

Principles of Robot Autonomy II

Learning from Diverse Sources of Data (1)



Today's itinerary

- Recap of imitation learning and inverse RL
- Learning from other sources of data – Pairwise Comparisons
- Learning from other sources of data – Foundation Models
- Learning from physical feedback
- Learning from gestures
- Learning from sketches
- Data Quality

Today's itinerary

- Recap of imitation learning and inverse RL
- Learning from other sources of data – Pairwise Comparisons
- Learning from other sources of data – Foundation Models
- Learning from physical feedback
- Learning from gestures
- Learning from sketches
- Data Quality

Types of Imitation Learning

Behavioral Cloning

$$\arg \min_{\theta} \mathbb{E}_{(s, a^*) \sim P^*} L(a^*, \pi_{\theta}(s))$$

Works well when P^* is close to P_{θ}

Direct Policy Learning (via Interactive Demonstrator)

Requires Interactive Demonstrator (BC is a 1-step special case)

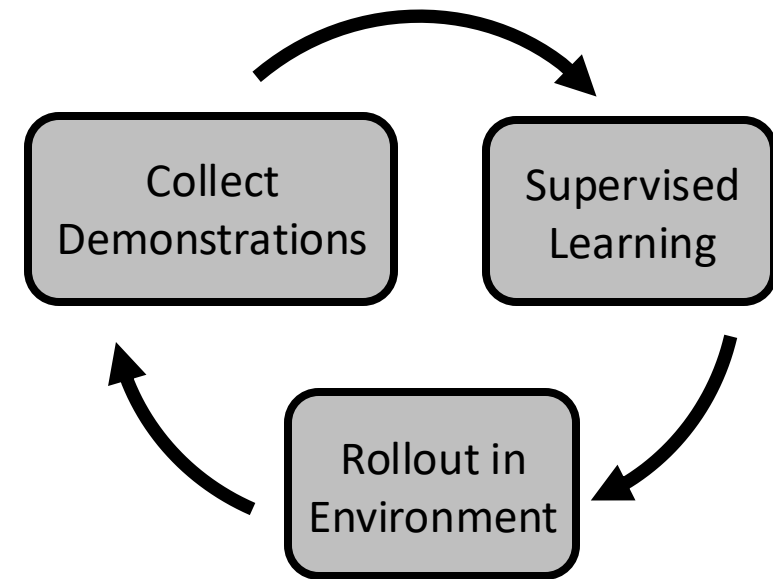
Inverse RL

Learn r such that:

$$\pi^* = \arg \max_{\theta} \mathbb{E}_{s \sim P(s|\theta)} r(s, \pi_{\theta}(s))$$

RL problem

Assume learning r is statistically easier than directly learning π^*



How to deal with reward ambiguity?

Reward ambiguity: There are many reward functions under which the expert demonstrations are optimal!!

Which reward function should we pick?

- **Maximum Margin Planning:** Looks for the one that separates the optimal policy best.

- **Maximum Entropy IRL:** Looks for the one where expert demonstrations are drawn from a high entropy distribution.

Today's itinerary

- Recap of imitation learning and inverse RL
- Learning from other sources of data – Pairwise Comparisons
- Learning from other sources of data – Foundation Models
- Learning from physical feedback
- Learning from gestures
- Learning from sketches
- Data Quality

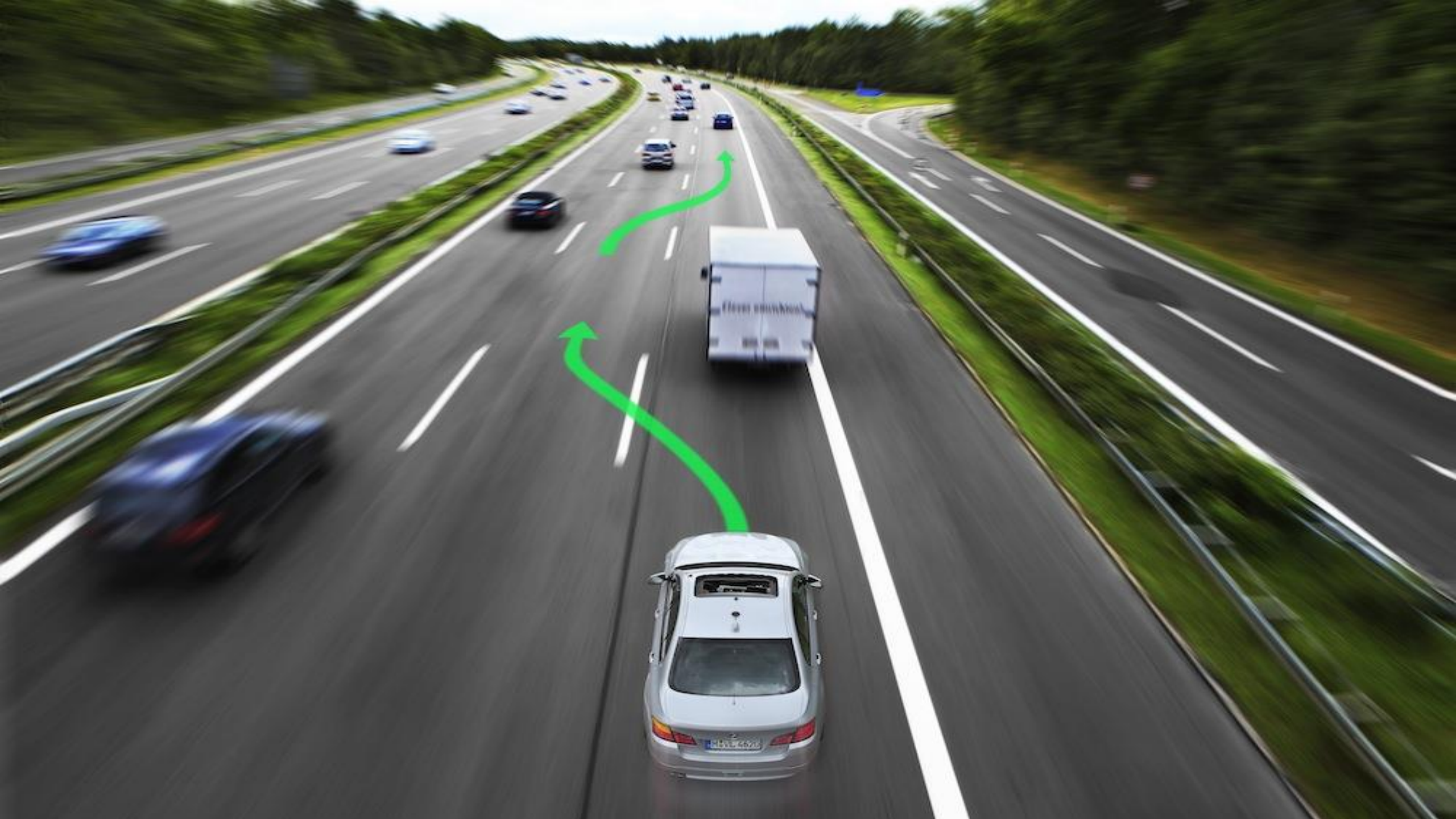
Expert demonstrations are difficult to collect, variable, and suboptimal!

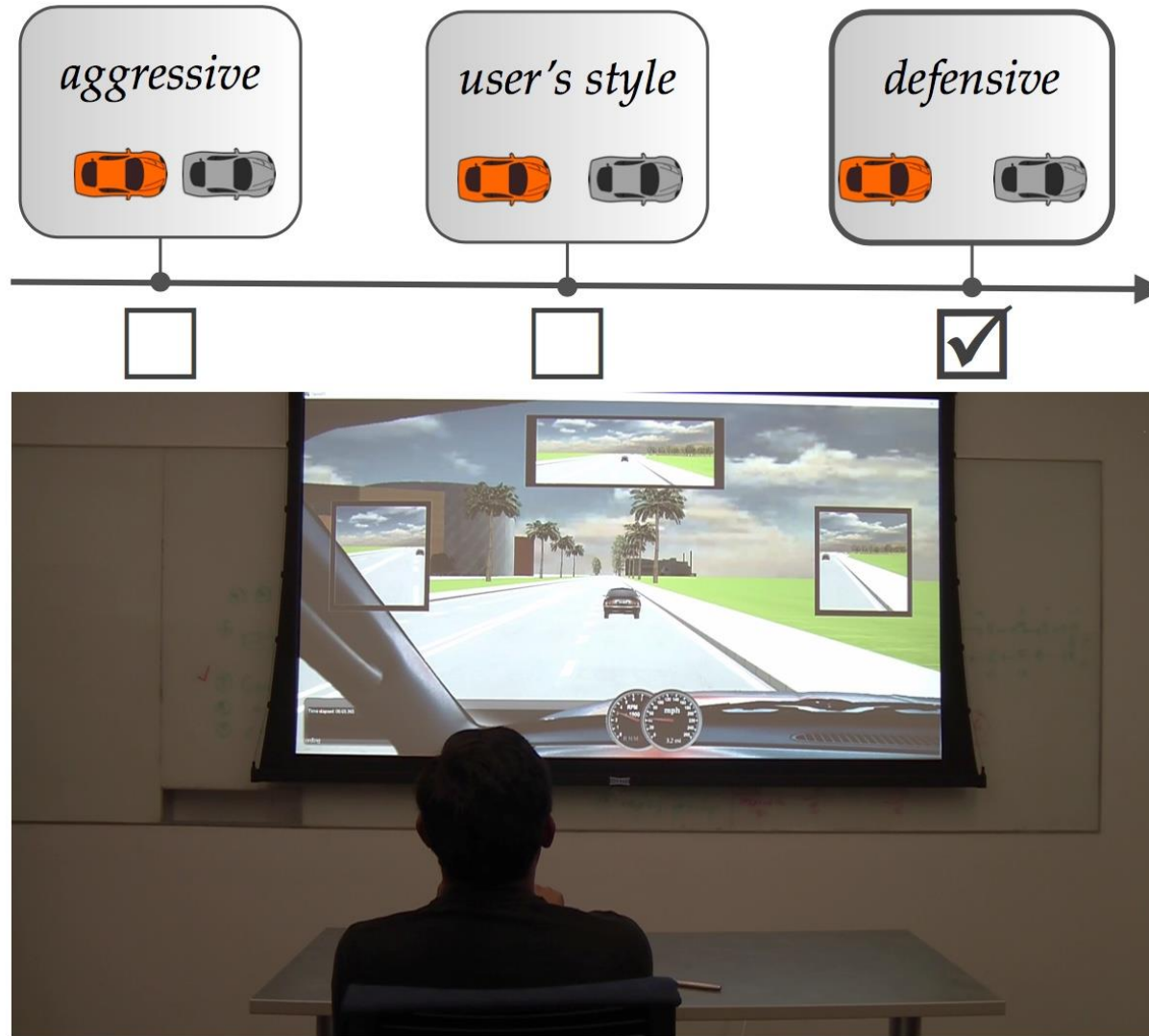


difficult to collect



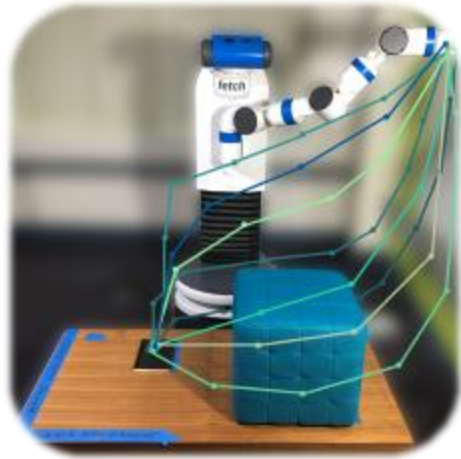
suboptimal and variable





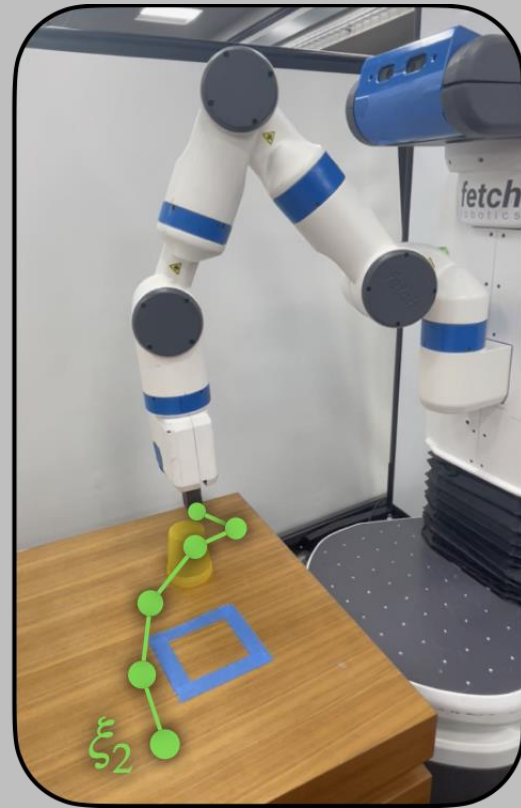
Demonstrations can be uninformative.

