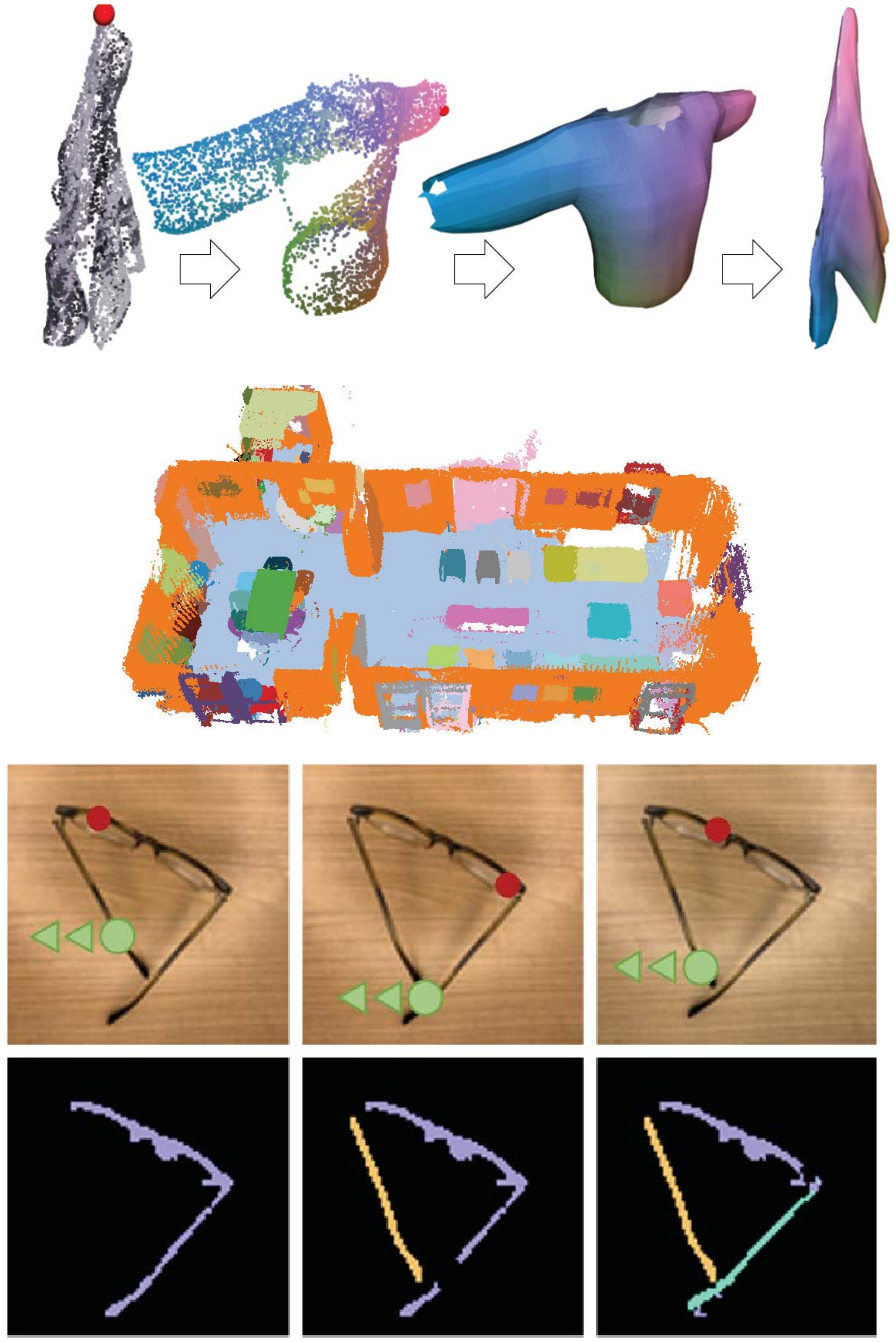
Active Scene Understanding with Robot Interactions

