Prediction in goal-directed action
Fiehler et al

- **Prediction allows humans and other animals to prepare for future interactions with their environment. This is important in our dynamically changing world that requires fast and accurate reactions to external events.**
- **Knowing when and where an event is likely to occur allows us to plan eye, hand, and body movements that are suitable for the circumstances.**
- **Predicting the sensory consequences of such movements helps to differentiate between self-produced and externally generated movements.**

From: Prediction in goal-directed action
Journal of Vision. 2019;19(9):10. doi:10.1167/19.9.10

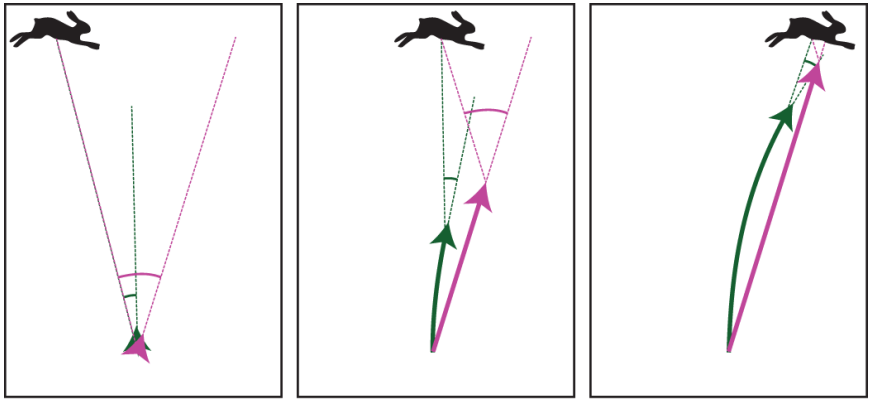


Figure Legend:

More ways in which interception could be controlled.

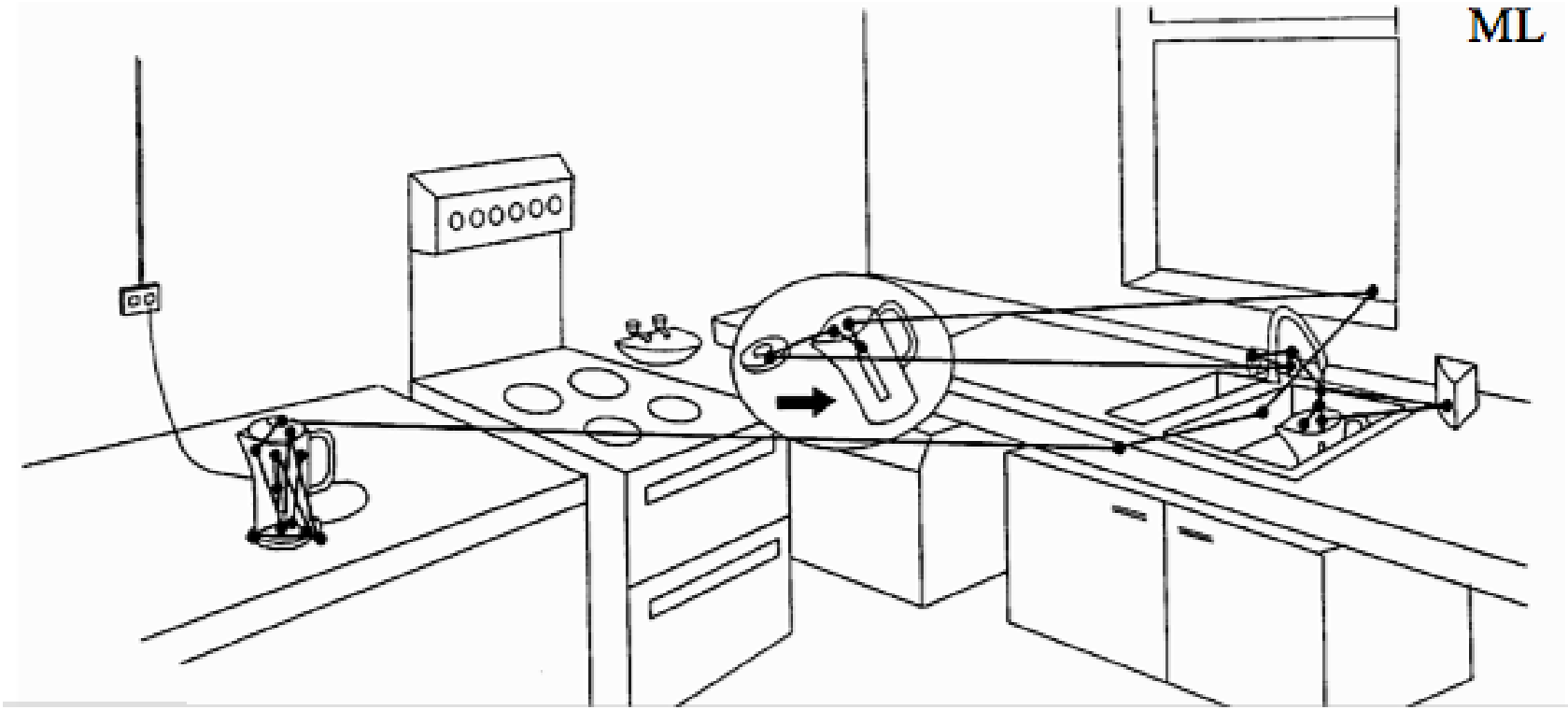
Efference copy as a form of prediction

- Imagine watching a train pass by. If you do not follow the train with your eyes, its image will sweep across the retina while the image of the context surrounding the train will not. If you do follow the train with your eyes, its image will be (more or less) stable on the retina, while the context's image will produce a motion sweep.
- How does the brain figure out when to attribute retinal motion to object motion and when to attribute it to movements of the eyes?
- Von Holst and Mittelstaedt (1950) proposed that when a motor command is sent to the muscles that move the eyes, a copy of the efferent signal is simultaneously sent to visual areas of the brain.

Vision for Manipulation

Jitendra Malik

Eye Movements while making a cup of tea
Land, Mennie and Rusted (1999)