Learn from Different Sources of Data

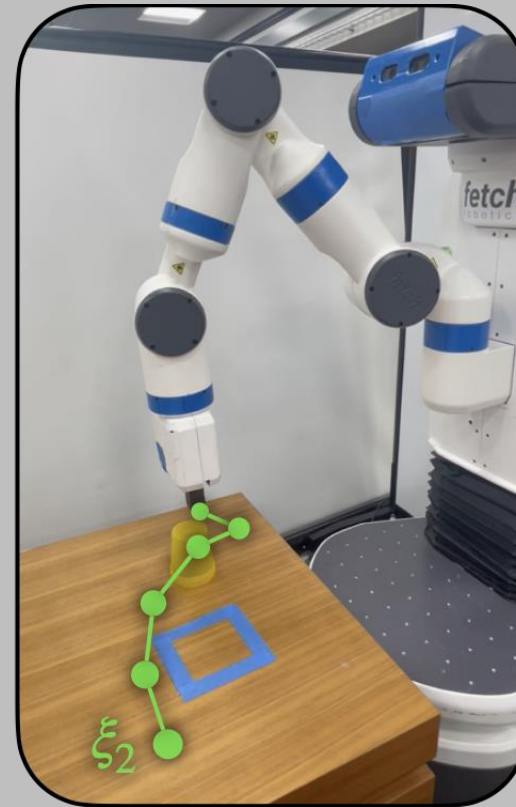
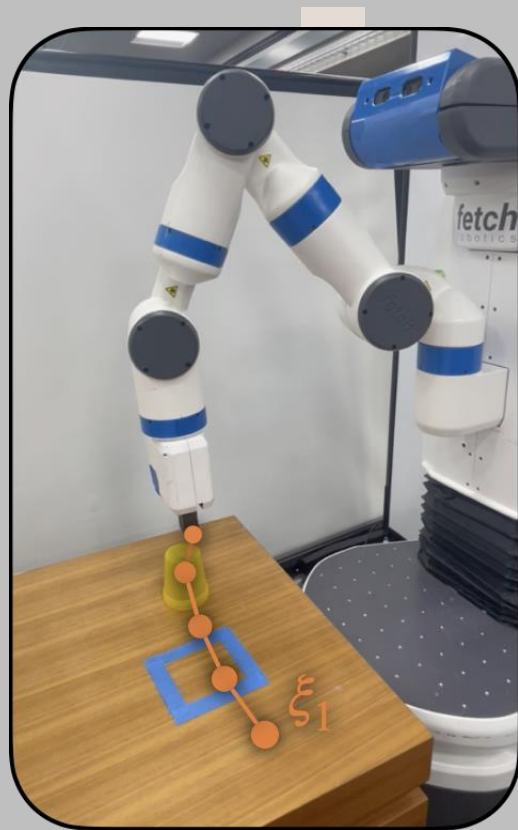


Expert demonstrations





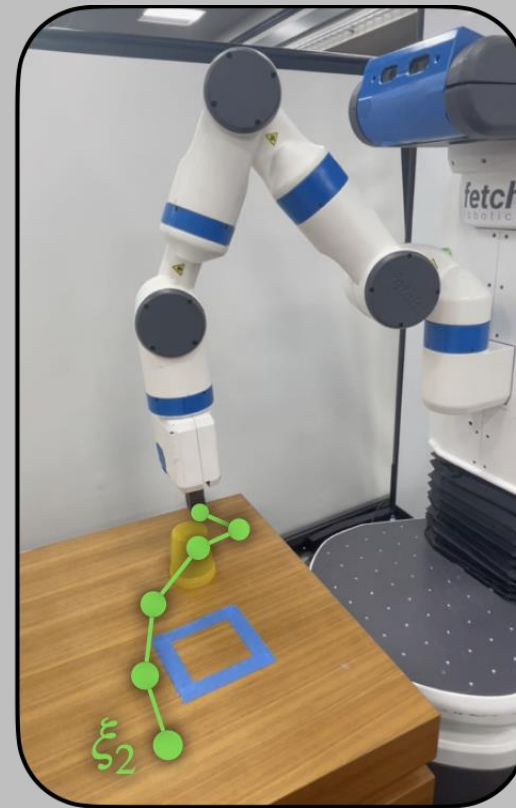
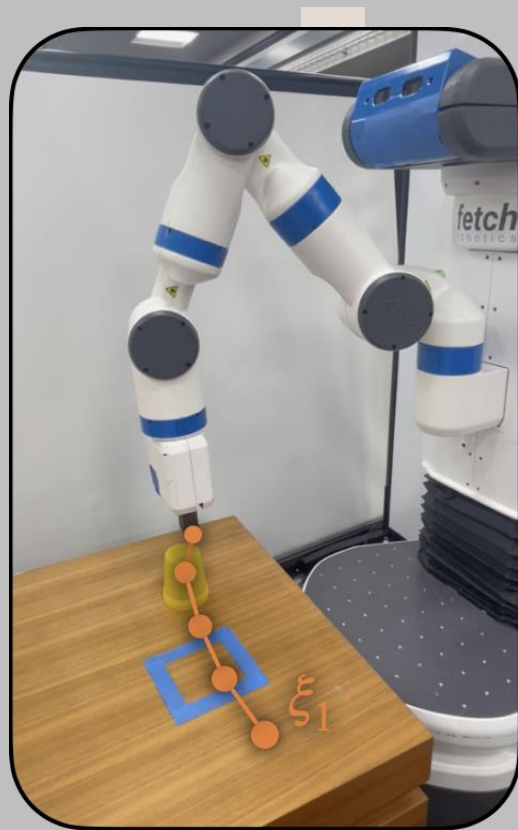
ξ_A or ξ_B ?



R_A or R_B ?



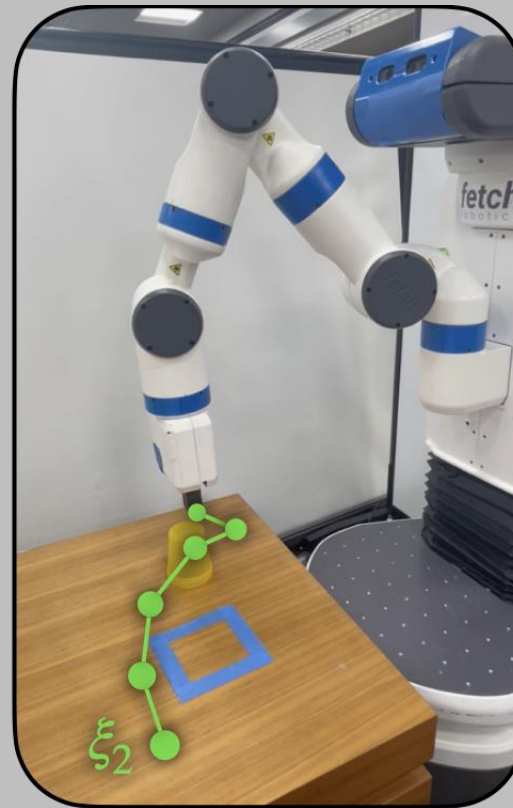
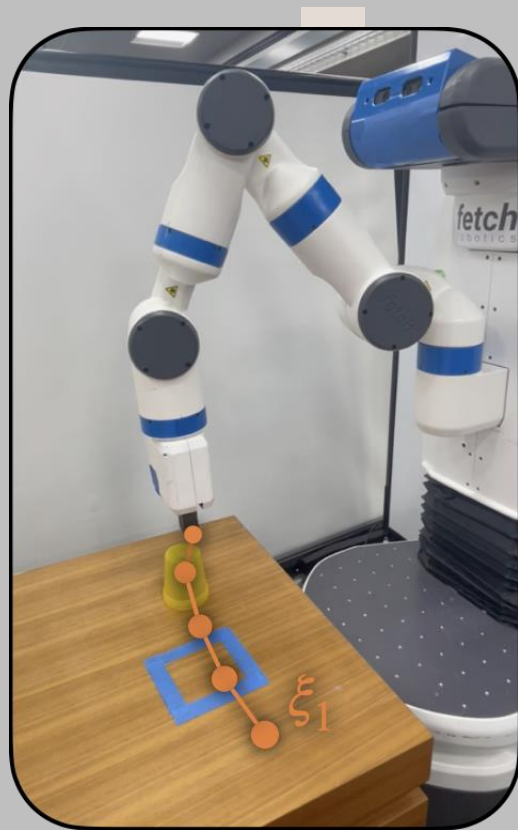
$$R(\xi) = w \cdot \phi(\xi)$$



R_A or R_B ?



or $w \cdot \phi_A > w \cdot \phi_B$
 $w \cdot \phi_A < w \cdot \phi_B$



$$\varphi = (\phi_A - \phi_B)$$

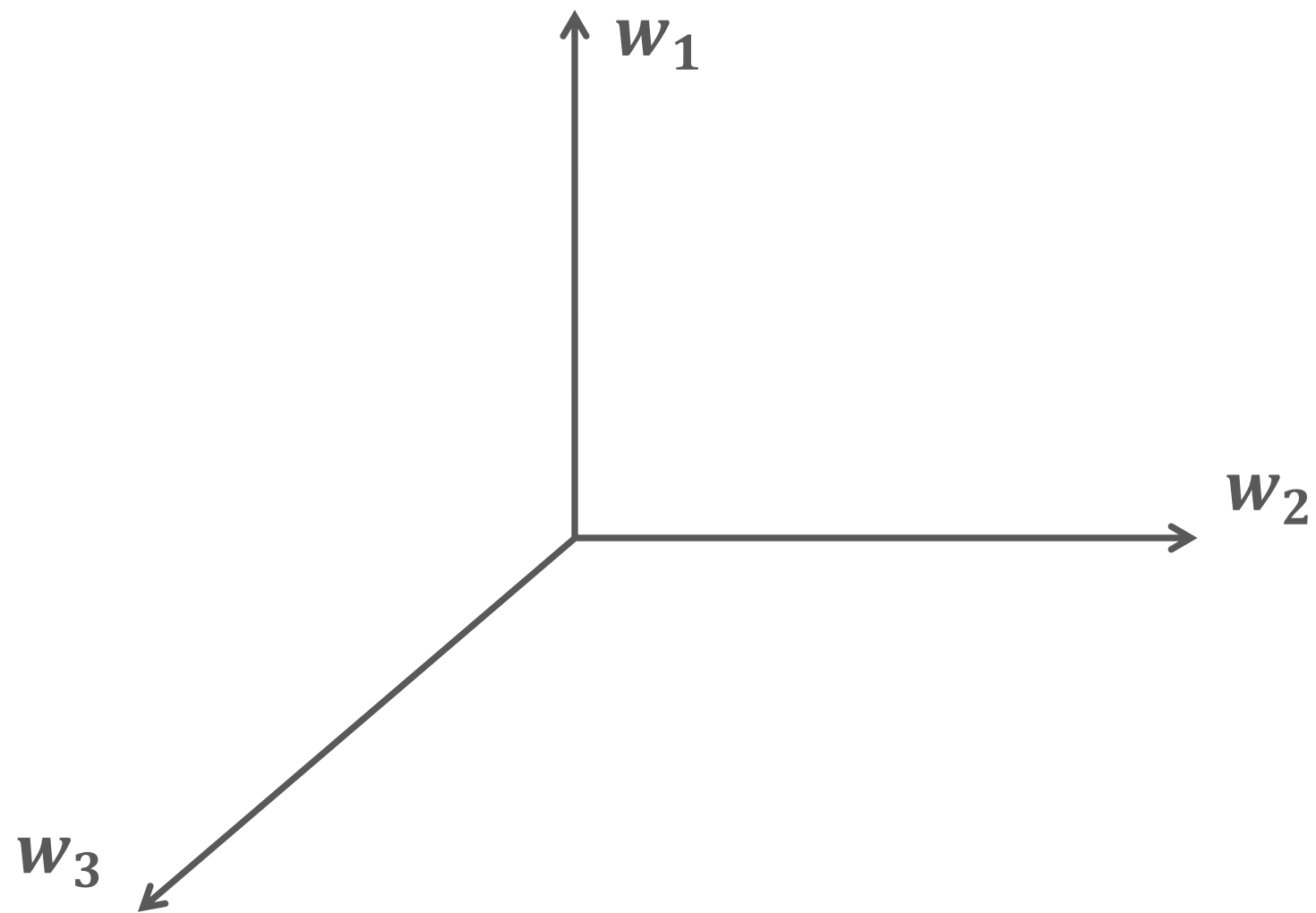


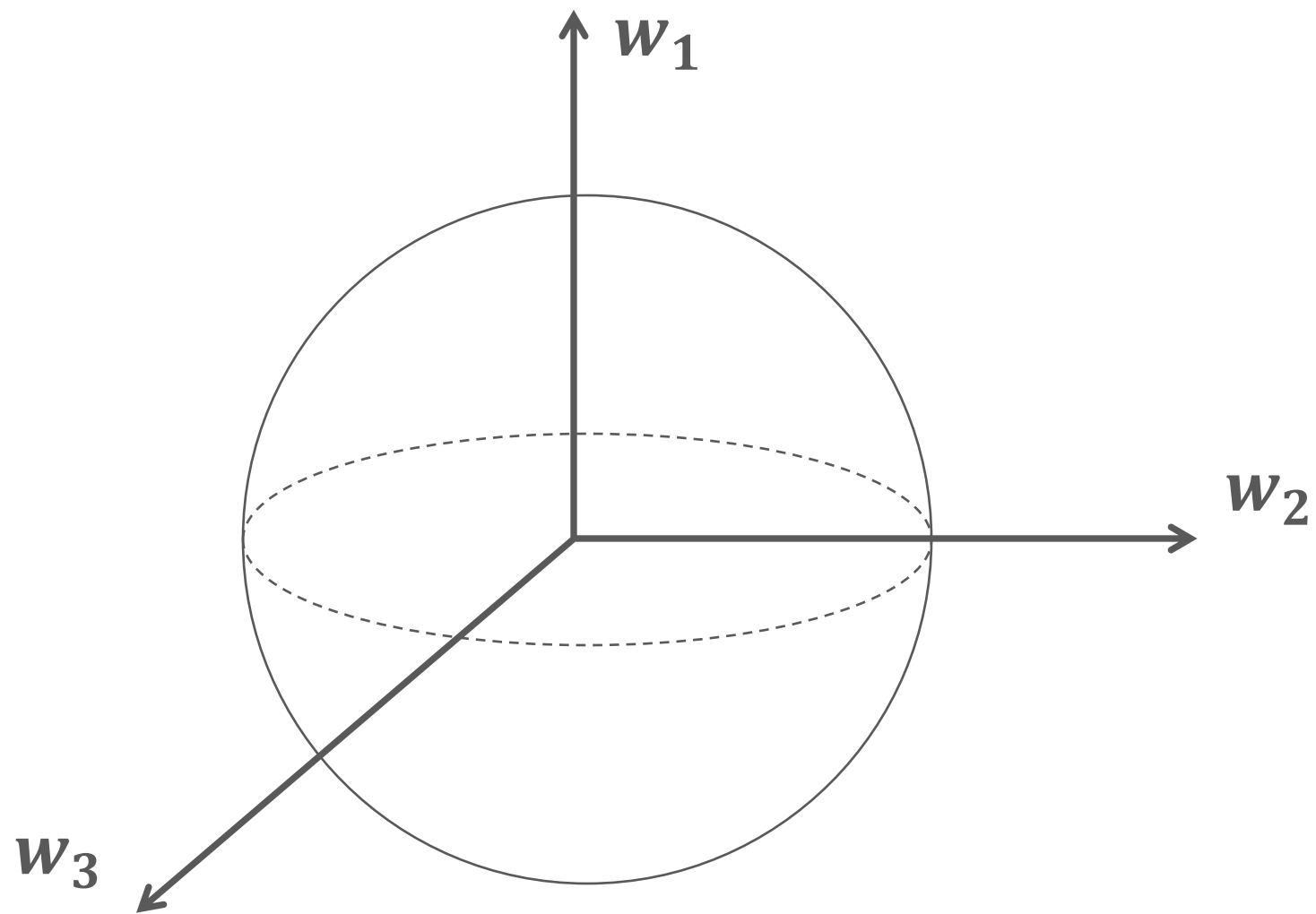
R_A or R_B ?

