MODERN RL LANDSCAPE: PART 1

Scott Niekum

Assistant Professor, Department of Computer Science

The University of Texas at Austin

The University of Texas at Austin

Personal Autonomous Robotics Lab

Distributional RL (Bellemare et al. 2017)

$$Q(x,a) = \mathbb{E}\, R(x,a) + \gamma\, \mathbb{E}\, Q(X', A').$$

vs.

$$Z(x,a) \overset{D}{=} R(x,a) + \gamma Z(X', A').$$

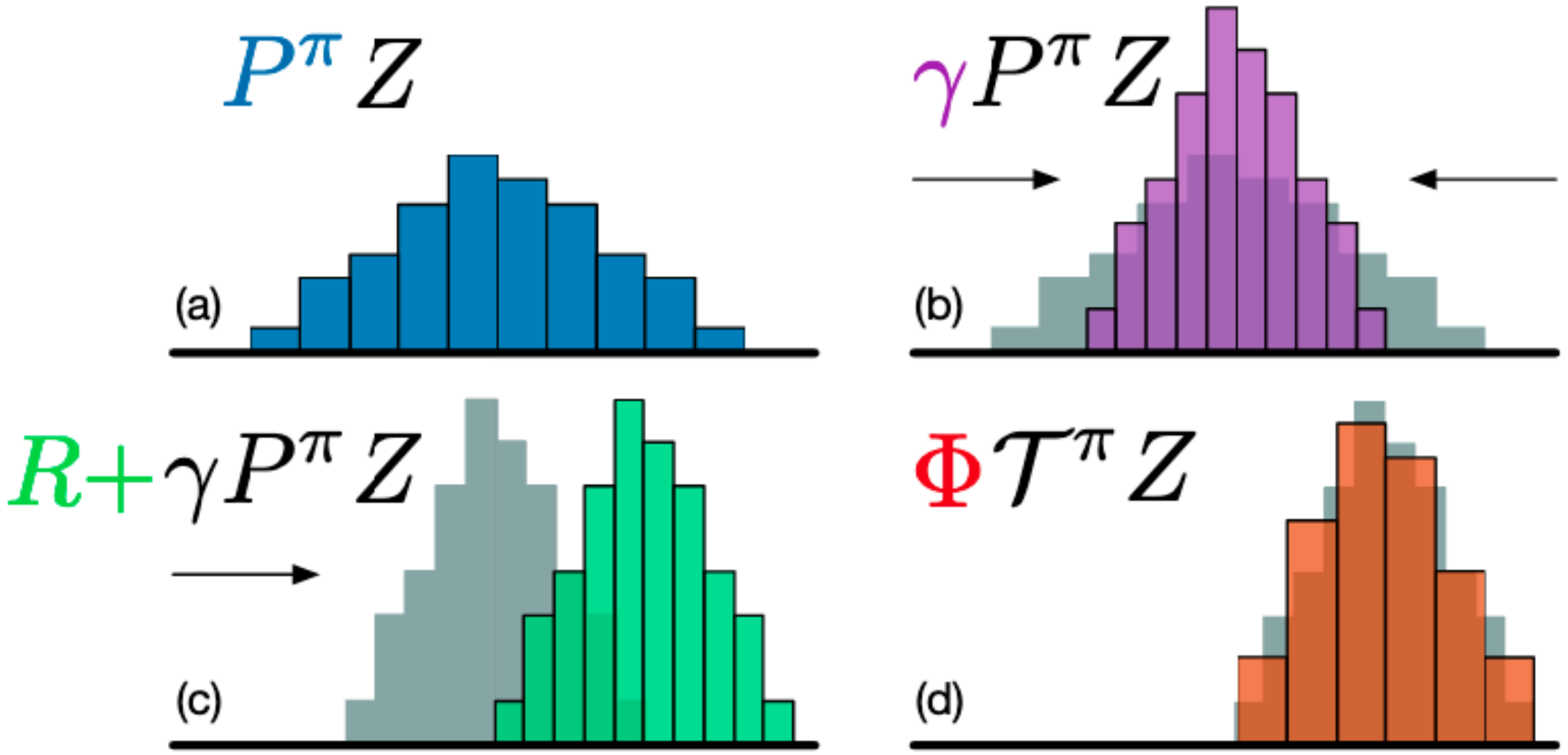


Figure 1. A distributional Bellman operator with a deterministic reward function: (a) Next state distribution under policy π, (b) Discounting shrinks the distribution towards 0, (c) The reward shifts it, and (d) Projection step (Section 4).

Distributional RL (Bellemare et al. 2017)

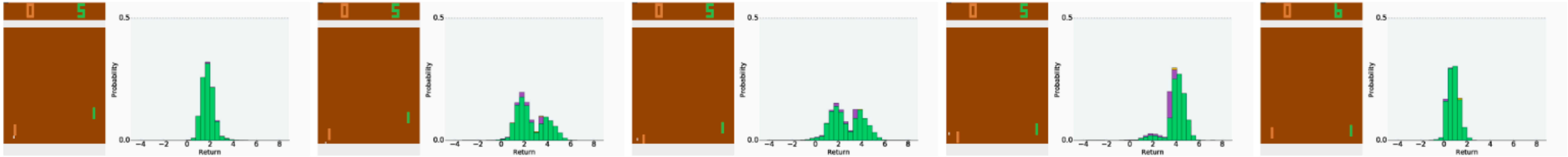


Figure 5. Intrinsic stochasticity in PONG.

Distributional RL (Bellemare et al. 2017)

	Mean	Median	> H.B.	> DQN
DQN	228%	79%	24	0
DDQN	307%	118%	33	43
DUEL.	373%	151%	37	50
PRIOR.	434%	124%	39	48
PR. DUEL.	592%	172%	39	44
C51	**701%**	**178%**	**40**	**50**
UNREAL†	880%	250%	-	-

Figure 6. Mean and median scores across 57 Atari games, measured as percentages of human baseline (H.B., Nair et al., 2015).

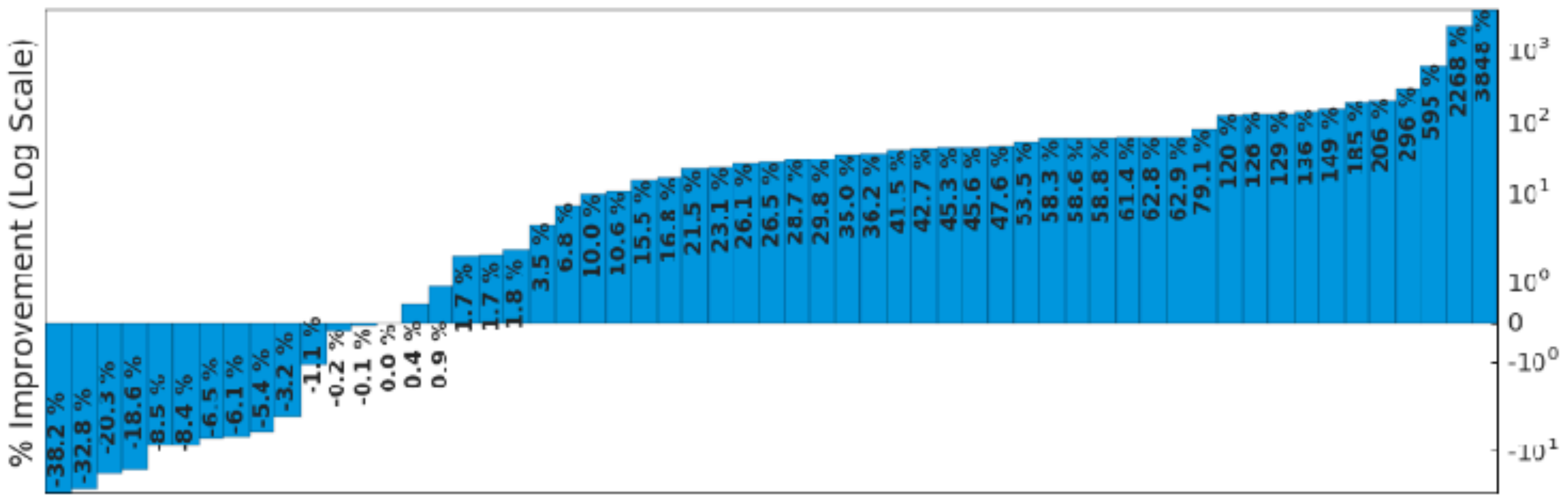


Figure 7. Percentage improvement, per-game, of C51 over Double DQN, computed using van Hasselt et al.'s method.

What is distributional RL doing? (Lyle et al. 2019)

- Reduces chattering?
- Stabilizes updates, handles nonstationarity?
- Good auxiliary task?

What is distributional RL doing? (Lyle et al. 2019)

- Identical expectations computed in most tabular and linear approx cases
- And when predictions are different, actually hurts performance often!
- But usually helps with nonlinear function approximation (e.g. DNN)
- Good auxiliary task for representation learning /regularization?

What is meta-learning?

- If you've learned 100 tasks already, can you figure out how to *learn* more efficiently?
  - Now having multiple tasks is a huge advantage!
- Meta-learning = *learning to learn*
- In practice, very closely related to multi-task learning
- Many formulations
  - Learning an optimizer
  - Learning an RNN that ingests experience
  - Learning a representation

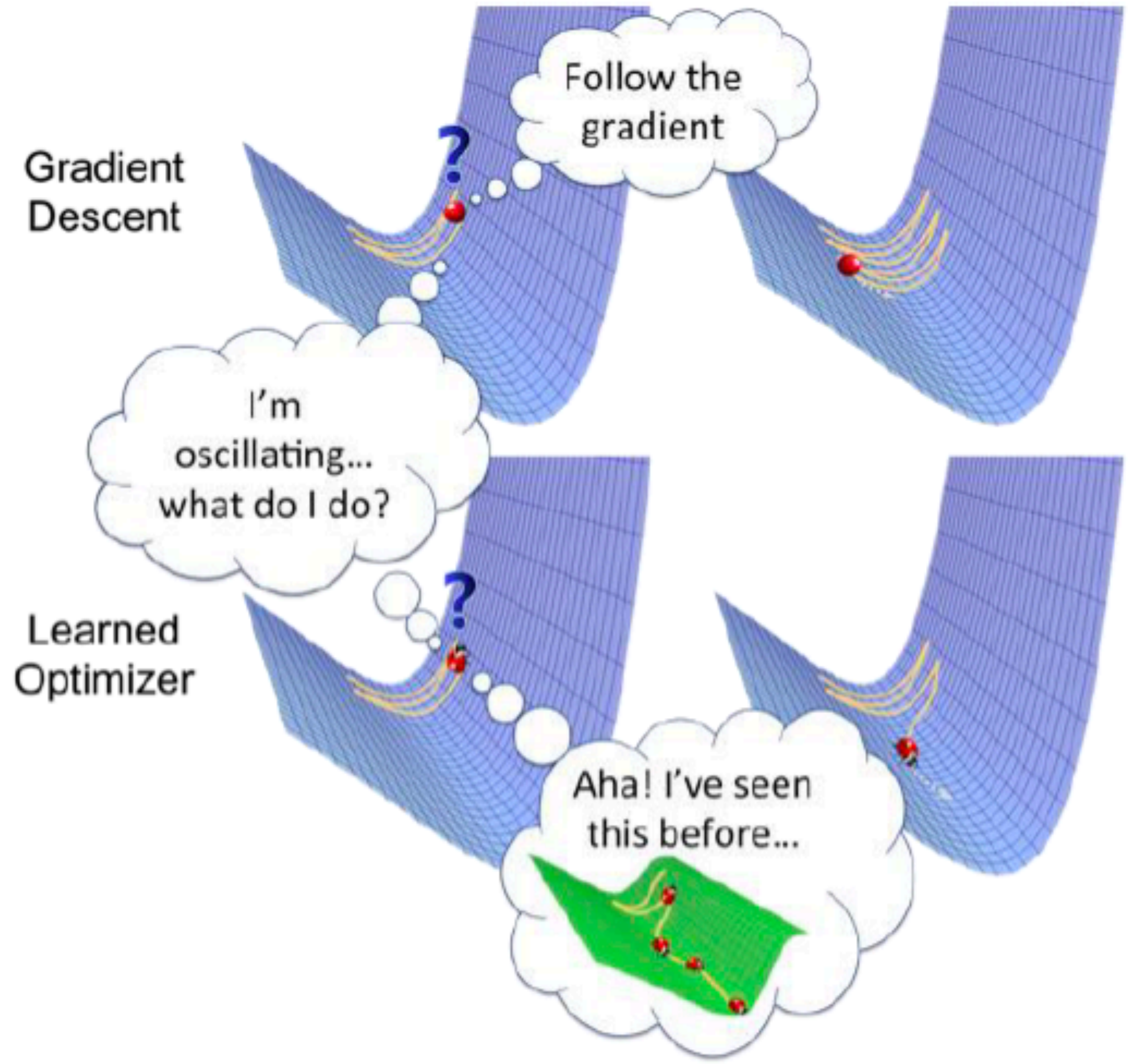


image credit: Ke Li

Slide credit: Sergey Levine

Why is meta-learning a good idea?

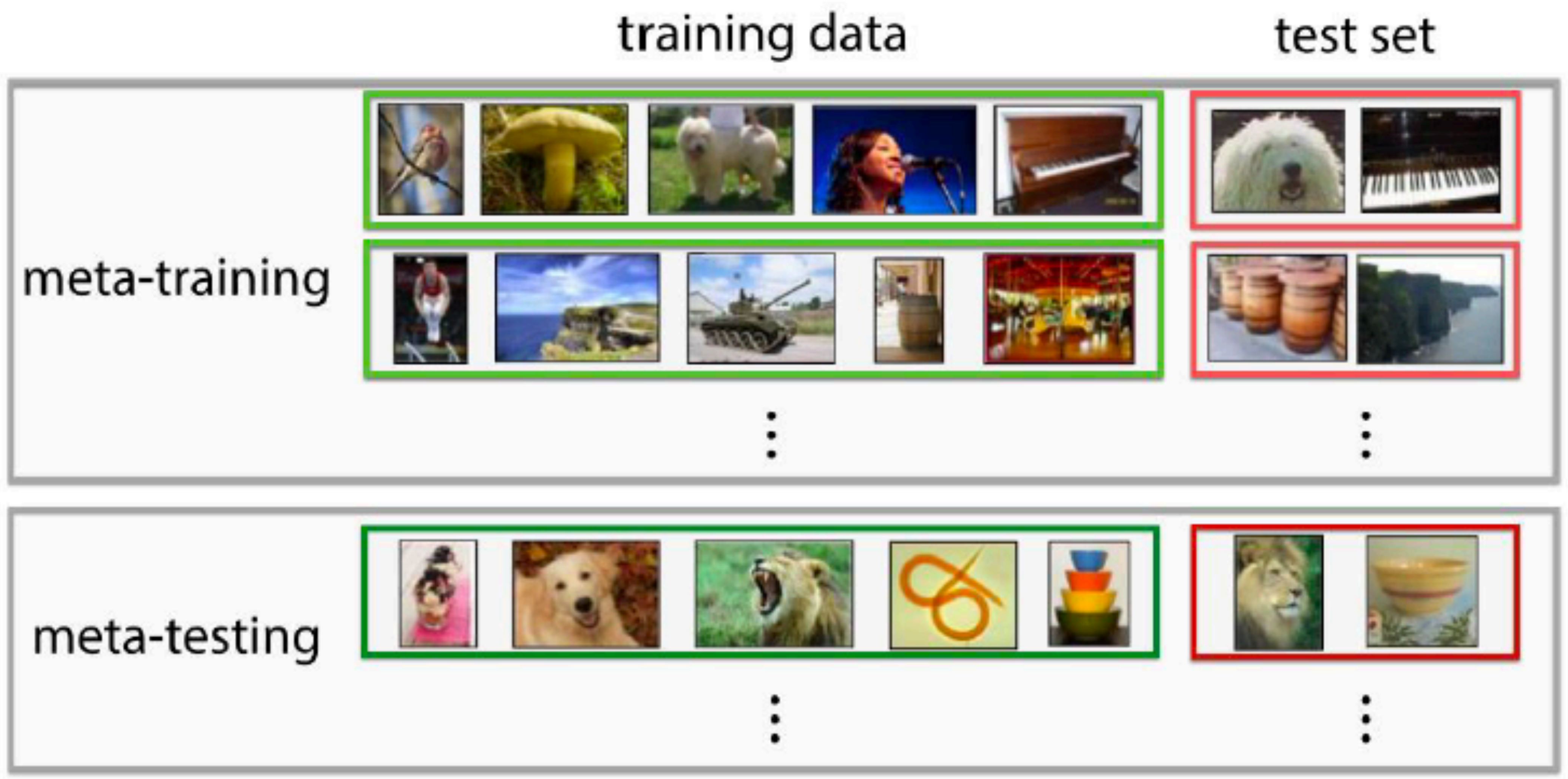
- Deep reinforcement learning, especially model-free, requires a huge number of samples
- If we can *meta-learn* a faster reinforcement learner, we can learn new tasks efficiently!
- What can a *meta-learned* learner do differently?
  - Explore more intelligently
  - Avoid trying actions that are know to be useless
  - Acquire the right features more quickly