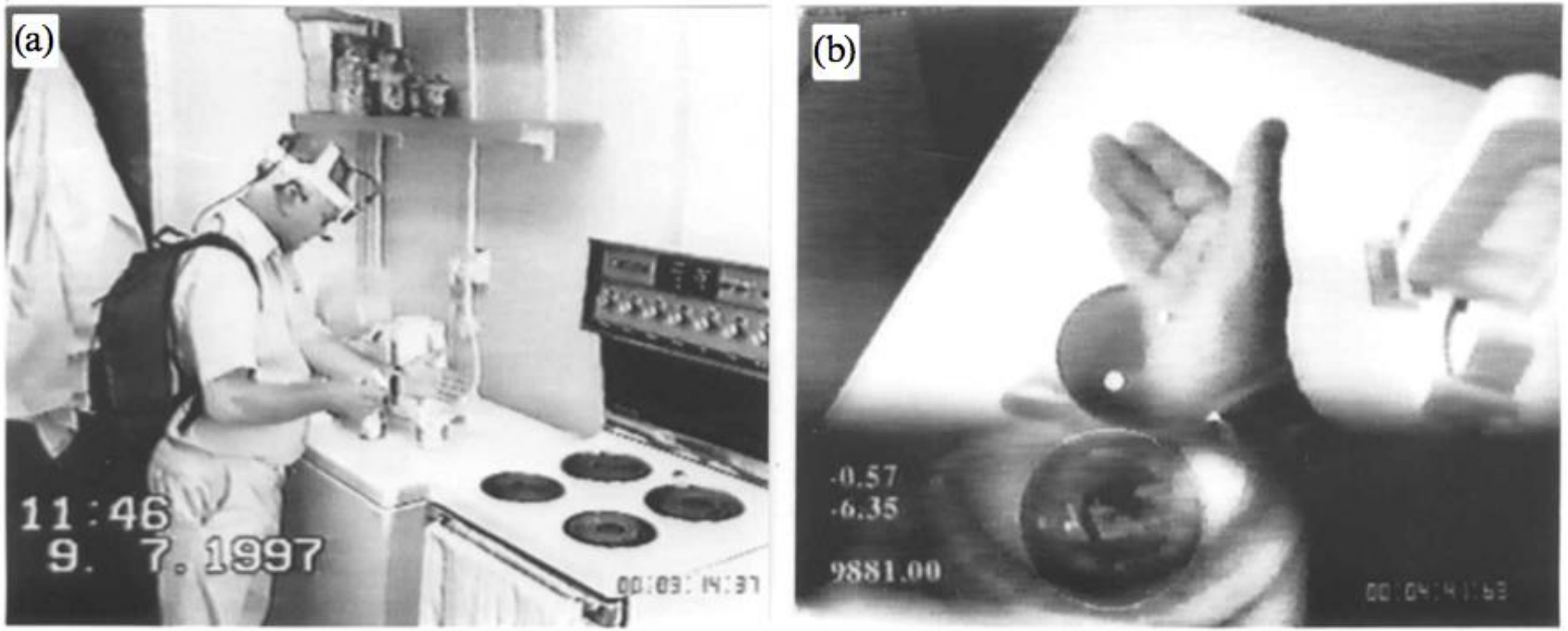
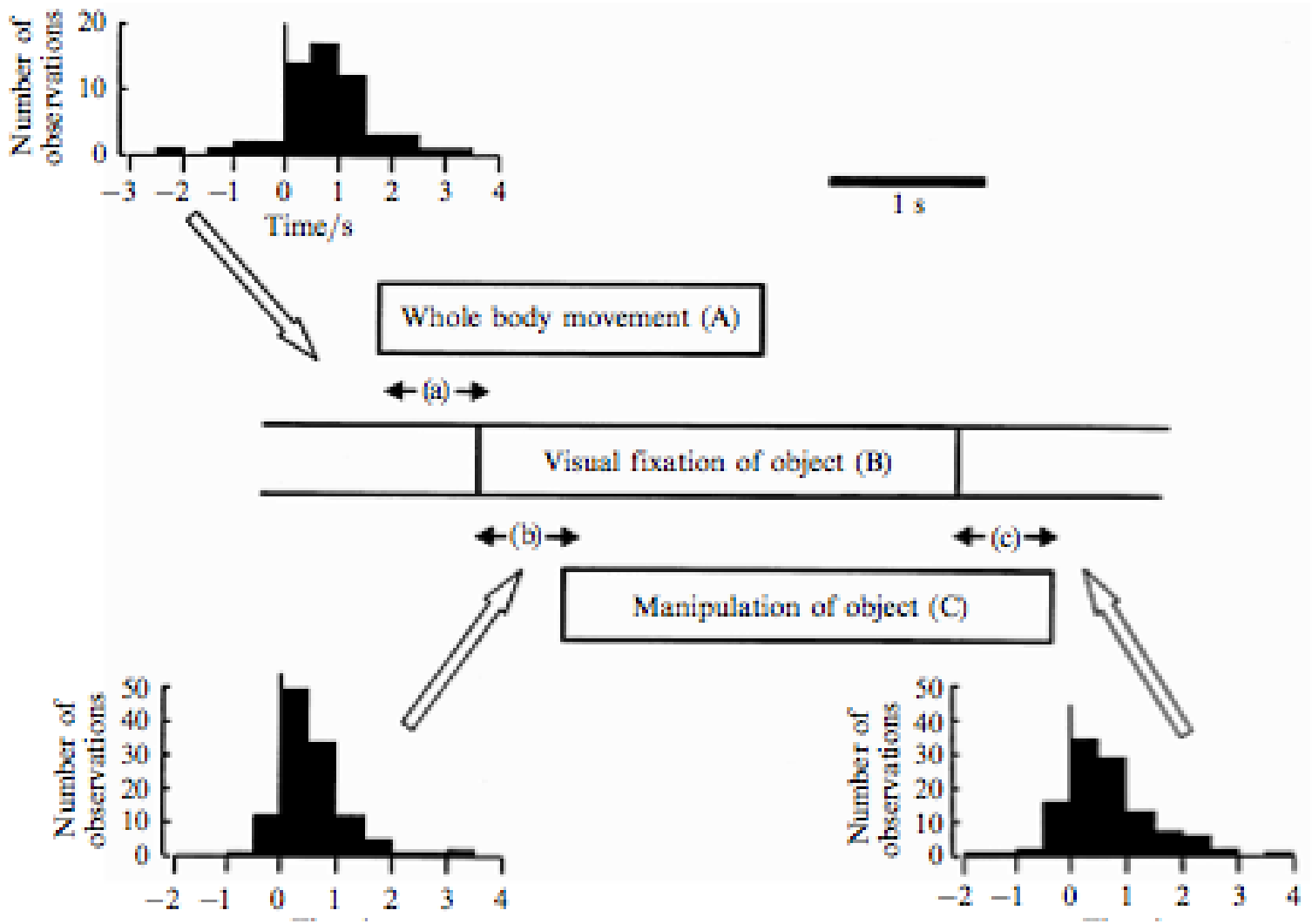


Figure 1. Prints from (a) the activity video, and (b) eye-movement video of the same instant, when the sweetener is dropped into the mug (3.14 on figure 3). The head-mounted camera and

This may be the first example of ego-exo video


Number of observations
20
10
0
−3 −2 −1 0 1 2 3 4
Time/s
1 s
Whole body movement (A)
(a)
Visual fixation of object (B)
(b)
(c)
Manipulation of object (C)
Number of observations
50
40
30
20
10
0
−2 −1 0 1 2 3 4
Number of observations
50
40
30
20
10
0
−2 −1 0 1 2 3 4

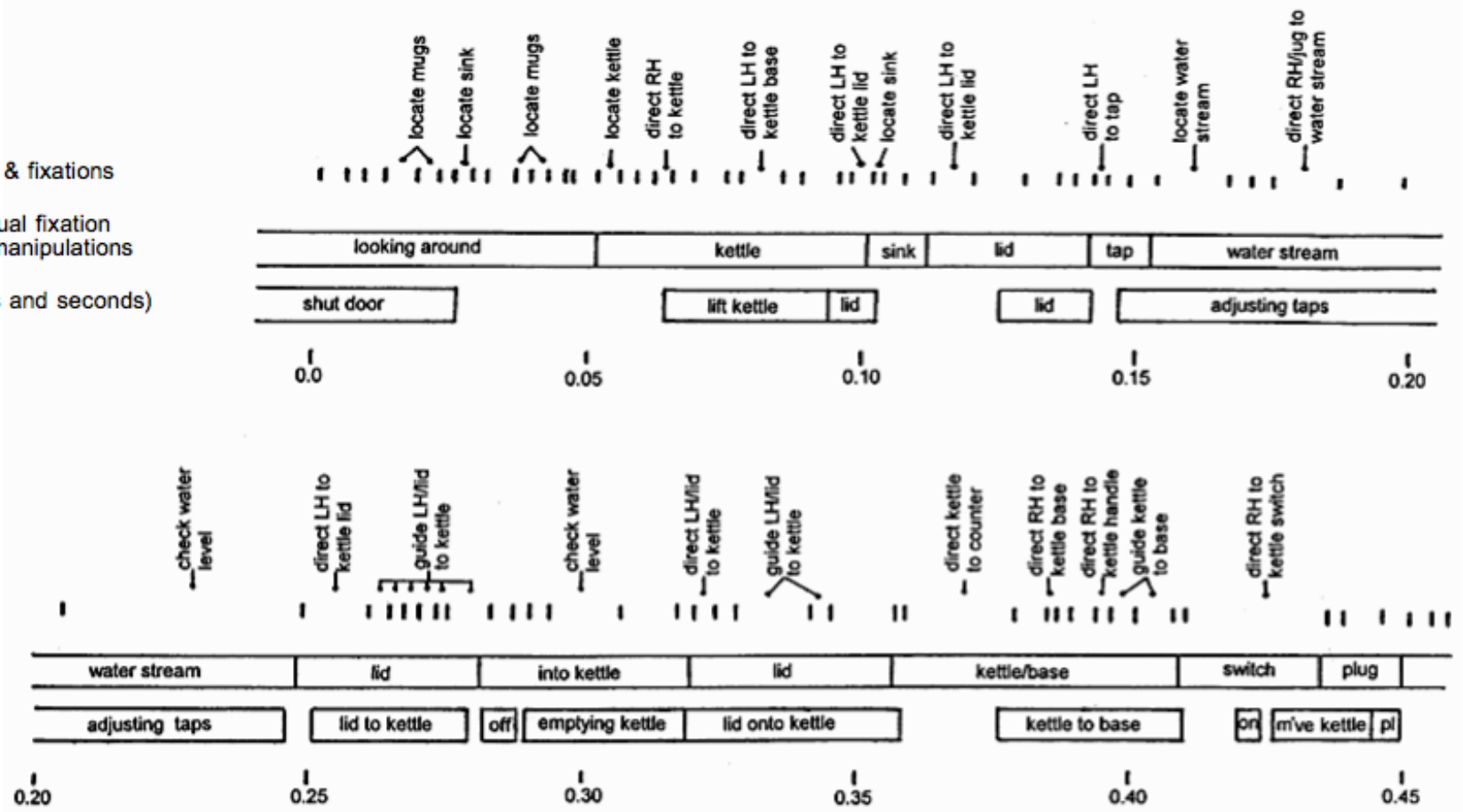


Figure 7. The specific roles of individual fixations in the first level-2 subtask ('fill the kettle') shown in figure 3. The pattern of saccades and intervening fixations are shown, and labels are