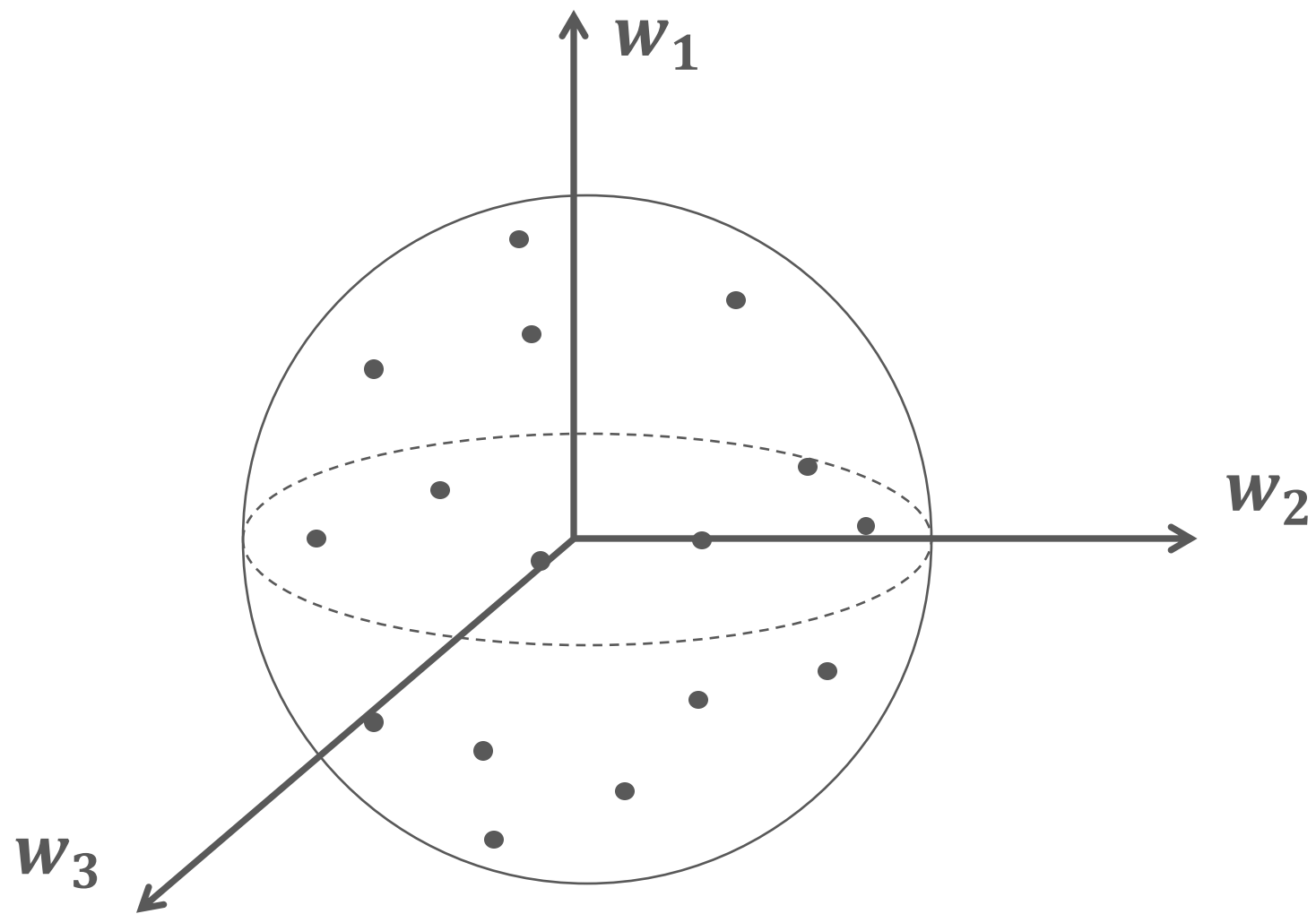


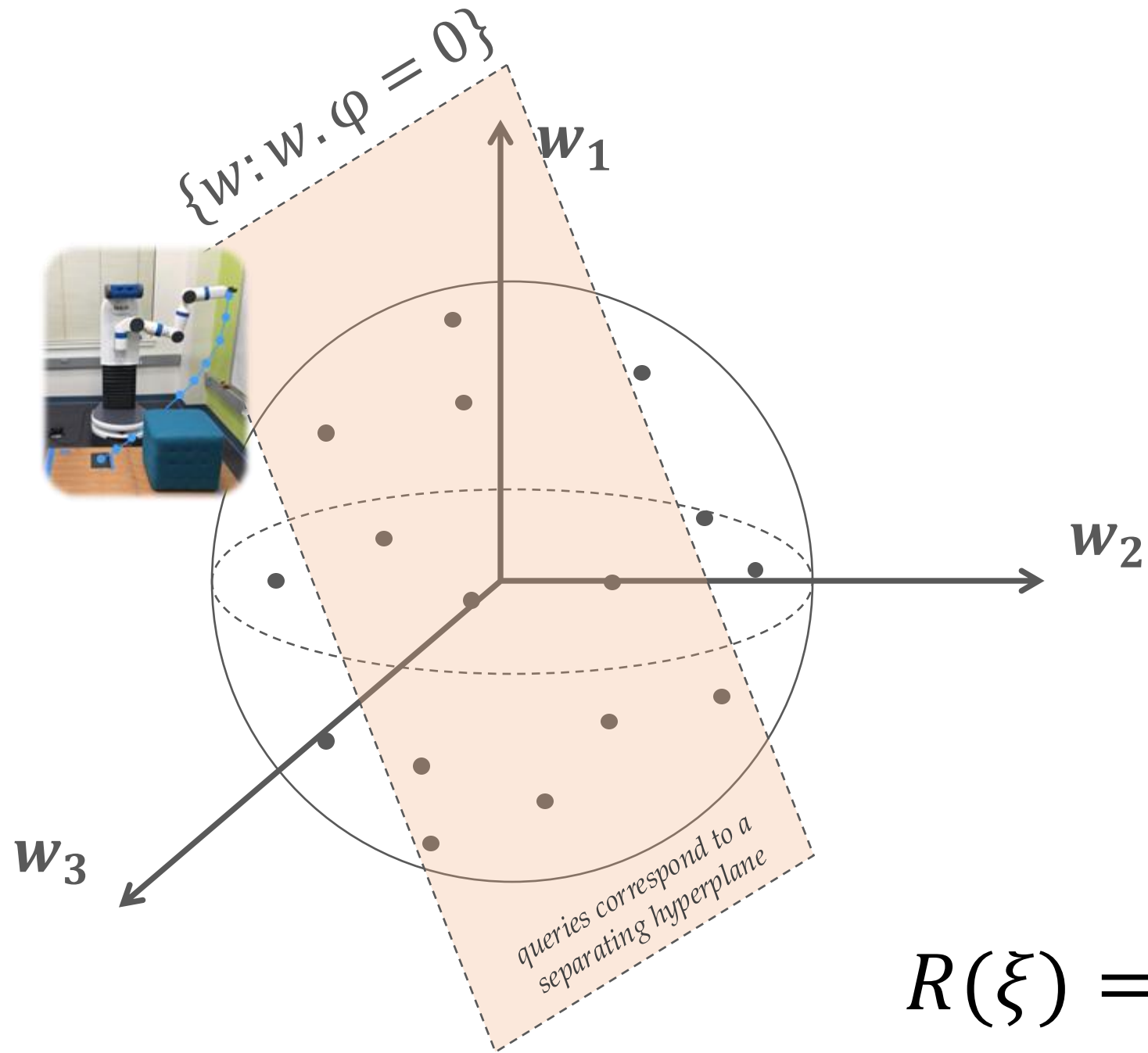
$$w \cdot \varphi > 0$$

or $w \cdot \varphi < 0$

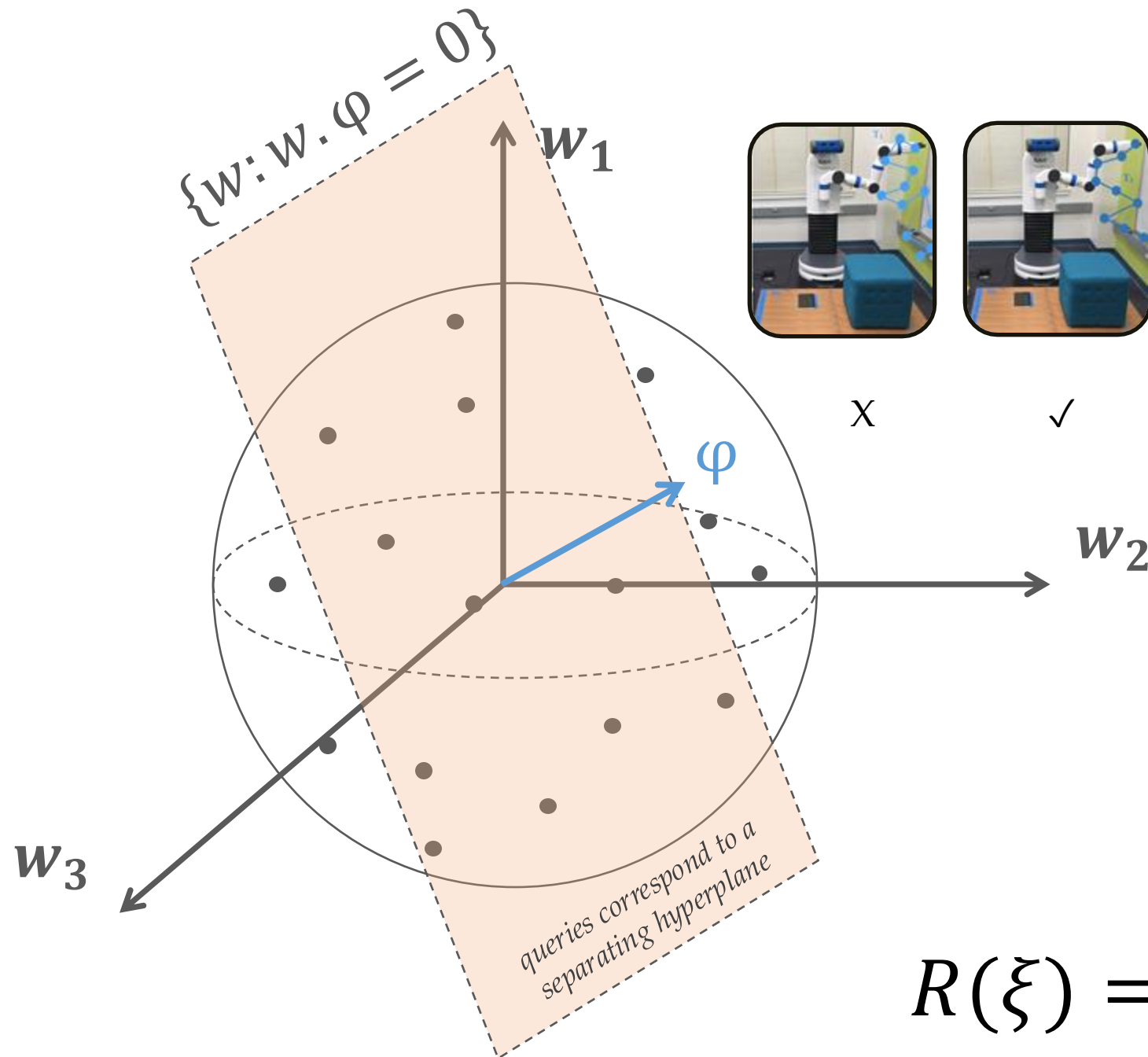








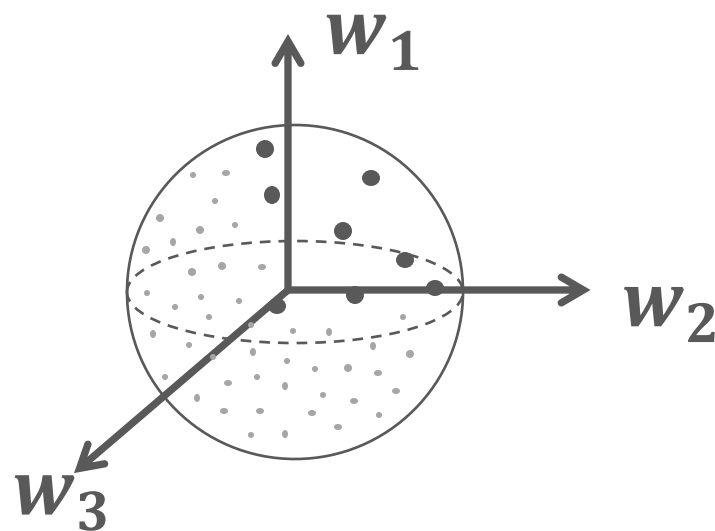
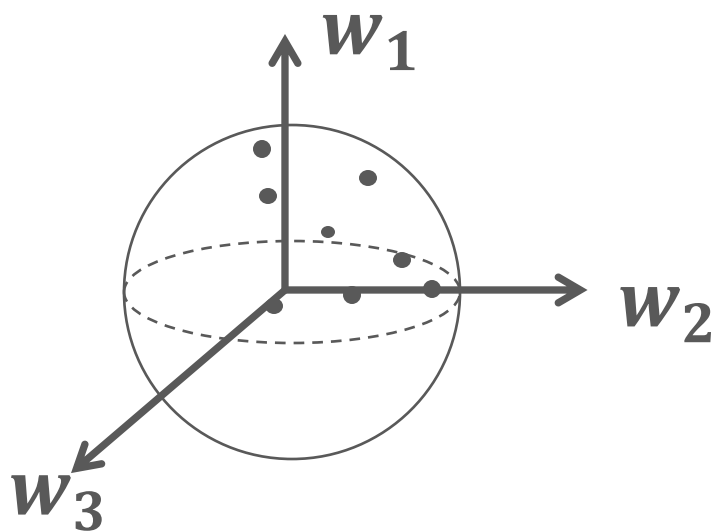
$$R(\xi) = w \cdot \phi(\xi)$$