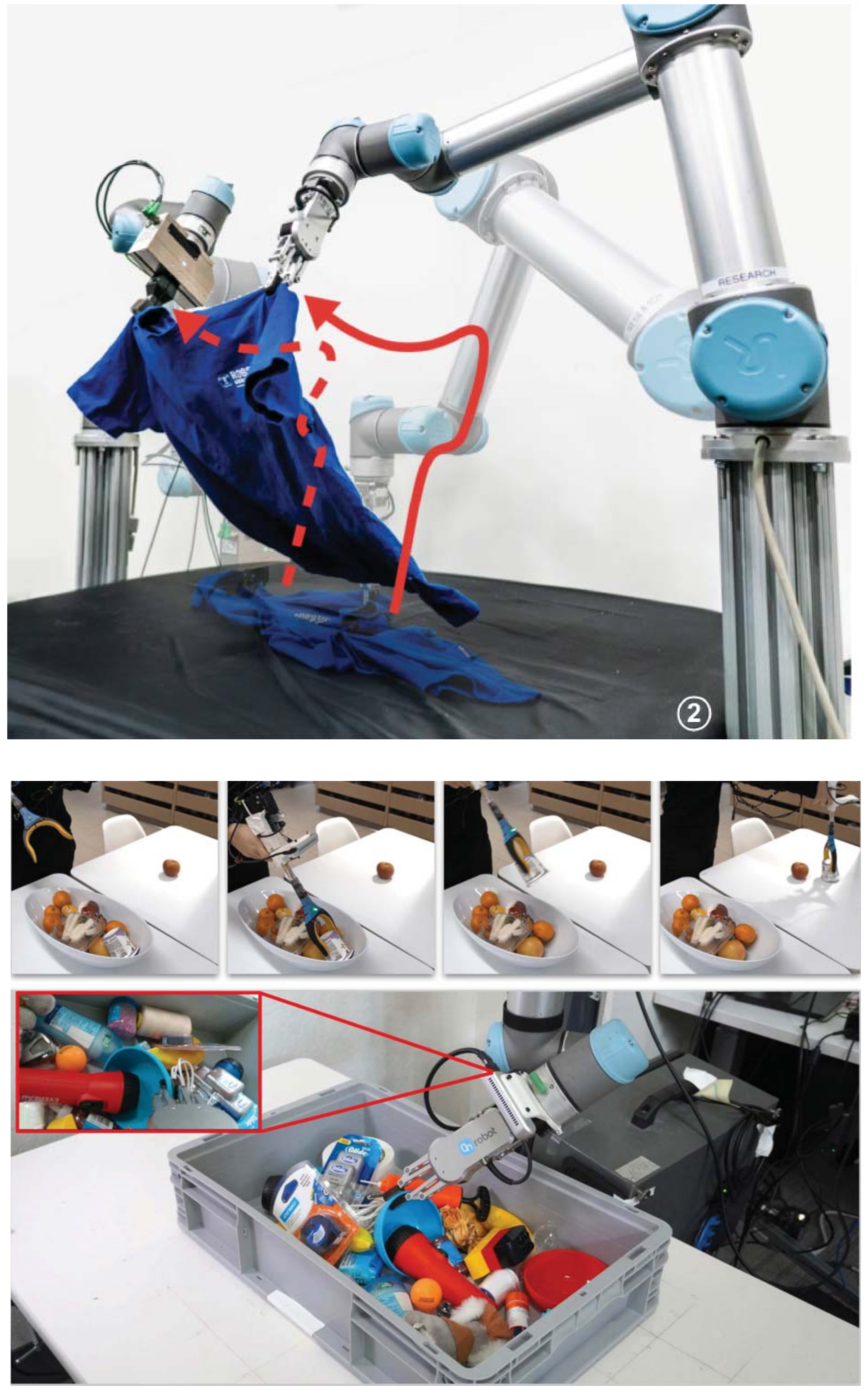
Shuran Song

Columbia University
Artificial Intelligence & Robotics Lab

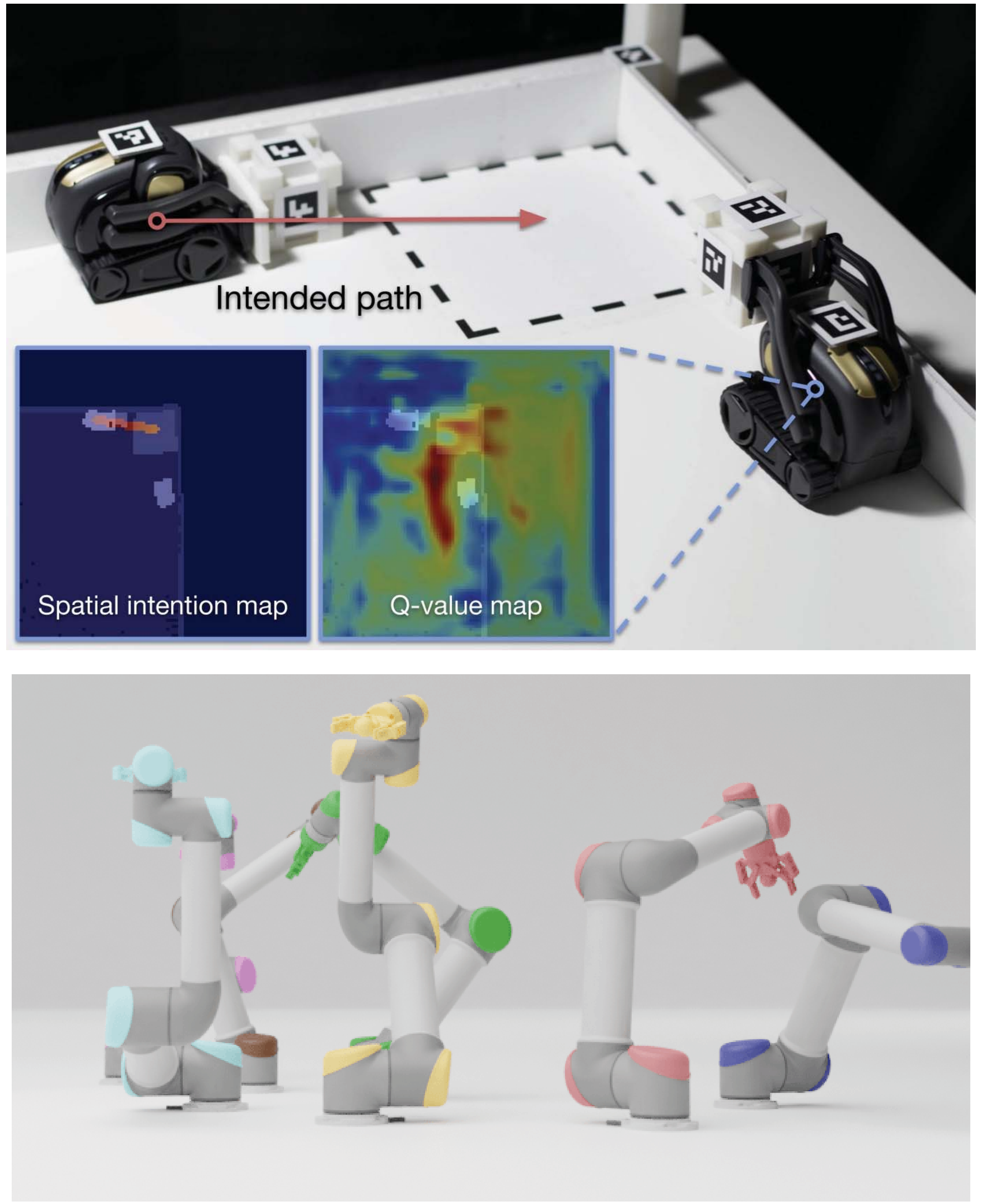
See, Understand, Act



Perception

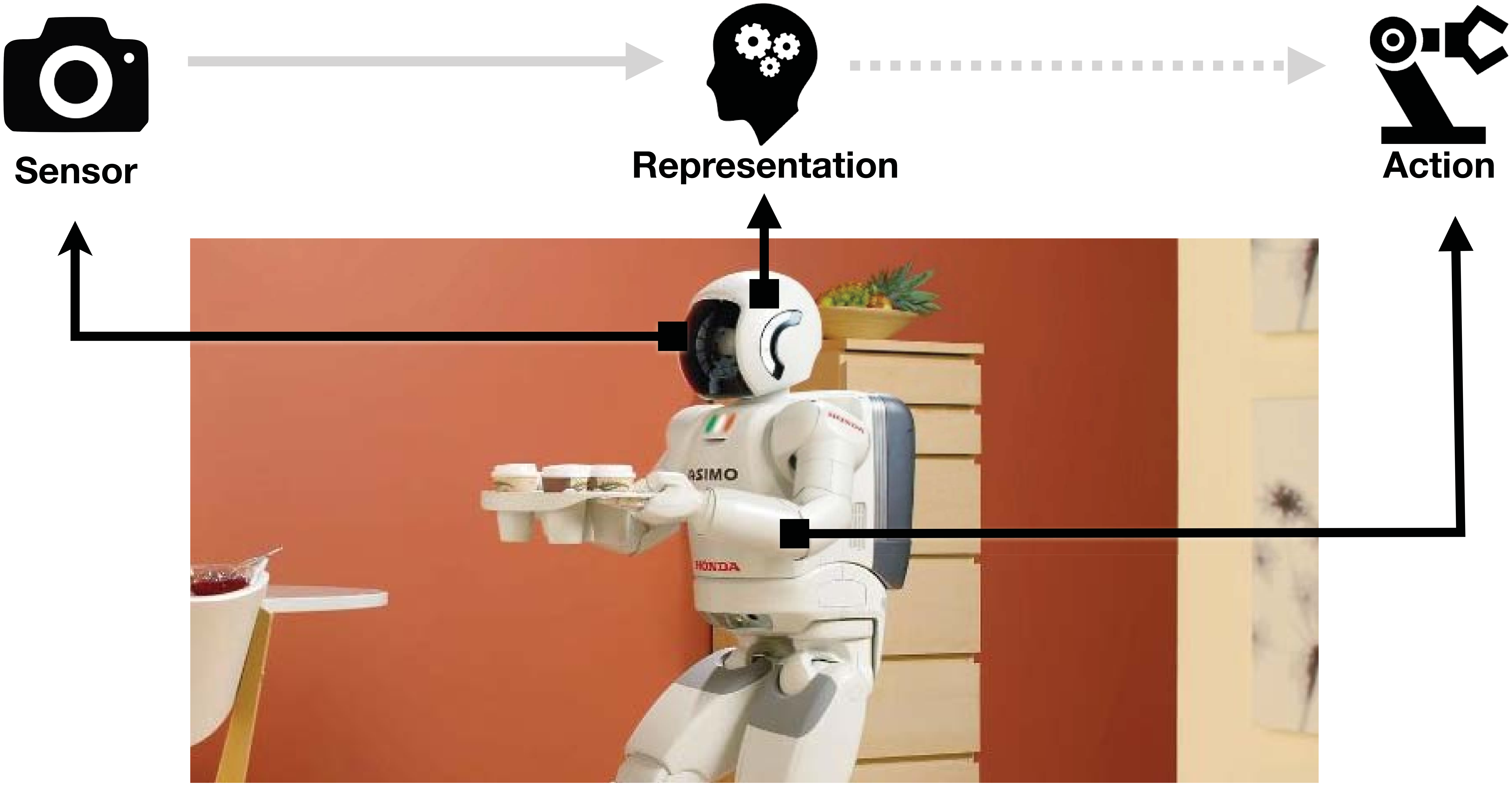


Manipulation



Collaboration

See, Understand, Act

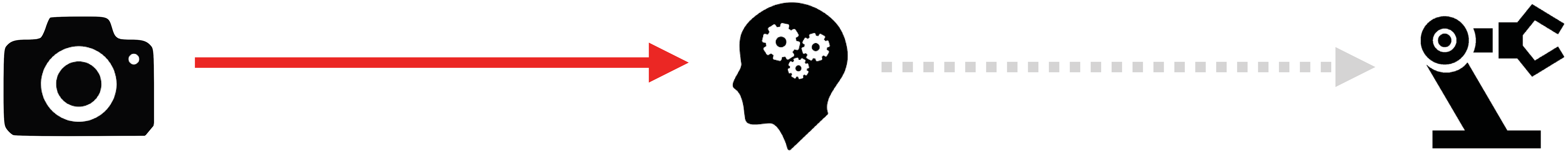


Sensor

Representation

Action

Scene Understanding

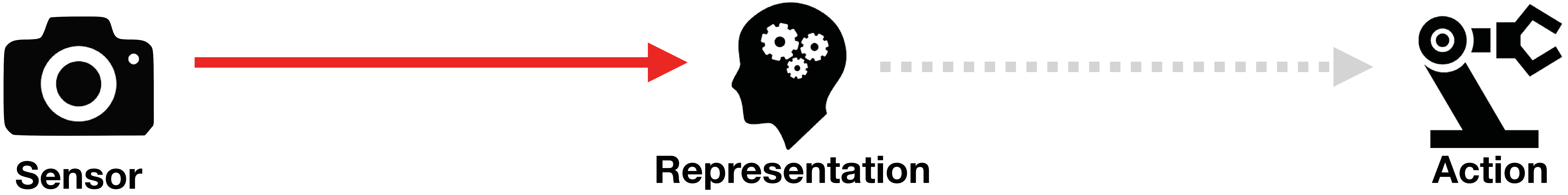


Sensor

Representation

Action

Scene Understanding



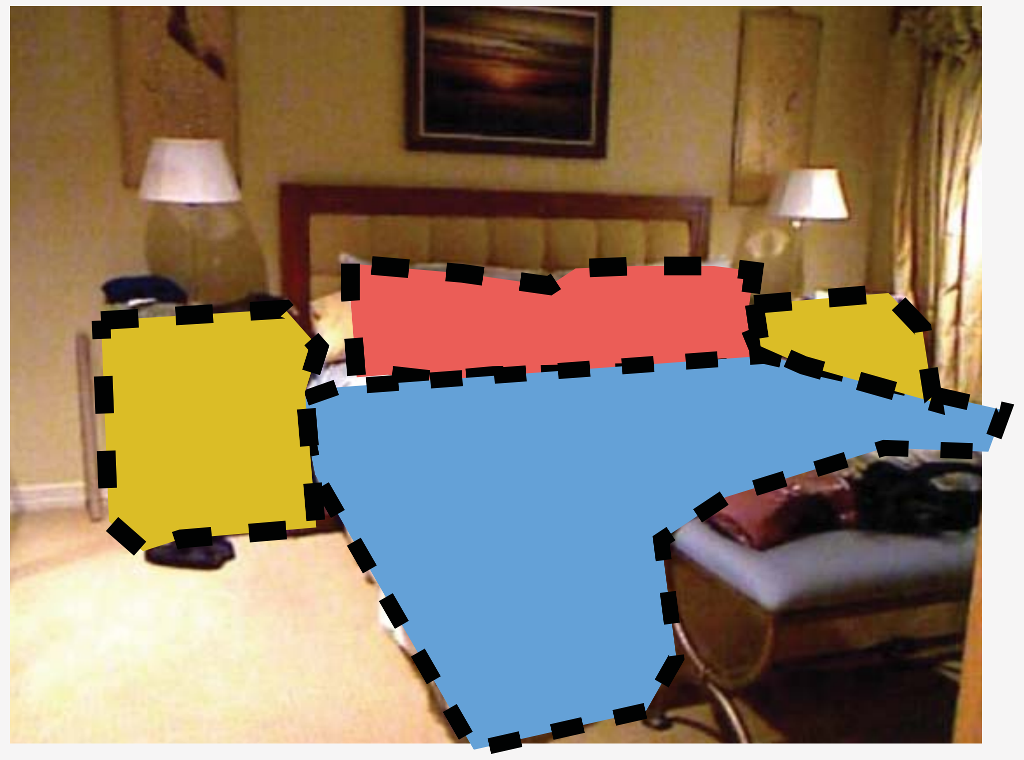
Sensor

Representation

Action

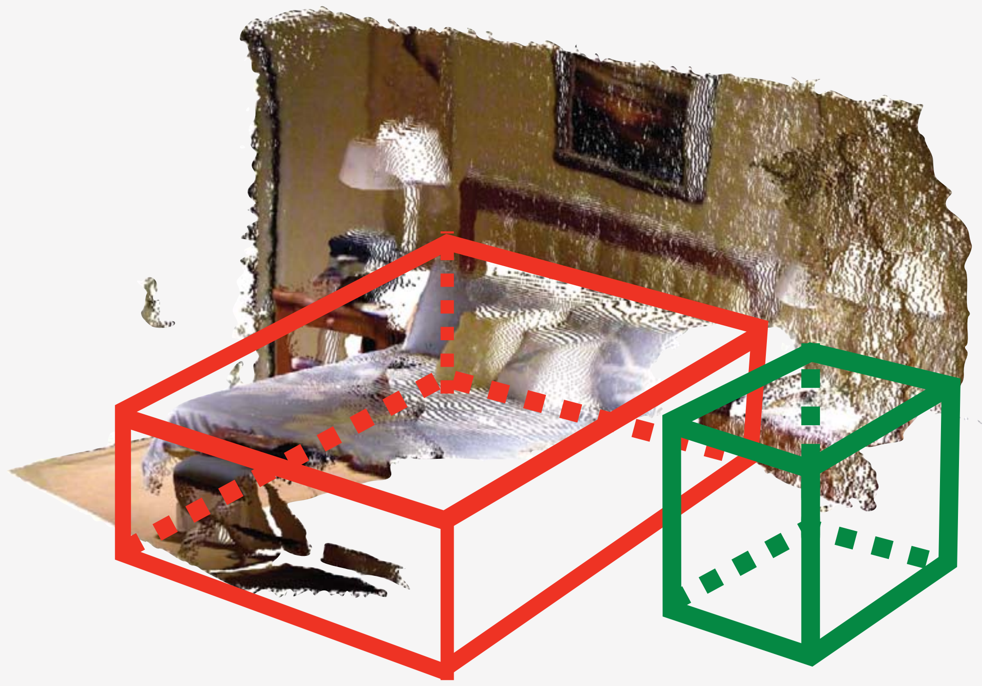
Scene Representations

Segmentation



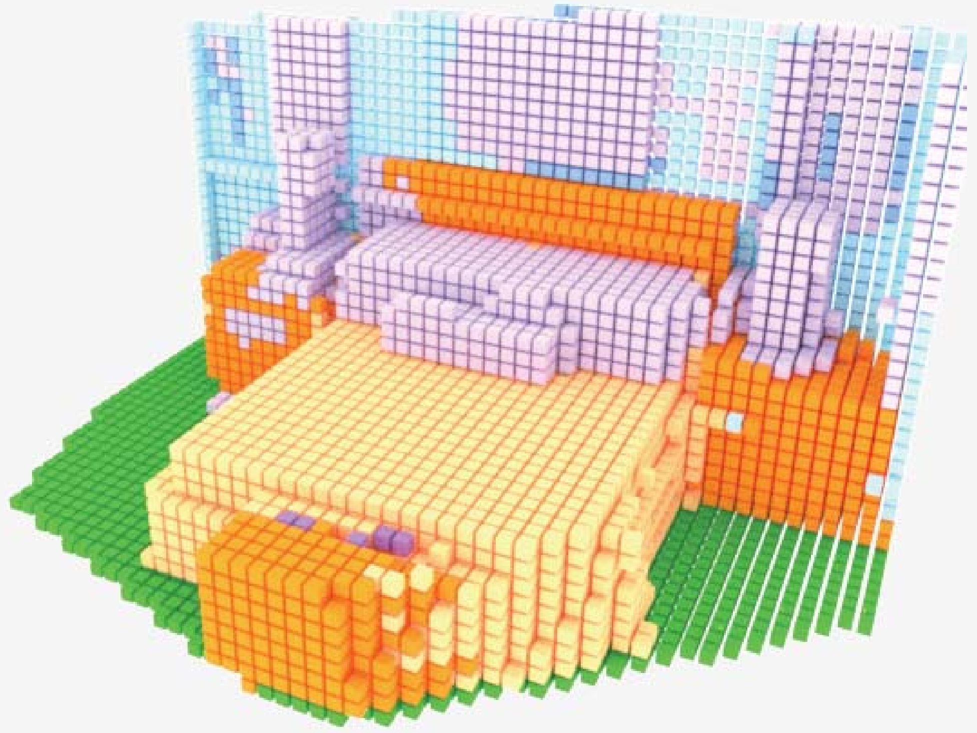
SUNRGB-D
CVPR’15

3D Object



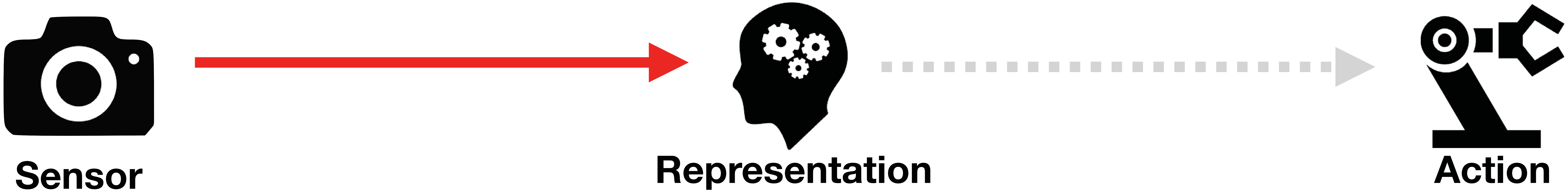
SlidingShapes
ECCV’14,CVPR’16

Semantic Scene Completion



SSCNet
CVPR'17

Scene Understanding

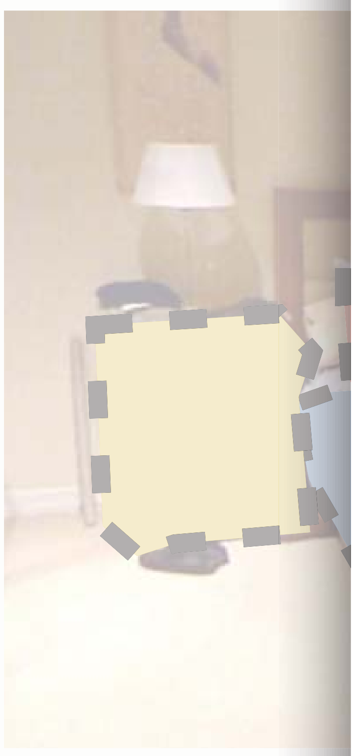


Sensor

Representation

Action

Segm



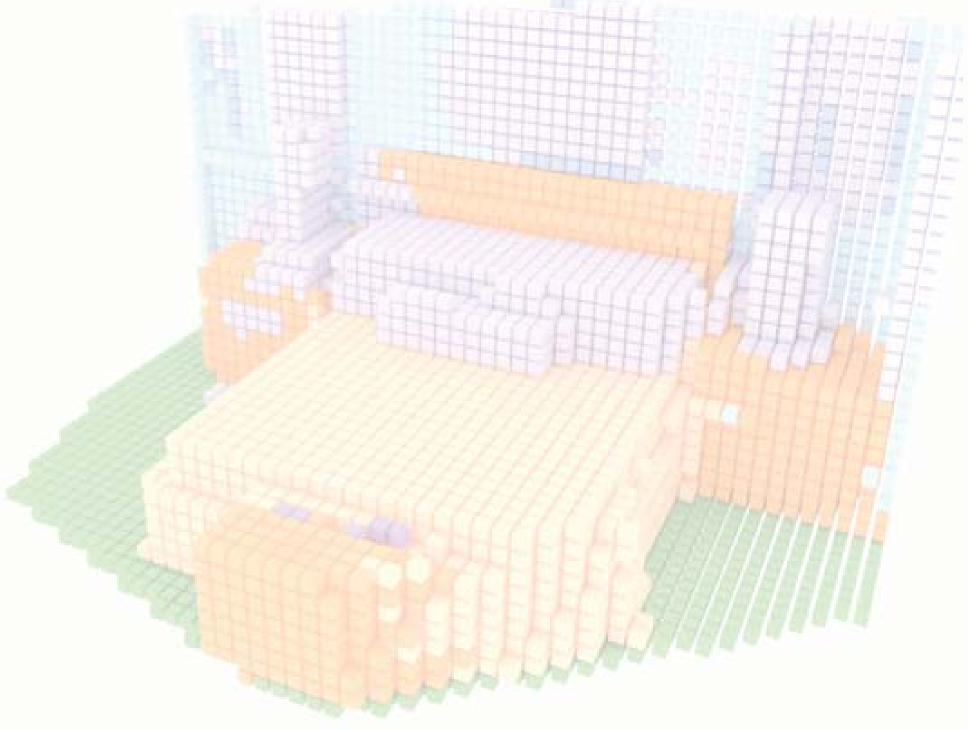
S

Passive Observers

s

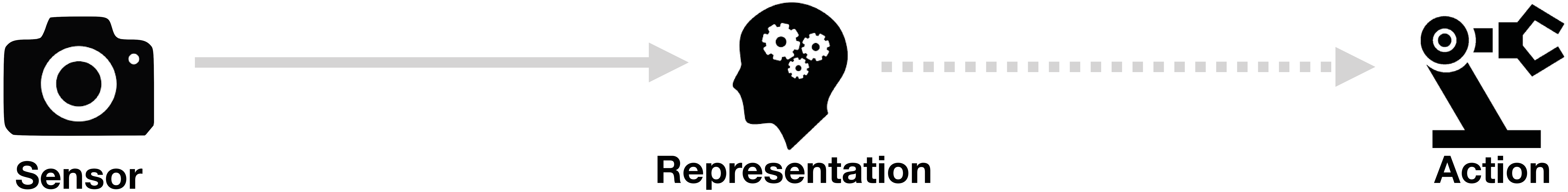
'16

Semantic Scene Completion

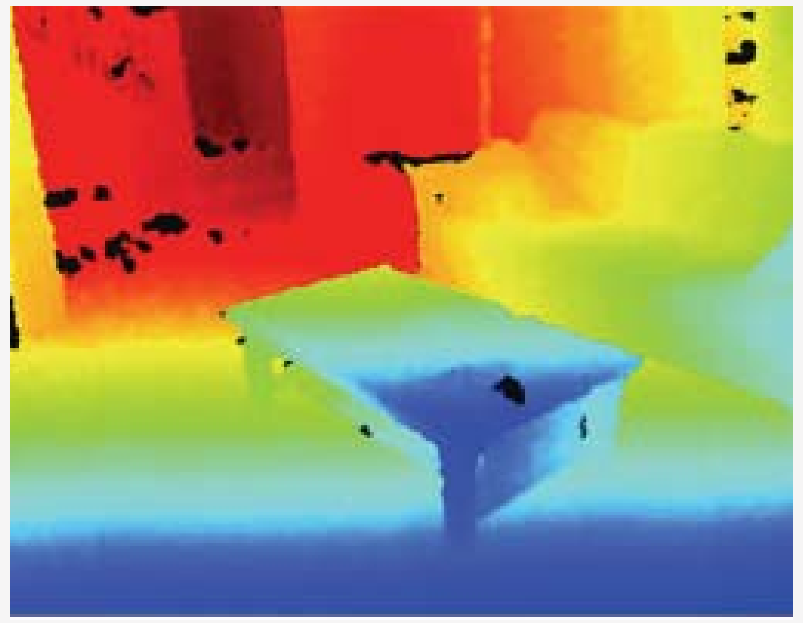


SSCNet
CVPR'17

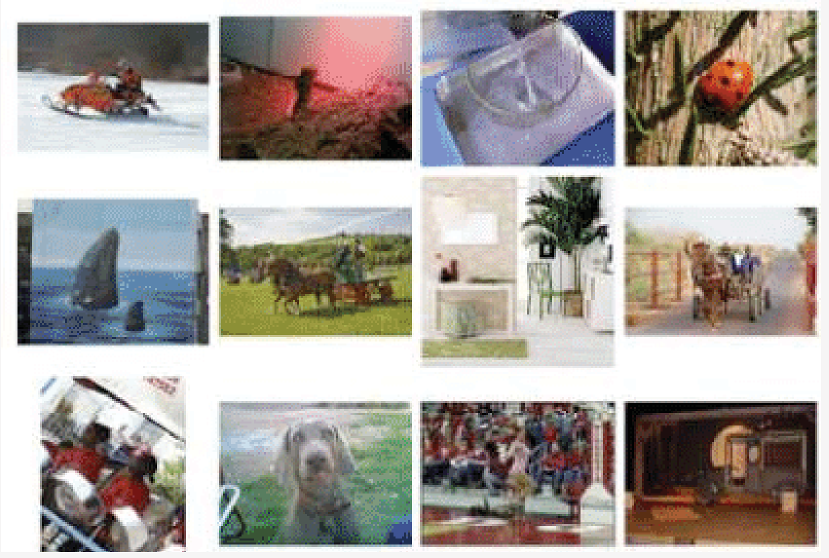
Scene Understanding



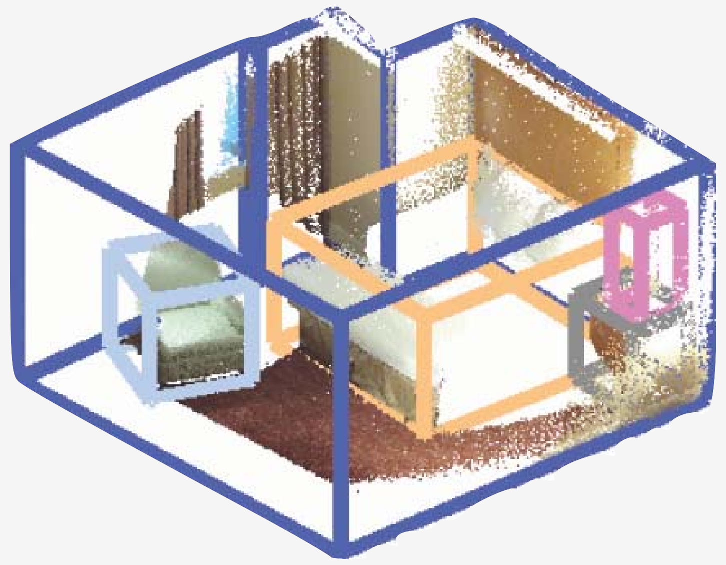
PASCAL VOC



NYU depth



ImageNet

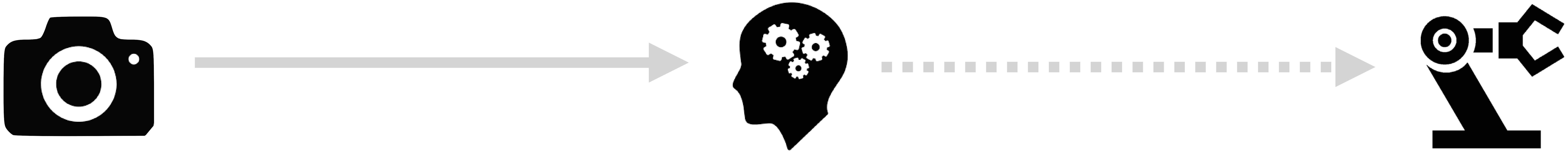


SUN RGB-D

Computer Vision Benchmarks

- Static images

Scene Understanding

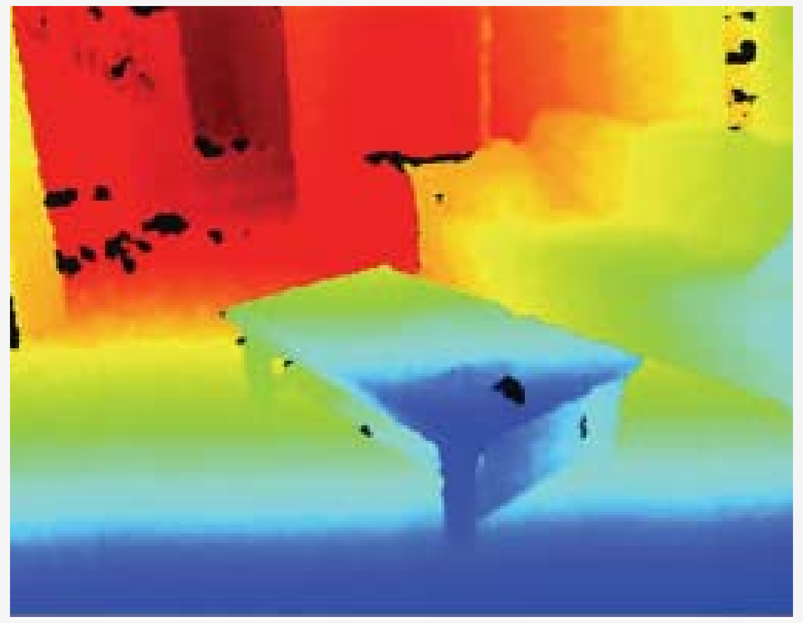


Sensor

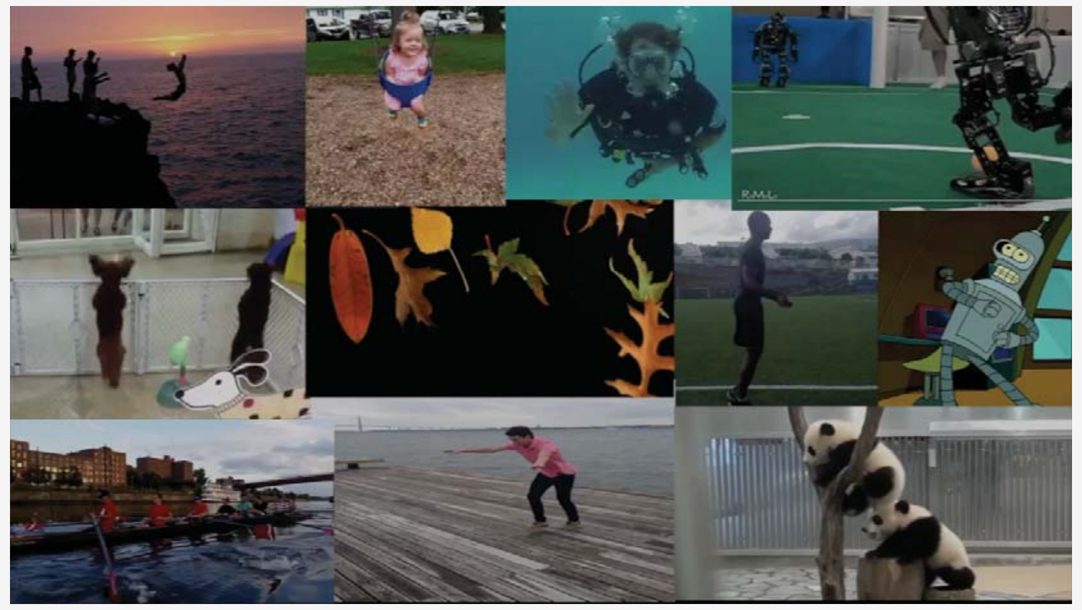
Representation

Action

PASCAL VOC

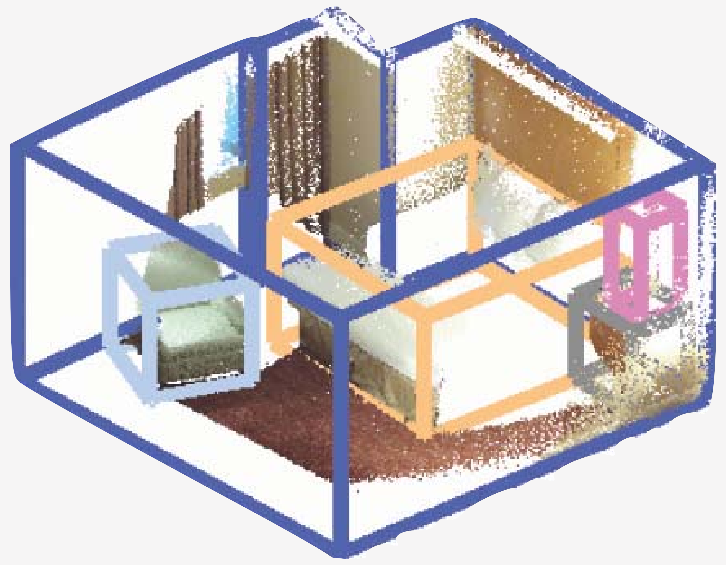


NYU depth

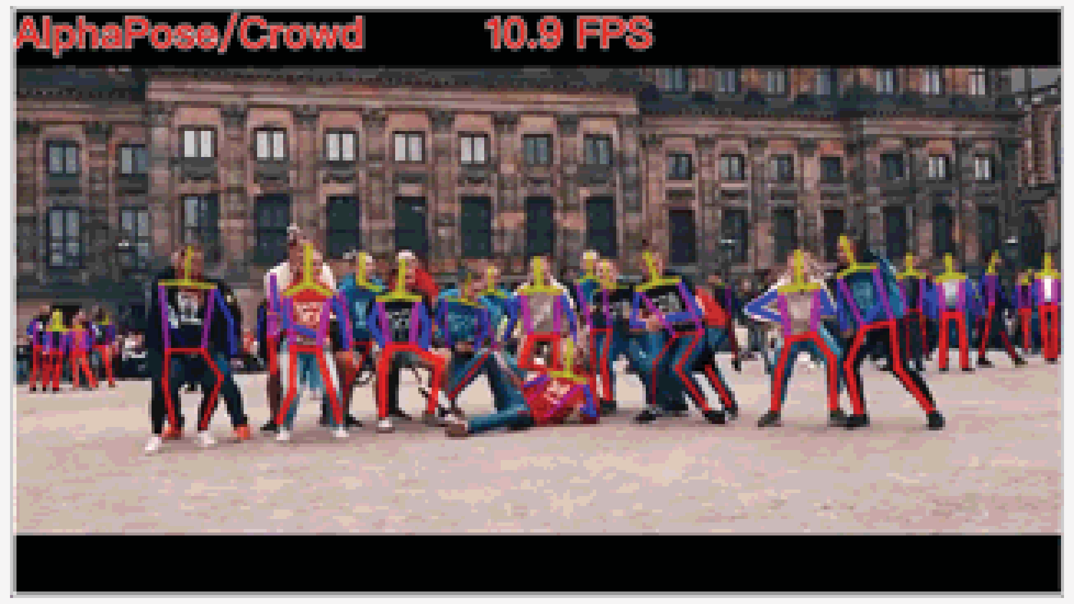


Moment in Time

ImageNet



SUN RGB-D

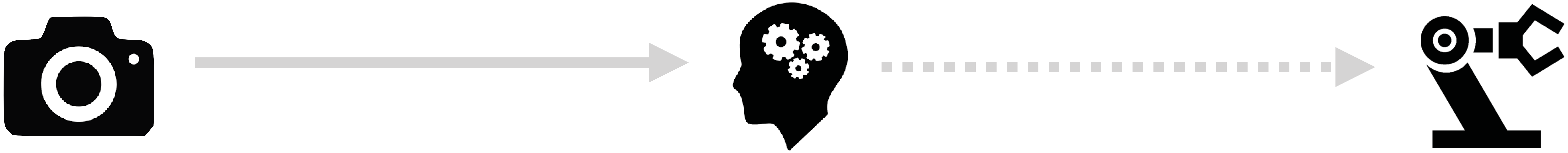


CrowdPose

Computer Vision Benchmarks

- Static images
- Passive video

Scene Understanding

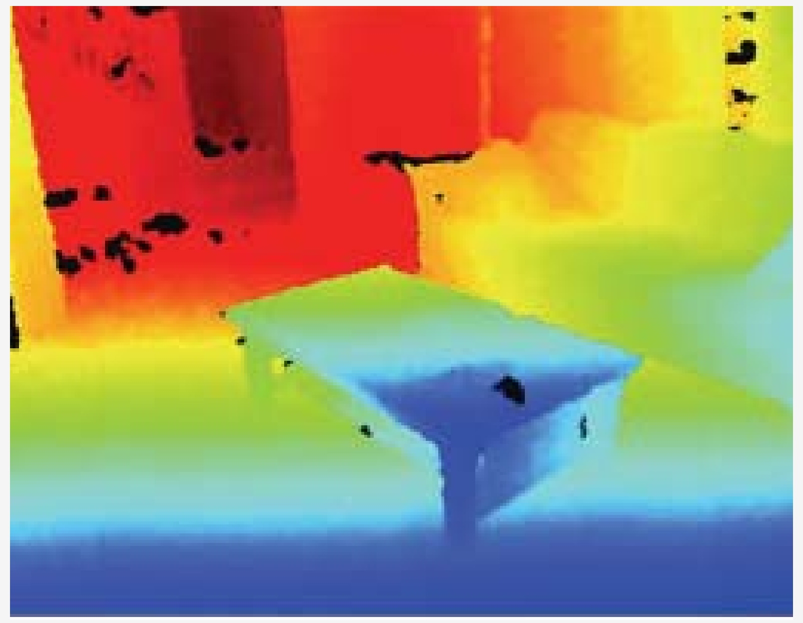


Sensor

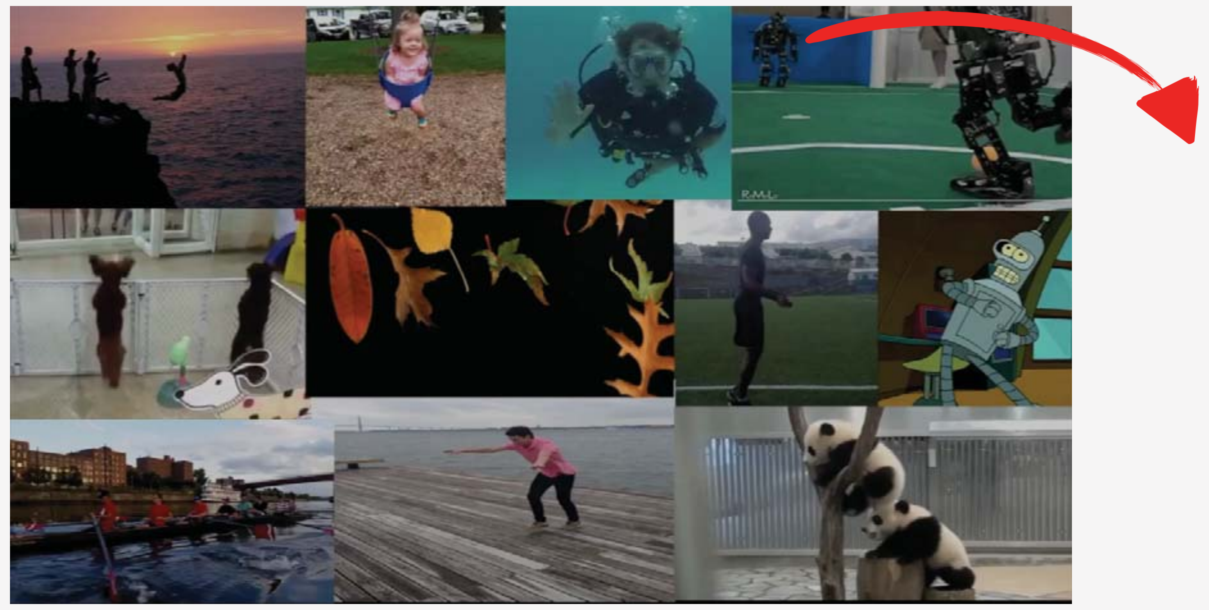
Representation

Action

PASCAL VOC

