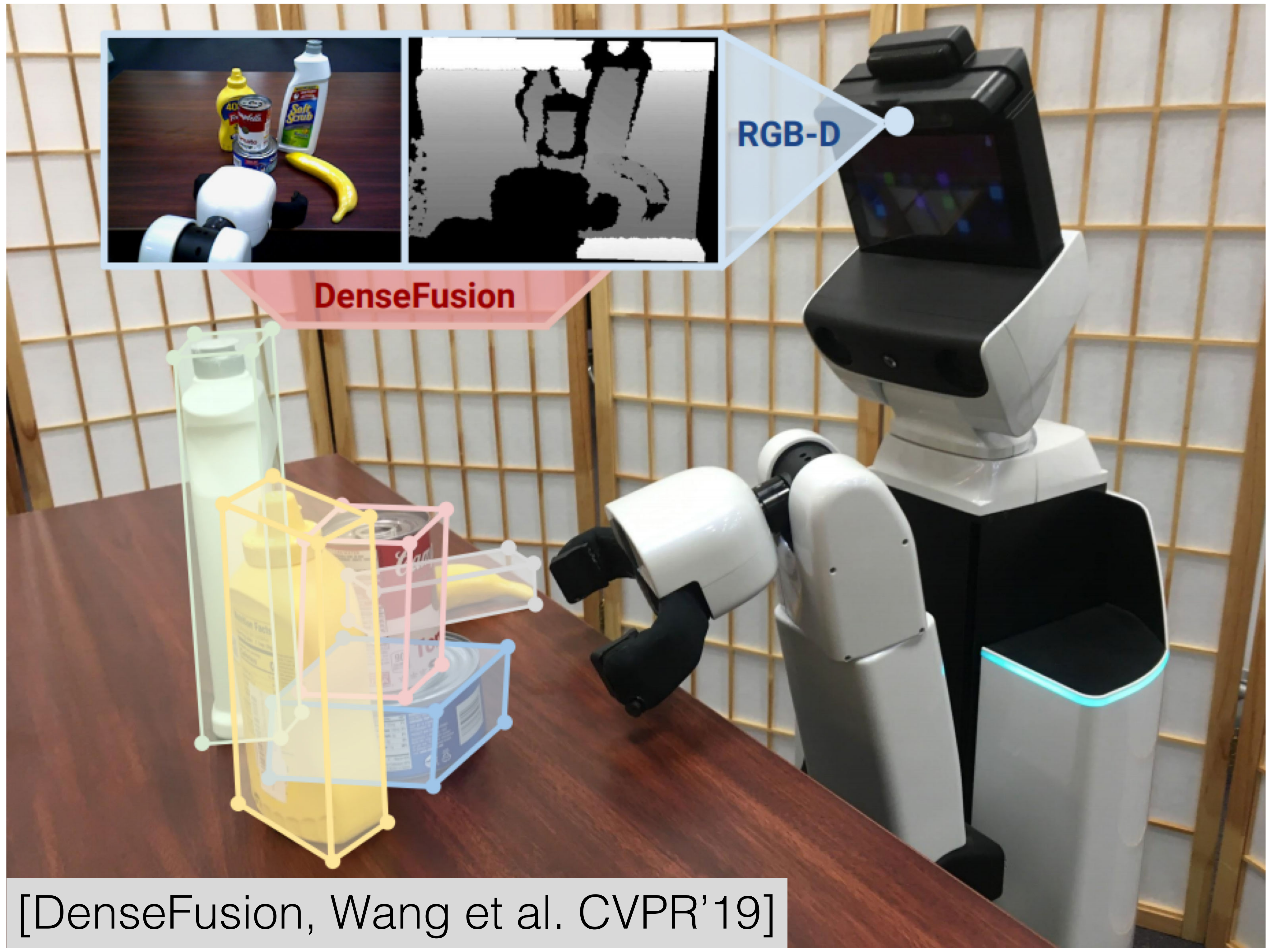
Applications

Activity understanding

Motion planning

Augmented Reality

Related Work



[DenseFusion, Wang et al. CVPR'19]

Traditional methods

- Hinterstoisser et al. ACCV' 12
- Choi & Christensen IROS' 13
- Collet et al. ICRA' 11
- Lepetit et al. TPAMI' 04

More robust against occlusion and illumination changes

Learning-based methods

- DOPE [Tremblay et al. CoRL'18]
- PoseCNN [Xiang et al. RSS'18]
- PoseRBPF [Deng et al. RSS'19]
- DeepIM [Li et al. ECCV' 18]
- DenseFusion [Wang et al. CVPR'19]

However, model-based 6D tracker assumes known 3D model of the object and fails to generalize to unseen objects.

Pose estimation without known model?

Category-level canonical pose

Category-Level 6D Pose Estimation

Training

Laptop Category

Synthetic data with ShapeNetCore** models (90% of the training data)

+

Real data with 3 objects (10% of the training data)

Testing

Real data with unseen objects

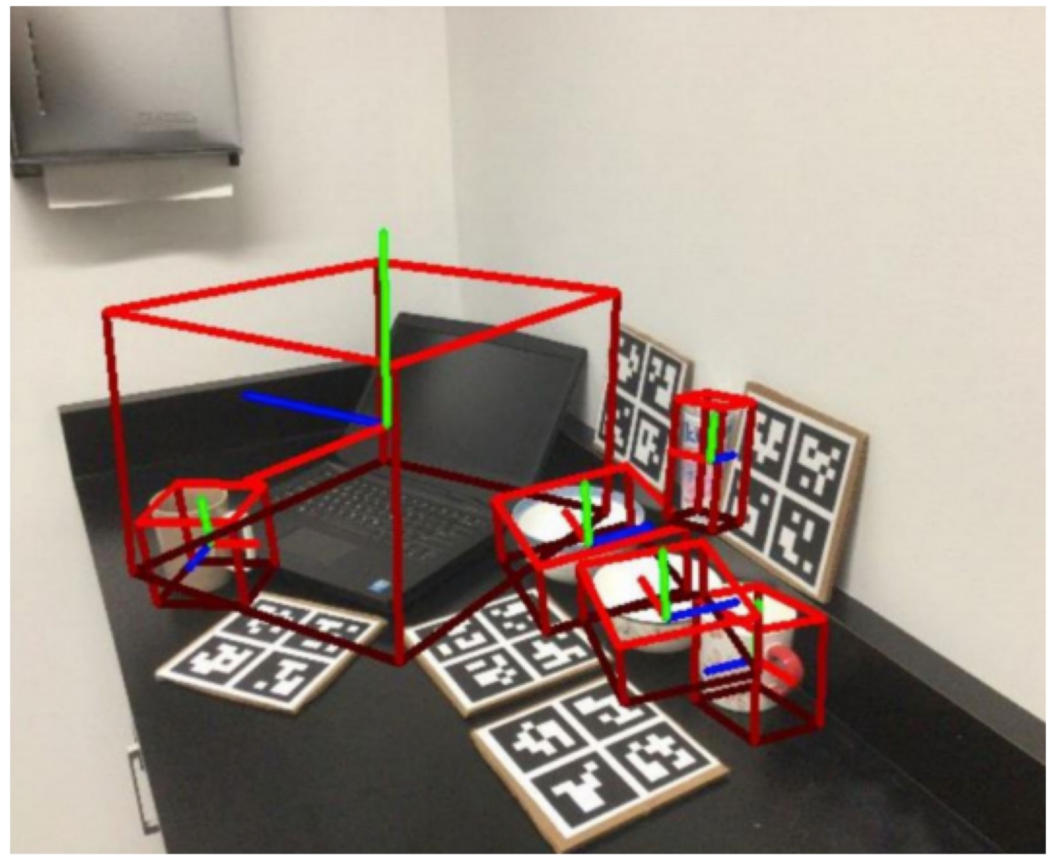
* Wang et al. 2019, “Normalized Object Coordinate Space for Category-Level 6D Object Pose and Size Estimation” CVPR2019

Category-Level 6D Pose Tracking

6D Pose Estimation

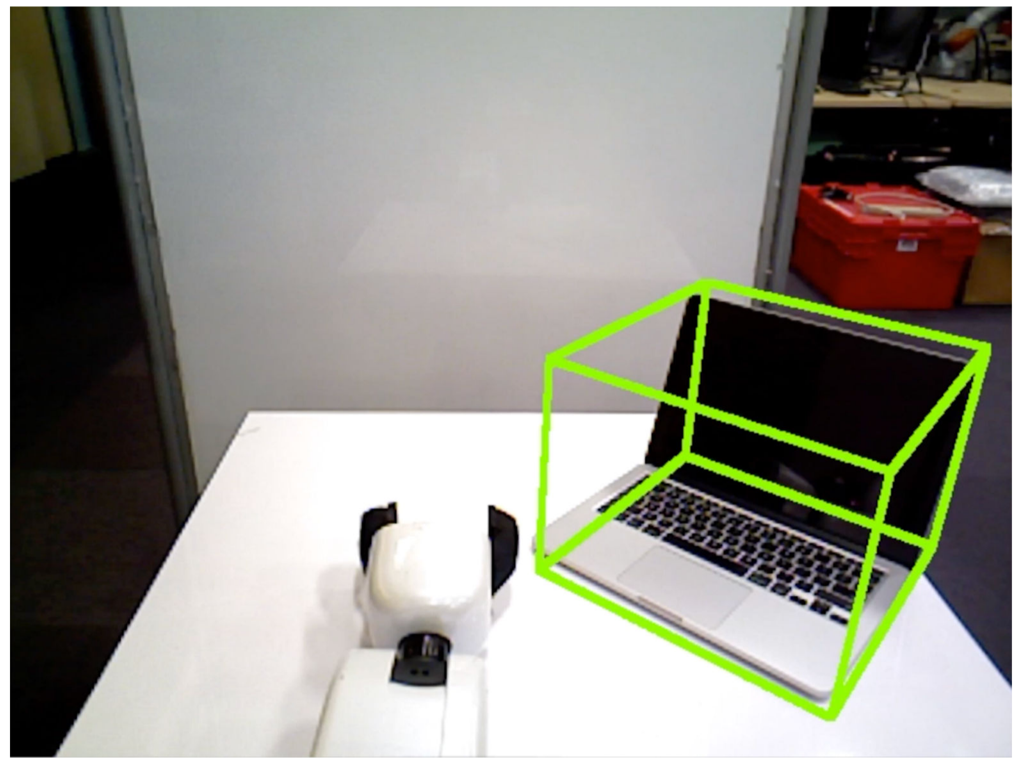
input: image frame

per-frame prediction



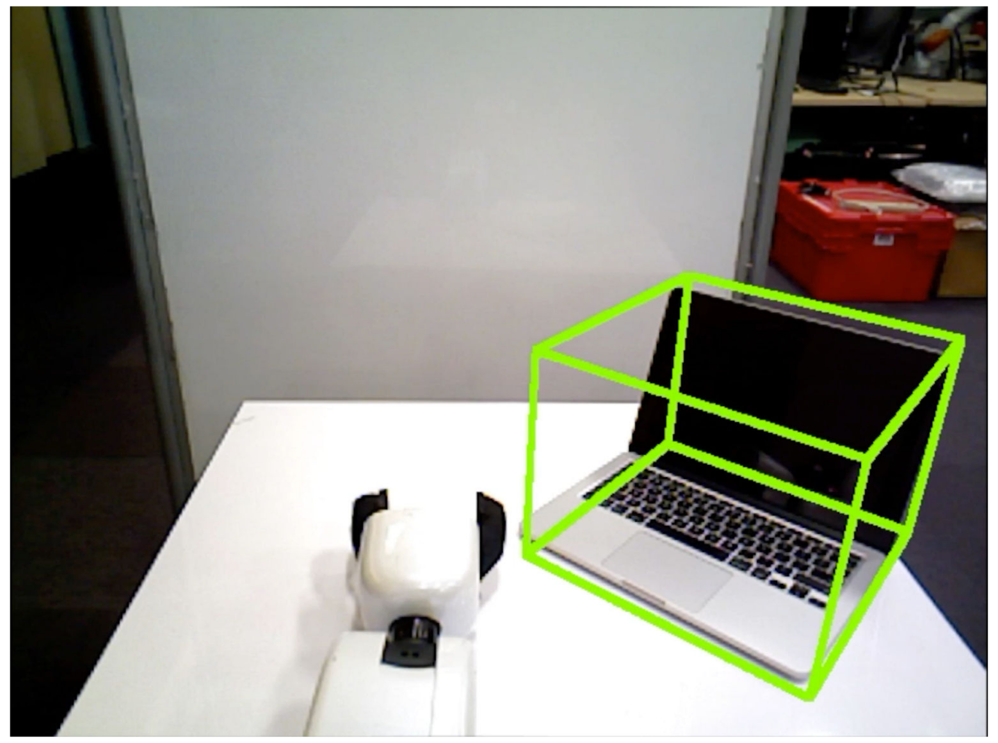
output: 6D pose

6D Pose Tracking



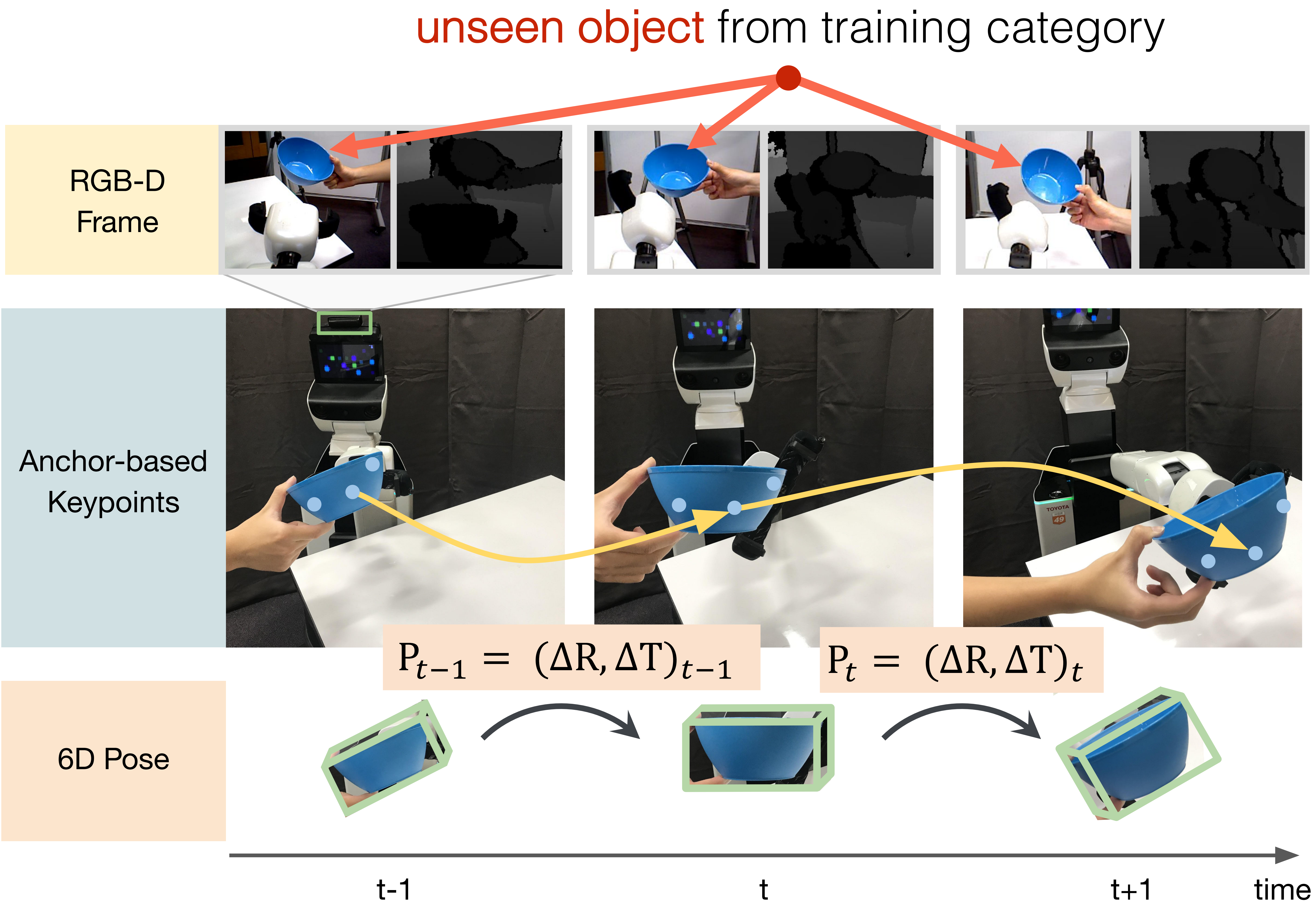
input: video + initial bbox

tracking over time



output: 6D pose

Category-Level 6D Pose Tracking



6D pose estimation of current frame

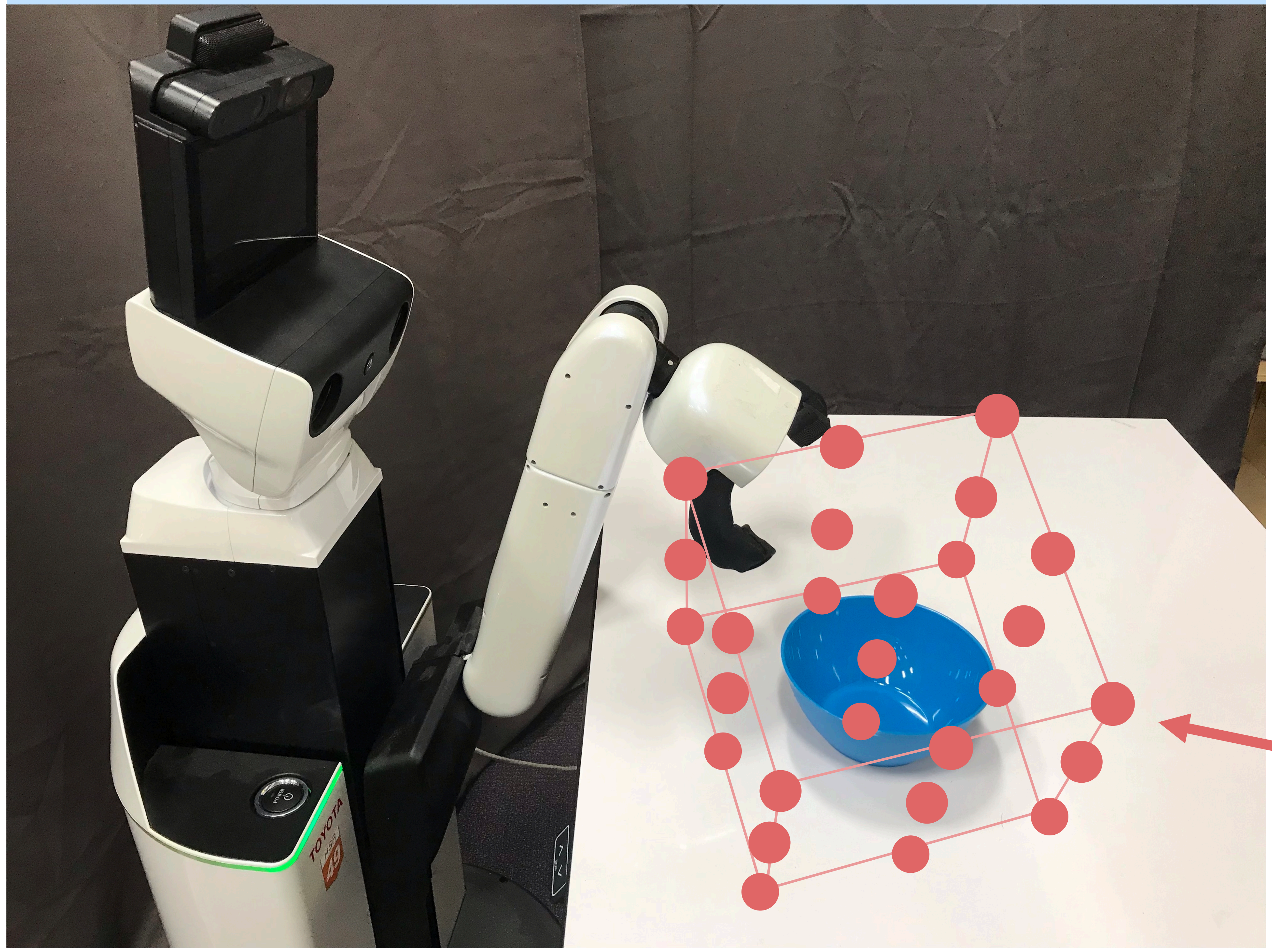
$$P_t = P_0[\Delta R|\Delta T]_1 \cdots [\Delta R|\Delta T]_{t-1}$$