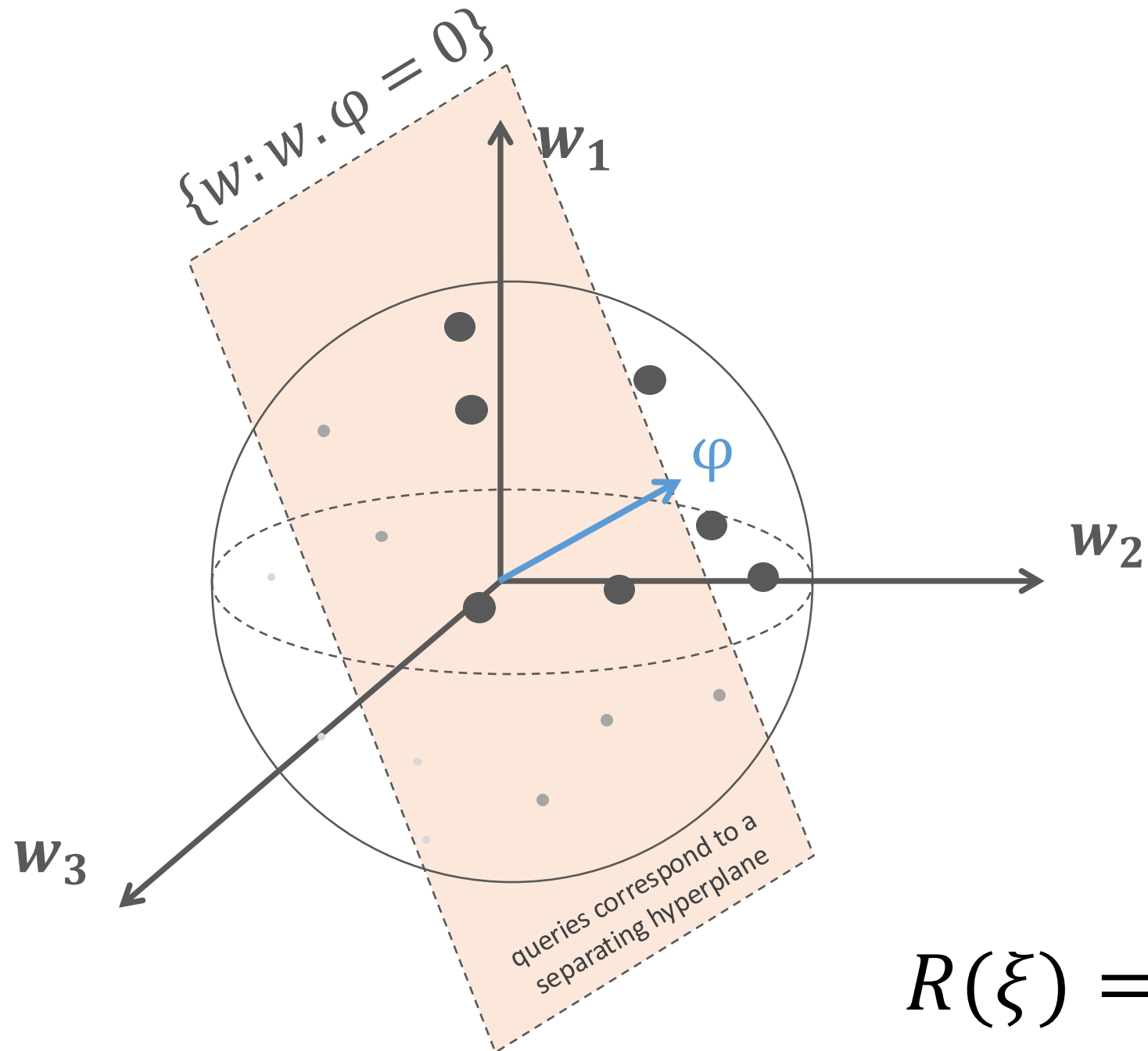
$$R(\xi) = w \cdot \phi(\xi)$$



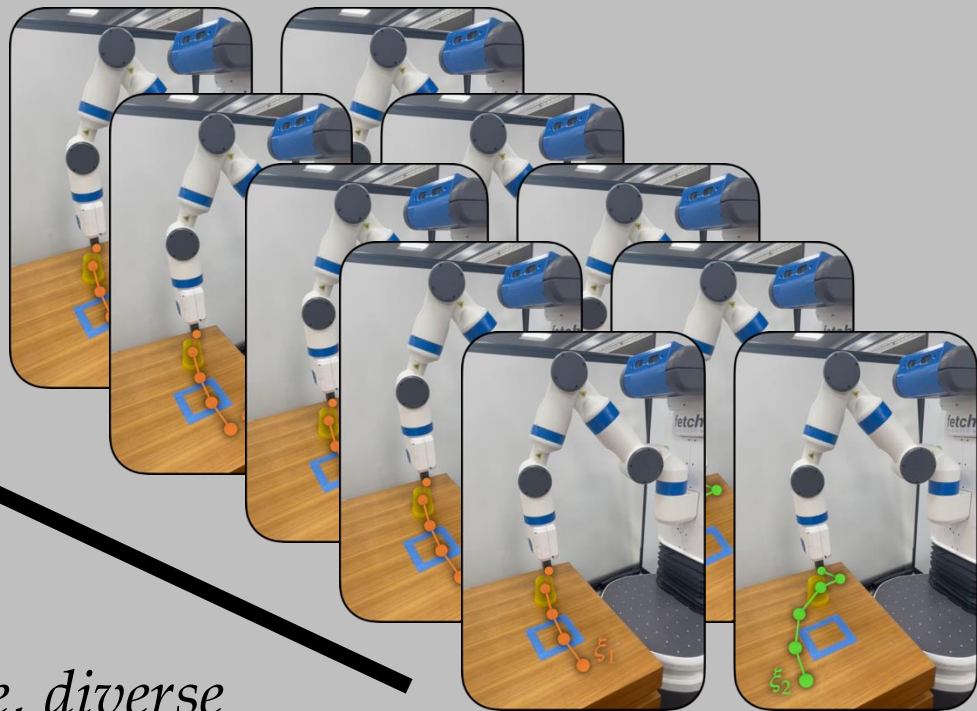
can be noisy!



$$f_{\varphi}(\mathbf{w}) = p(I_t | \mathbf{w}) = \frac{1}{1 + \exp(-I_t \mathbf{w}^T \varphi)}$$



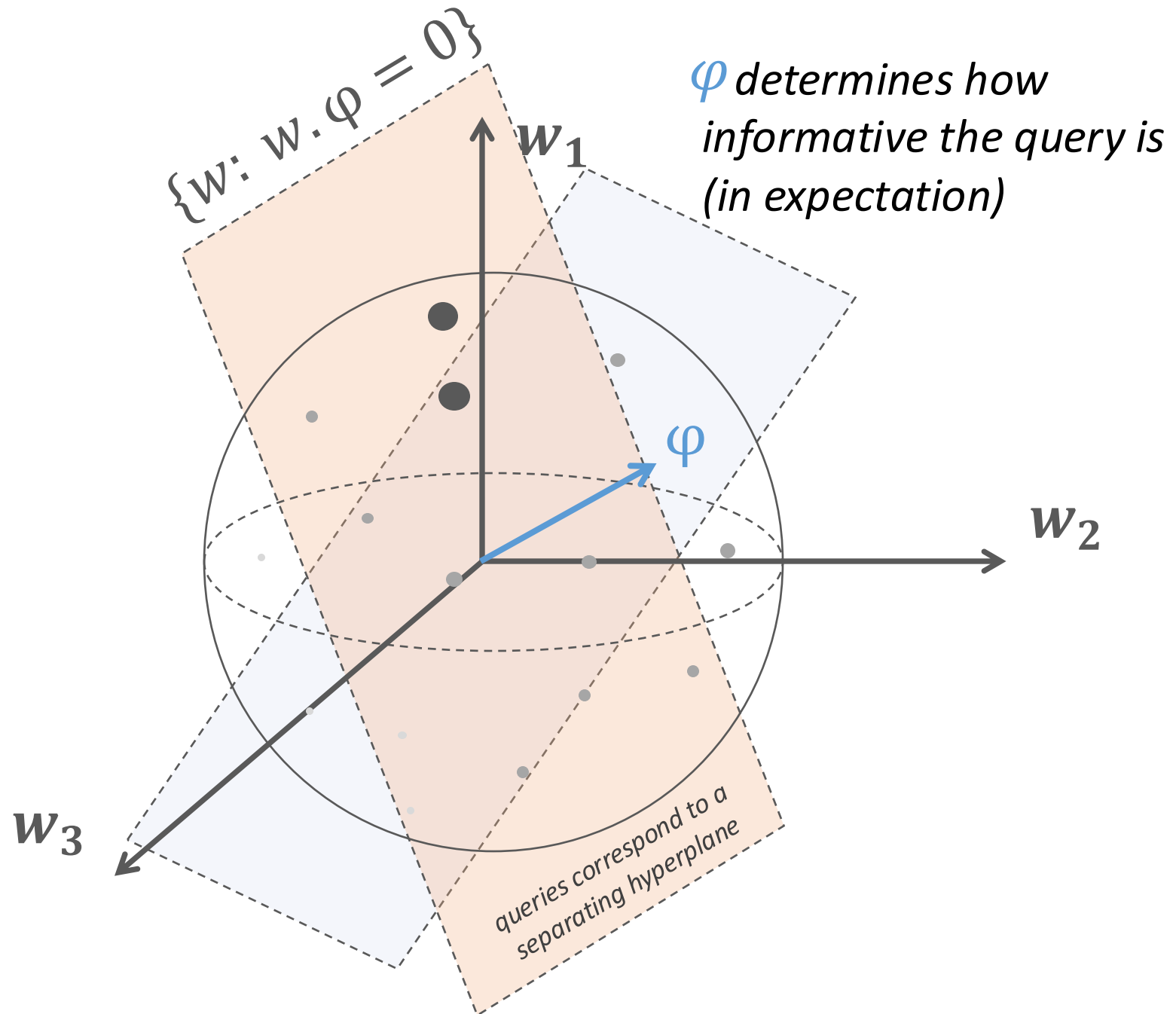
$$R(\xi) = w \cdot \phi(\xi)$$

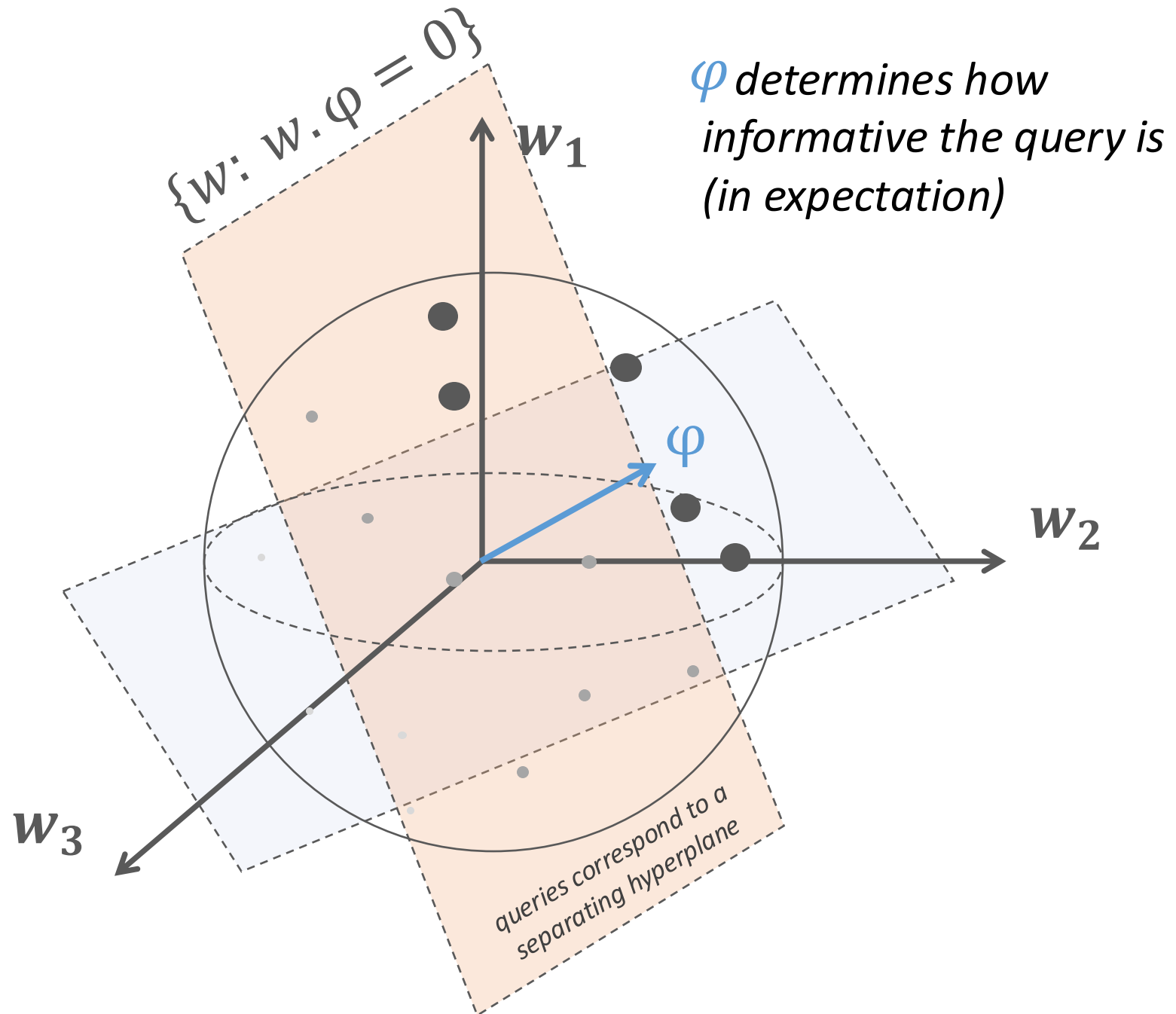


*Most informative, diverse
sequence of queries*



ξ_A or ξ_B ?





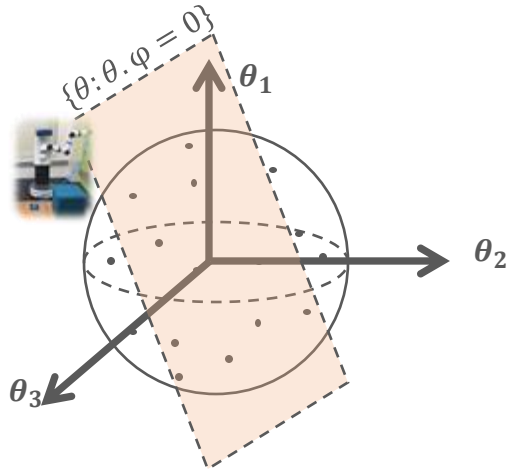
Queries should be actively synthesized



Erdem Biyik

Actively synthesizing queries

minimum volume removed

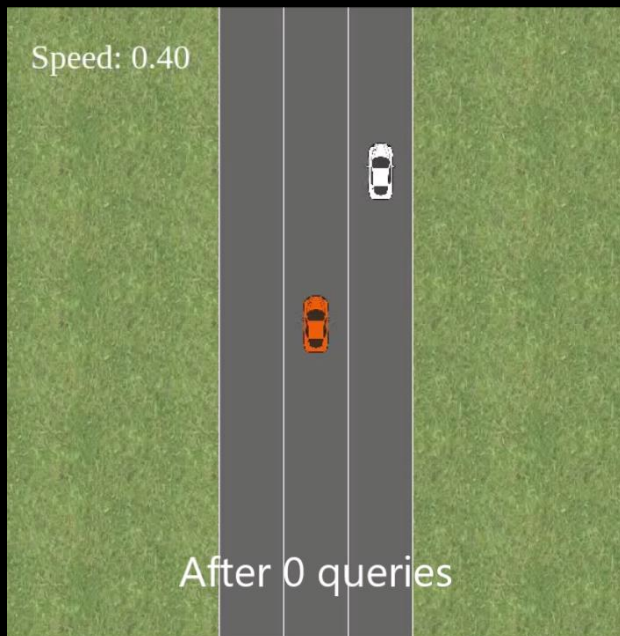


$$\max_{\varphi} \min\{\mathbb{E}[1 - f_{\varphi}(w)], \mathbb{E}[1 - f_{-\varphi}(w)]\}$$

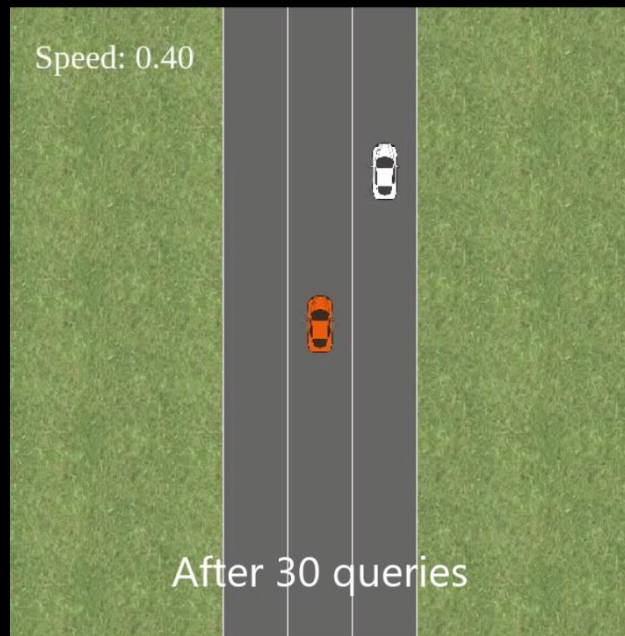
Subject to $\varphi \in \mathbb{F}$

$$\mathbb{F} = \{\varphi: \varphi = \Phi(\xi_A) - \Phi(\xi_B), \xi_A, \xi_B \in \Xi\}$$