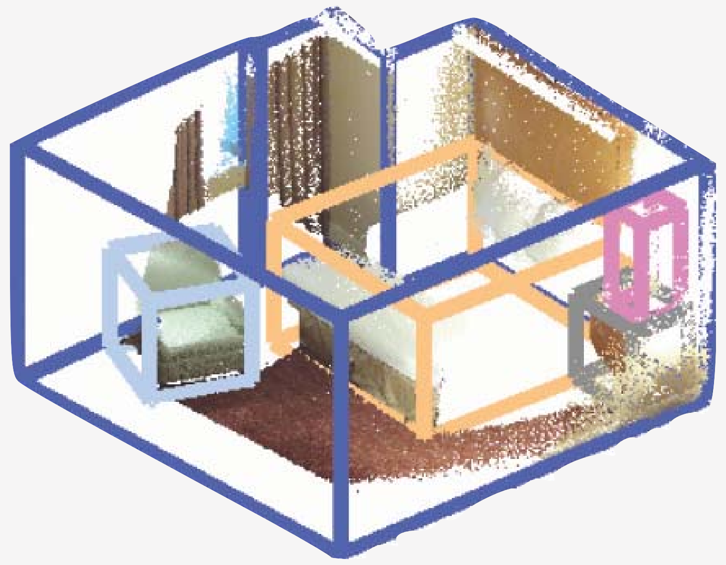


NYU depth

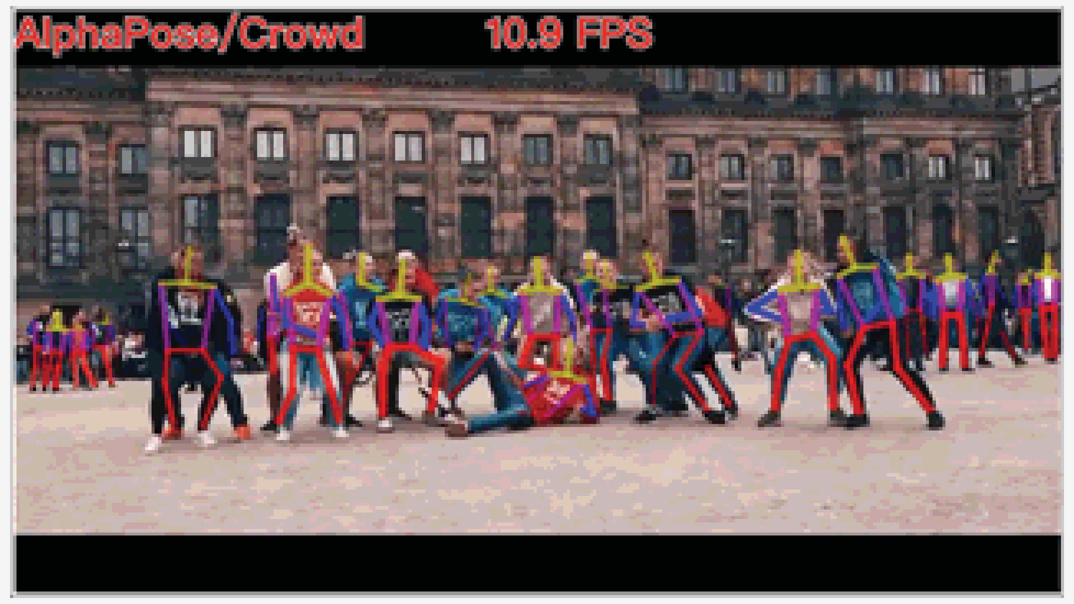


Moment in Time

ImageNet



SUN RGB-D



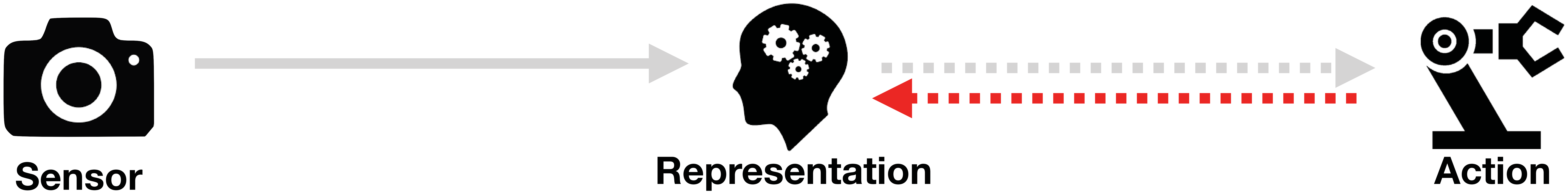
CrowdPose

What causes all the motions?

✘ Agent cannot actively choose or apply the action.

✘ Casual relationship between action and motion.

Scene Understanding



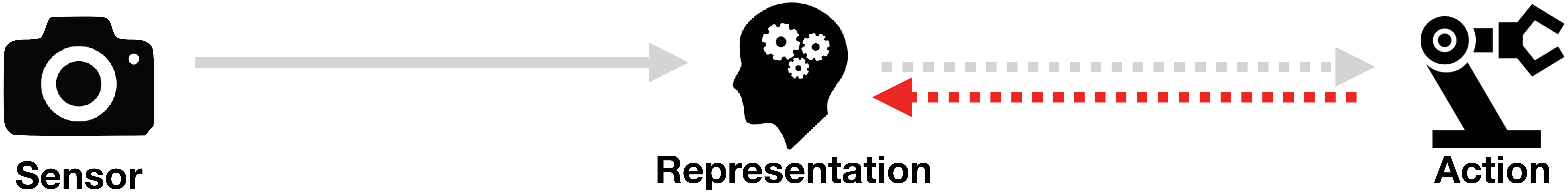
Using active exploration to retrieve useful information

Dip our toes into the water to sense its temperature

Push a large box to sense its weight

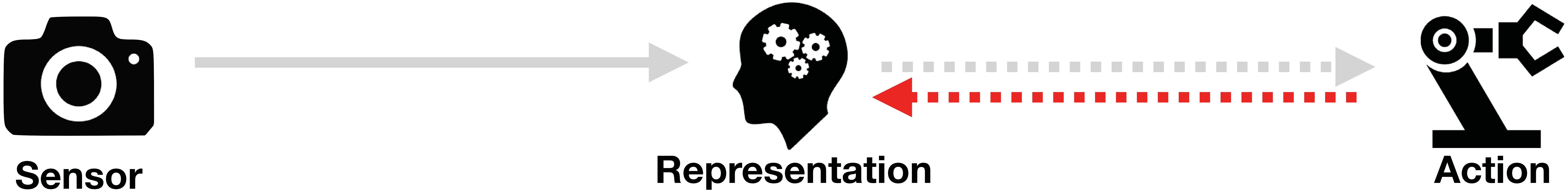
Pick up a t-shirt from a pile to recognize it

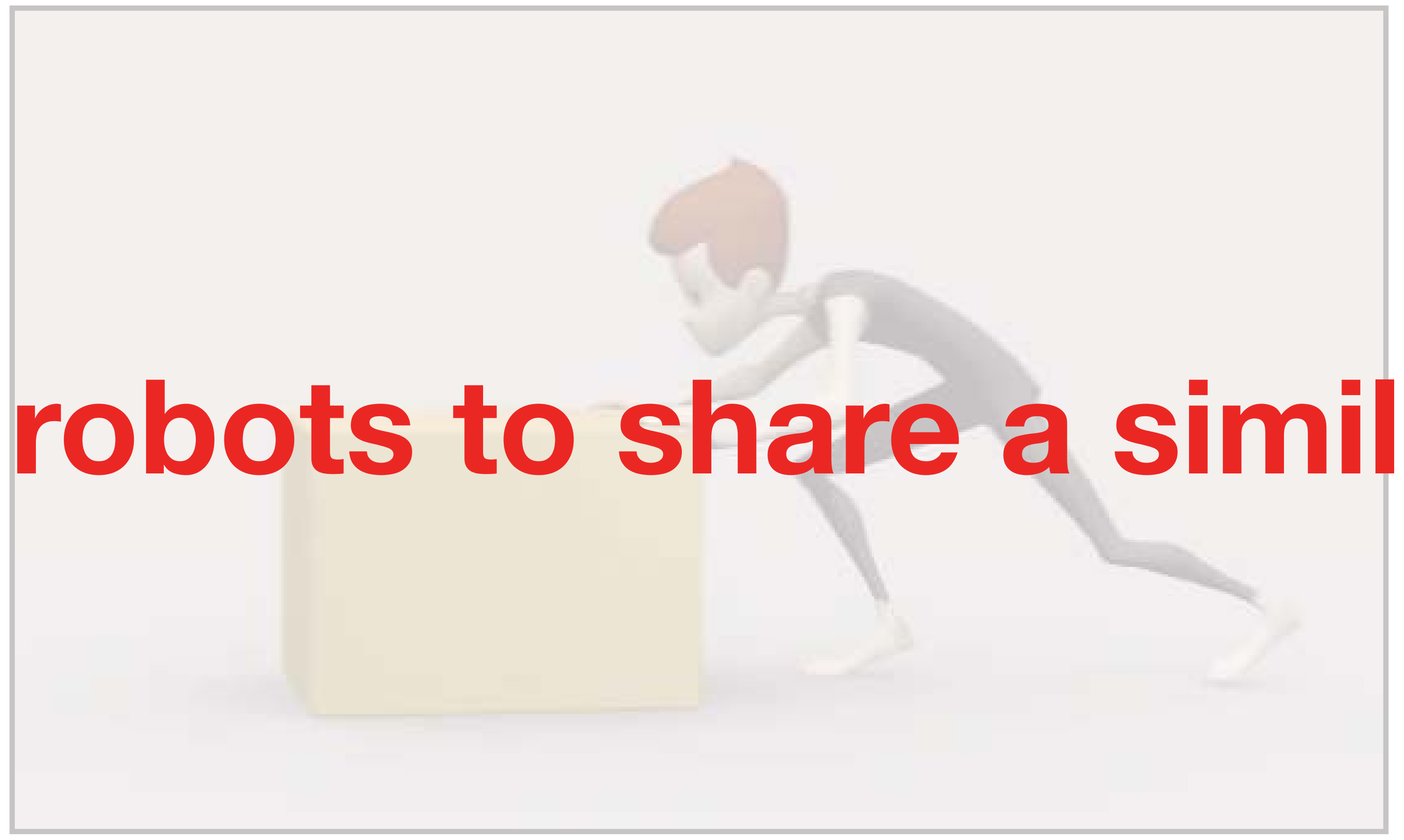
Scene Understanding



Action	Dipping	Pushing	Lifting
Information	Temperature	Weight	Identity
Planing	Swim	Lift up the box	Wear the T-Shirt

Scene Understanding

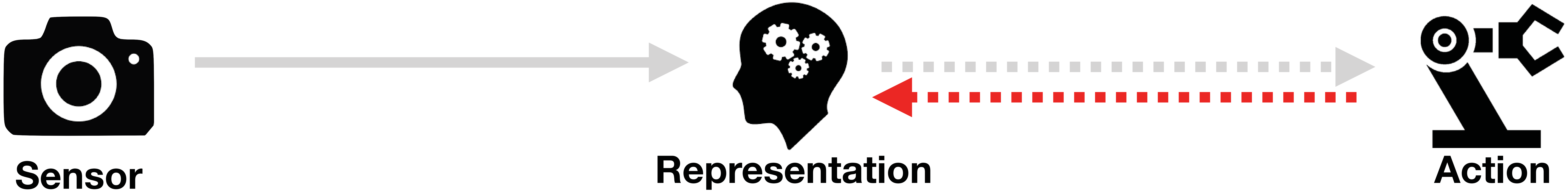


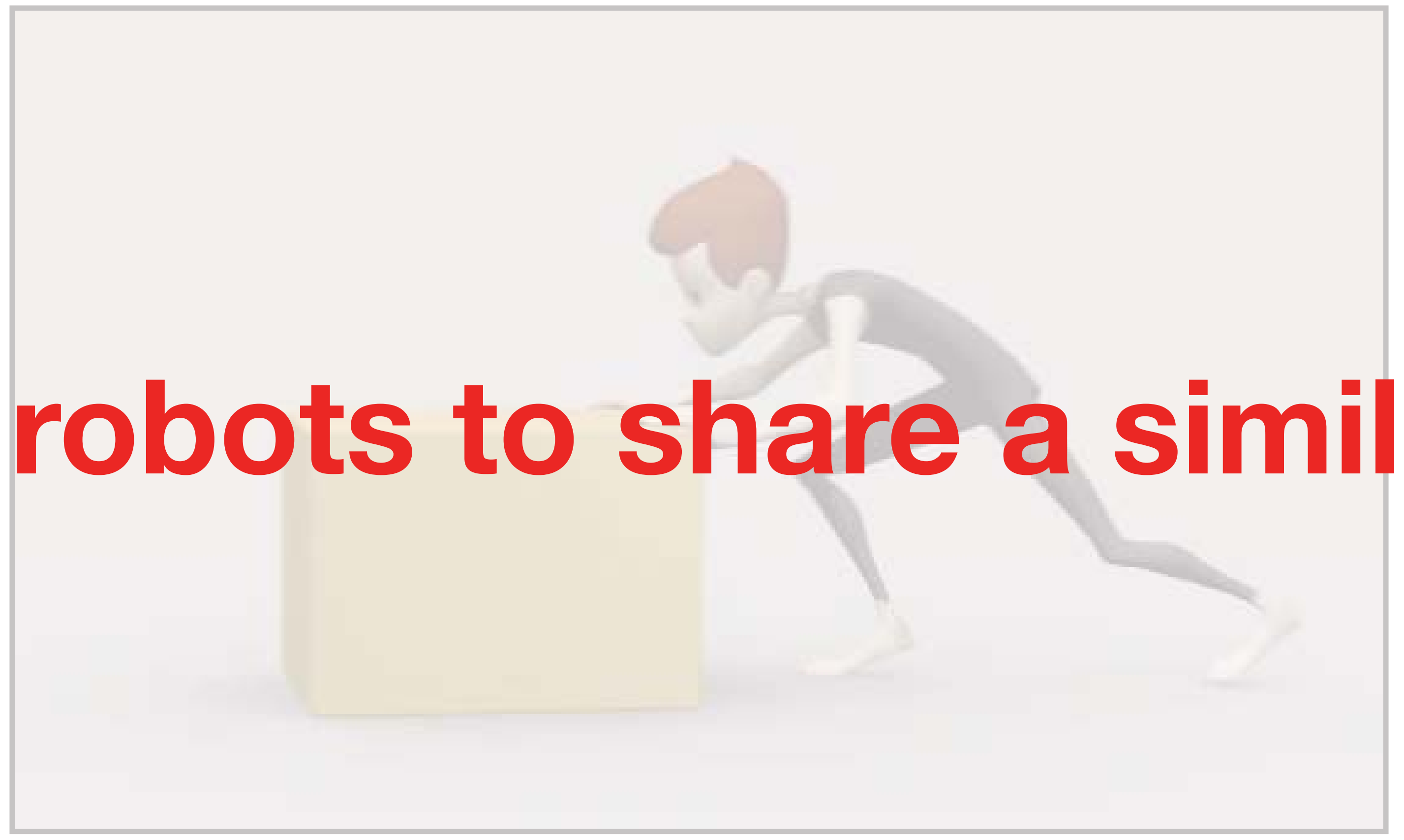


Can we enable robots to share a similar capability?

Action	Dipping	Pushing	Lifting
Information	Temperature	Weight	Identity
Planing	Swim	Lift up the box	Wear the T-Shirt

Active Scene Understanding

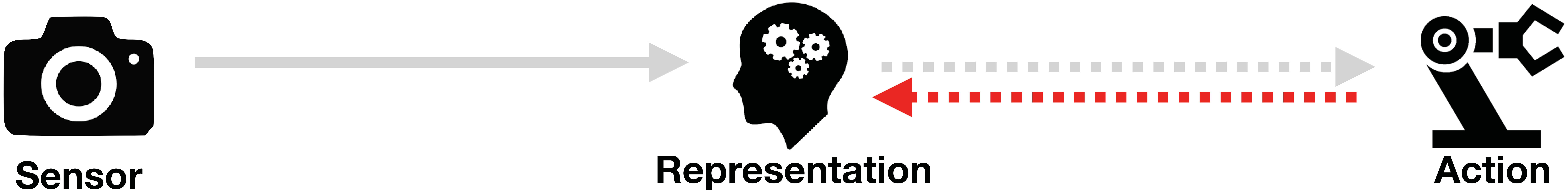




Can we enable robots to share a similar capability?

Action	Dipping	Pushing	Lifting
Information	Temperature	Weight	Identity
Planing	Swim	Lift up the box	Wear the T-Shirt

Active Scene Understanding



Sensor

Representation

Action

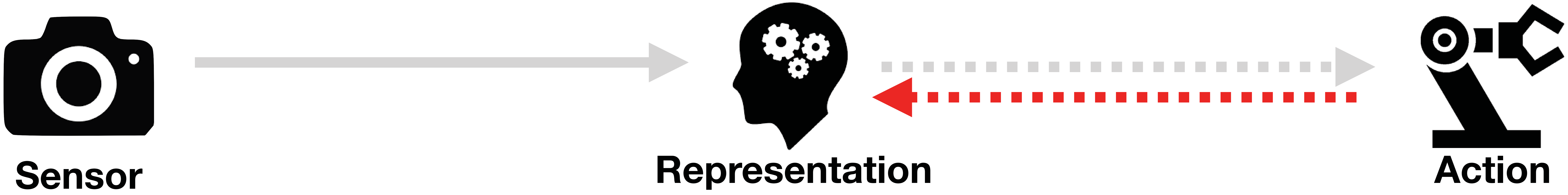
1. Obtain additional observations that hard to obtain passively

2. Discover objects physical properties beyond visual appearance

3. Provide opportunities for self-supervised learning

Advantages?

Active Scene Understanding

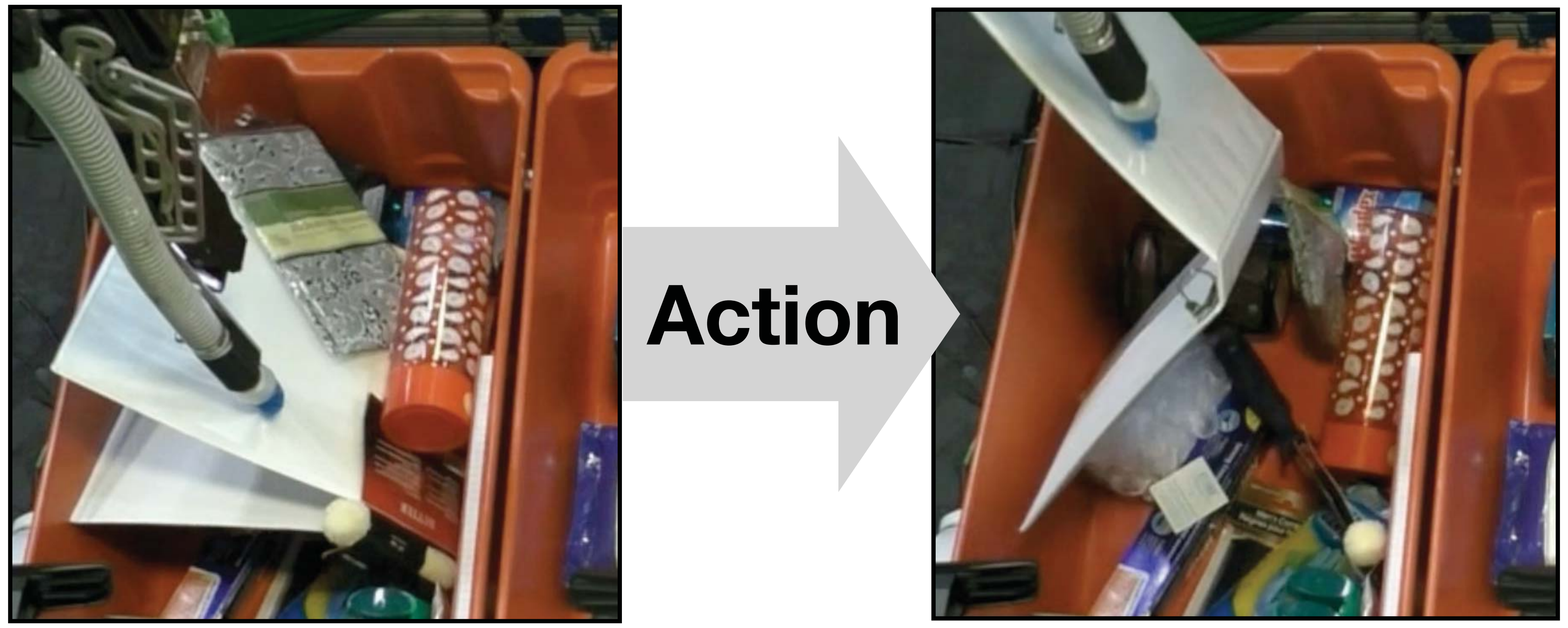


Sensor

Representation

Action

1. Obtain additional observations that are hard to obtain passively
2. Discover objects physical properties beyond visual appearance
3. Provide opportunities for self-supervised learning