6-PACK: **6**D **P**ose **A**nchor-based **C**ategory-level **K**eypoint tracker

Previous Frame

6-PACK: **6**D **P**ose **A**nchor-based **C**ategory-level **K**eypoint tracker

Previous Frame



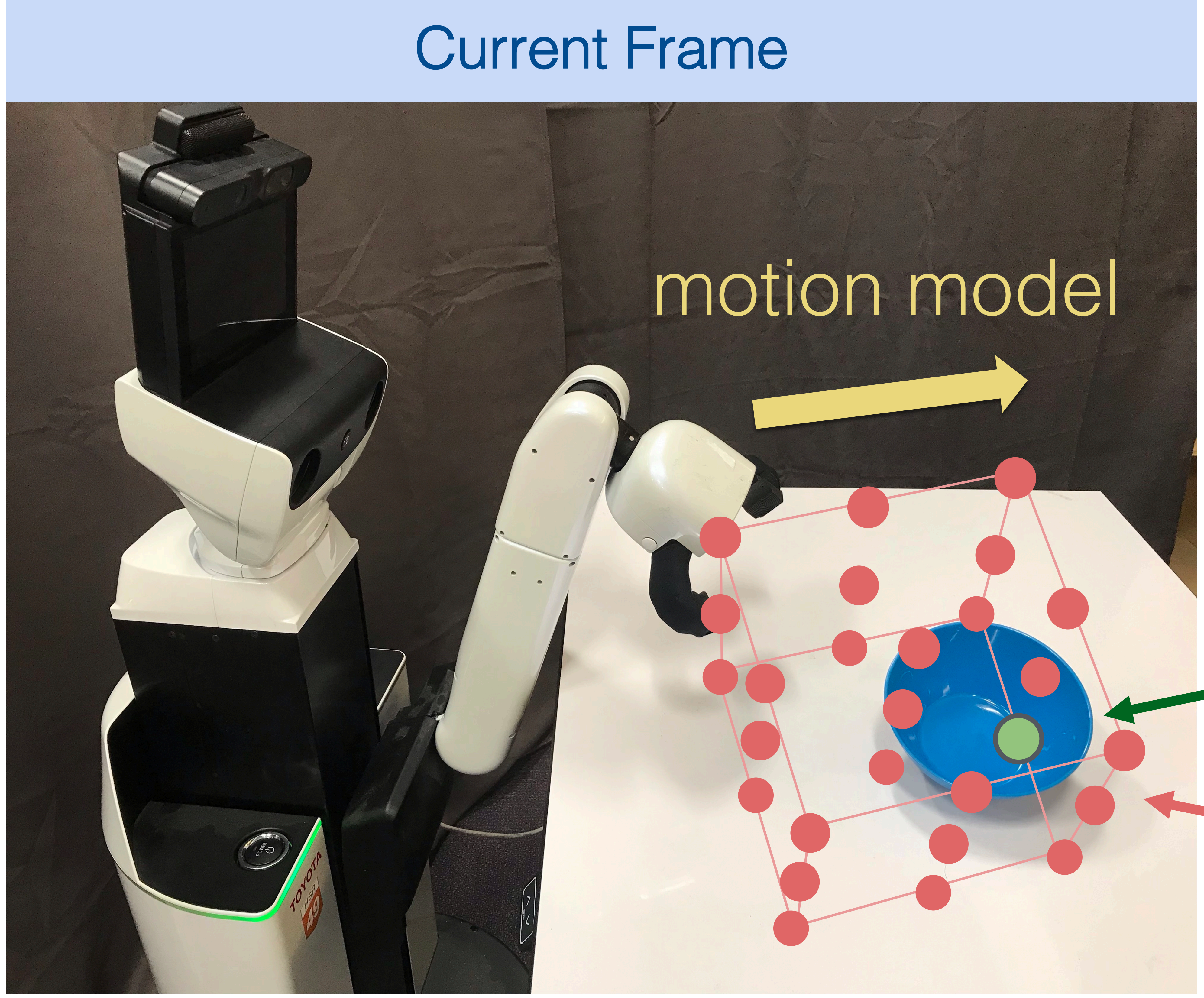
Challenge: No ground-truth object model

Idea: Use anchor box as a scaffold

each anchor point captures local information

3D anchors around the previous pose

6-PACK: **6**D **P**ose **A**nchor-based **C**ategory-level **K**eypoint tracker



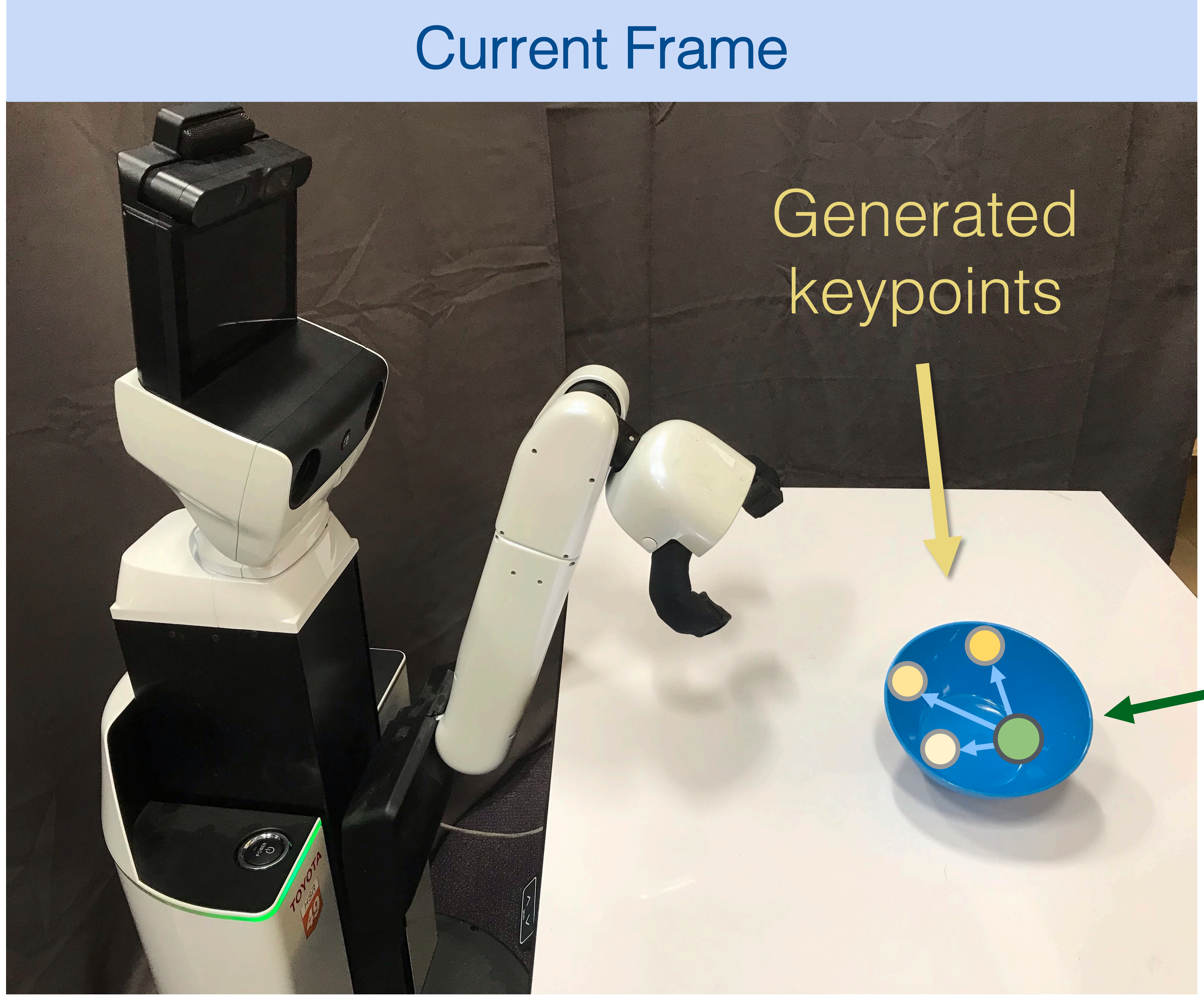
Challenge: Incorporate temporal information

