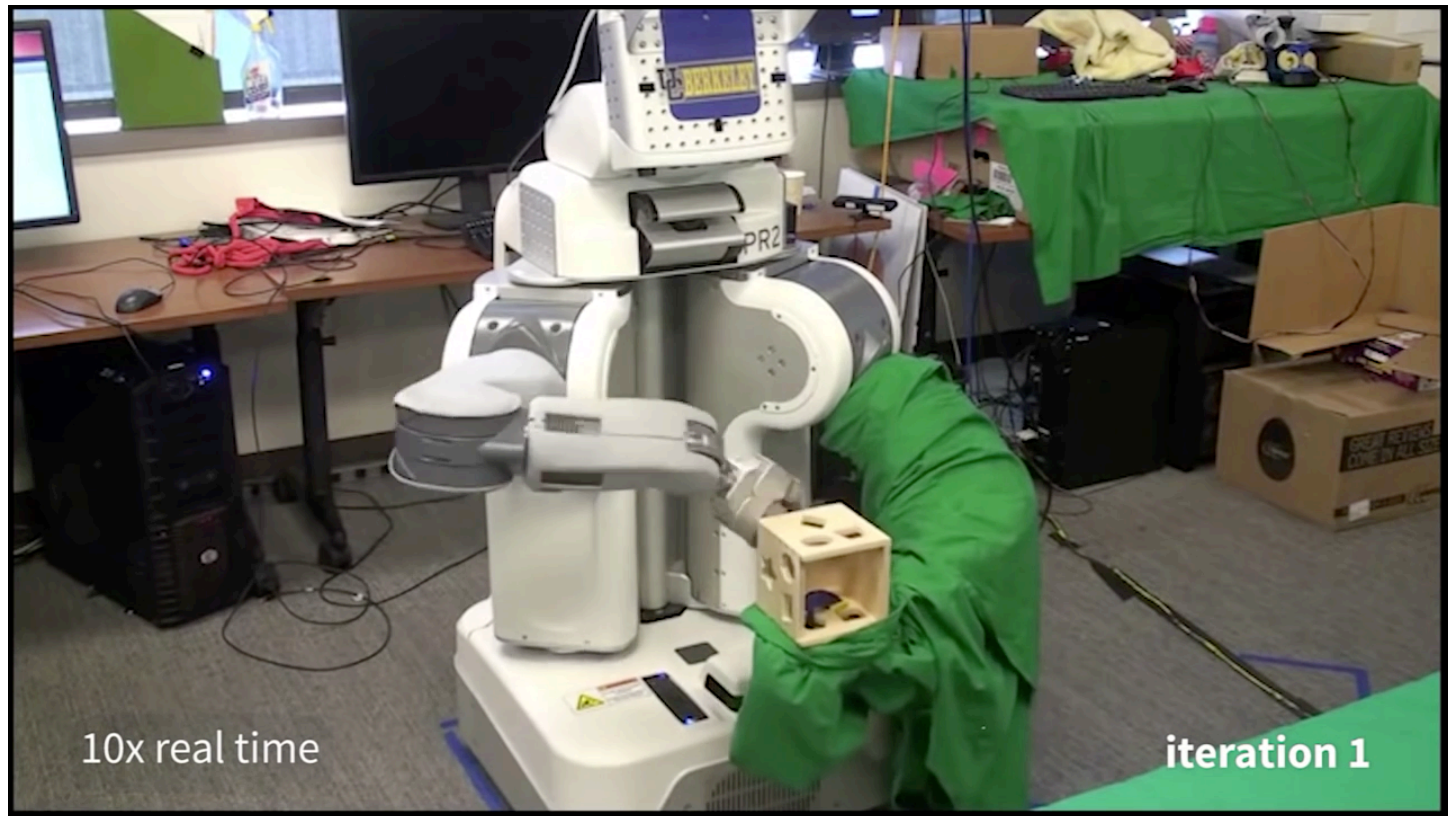
Robot Learning by Understanding Egocentric Videos

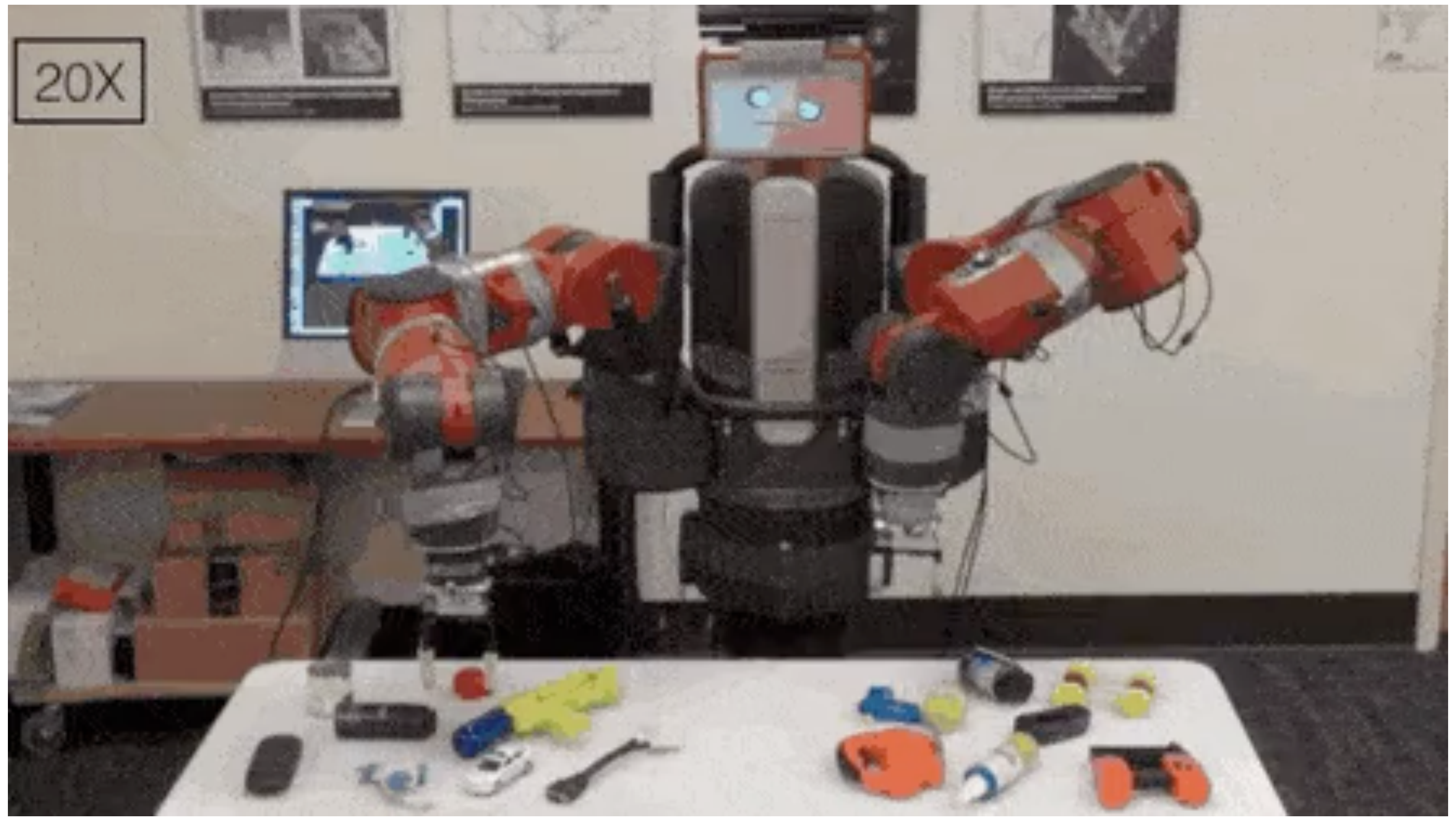
Saurabh Gupta
UIUC

Learning in Computer Vision / NLP

All the text on the Internet

Policy Learning in Robotics

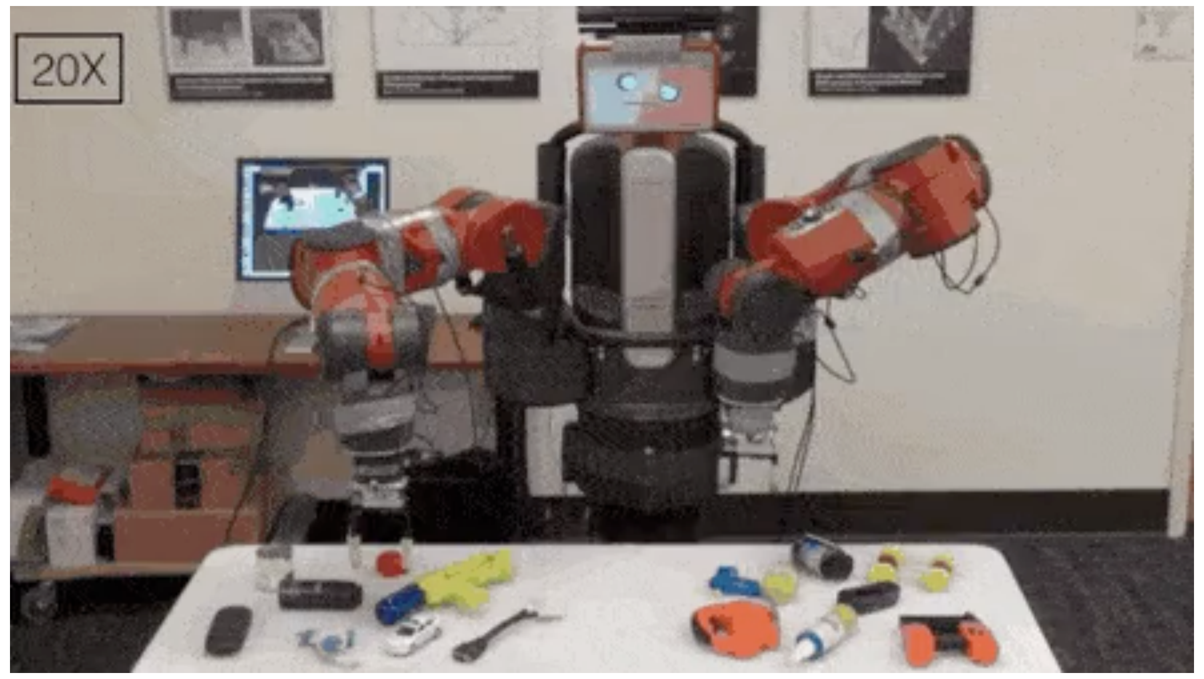




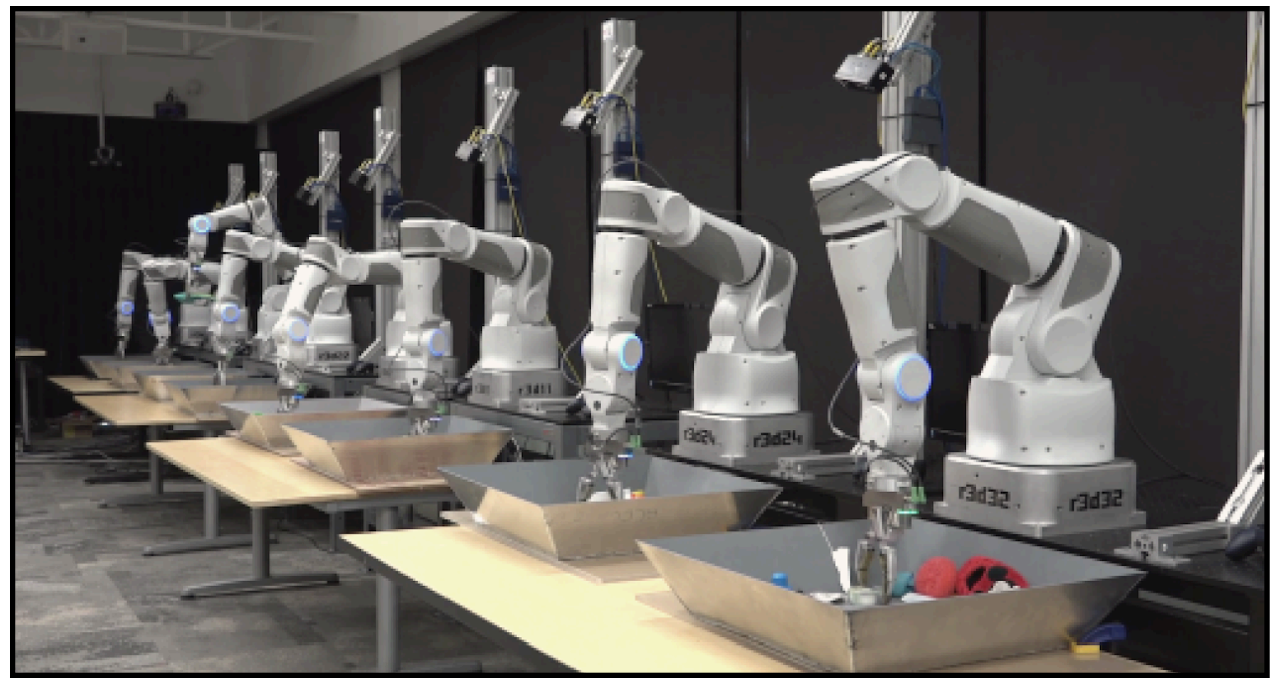
How do we scale up learning for robotics?

Scaling up Learning in Robotics

Many Answers

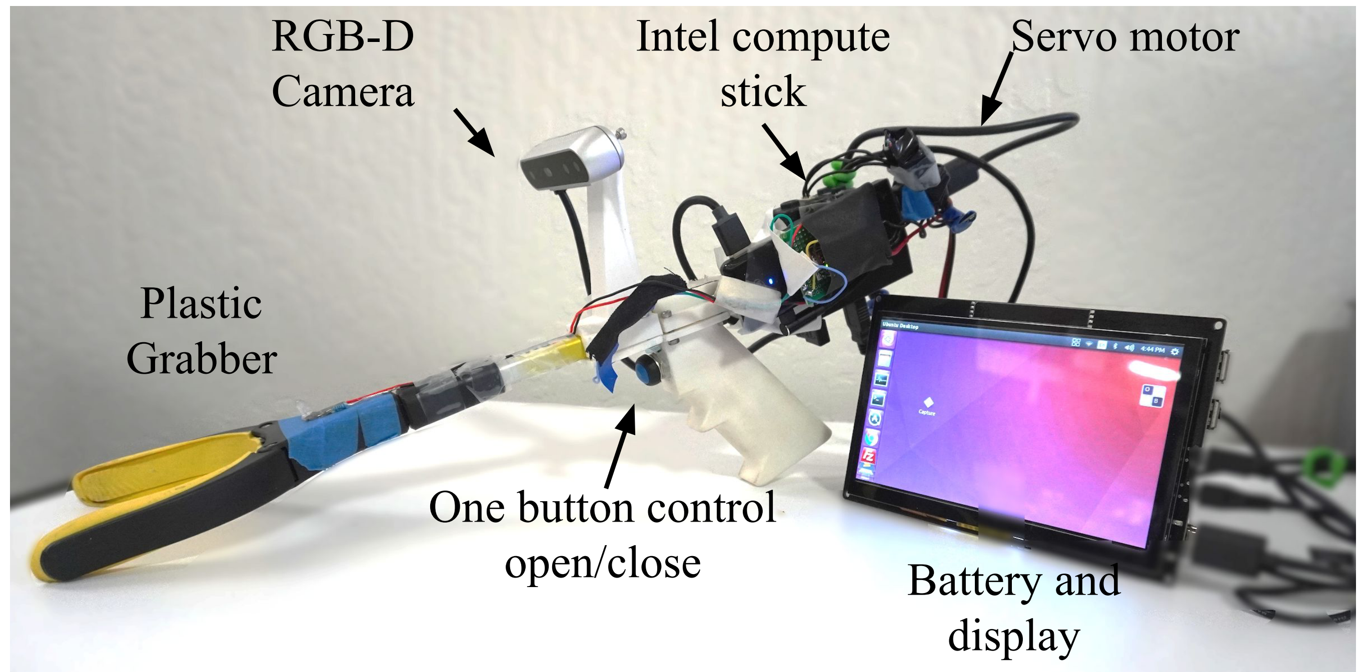


Self-supervision
[Pinto et al.]

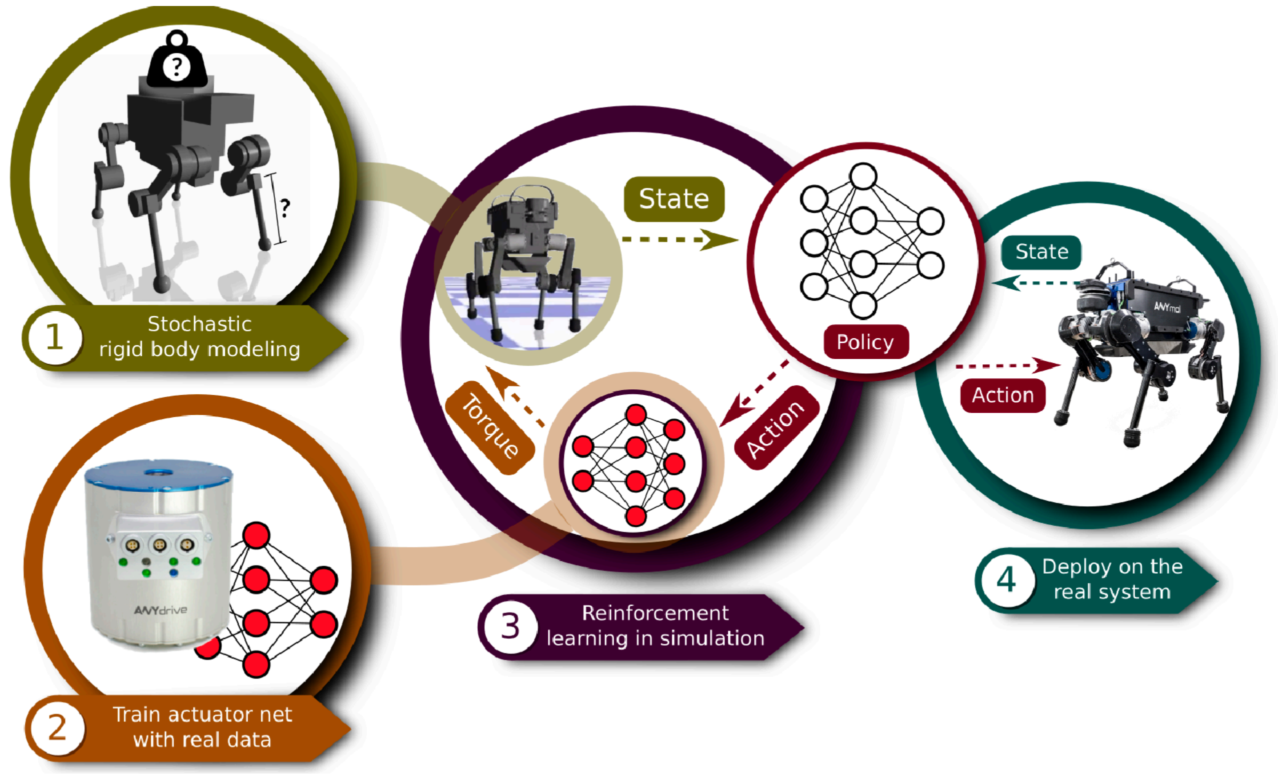


Arm farms
[Levine et al.]