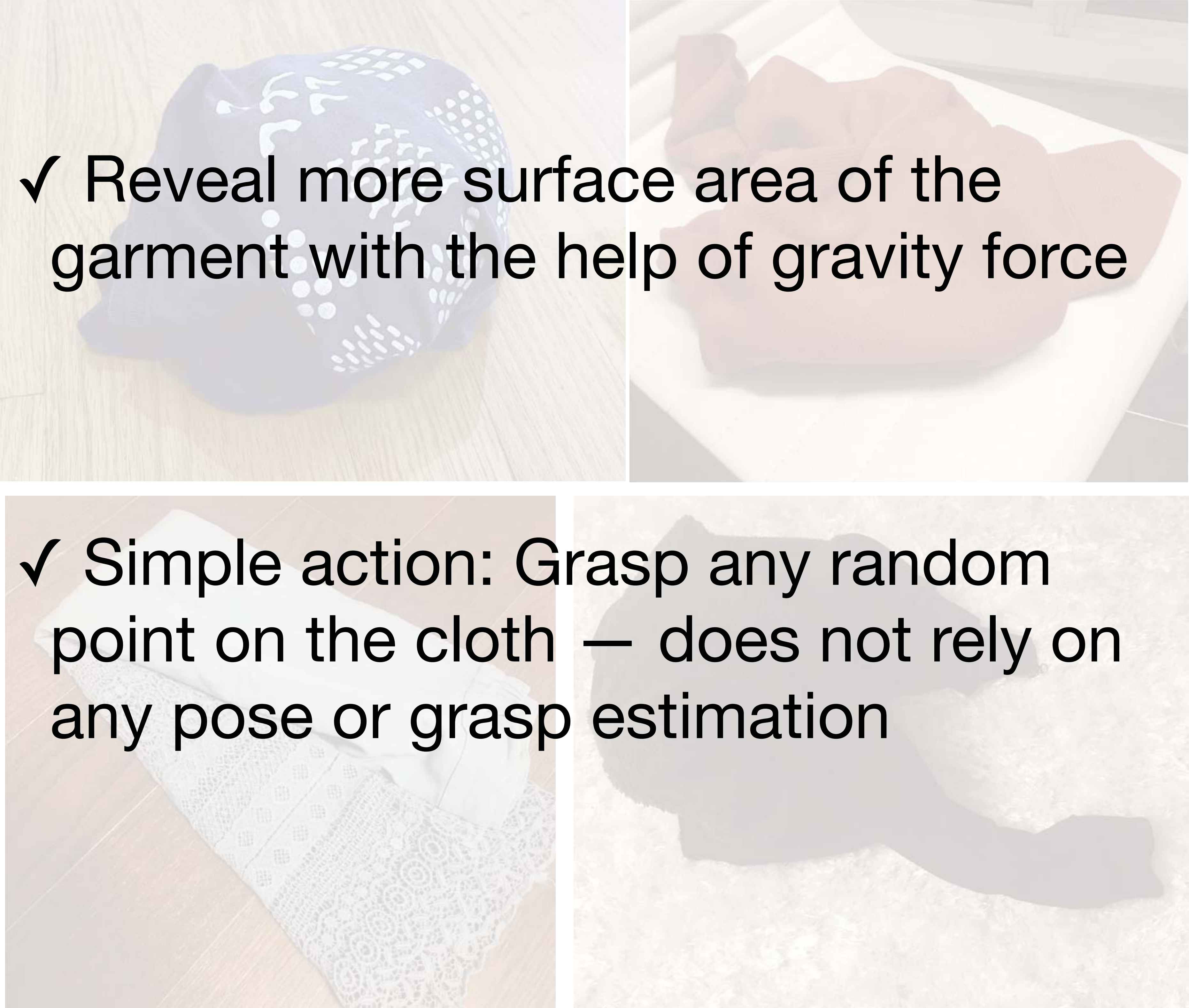
Additional Observation

for deformable objects …

Severe self-occlusion: Observed surface can be as little as 10% of total surface!

Additional Observation

for deformable objects …



✓ Reveal more surface area of the garment with the help of gravity force

✓ Simple action: Grasp any random point on the cloth — does not rely on any pose or grasp estimation

Severe self-occlusion: Observed surface can be as little as 10% of total surface!

Use simple robot interaction to help perception

Additional Observation

for deformable objects …

Interaction

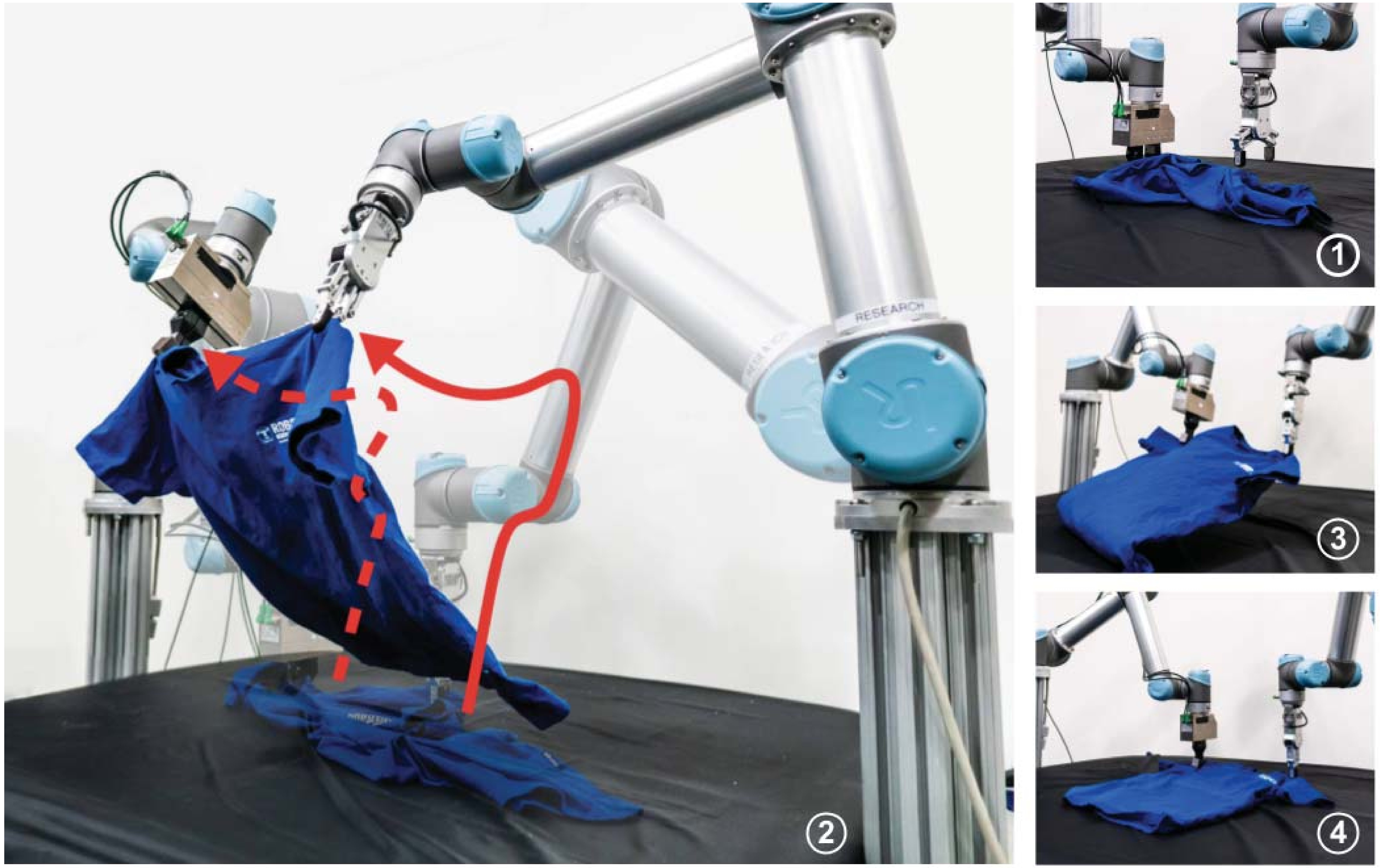
Input Observation

Completed Mesh

GarmentNets: Category-Level Pose Estimation for Garments via Canonical Space Shape Completion.

Cheng Chi and Shuran Song ICCV 2021

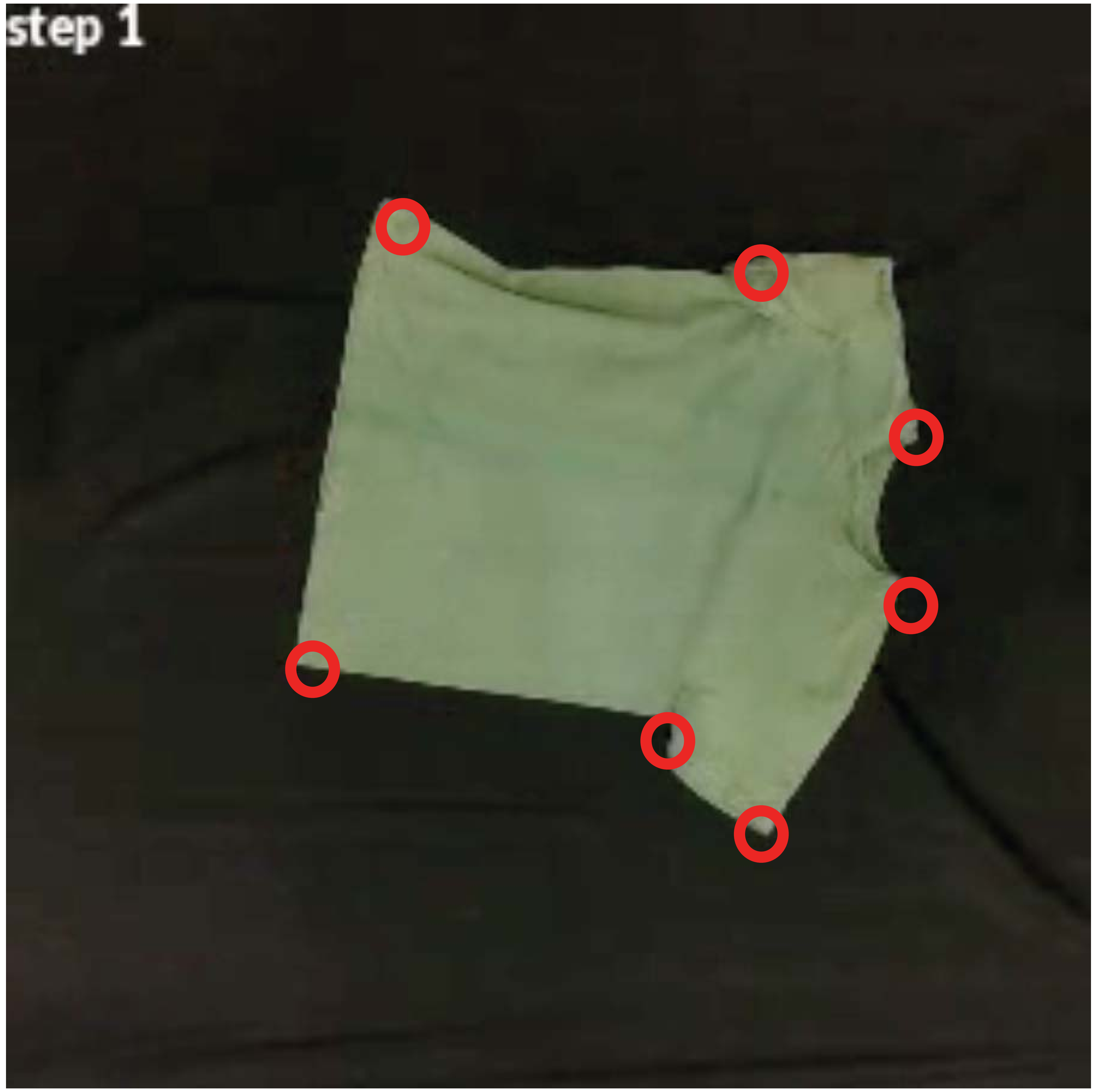
Additional Observation



FlingBot: The Unreasonable Effectiveness of Dynamic Manipulation for Cloth Unfolding

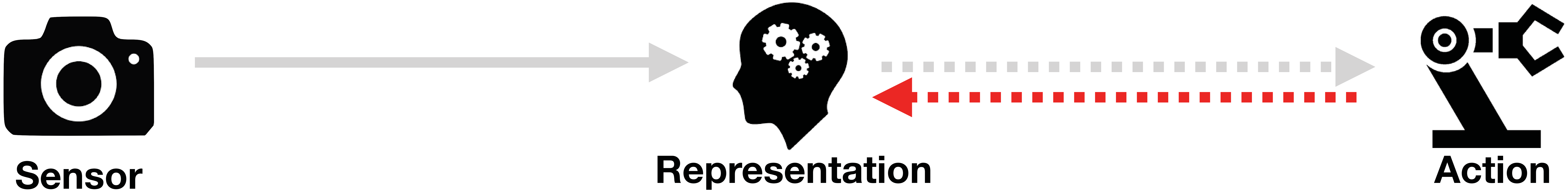
Huy Ha and Shuran Song CORL 2021

Additional Observation



FlingBot: The Unreasonable Effectiveness of Dynamic Manipulation for Cloth Unfolding
Huy Ha and Shuran Song

Active Scene Understanding

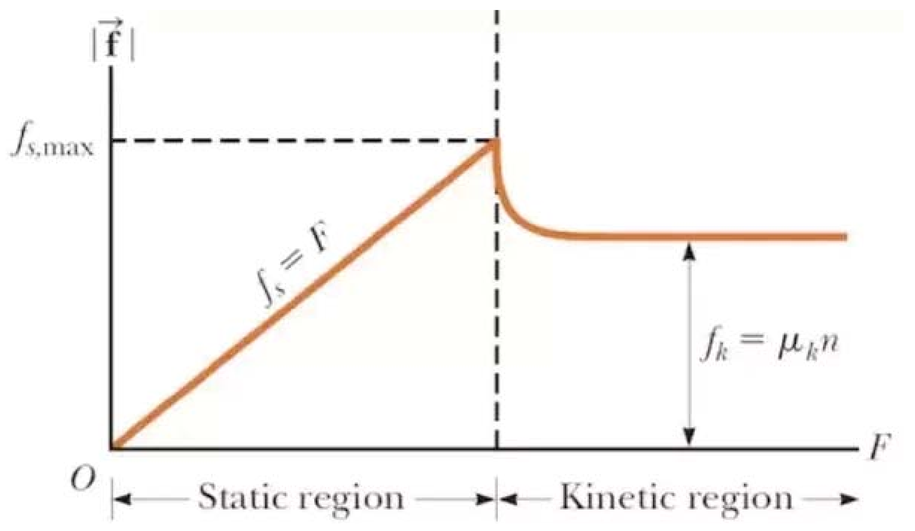


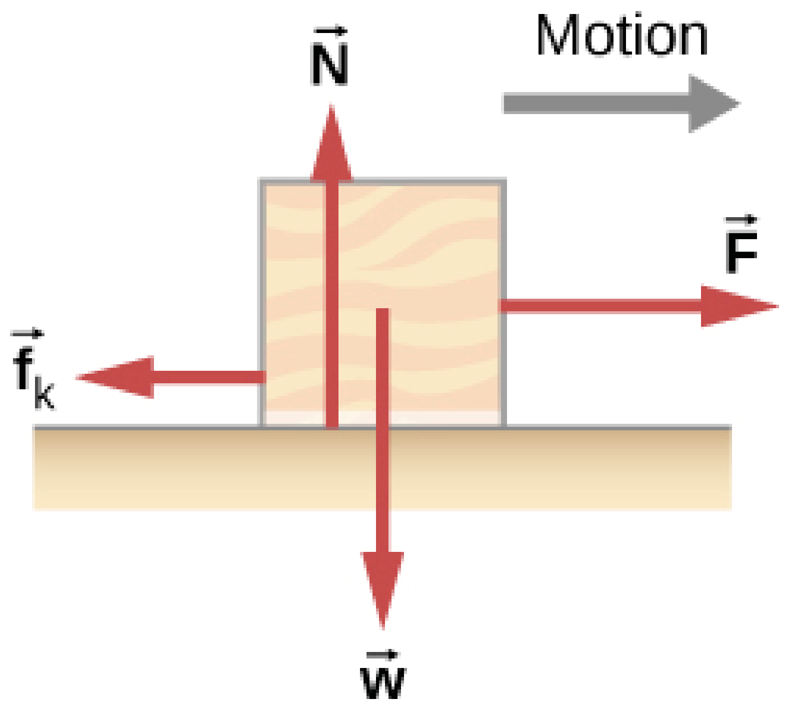
Sensor

Representation

Action

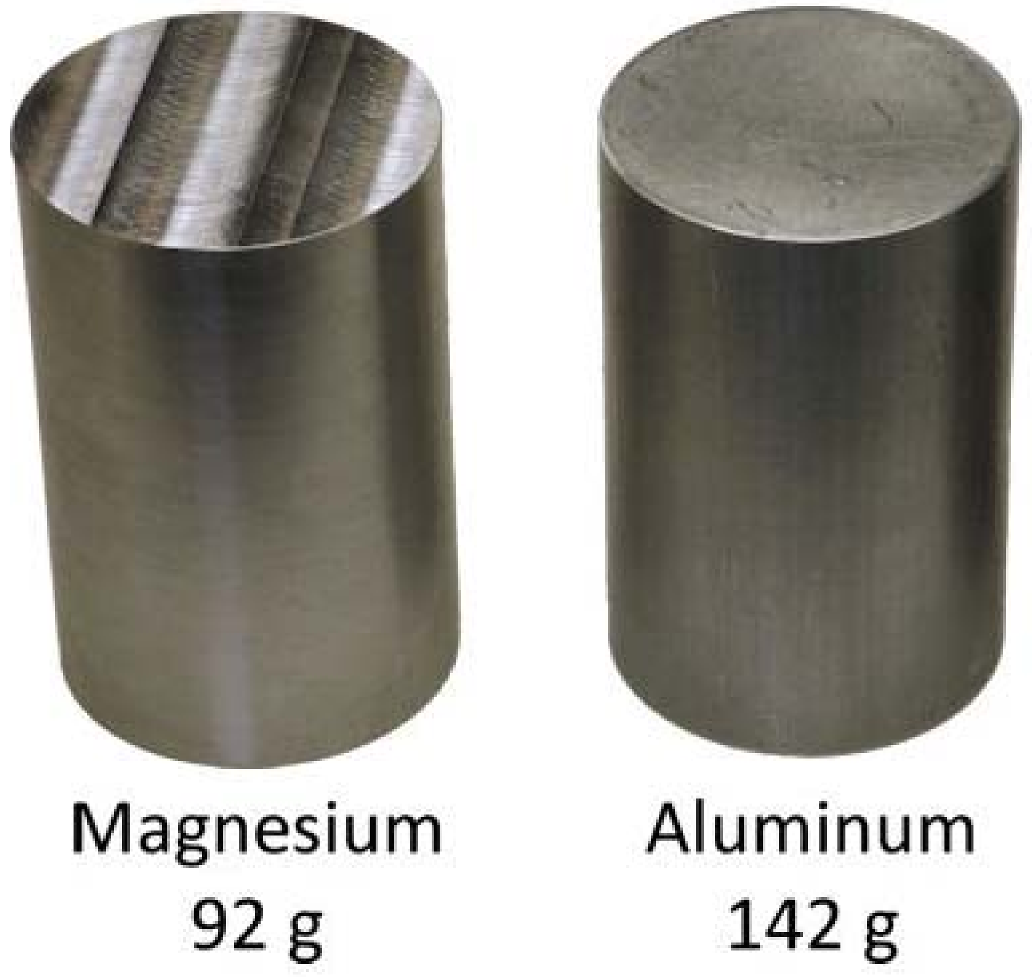
1. Obtain additional observations that hard to obtain passively
2. Discover objects physical properties beyond visual appearance
3. Provide opportunities for self-supervised learning





Why it is hard?