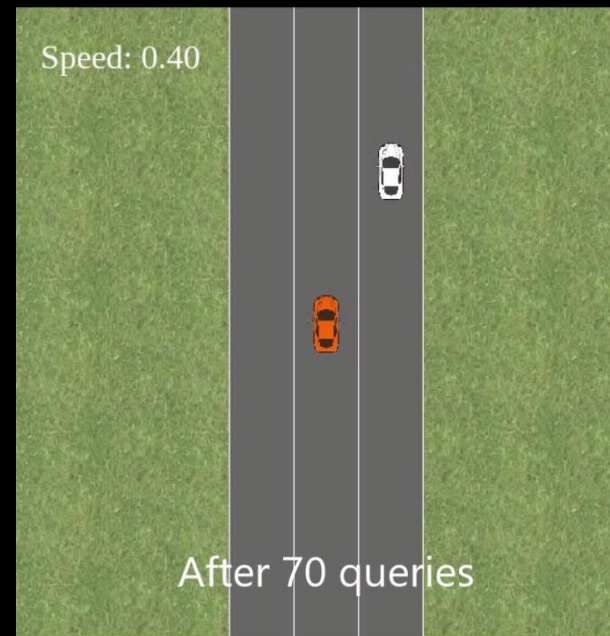
Human update function $f_{\varphi}(w) = \min(1, \exp(I_t w^T \varphi))$



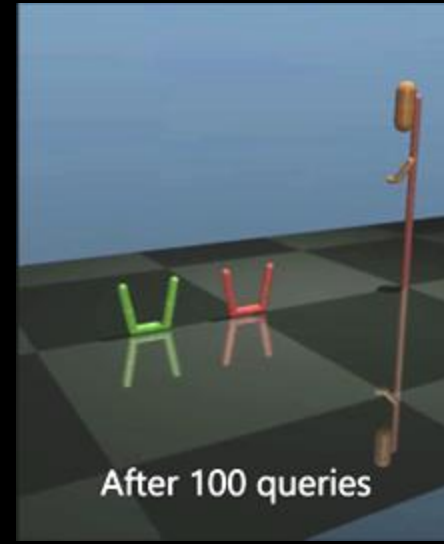
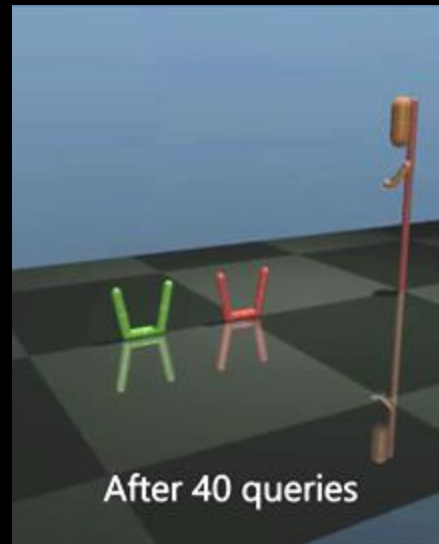
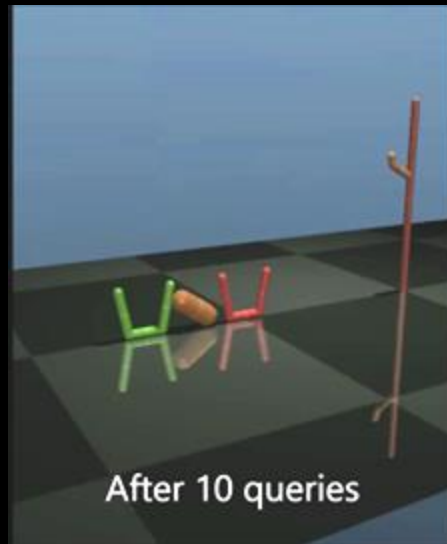
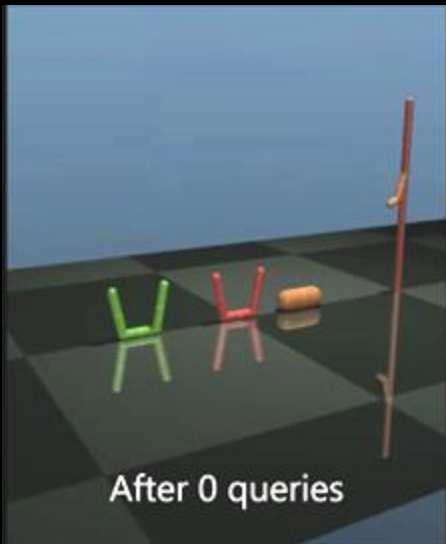
No prior preference



Learns *heading* preferences



Learns *collision avoidance* preferences

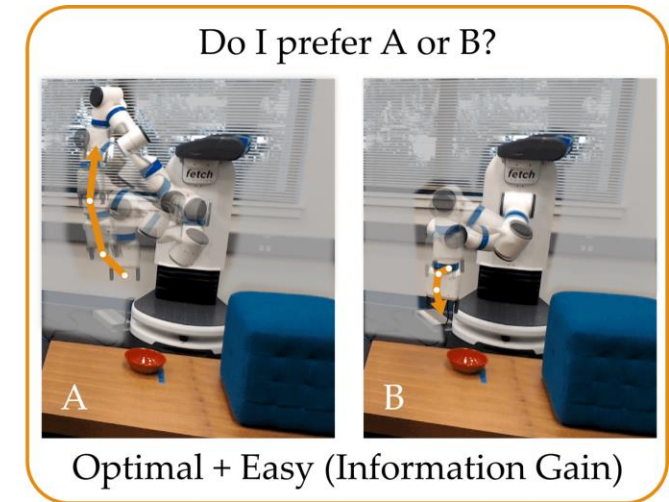
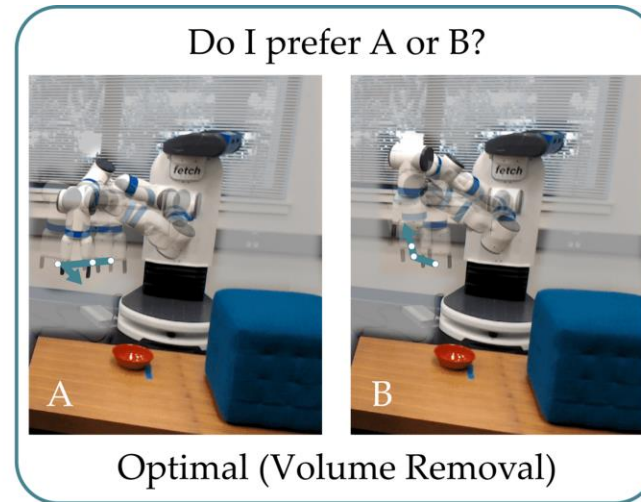
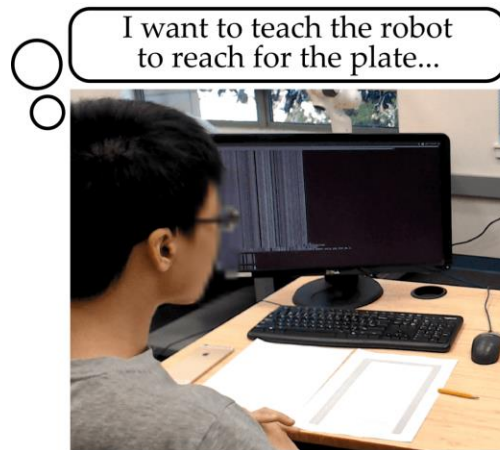


No prior preference

Preferring *green* basket over the *red* one.

Features: *max altitude, final distance to the closest basket, max horizontal range, total angular displacement*

Actively synthesizing queries

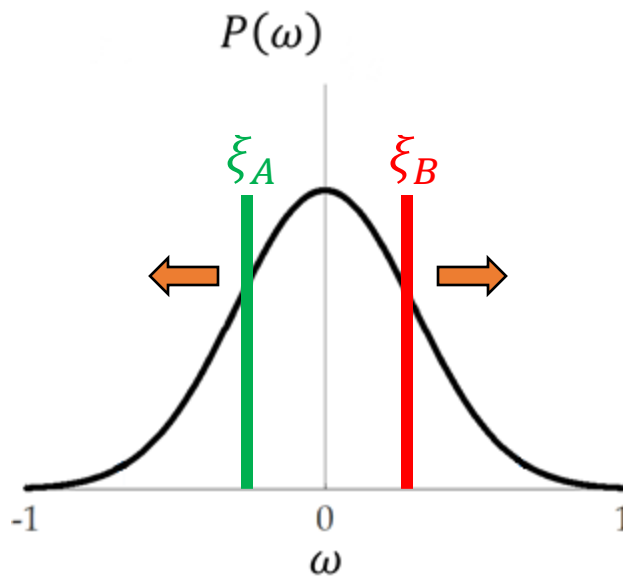


- Queries that are *easy to answer* for the human?

$$\begin{aligned} & \max I(\text{response}; \omega) \\ & = \max H(\text{response}) - H(\text{response}|\omega) \end{aligned}$$

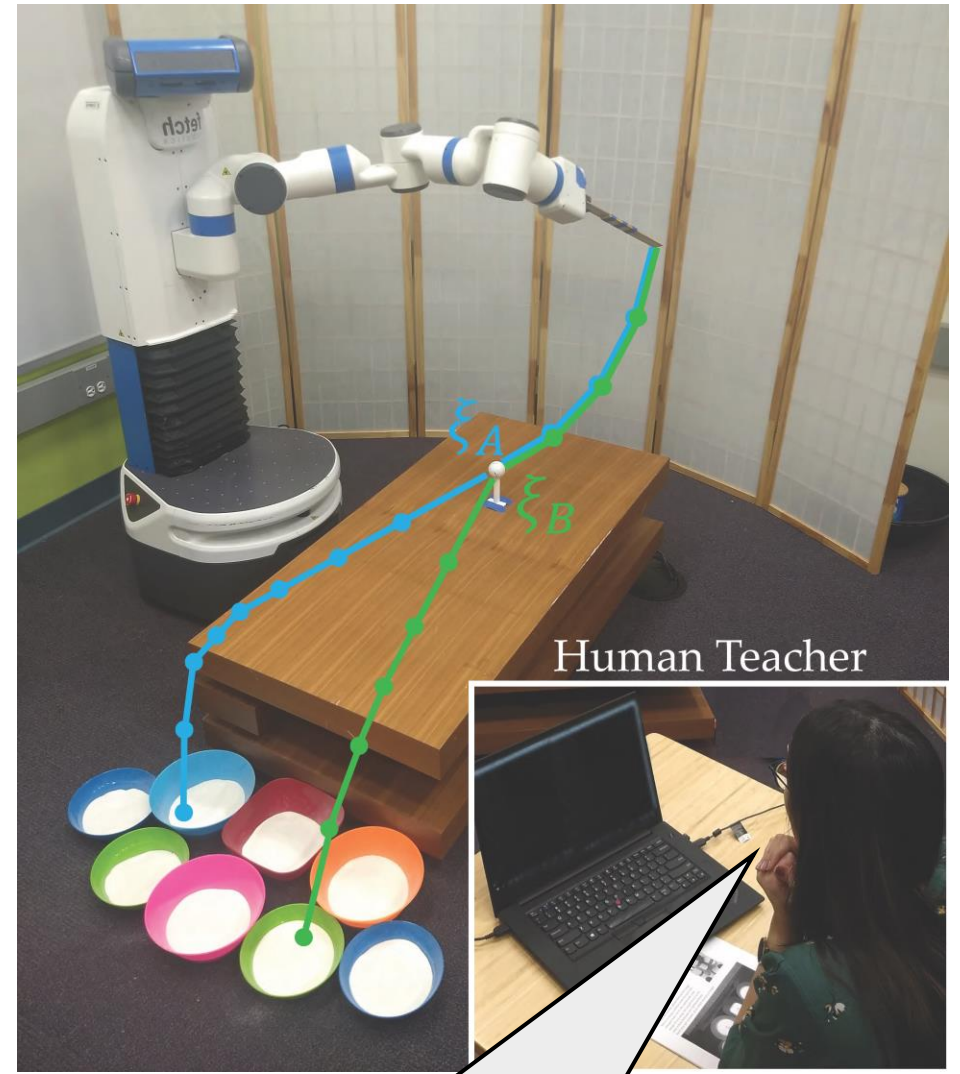
Robot's
uncertainty

Human's
uncertainty



$$R(\xi) = \theta \cdot \phi(\xi)$$





$$R(\xi_A) > R(\xi_B)$$

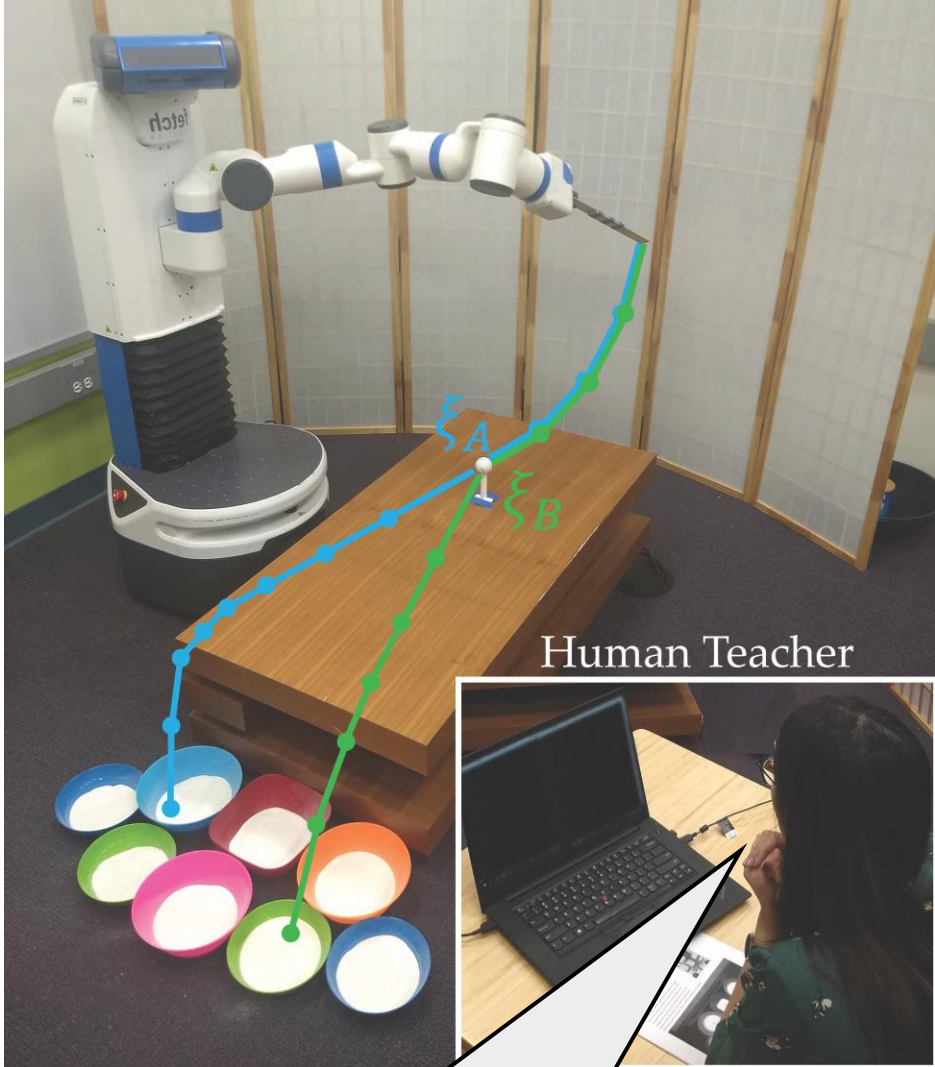
$$R(\xi_A) = \theta \cdot \phi(\xi_A)$$

 Designing features is hard.