Robot in homes
[Gupta et al.]



Simplifying data collection [Song et al.]



Sim2Real
[Hwangbo et al.]

But today, scaling up robot learning through observation of other agents solving tasks.

- We do it as adults
- Critical part of child development [1]:
  - Early imitation in children, as young as a few hours / days

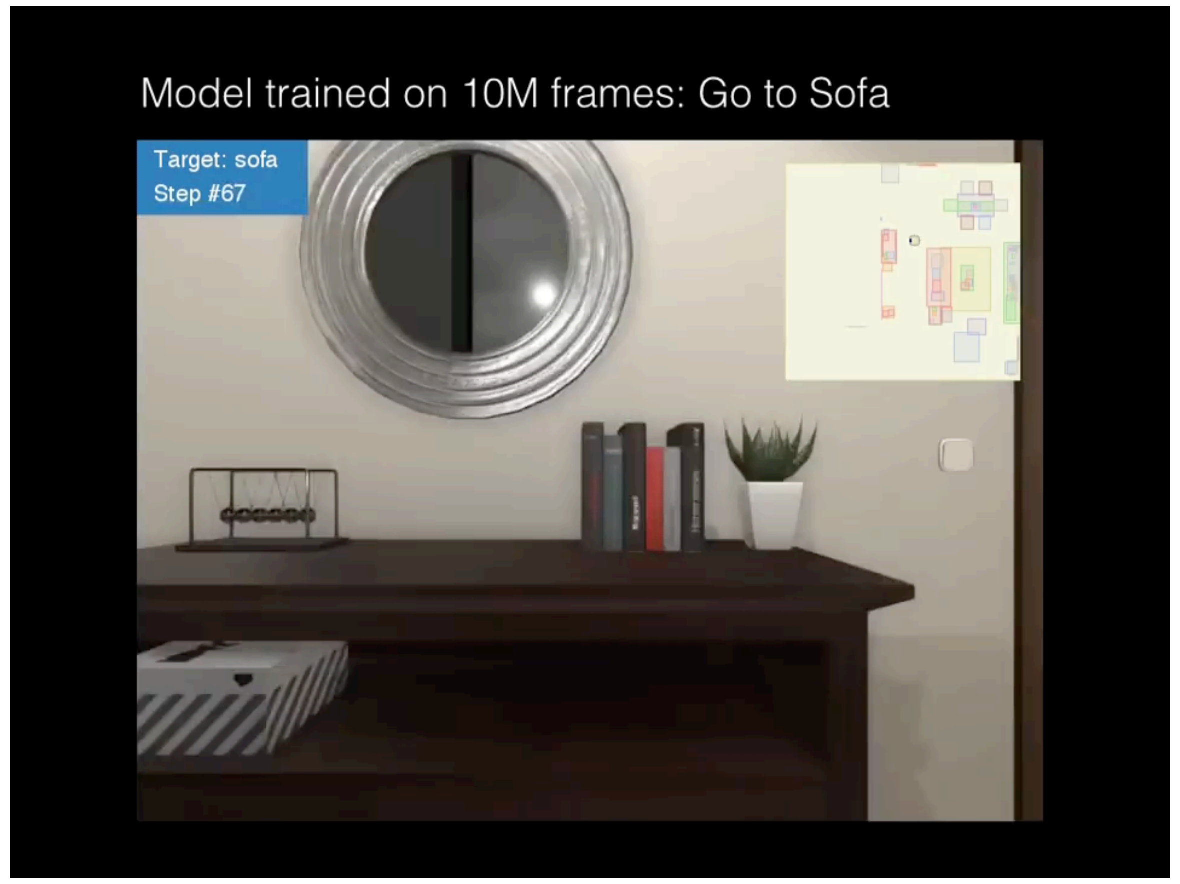
[1] Andrew Meltzoff and Alison Gopnik. The role of imitation in understanding persons and developing a theory of mind.

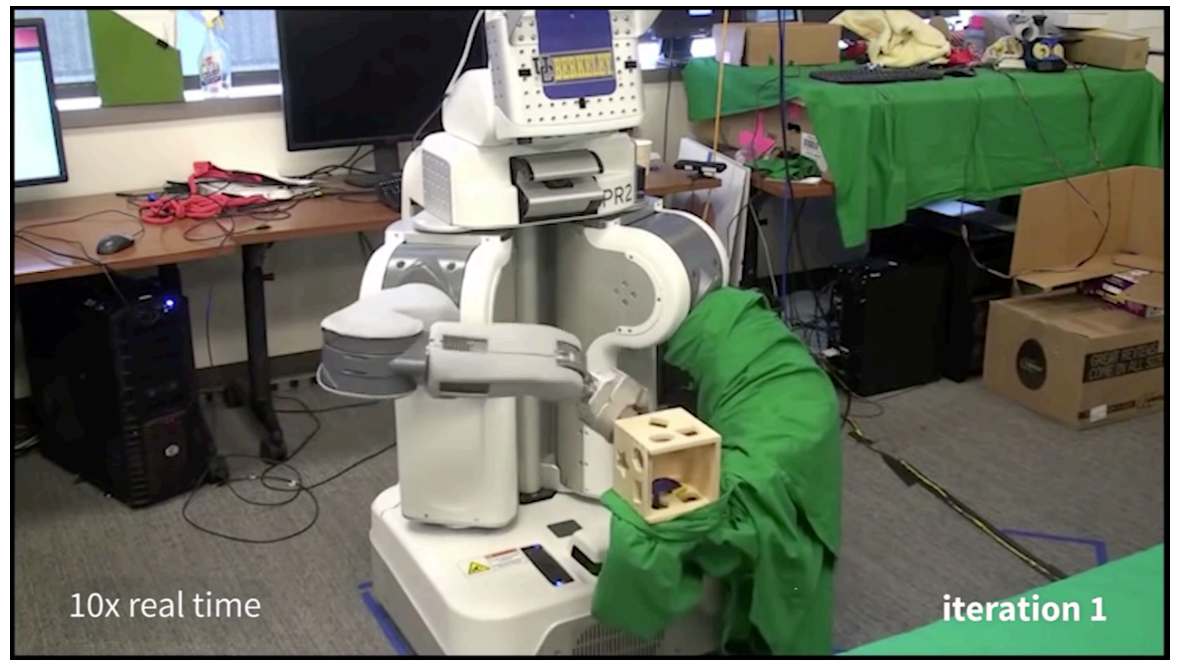
In particular, we will focus on egocentric videos

Why would such videos be useful for robot learning, and how can we use them?

Motivation

Policy Learning from Interaction





- Challenging to specify reward functions
- Impractically large sample complexity
- Learning signal derived solely from interaction
- Poor generalization due to lack of visual diversity in training, sim2real transfer

How can egocentric videos aid?

- Large diversity may provide good generalization.
- Demonstrations may directly show how to solve long horizon tasks.
- Depict what the world is like, and how it works.

Motivation

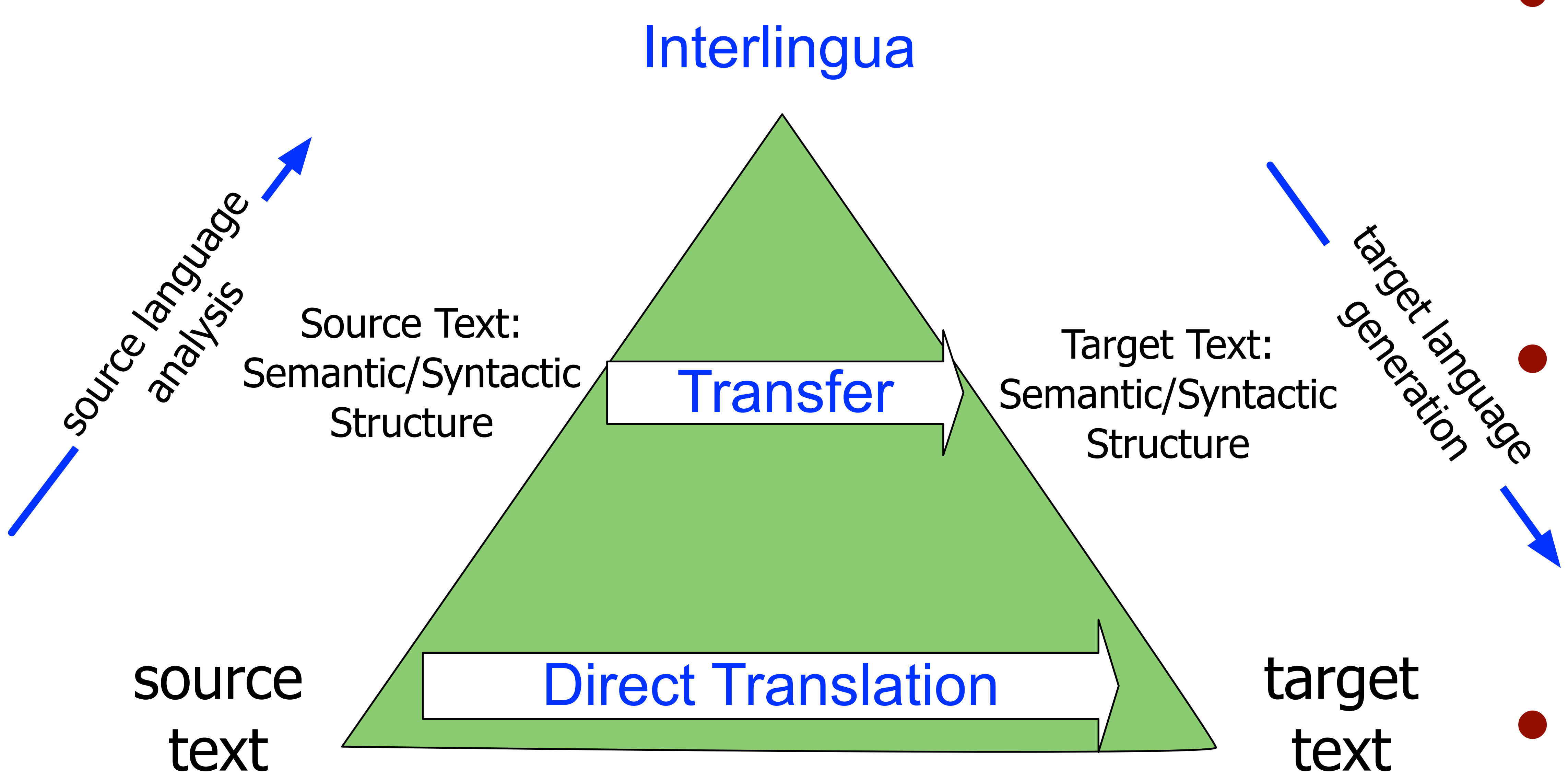
How can egocentric videos aid?

- Large diversity may provide good generalization.
- Demonstrations may directly show how to solve long horizon tasks.
- Depict what the world is like, and how it works.

However,

- Videos don't come with action labels
- Goals and intents are not known
- Depicted trajectories may be sub-optimal
- Embodiment gap (sensors / actions / capabilities)
- Only showcase positive data

Learning at different abstraction levels



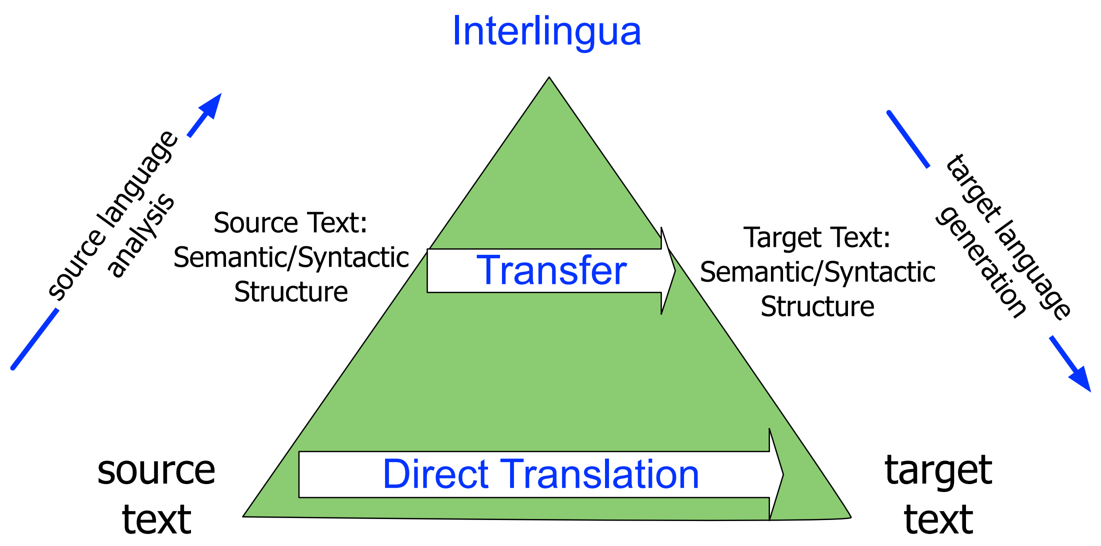
The Vauquois triangle for machine translation

- *Interlingua* (text to abstract meaning space, back to text)
- *Transfer* (text to parse tree, transform tree, generate text)
- *Direct Translation* (translate word-by-word or phrase-by-phrase)

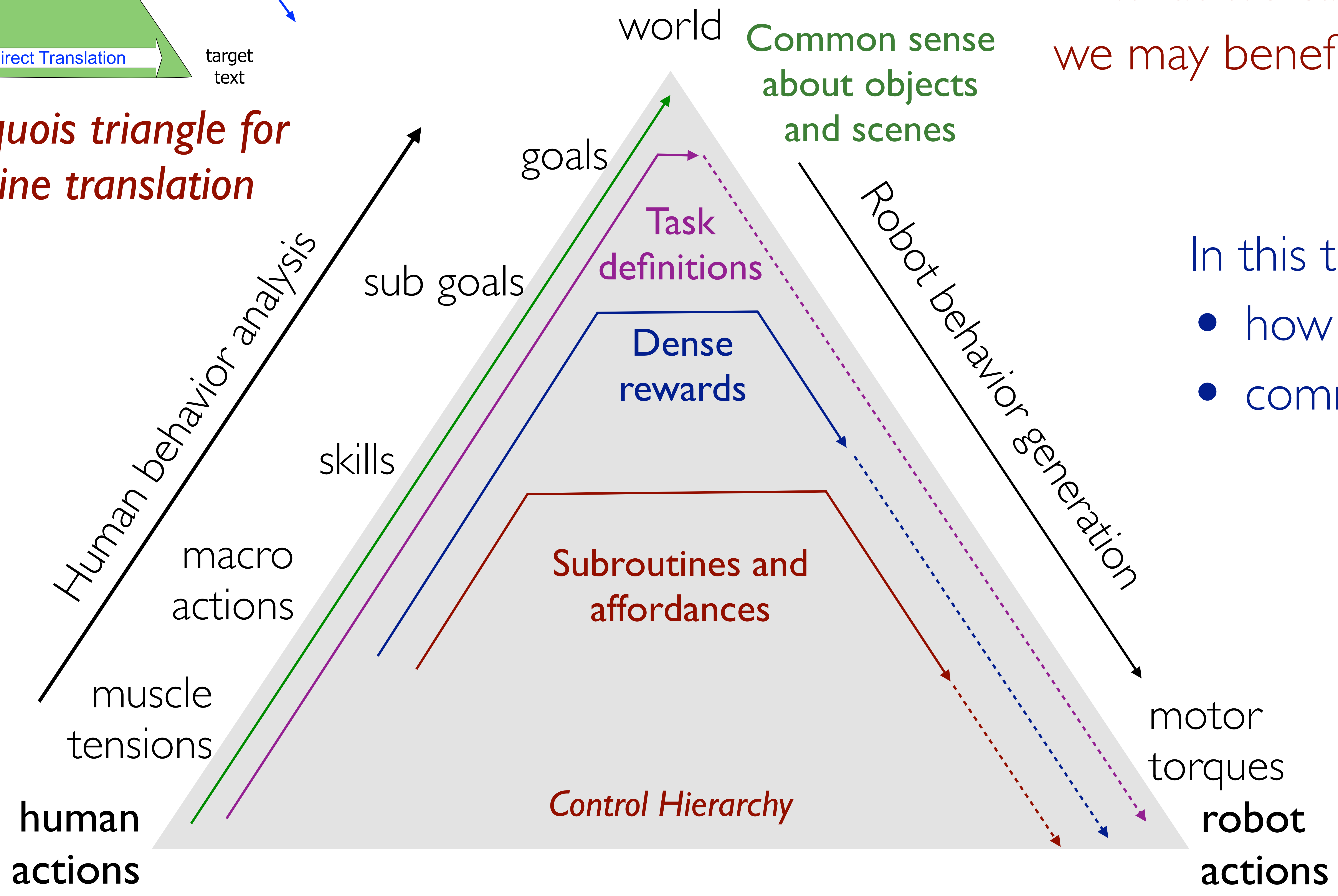
- *Increasing depth of understanding for analysis and generation*
- *Decreasing amount of transfer knowledge needed*

Figure from Jurafsky and Martin

Learning at different abstraction levels



The Vauquois triangle for machine translation



Depending on the amount of gap between:

- goals,
- embodiment,
- what we can observe in videos

we may benefit from transfer at different levels.