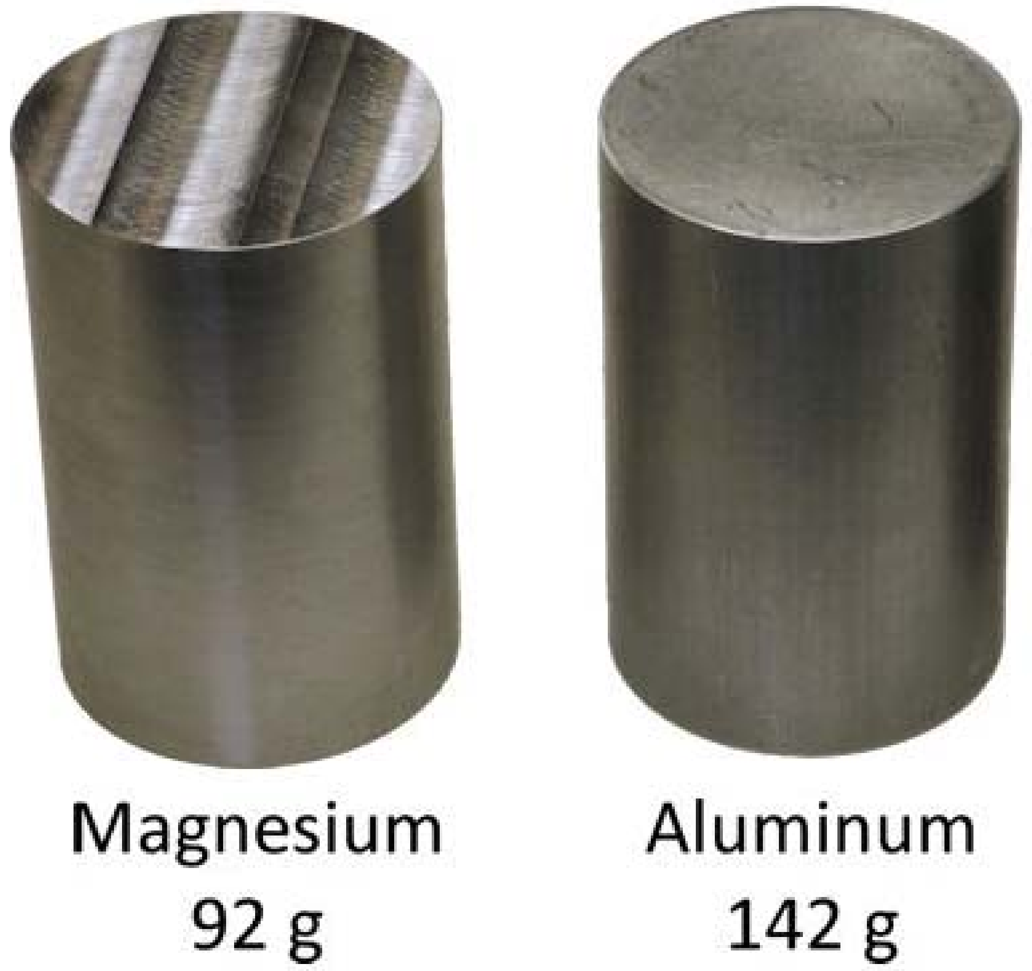
Learning object physical properties though vision

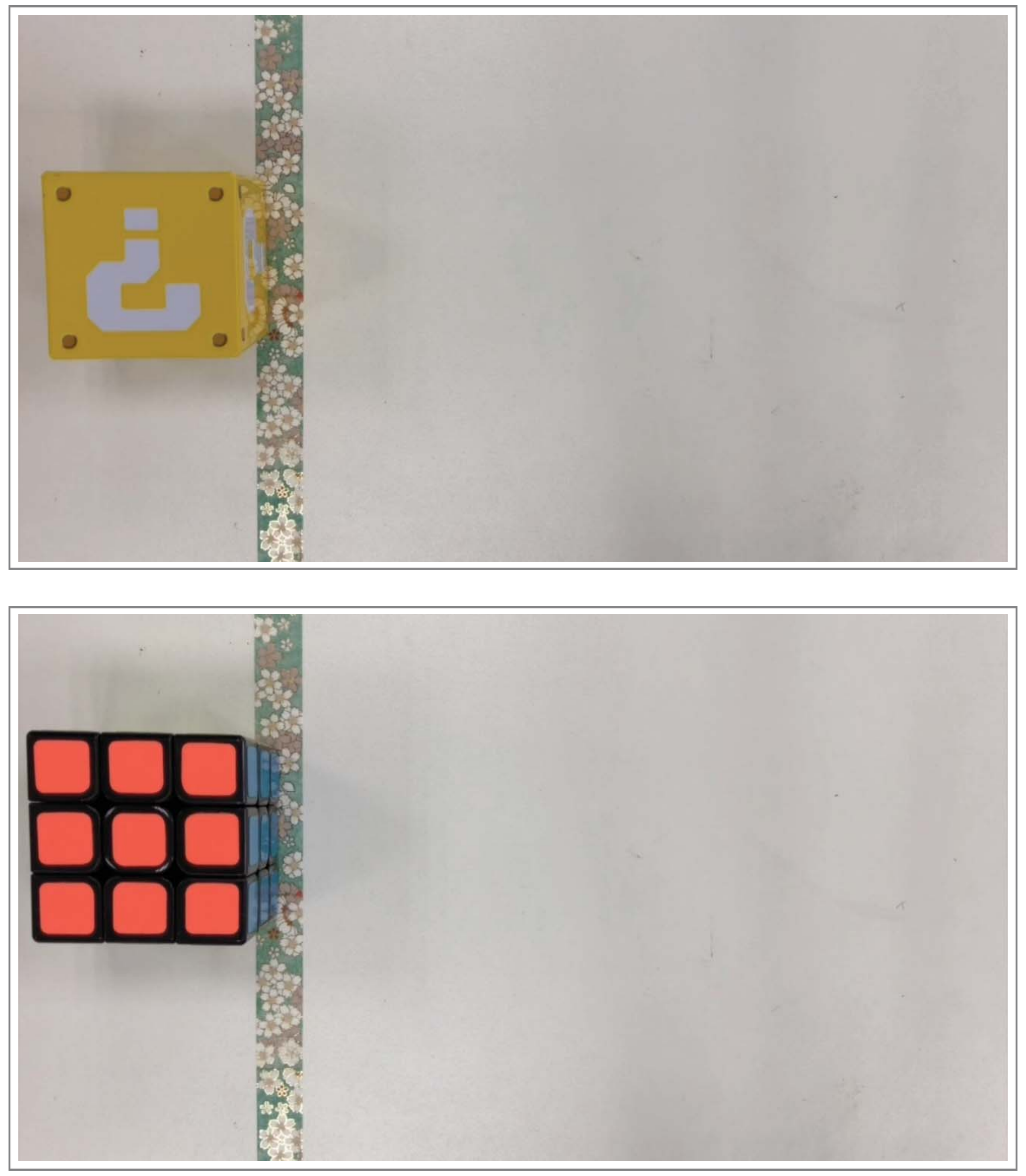


Cannot be inferred from appearance alone

Why it is hard?

Learning object physical properties though vision



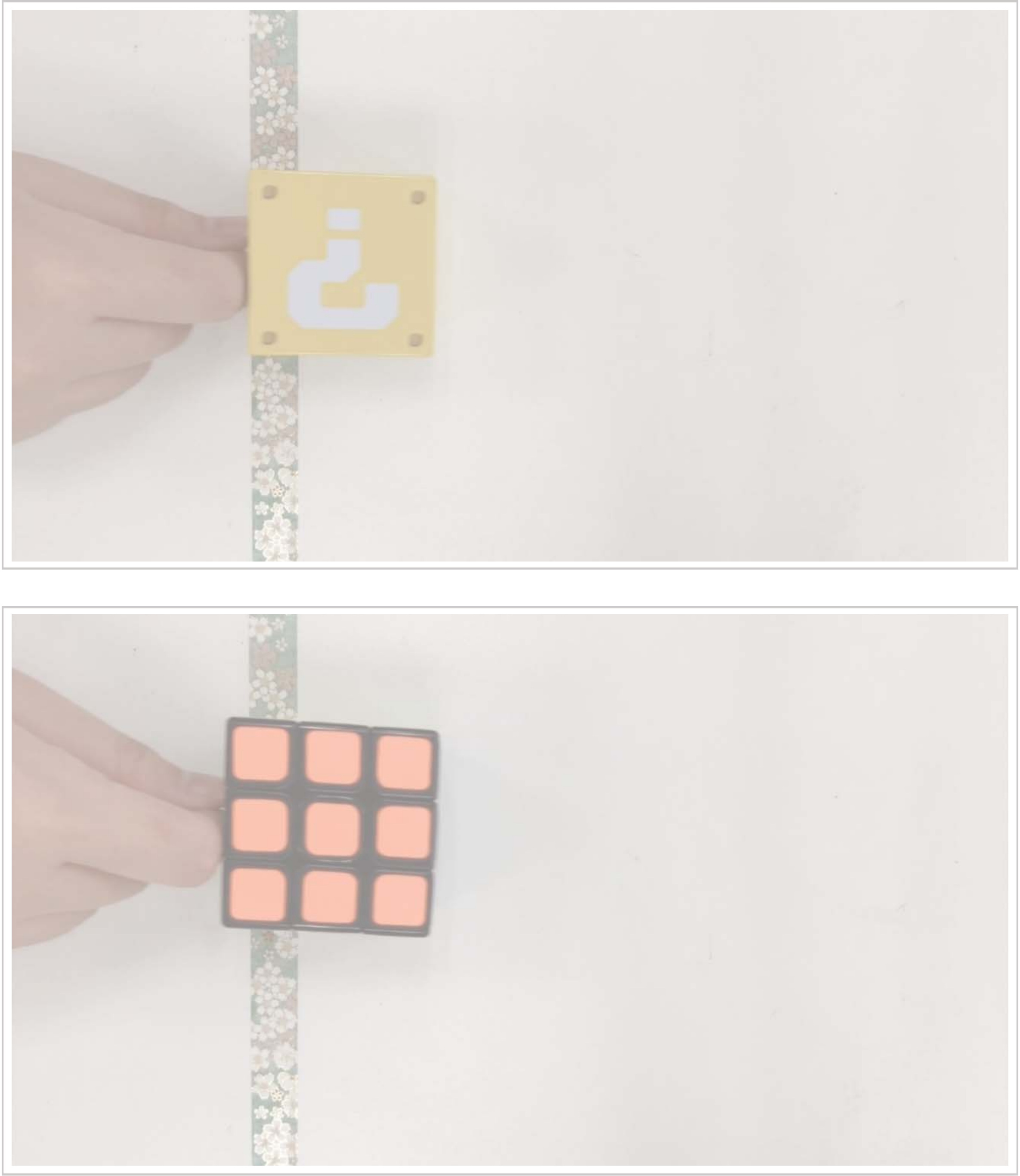


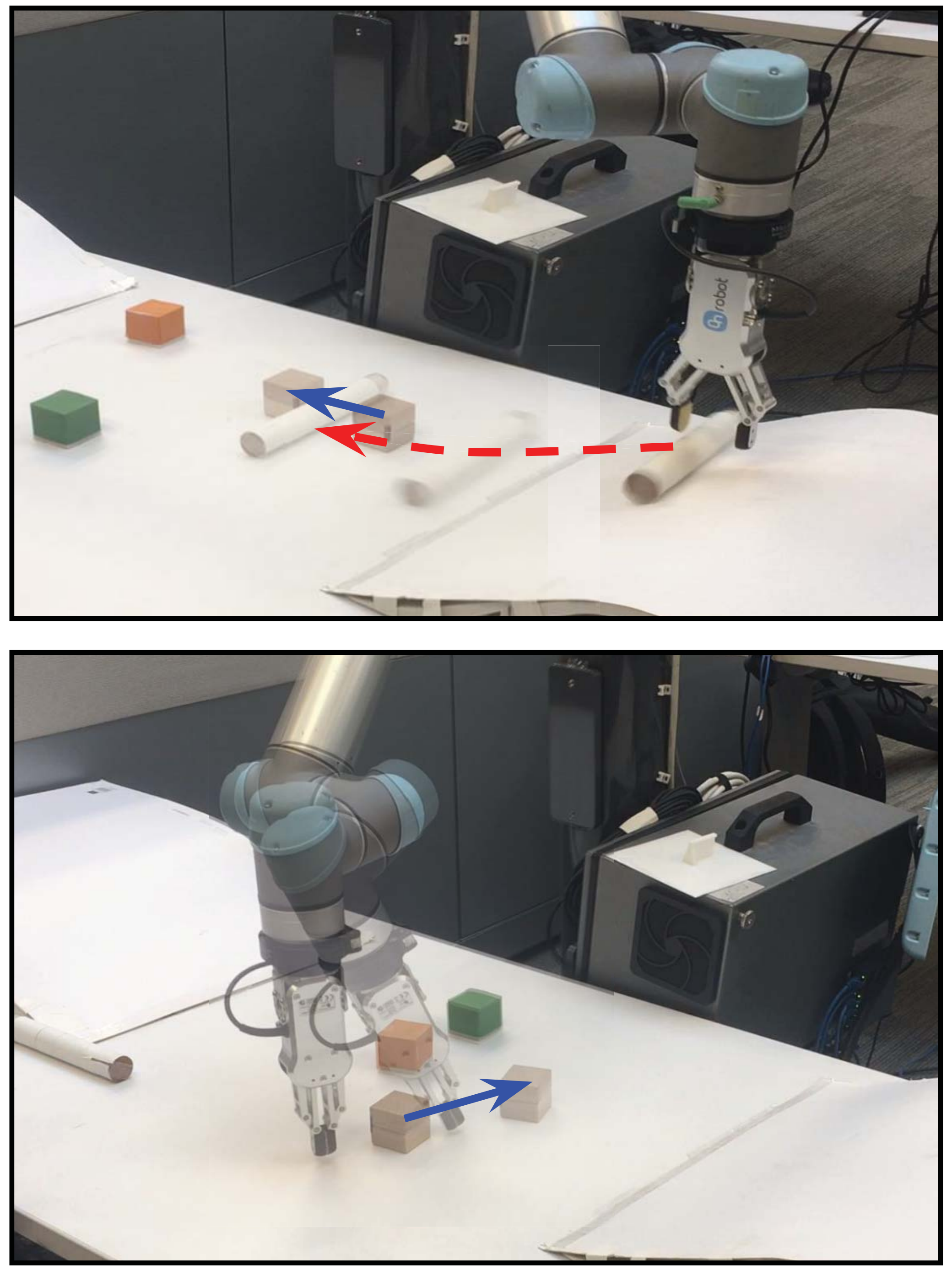
Cannot be inferred from appearance alone

Not salient under quasi-static interactions

Why it is hard?

Learning object physical properties though vision



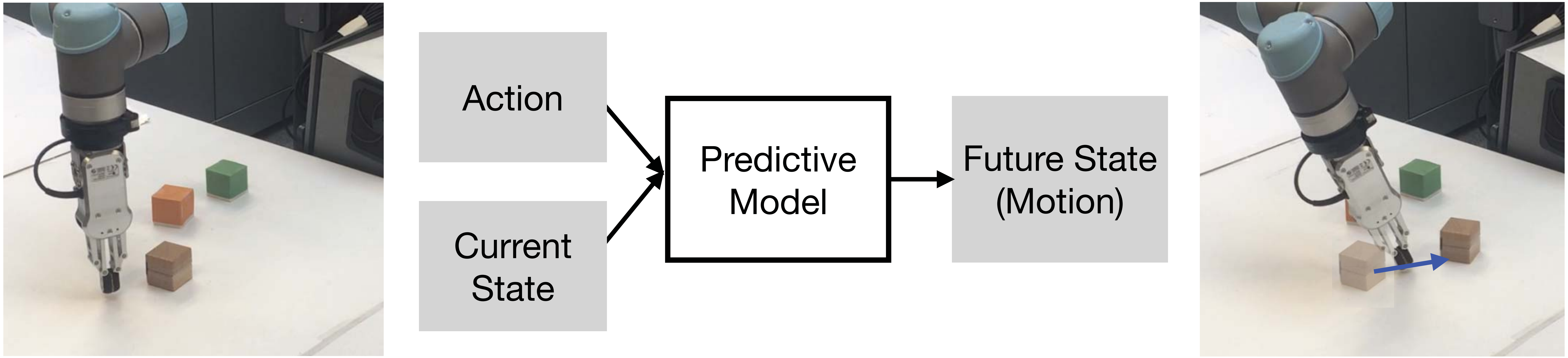


Cannot be inferred from appearance alone

Not salient under quasi-static interactions

Need multiple interactions to decouple the properties

DensePhysNet



Action

Current State

Predictive Model

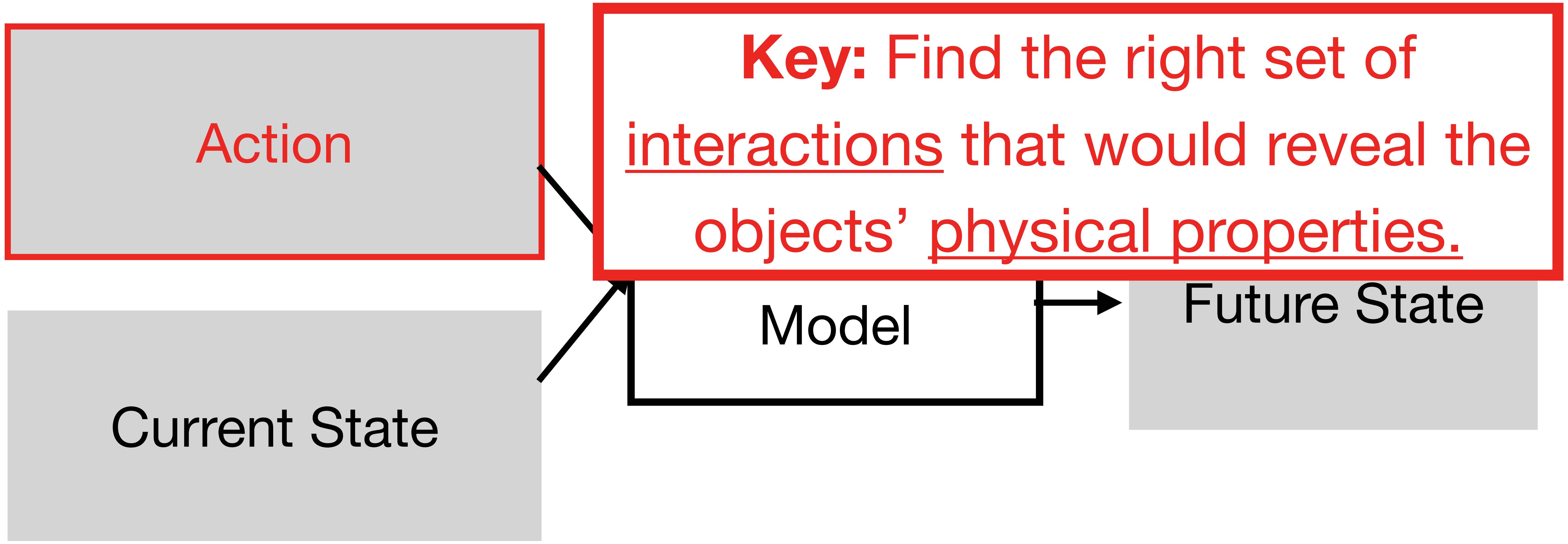
Future State (Motion)

Hypothesis:

In order to accurately predict the future states, the system will need to acquire an implicit understanding of objects' physical properties and how they influence objects' motion.

Zhenjia Xu

DensePhysNet

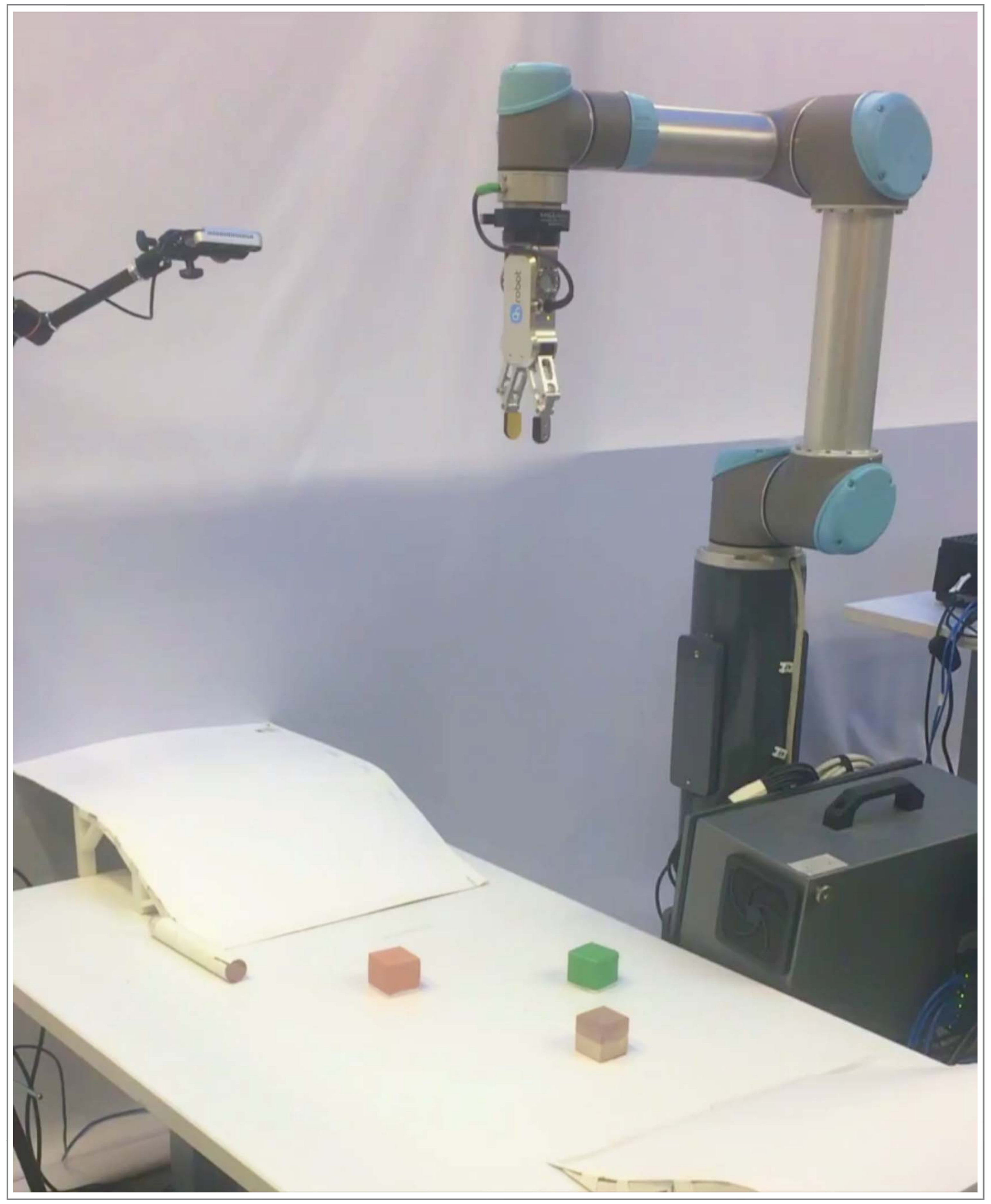


Key: Find the right set of interactions that would reveal the objects' physical properties.

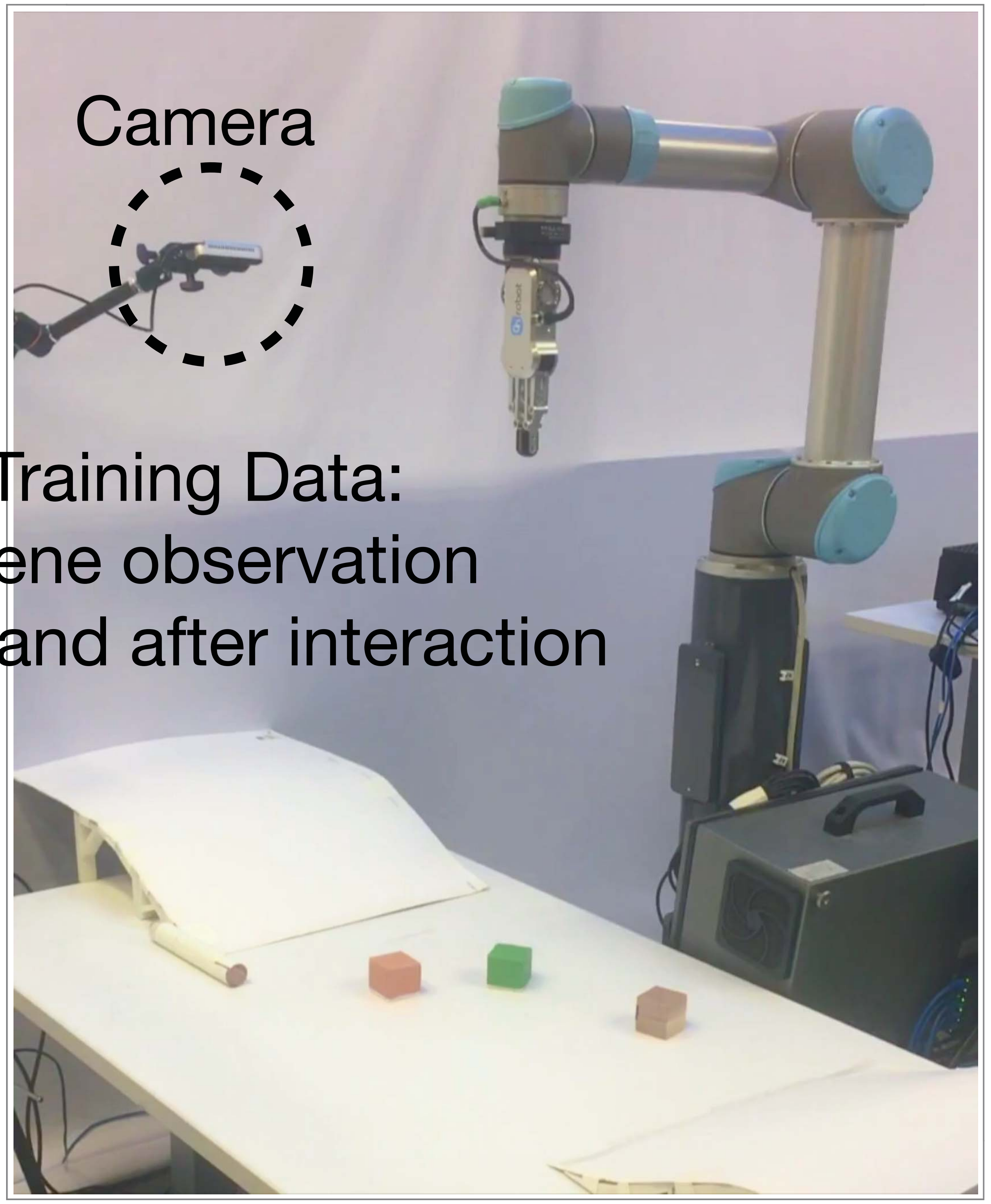
Hypothesis:

To accurately predict the future states, the system will need to acquire an implicit understanding of objects' physical properties and how they influence objects' motion.