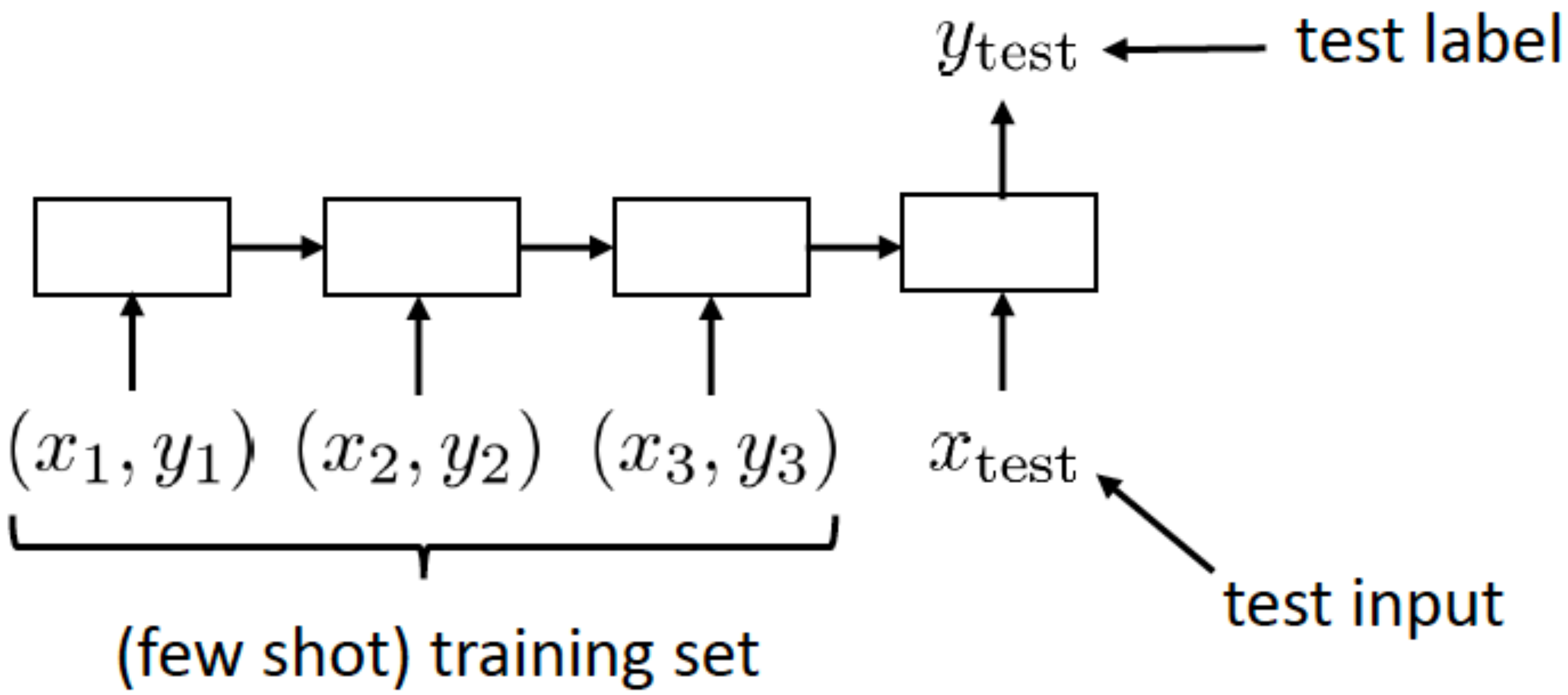
Slide credit: Sergey Levine

Meta-learning with supervised learning

image credit: Ravi & Larochelle '17

Meta-learning with supervised learning





supervised learning: $f(x) \rightarrow y$

input (e.g., image) output (e.g., label)

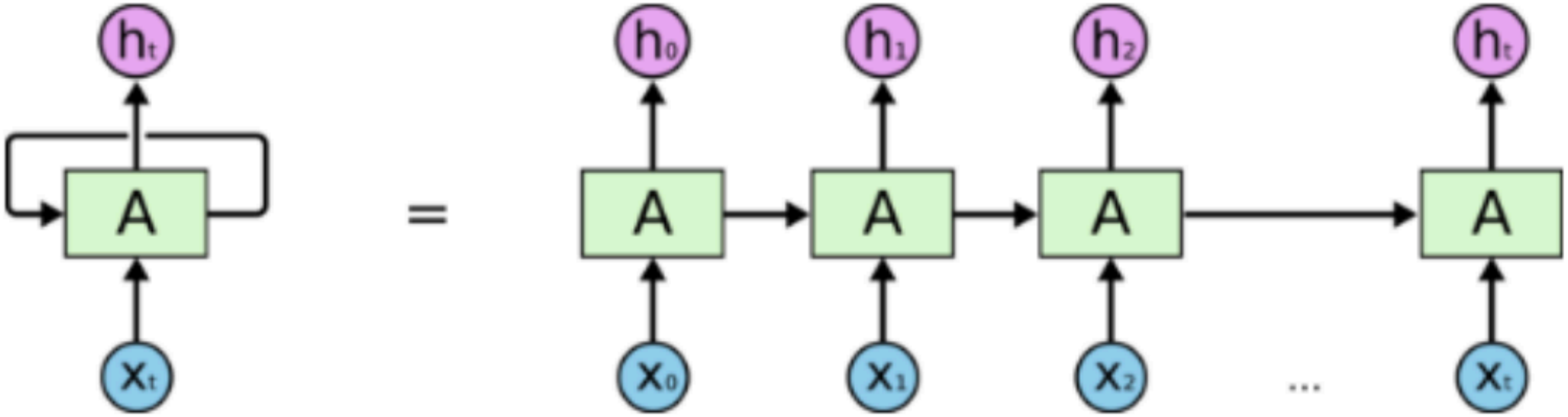
supervised meta-learning: $f(\mathcal{D}_{\text{train}}, x) \rightarrow y$

training set

- How to read in training set?
  - Many options, RNNs can work
  - More on this later

Slide credit: Sergey Levine

RNN-based meta-learning



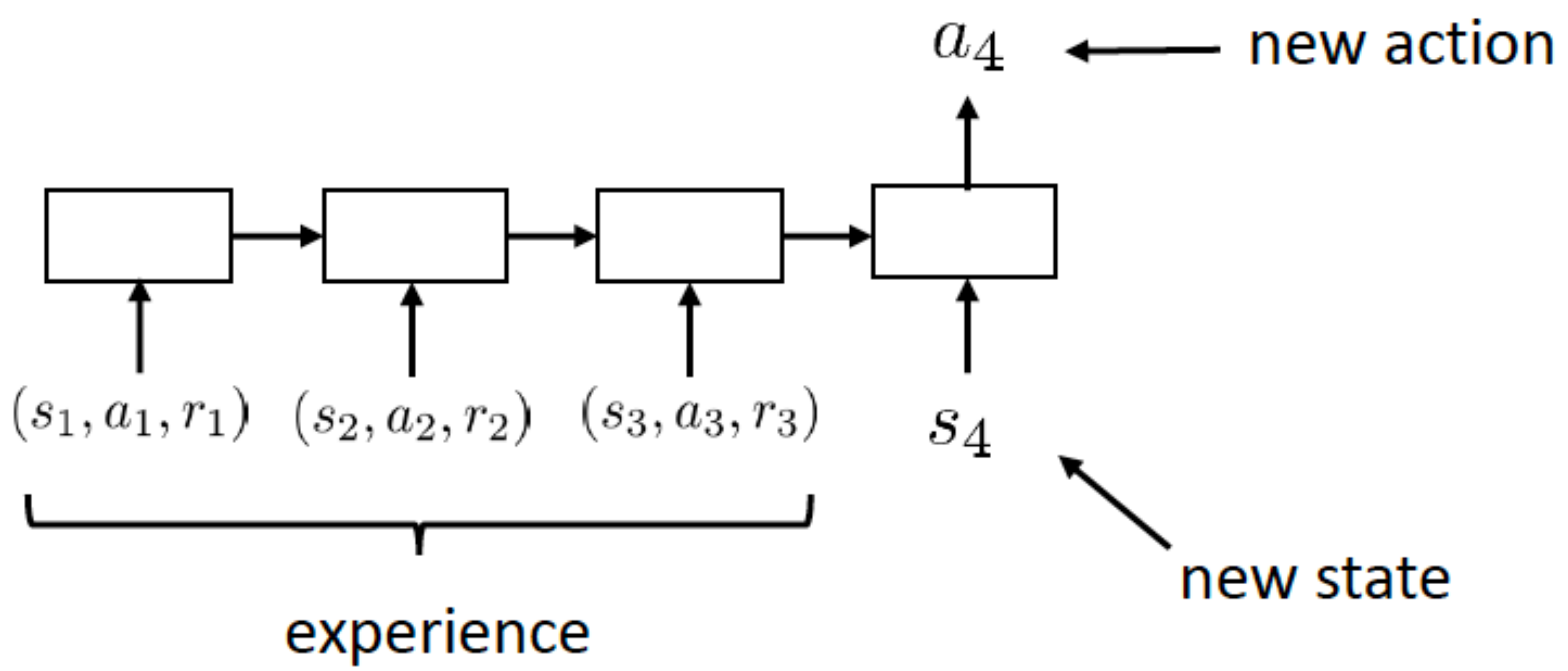
The meta-learning problem in RL

supervised meta-learning: $f(\mathcal{D}_{\text{train}}, x) \rightarrow y$

reinforcement meta-learning (for example...): $f(\mathcal{D}_{\text{train}}, s) \rightarrow a$

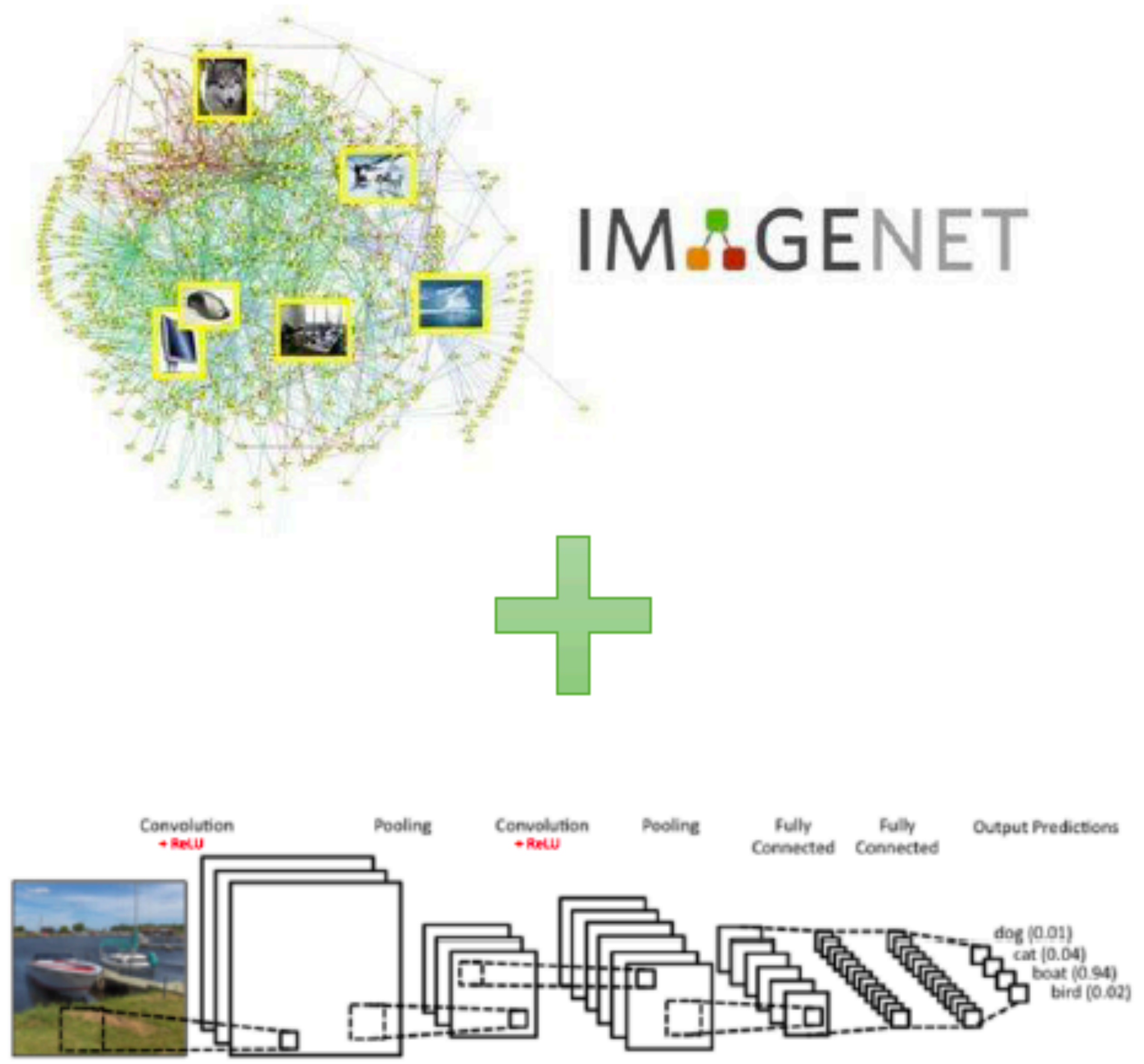
recent experience state output (e.g., action)

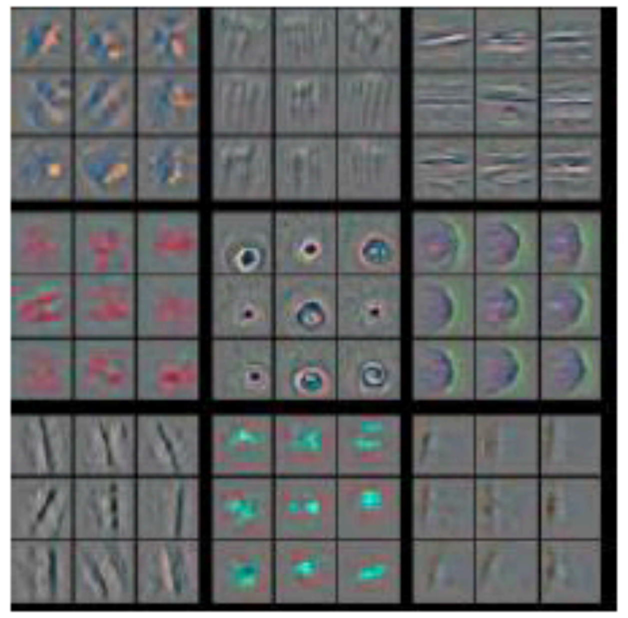
$$\mathcal{D}_{\text{train}} = \{s_1, a_1, r_1, \ldots, a_N, s_N, r_N\}$$