In this talk, using video to learn,

- how to interact with objects
- common sense about scenes

Human Hands as Probes for Interactive Object Understanding

Mohit Goyal Sahil Modi Rishabh Goyal Saurabh Gupta

CVPR 2022

Mohit Goyal

Sahil Modi

Rishabh Goyal

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Interactive Object Understanding

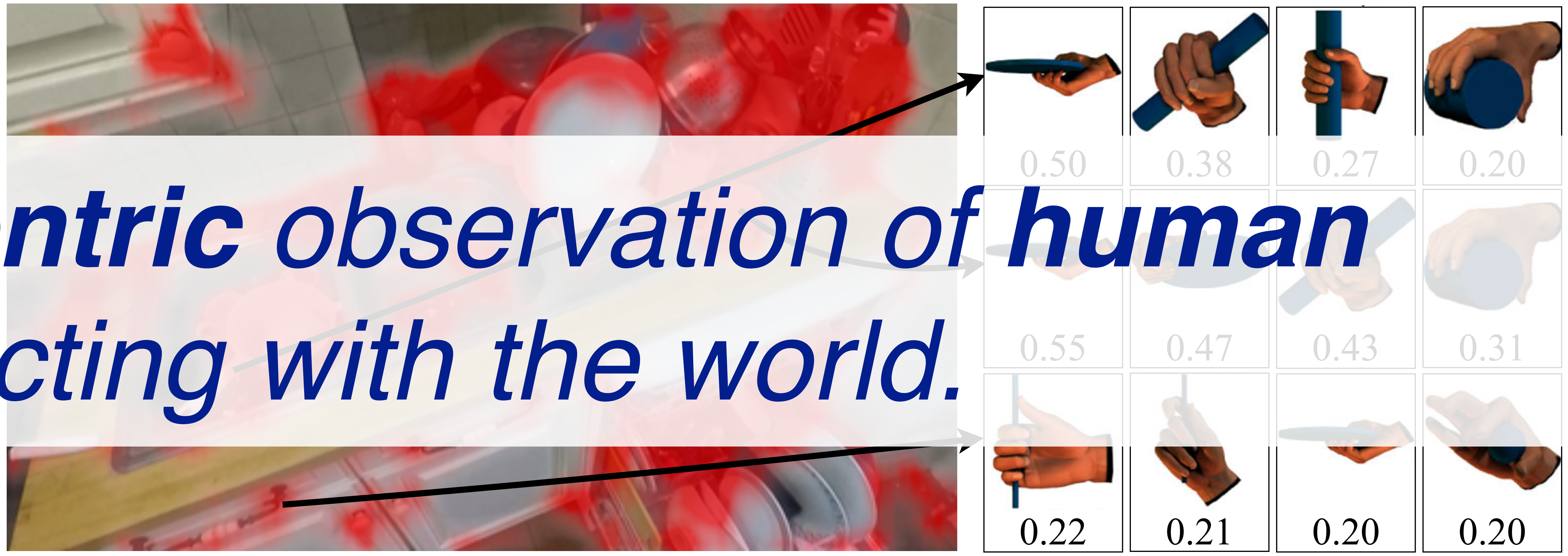
A) Which sites can we interact at?
(cupboard handles)

B) How to interact with those sites?
(using adducted thumb grasp)

C) What happens when we do?
(cupboard undergoes state transition)

Learn through ***egocentric*** *observation of* ***human hands*** *interacting with the world.*

1) State Sensitive Features (C)

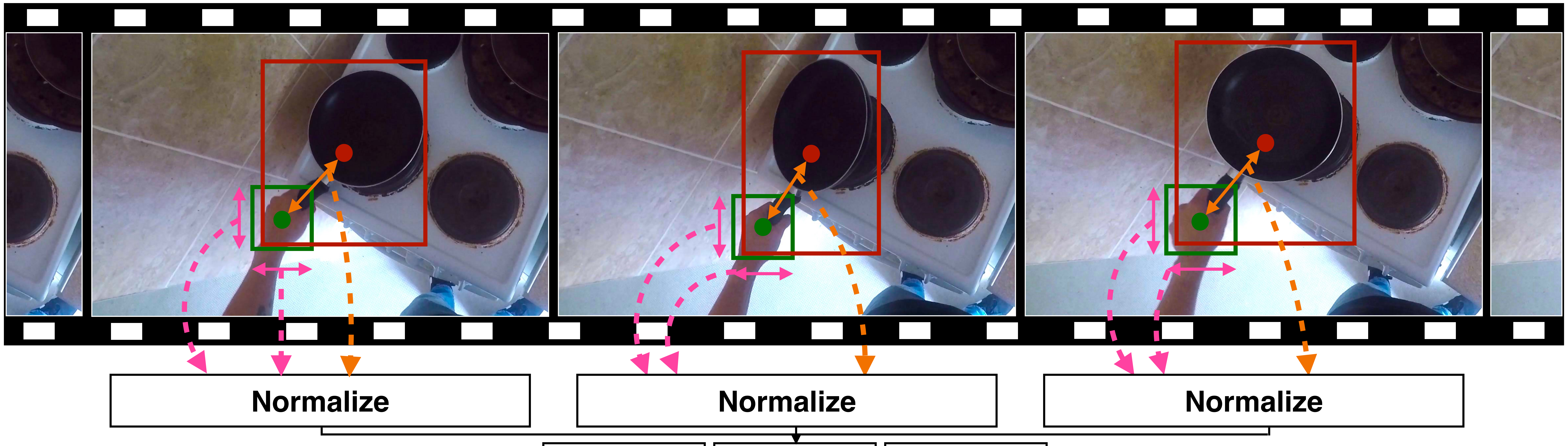


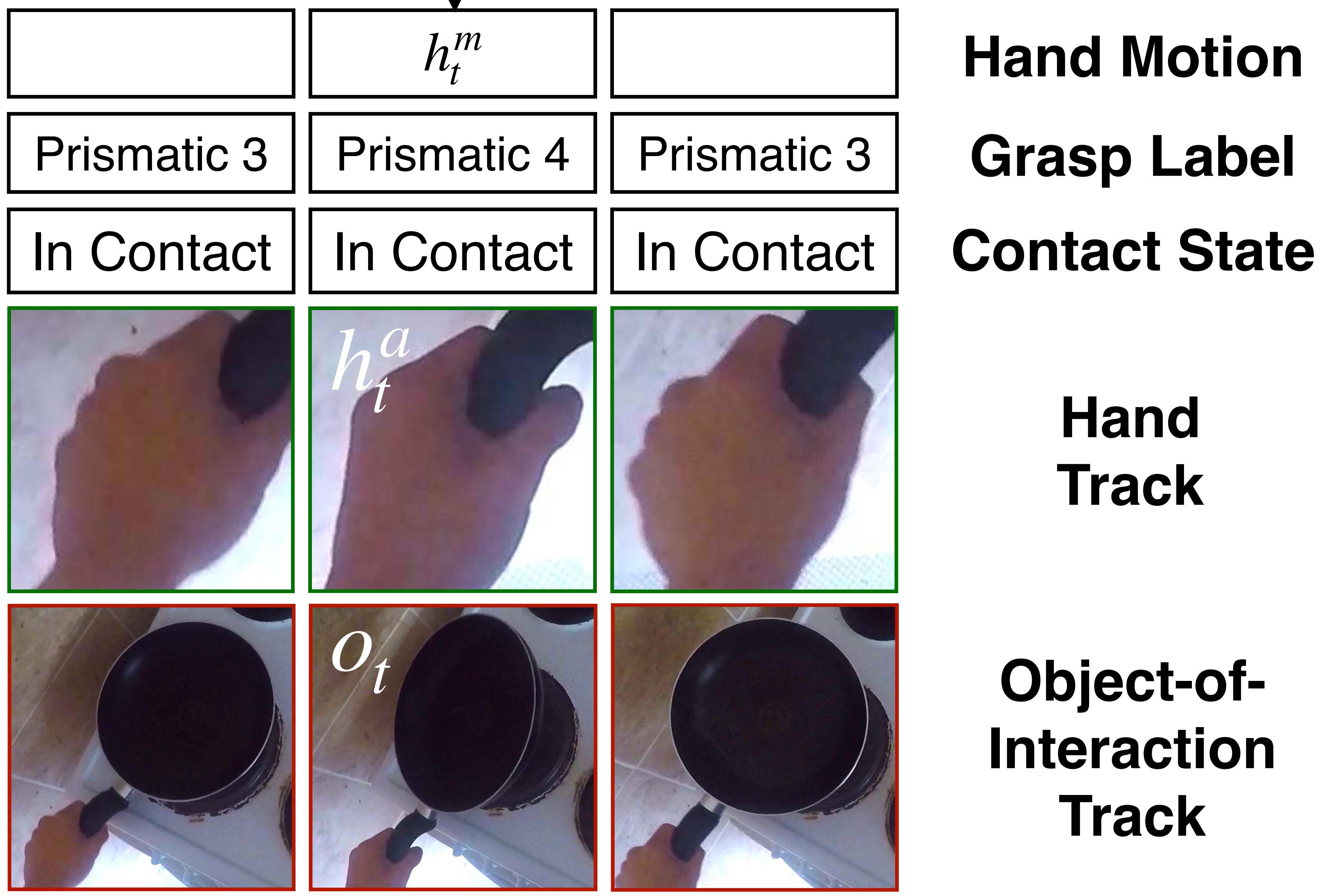
2) Object Affordance Prediction (A,B)

Human Hands in Egocentric Videos are Informative

1. In-the-wild egocentric videos focus upon natural ways of hand-object interaction.
2. Attending to hands localizes and stabilizes active objects.
3. Hands show where all we can interact in the scene.
4. Analyzing hands reveals information about objects: their state and how to interact with them.

Data Preparation using Off-the-shelf models





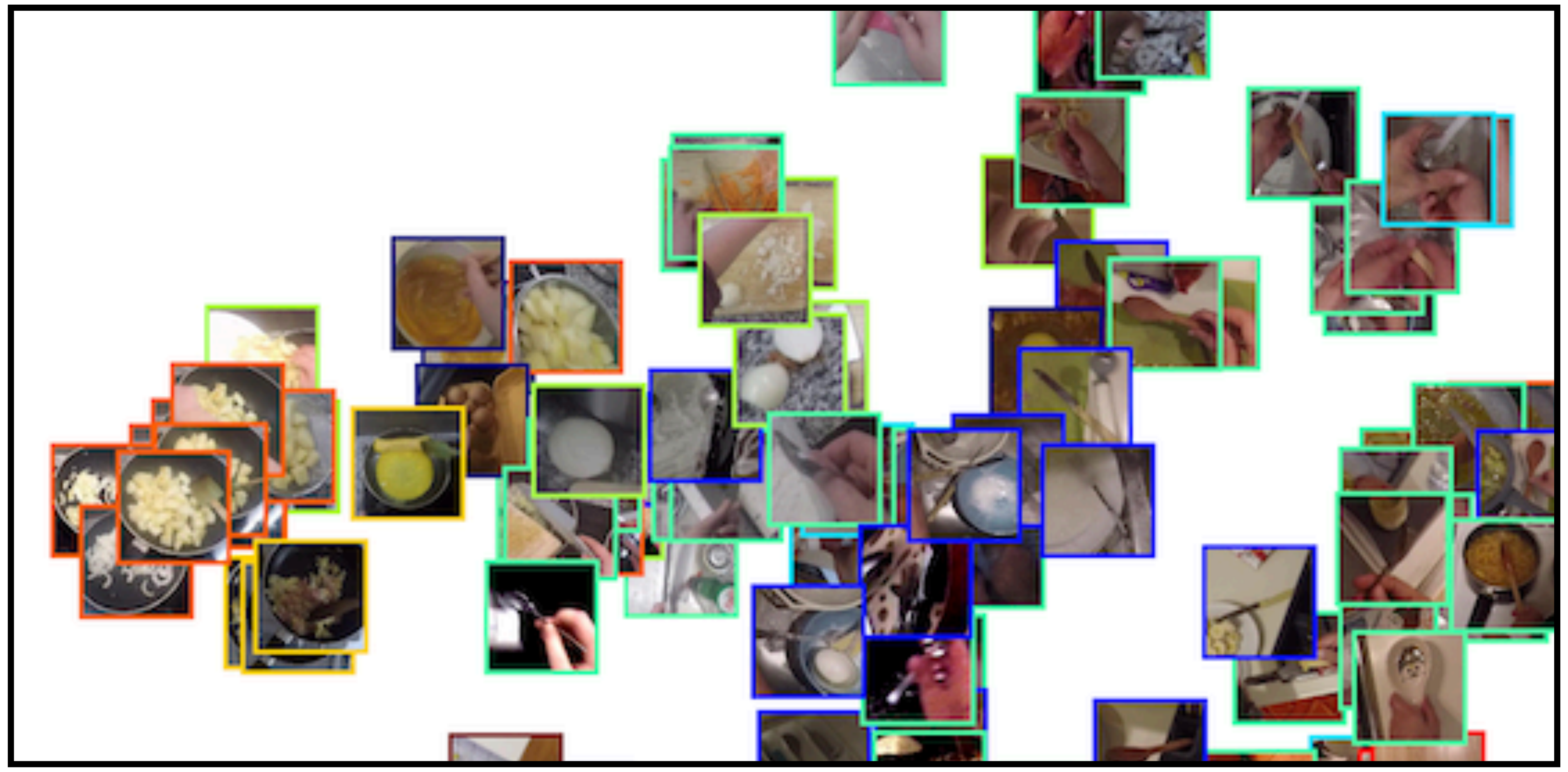
Shan et al. CVPR 2020.
Understanding Human Hands in Contact at Internet Scale.

Interactive Object Understanding

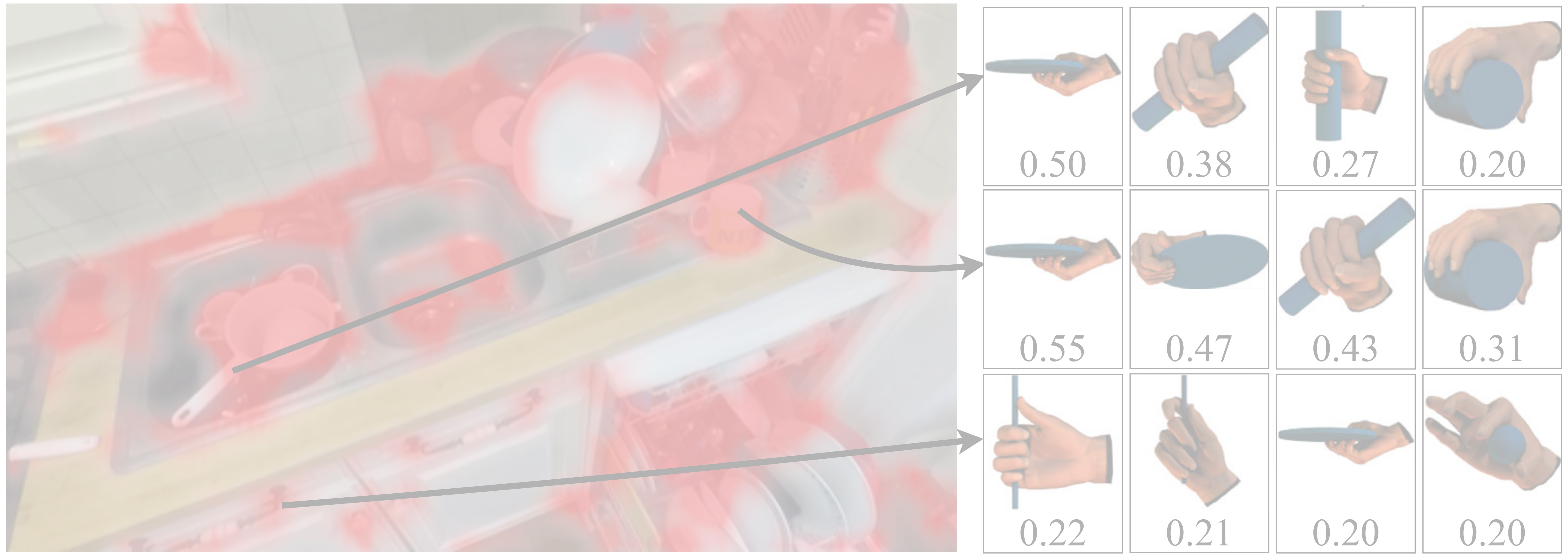
A) Which sites can we interact at?
(cupboard handles)

B) How to interact with those sites?
(using adducted thumb grasp)

C) What happens when we do?
(cupboard undergoes state transition)



1) State Sensitive Features (C)



2) Object Affordance Prediction (A,B)

Task 1. Learning State Sensitive Features: Approach

Temporal SimCLR with Object-Hand Consistency (TSC + OHC)

1. Leverage Temporal Consistency in States

Task 1. Learning State Sensitive Features: Approach

Temporal SimCLR with Object-Hand Consistency (TSC + OHC)

1. Leverage Temporal Consistency in States

2. Using Object-Hand Consistency: Similarity in states through similarity in interaction

Task 1. Learning State Sensitive Features: Results

Evaluation on EPIC-STATES Dataset

EPIC-STATES Evaluation (mAP)

Methods	All Objects
ImageNet Pre-trained	83.0
SimCLR [3]	79.9
EPIC Action Classification	77.9
MIT States [4] (Internet Images)	81.5
TSC (Ours)	83.6
TSC+OHC (Ours)	**84.9**

TSC improves over ImageNet features

SimCLR features perform worse

TSC improves over semantic supervision

Object-hand consistency further helps

[3] Chen et al. ICML 2020. A simple framework for contrastive learning of visual representations.
[4] Isola et al. CVPR 2015. Discovering states and transformations in image collections.

Task 1. Learning State Sensitive Features: Results

Evaluation on EPIC-STATES Dataset

EPIC-STATES Evaluation (mAP)

Methods	All Objects	Novel Objects
ImageNet Pre-trained	83.0	74.5
SimCLR [3]	79.9	74.4
EPIC Action Classification	77.9	77.0
MIT States [4] (Internet Images)	81.5	73.9
TSC (Ours)	83.6	80.2
TSC+OHC (Ours)	**84.9**	**81.8**

+1.9 %

+7.3 %

TSC improves over ImageNet features

SimCLR features perform worse

TSC improves over semantic supervision

Object-hand consistency further helps

[3] Chen et al. ICML 2020. A simple framework for contrastive learning of visual representations.

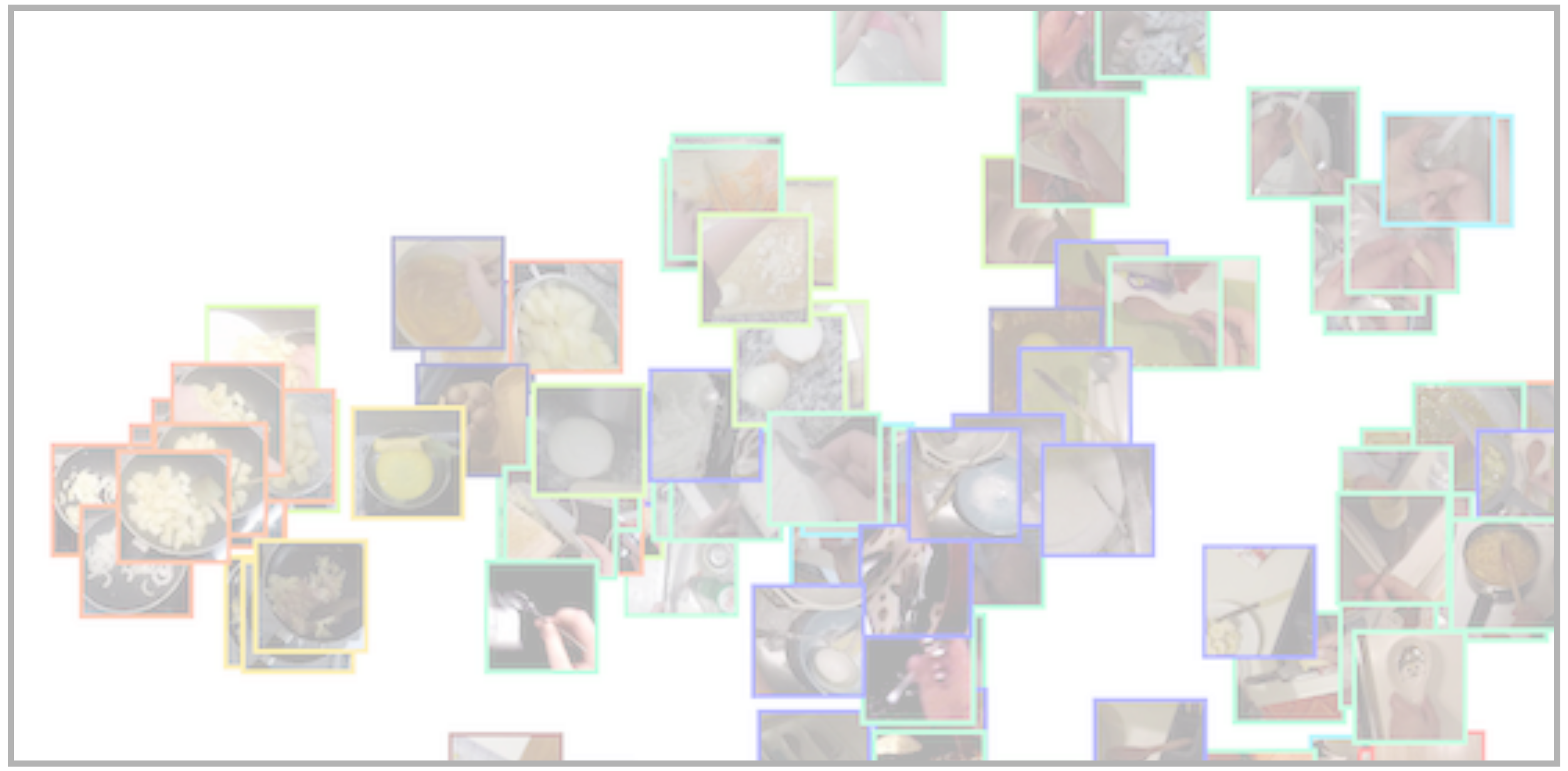
[4] Isola et al. CVPR 2015. Discovering states and transformations in image collections.