$e^{-c_4 d_4}$ where d_4 is the final horizontal distance between the object and the center of the closest basket, and $c_4 = 3$.

The average of $e^{-c_2 d_2^2 - c_3 d_3^2}$ over the trajectory, where d_2 and d_3 are the horizontal and vertical distances between the ego car and the other car, respectively; and $c_2 = 7, c_3 = 3$

Feature generation?
Deep learning?

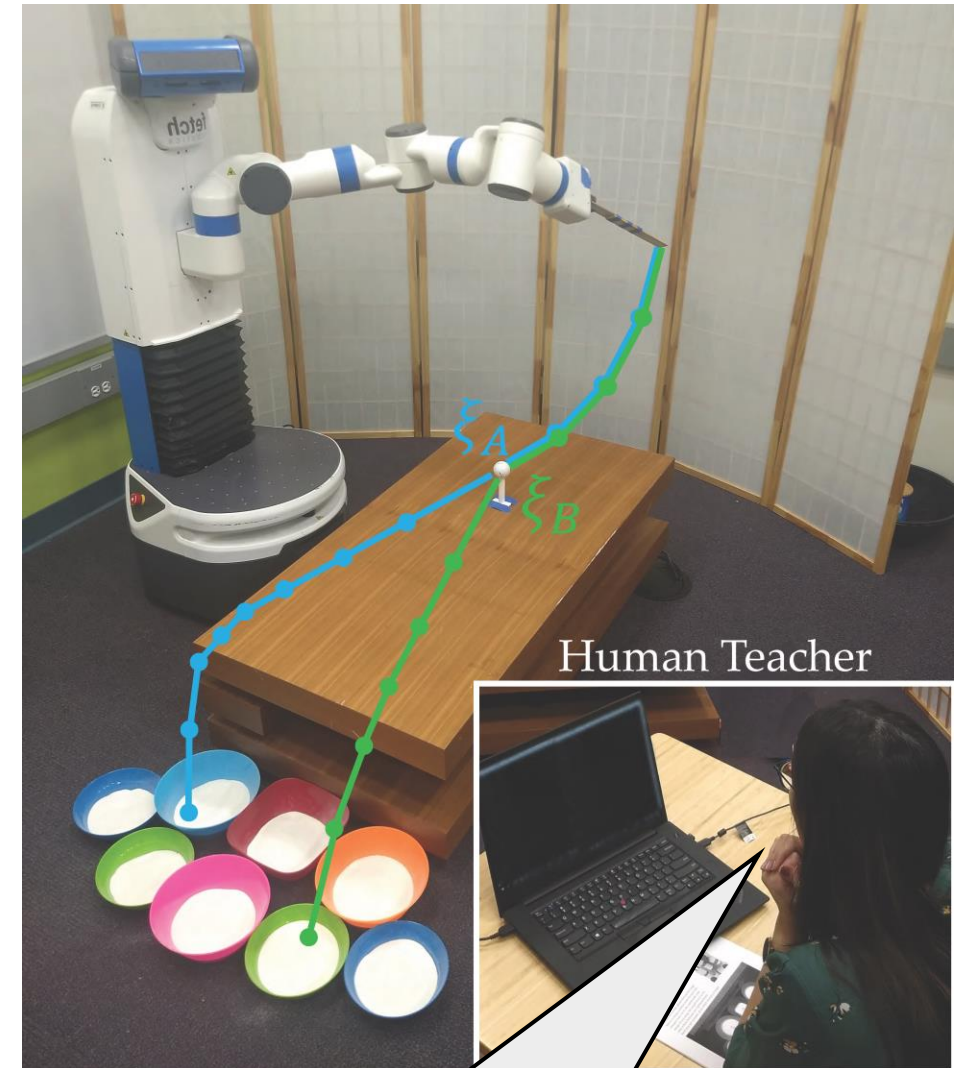
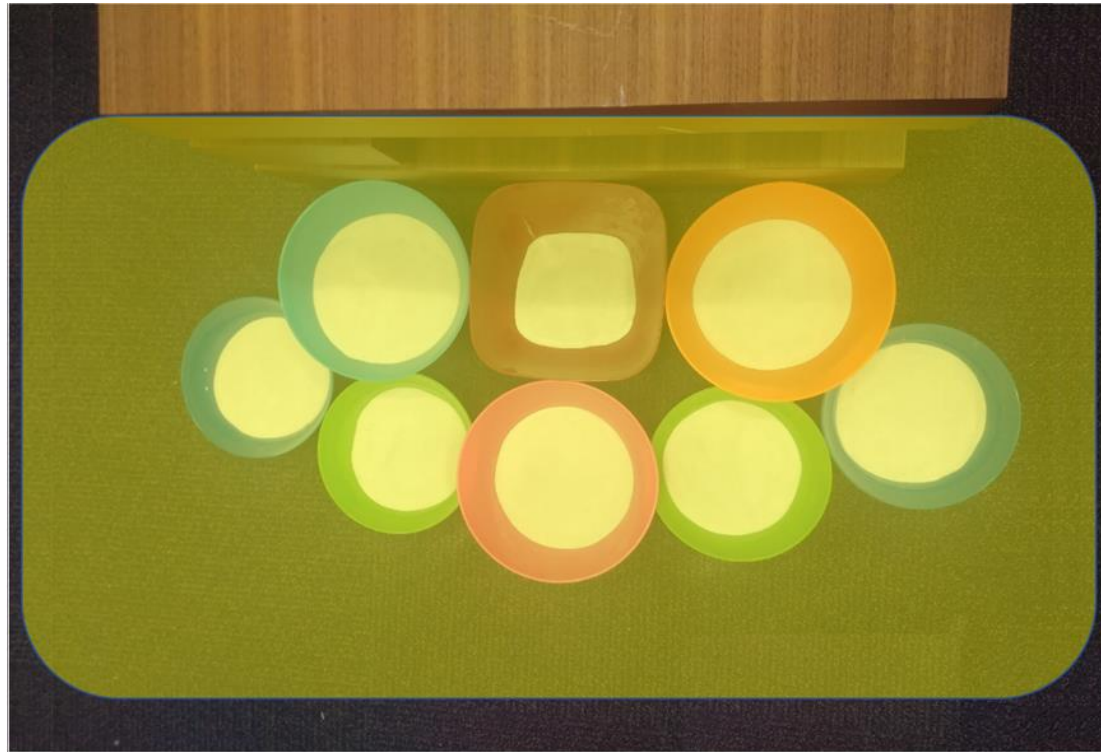


$$R(\xi_A) > R(\xi_B)$$

- [Sadigh'17]
- [Basu'18]
- [Bryk'18, 19]
- [Katz'19]
- [Chen'19]
- [Wilde'20]

Trajectory Features: Shot Speed, Shot Angle

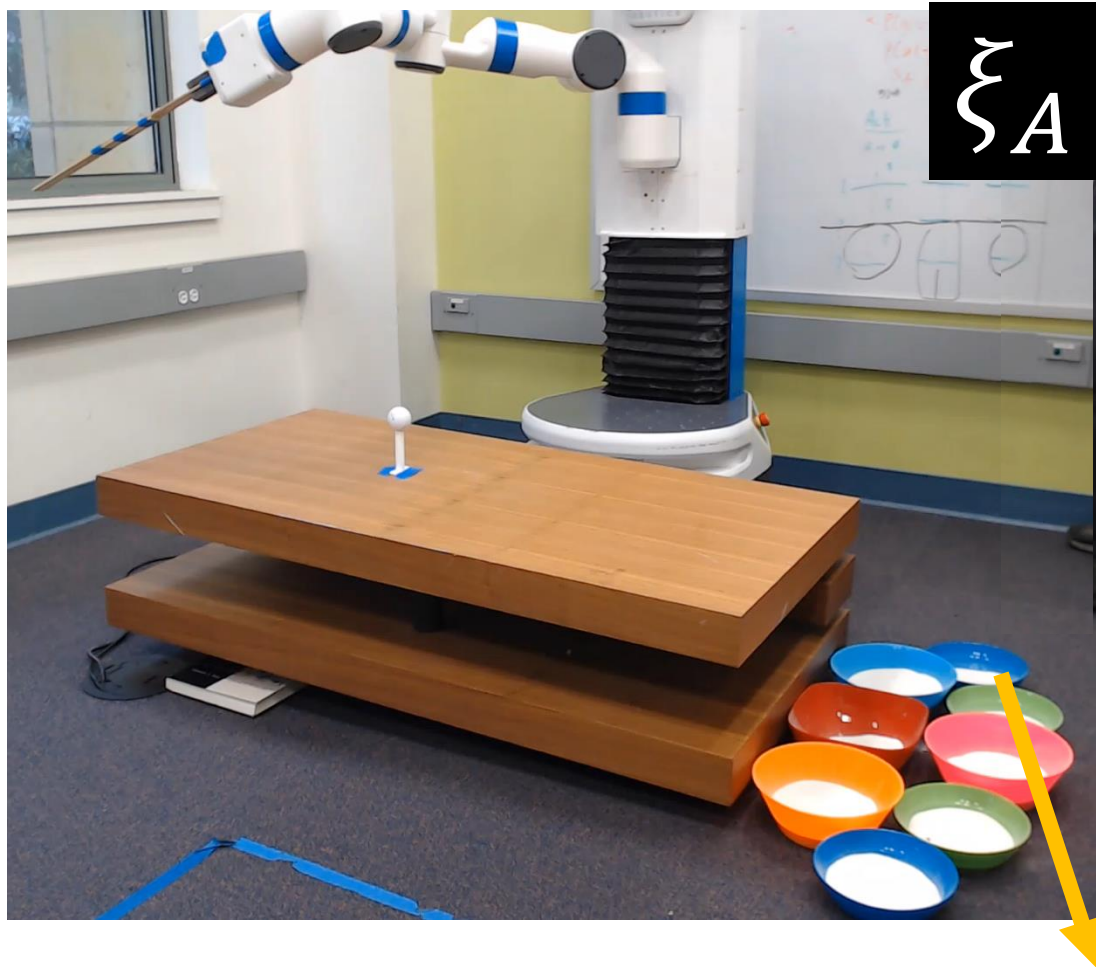
$$R(\xi_A) = \theta(\phi(\xi_A))$$



Human Teacher

$$R(\xi_A) > R(\xi_B)$$

Training: An Optimal Query with GP Reward



GP Reward enables the exploration of different trajectories (not just the boundaries).

Online: Final policy based on learned reward

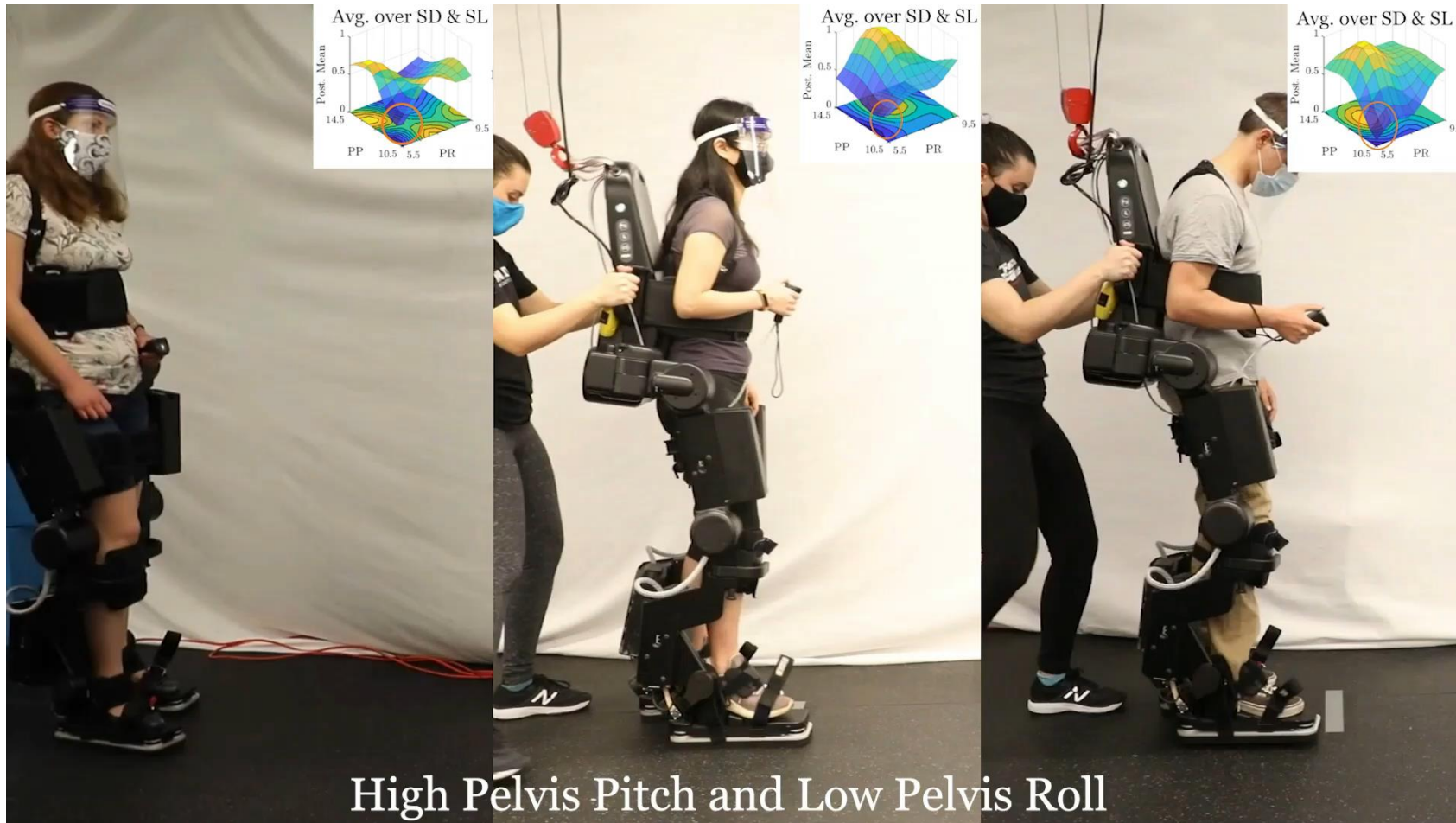


Linear Reward



GP Reward

Nonlinear Rewards for Exoskeletons

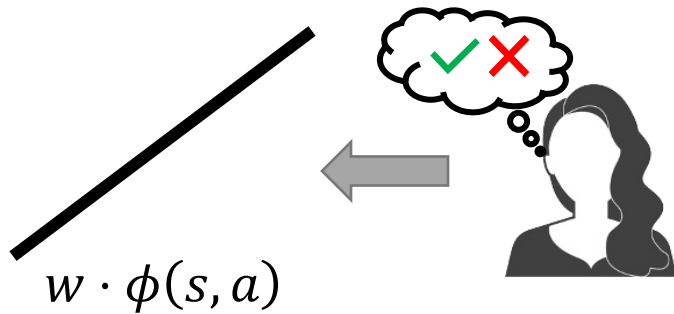


High Pelvis Pitch and Low Pelvis Roll

Linear Models:

- Inexpressive

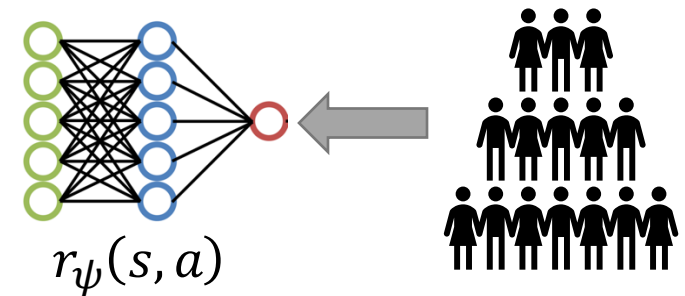
+ Feedback efficient



Neural Models:

- Thousands of Queries

+ Highly expressive



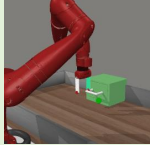
Few-Shot Preference Learning for Human-in-the-Loop RL

Pre-training

Prior Tasks



Door Open



Window Open



Button Press

Few-Shot Preference Learning for Human-in-the-Loop RL

Pre-training

Prior Tasks



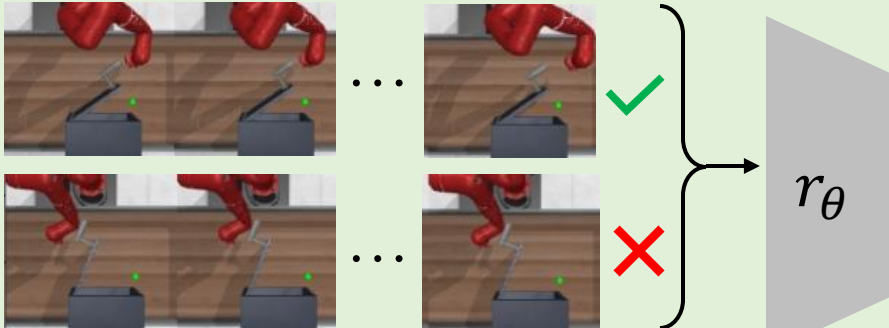
Door Open



Window Open



Button Press



Segments σ_1, σ_2

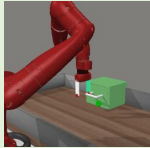
Few-Shot Preference Learning for Human-in-the-Loop RL

Pre-training

Prior Tasks



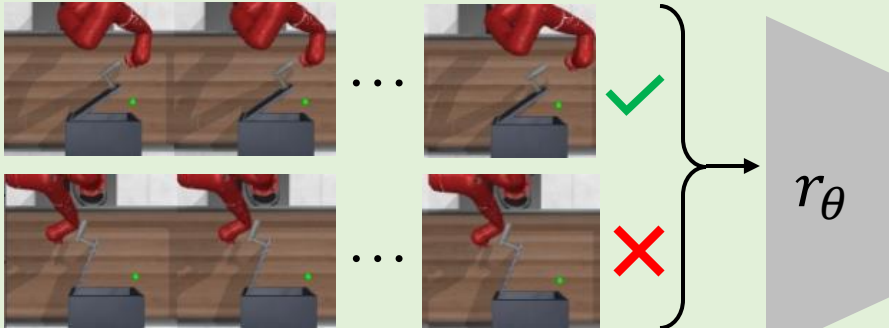
Door Open



Window Open



Button Press



Segments σ_1, σ_2

Online Adaptation

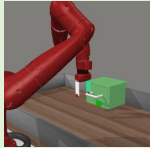
Few-Shot Preference Learning for Human-in-the-Loop RL

Pre-training

Prior Tasks



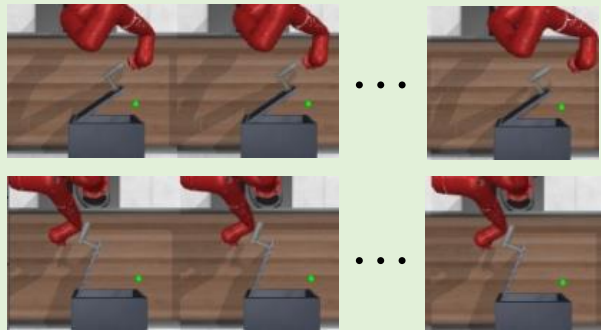
Door Open



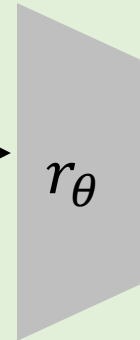
Window Open



Button Press

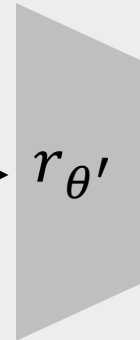


Segments σ_1, σ_2



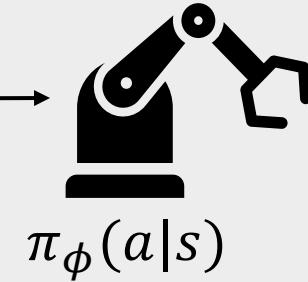
r_θ

Online Adaptation



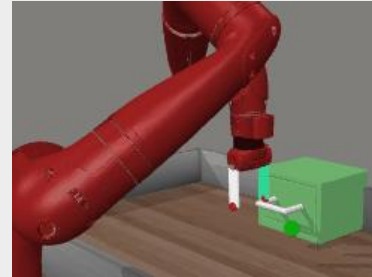
$r_{\theta'}$

$r_{\theta'}(s, a)$



$\pi_\phi(a|s)$

a



New Task

Drawer Open

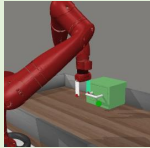
Few-Shot Preference Learning for Human-in-the-Loop RL

Pre-training

Prior Tasks



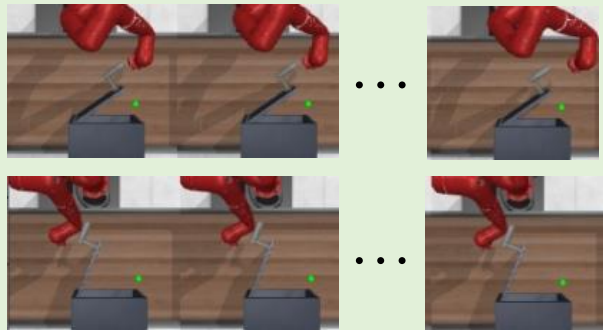
Door Open



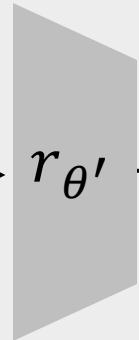
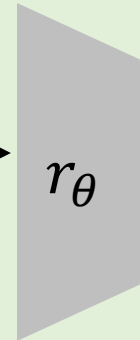
Window Open



Button Press

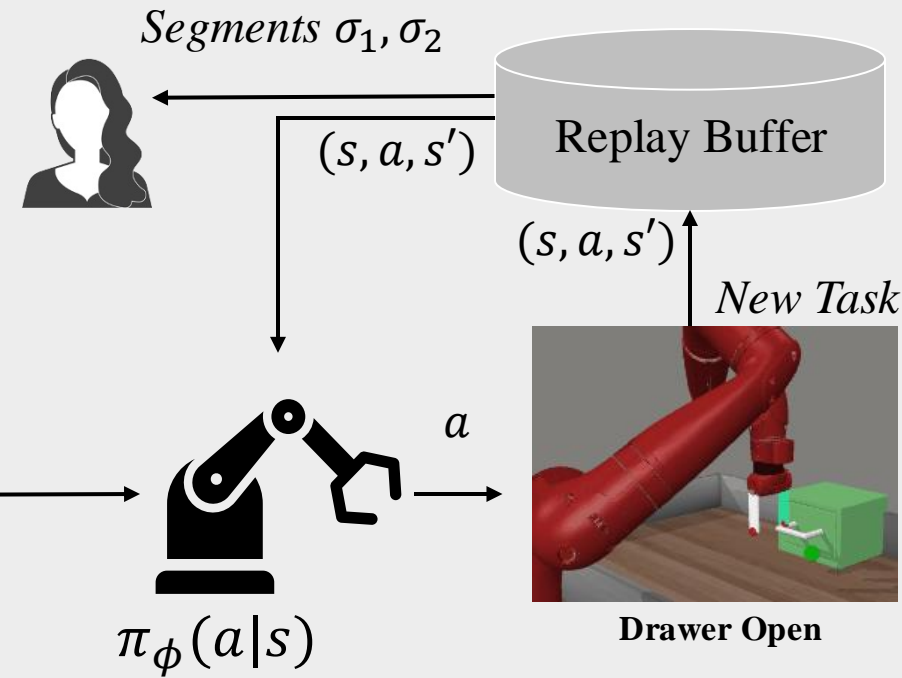


Segments σ_1, σ_2

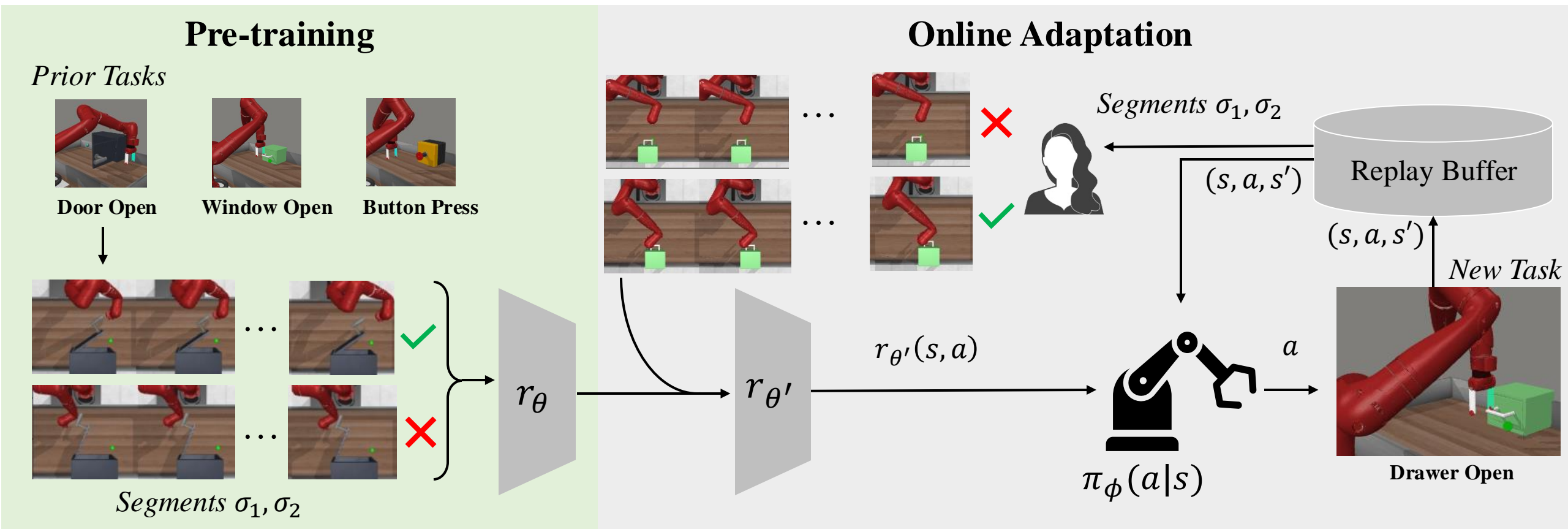


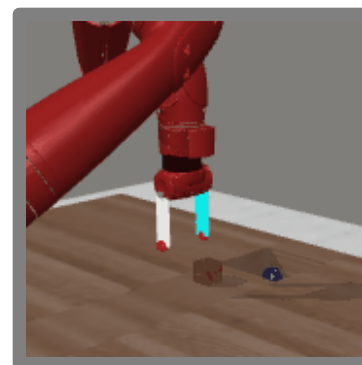
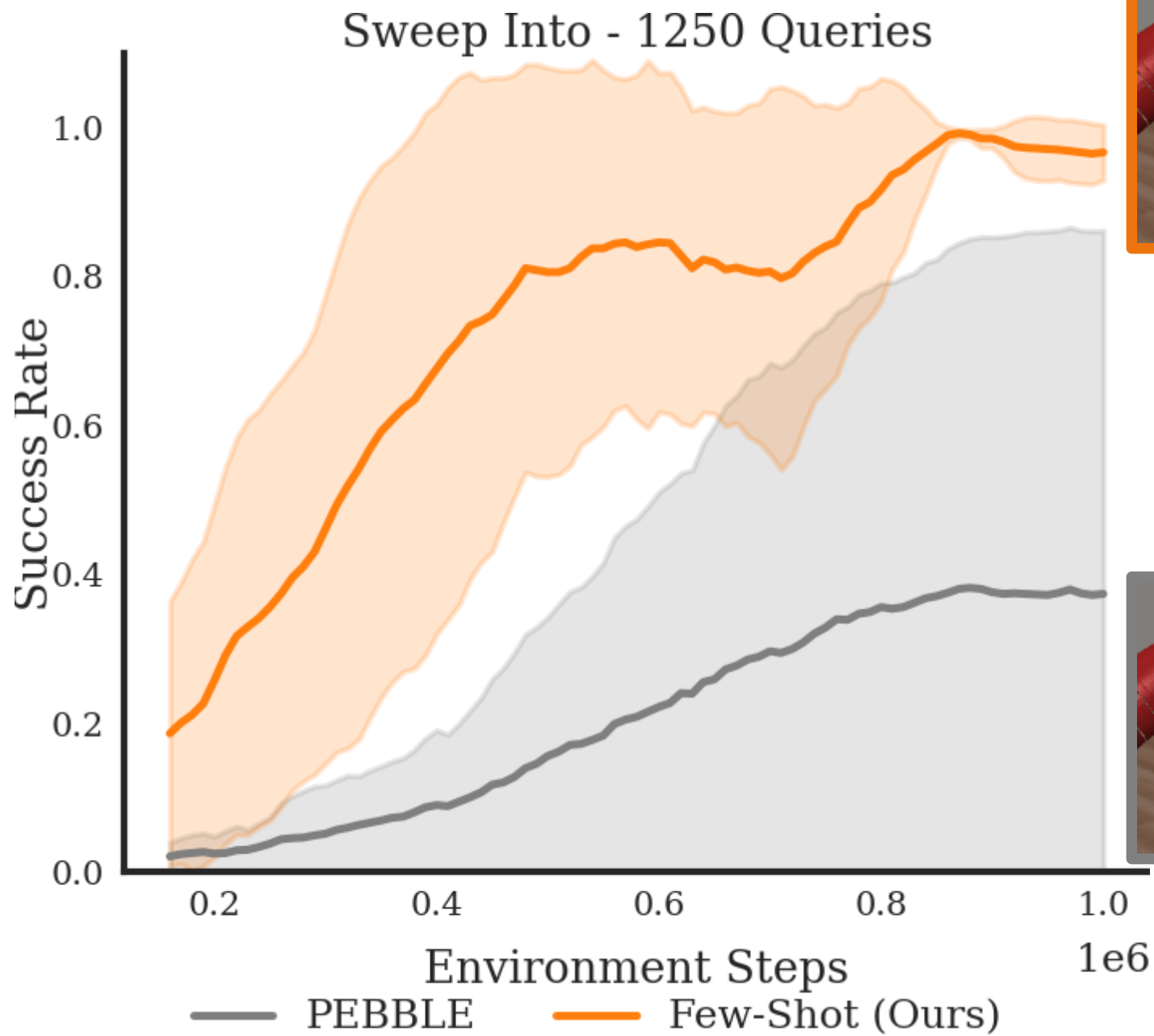
$r_{\theta'}(s, a)$