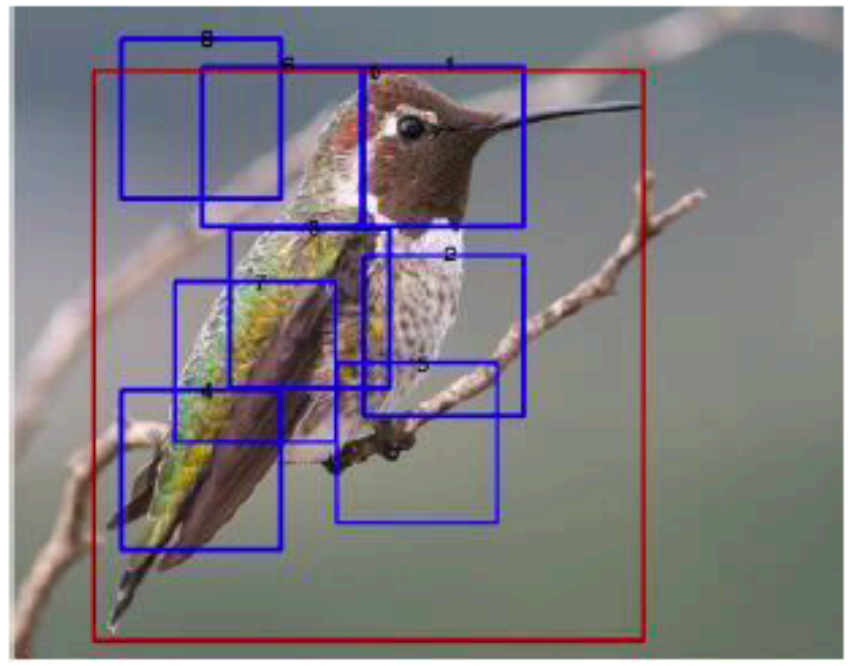


Slide credit: Sergey Levine

Back to representations…





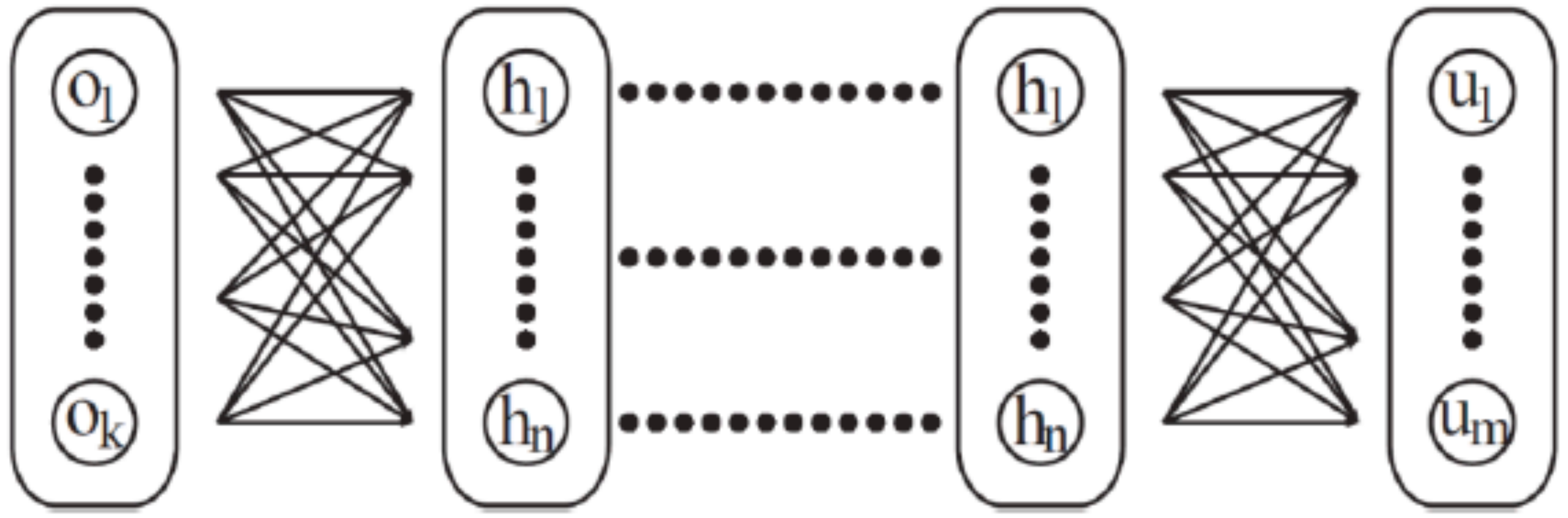


is pretraining a *type* of meta-learning?

better features = faster learning of new task!

Slide credit: Sergey Levine

Preparing a model for faster learning

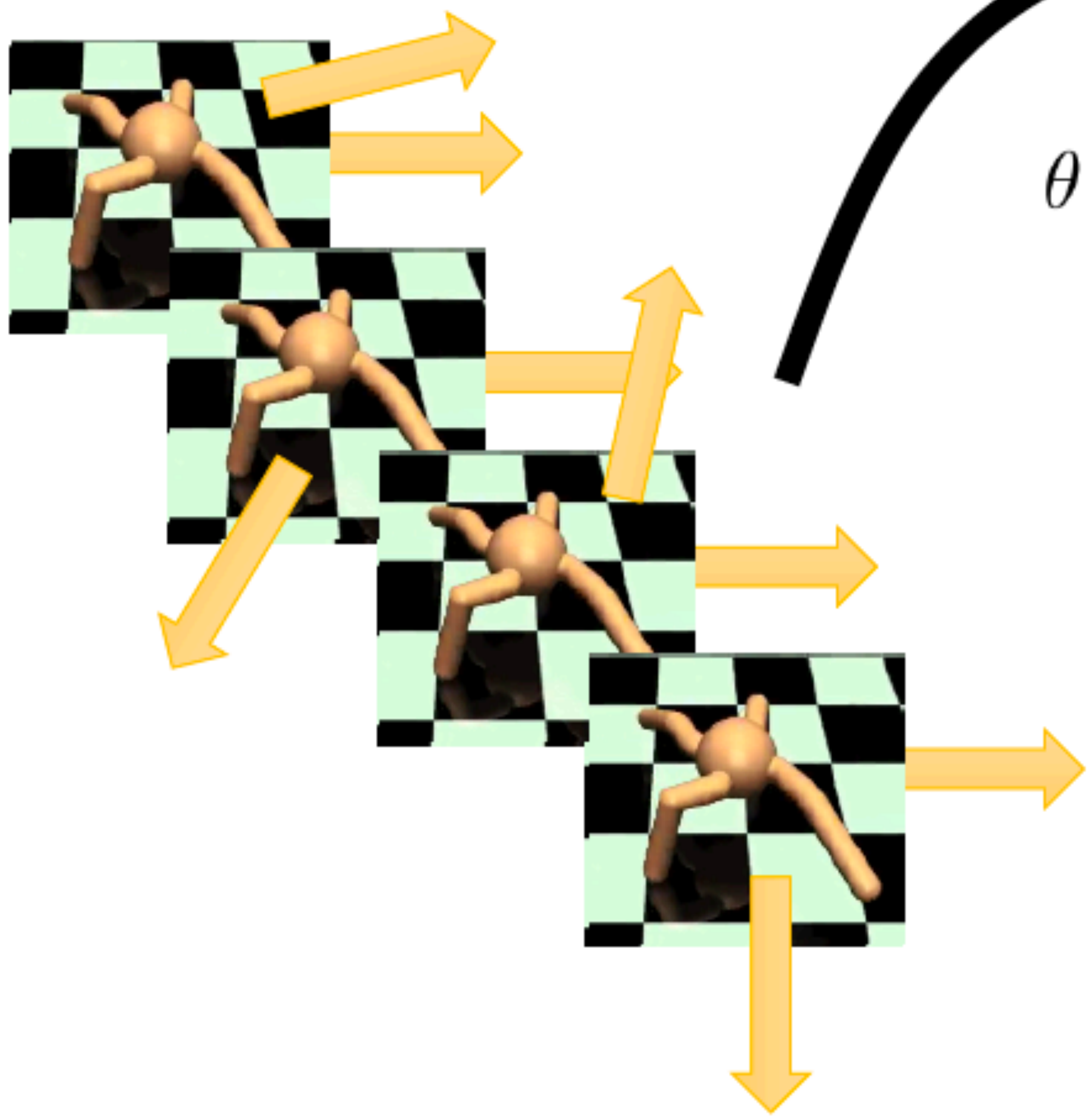


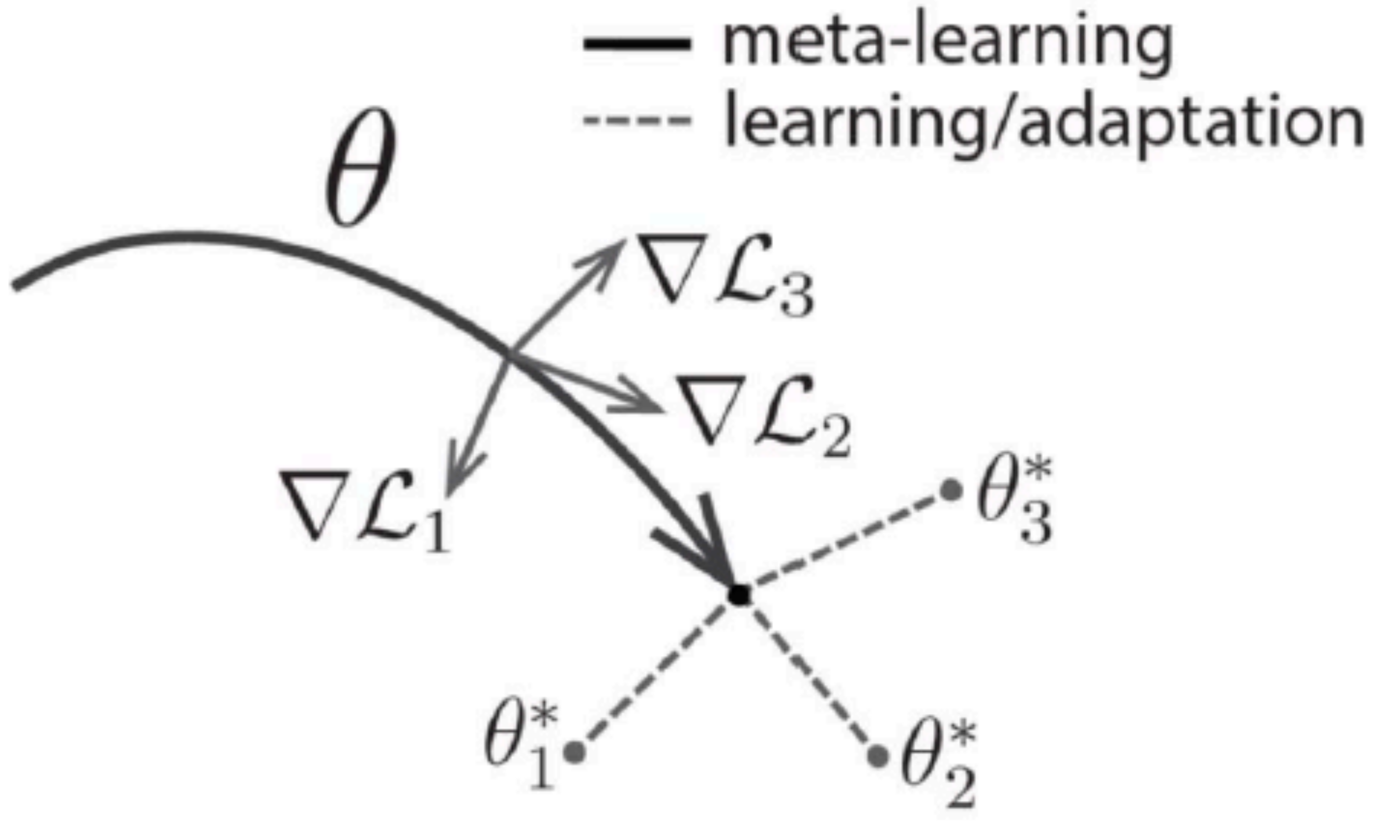
θ

$$\cancel{\theta \leftarrow \theta + \alpha \nabla_\theta R(\theta)}$$

$$\theta \leftarrow \theta + \alpha \sum_i \nabla_\theta R_i[\theta + \alpha \nabla_\theta R_i(\theta)]$$

θ'





Finn et al., “Model-Agnostic Meta-Learning”

Slide credit: Sergey Levine

Meta-learning summary & open problems

- Meta-learning = learning to learn
- Supervised meta-learning = supervised learning with datapoints that are entire datasets
- RL meta-learning with RNN policies
  - Ingest past experience with RNN
  - Simply run forward pass at test time to “learn”
  - Just contextual policies (no actual learning)
- Model-agnostic meta-learning
  - Use gradient descent (e.g., policy gradient) learning rule
  - Conceptually not that different
  - …but can accelerate standard RL algorithms (e.g., learn in one iteration of PG)

Slide credit: Sergey Levine

Meta-learning summary & open problems

- The promise of meta-learning: use past experience to simply acquire a much more efficient deep RL algorithm
- The reality of meta-learning: mostly works well on smaller problems
- …but getting better all the time
- Main limitations
  - RNN policies are extremely hard to train, and likely not scalable
  - Model-agnostic meta-learning presents a tough optimization problem
  - Designing the right task distribution is hard
  - Generally very sensitive to task distribution (meta-overfitting)

Slide credit: Sergey Levine

Why not just initialize parameters to those that give the best average performance across tasks?