Interactive Object Understanding

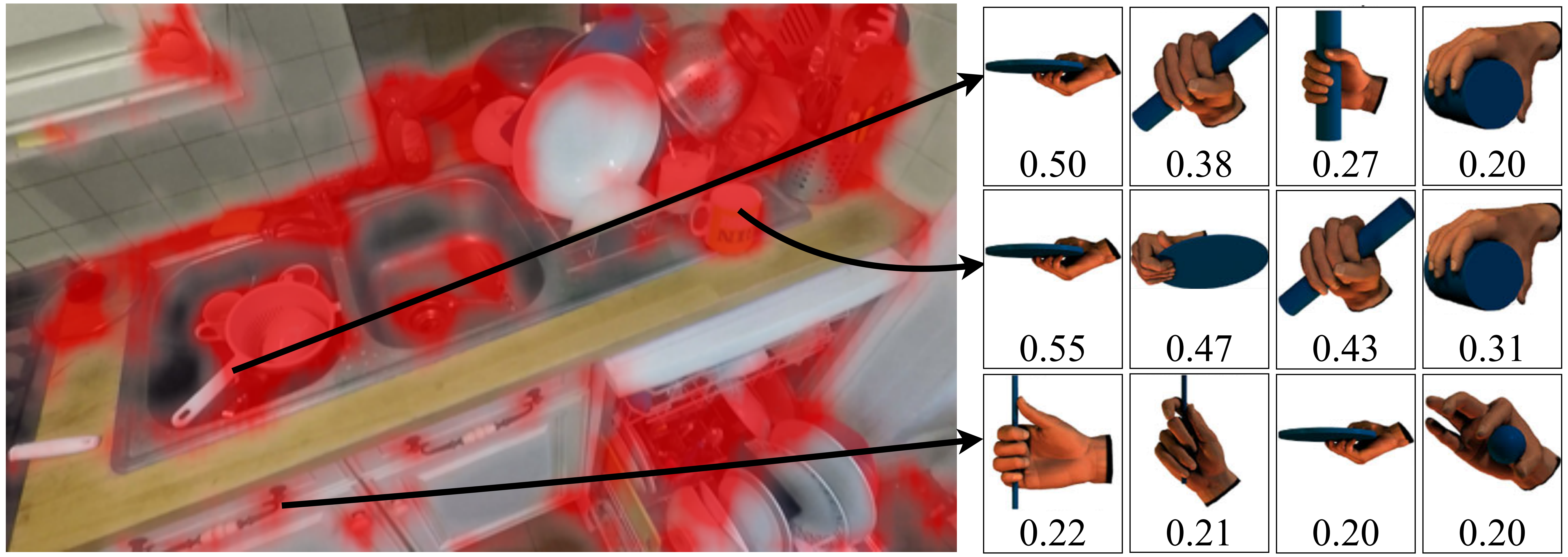
A) Which sites can we interact at? (cupboard handles)

B) How to interact with those sites? (using adducted thumb grasp)

C) What happens when we do? (the cupboard opens)



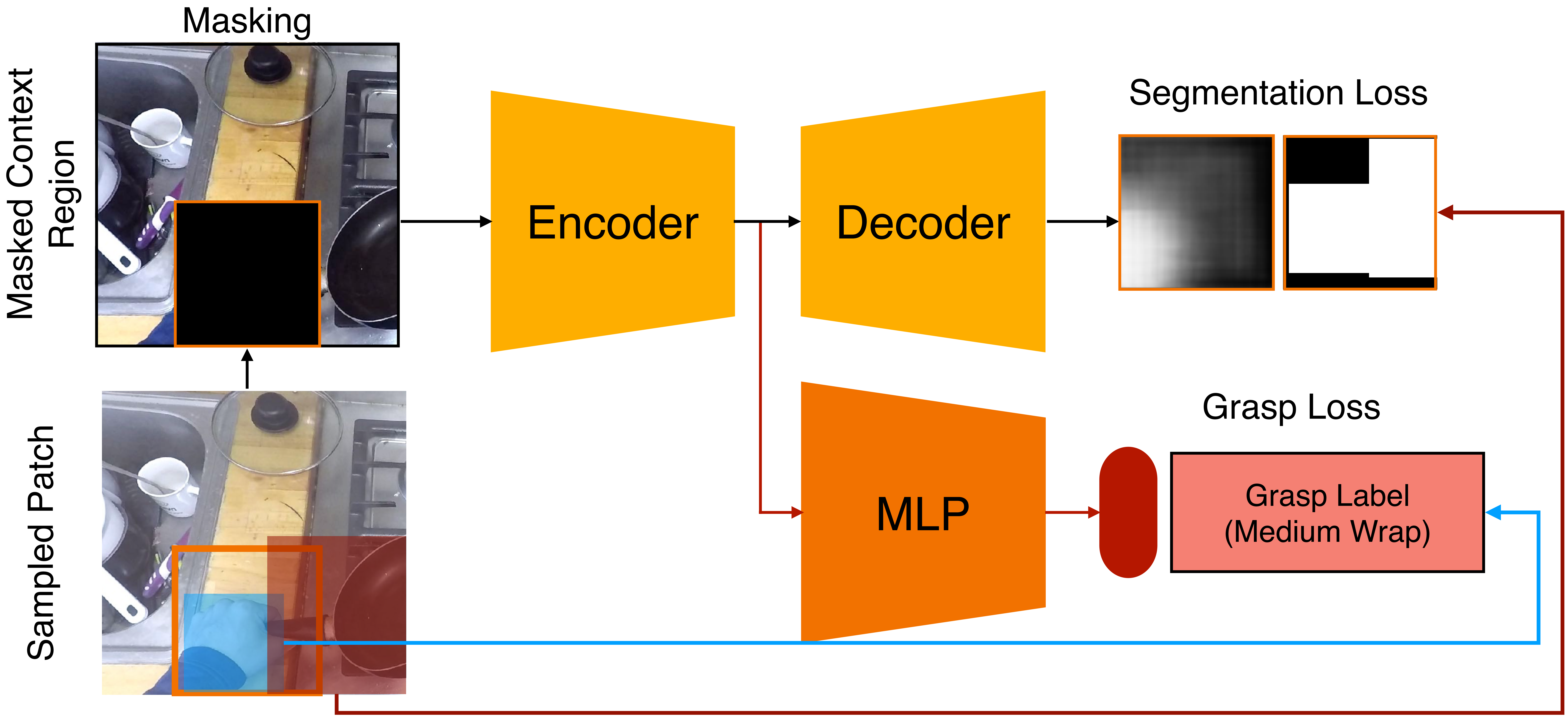
1) State Sensitive Features (C)



2) Object Affordance Prediction (A,B)

Task 2. Learning Object Affordances: Approach

Affordances via **C**ontext **P**rediction (ACP)



Task 2a. Learning Learning Object Affordances: Results

Evaluating Region-of-Interaction Prediction

Task 2a. Learning Learning Object Affordances: Results

Evaluation on EPIC-ROI Dataset

RoI-prediction Quantitative Comparison

Methods	Supervision	AP
MaskRCNN	MSCOCO	64.0
IHOTSPOT [5]	Action and Object Labels	43.8
DEEPGAZE2 [6]	Recorded Eye Fixations	55.7
ACP (Ours)	Hand-Object detections	57.0
MaskRCNN + DEEPGAZE2	**Adding the predictions**	**66.6**
MaskRCNN + ACP (Ours)	**Adding the predictions**	**68.6**

Supervised MaskRCNN does better than ACP

ACP improves over action-classification and objectness methods

ACP combined with MaskRCNN performs the best

[5] Nagarajan et al. CVPR 2019. Grounded human-object interaction hotspots from video.
[6] Kummerer et al. ICCV 2017, Understanding low- and high-level contributions to fixation prediction

Task 2a. Learning Learning Object Affordances: Results

Evaluation on EPIC-ROI Dataset (Non-COCO Objects)

RoI Quantitative Comparison

Methods	Supervision	AP
MaskRCNN	MSCOCO	22.8
MaskRCNN + DEEPGAZE2	Recorded Eye-fixations	26.2
MaskRCNN + ACP (Ours)	Hand-Object Detections	**30.5**

MaskRCNN performance is low
On Non-COCO Categories

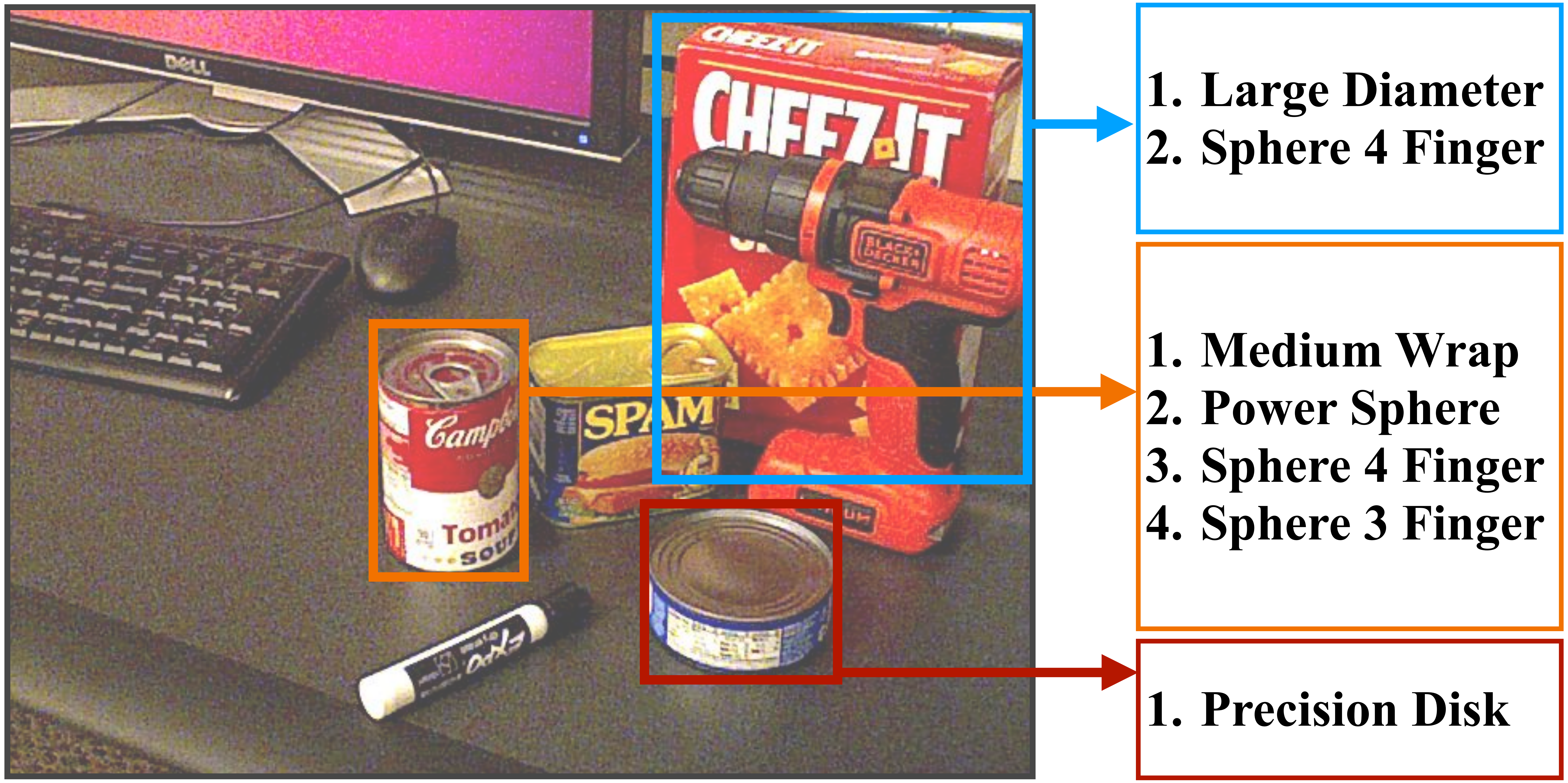
MaskRCNN+ACP improves by 7.7%

ACP better than Deepgaze2

Task 2b. Learning Learning Object Affordances: Results

Grasps Afforded by Objects (GAO) Task

YCBAffordance Dataset [7]



[7] Corona et al. CVPR 2020. Ganhand: Predicting human grasp affordances in multi-object-scenes

Task 2b. Learning Learning Object Affordances: Results

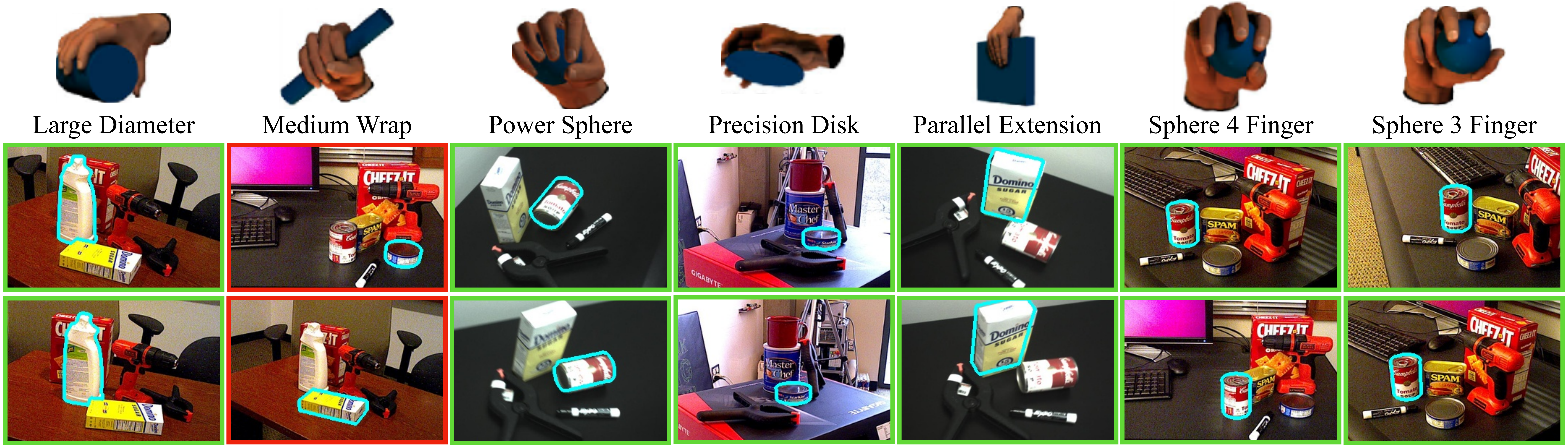
Grasps Afforded by Objects (GAO) Task

Chance - 30 % mAP

ACP (Ours) - 38 % mAP

Supervised - 50 % mAP

Top Ranking Objects predicted by ACP



Human Hands as Probes for Interactive Object Understanding

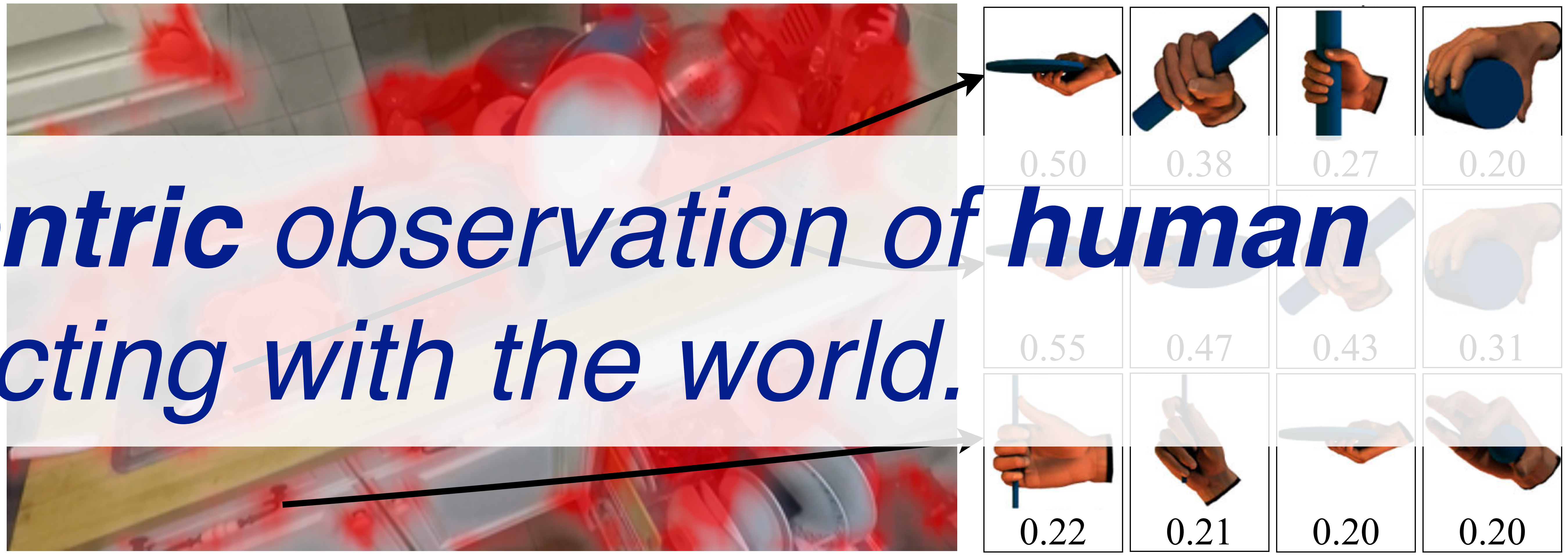
A) Which sites can we interact at? (cupboard handles)

B) How to interact with those sites? (using adducted thumb grasp)

C) What happens when we do? (the cupboard opens)