Idea: Use motion model to predict the likely position of the object

Selected anchor

3D anchors around the next (likely) pose

6-PACK: **6**D **P**ose **A**nchor-based **C**ategory-level **K**eypoint tracker

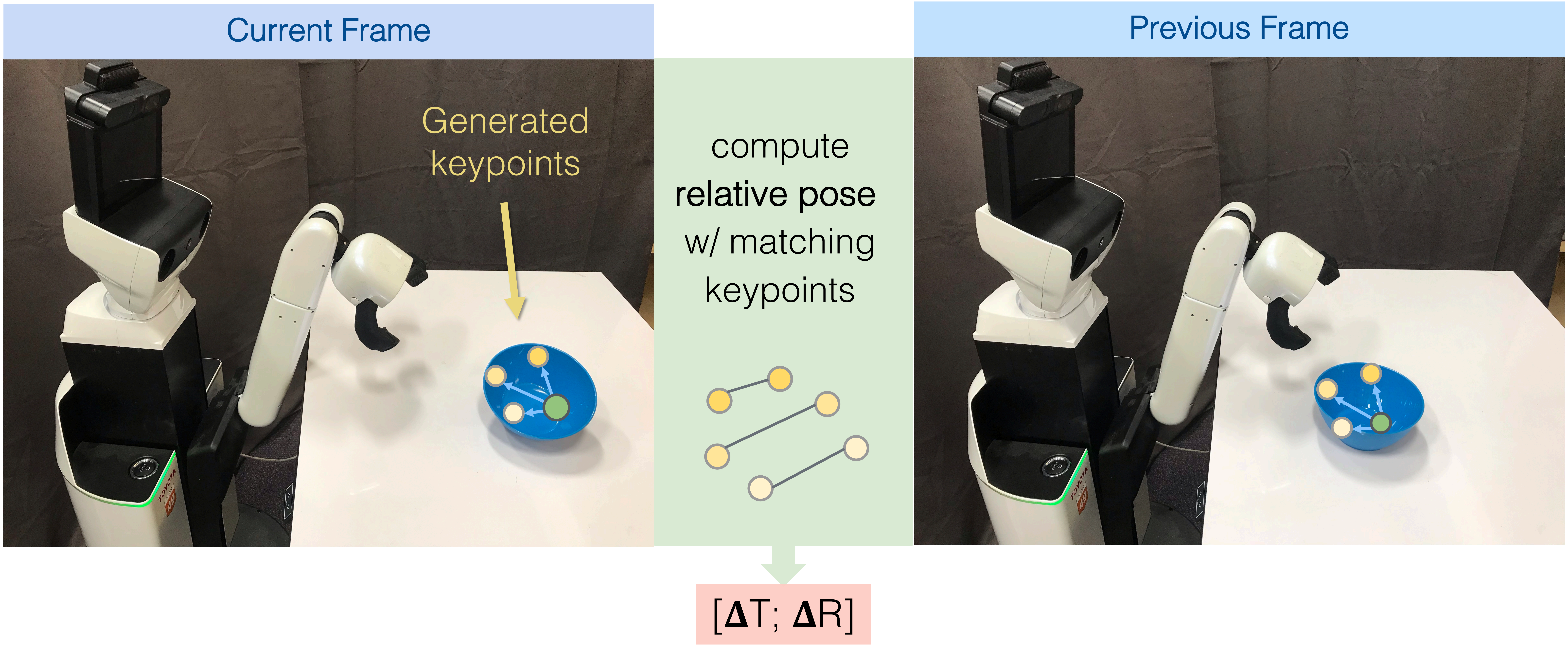


Selected
anchor

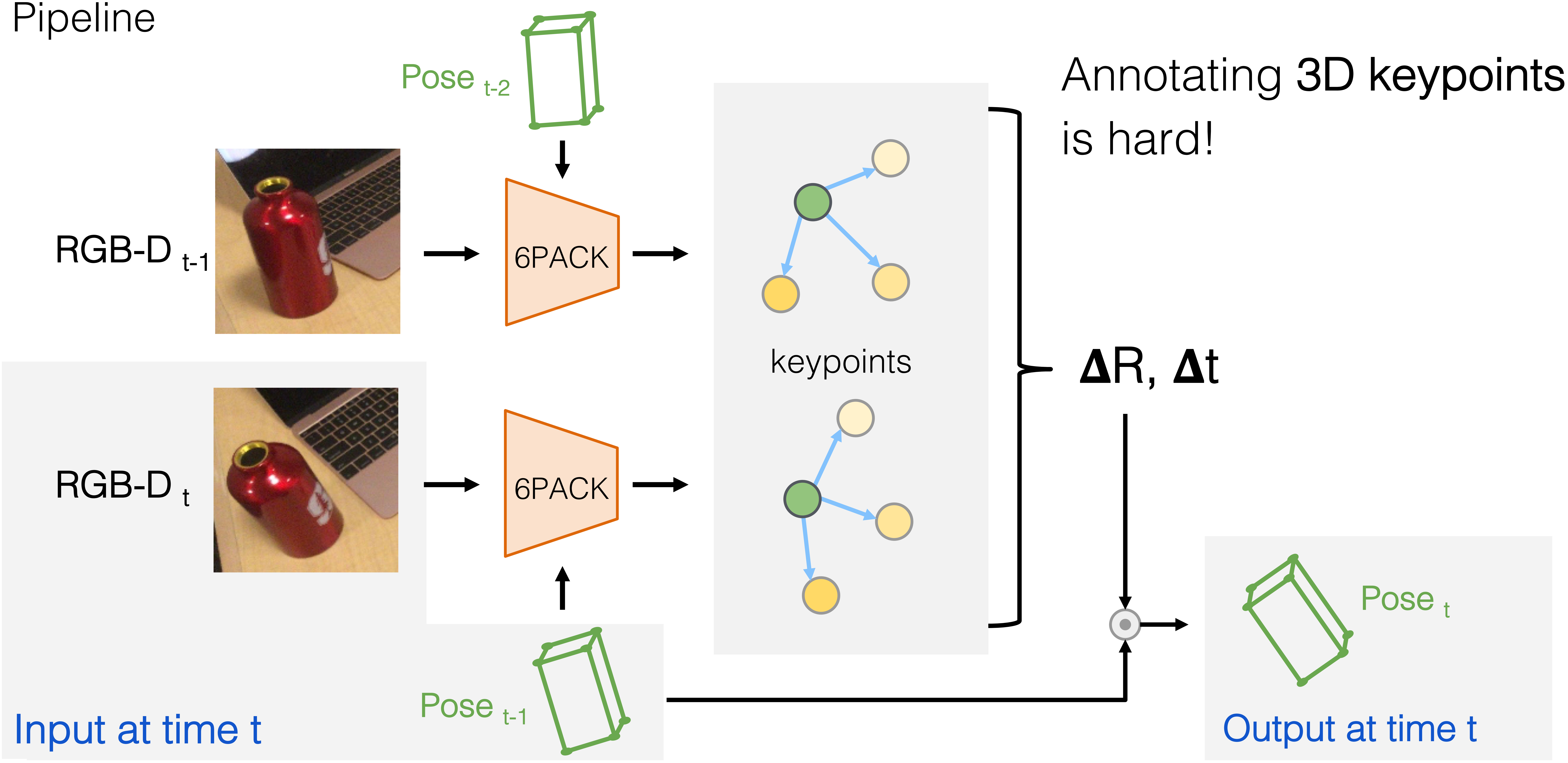
Challenge: Generate 3D keypoints

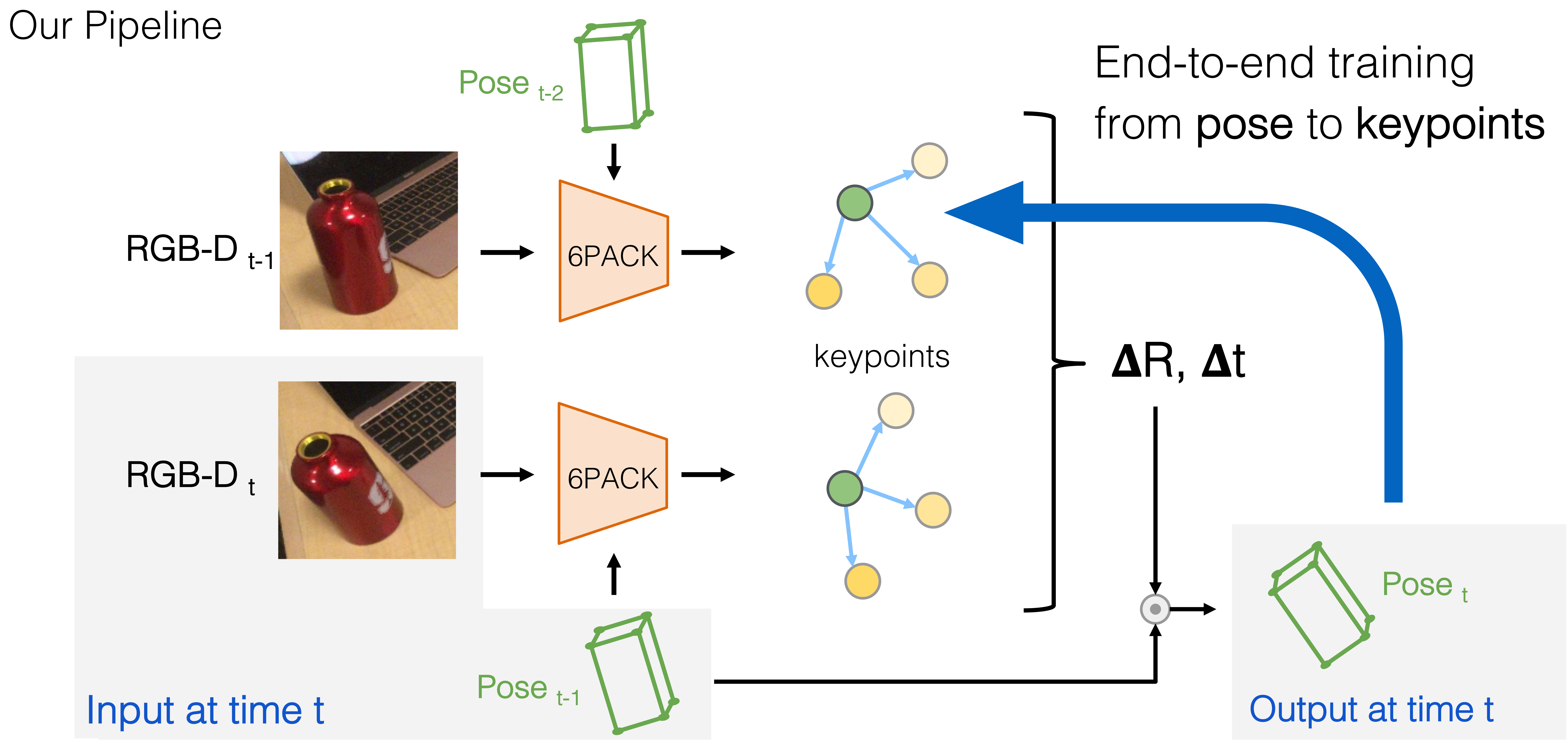
Idea: Predict offsets from selected anchor

6-PACK: **6**D **P**ose **A**nchor-based **C**ategory-level **K**eypoint tracker



Our Pipeline

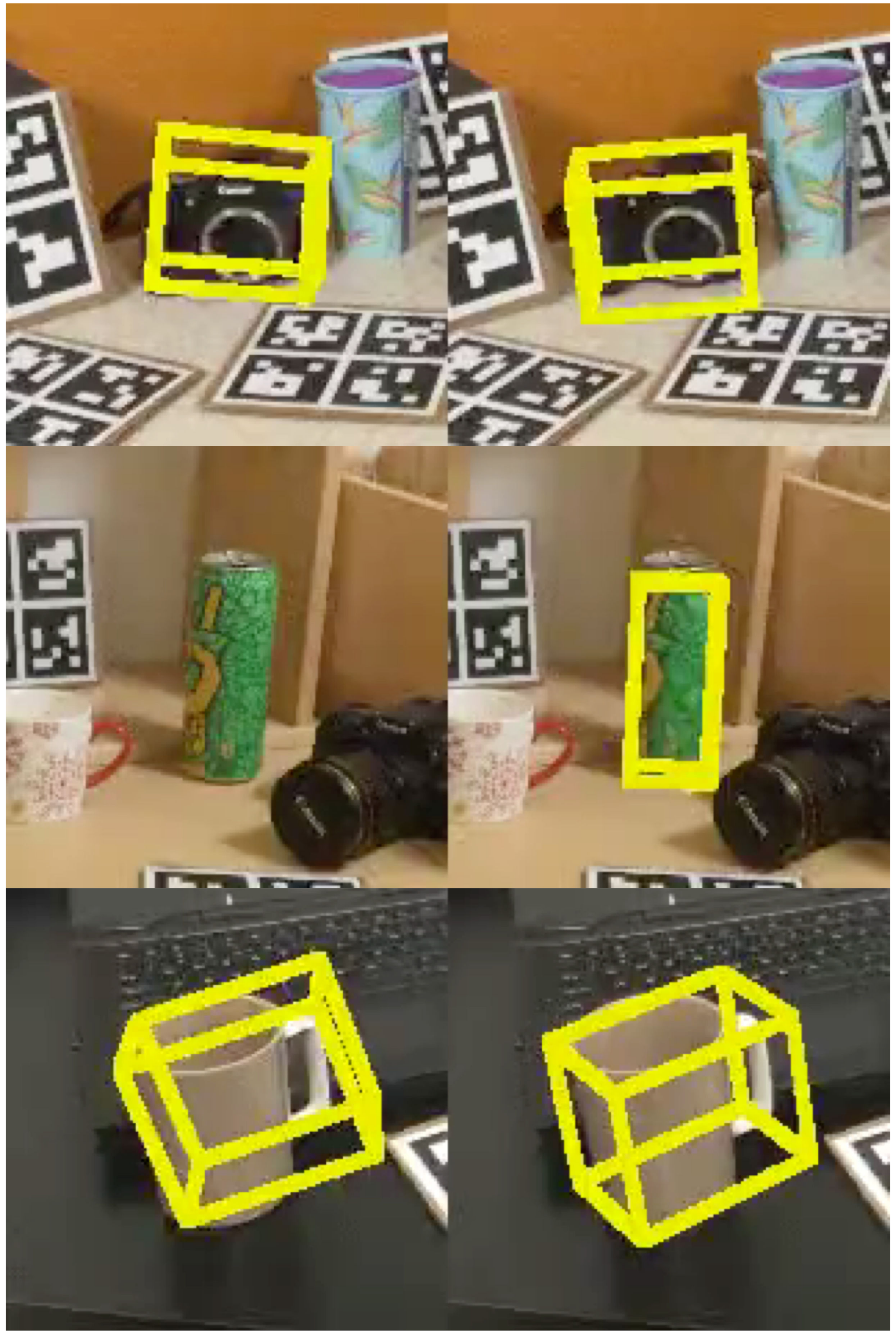



Our Pipeline
Pose t-2
RGB-D t-1
6PACK
keypoints
End-to-end training
from **pose** to **keypoints**
ΔR, Δt
RGB-D t
6PACK
Pose t-1
Input at time t
Pose t
Output at time t

Evaluation Results

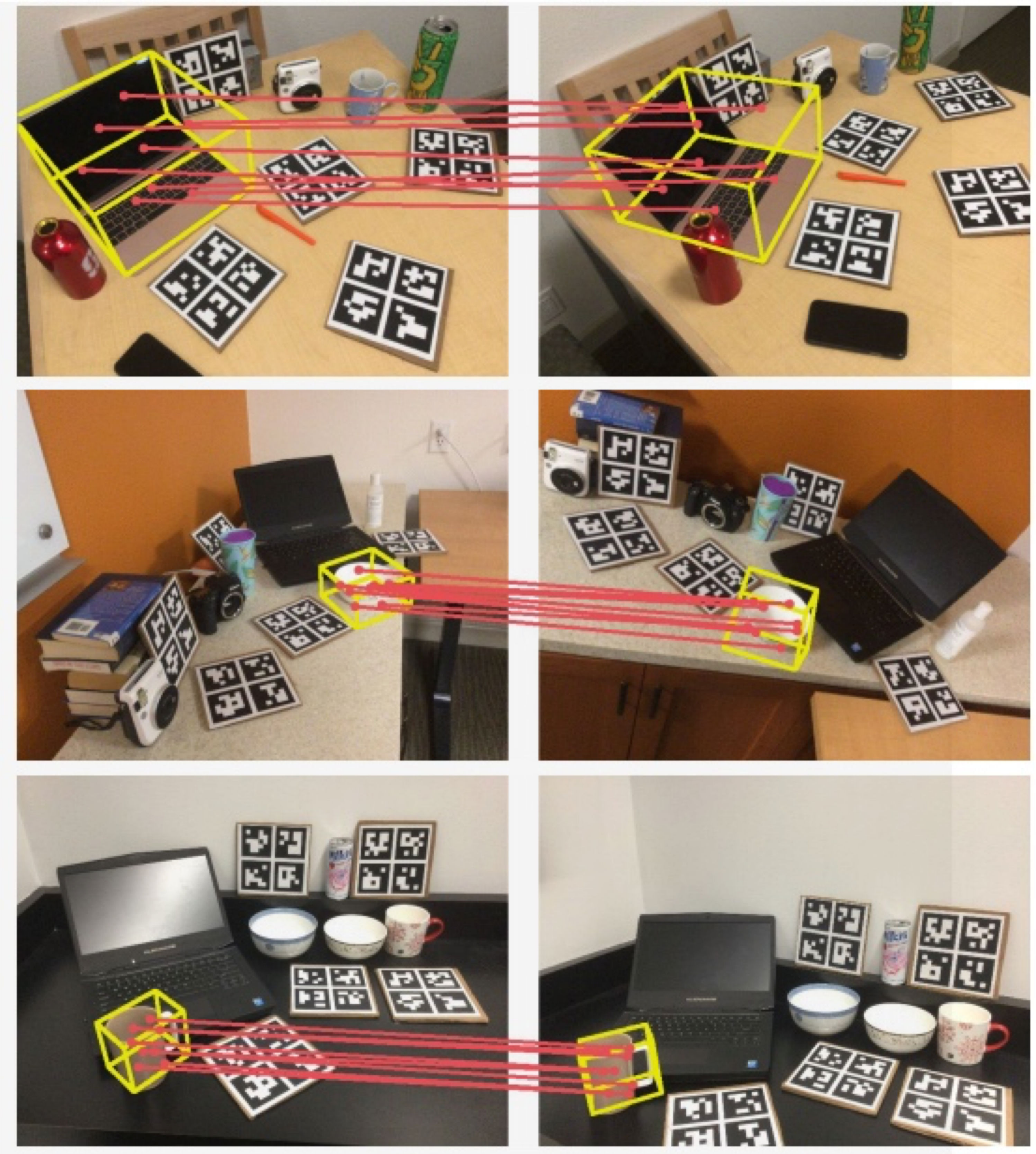
		NOCS [46]	ICP [50]	Keypoint Net [41]	Ours w/o temporal	Ours
bottle	5°5cm	5.5	10.1	5.9	23.7	**24.5**
	IoU25	48.7	29.9	23.1	**92.0**	91.1
	R_{err}	25.6	48.0	28.5	15.7	**15.6**
	T_{err}	14.4	15.7	9.5	4.2	**4.0**
bowl	5°5cm	**62.2**	40.3	16.8	53.0	55.0
	IoU25	99.6	79.7	74.7	**100.0**	**100.0**
	R_{err}	**4.7**	19.0	9.8	5.3	5.2
	T_{err}	**1.2**	4.7	8.2	1.6	1.7
camera	5°5cm	0.6	**12.6**	1.8	8.4	10.1
	IoU25	90.6	53.1	30.9	**91.0**	87.6
	R_{err}	**33.8**	80.5	45.2	43.9	35.7
	T_{err}	**3.1**	12.2	8.5	5.5	5.6
can	5°5cm	7.1	17.2	4.3	**25.0**	22.6
	IoU25	77.0	40.5	42.6	89.9	**92.6**
	R_{err}	16.9	47.1	28.8	**12.5**	13.9
	T_{err}	**4.0**	9.4	13.1	5.0	4.8
laptop	5°5cm	25.5	14.8	49.2	62.4	**63.5**
	IoU25	94.7	50.9	94.6	97.8	**98.1**
	R_{err}	8.6	37.7	6.5	4.9	**4.7**
	T_{err}	**2.4**	9.2	4.4	2.5	2.5
mug	5°5cm	0.9	6.2	3.1	22.4	**24.1**
	IoU25	82.8	27.7	52.0	**100.0**	95.2
	R_{err}	31.5	56.3	61.2	**20.3**	21.3
	T_{err}	4.0	9.2	6.7	**1.8**	2.3
Overall	5°5cm	17.0	16.9	13.5	32.5	**33.3**
	IoU25	82.2	47.0	53.0	**95.1**	94.2
	R_{err}	20.2	48.1	30.0	17.1	**16.0**
	T_{err}	4.9	10.5	8.4	**3.4**	3.5

6-PACK's 6D pose tracking accuracy is still higher than NOCS for more than 12%.

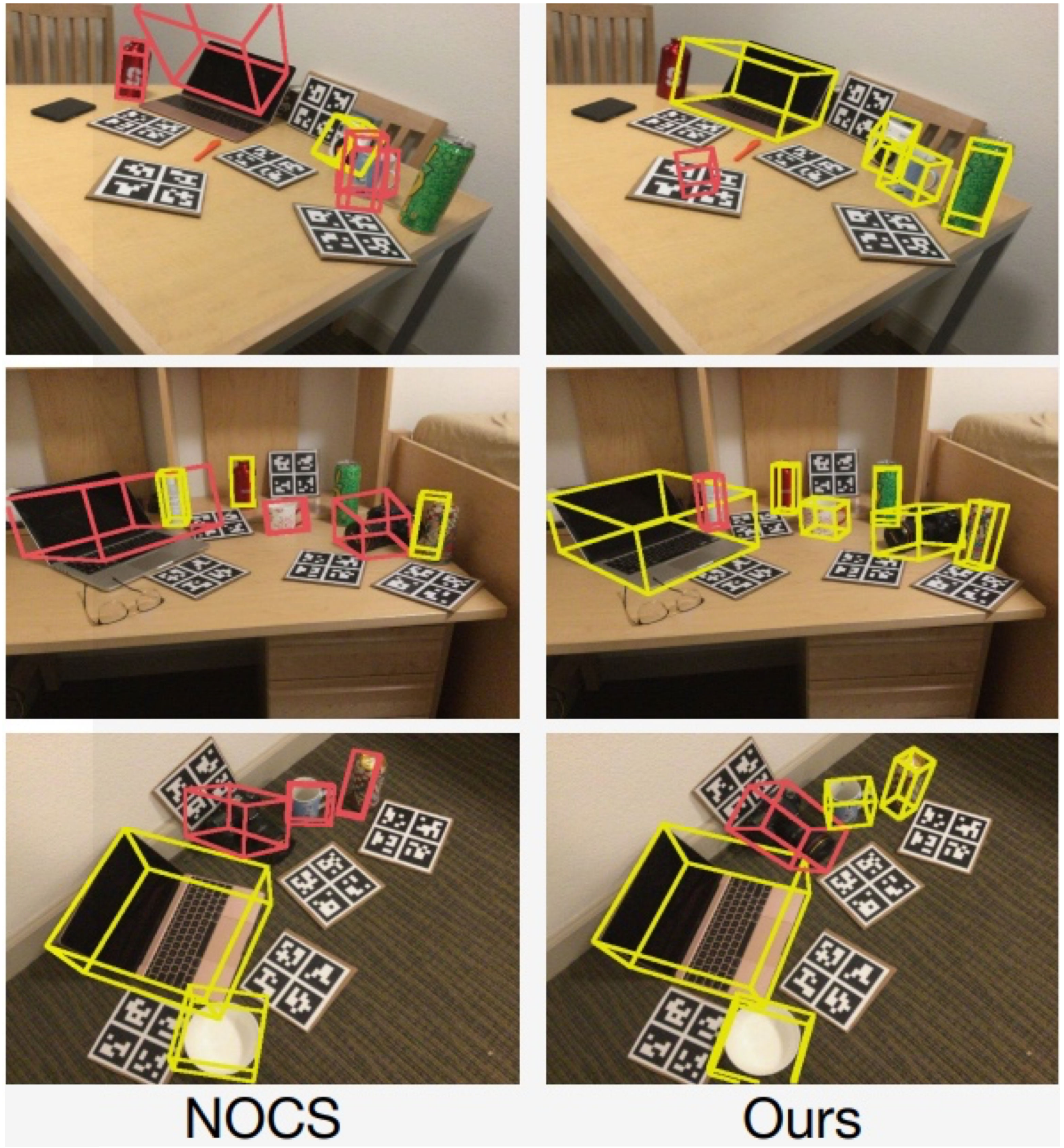


NOCS Ours

Qualitative Evaluation Results



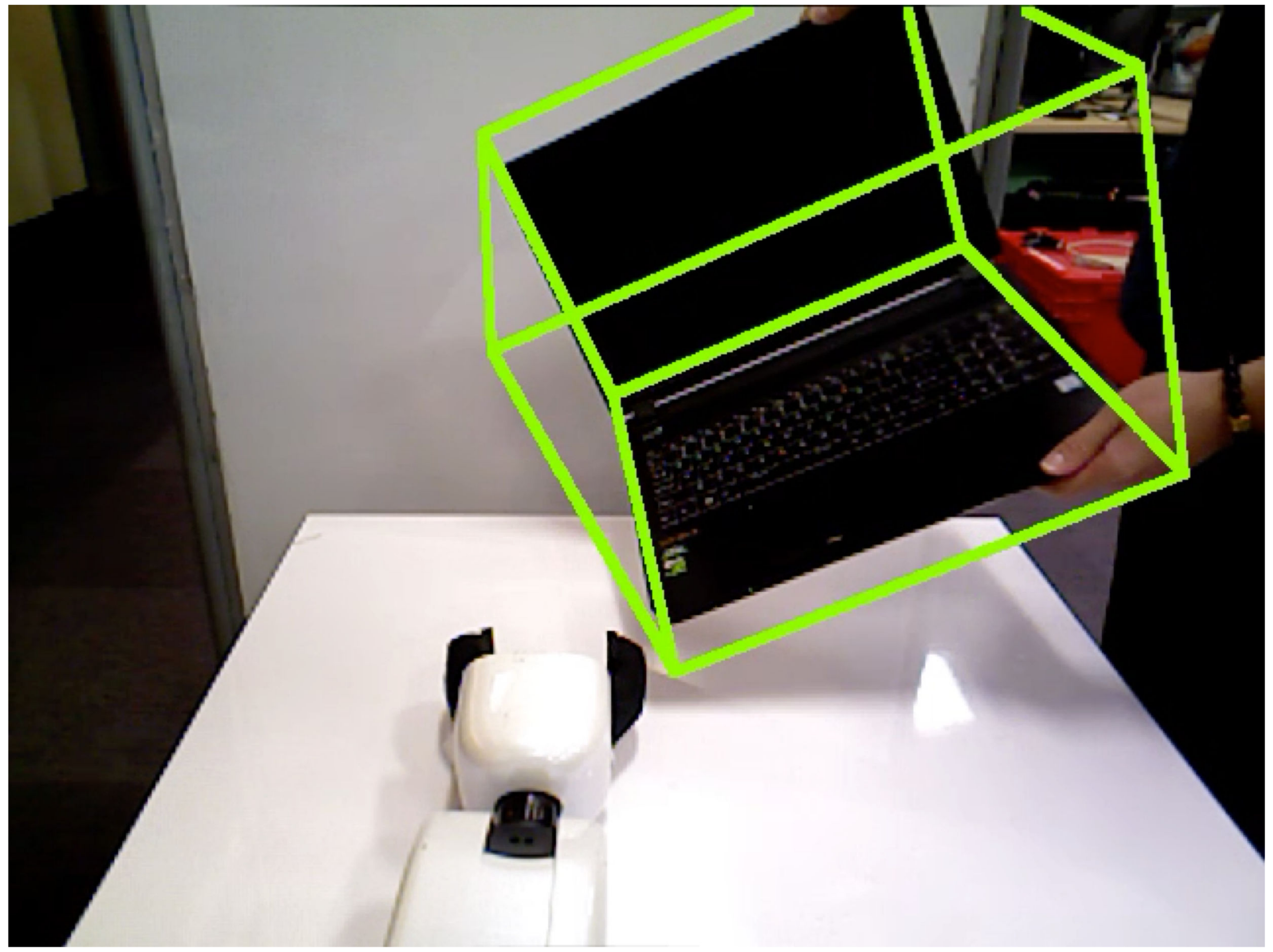
Keypoints generation results
(Matchings of each keypoint from different view)