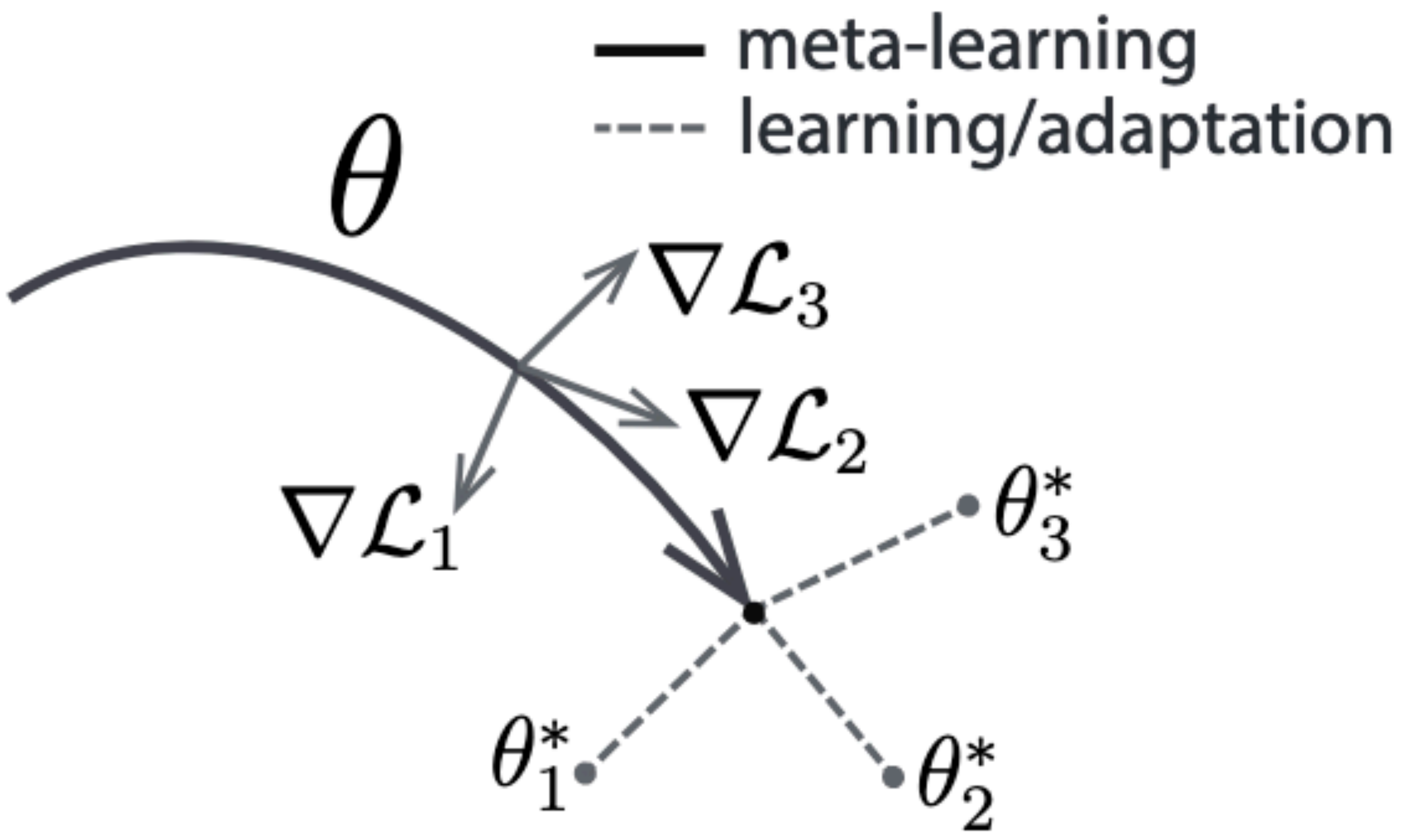


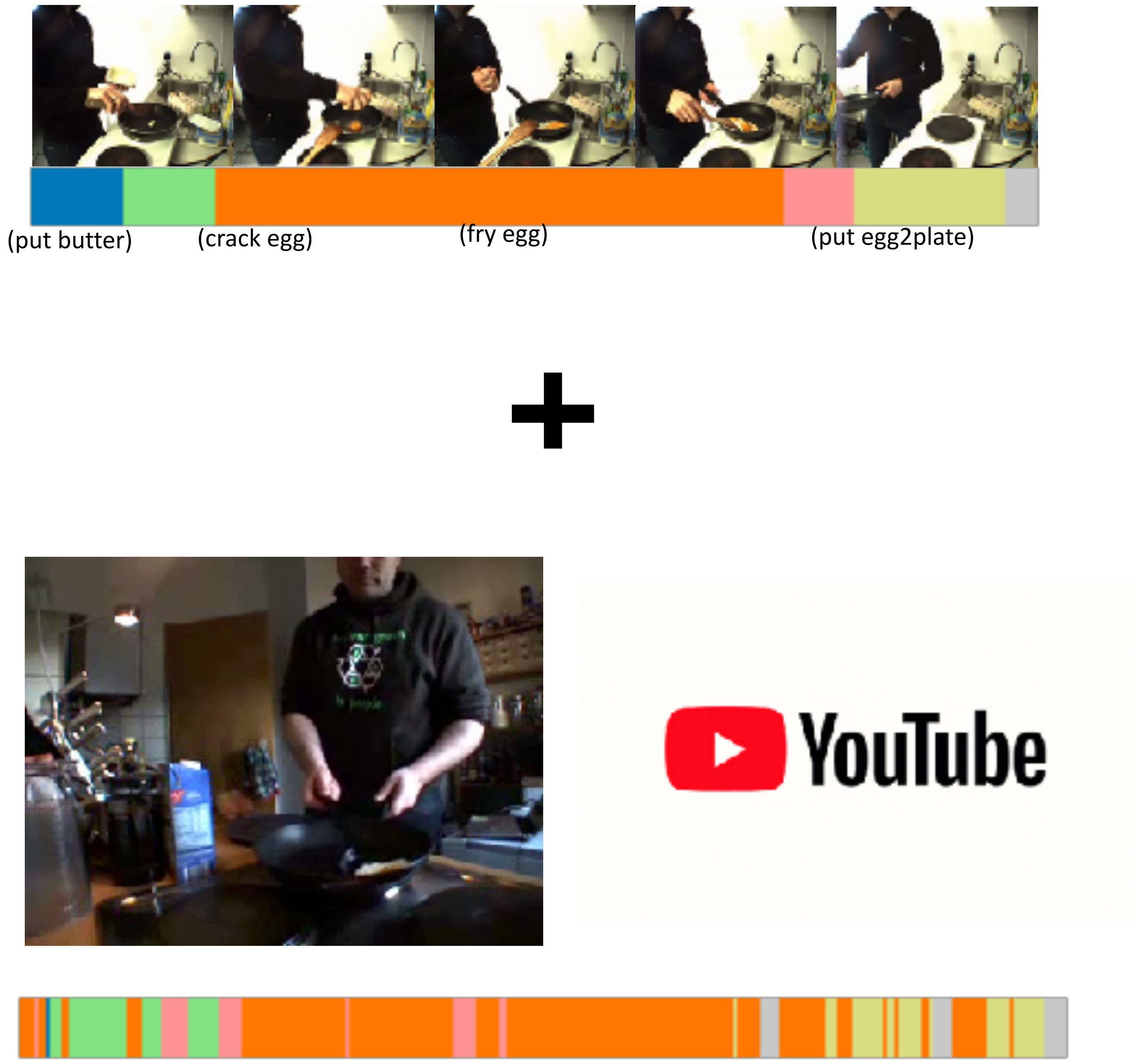
Figure 1. Diagram of our model-agnostic meta-learning algorithm (MAML), which optimizes for a representation θ that can quickly adapt to new tasks.

Isn't MAML just parameter initialization?

No! Surprisingly, MAML is *universal*:

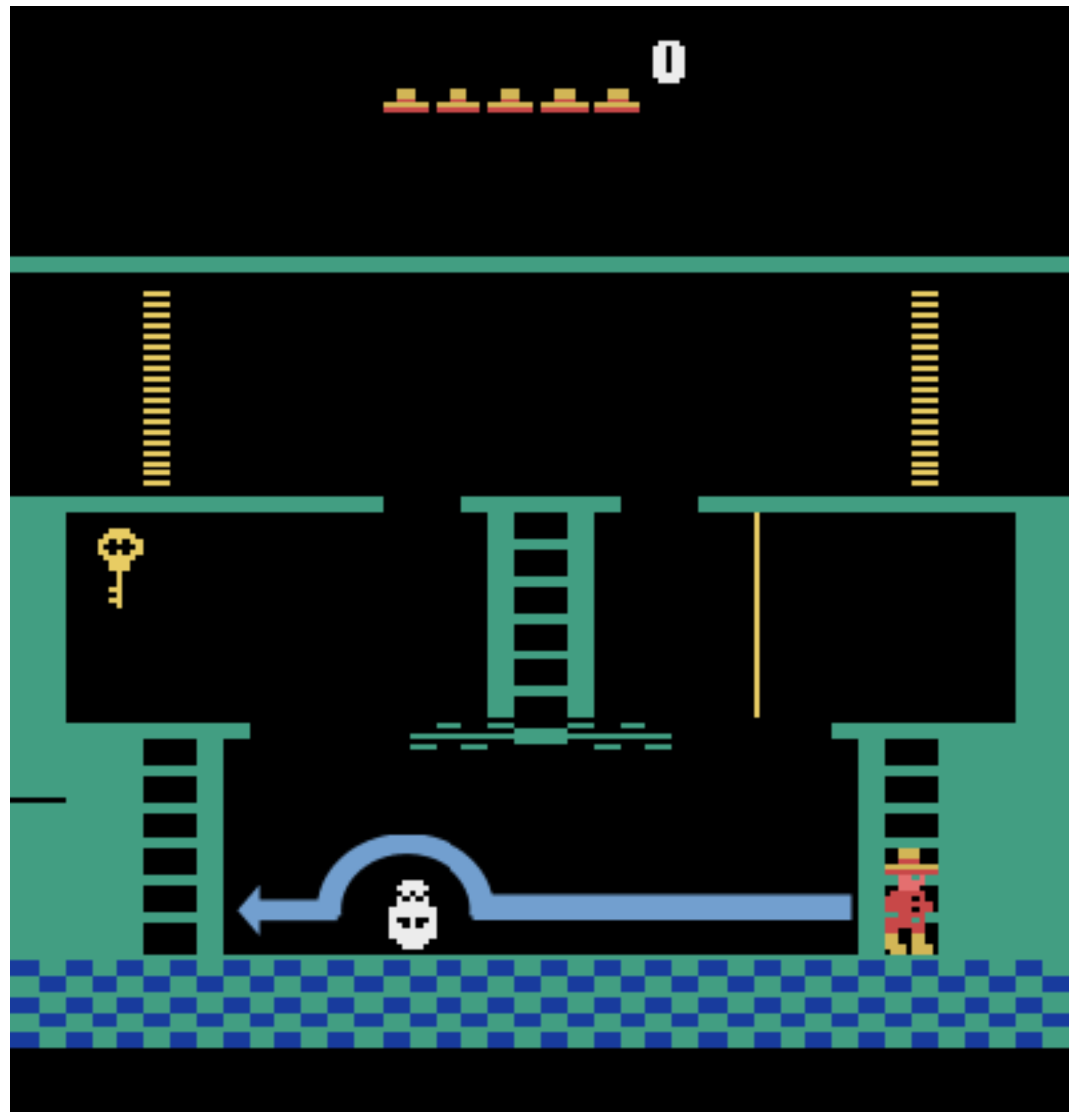
it can learn any update rule, in principle

Leveraging auxiliary data sources and multiple data modalities for increased efficiency

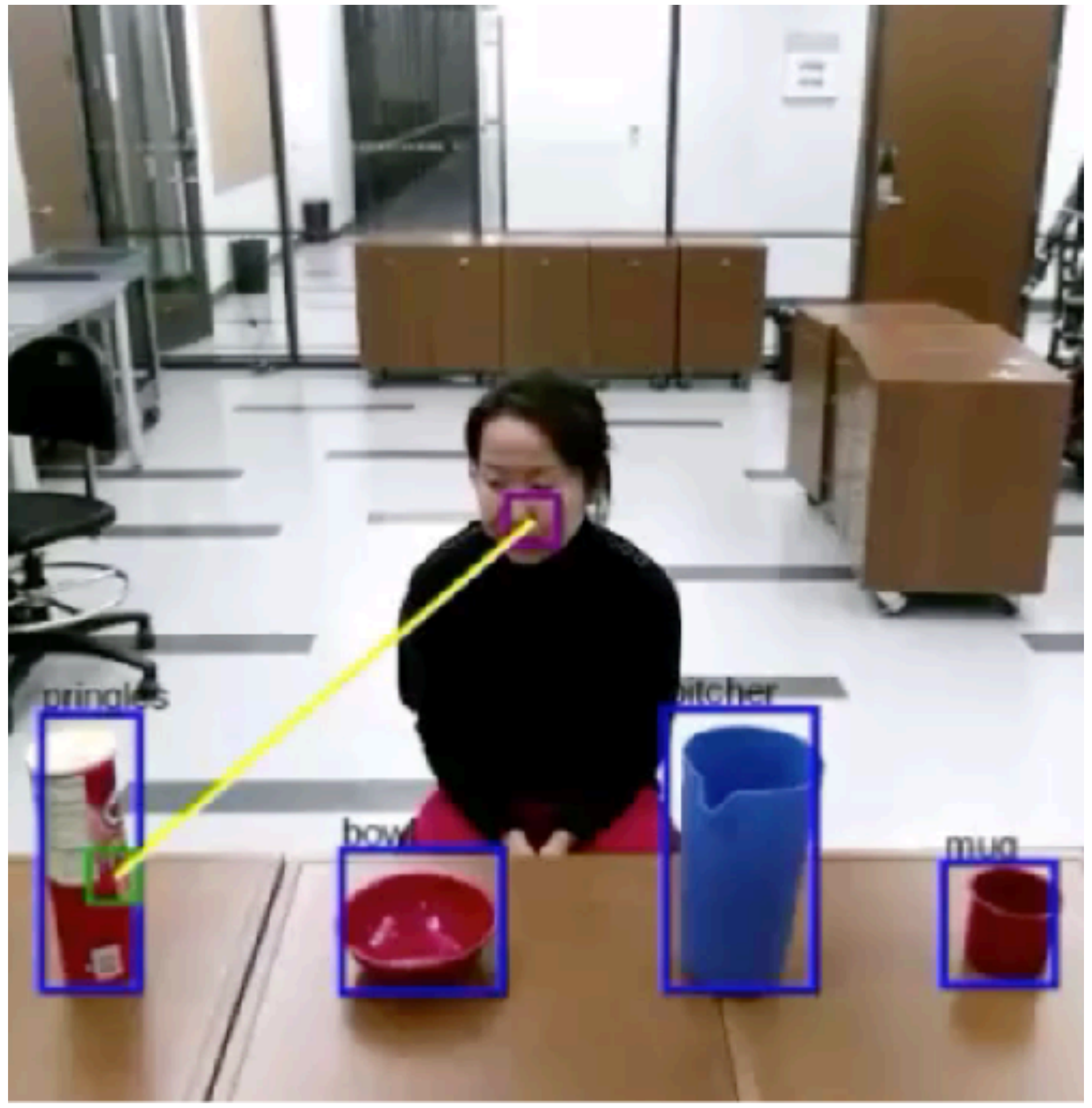


Auxiliary video alignment

"Jump over the skull while going to the left"



Natural language narration



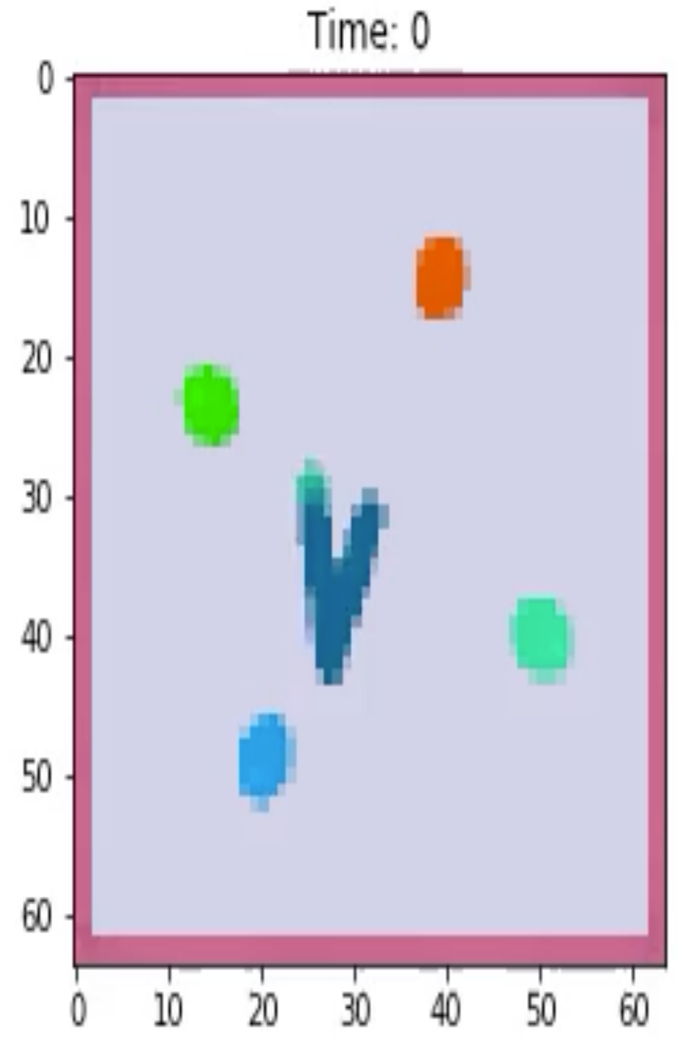
Gaze and facial expressions

W. Goo and S. Niekum.
One Shot Learning of Multi-Step Tasks from Observation via Activity Localization in Auxiliary Video
International Conference on Robotics and Automation, May 2019.

P. Goyal, S. Niekum, and R. Mooney.
Using Natural Language for Reward Shaping in RL
International Joint Conference on AI, August 2019.

A. Saran, E.S. Short, A.L. Thomaz, and S. Niekum.
Understanding Teacher Gaze Patterns for Robot Learning.
Conference on Robot Learning (CoRL), October 2019.

Colored Target Reaching Task



Subtask A:
Reaching to an orange target

→

Subtask B:
Reaching to a green target

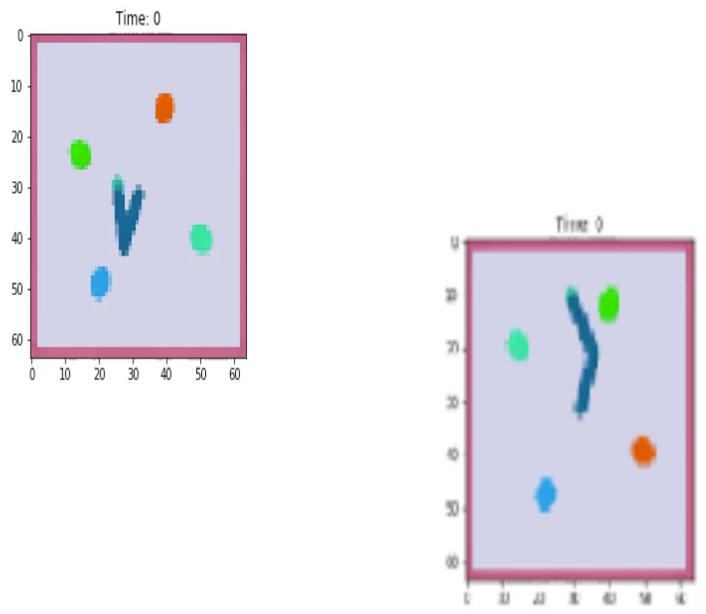
→

One-Shot Learning from Observation for Multi-Step Tasks **via Activity Localization in Auxiliary Video**

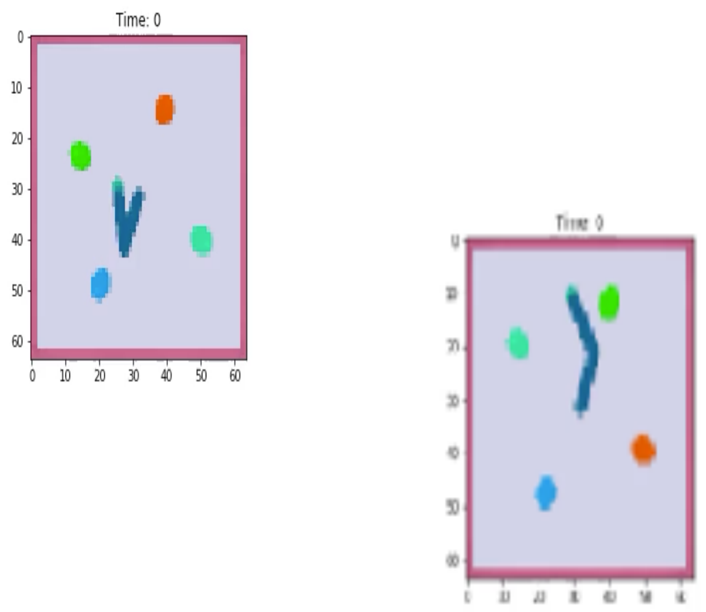
Meta-learn a low-shot activity classifier