*Learn through **egocentric** observation of **human hands** interacting with the world.*

1) State Sensitive Features (C)



2) Object Affordance Prediction (A,B)

Hands were useful, but they are also a nuisance…

1) State-sensitive features

2) Affordances

Look Ma, No Hands! Agent-Environment Factorization of Egocentric Videos

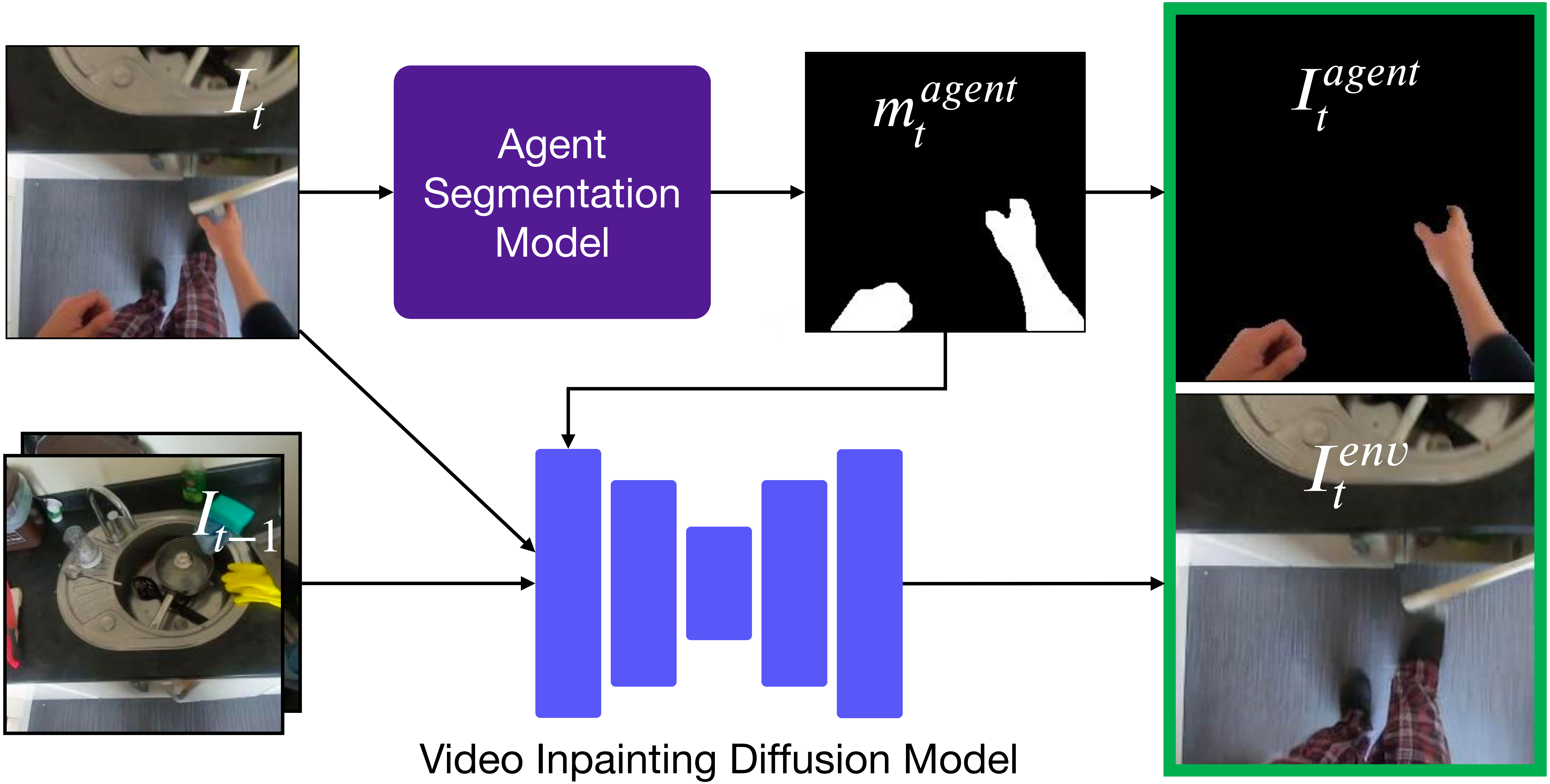
Matthew Chang Aditya Prakash Saurabh Gupta

arXiv 2023

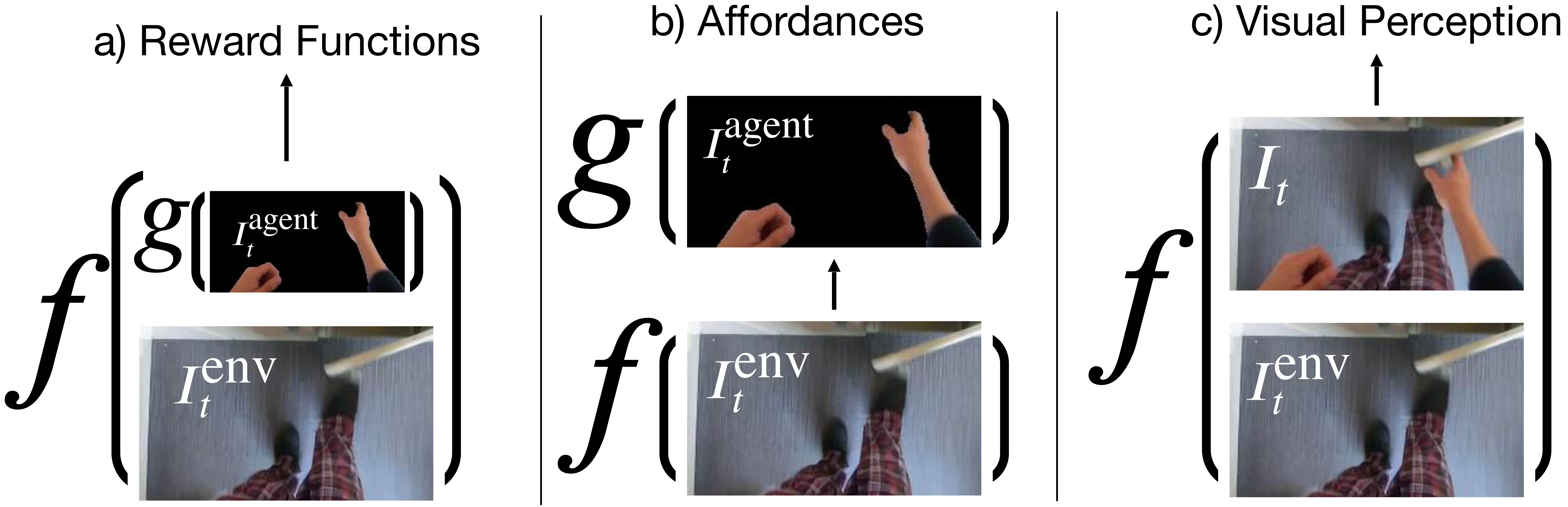
Find Aditya at the poster session

ILLINOIS

Agent-Environment Factored Representations



Applications of Factorization



Video Inpainting Diffusion Model (VIDM)

1. Leverage priors on how object are

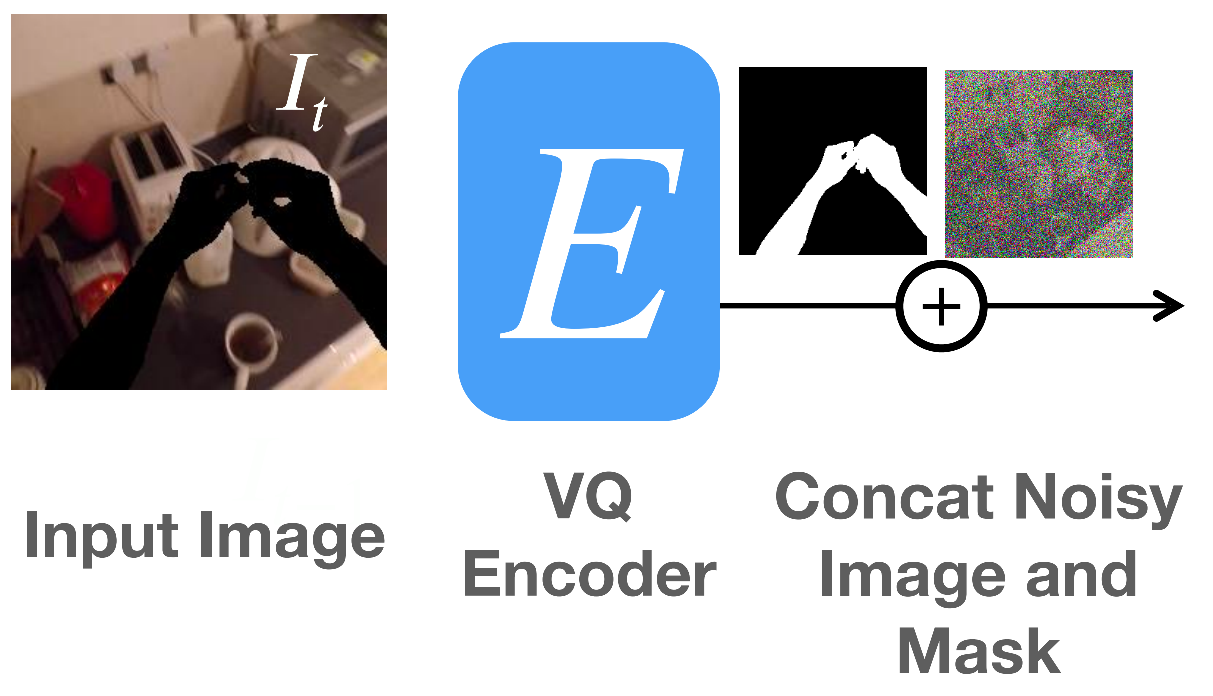
2. Leverage past information in the video

Input Image

Video Inpainting Diffusion Model (VIDM)

1. Leverage priors on how object are

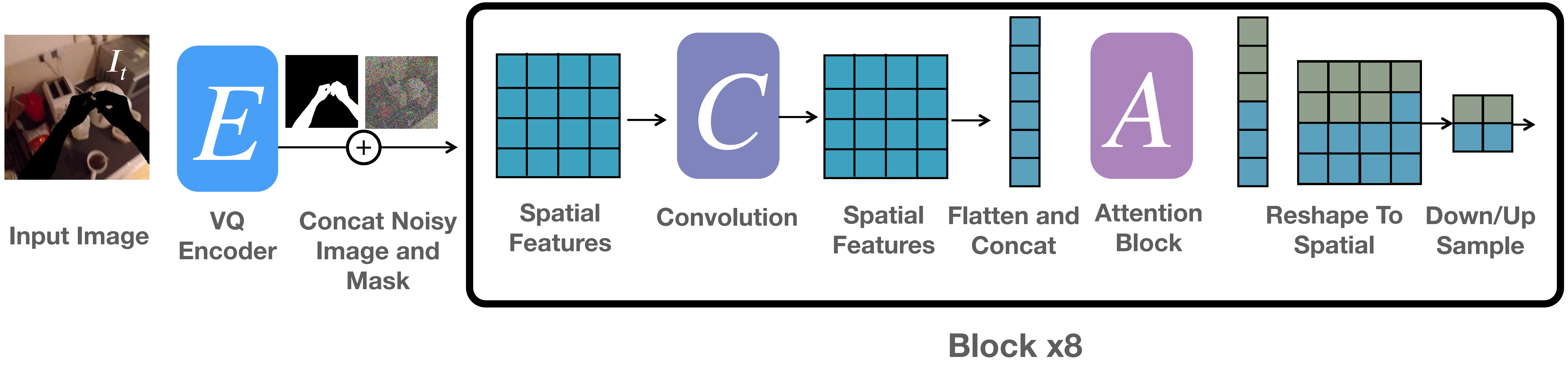
2. Leverage past information in the video



Video Inpainting Diffusion Model (VIDM)

1. Leverage priors on how object are

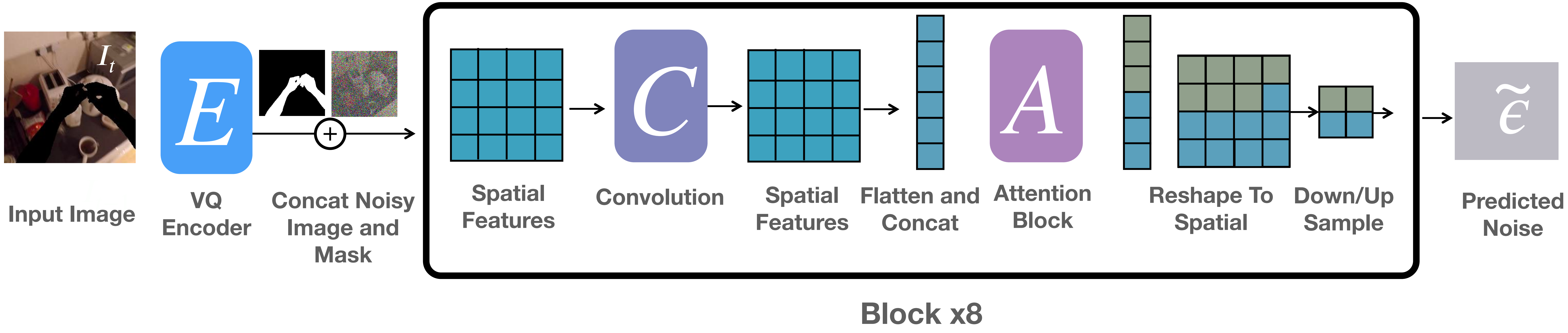
2. Leverage past information in the video



Video Inpainting Diffusion Model (VIDM)

1. *Leverage priors on how object are*

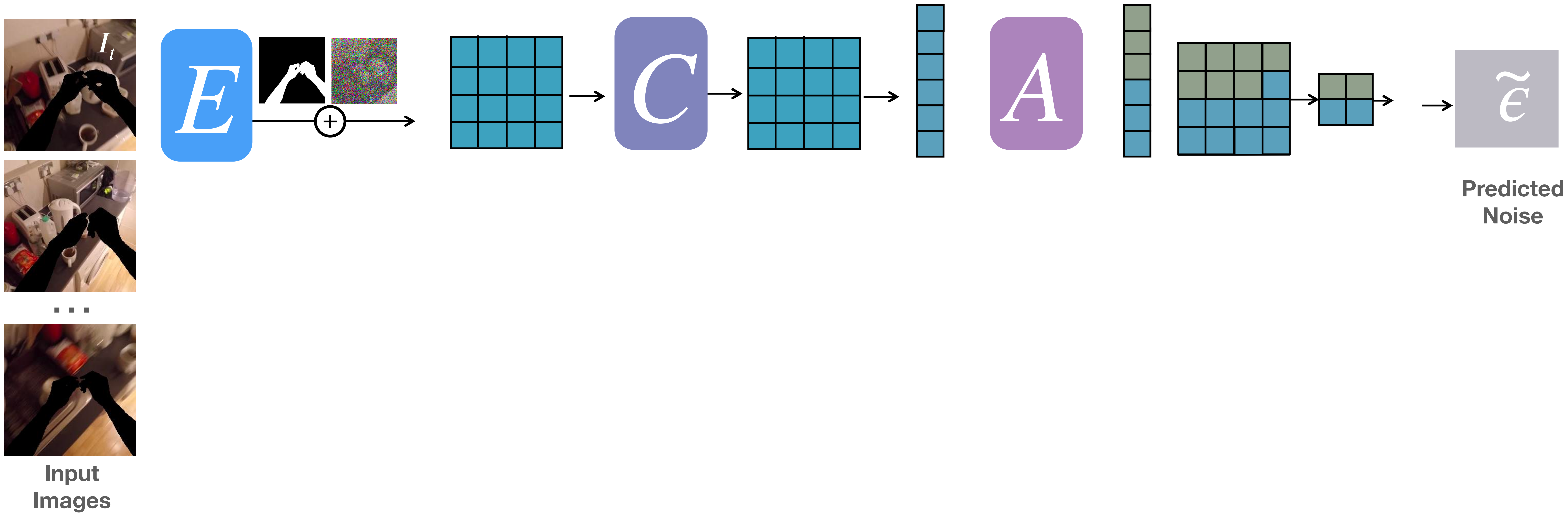
2. *Leverage past information in the video*



Video Inpainting Diffusion Model (VIDM)

1. Leverage priors on how object are

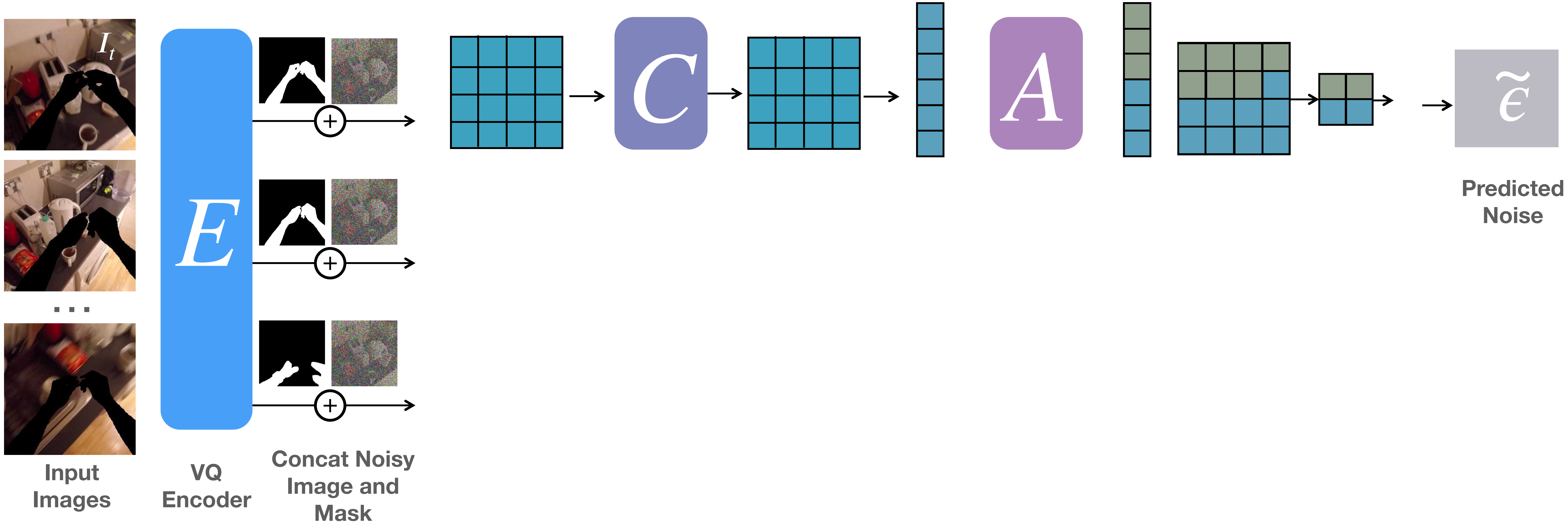
2. Leverage past information in the video



Video Inpainting Diffusion Model (VIDM)