Dynamic Interactions



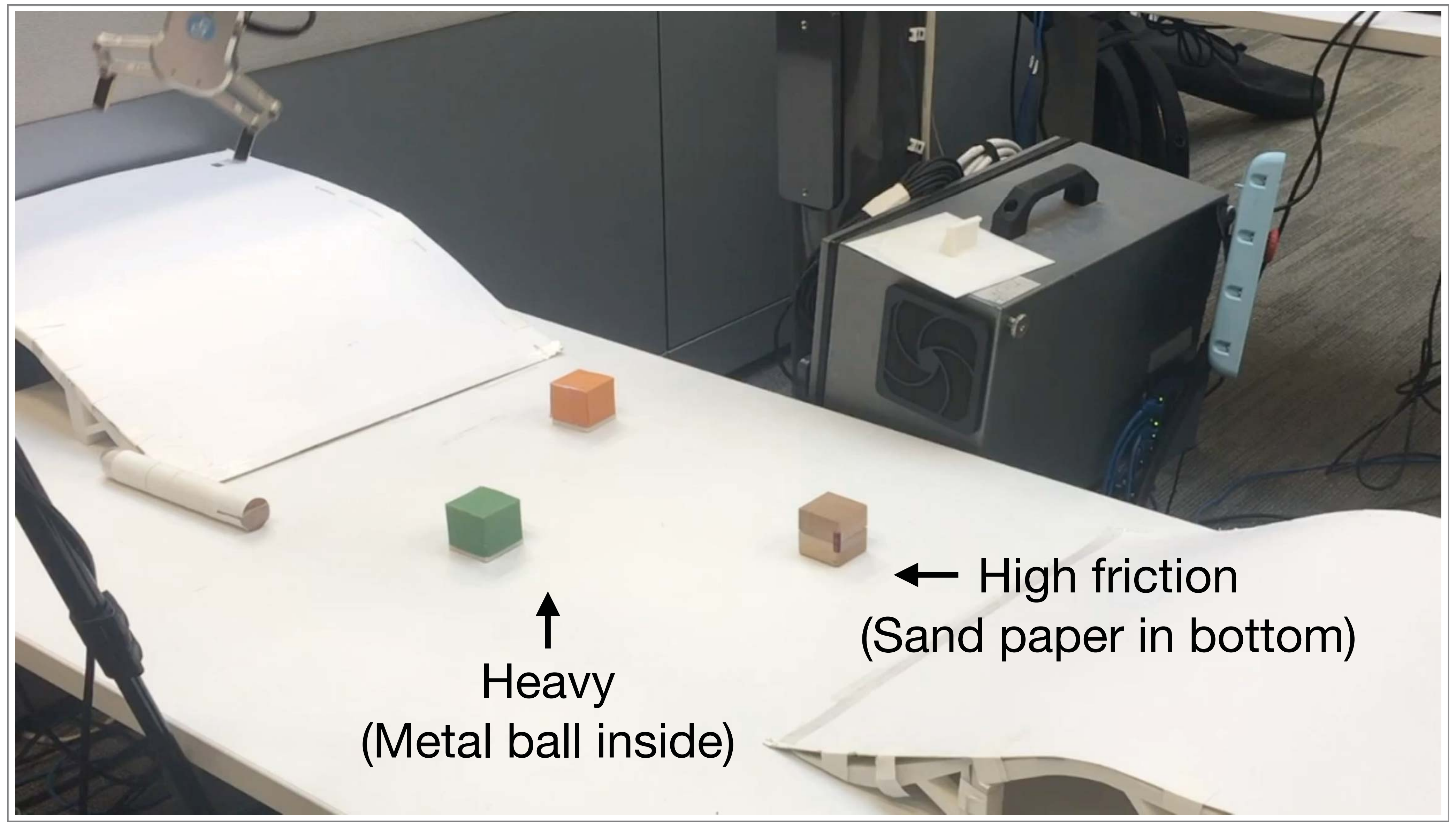
Sliding



Training Data:
Scene observation
before and after interaction

Collision

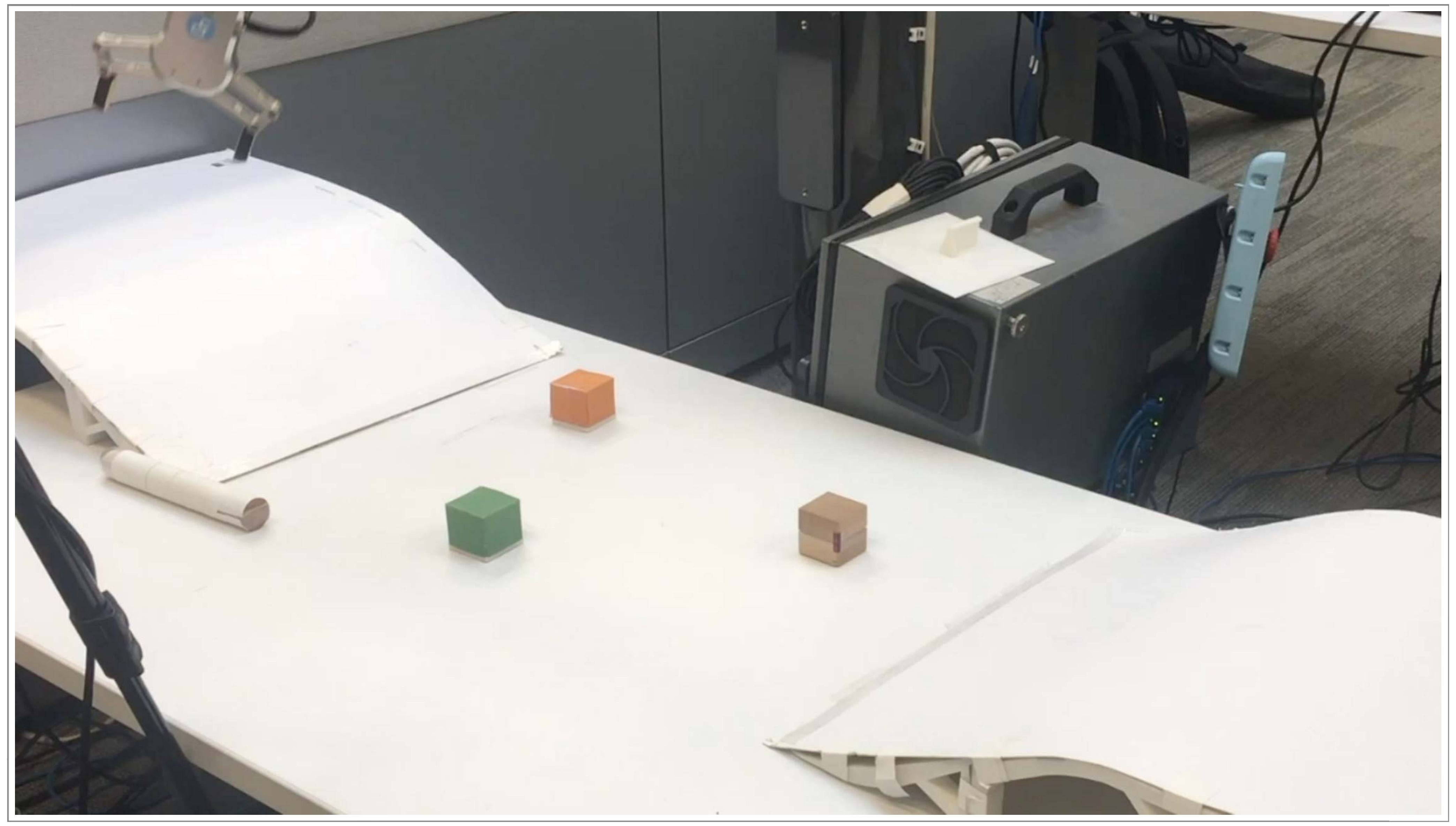
DensePhysNet



Interaction Video

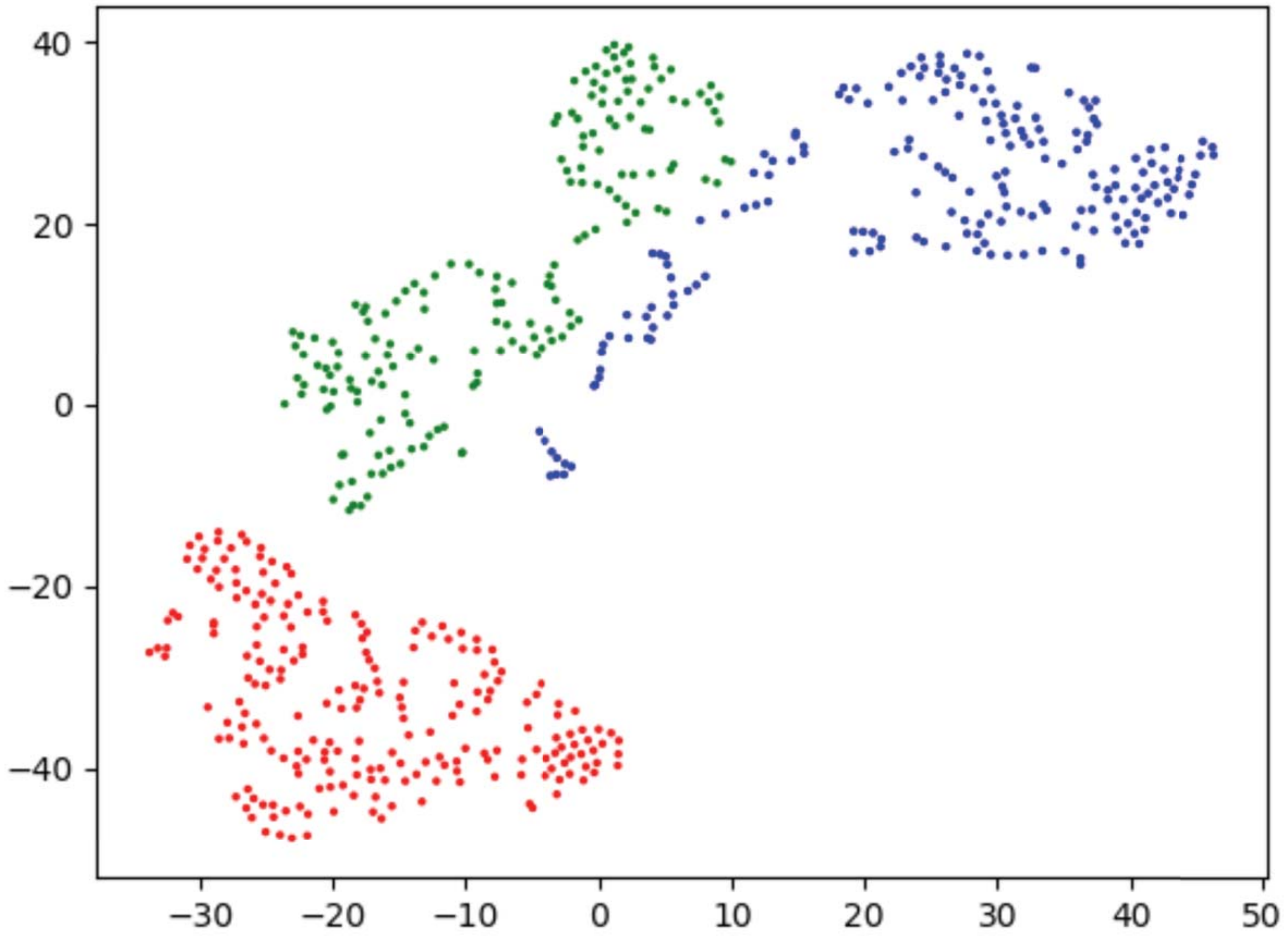
Action: Sliding or Collision

DensePhysNet



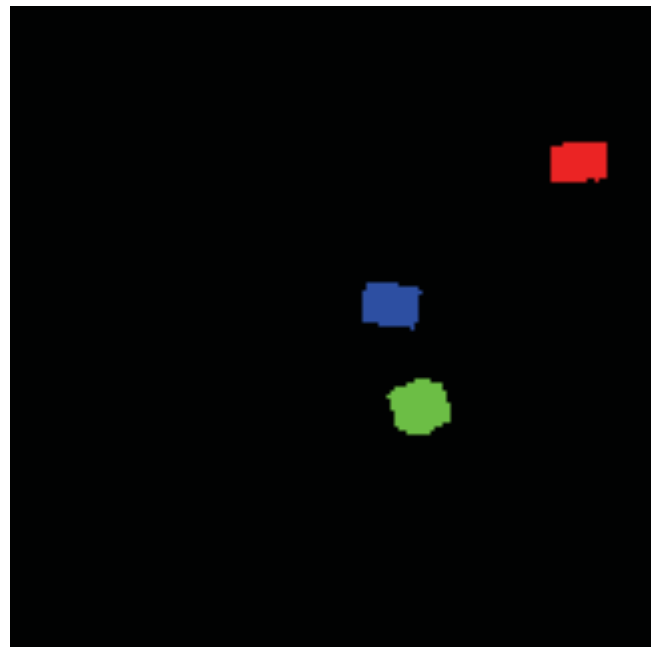
Interaction Video

Action: Sliding or Collision



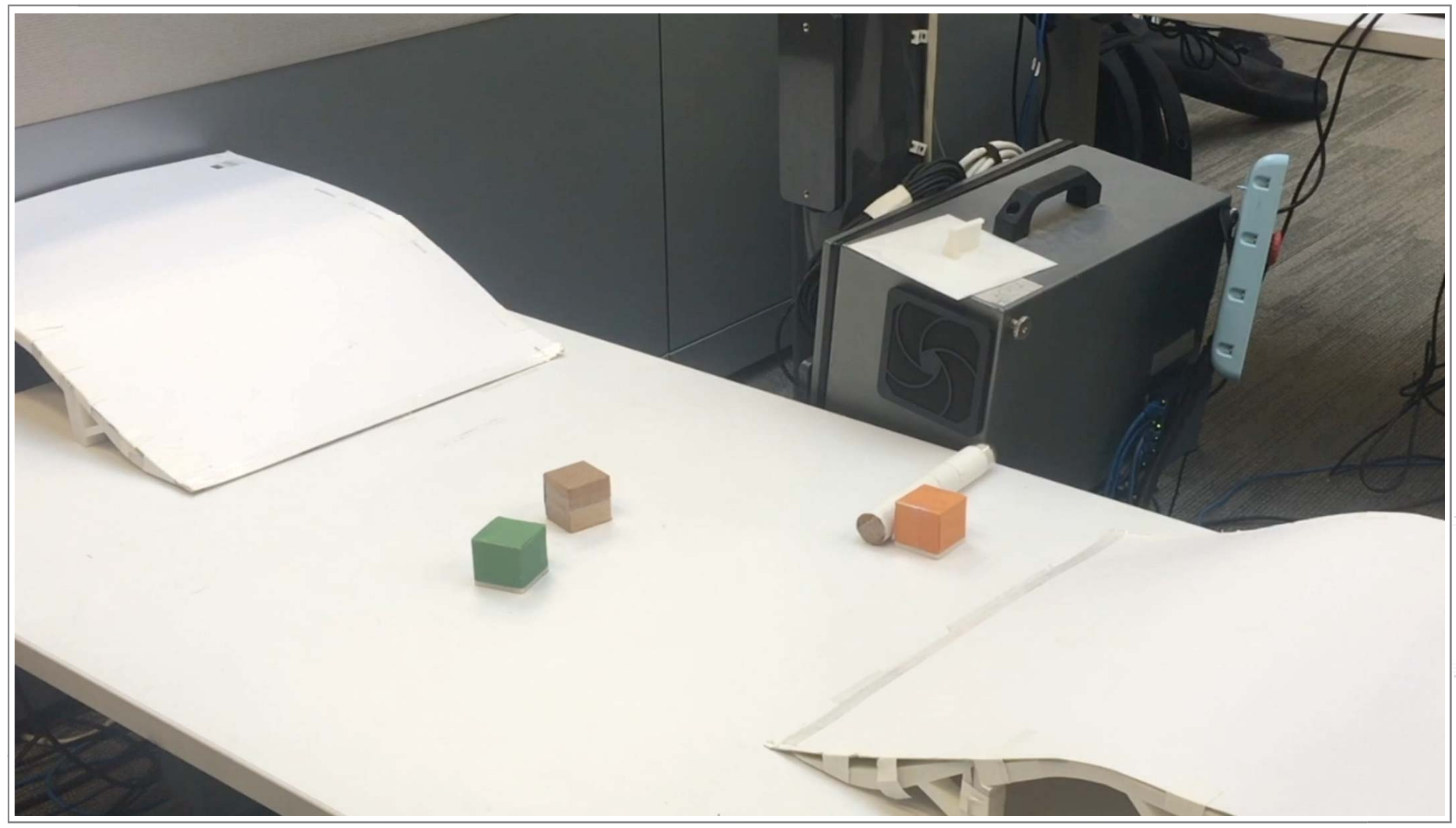
Feature Embedding Space

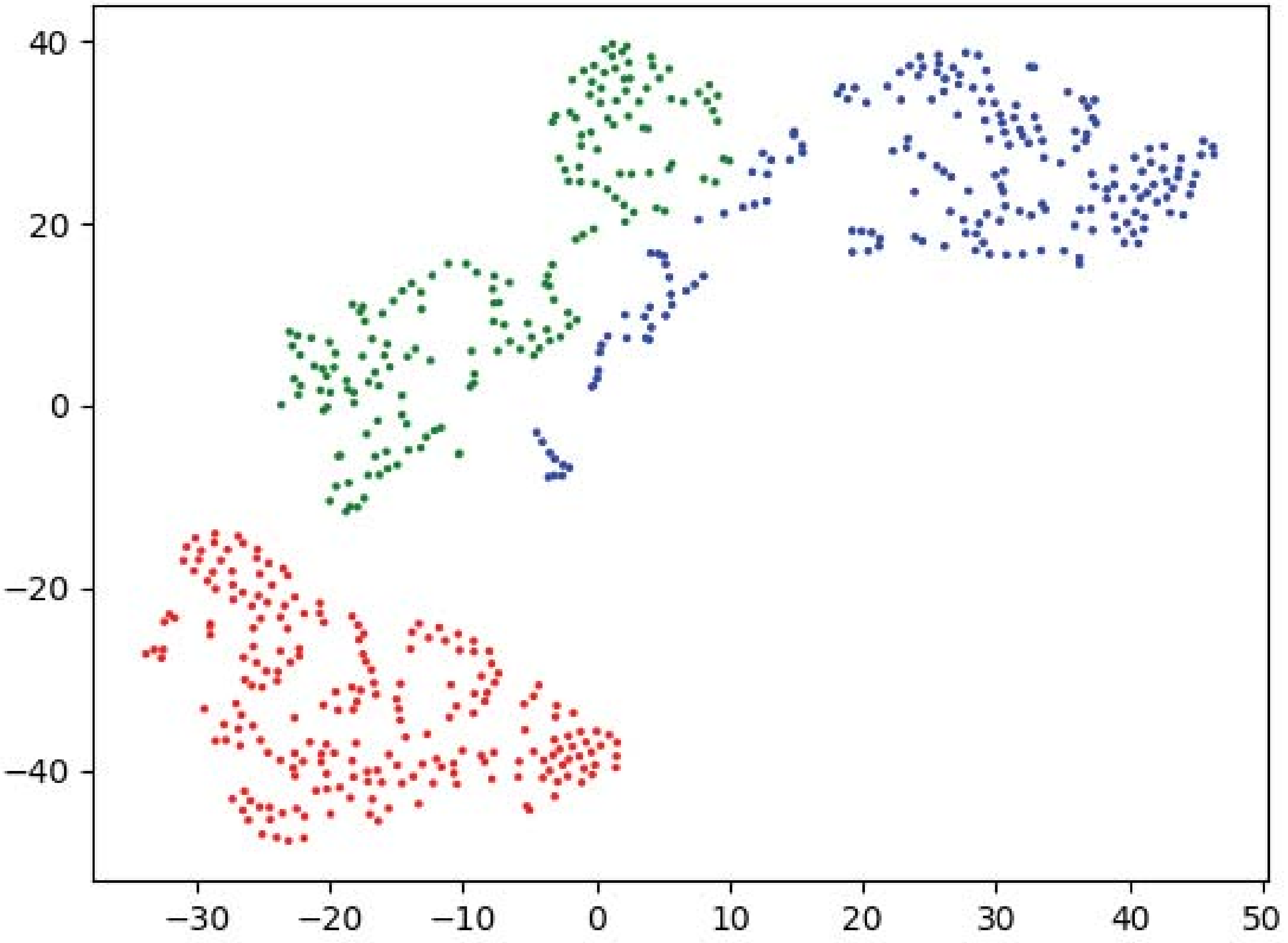
Larger distance indicate larger feature distance