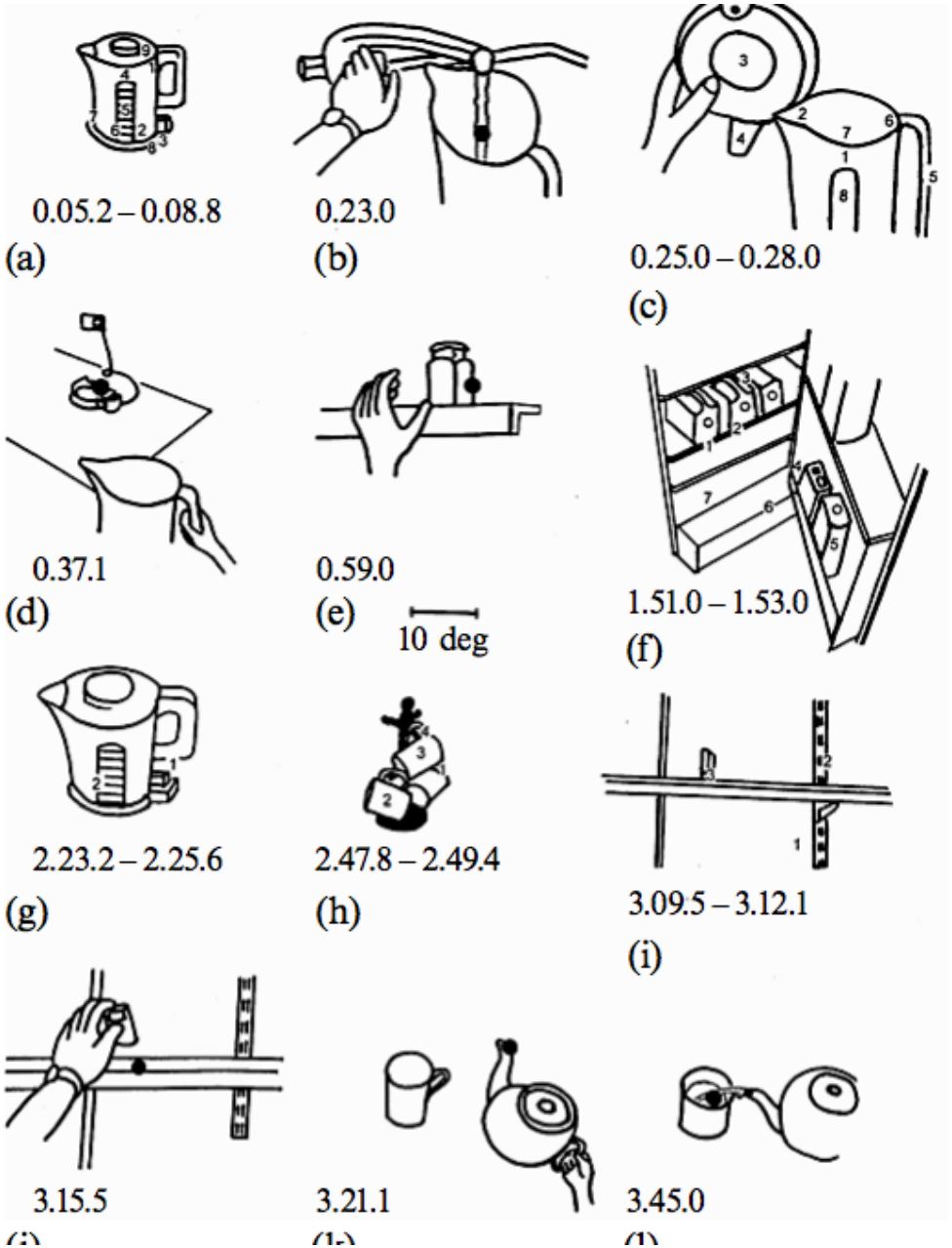
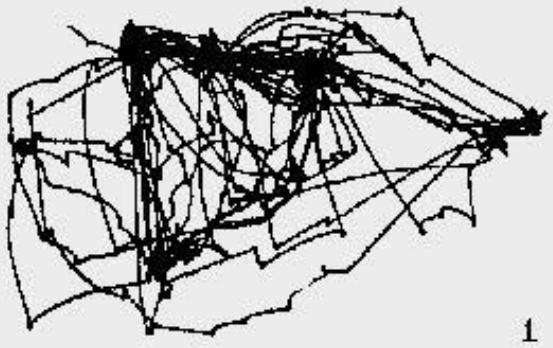


Figure 8. Examples of fixation patterns drawn from the eye-movement videotape. Sequences of successive fixation positions are indicated by numbers on the figures, and single fixations by single black dots. Numbers beneath each figure refer to timings in figure 3. 10 deg scale in centre applies to all figures. (a) Initial examination of kettle. (b) Tap control via water stream. (c) Fitting lid to kettle (drawing made at fixation 4). (d) Moving kettle to base: base is fixated. (e) Hand being directed to the tea-caddy. (f) Search around the inside of fridge 2. The tea-making milk is located at fixation 5. (g) Fixations checking the switch and gauge of the kettle when waiting for it to boil. (h) Selecting a mug. Hand goes to fixation 4. (i) Relocating sweetener prior to use requires 3 fixations. Sweetener last seen 68 s earlier. (j) Replacing sweetener 5 s after (i). Location on shelf is fixated first. (k) Swirling teapot: checking spout. (l) Pouring tea: receiving vessel fixated.

Eyes are directed towards where information that will be useful in the near future is likely to be found

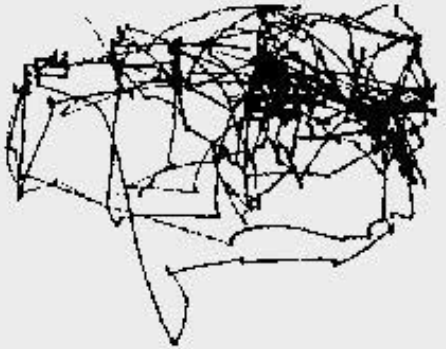
- This has been demonstrated for various everyday tasks such as making tea (Land, Mennie, & Rusted, 1999), making a sandwich (Hayhoe, Shrivastava, Mruczek, & Pelz, 2003), walking (Matthis, Yates, & Hayhoe, 2018) and driving (Land & Lee, 1994; Wilkie & Wann, 2003). It has also been demonstrated for more specialized activities such as reading the score when playing music (Furneaux & Land, 1999).
- In all these cases, gaze precedes and guides movements of the arm or leg. Similar eye movements guide tasks that do not involve such movements. For instance, when reading (Rayner, 1998), searching for an object (Eckstein, 2011), or identifying faces (Peterson & Eckstein, 2013), observers move their eyes to where they anticipate to find the most relevant information at each moment. In general, people anticipate when they will need certain information

Ilya Repin, An Unexpected Visitor, 1884.



1

Free examination.



2

Estimate material circumstances of the family

3

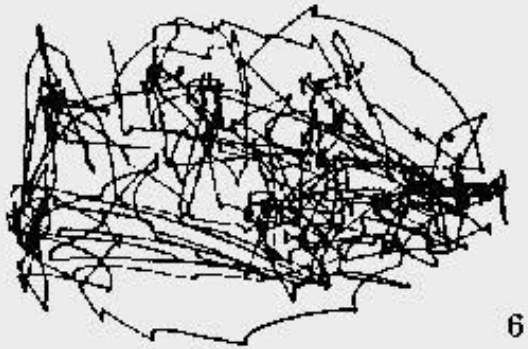
Give the ages of the people.

4

Surmise what the family had been doing before the arrival of the unexpected visitor.

5

Remember the clothes worn by the people.



6

Remember positions of people and objects in the room.