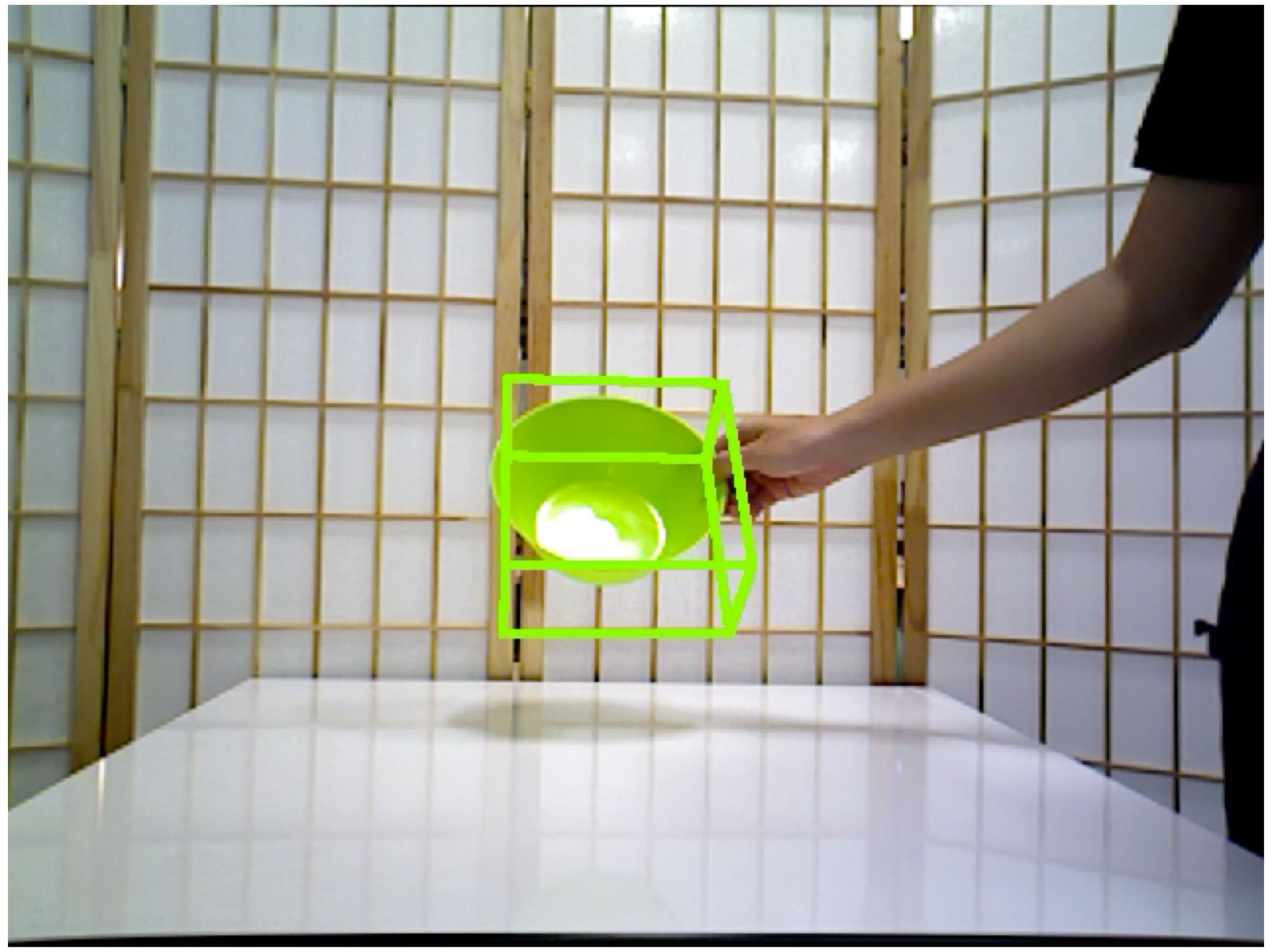
Qualitative results
(Red bounding box refers to pose error larger than 5cm 5°)

Real-time testing results on unseen objects

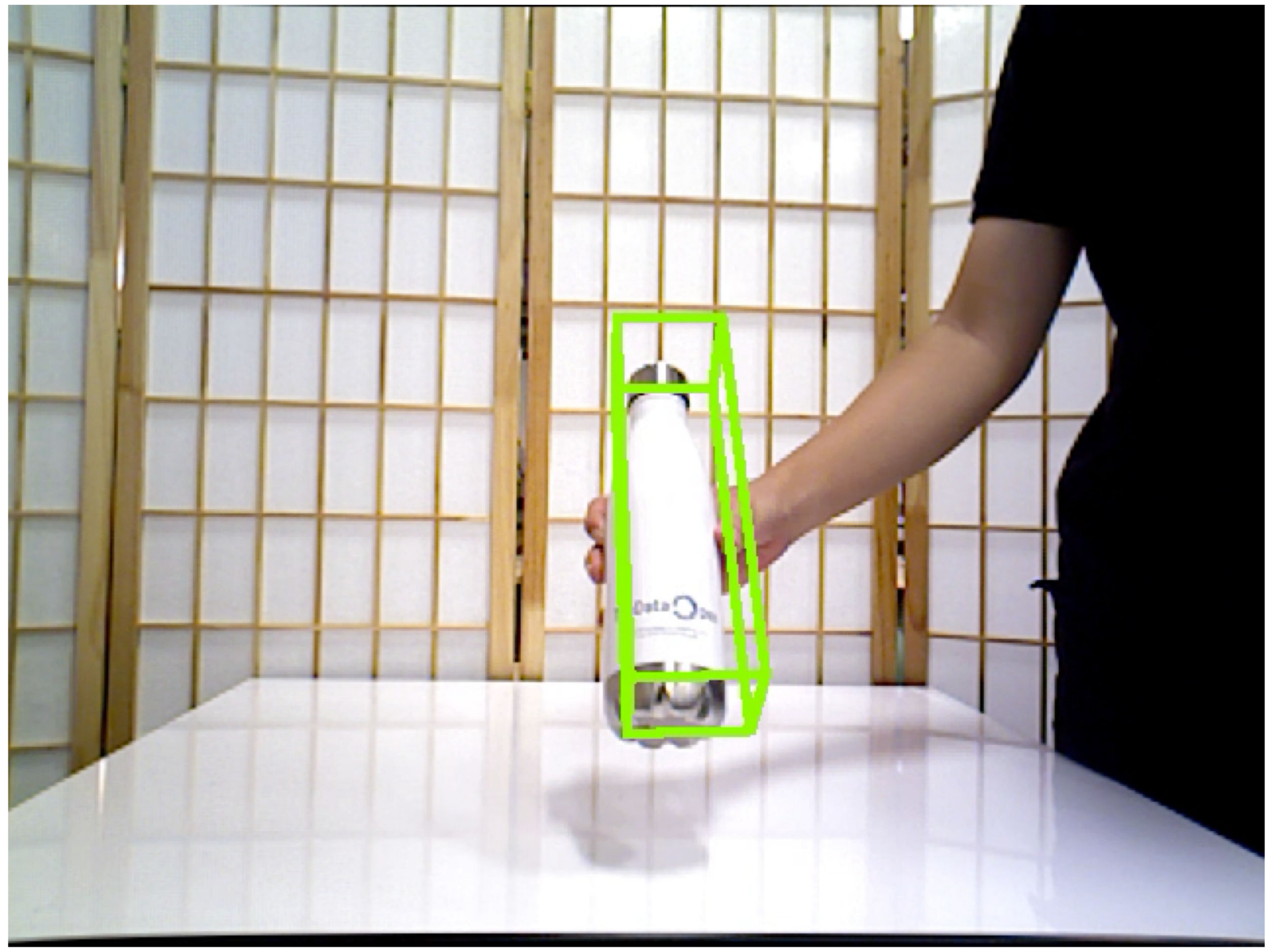
Tracker runs at 10 fps on an NVIDIA GTX1070 GPU



laptop



bowl



bottle


Robot view
real-time tracking on unseen object
x4

Summary

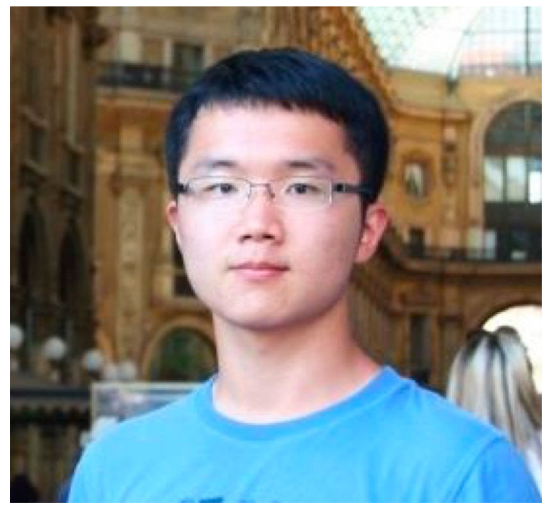
- 3D keypoints are compact object representations for 6D tracking
- End-to-end learning without manual keypoint annotations
- Real-time category-level tracking for robot interaction

Code: github.com/j96w/6PACK

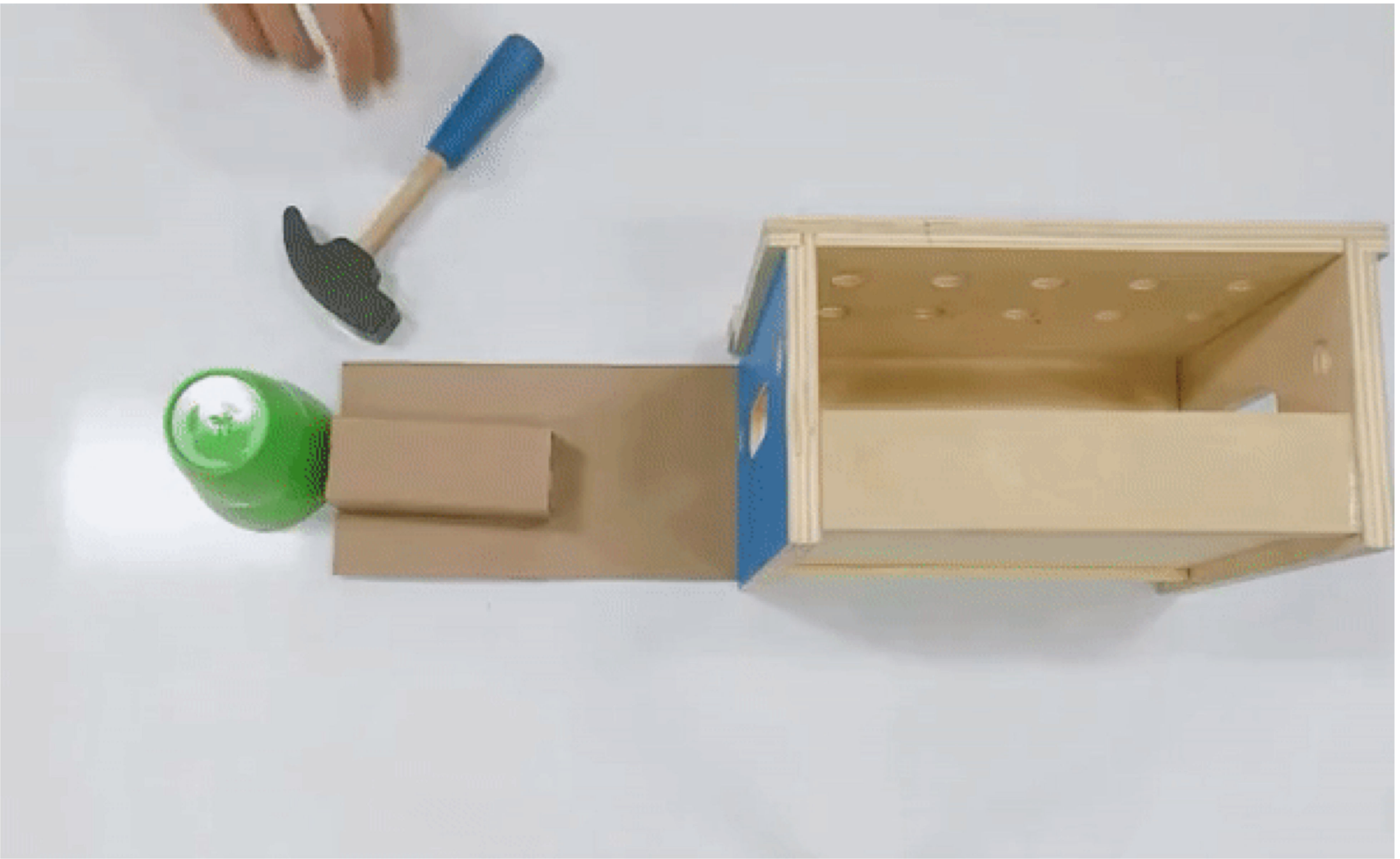
KETO: Learning Keypoint Representations for Tool Manipulation

Zengyi Qin, Kuan Fang, **Yuke Zhu**,

Li Fei-Fei, Silvio Savarese



Vision-Based Tool Manipulation



Recognizing tools

Understanding tools

Manipulating tools

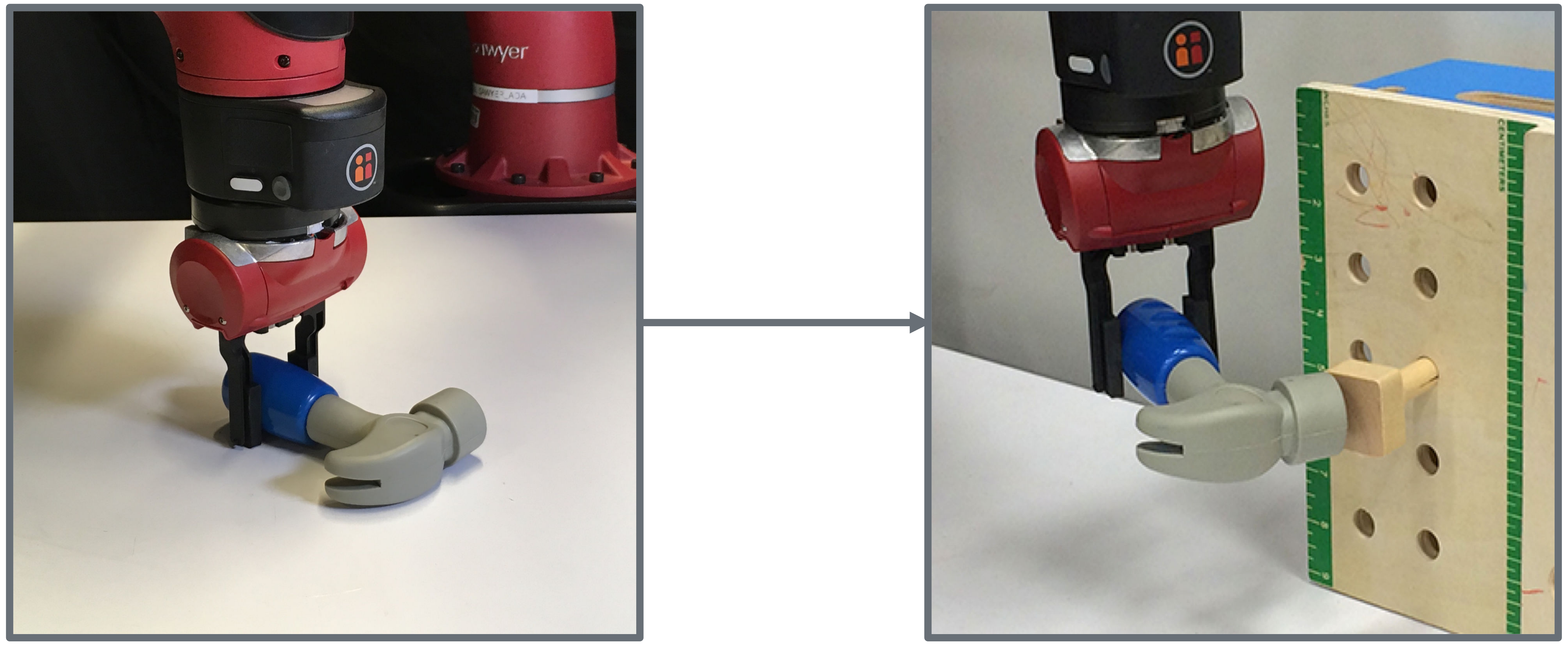
Vision-Based Tool Manipulation

[Fang, Zhu, et al., RSS'18]

Vision-Based Tool Manipulation

Tool manipulation as a two-stage process

- Stage 1: Grasp an object as a tool.
- Stage 2: Use the grasped tool to complete the goal of the task.