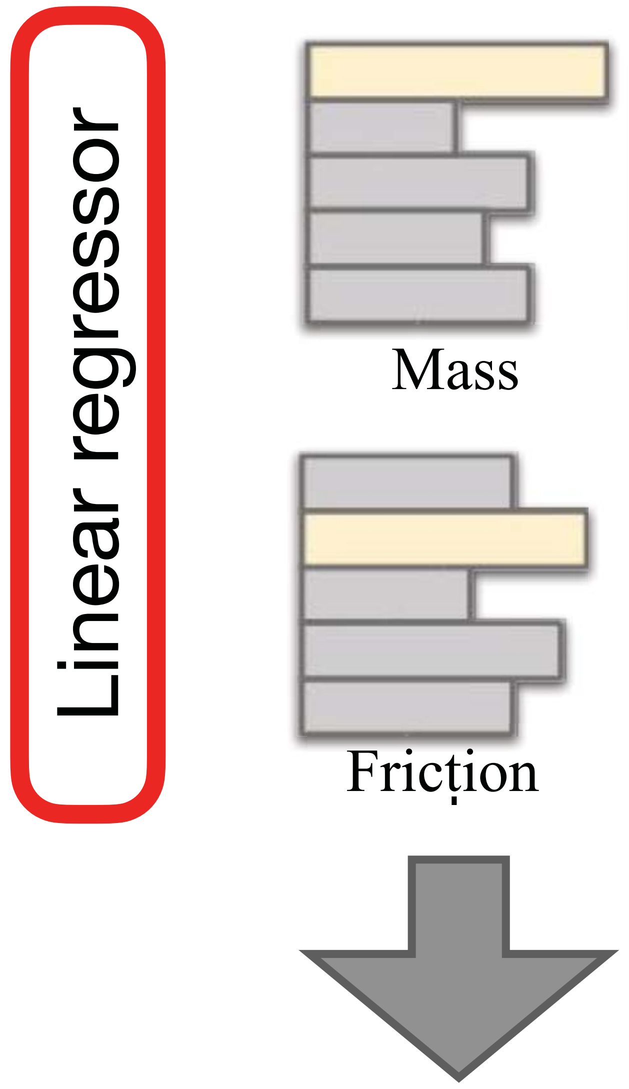


Object Instance Mask

DensePhysNet

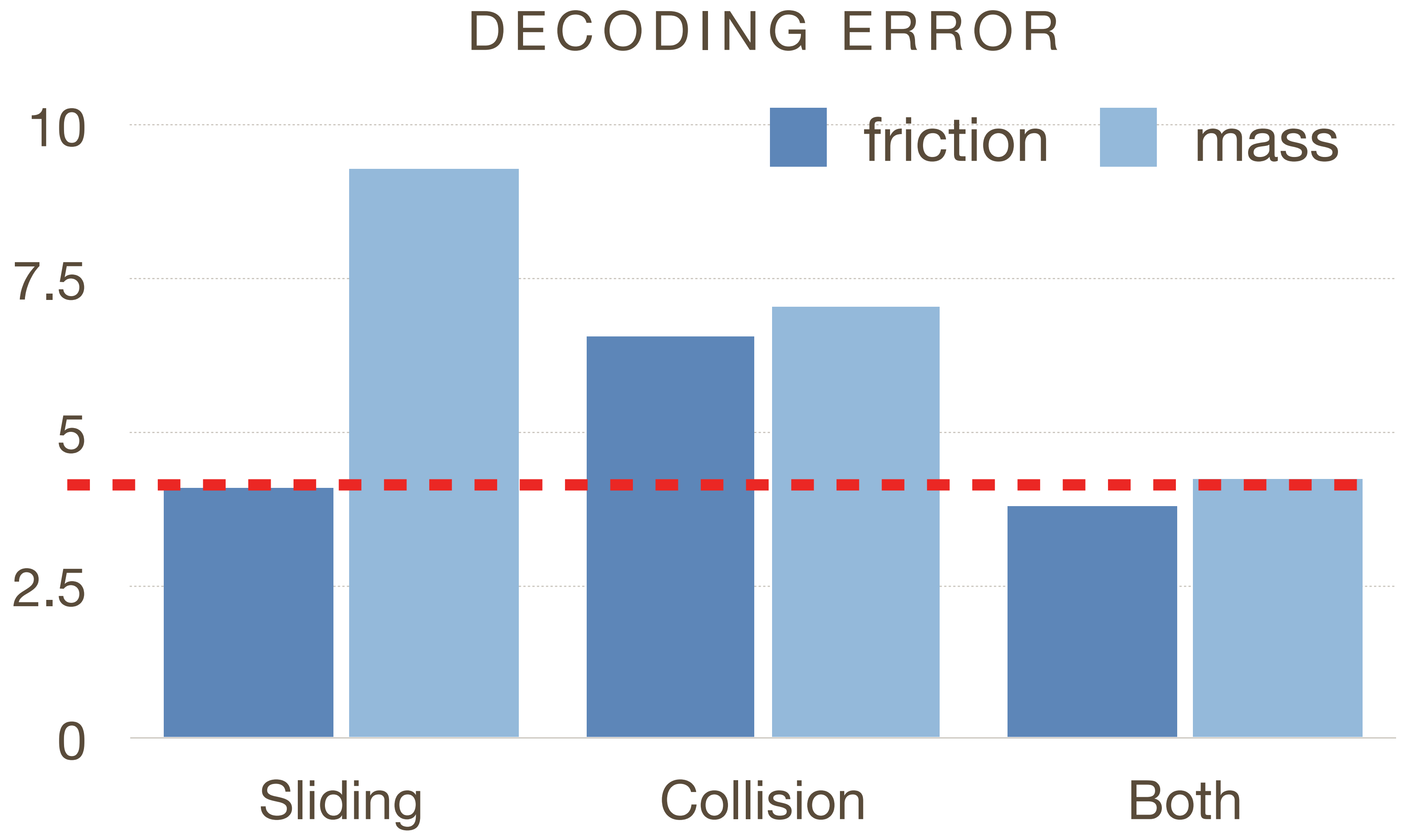




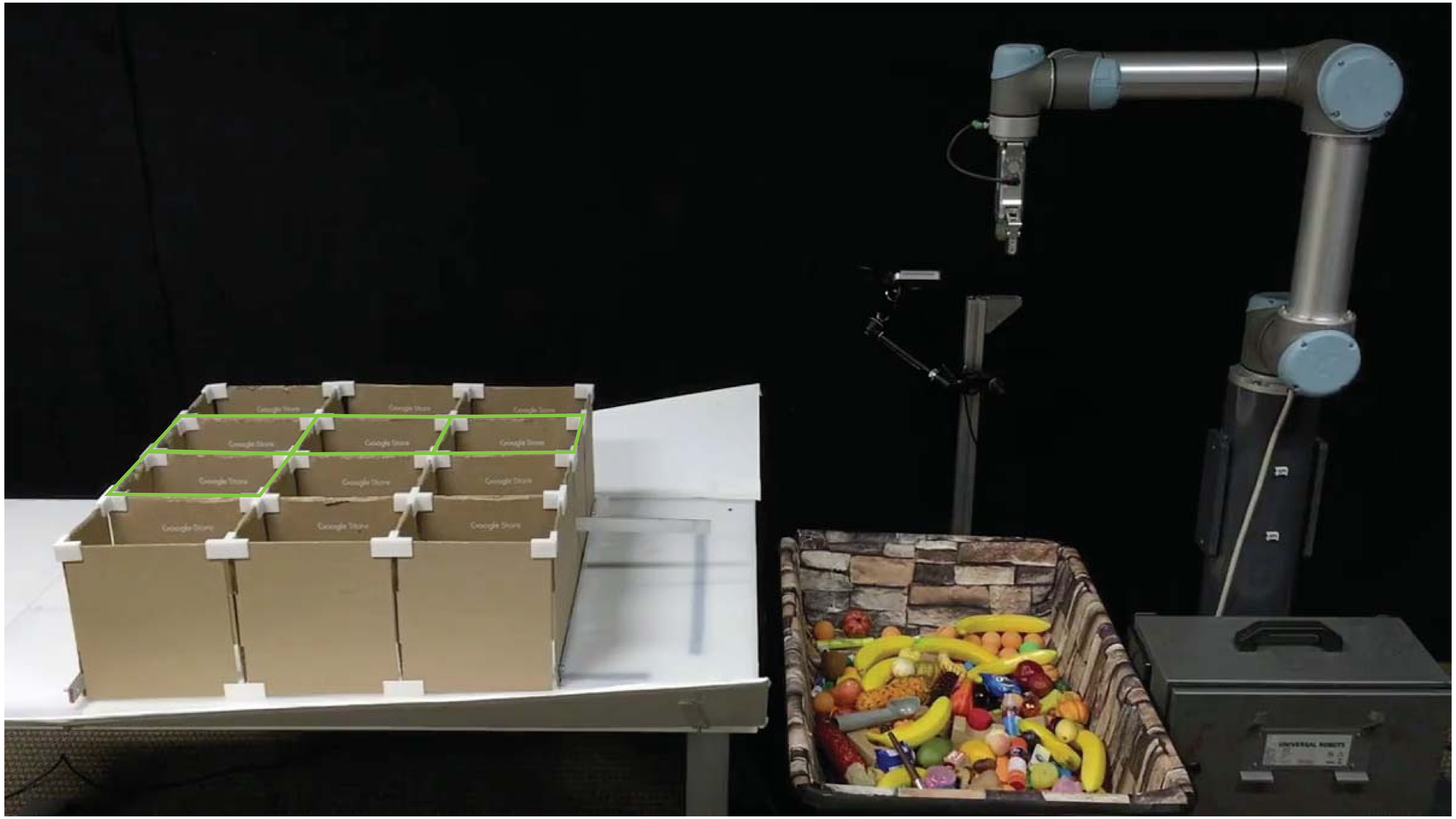


How about using other Type of Actions?

Diverse Action Types Helps!



TossingBot



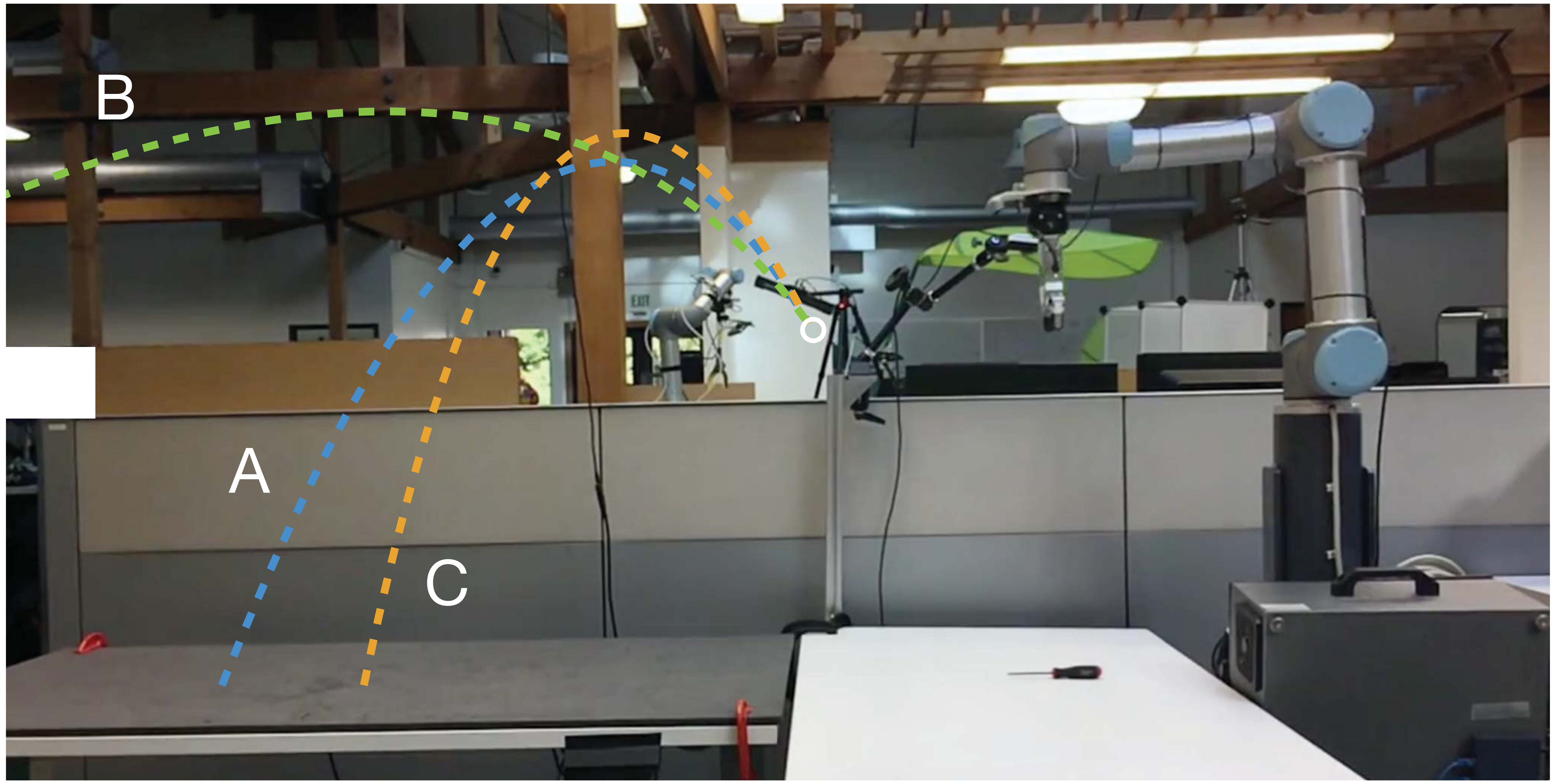
Side View

TossingBot: Learning to Throw Arbitrary Objects with Residual Physics

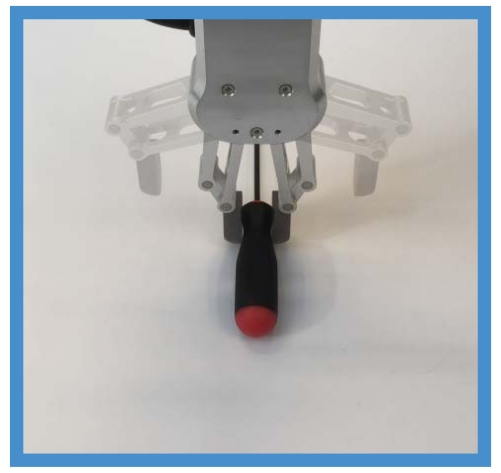
A. Zeng, S. Song, J. Lee, A. Rodriguez, T. Funkhouser

RSS Best System Paper. TR-O Best Paper

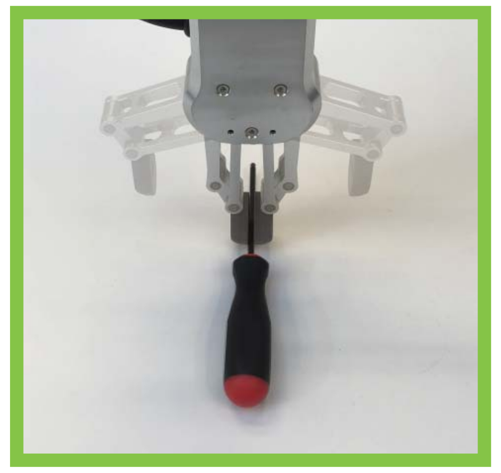
What does TossingBot learn?



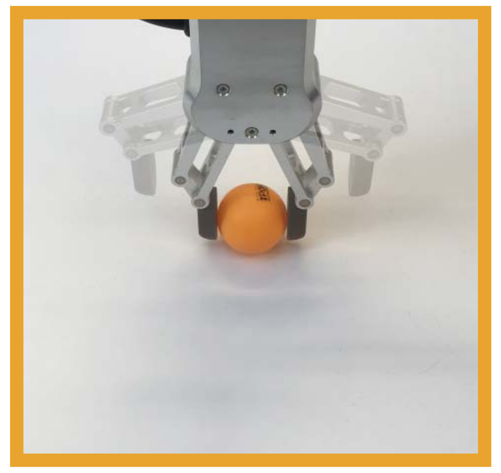
A



B



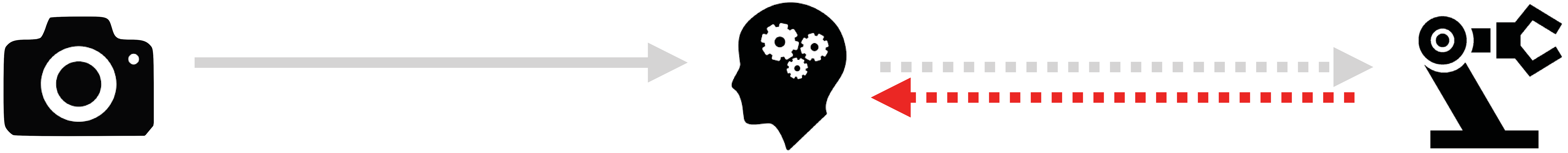
C



Mass Distribution

Varying Dynamics

Active Scene Understanding



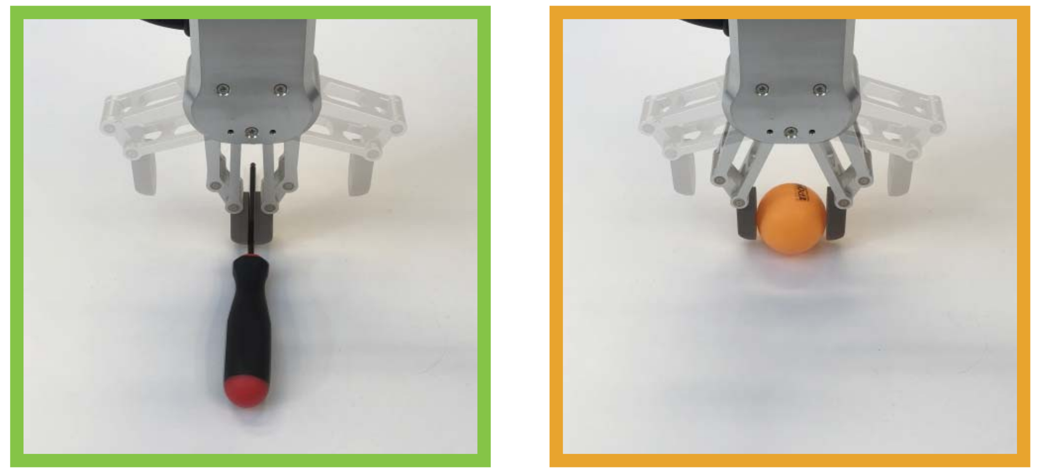
Sensor

Representation

Action

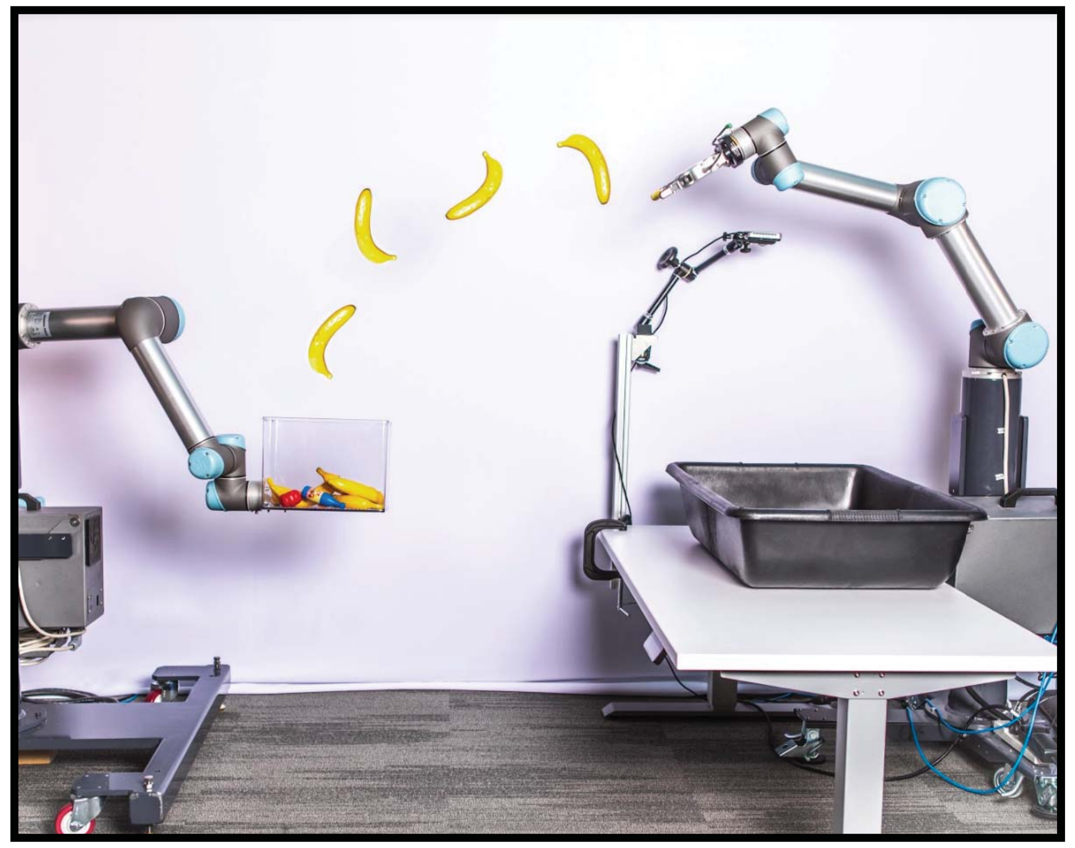
Mass
Friction

Sliding
Colliding



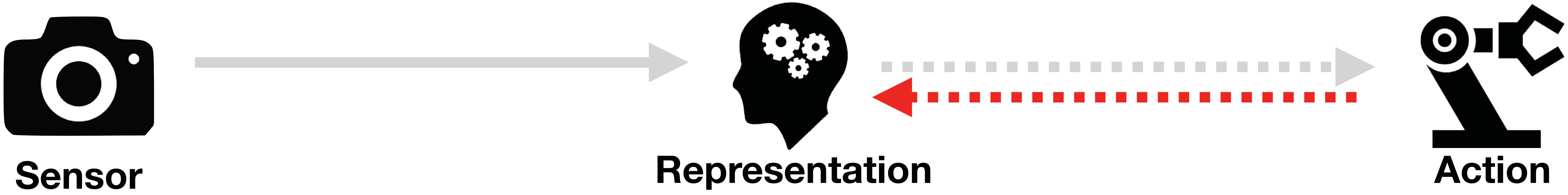
Mass distribution
Aerodynamics

Accurate Throwing



Throwing

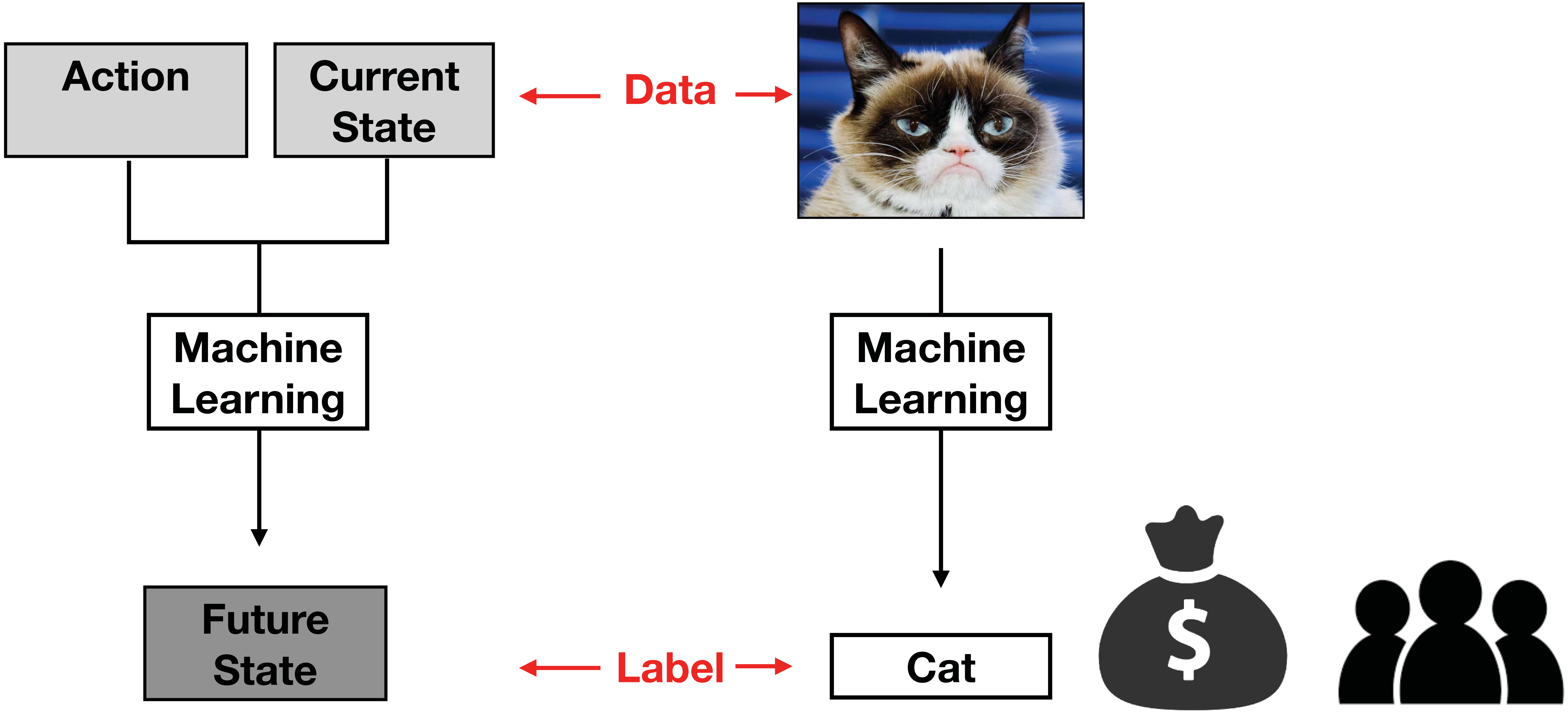
Active Scene Understanding



1. Obtain additional observations that hard to obtain passively
2. Discover objects physical properties beyond visual appearance
3. Provide opportunities for self-supervised learning

Self-Supervised Learning

Why it matters?