1. Leverage priors on how object are

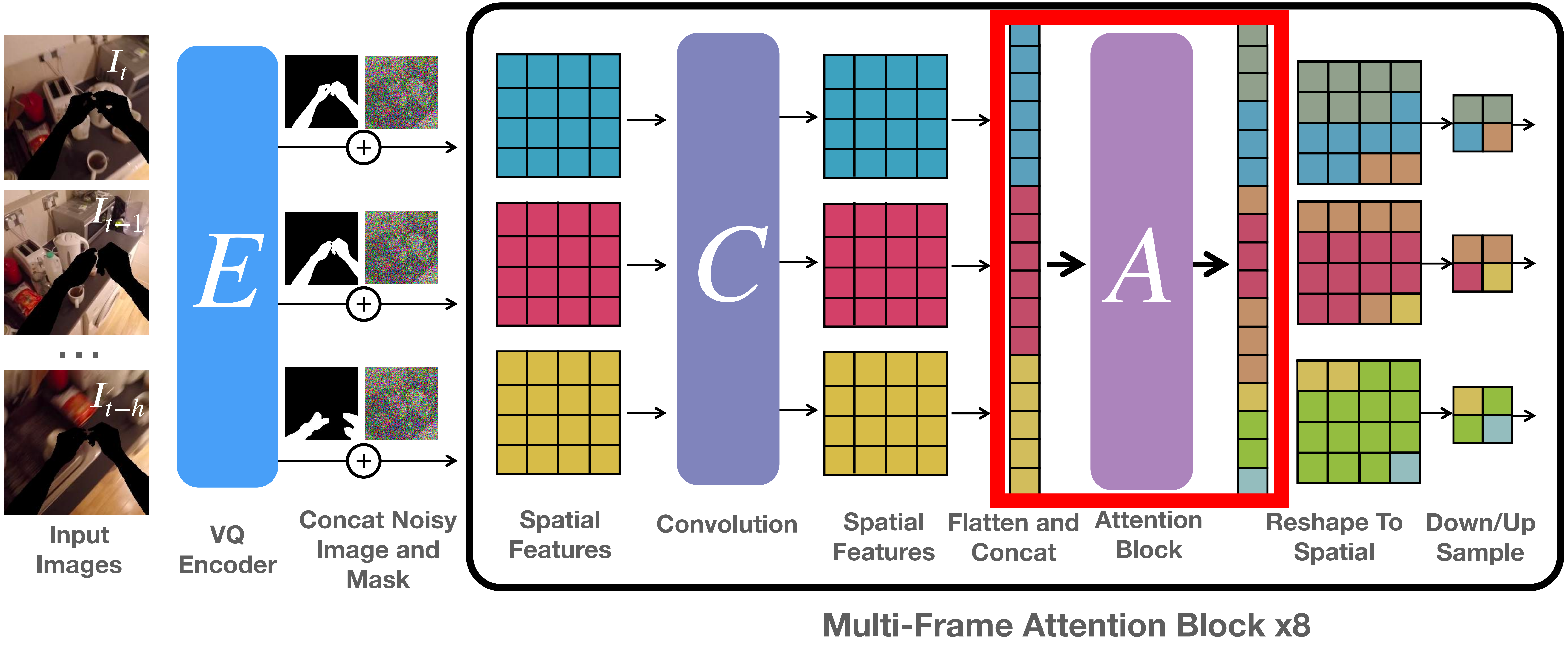
2. Leverage past information in the video



Video Inpainting Diffusion Model (VIDM)

1. Leverage priors on how object are

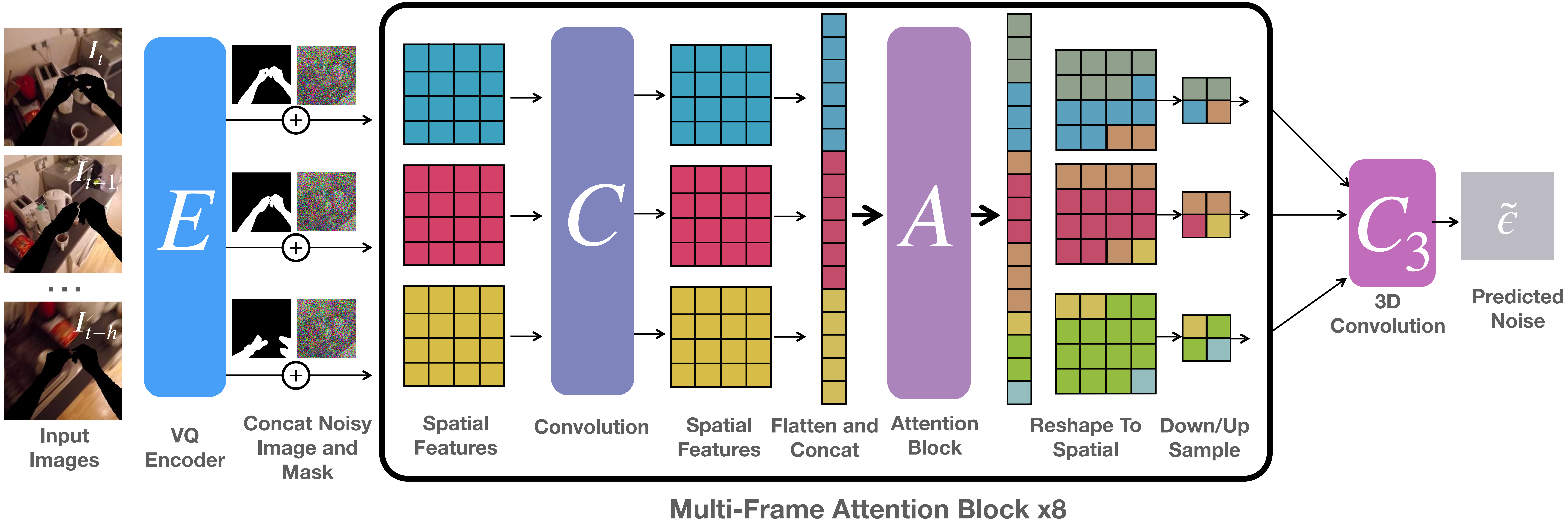
2. Leverage past information in the video



Video Inpainting Diffusion Model (VIDM)

1. Leverage priors on how object are

2. Leverage past information in the video



Reconstruction Evaluation

a) Original Image

Effectively leverages prior frames

… while also using priors learned on large scale image datasets

Inpainting Method	PSNR↑	SSIM↑	FID↓	Runtime ↓

Visualizations

Frame-by-frame results, no temporal smoothing

Visualizations

Visualizations

Visualizations

Results

Reward Functions

Human-Robot Domain Gap

ρ: 0.56 → 0.61

Affordances

Data Mismatch

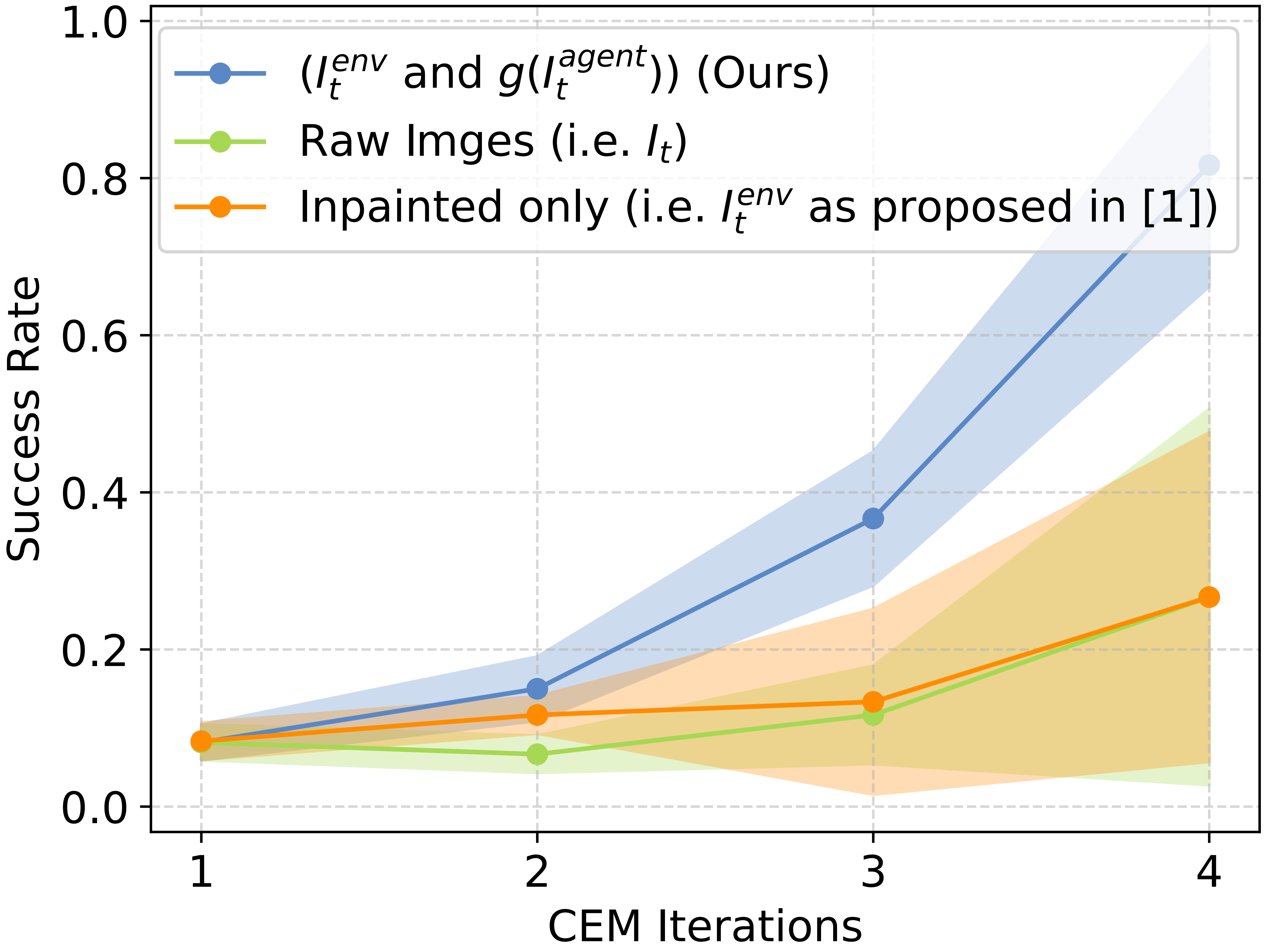
Acc: 0.35 → 0.41

Detection

Occlusion

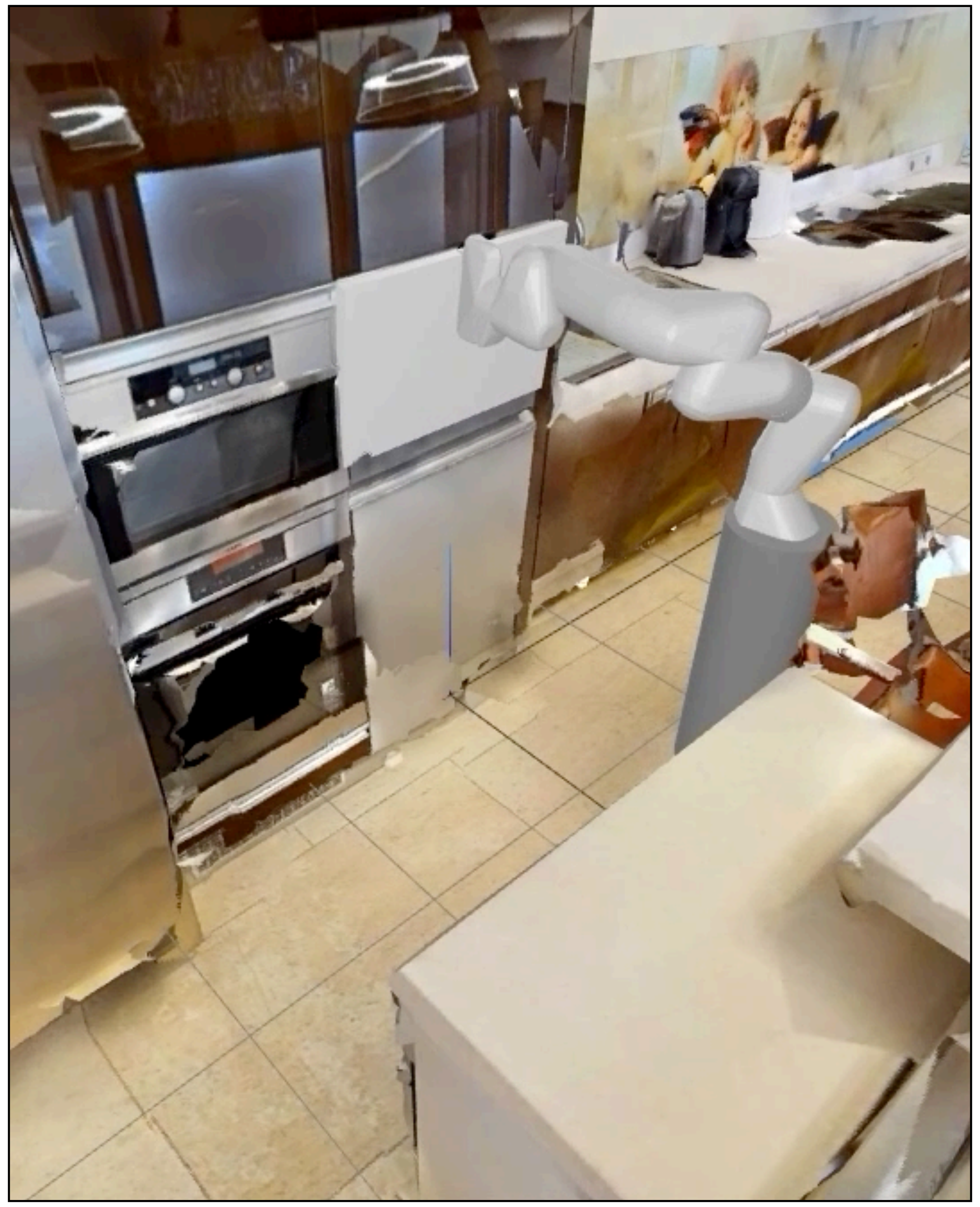
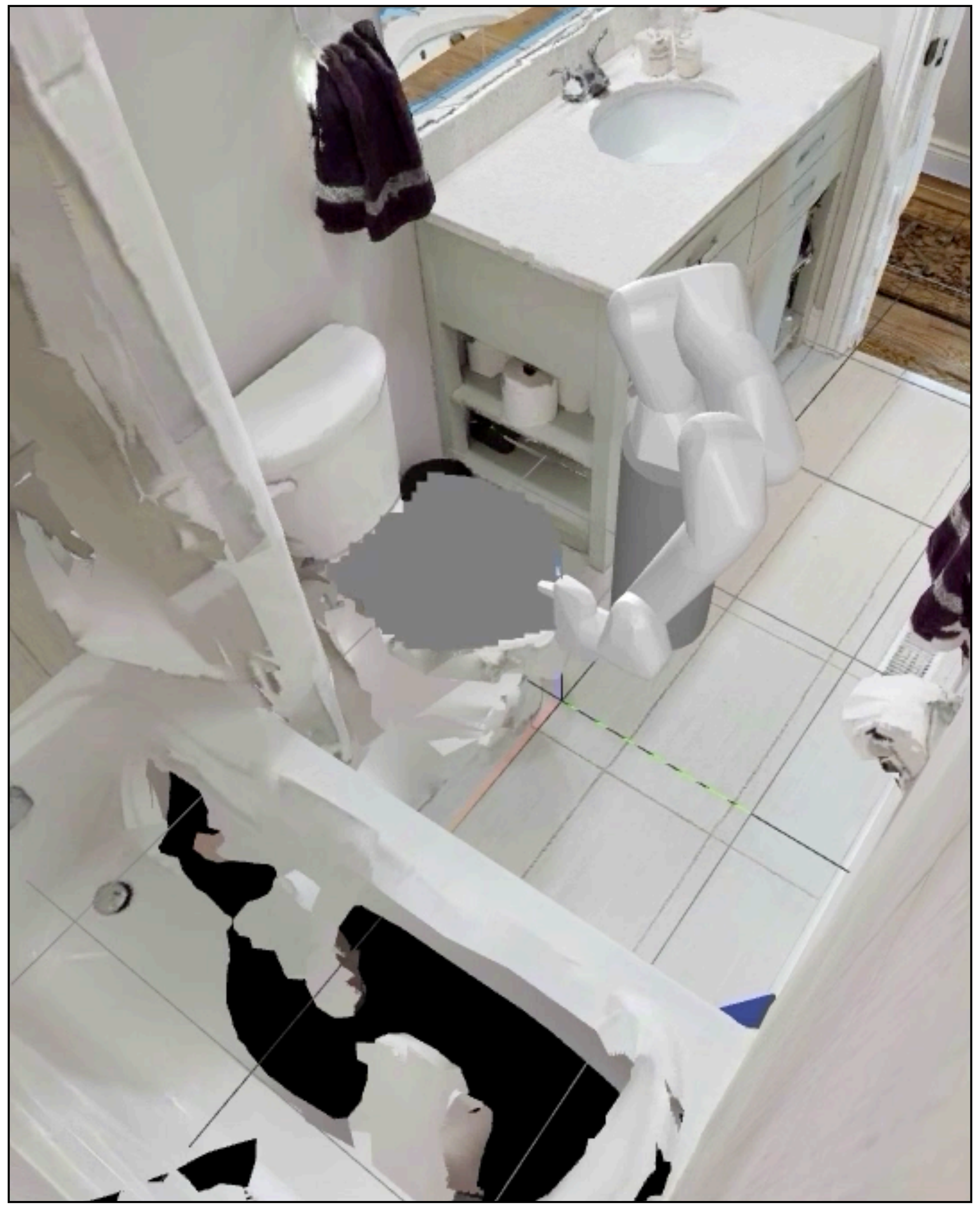
mAR: 0.26 → 0.38