…then perform IRL

Experiment – Meta-Training

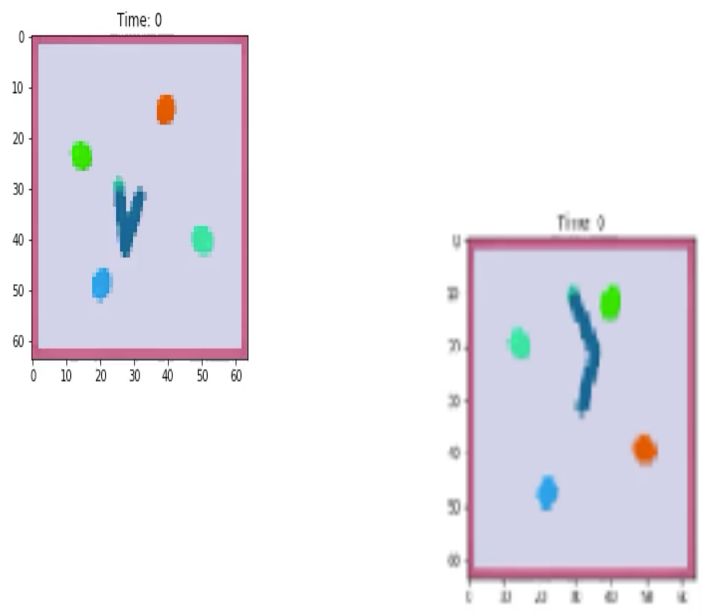


$\mathcal{T}_1$; target orange and green



$\mathcal{T}_2$; target blue and yellow

……

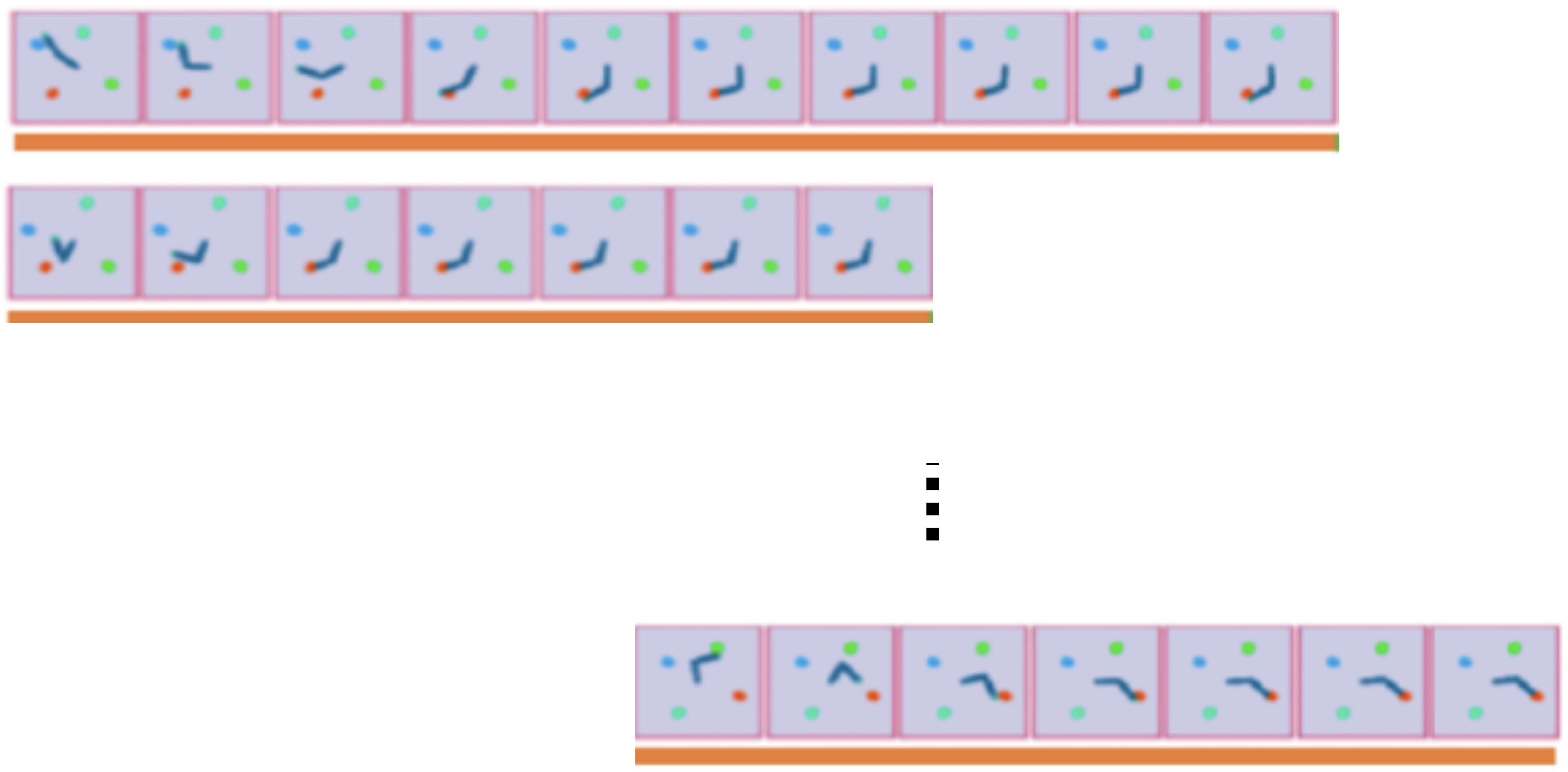


$\mathcal{T}_n$; target purple and red

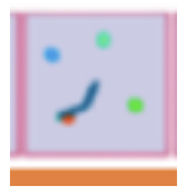
Meta-Training dataset;
videos with preselected 36 target colors, 100 videos per each task

Learning from Observation (LfO) – Approach

- Learning a notion of ***progress***
  - Shuffle-and-Learn loss



Are the frames in order?

$g($  ,  $)$ = 1; in order

$g($  ,  $)$ = 0; out of order

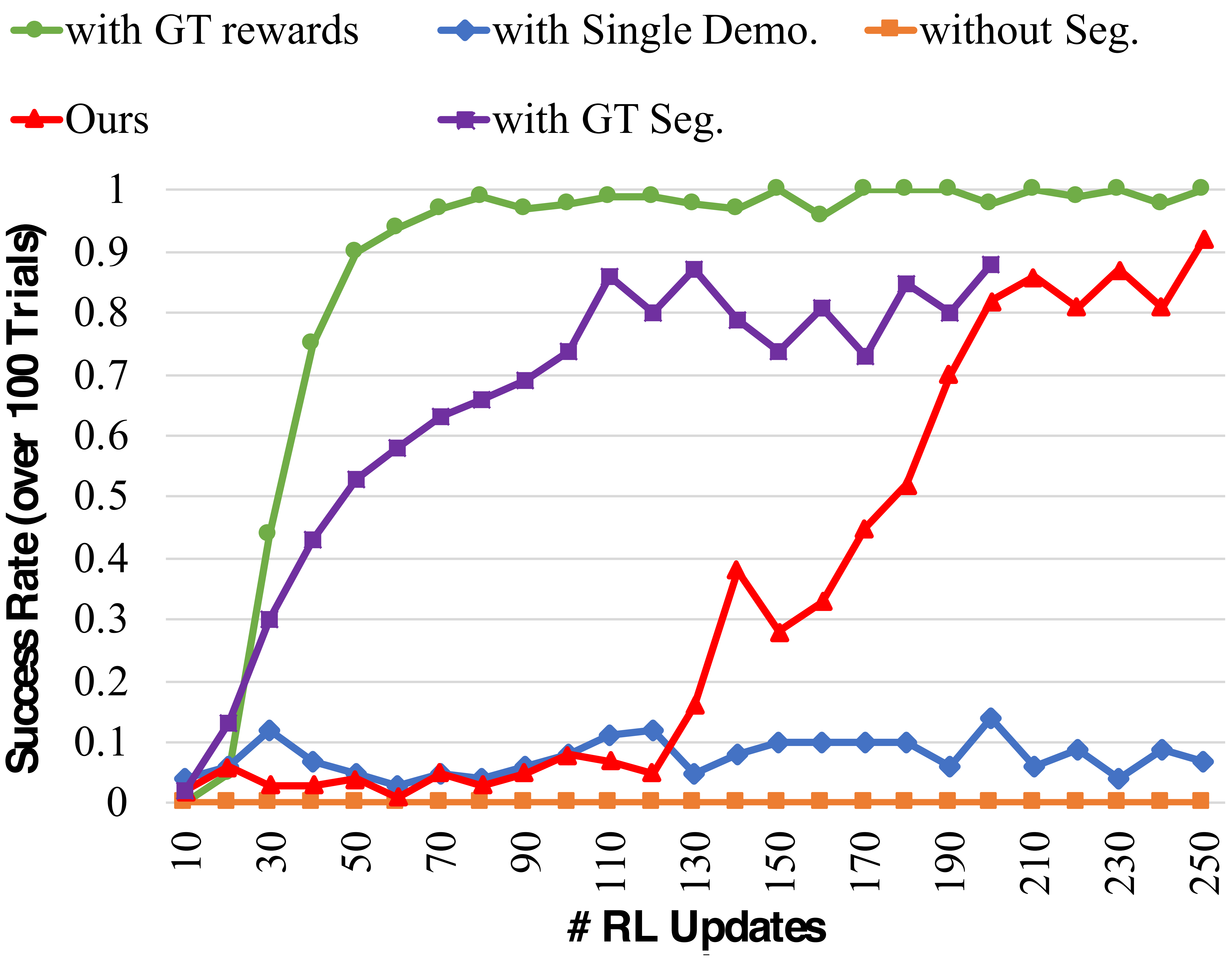
⋮

For all possible pairs,

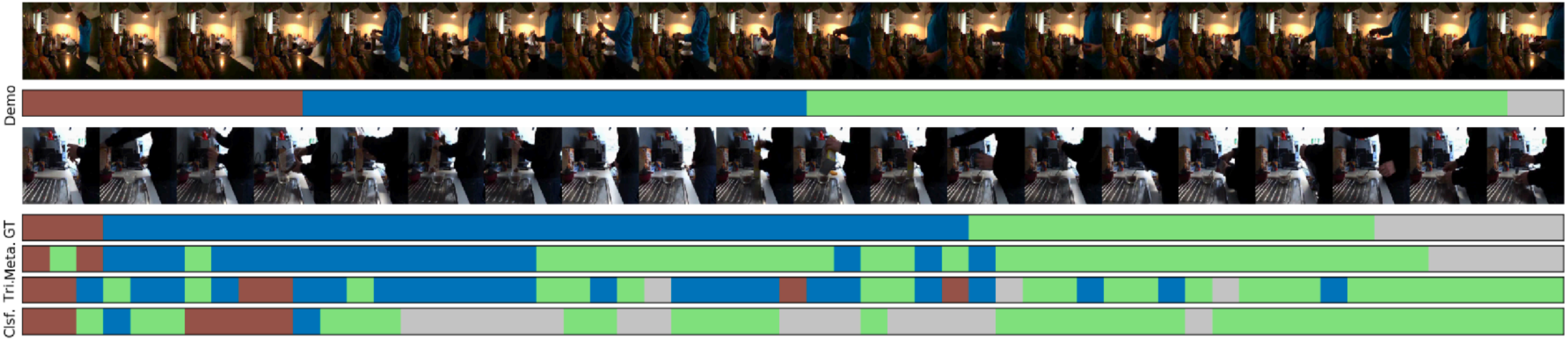
$$Loss = L_{ce}(sigmoid(g(o_t, o_{t'})), \mathbb{1}(t < t')),$$

Learning from Observation (LfO) – Result

Result – the whole pipeline



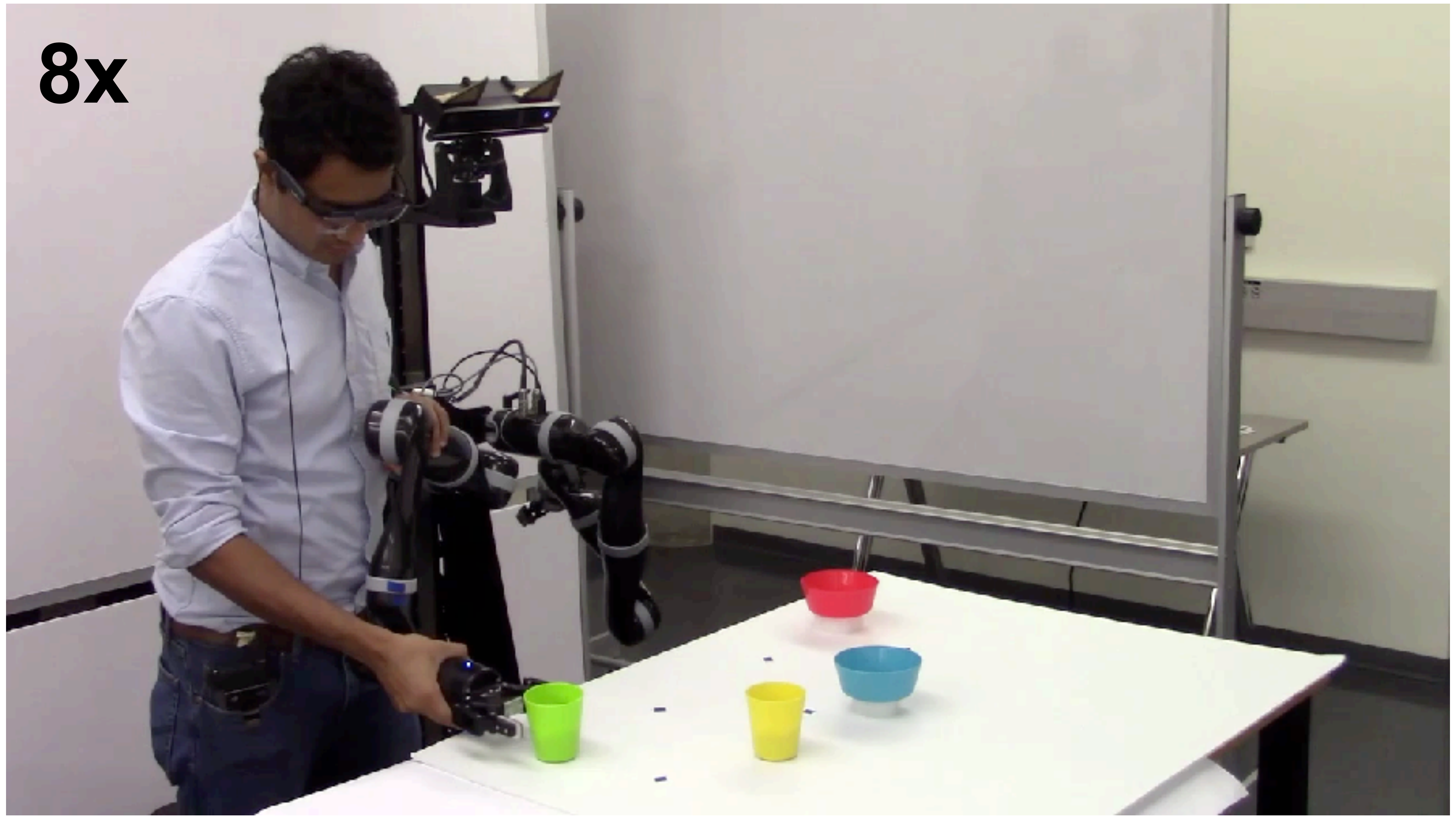
Results – Breakfast dataset



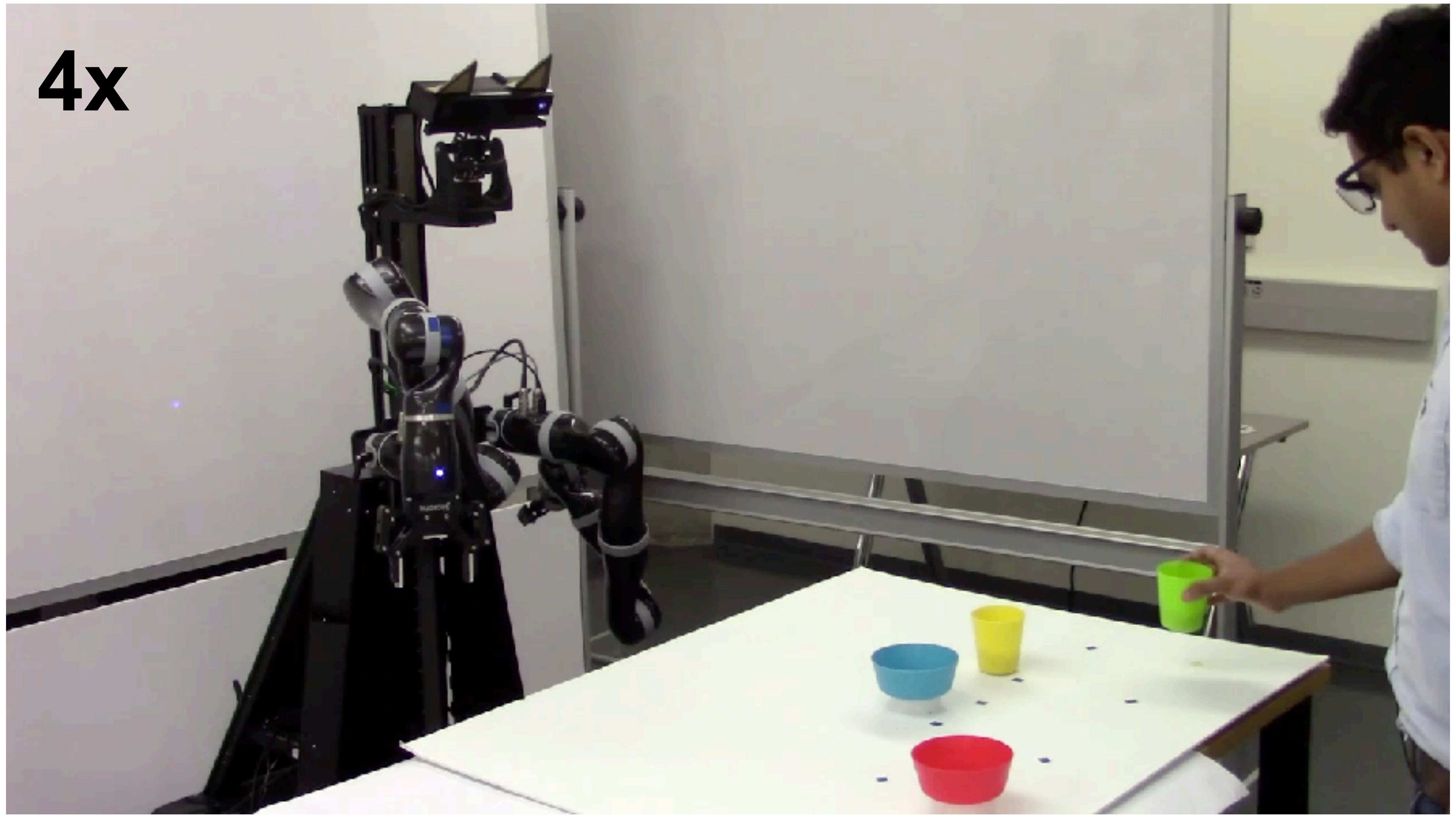
Gaze – a signal of Human Intent

Gaze Patterns in Human Demonstrations for Robots

Keyframe-based Kinesthetic Teaching (KT)