Action

Current State

Data

Machine Learning

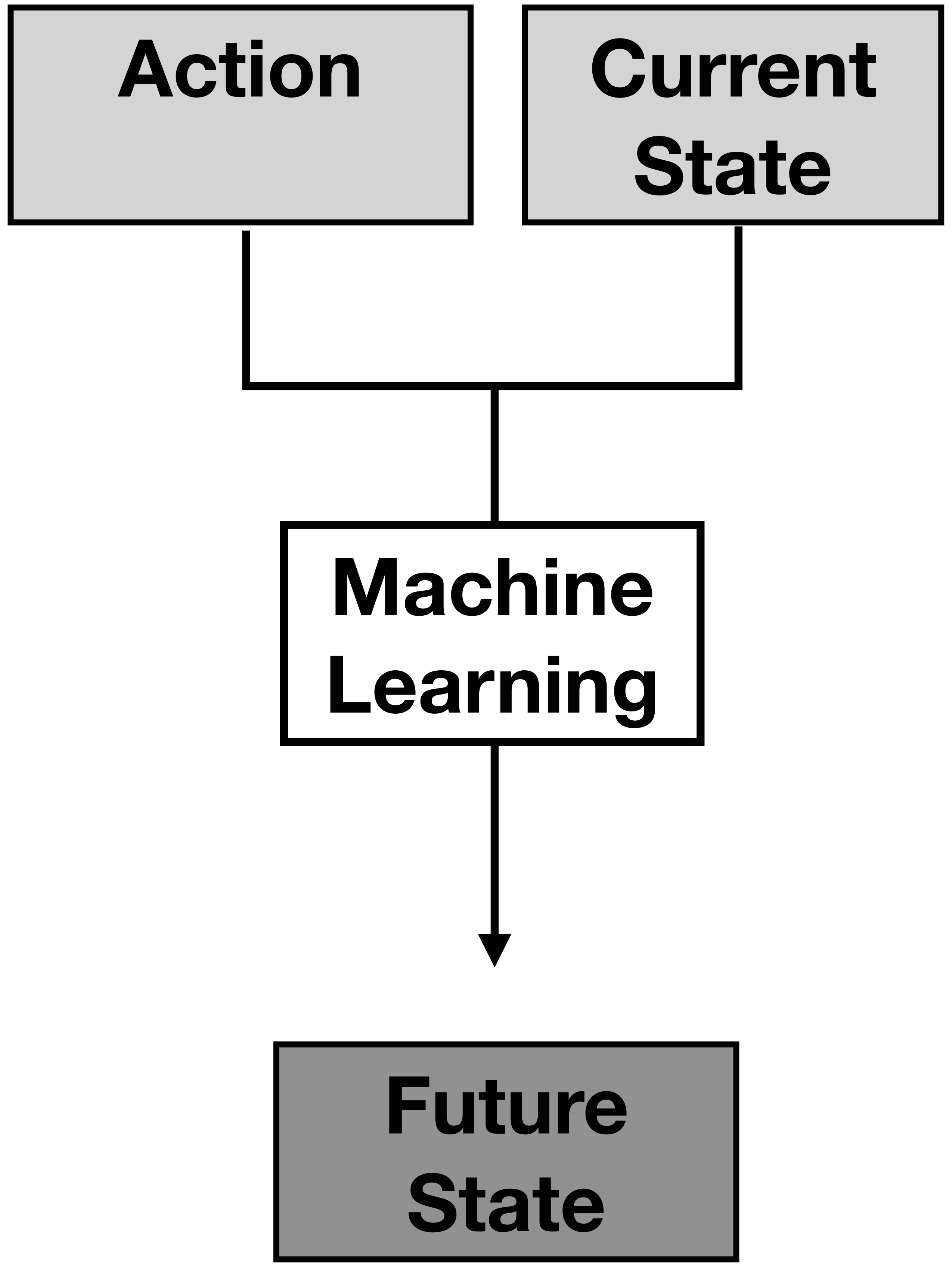
Future State

Label

Machine Learning

Cat

Self-Supervised Learning



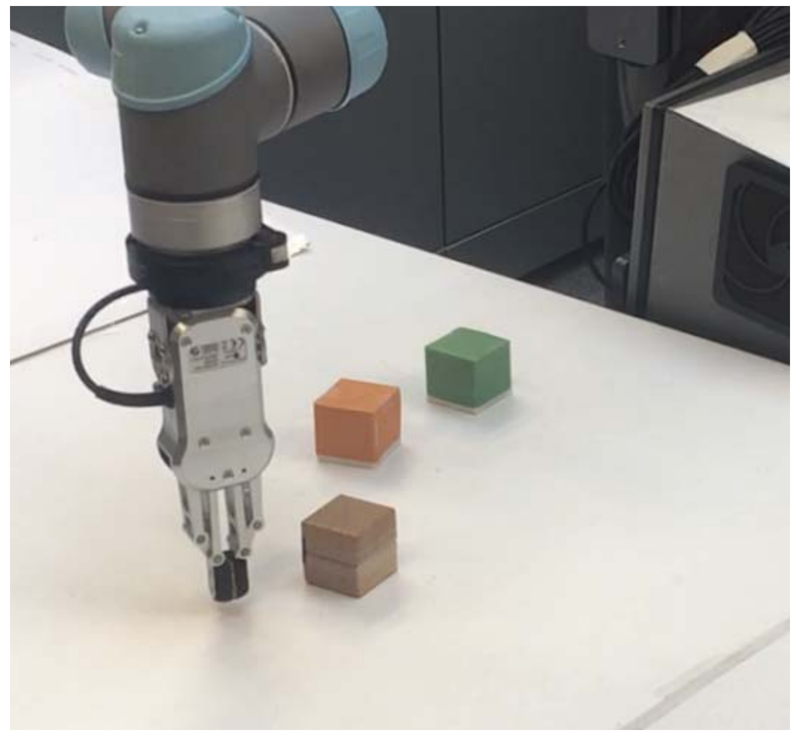
Action

Current State

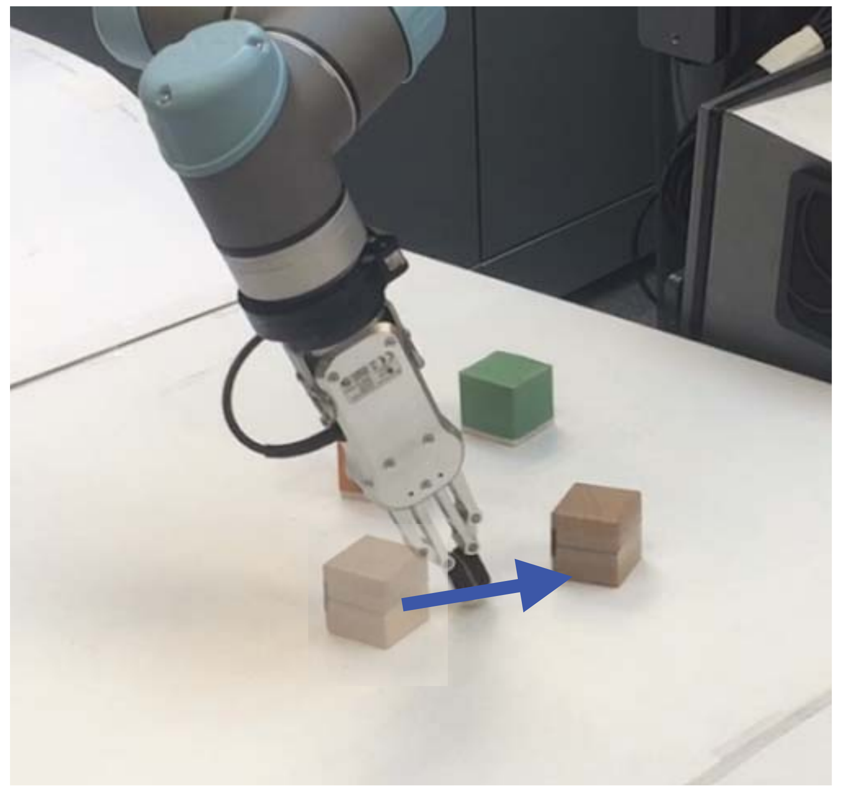
Machine Learning

Future State

DensePhysNet

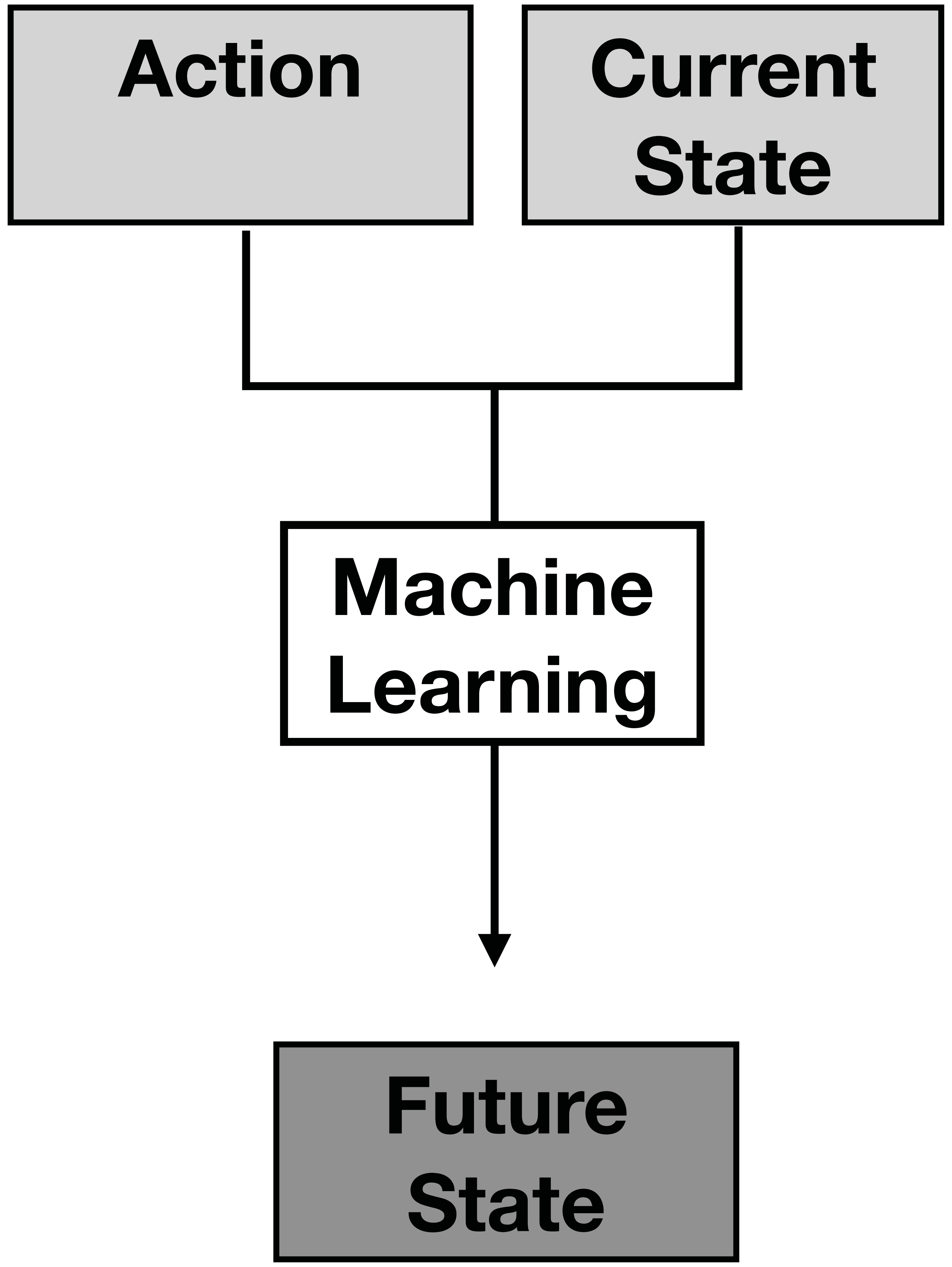


Push



Object after push

Self-Supervised Learning

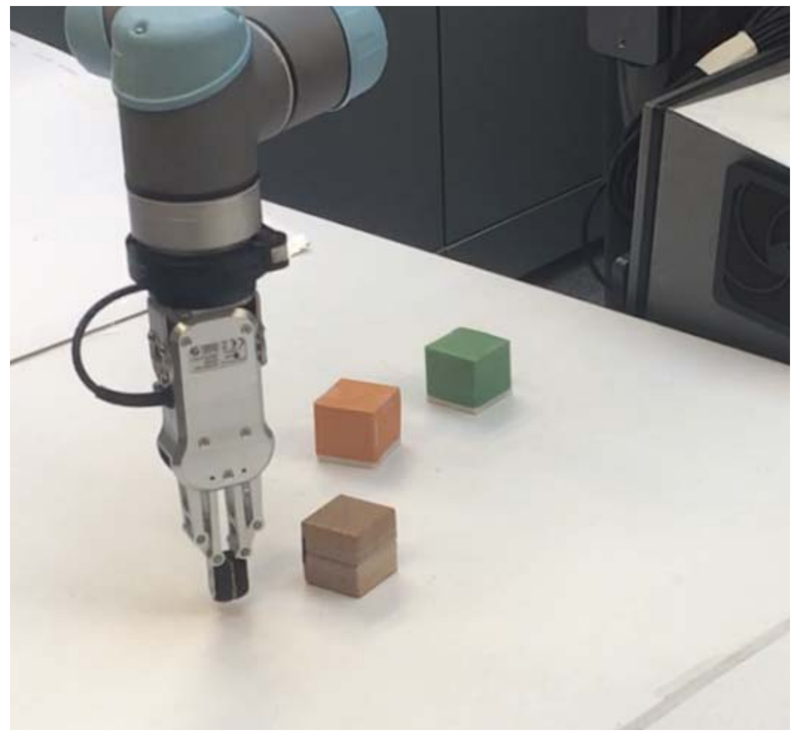

Action

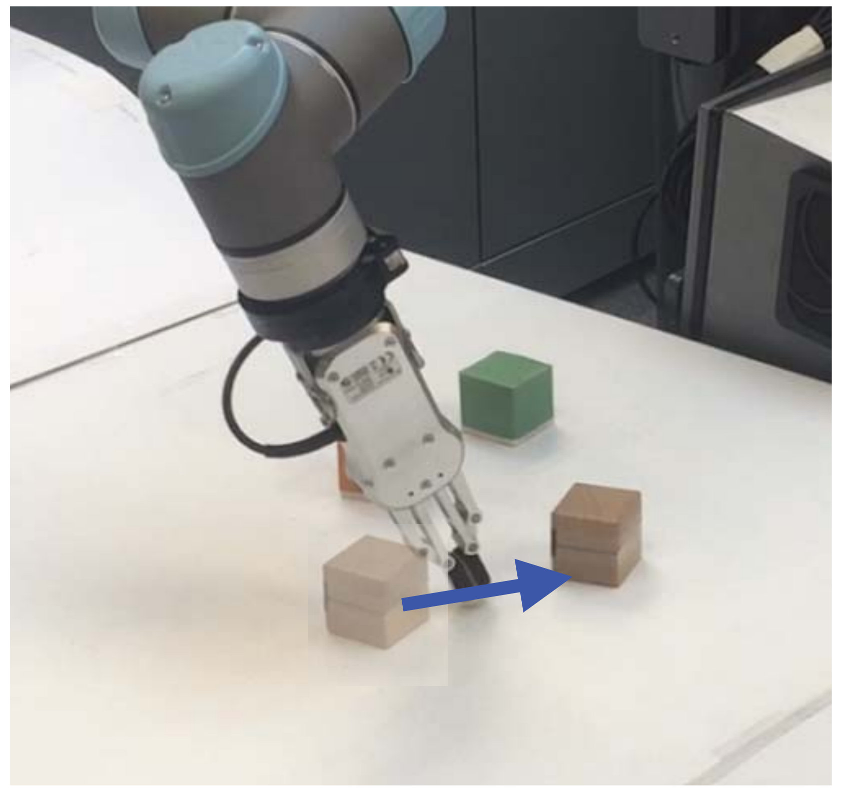
Current State

Machine Learning

Future State

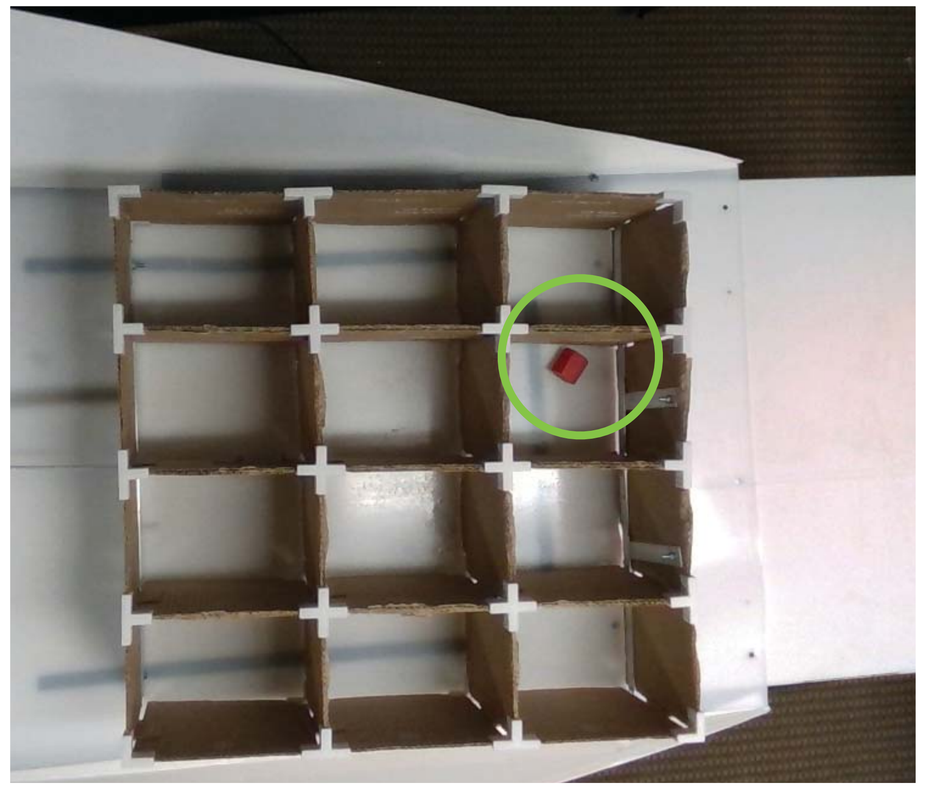
DensePhysNet


Push


Object after push

TossingBot

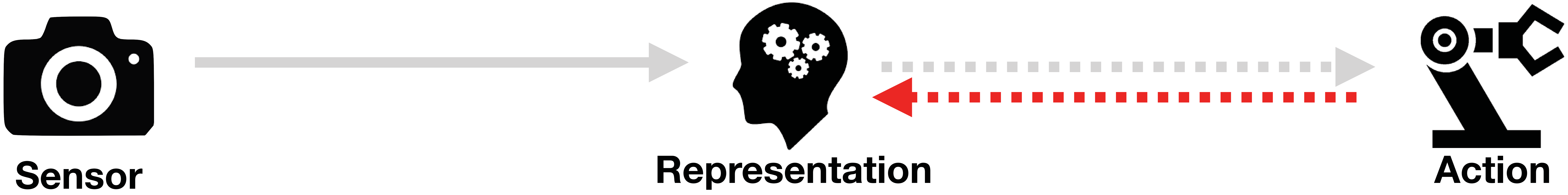
Tossing


Object landing location

State Reset

Continuously gather training data without human intervention.

Active Scene Understanding



Sensor

Representation

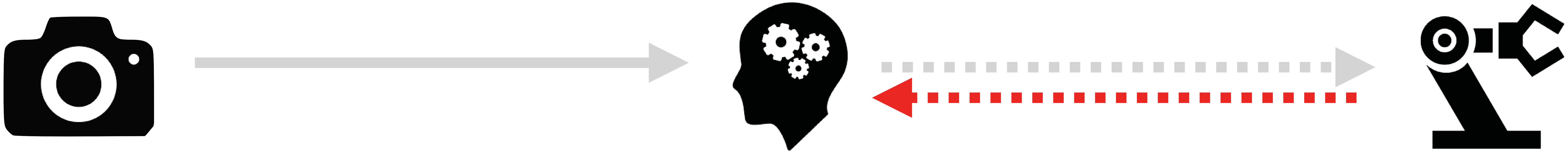
Action

1. Obtain additional observations that hard to obtain passively

2. Discover objects physical properties beyond visual appearance

3. Provide opportunities for self-supervised learning

Active Scene Understanding



Sensor

Representation

Action

Active Explorers

Passive Observers