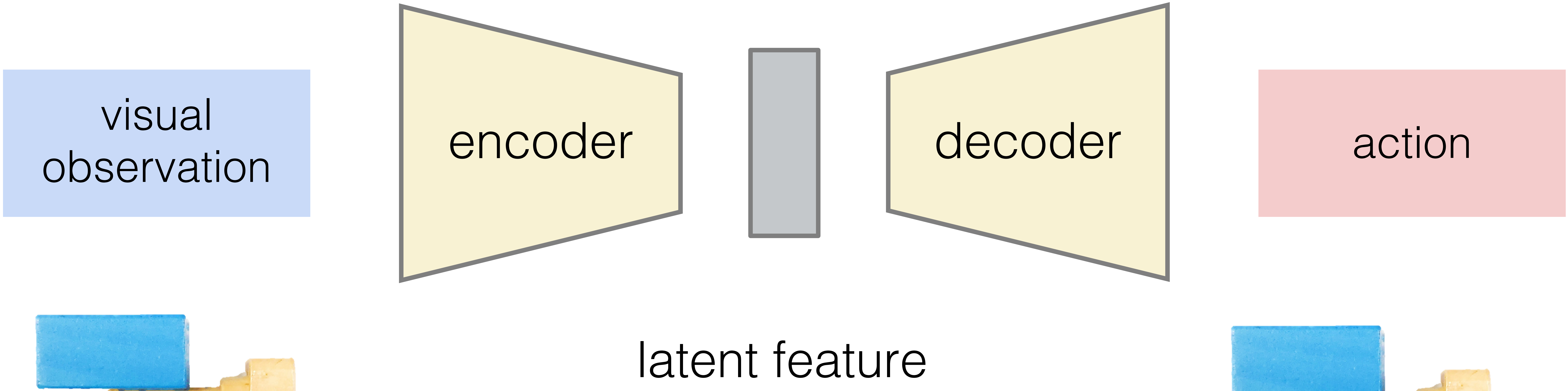


[Fang, Zhu, et al., RSS'18]

Vision-Based Tool Manipulation

visual observation → representation → action

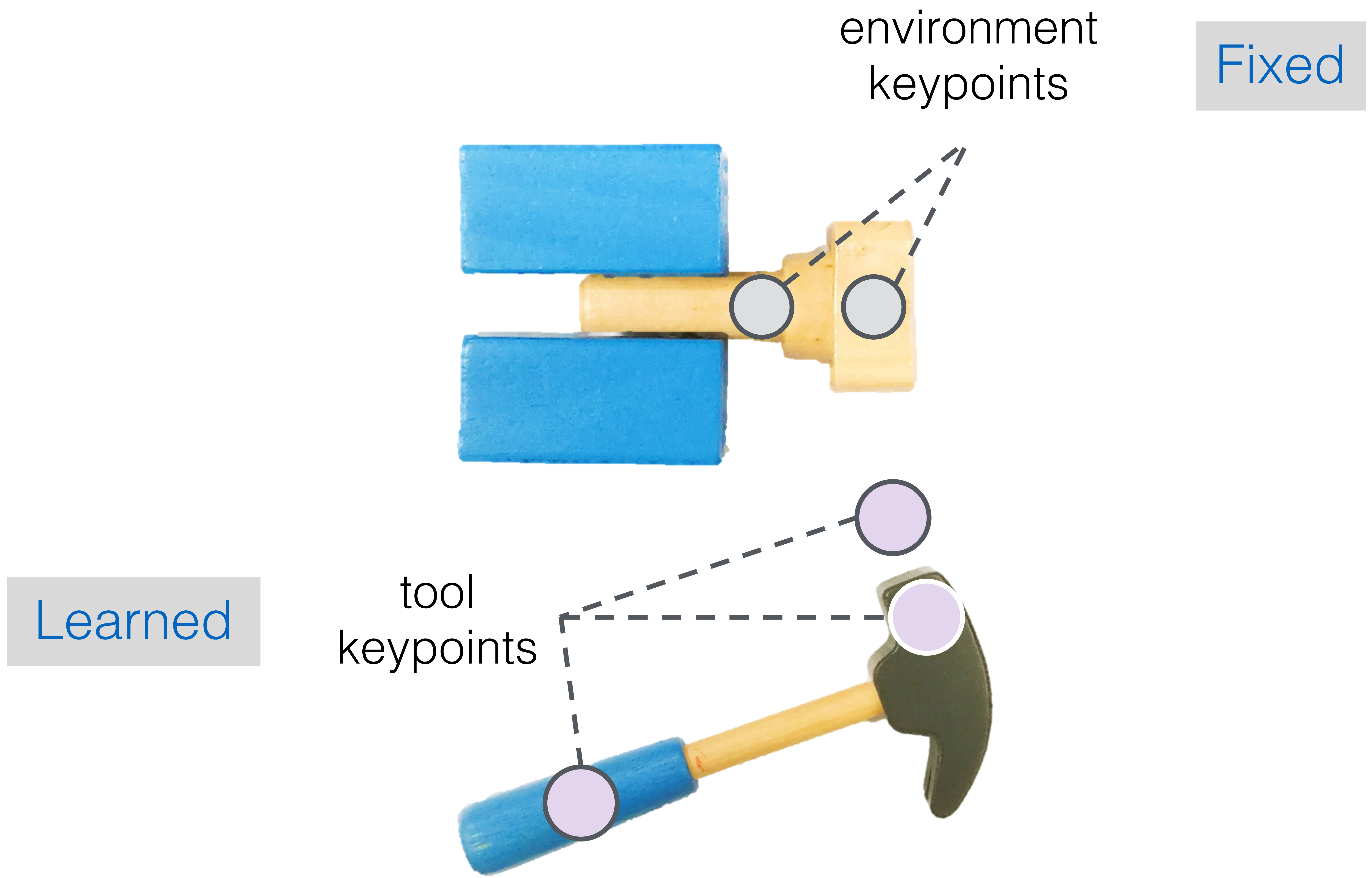
Vision-Based Tool Manipulation



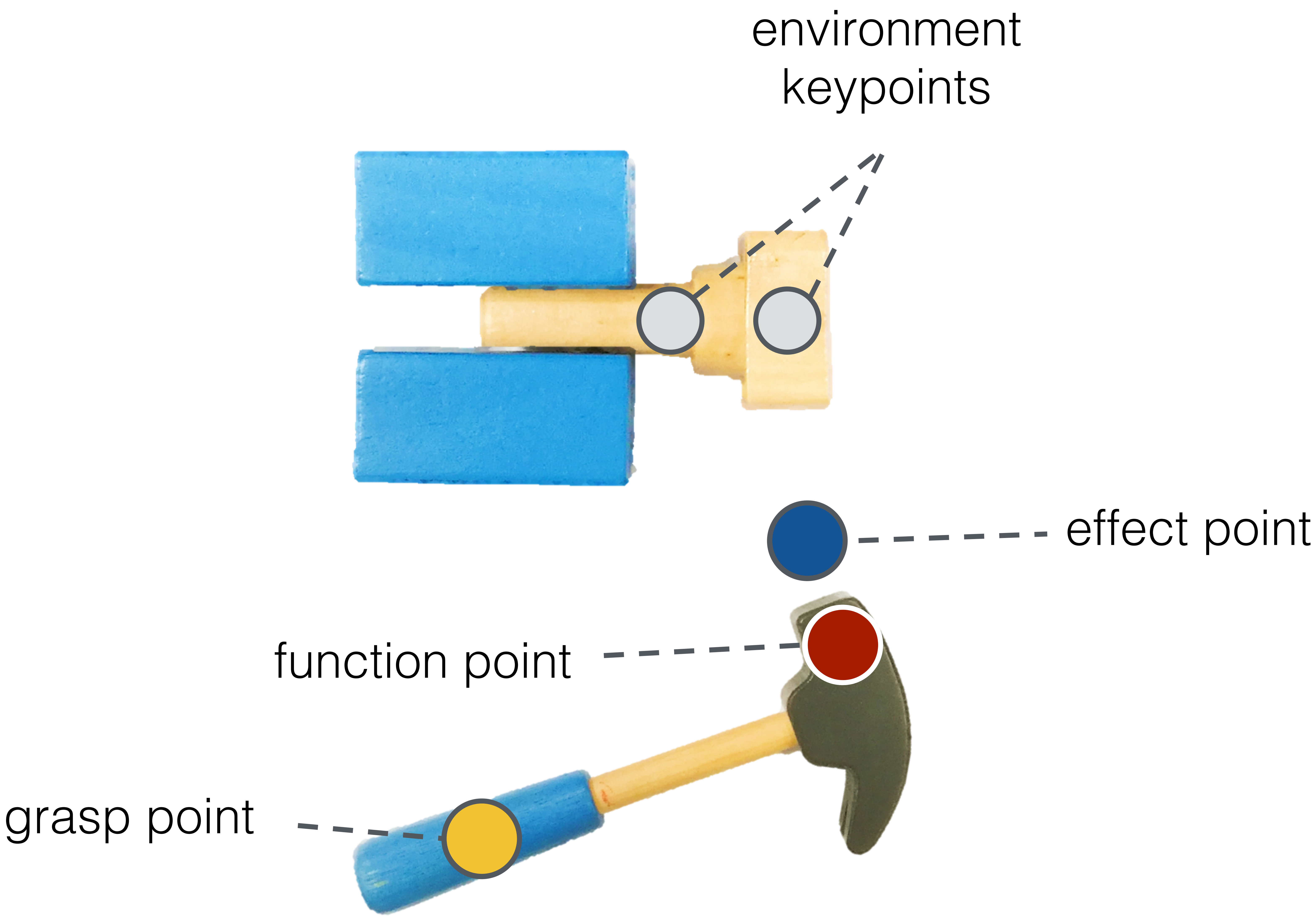
[Fang, Zhu, et al., RSS'18]

- High-dimensionality
- Lack of interpretability

KETO: **Ke**ypoint Representations for **To**ol Manipulation



KETO: **Ke**ypoint Representations for **To**ol Manipulation



KETO: Keypoint Representations for Tool Manipulation

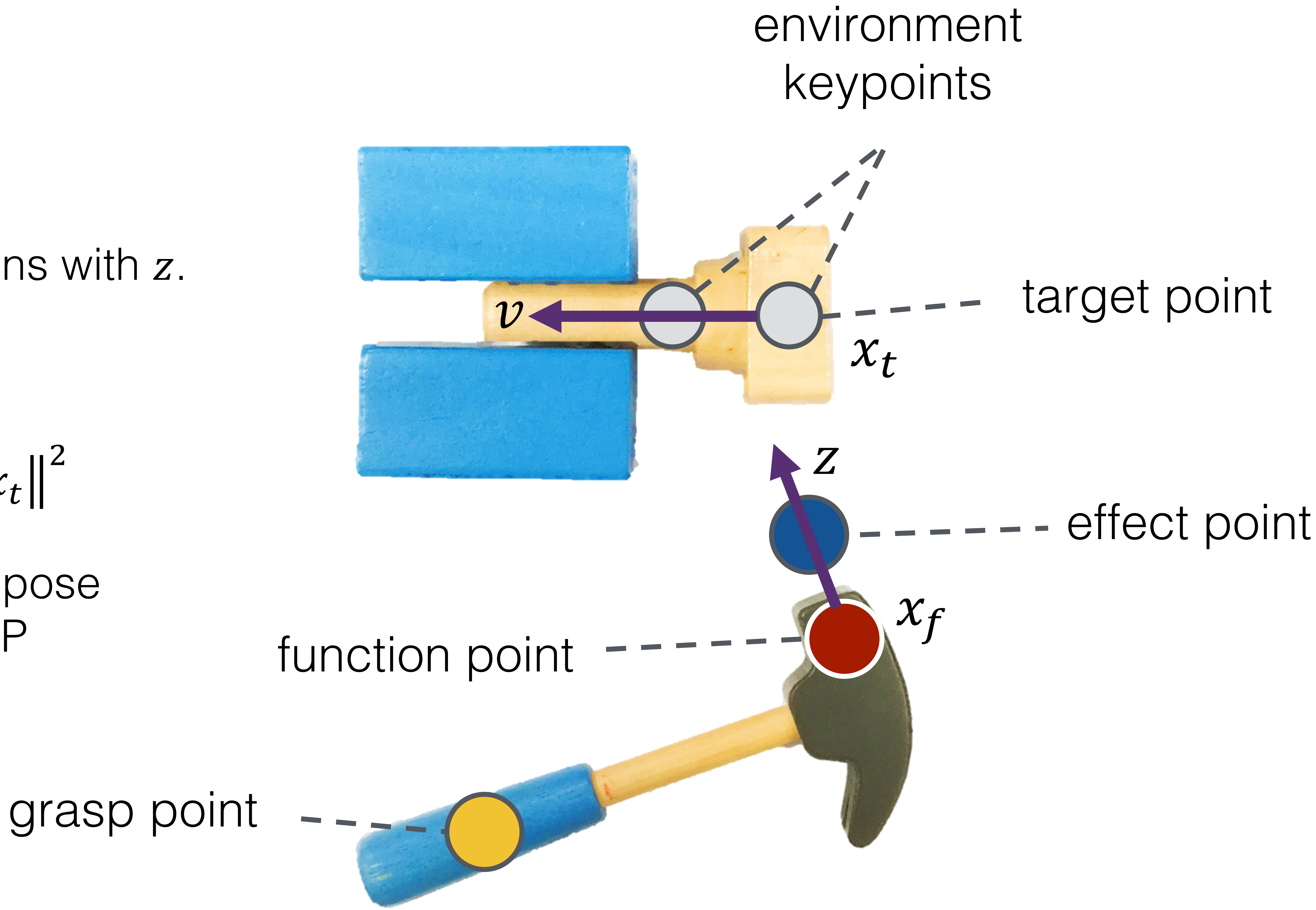
For **hammering**

1. x_t is close to x_f

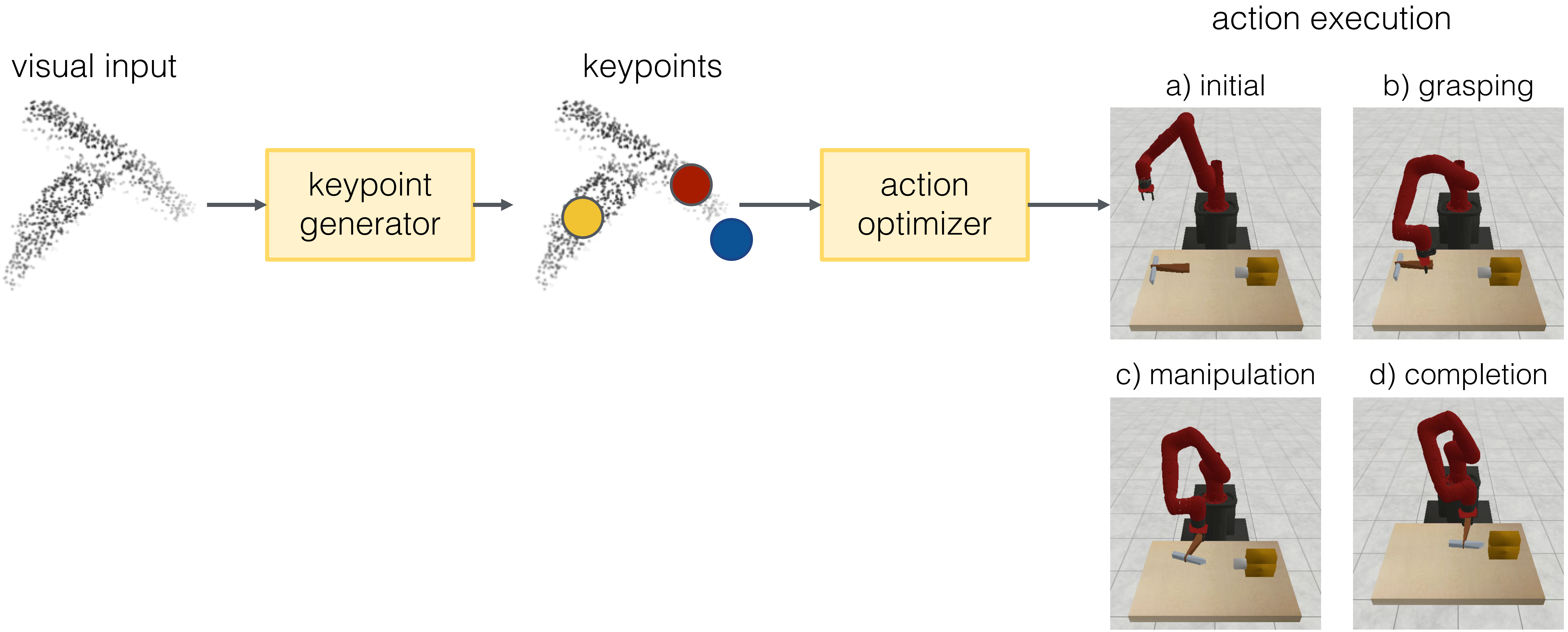
2. Direction of v aligns with z.