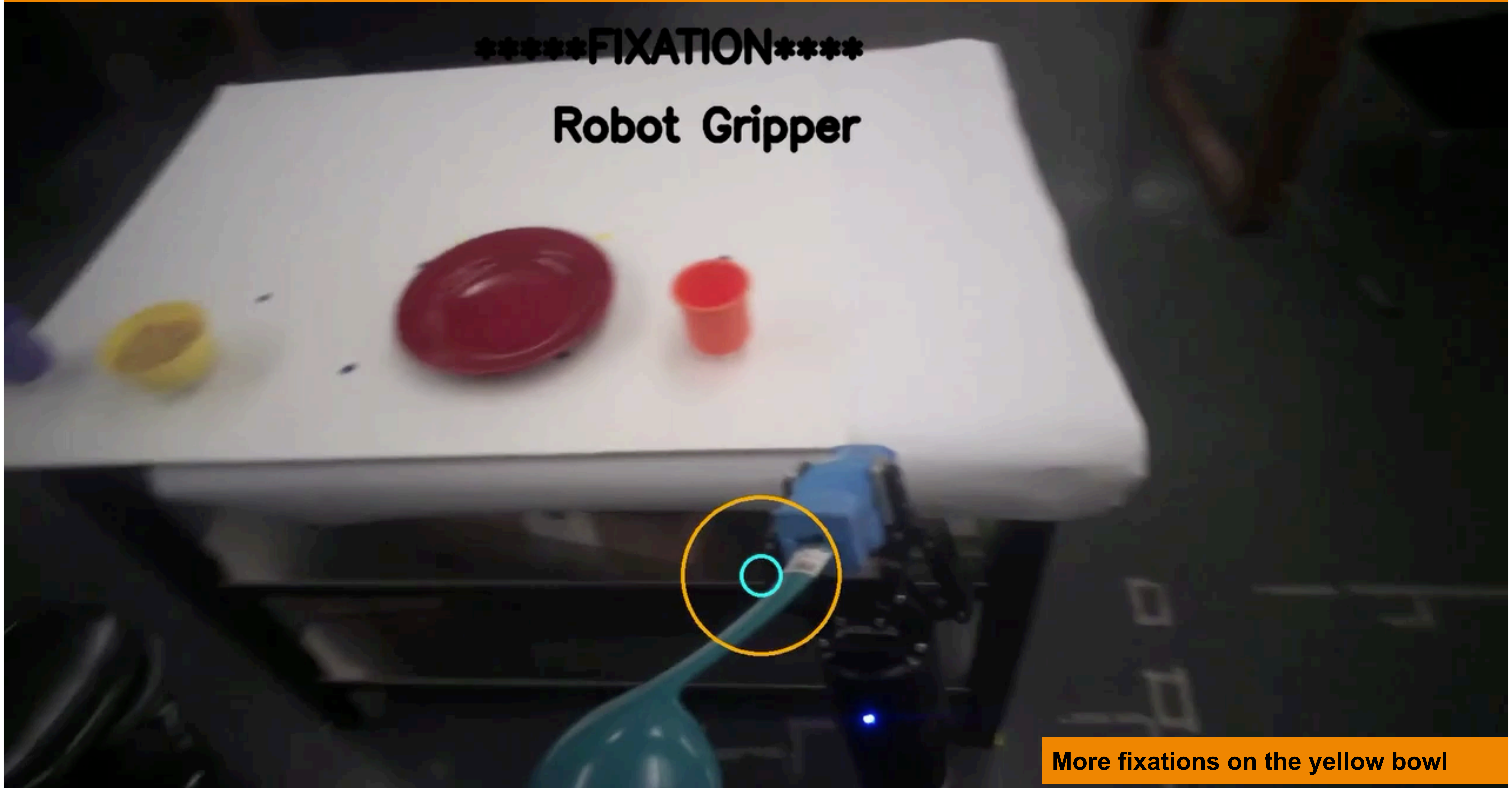


Observational/Video Demonstrations



Gaze Fixations during Ambiguous Placement Demonstrations

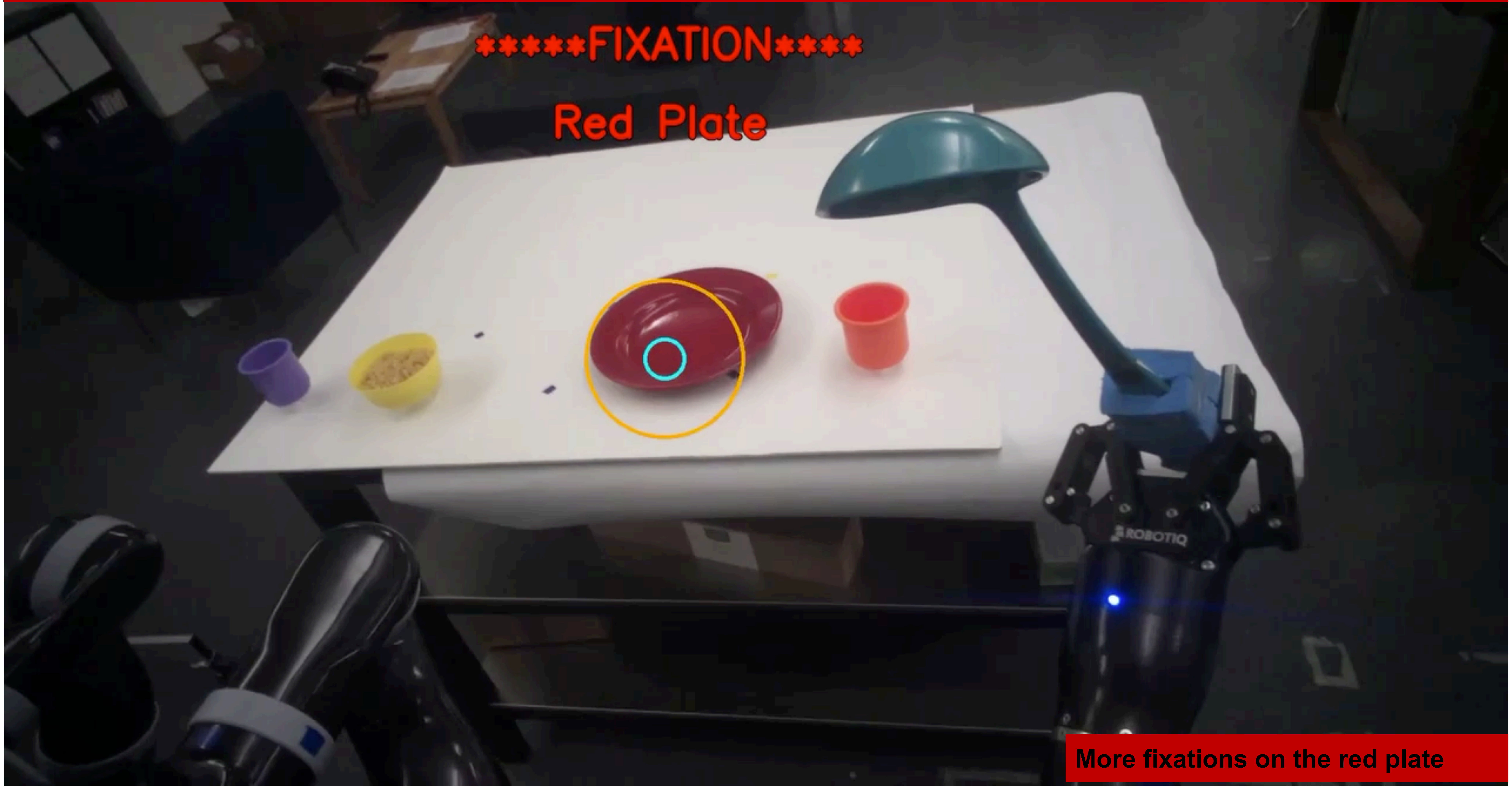
Instruction: Place Green Ladle to the **right of Yellow Bowl**



More fixations on the yellow bowl

Gaze Fixations during Ambiguous Placement Demonstrations

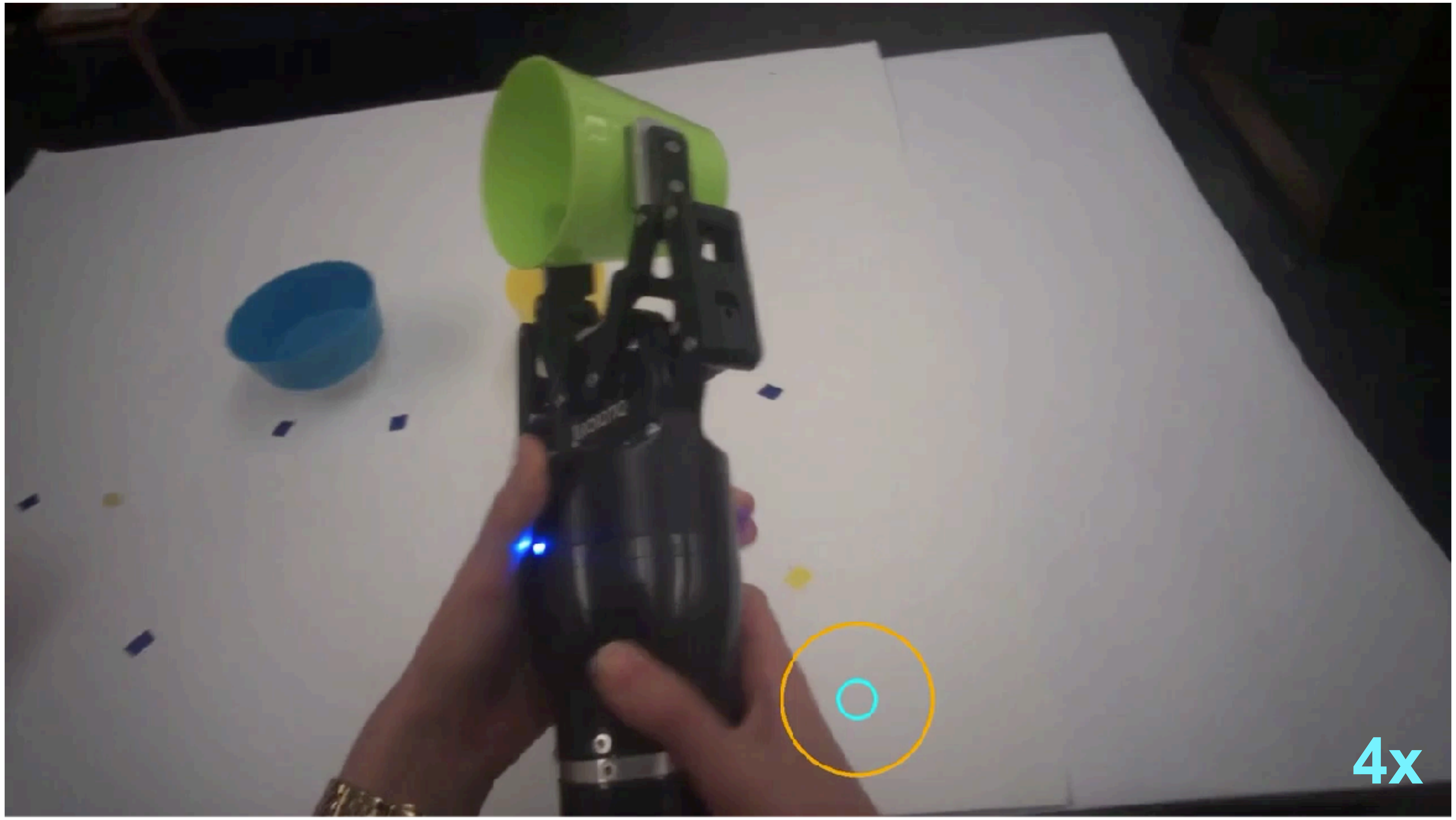
Instruction: Place Green Ladle to the **left of Red Plate**

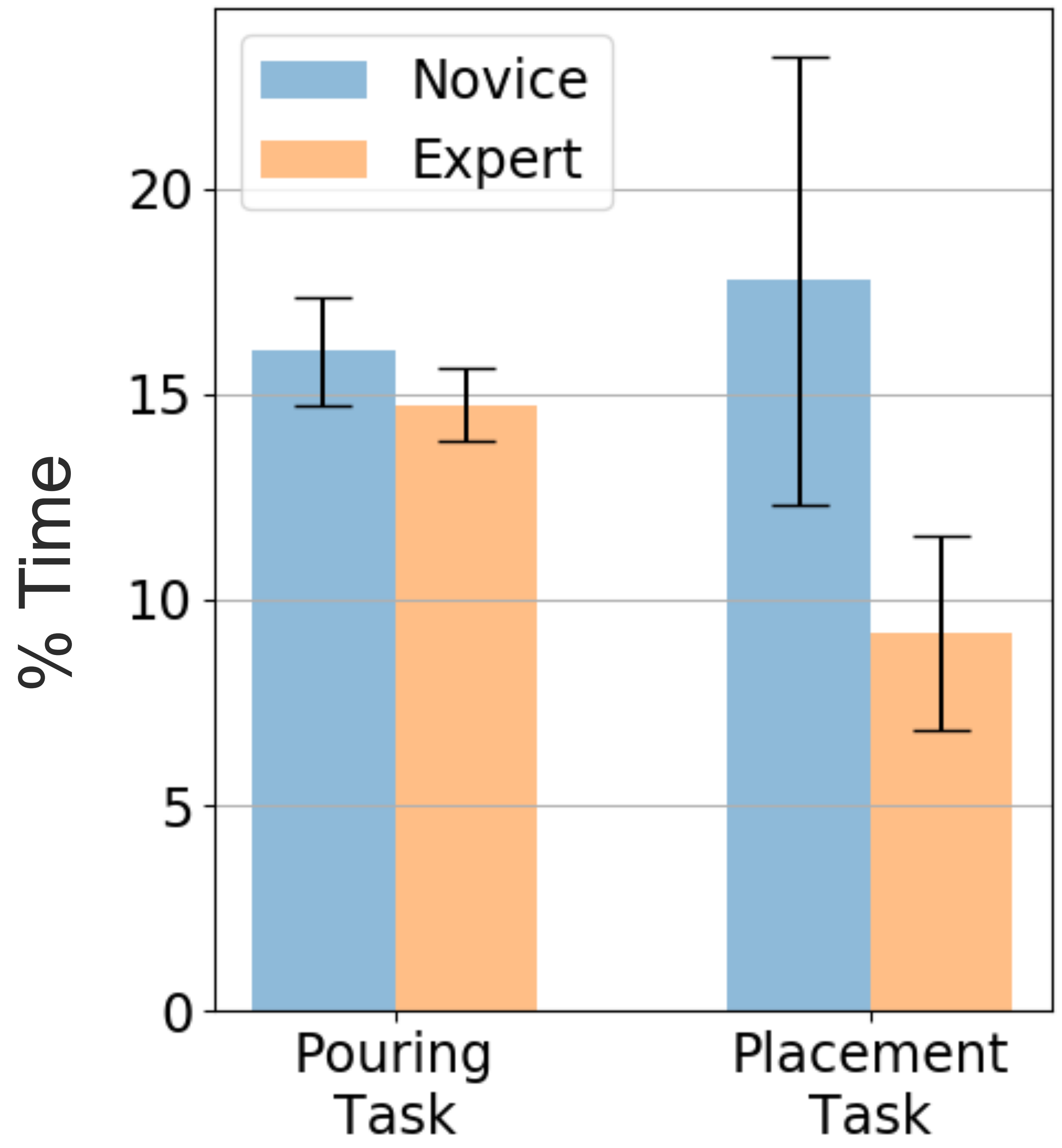


More fixations on the red plate

Kinesthetic Demos:

Novice Users focus more on the Robot's Gripper





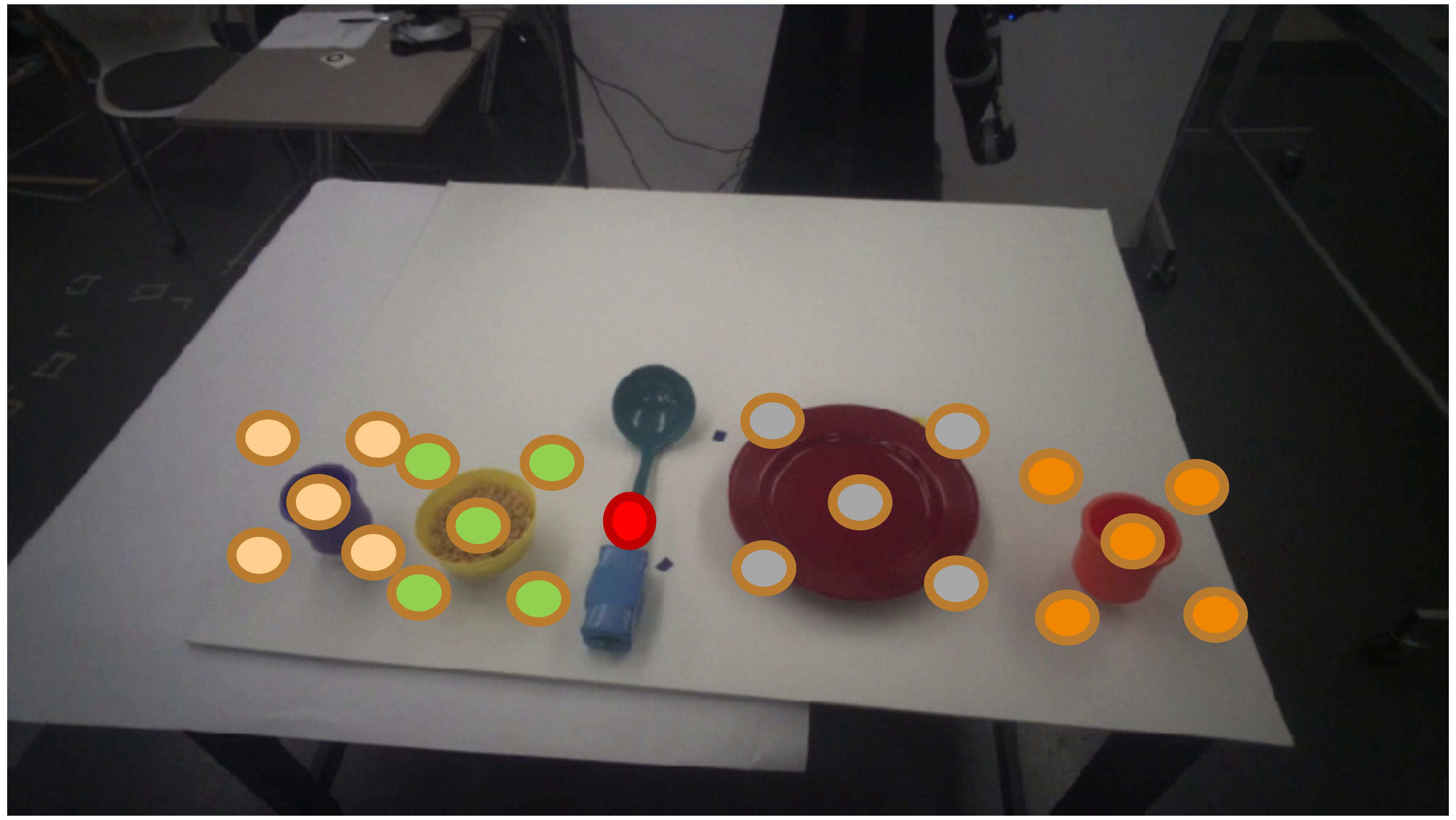
Reward Learning for the Placement Task

Gaze augmented Bayesian IRL for Placement Task

$$P(R|D,G) \propto P(D|R)\underbrace{P(R|G)}$$

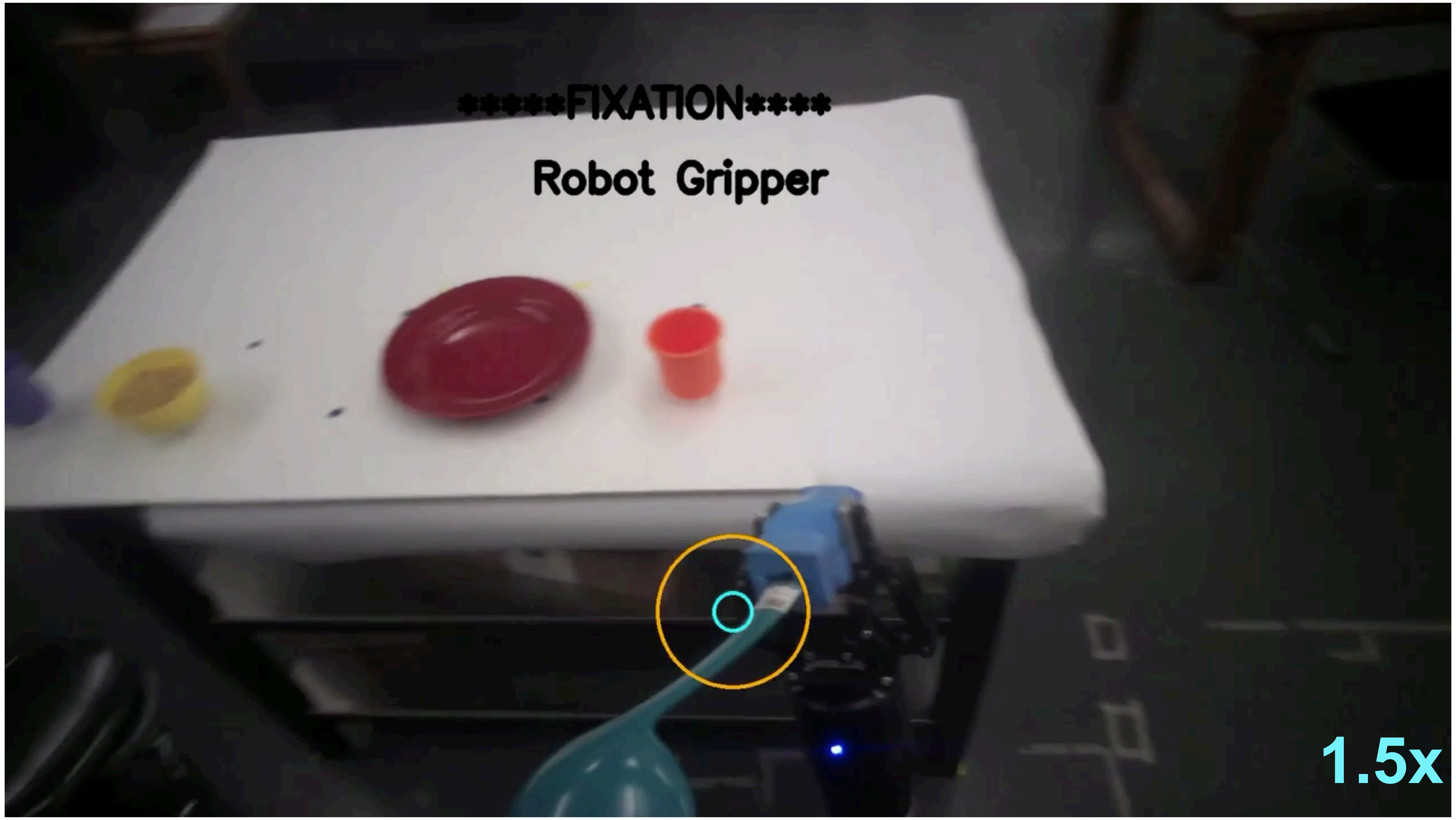
Penalize reward functions for which pairwise gaze fixation times are not ranked according to corresponding object weights

Reward functions modeled as weighted RBF kernels near objects



Bayesian IRL using Gaze from Ambiguous Demonstrations

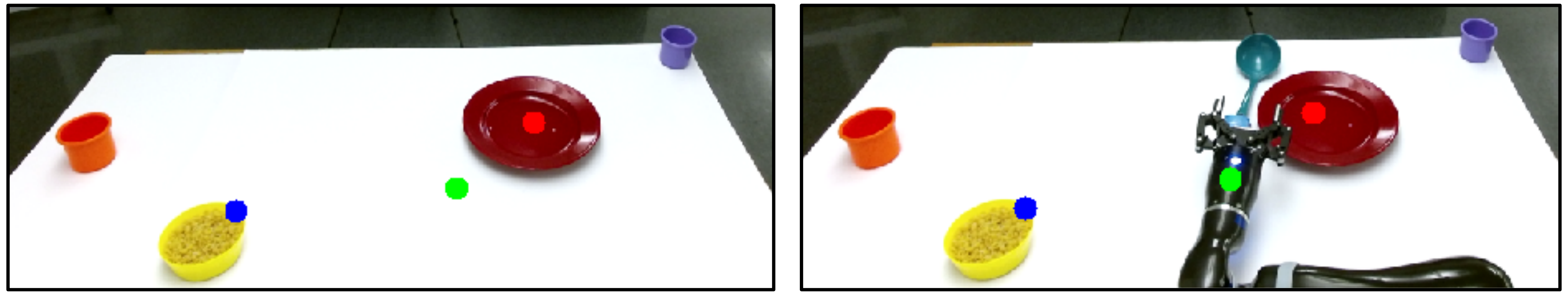
"Place green ladle to the **right of the yellow bowl**"



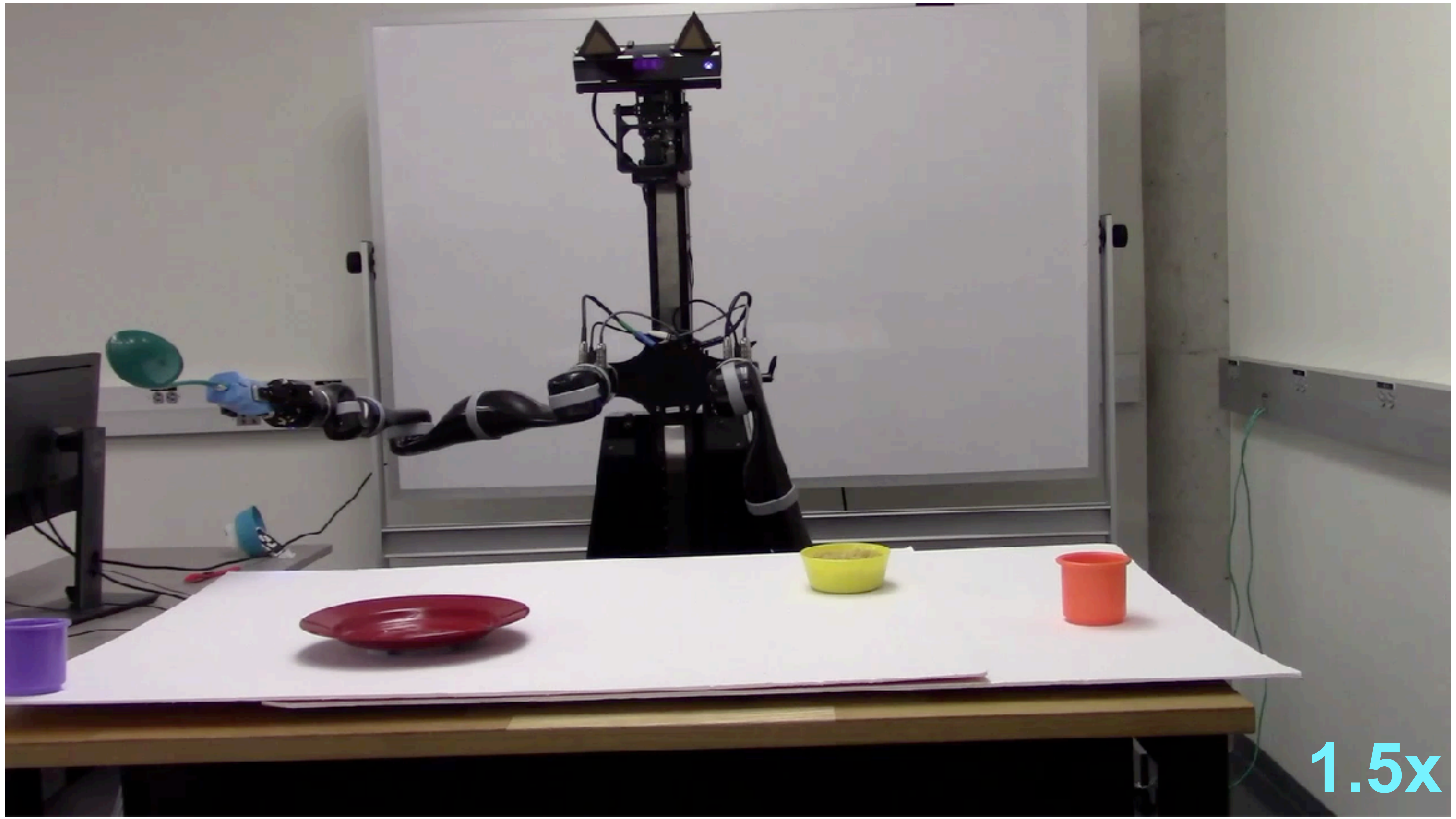
DEMONSTRATION

BIRL **without** Gaze Information

“Place green ladle to the **right of the yellow bowl**”