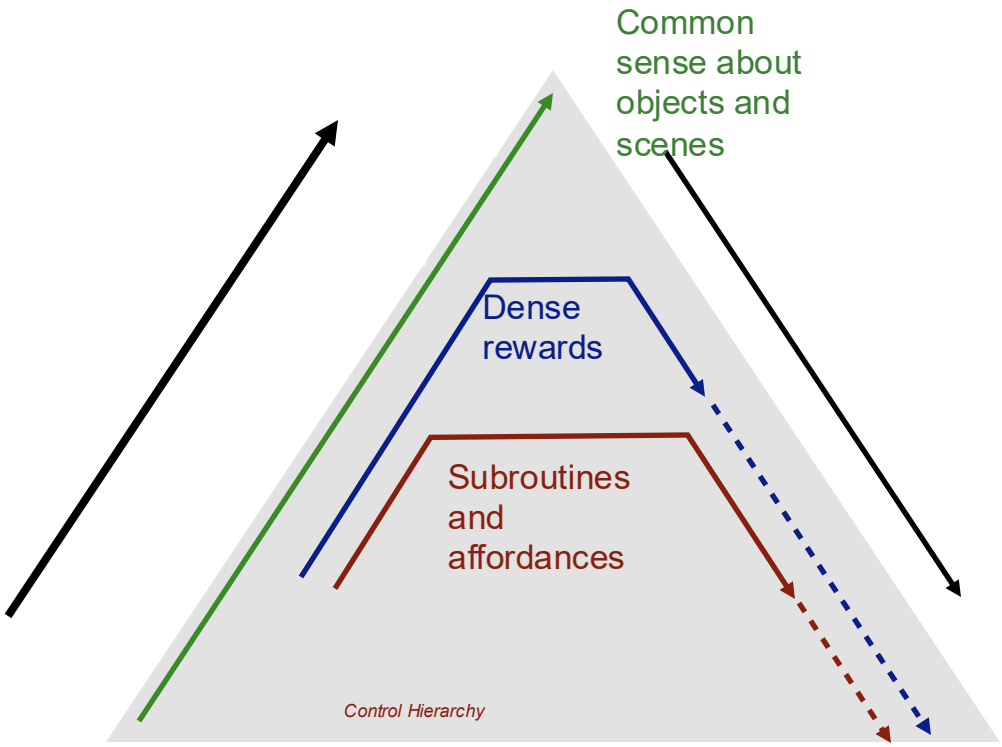
7

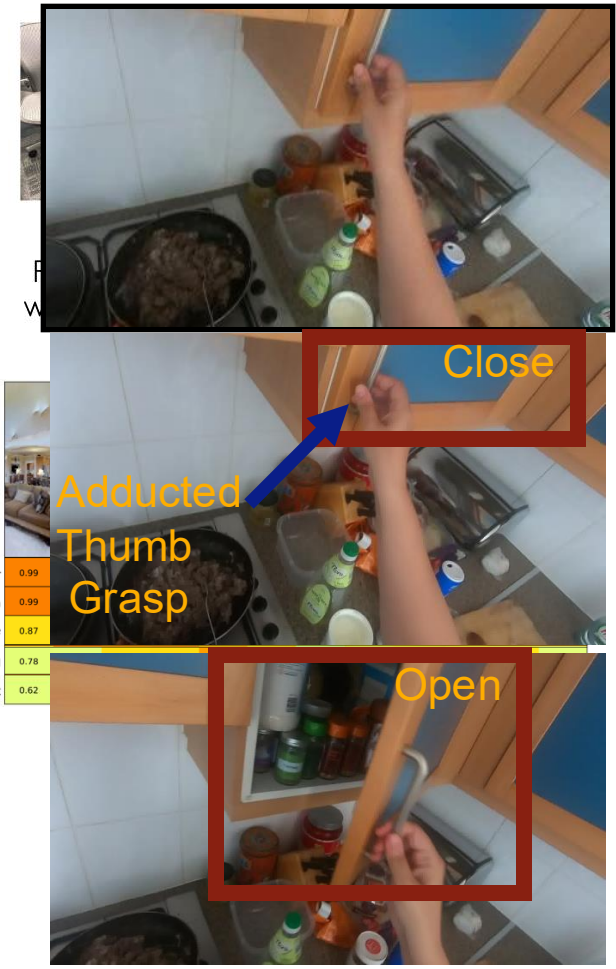
Estimate how long the visitor had been away from the family.

3 min. recordings of the same subject

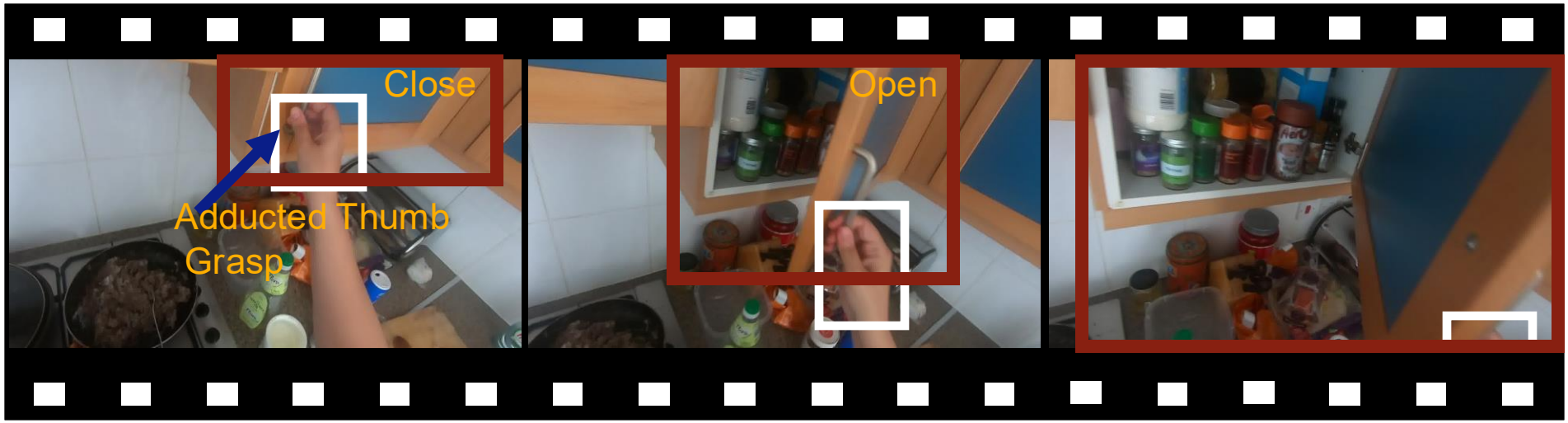
Transferring at different abstraction levels

Credit: Saurabh Gupta, UIUC





Human Hands in Egocentric Videos are Informative



1. Attending to hands localizes and stabilizes active objects.
2. Hands show where all we can interact in the scene.
3. Analyzing hands reveals information about objects: their state and how to interact with them.

Slide credit: Saurabh Gupta

Interactive Object Understanding

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

Goyal et. al. Human Hands as Probes for Interactive Object Understanding. CVPR 2022

Imitation Learning for Robot Manipulation


1 2 3 4 5
6 7 8 9 10
1 2 3 4 5
6 7 8 9 10


HaMeR Results
Contact Juggling

Reconstructing Hand-Held Objects in 3D

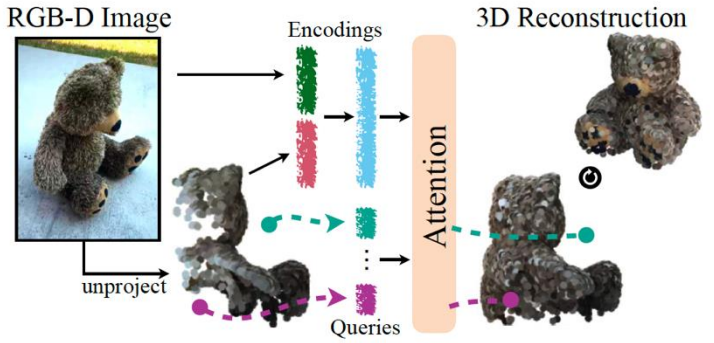
Jane Wu, Georgios Pavlakos, Georgia Gkioxari, Jitendra Malik

https://arxiv.org/abs/2404.06507

Builds on two previous foundation models : HaMeR (Berkeley) and MCC (Meta)