$$\max_{p} \; v^T z - \|x_f - x_t\|^2$$

Solving the optimal pose of object as a QP



KETO: Keypoint Representations for Tool Manipulation

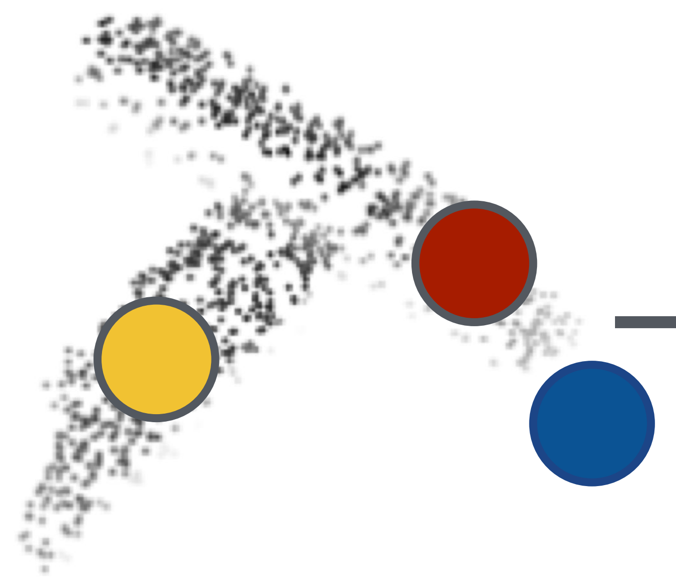


KETO: **Ke**ypoint Representations for **To**ol Manipulation

visual input

keypoint generator

keypoints



action optimizer

Conditional VAE + Discriminator

QP Solver

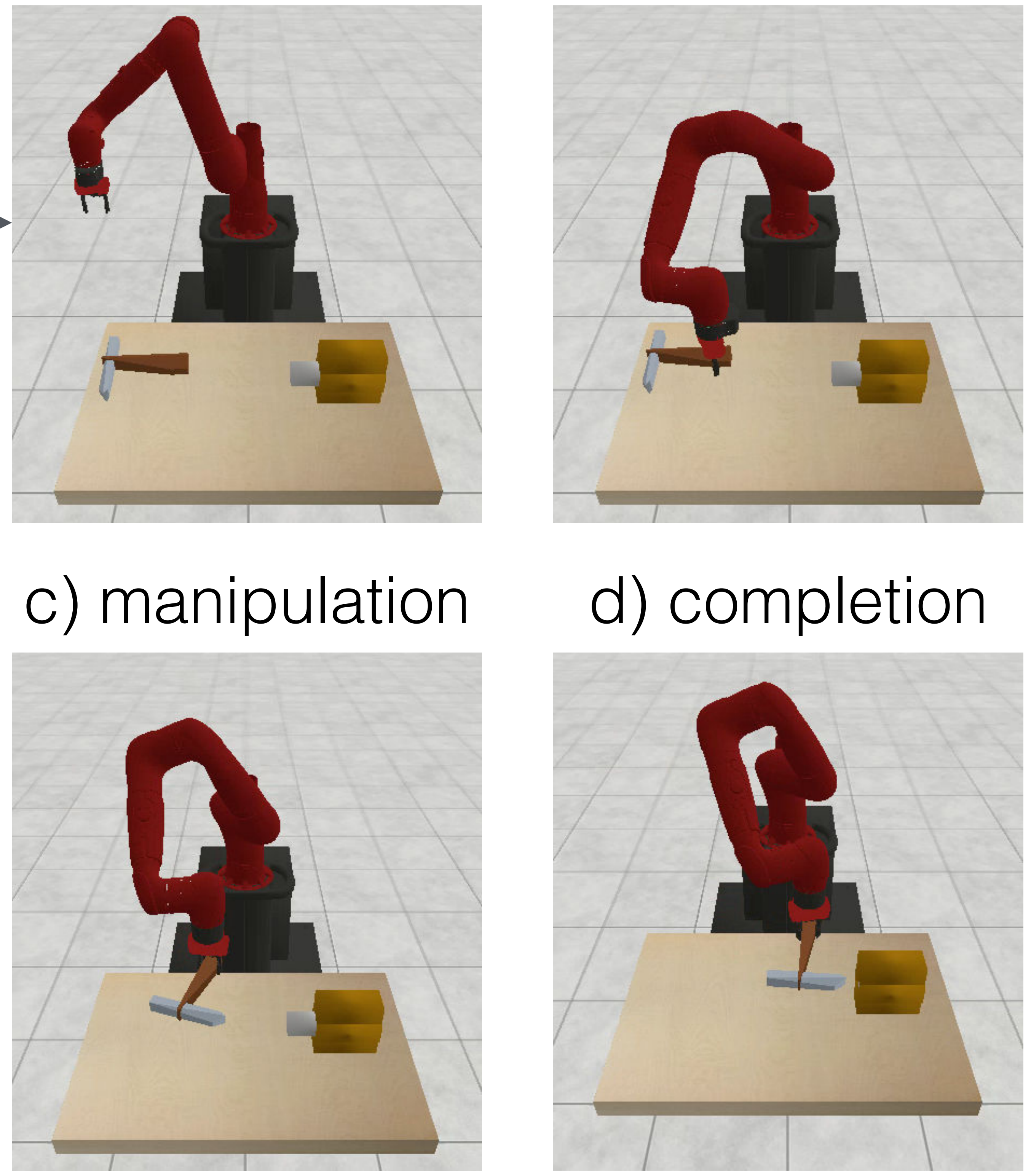
action execution

a) initial

b) grasping

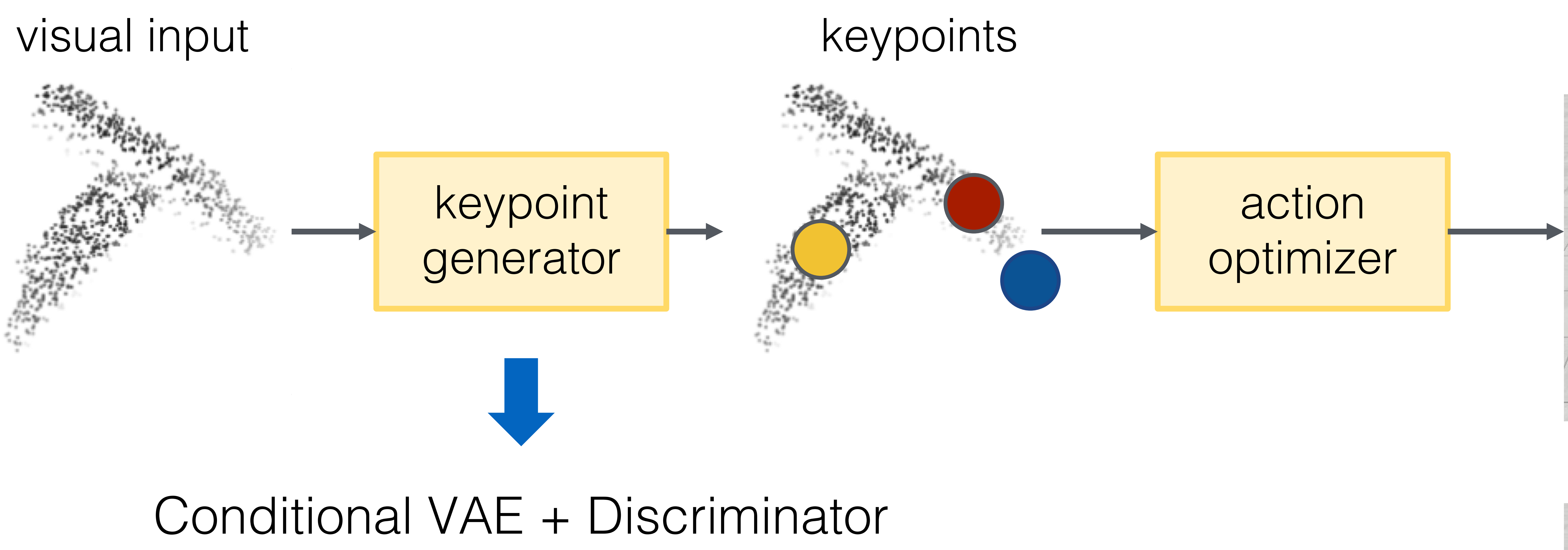
c) manipulation

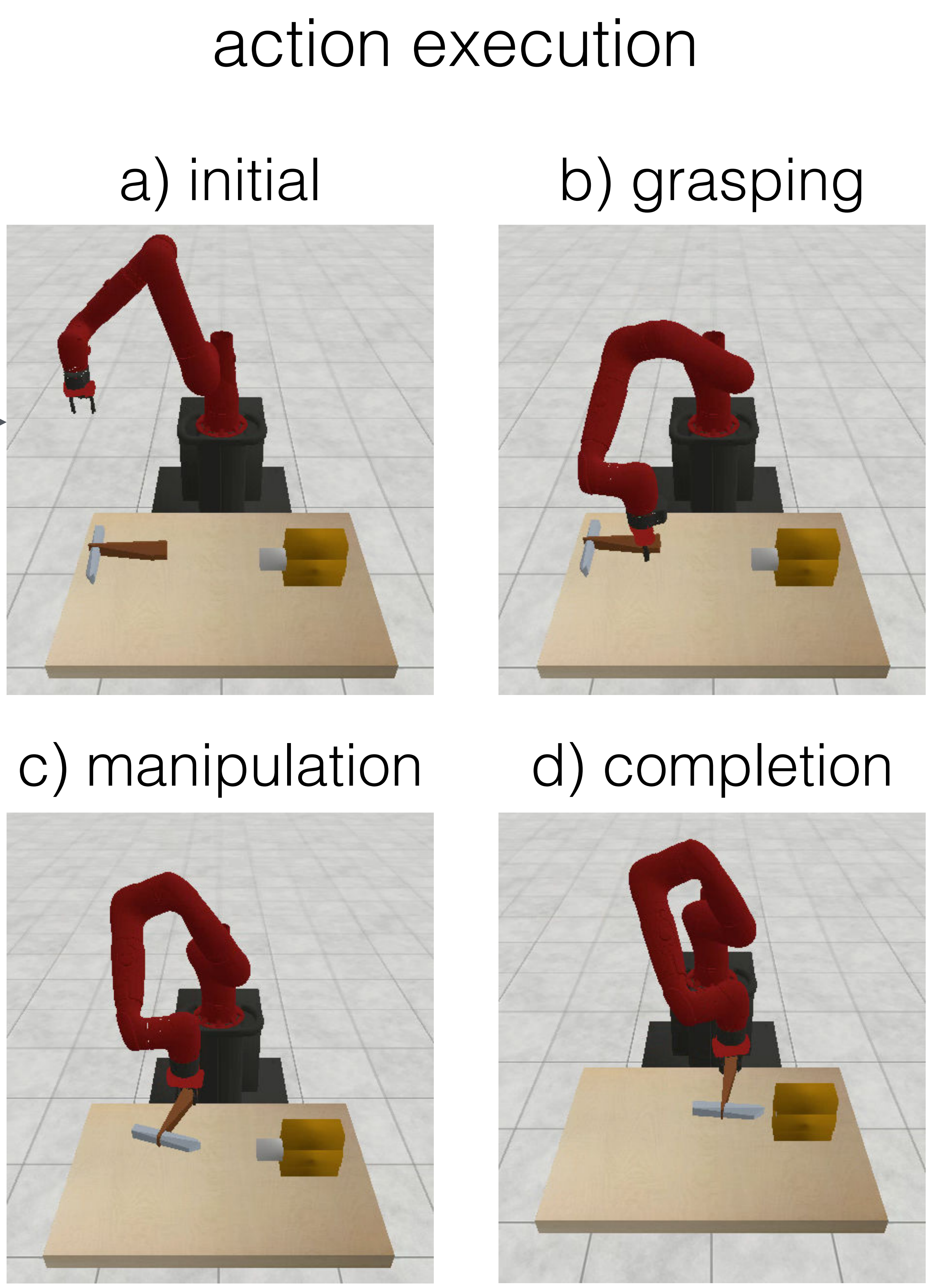
d) completion



Decoupling perception problem and control problem

KETO: Keypoint Representations for Tool Manipulation

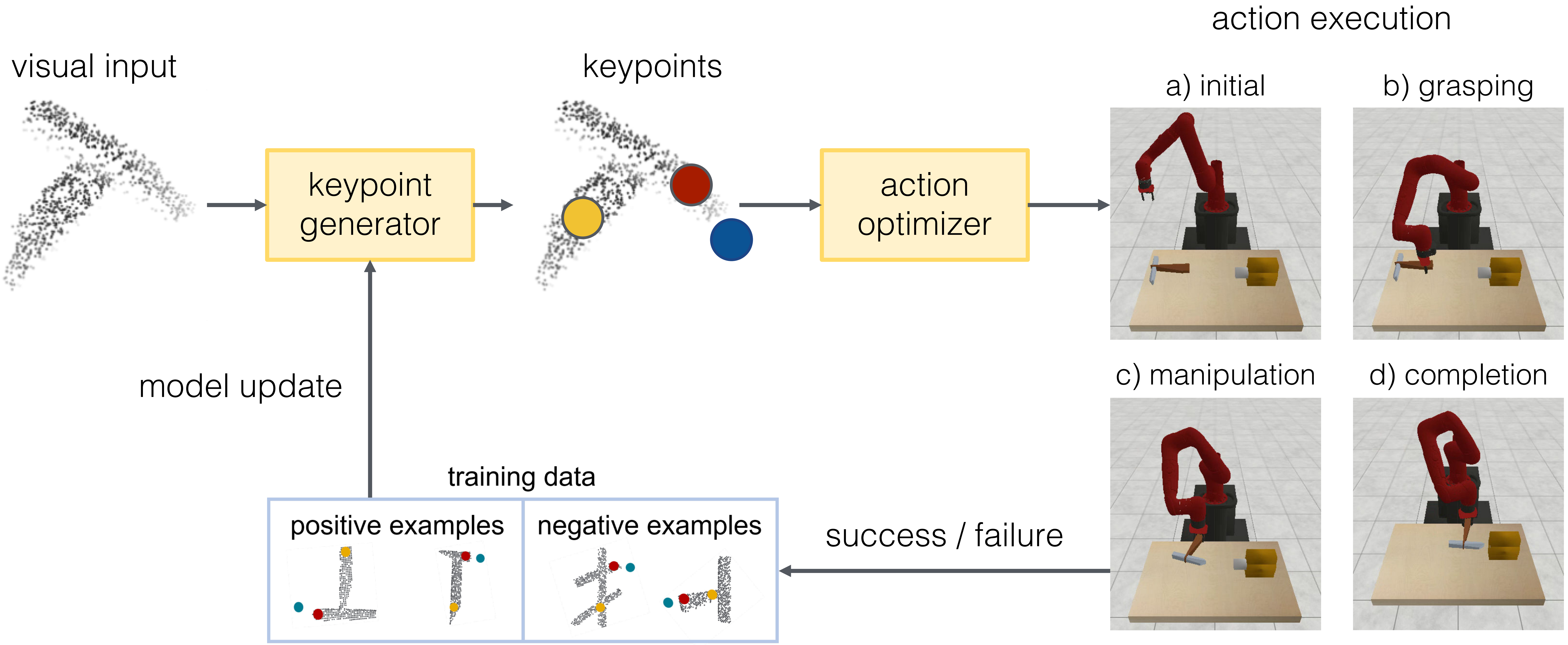




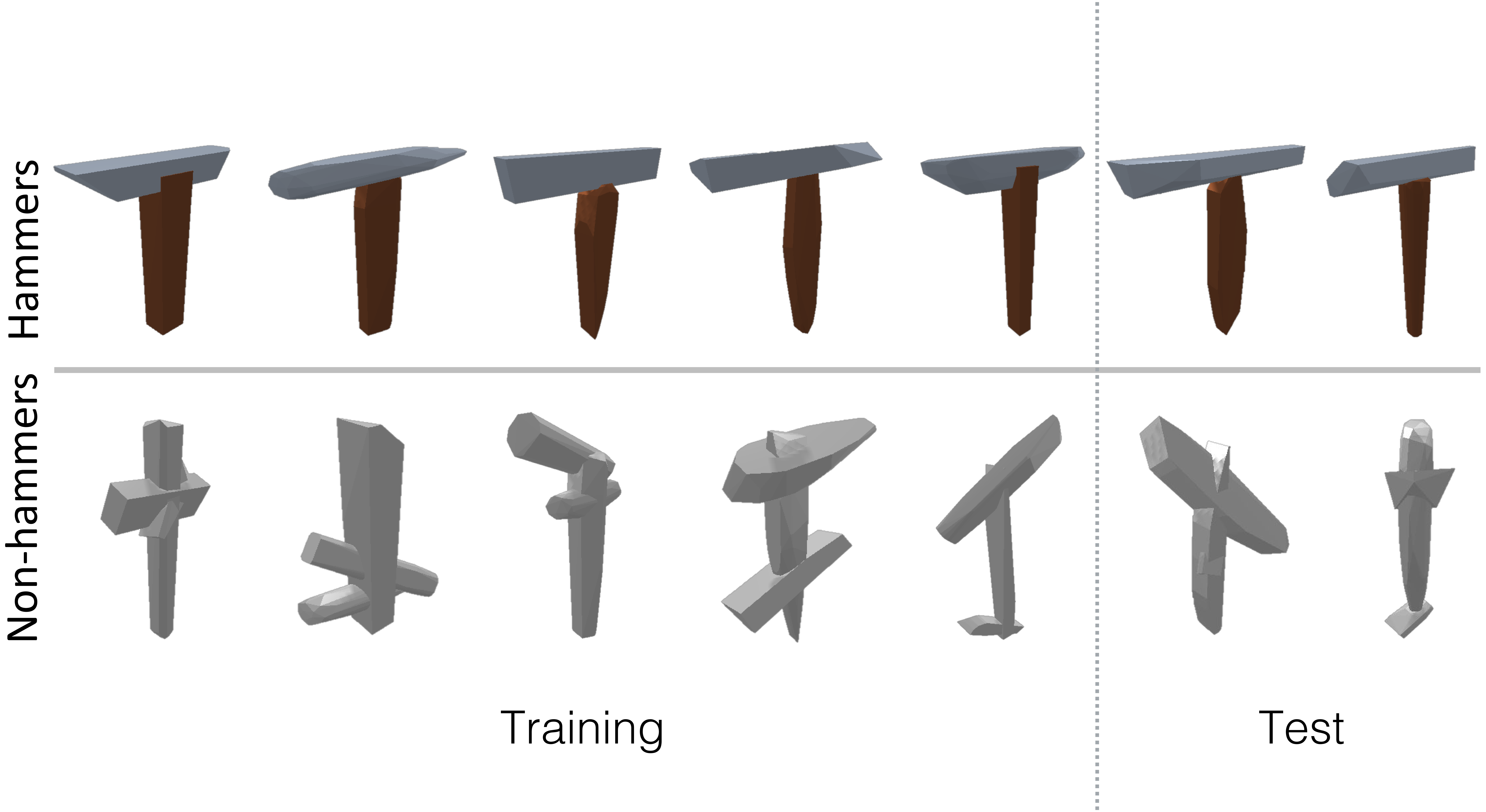
How do we get supervision?