Faster Real-world Robot Learning



[1] Bahl et al. RSS 2022, Human-to-Robot Imitation in the Wild

Aside: Precise Motion Plans to Articulate Articulated Objects

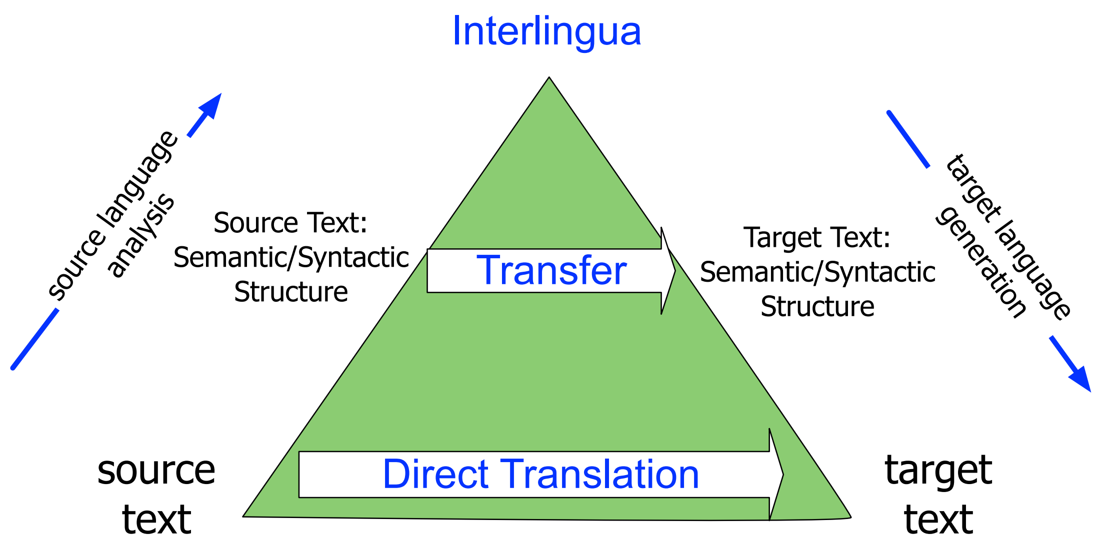




Talk to me at the poster session

Arjun Gupta, Max Shpeherd, Saurabh Gupta. In *ICRA 2023.*

Predicting Motion Plans for Articulating Everyday Objects

Learning at different abstraction levels

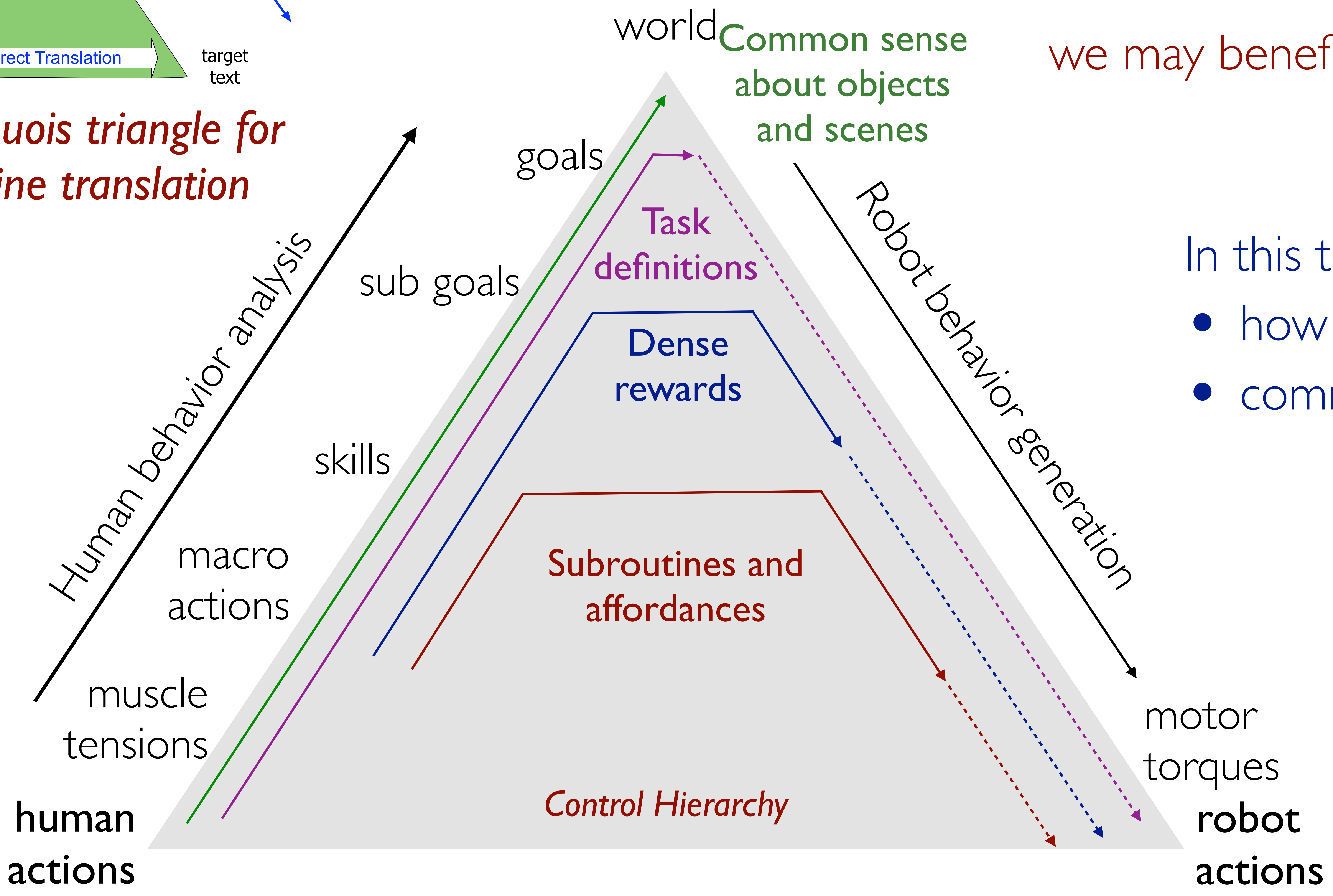


The Vauquois triangle for machine translation

Depending on the amount of gap between:

- goals,
- embodiment,
- what we can observe in videos

we may benefit from transfer at different levels.



In this talk, using video to learn,

- how to interact with objects
- common sense about scenes

Semantic Visual Navigation by Watching YouTube Videos

Matthew Chang Arjun Gupta Saurabh Gupta

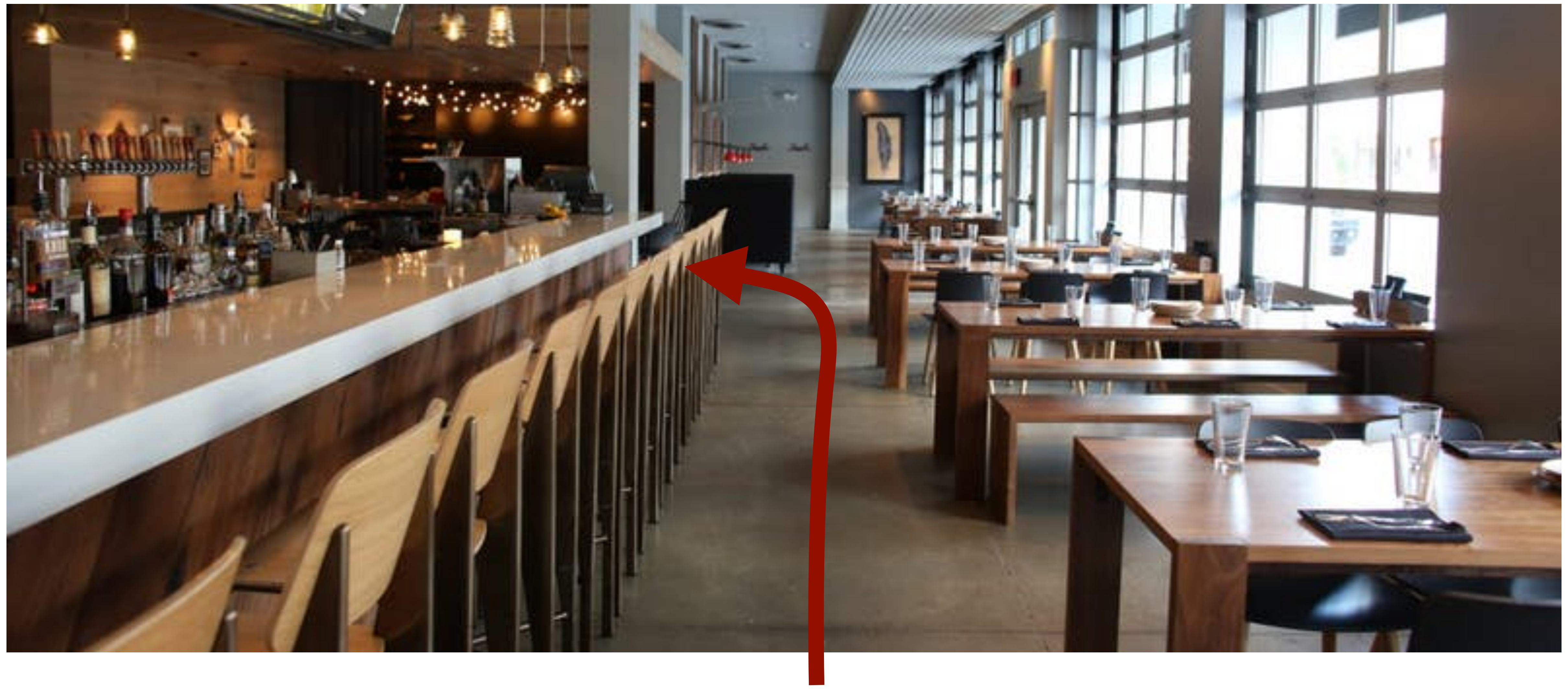
University of Illinois at Urbana-Champaign


NeurIPS 2020

Problem Statement

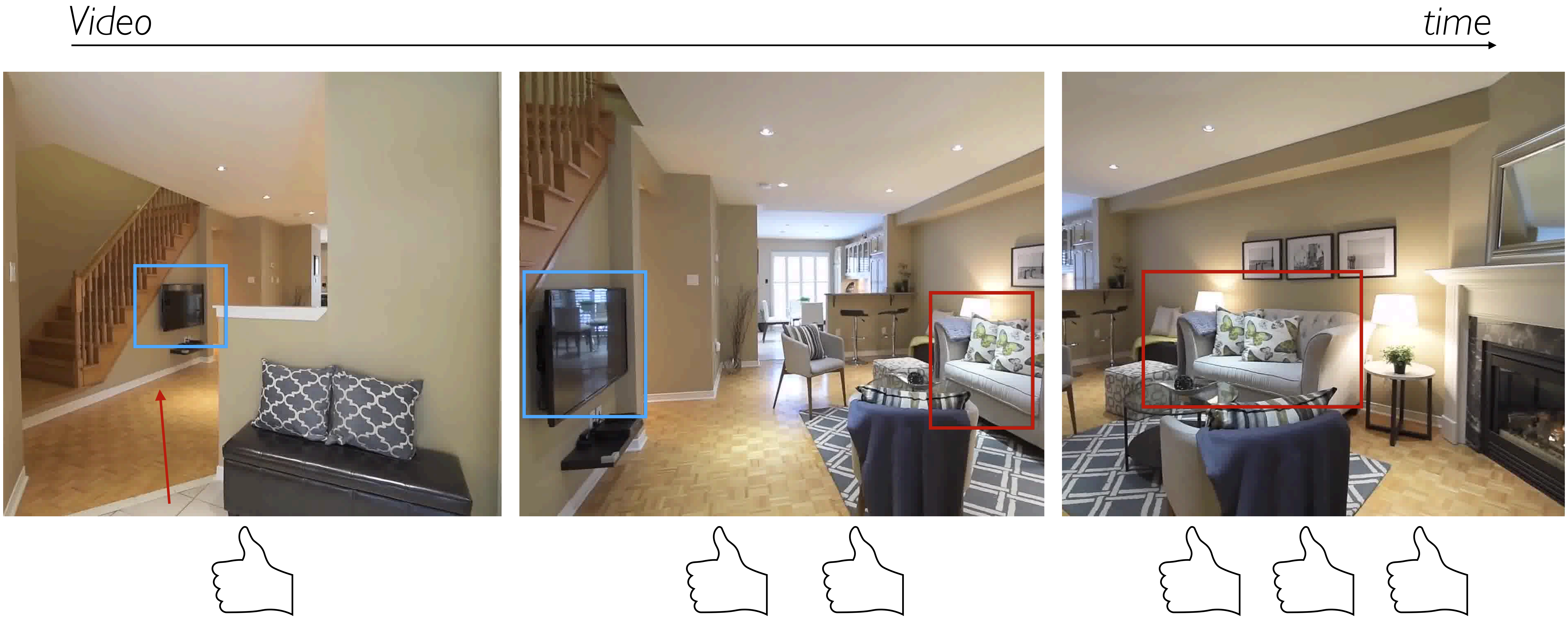
Input: *Egocentric videos (real estate tours from YouTube)*

Output: *Semantic cues to efficiently find objects in novel indoor environments, e.g. finding a restroom*



Some Intuition

Mine for spatial co-occurrences



e.g. cues for finding a couch

Value Learning from Videos (VLV)

Learned Value Function

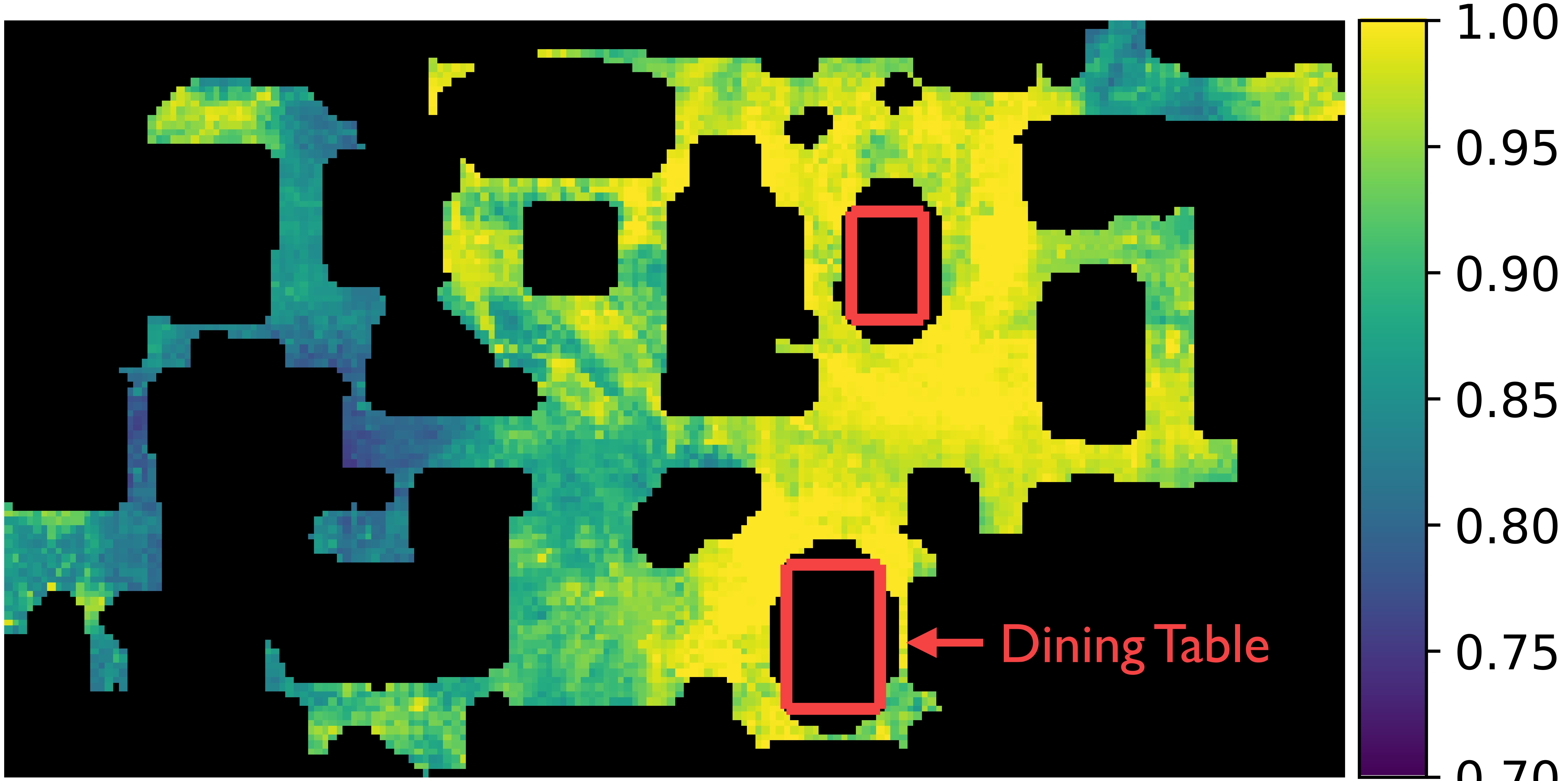
$f(I, c) \approx$ nearness to goal

Value function predicts a proxy for nearness to a goal object for a given image

Learned Value Function

$f(I, c) \approx$ nearness to goal

Value function predicts a proxy for nearness to a goal object for a given image

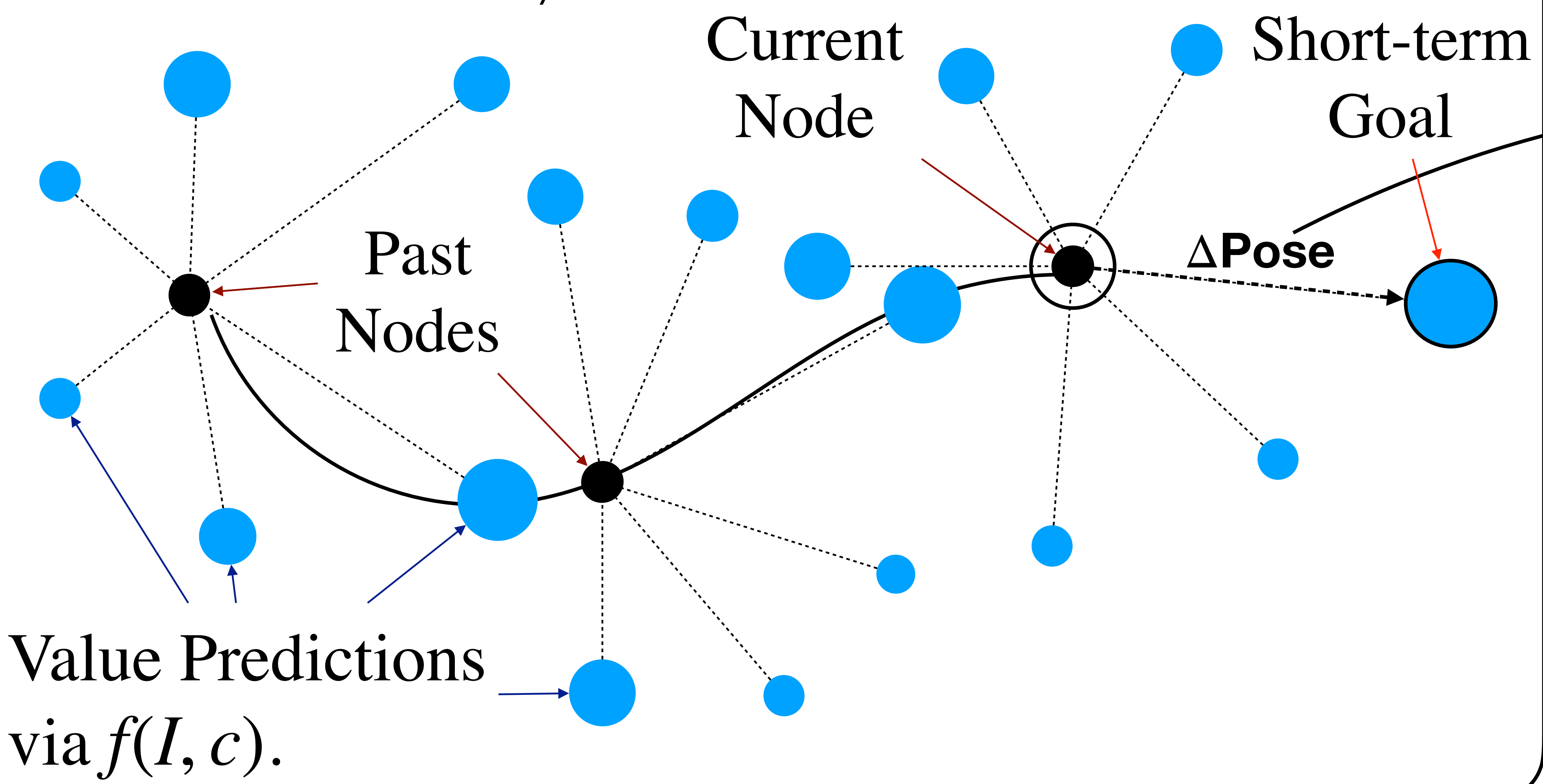


Using Learned Values for Semantic Navigation

Hierarchical Policy

High-Level Policy

- Decides where to go next and emits short-term goal
- Builds a topological map [1] that stores values predicted by $f(I, c)$ at different locations in different directions
- Samples most promising direction, and passes ΔPose to Low-Level Policy