Multiview Compressive Coding for 3D Reconstruction

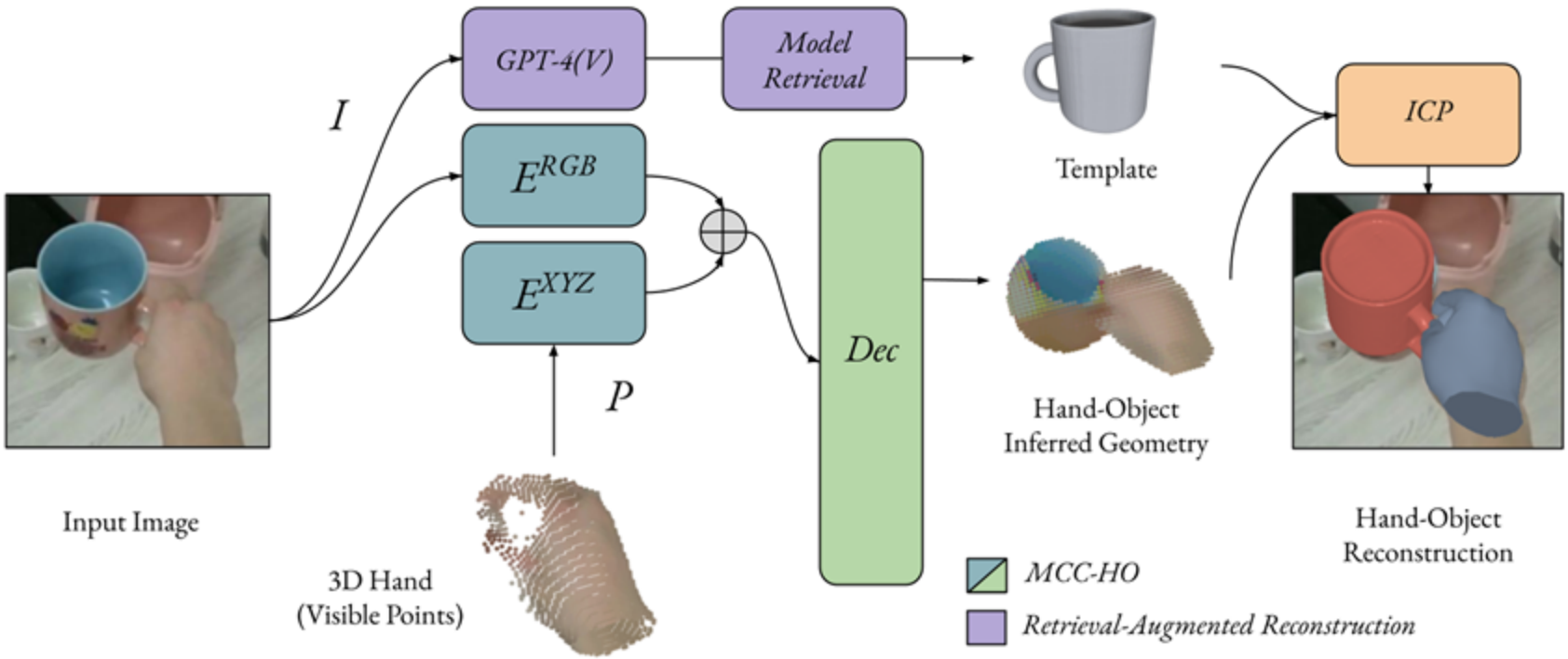
CY Wu, J Johnson, J Malik, C Feichtenhofer, G. Gkioxari

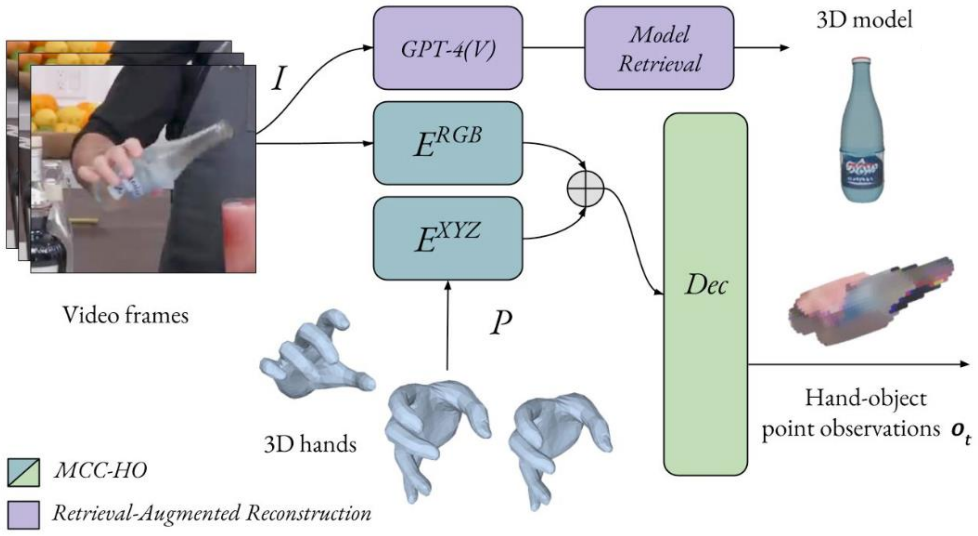


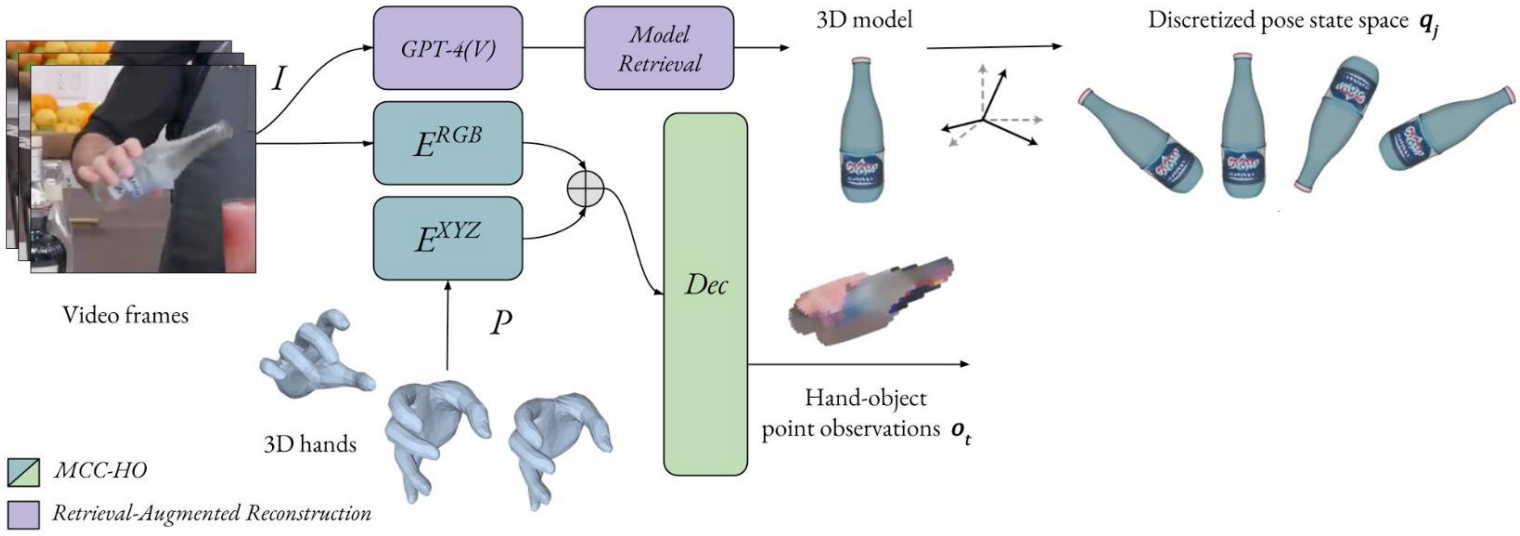
(a) **MCC Overview**

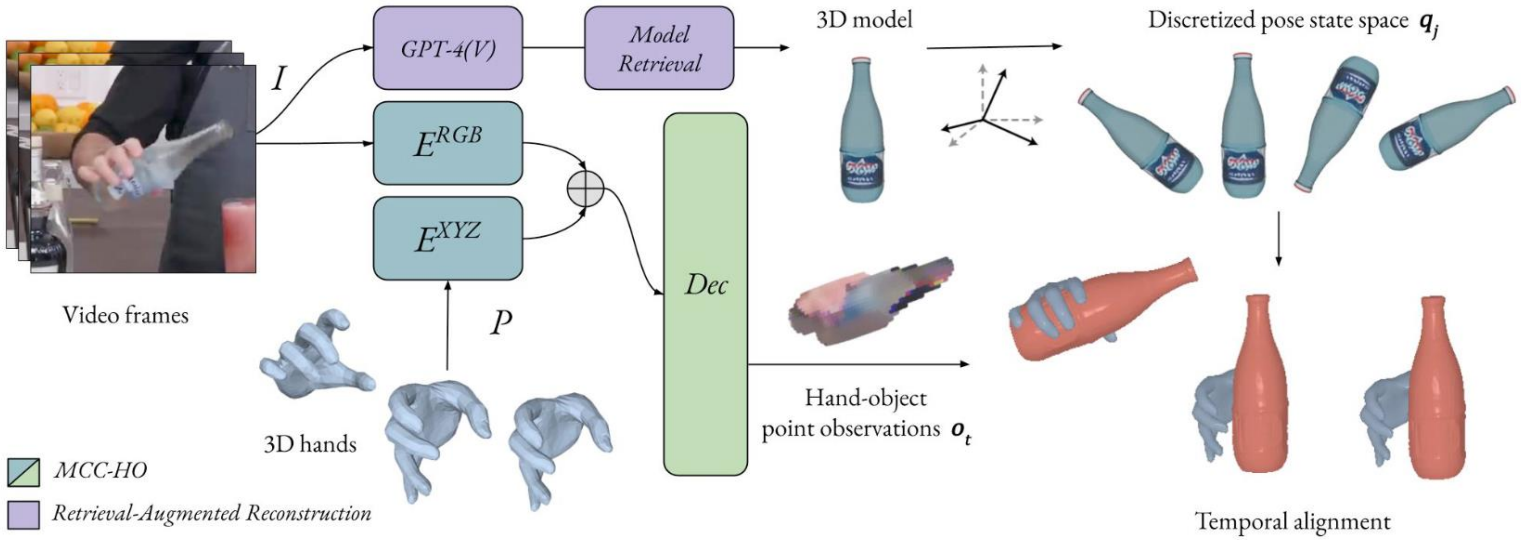
(b) **3D Reconstructions by MCC**

Figure 1. **Multiview Compressive Coding (MCC).** (a): MCC encodes an input RGB-D image and uses an attention-based model to predict the occupancy and color of query points to form the final 3D reconstruction. (b): MCC generalizes to novel objects captured with iPhones (left) or imagined by DALL·E 2 [47] (middle). It is also general – it works not only on objects but also scenes (right).