Self-supervision from interaction

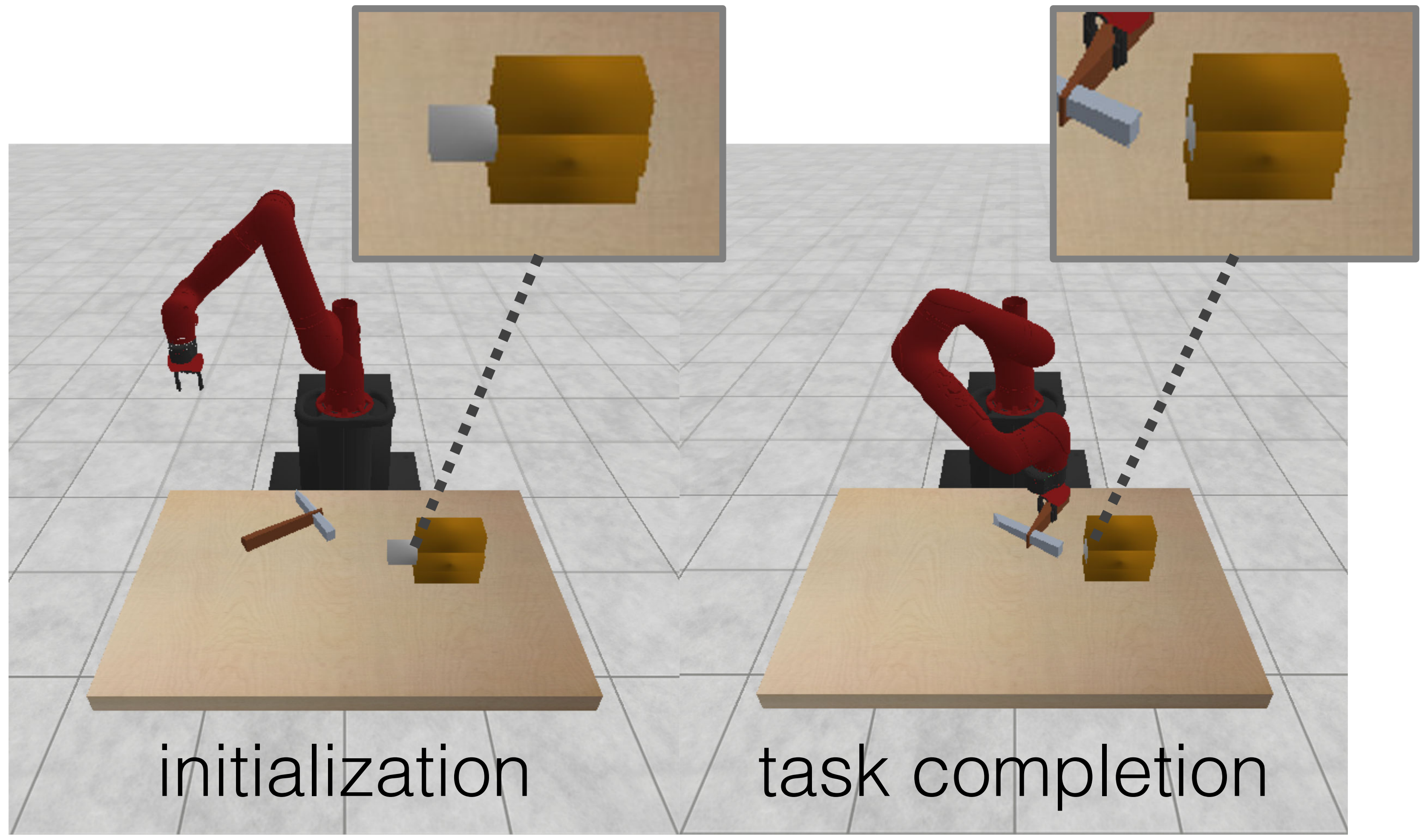
KETO: Keypoint Representations for Tool Manipulation