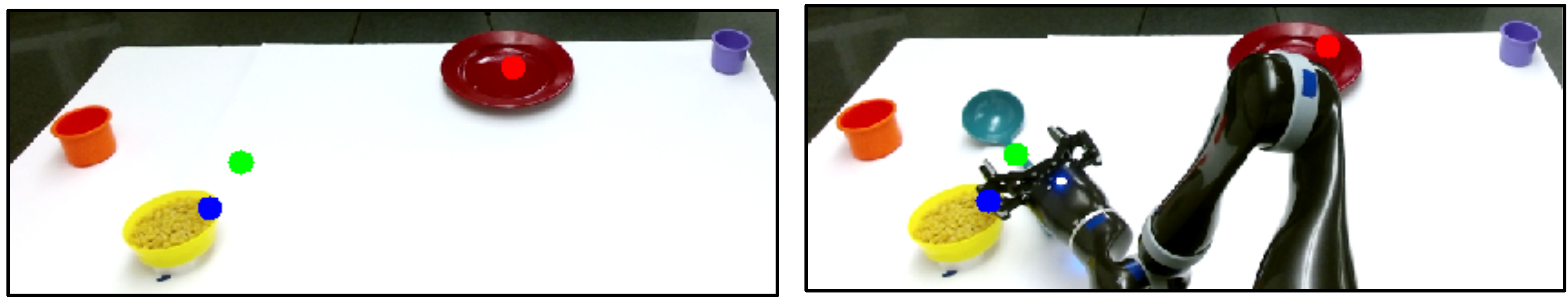


Proposed ladle location from learnt policy

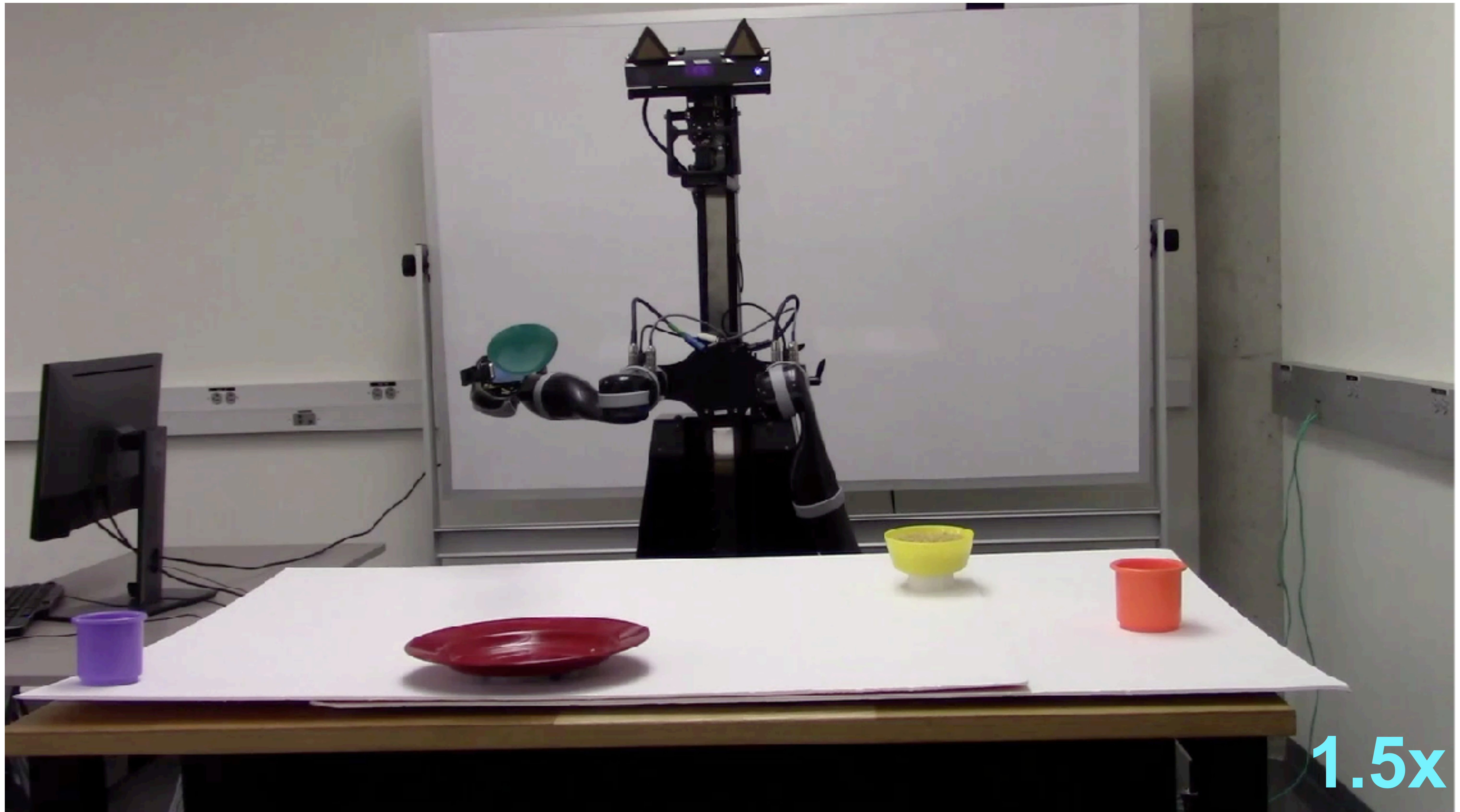


BIRL **with** Gaze Information

"Place green ladle to the **right of the yellow bowl**"

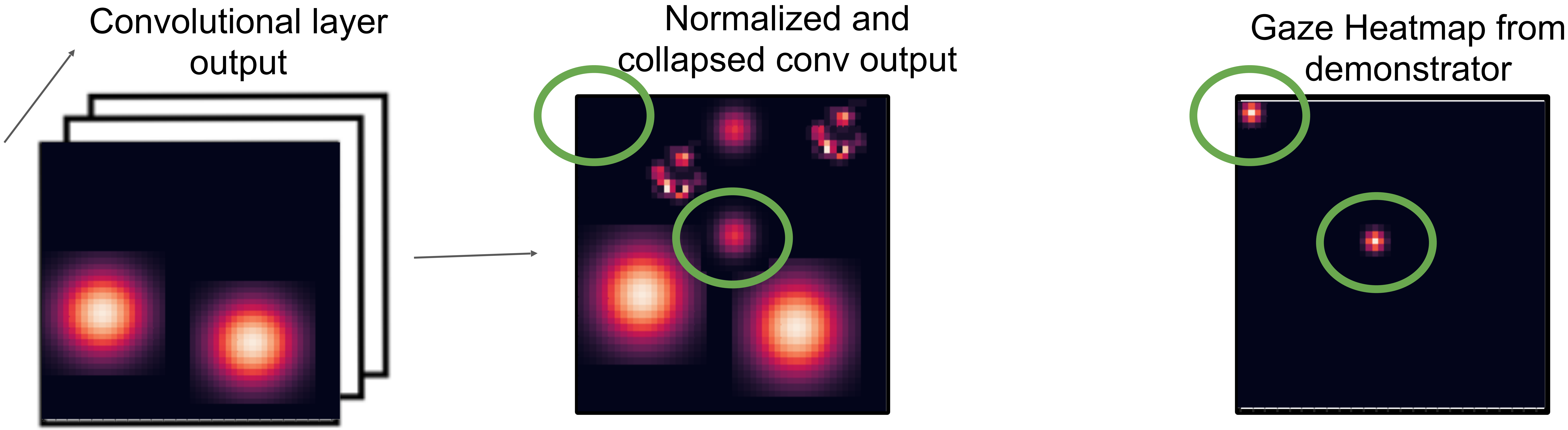


Proposed ladle location from learnt policy



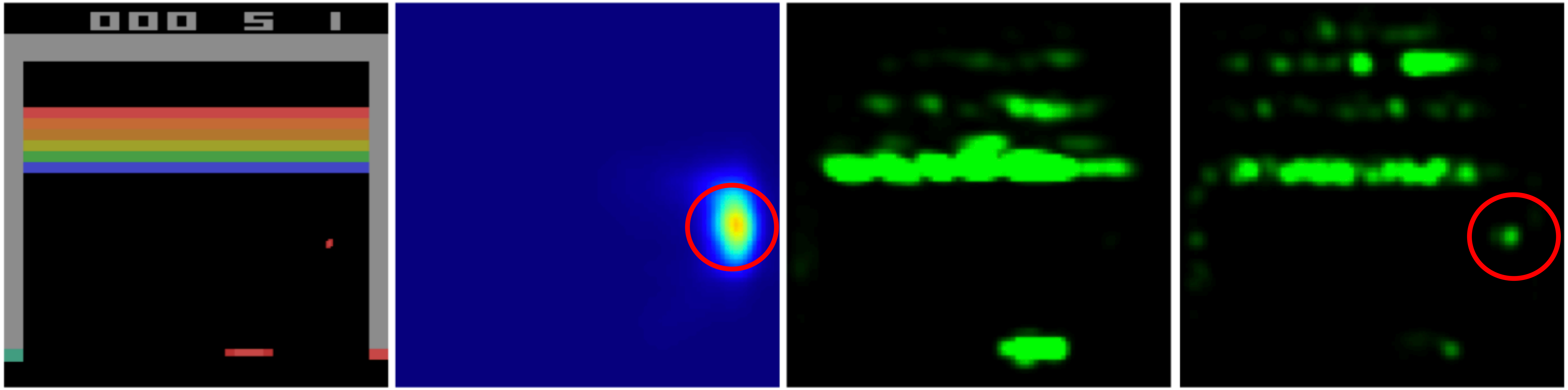
Coverage-based Gaze Loss (CGL)

- Only required during training as part of an auxiliary loss function
- Can be applied to any existing Imitation Learning network with convolutional layers
- Improved performance without varying model complexity



Intuition: Add a penalty for regions where gaze fixations are non-zero, but are not attended to by convolutional layers

CGL: Coverage-based Gaze Loss



(a) Input image (b) Human (c) T-REX (d) T-REX+CGL

A. Saran, R. Zhang, E.S. Short, and S. Niekum.
Efficiently Guiding Imitation Learning Algorithms with Human Gaze.
International Conference on Autonomous Agents and Multiagent Systems (AAMAS), May 2021.

BCO and T-REX + Gaze

Table 1: BCO performance with and without the usage of human demonstrators' gaze

Game	Human	BCO	BCO+GMD	BCO+CGL
Breakout	344 - 554	0.2	0.0	**0.6**
Hero	34305 - 50485	0.0	0.0	**1469.0**
MsPacman	17441 - 92610	90.0	70.0	**210.0**
Asterix	88000-537500	**650.0**	363.3	336.7
Phoenix	22410-27570	24.0	389.3	**656.3**
Space Invaders	845-2035	0.0	88.3	**311.2**
Enduro	278-742	0.0	0.0	**3.2**

Table 2: T-REX performance with and without the usage of expert human demonstrators' gaze

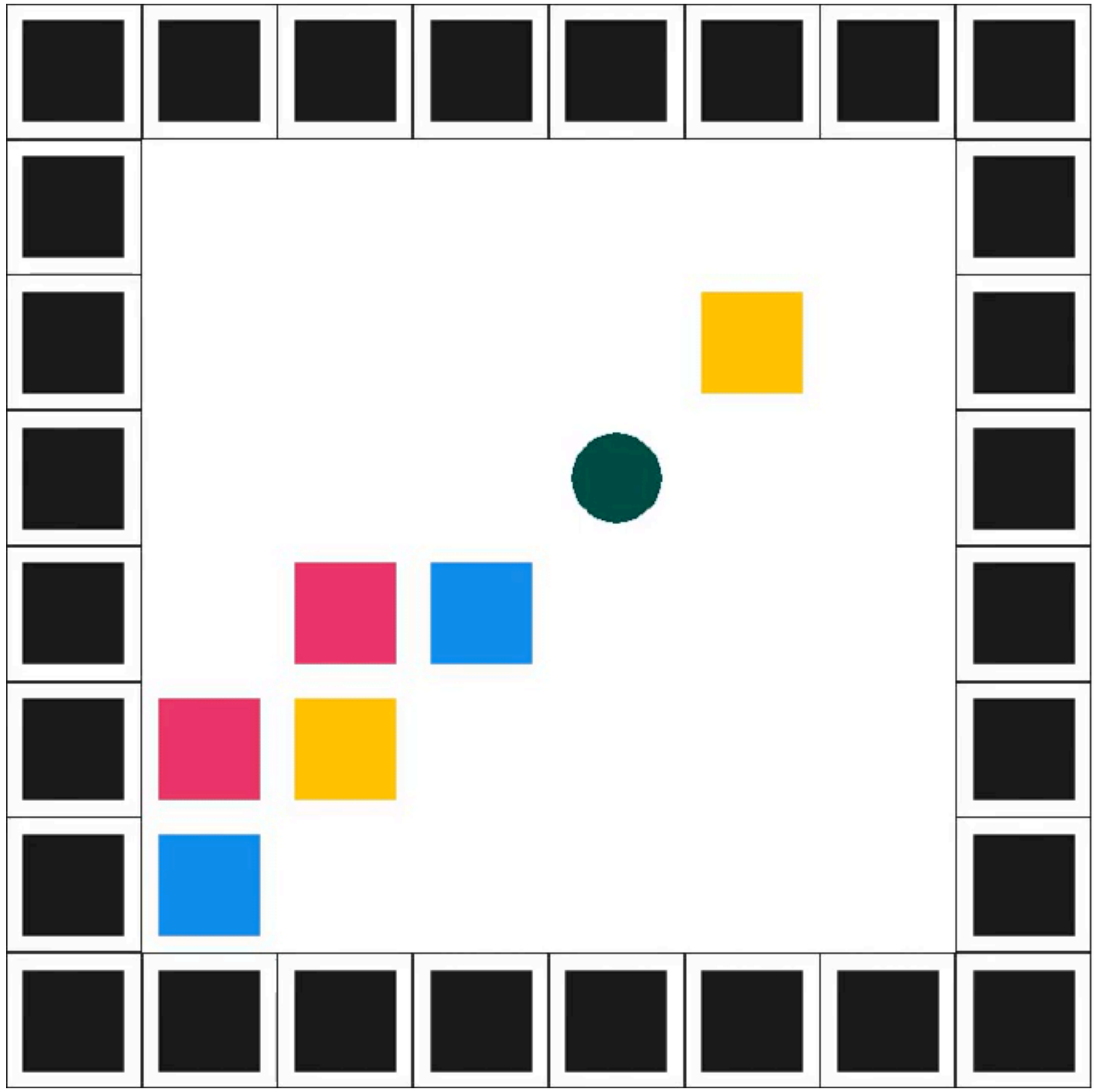
Game	Human	T-REX	T-REX+CGL
Asterix	88000-537500	23926.7	**99468.3**
Centipede	39737-251961	**12862.8**	8514.3
Phoenix	22410-27570	542.00	**669.7**
MsPacman	27731-36061	596.3	**625.7**

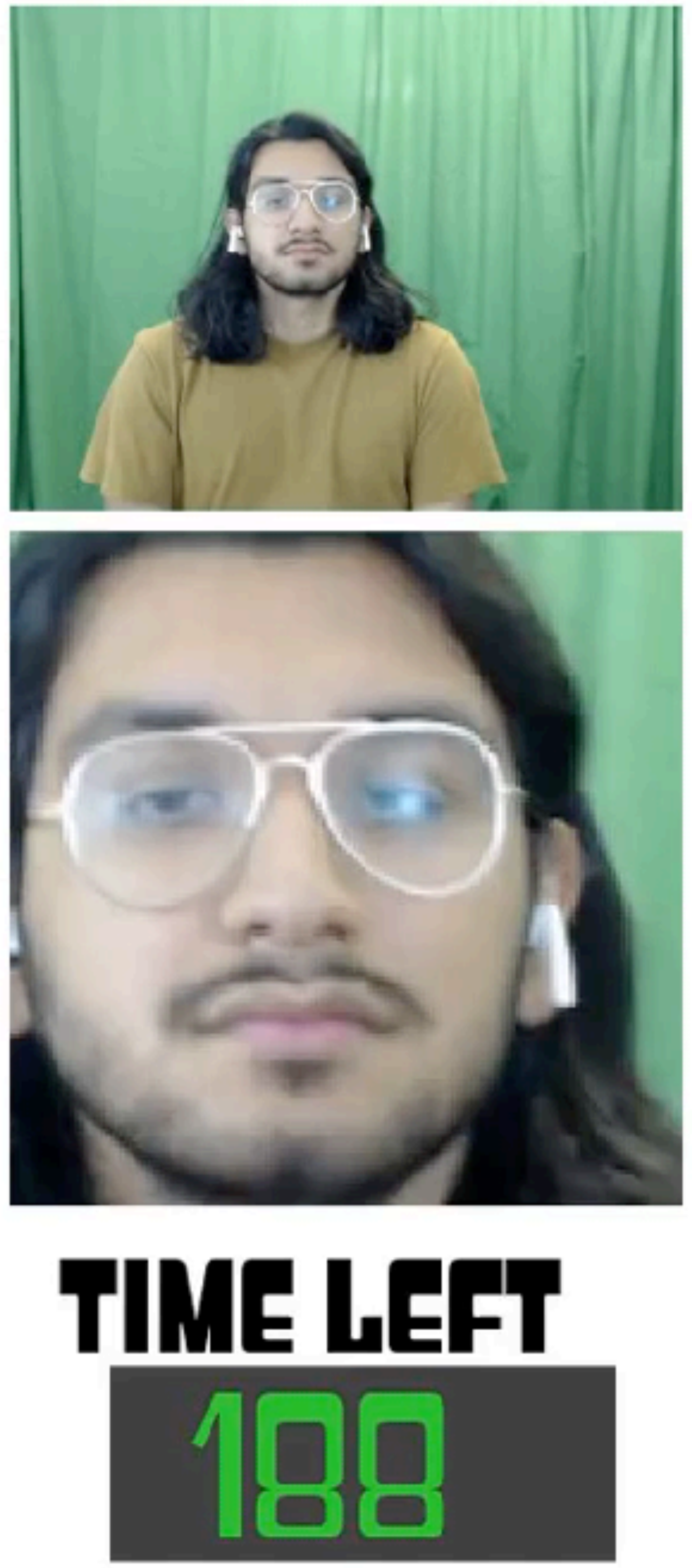
Multimodal data sources: Facial Reactions

Implicit human feedback:

- Occurs naturally
- Is not necessarily intended to influence behavior
- Can be used with no additional burden on user

EMPATHIC: Learning from implicit feedback — training





Y. Cui, Q. Zhang, A. Allievi, P. Stone, S. Niekum, and W. Knox.
The EMPATHIC Framework for Task Learning from Implicit Human Feedback.
Conference on Robot Learning (CoRL), November 2020.

EMPATHIC: Learning from implicit feedback — deployment

Y. Cui, Q. Zhang, A. Allievi, P. Stone, S. Niekum, and W. Knox.
[The EMPATHIC Framework for Task Learning from Implicit Human Feedback](#).
Conference on Robot Learning (CoRL), November 2020.