GPT-4(V)
Model Retrieval
I
E^{RGB}
E^{XYZ}
P
Dec
Template
ICP
Input Image
3D Hand
(Visible Points)
Hand-Object
Inferred Geometry
Hand-Object
Reconstruction
MCC-HO
Retrieval-Augmented Reconstruction


I
GPT-4(V)
Model Retrieval
3D model
E^{RGB}
E^{XYZ}
Dec
Video frames
P
3D hands
Hand-object point observations o_t
MCC-HO
Retrieval-Augmented Reconstruction


Video frames
I
GPT-4(V)
Model Retrieval
3D model
Discretized pose state space q_j
E^{RGB}
E^{XYZ}
P
Dec
3D hands
Hand-object point observations o_t
MCC-HO
Retrieval-Augmented Reconstruction


I
GPT-4(V)
Model Retrieval
3D model
Discretized pose state space q_j
E^{RGB}
E^{XYZ}
Dec
P
Video frames
3D hands
Hand-object point observations o_t
Temporal alignment
MCC-HO
Retrieval-Augmented Reconstruction

Inferred
Point Clouds

	DexYCB			MOW			HOI4D		
	F-5 (↑)	F-10 (↑)	CD (↓)	F-5 (↑)	F-10 (↑)	CD (↓)	F-5 (↑)	F-10 (↑)	CD (↓)
HO [24]	0.24	0.48	4.76	0.03	0.06	49.8	0.28	0.51	3.86
IHOI [90]	-	-	-	0.13	0.24	23.1	0.42	0.70	2.7
MCC-HO	**0.36**	**0.60**	**3.74**	**0.15**	**0.31**	**15.2**	**0.52**	**0.78**	**1.36**

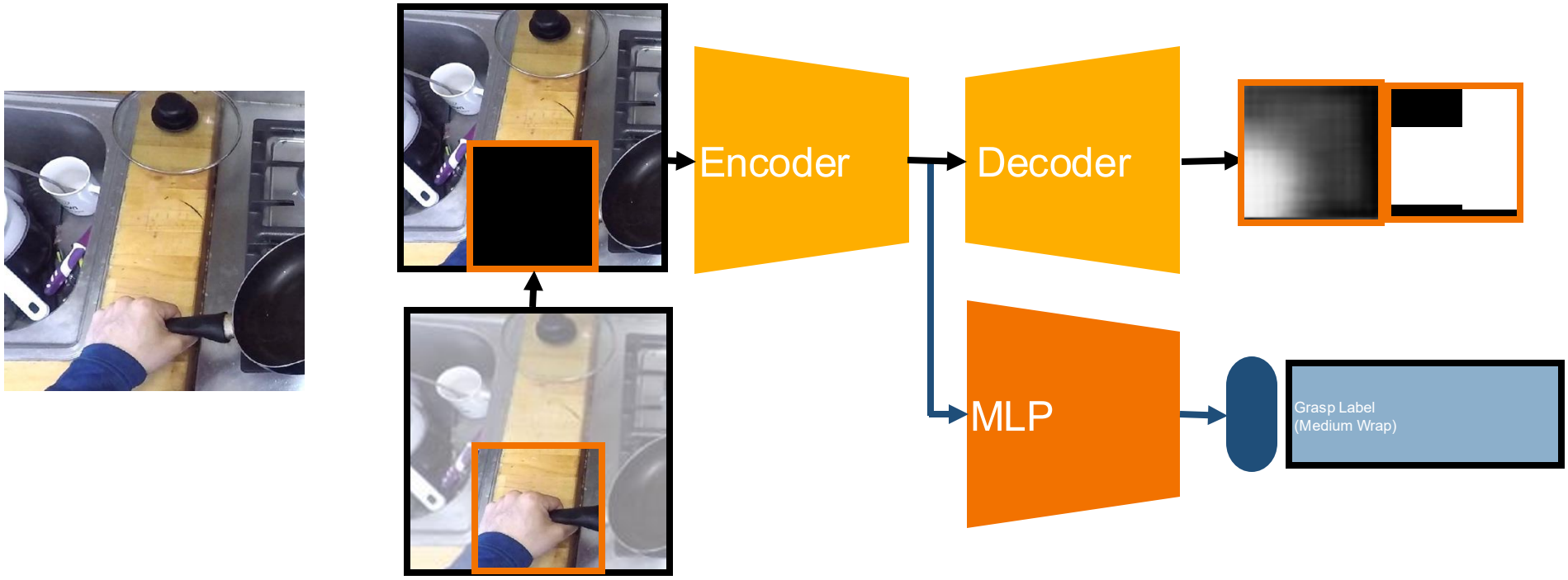
Table 2: We compare our method, MCC-HO, to prior works on held-out test images from DexYCB, MOW, and HOI4D. Chamfer Distance (cm^2) and F-score (5mm, 10mm) are reported.

Comparison to prior work. Chamfer Distance (cm^2) and F-score (5mm, 10mm).