Low-Level Policy

- Executes actions to achieve short-term goal
- Incrementally builds occupancy map from depth camera
- Uses Fast-Marching Method for path planning to get actions to execute
- Return control on success or failure

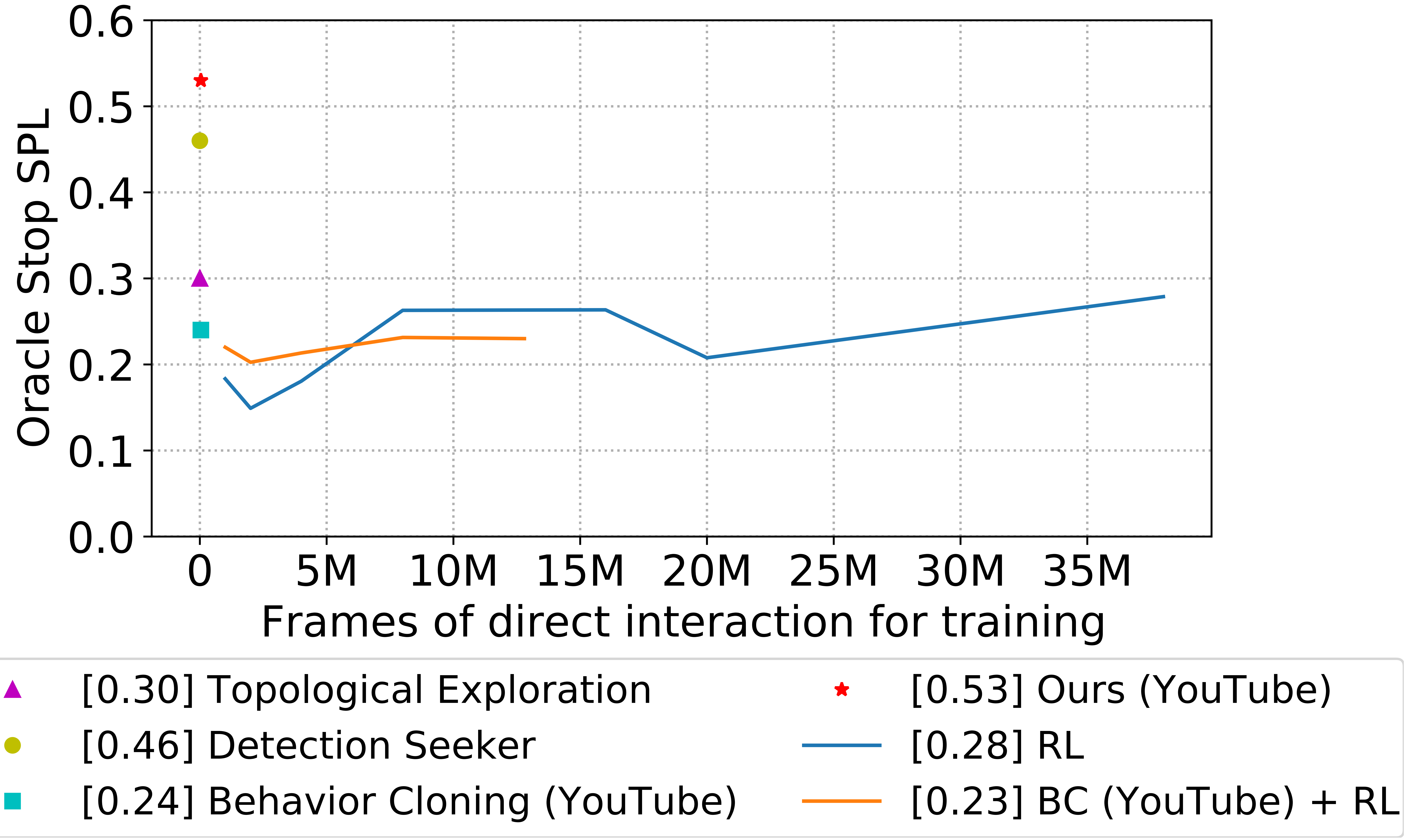
Occupancy Map FMM Cost Map

- Forward
- Left
- Right
- Stop

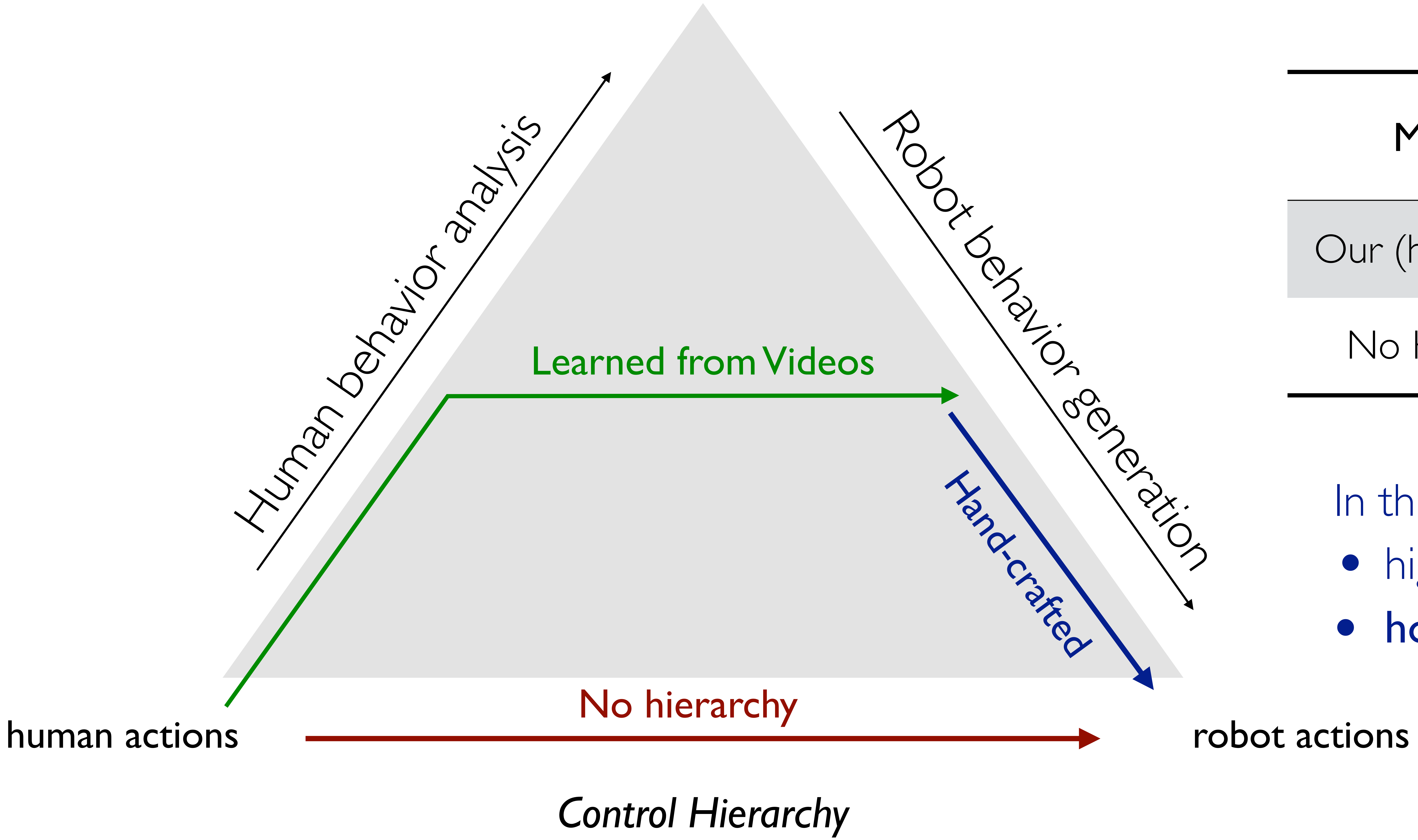
[1] D. Chaplot, R. Salakhutdinov, A. Gupta, S. Gupta. Neural topological slam for visual navigation. In *CVPR*, 2020.

Results (ObjectGoal Task)

Find object of interest (bed, chair, couches, tables, toilets) in novel indoor environments.



Transferring at appropriate level is important



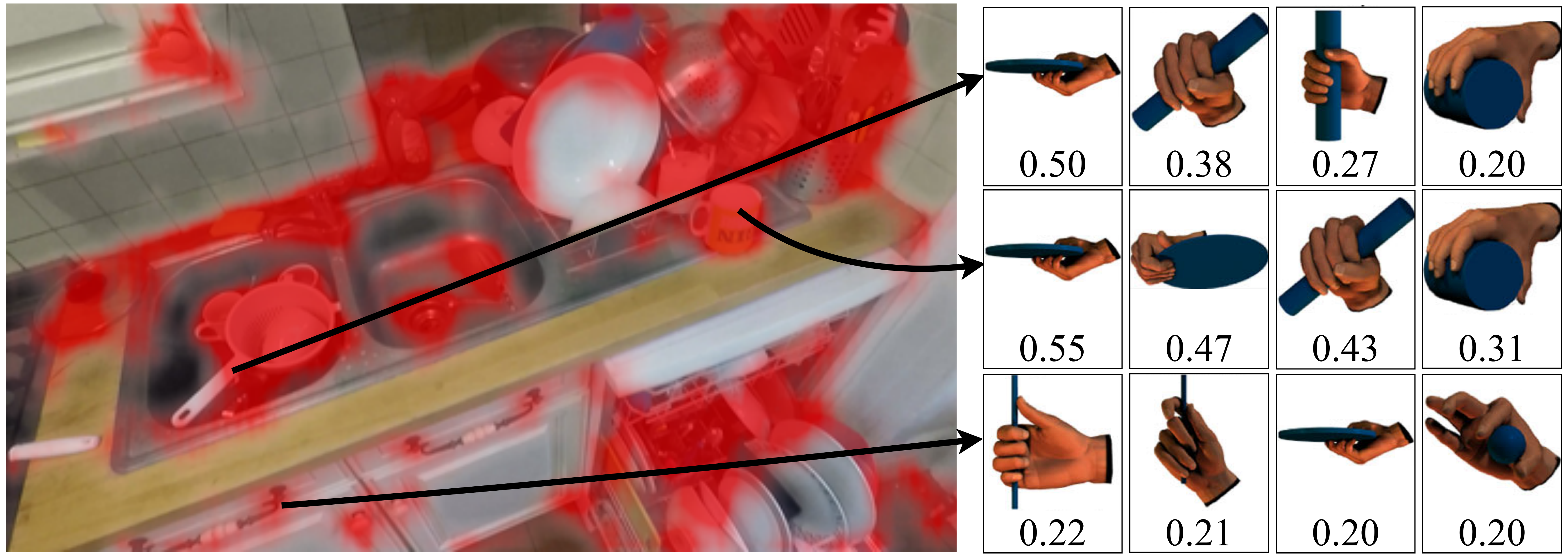
Method	Oracle Stop SPL (Valdiation Set)
Our (hierarchical)	**0.40**
No Hierarchy	0.15

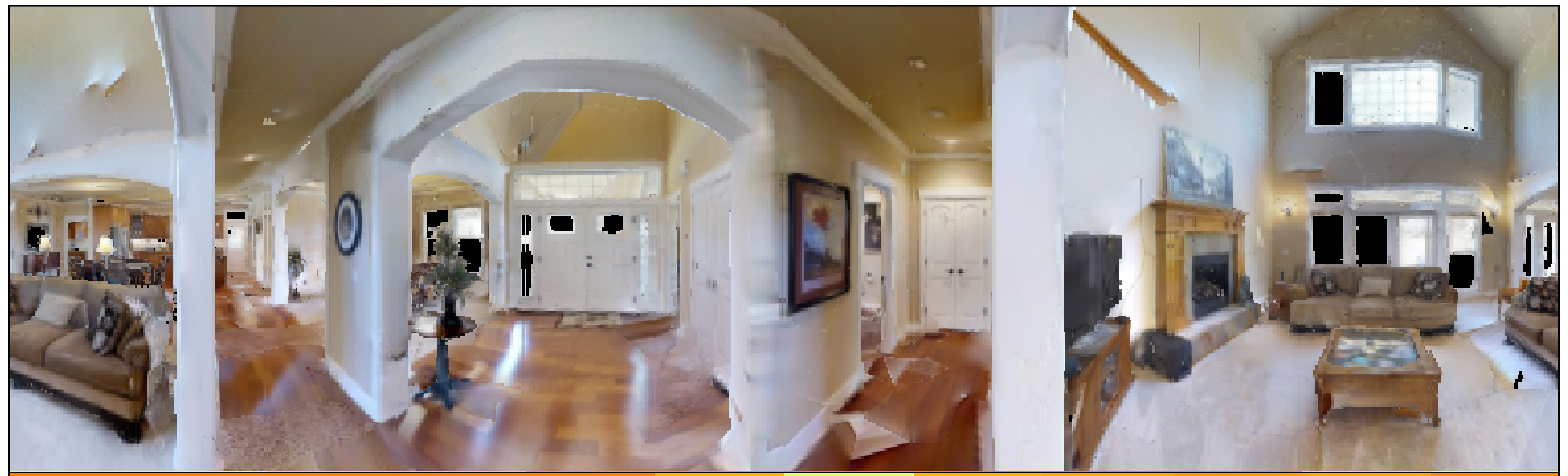
In this talk:

- high-level value functions
- **how to interact with objects**

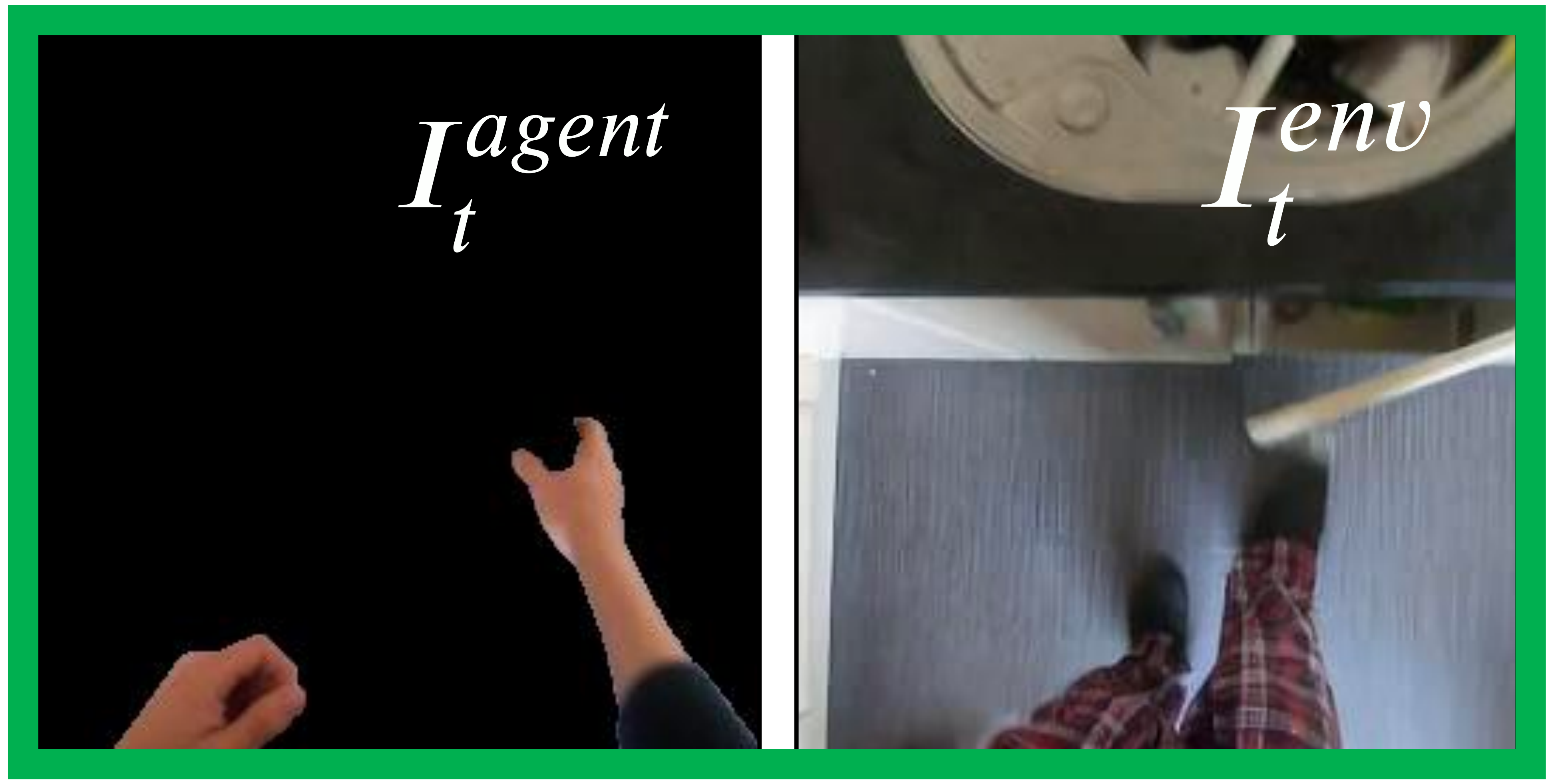
Summary

Transfer at the right level of abstraction





Chair	0.99	0.99	0.99	1.00	0.99	0.96	0.92	0.96	0.97	0.98	0.97	0.99
Couch	0.99	0.95	0.84	0.80	0.82	0.82	0.80	0.84	0.87	0.90	0.94	0.99
D. Table	0.87	0.97	0.99	1.01	0.92	0.88	0.82	0.84	0.85	0.85	0.84	0.83
Bed	0.78	0.78	0.80	0.80	0.83	0.83	0.84	0.84	0.84	0.83	0.80	0.78
Toilet	0.62	0.63	0.65	0.63	0.71	0.68	0.71	0.71	0.71	0.66	0.63	0.62




Matthew Chang


Arjun Gupta


Aditya Prakash


Mohit Goyal


Sahil Modi


Rishabh Goyal

Thank You!