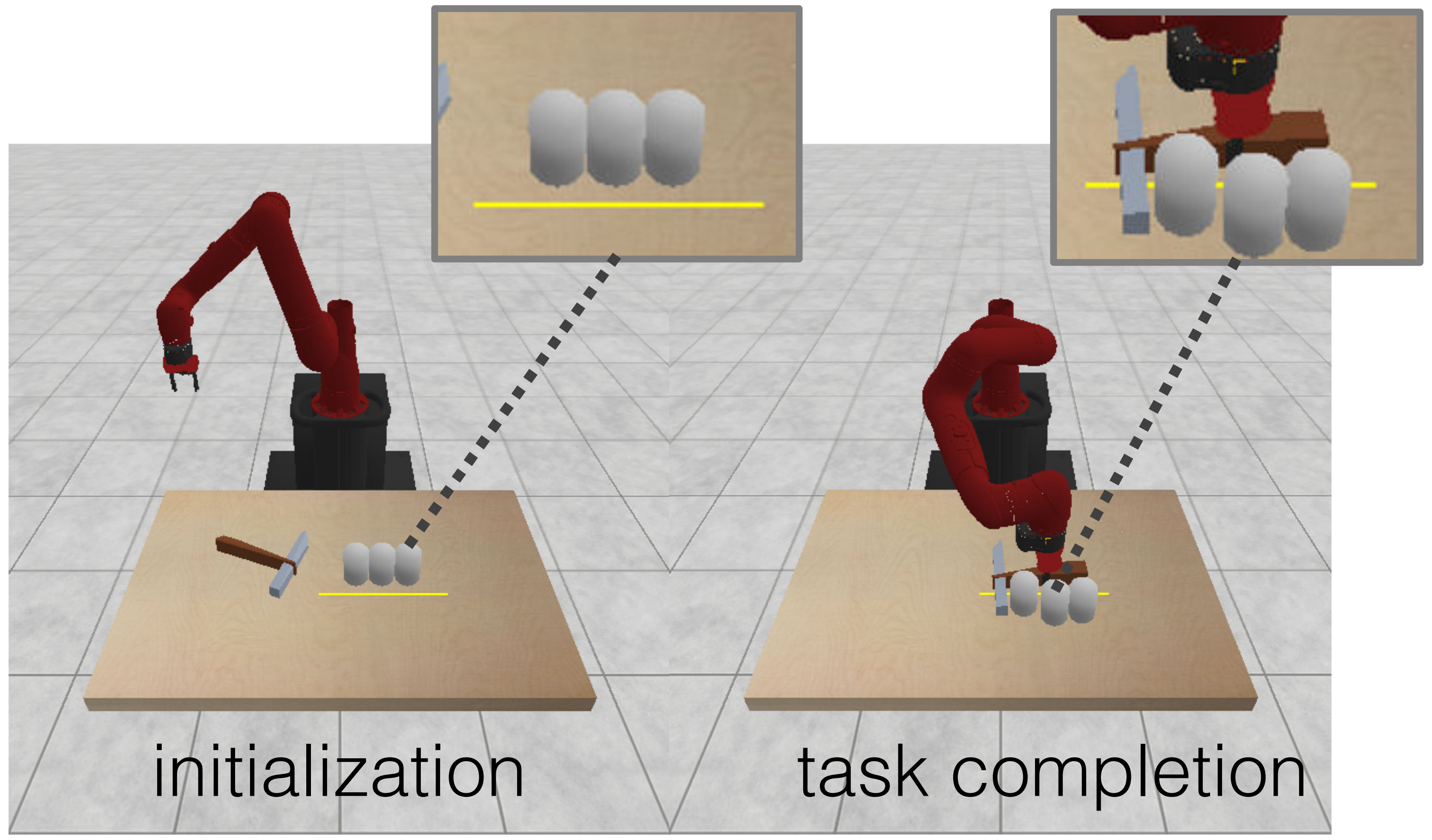
Procedural Generation of Tools



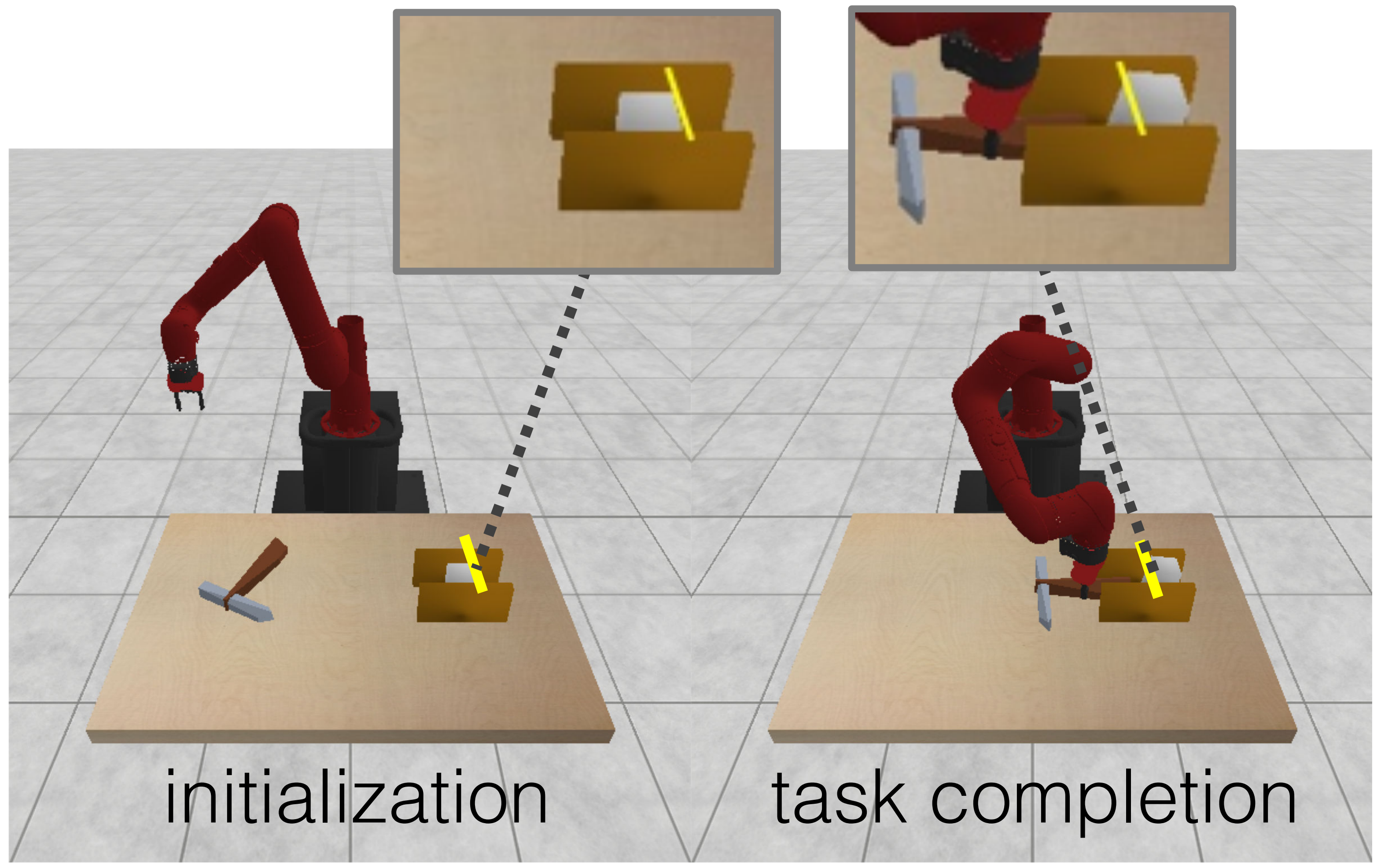
Tool Manipulation Tasks



(a) Hammering



(b) Pushing



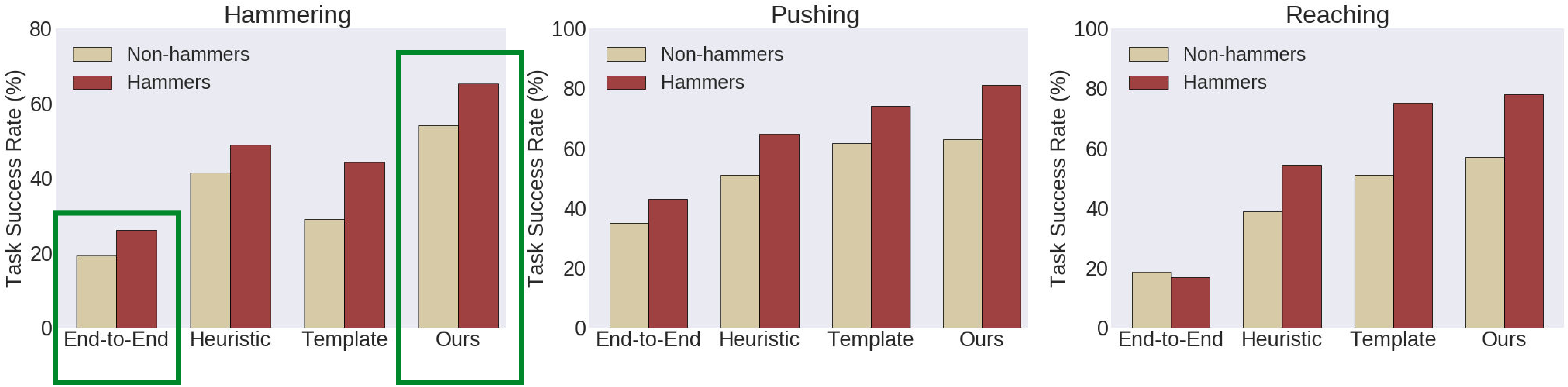
(c) Reaching

Results: Hammering Task

Results: Pushing Task

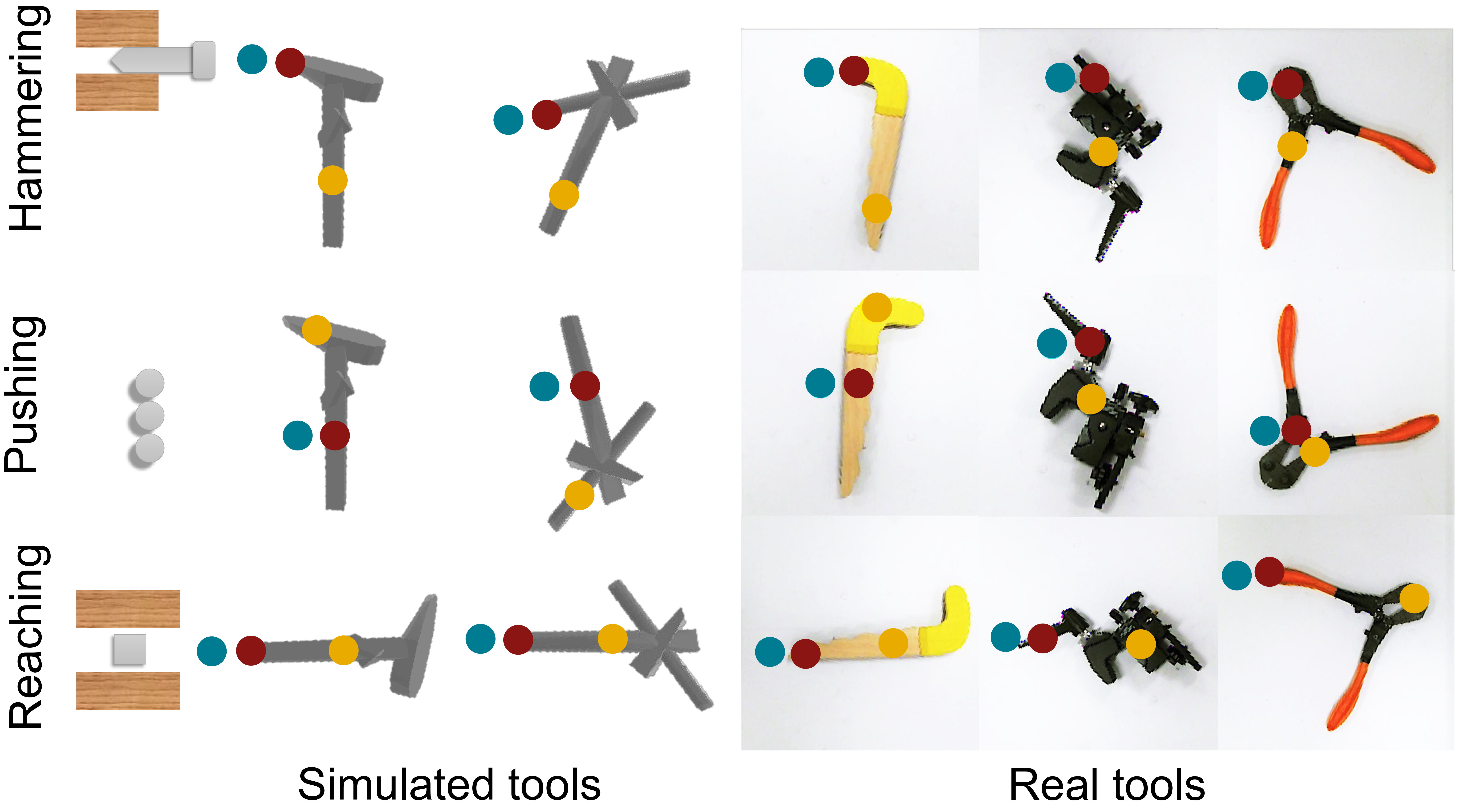
Results: Reaching Task

Results: Quantitative Evaluation

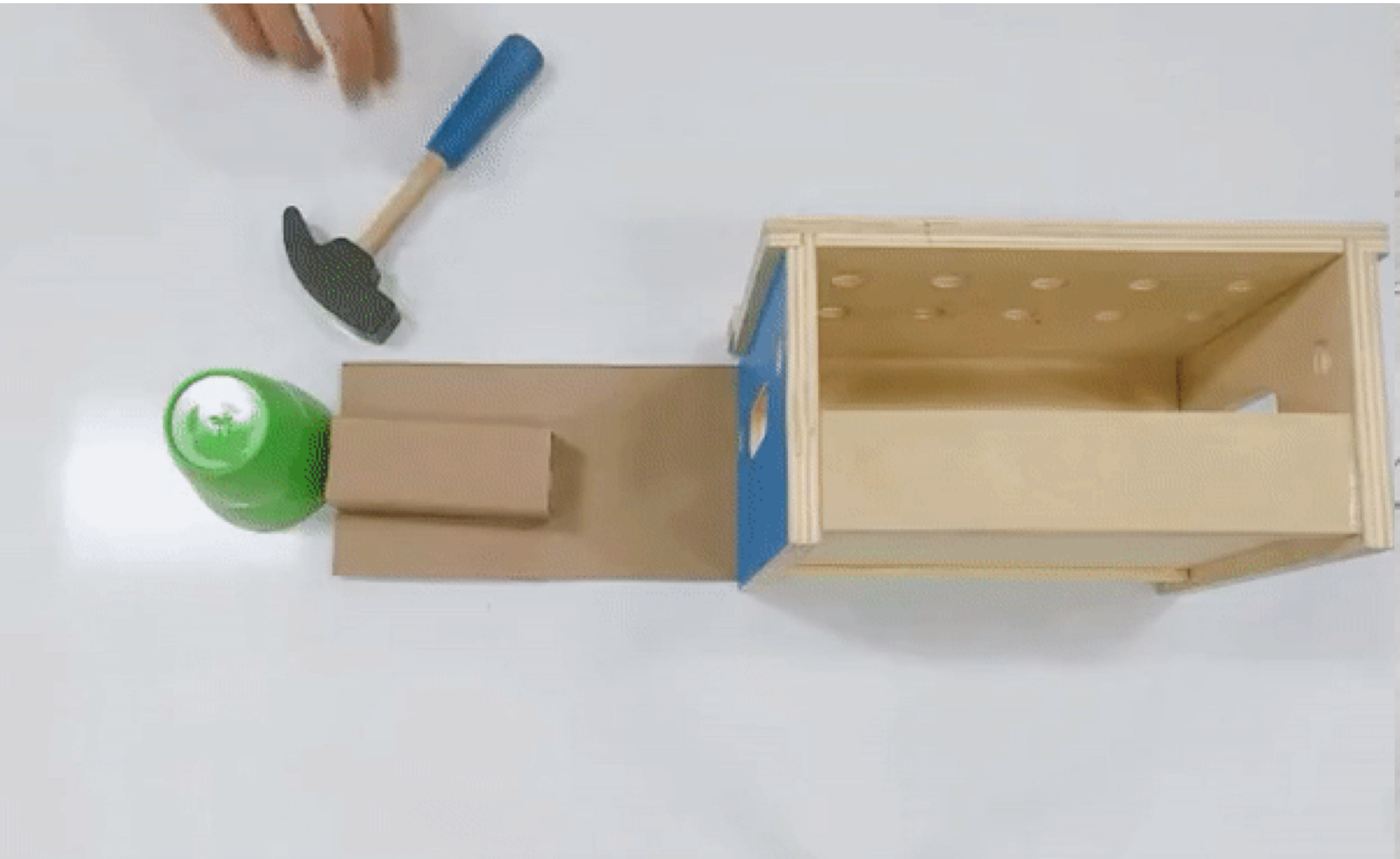


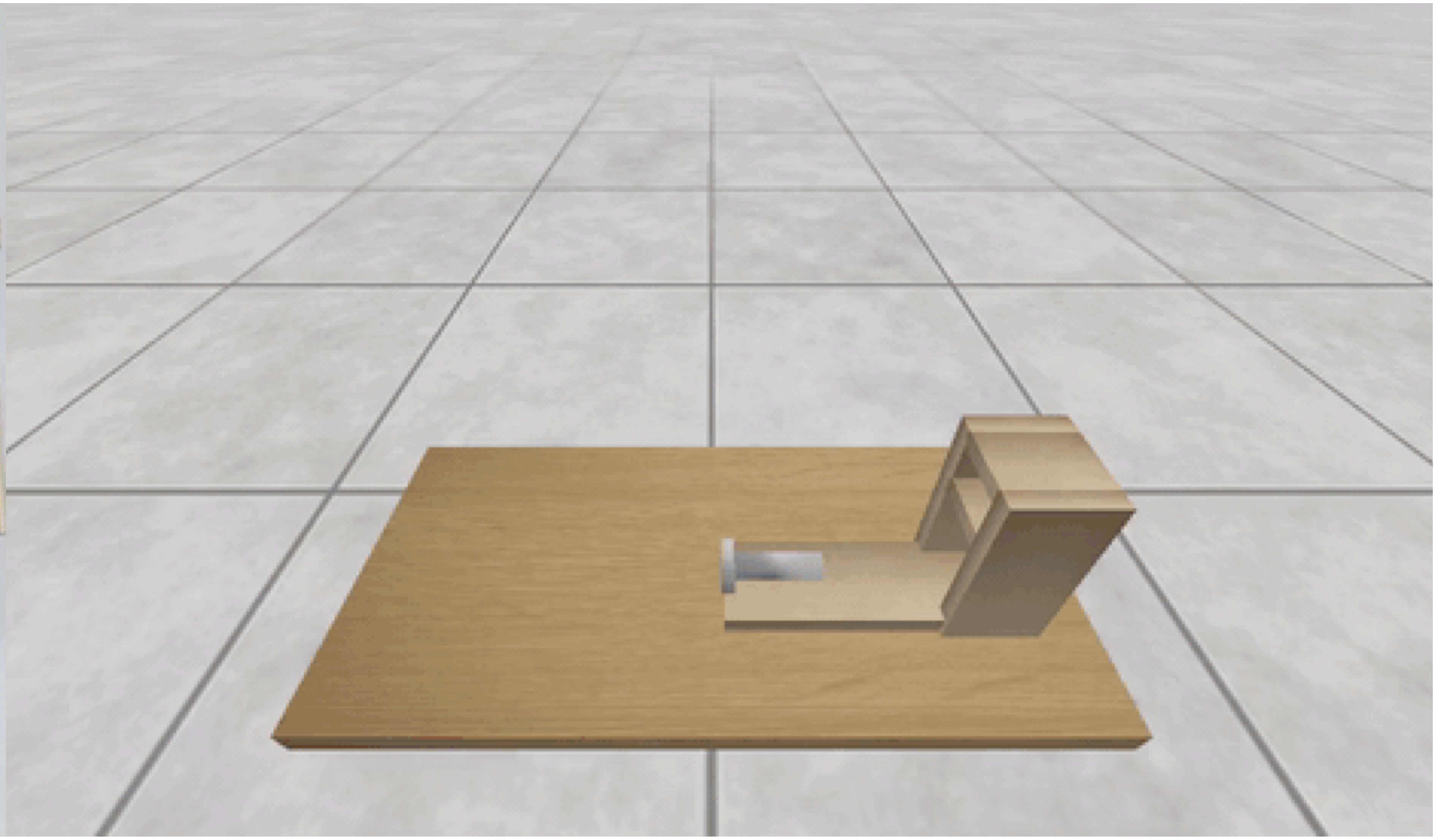
Keypoints as **intermediate representations** of tools are effective.

Results: Keypoint Prediction



Composite Task: Multi-stage Tool Use

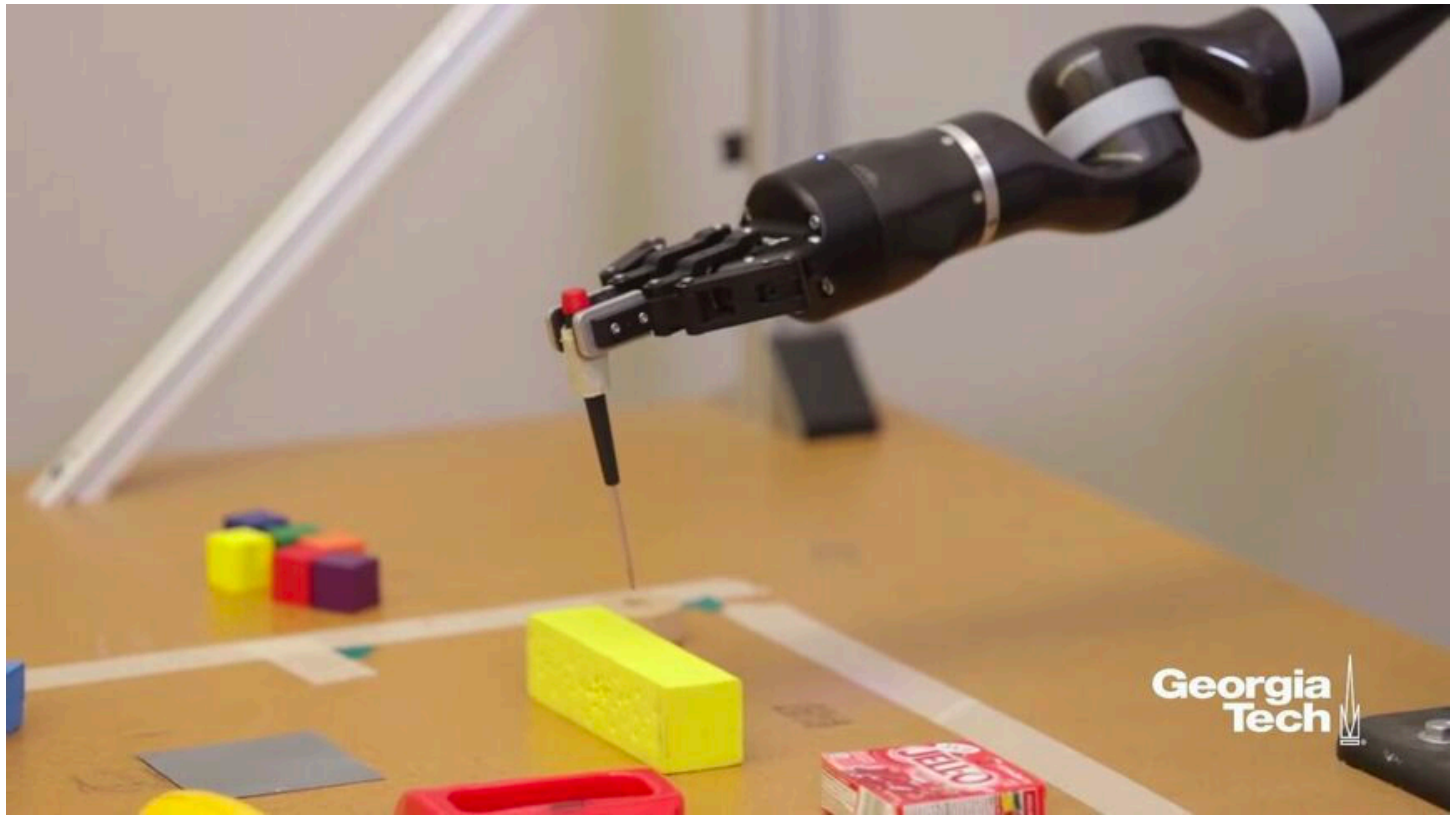




autonomous execution

Tool Creation: MacGyvering

Improvising tools for inventive problem solving



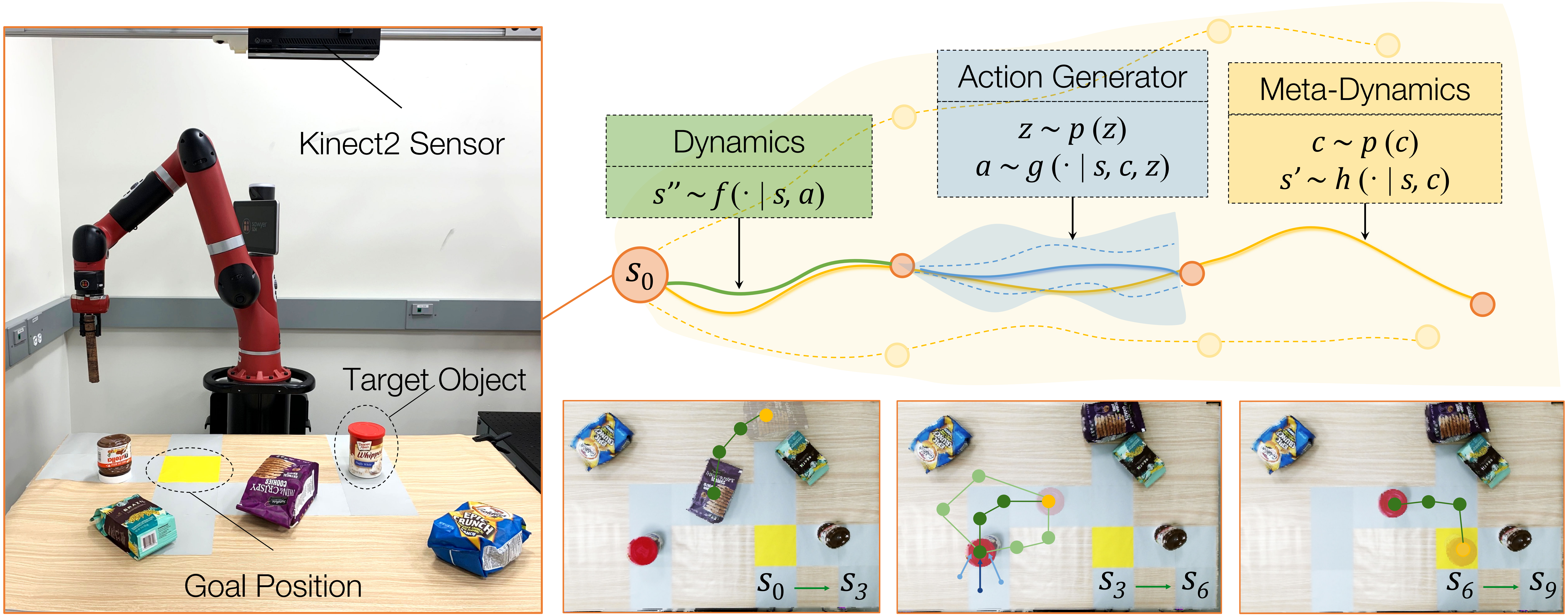
[Nair, Shrivatsav, Erickson, Chernova RSS'19]

Tool Creation: MacGyvering

Keypoints provides a scaffold for generating tools from object parts.

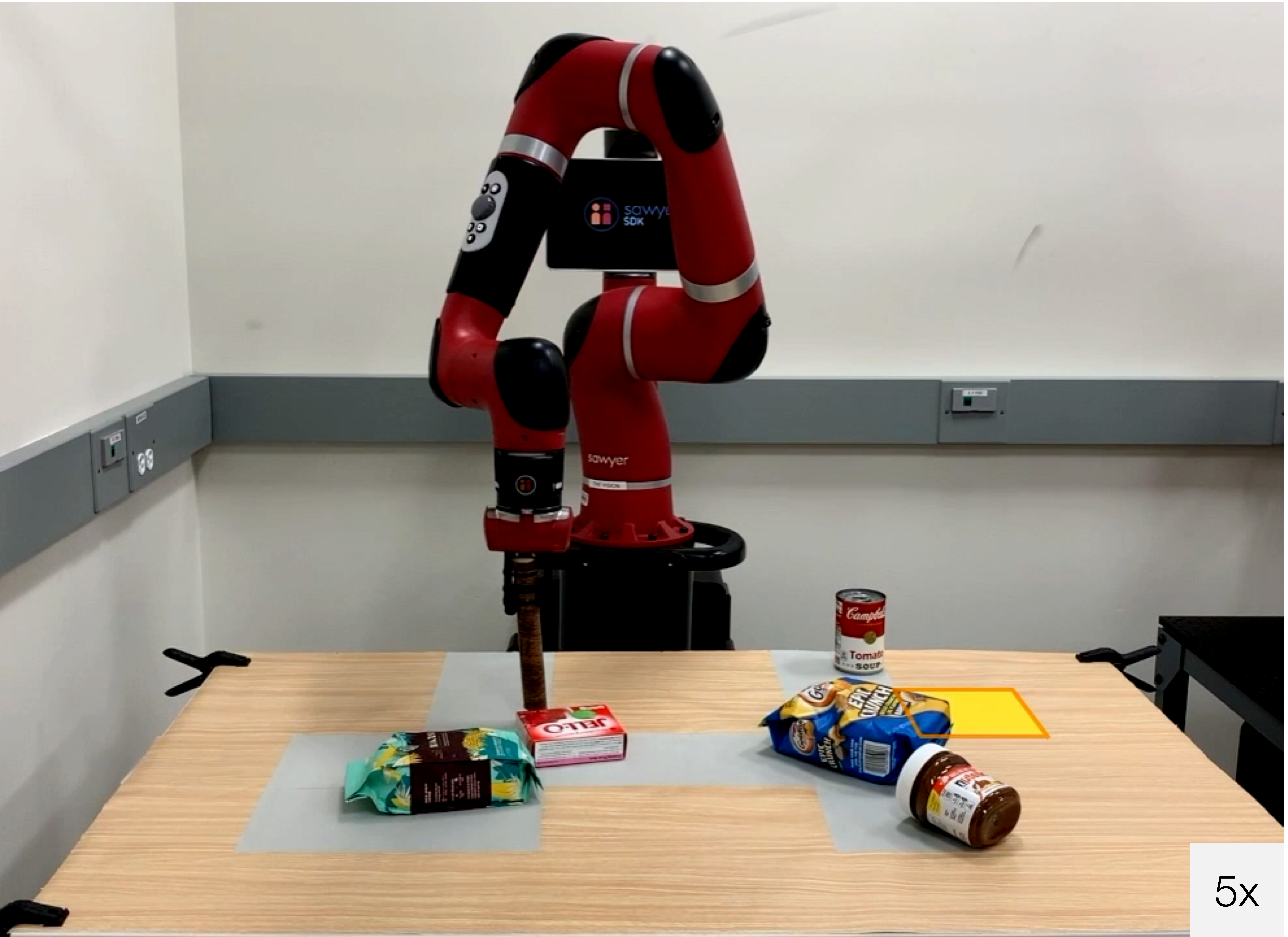
Hammering with the Created Tool

Hierarchical Planning with Cascade Variational Inference



Fang, Zhu, Garg, Savarese, Fei-Fei. CoRL'19

Hierarchical Planning with Cascade Variational Inference



Move away obstacles

Push target object to the goal

Fang, Zhu, Garg, Savarese, Fei-Fei. CoRL’19

Conclusions

3D Keypoints are compact and effective object representations for manipulation.

1. **6-PACK**: Keypoints for Category-level 6D Pose Tracking
2. **KETO**: Keypoints for Vision-based Tool Manipulation

Supervision can be acquired through object motion and robot interaction.

Open Question

How to integrate keypoints with other representations to incorporate fine-grained semantic, geometric, and physical information of objects and environments.


Questions?