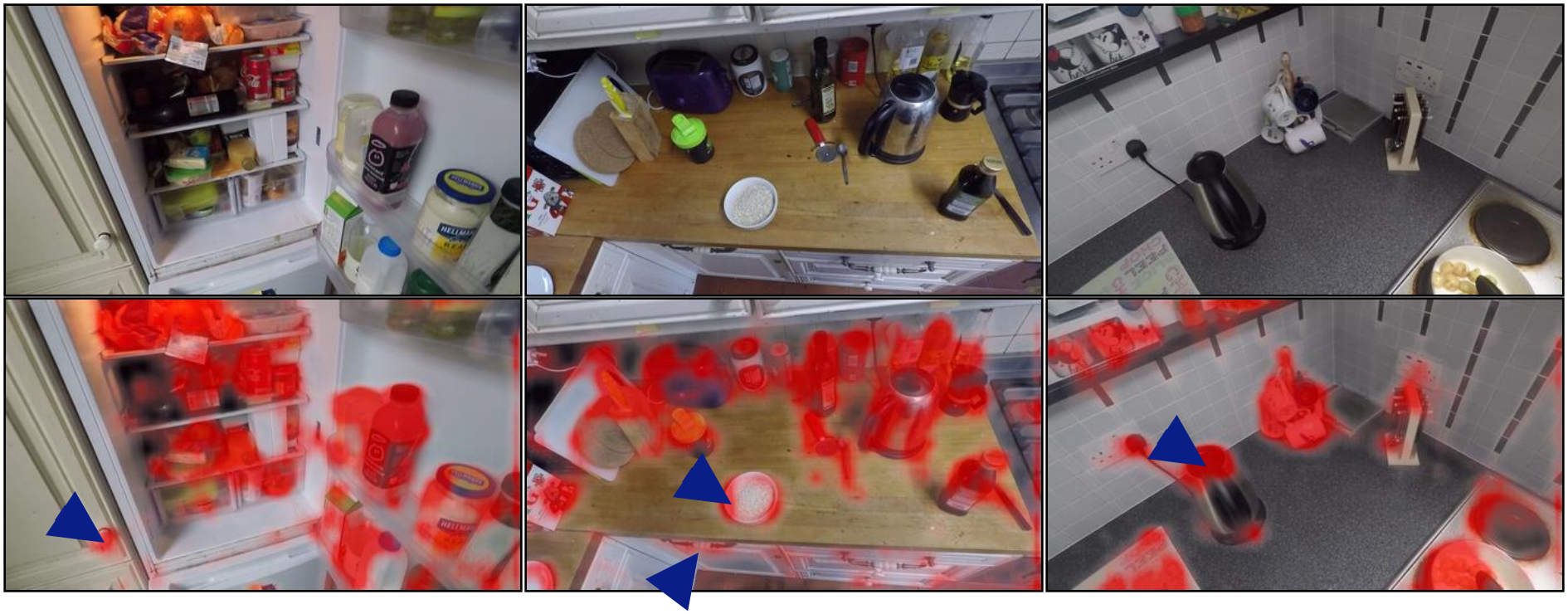

CTIVATOR SMOOTHIES
SMOOTHIEKING.COM

A,B) Learning Object Affordances

A,B) Learning Object Affordances



Learning Object Affordances: Results



Learning Object Affordances: Results

Interactive Object Understanding

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

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

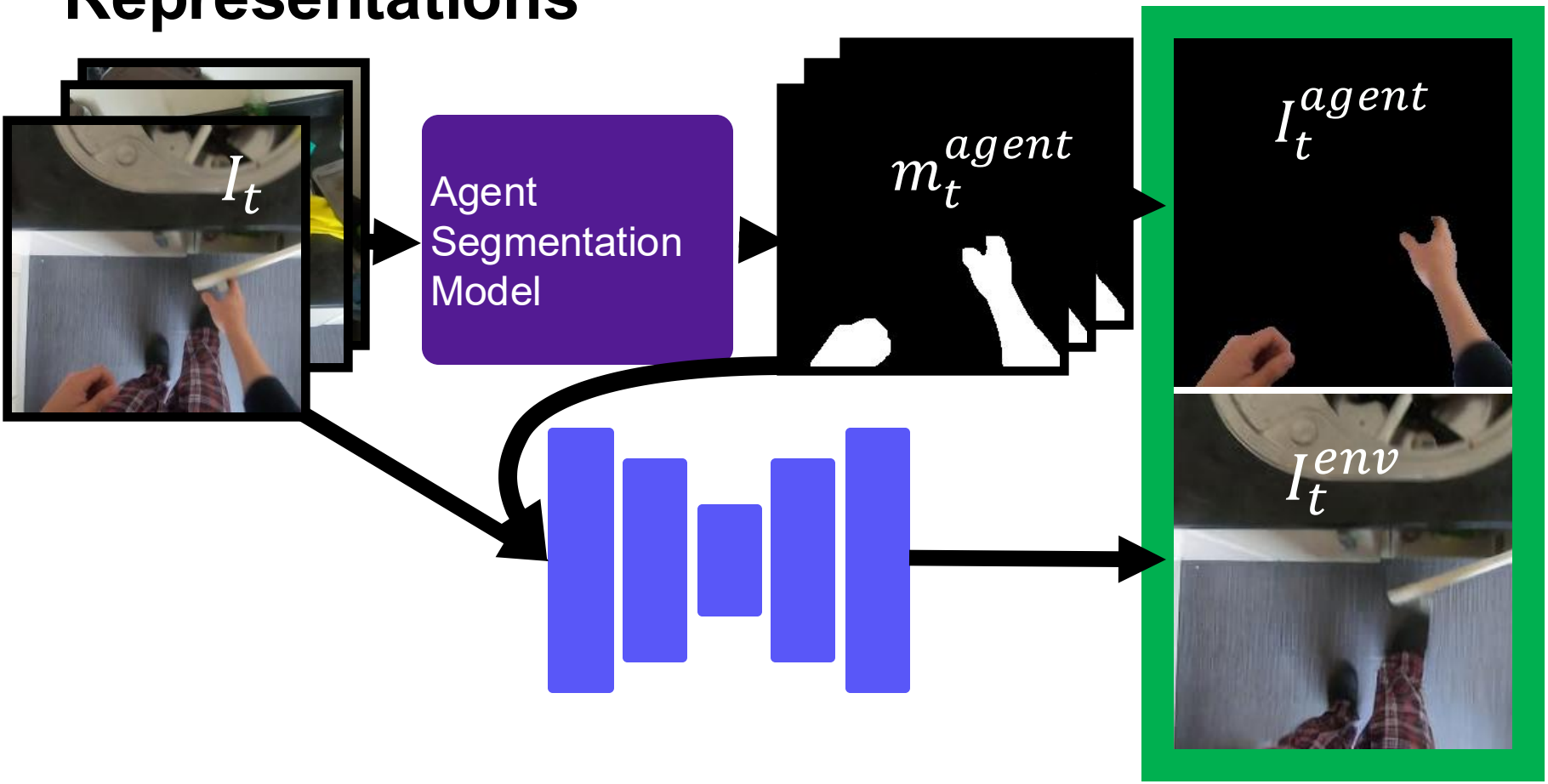
Slide credit: Saurabh Gupta

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

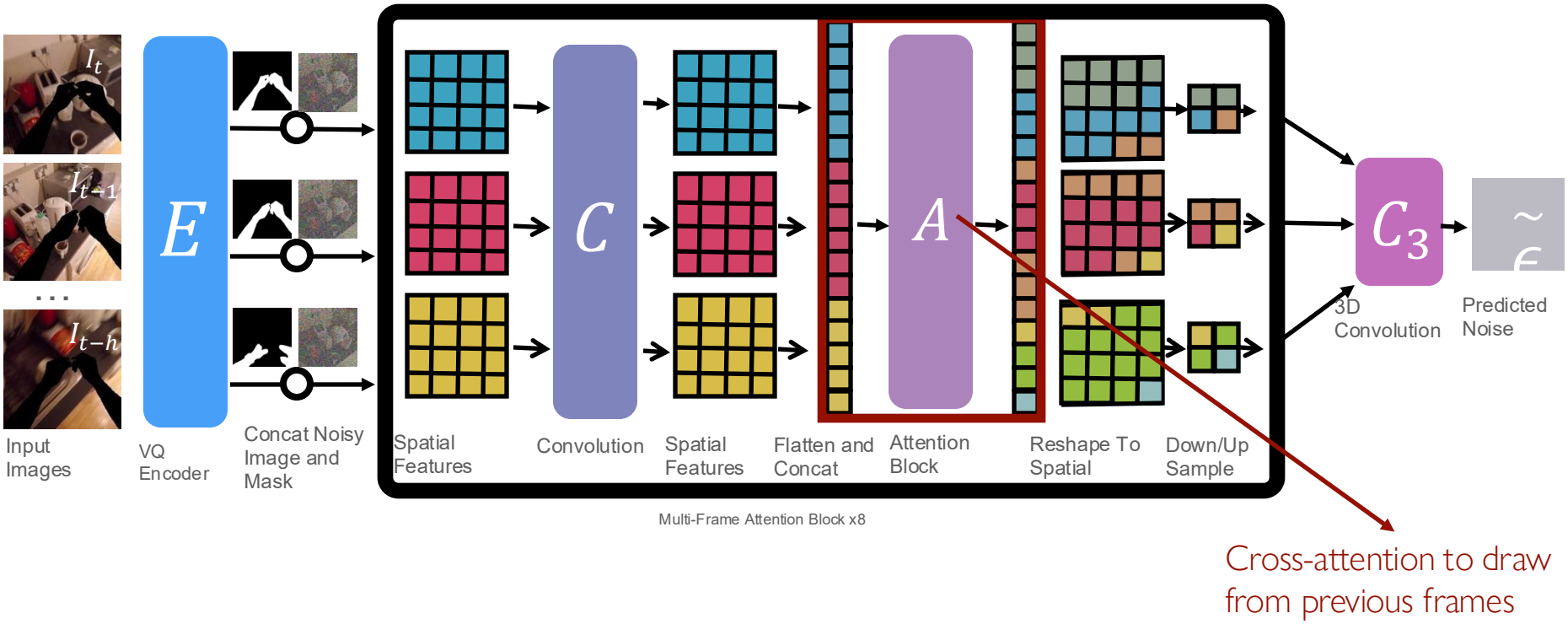
- Matthew Chang, Aditya Prakash, Saurabh Gupta

ILLINOIS

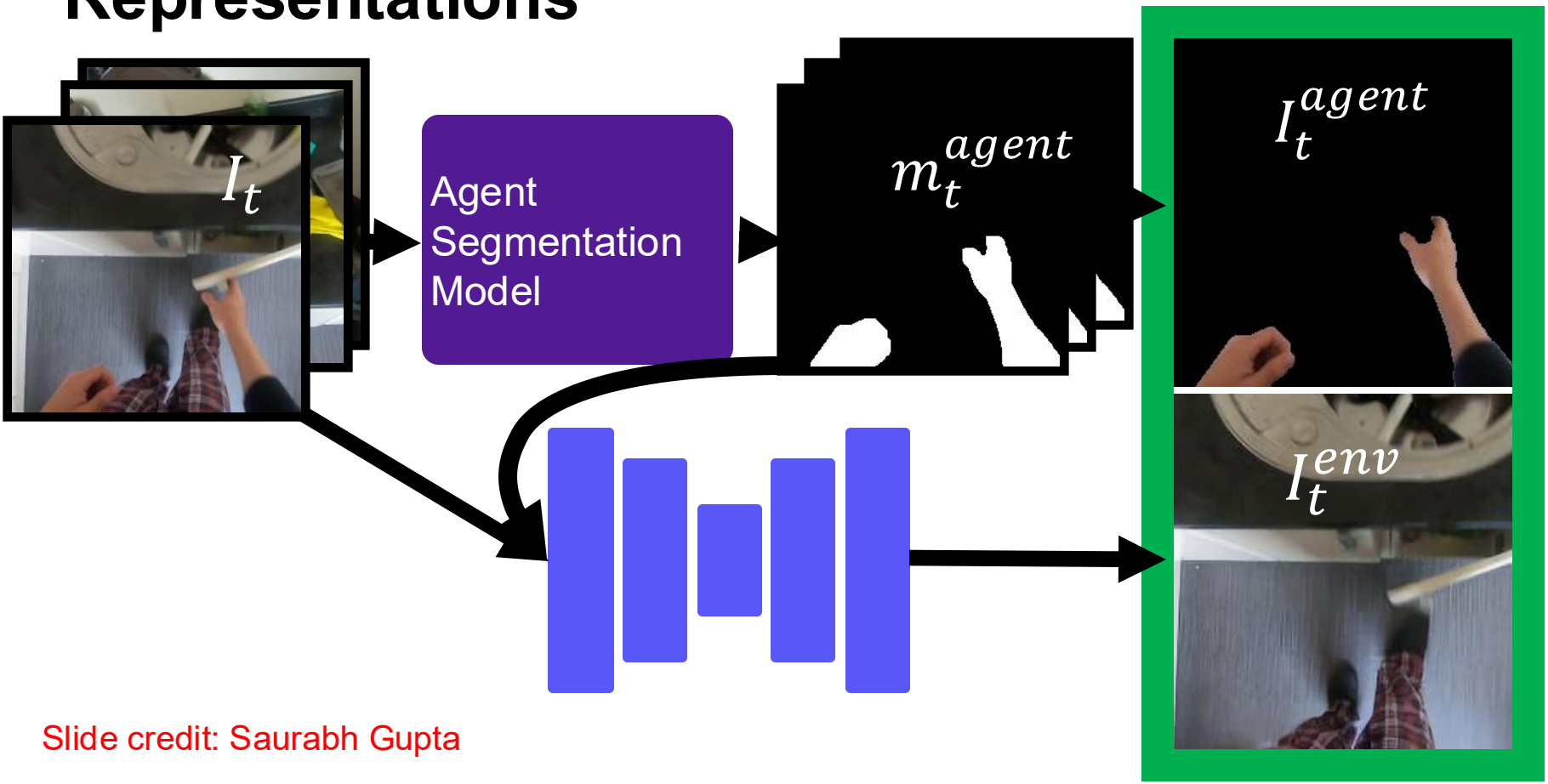
Agent-Environment *Factored* Representations



Video Inpainting Diffusion Model (VIDM)

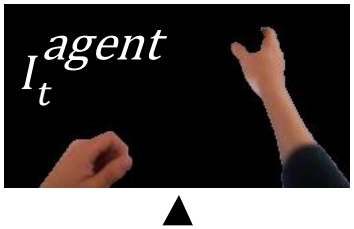


Agent-Environment *Factored* Representations



Slide credit: Saurabh Gupta

Using the *Factored* Representations

g 

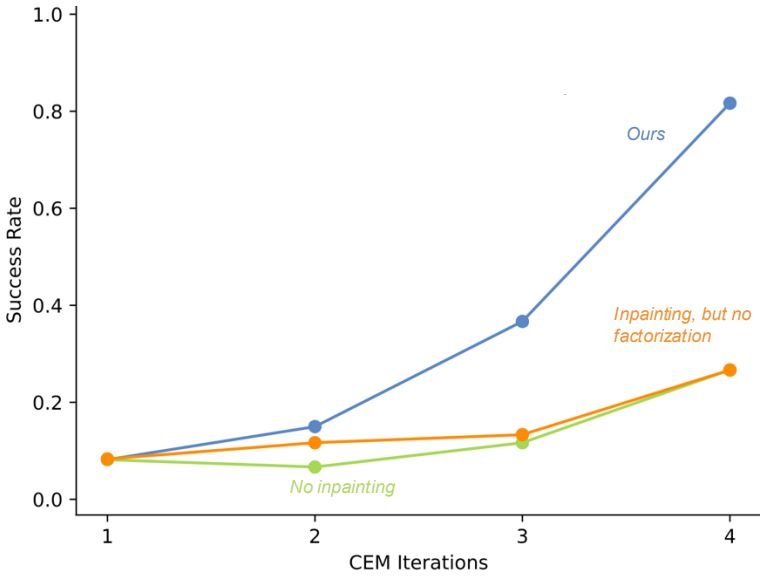
f 

Learn reward function from human data

$\Rightarrow$ Use it to train robots

Slide credit: Saurabh Gupta

Slide credit: Saurabh G



Affordances from Human Videos as a Versatile Representation for Robotics

Shikhar Bahl*[1,2] Russell Mendonca*[1] Lili Chen[1] Unnat Jain[1,2] Deepak Pathak[1]

[1]CMU [2]Meta AI

Figure 1. We leverage human videos to learn visual affordances that can be deployed on multiple real robot, in the wild, spanning several tasks and learning paradigms. Videos available at https://vision-robotics-bridge.github.io/.

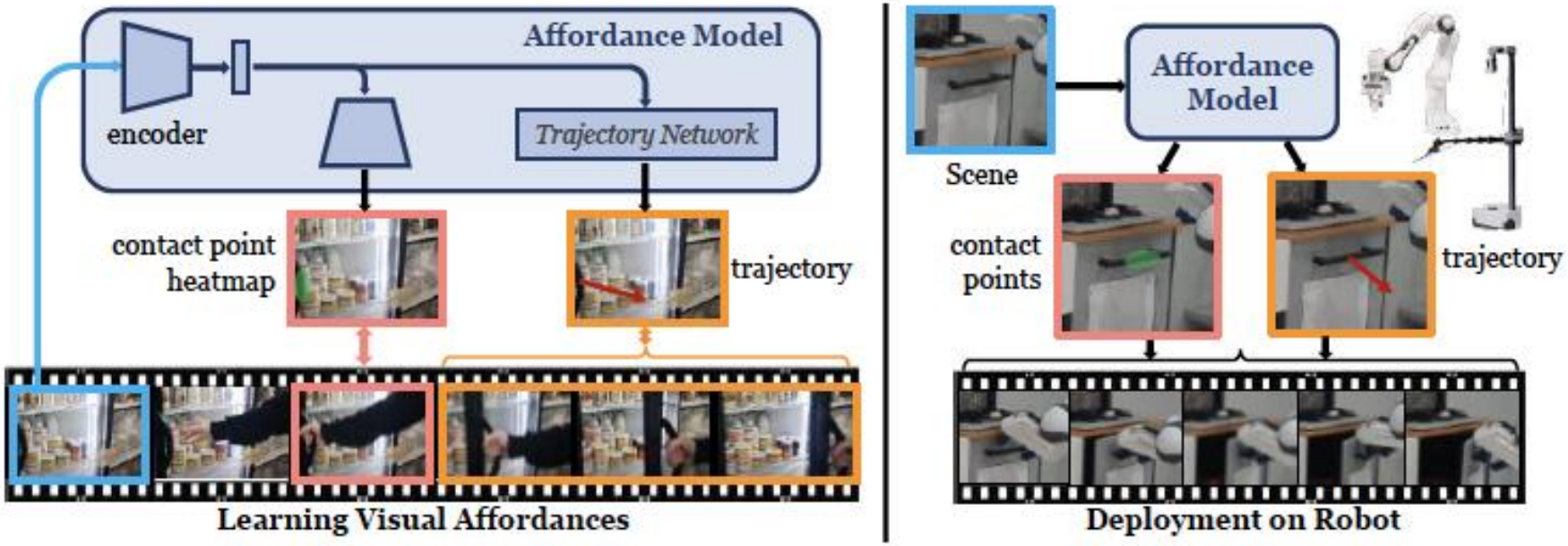


Figure 2. **VRB Overview**. First, we learn an actionable representation of visual affordances from human videos: the model predicts contact points and trajectory waypoints with supervision from future frames. For robot deployment, we query the affordance model and convert its outputs to 3D actions to execute.