Online Adaptation

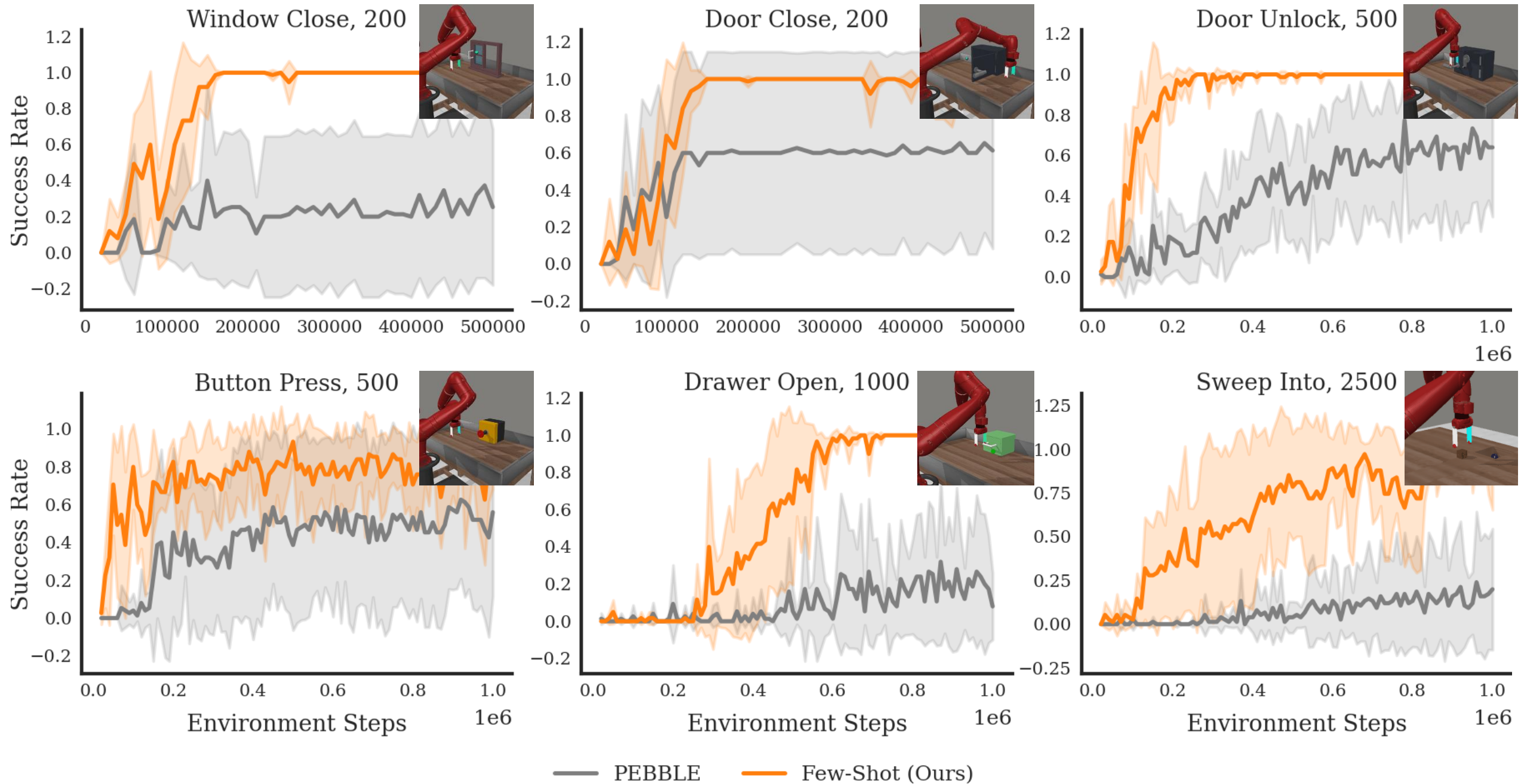


Few-Shot Preference Learning for Human-in-the-Loop RL





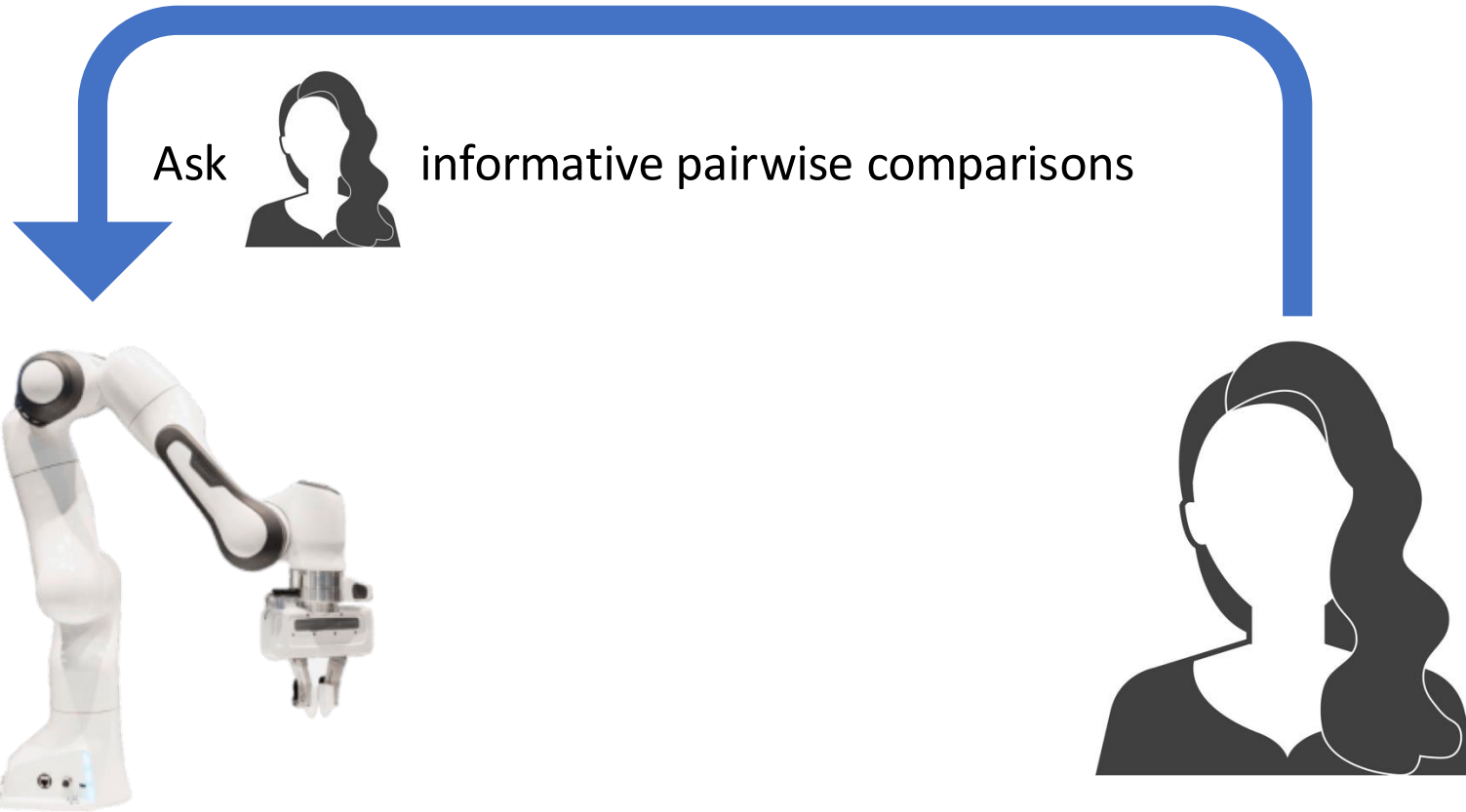
Test Tasks



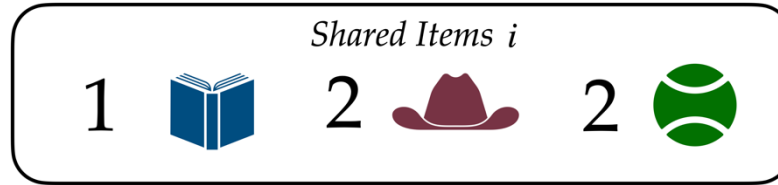
Today's itinerary

- Recap of imitation learning and inverse RL
- Learning from other sources of data – Pairwise Comparisons
- Learning from other sources of data – Foundation Models
- Learning from physical feedback
- Learning from gestures
- Learning from sketches
- Data Quality

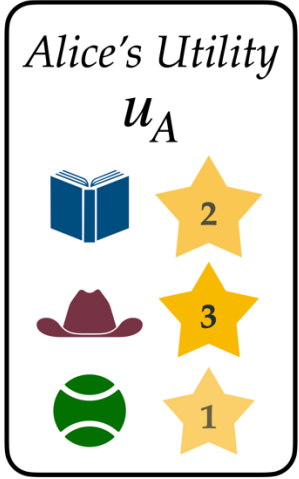
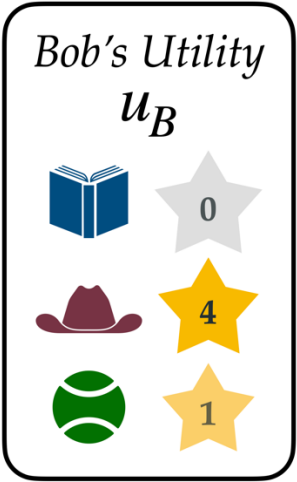
Learn Human Preferences



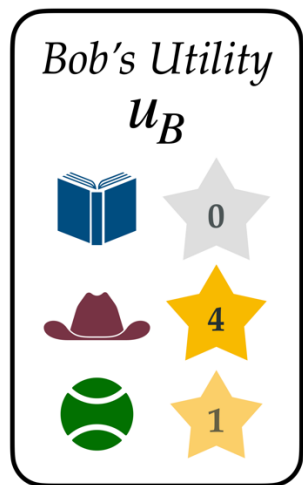
Negotiation Domain



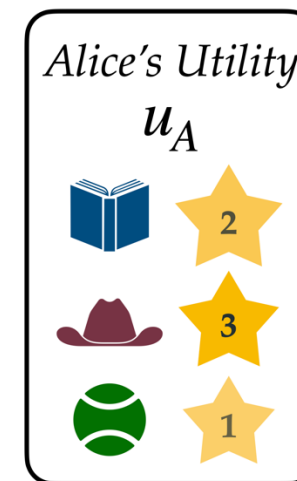
Negotiation Domain



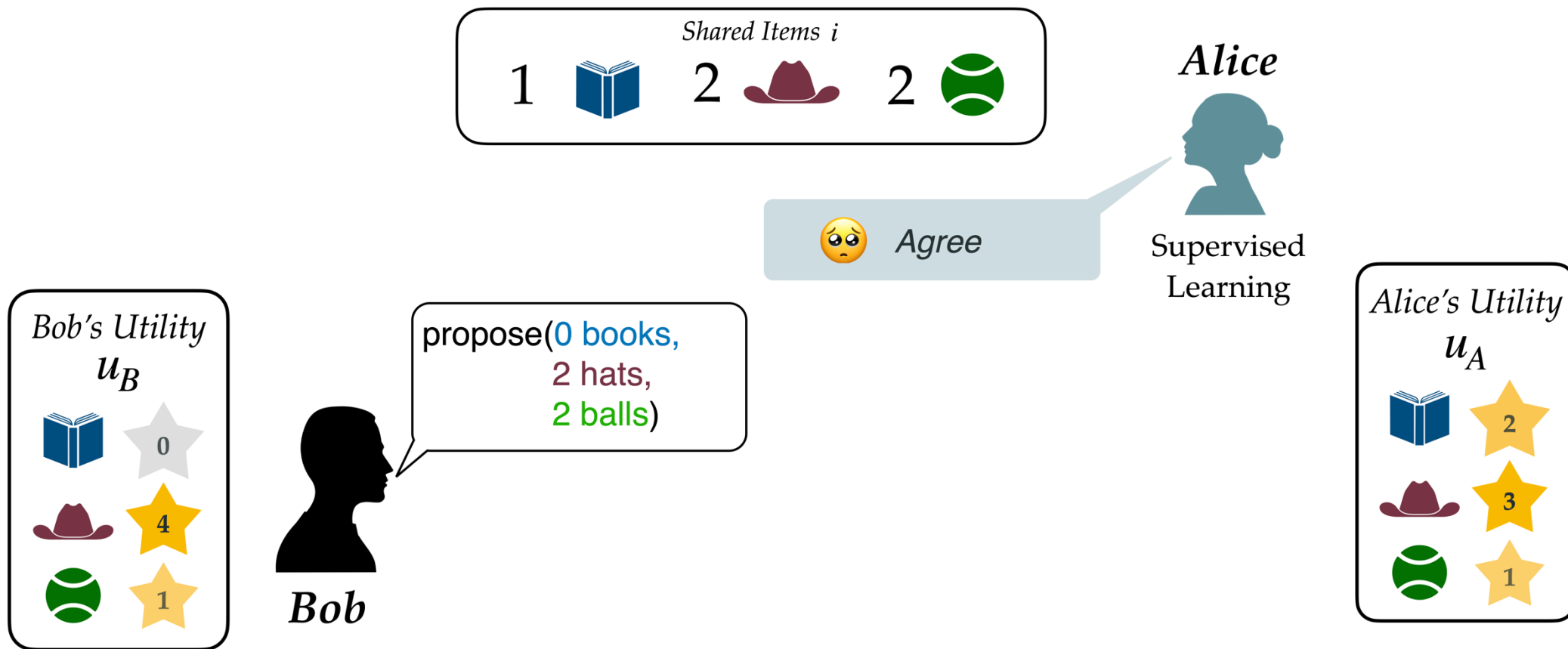
Negotiation Domain



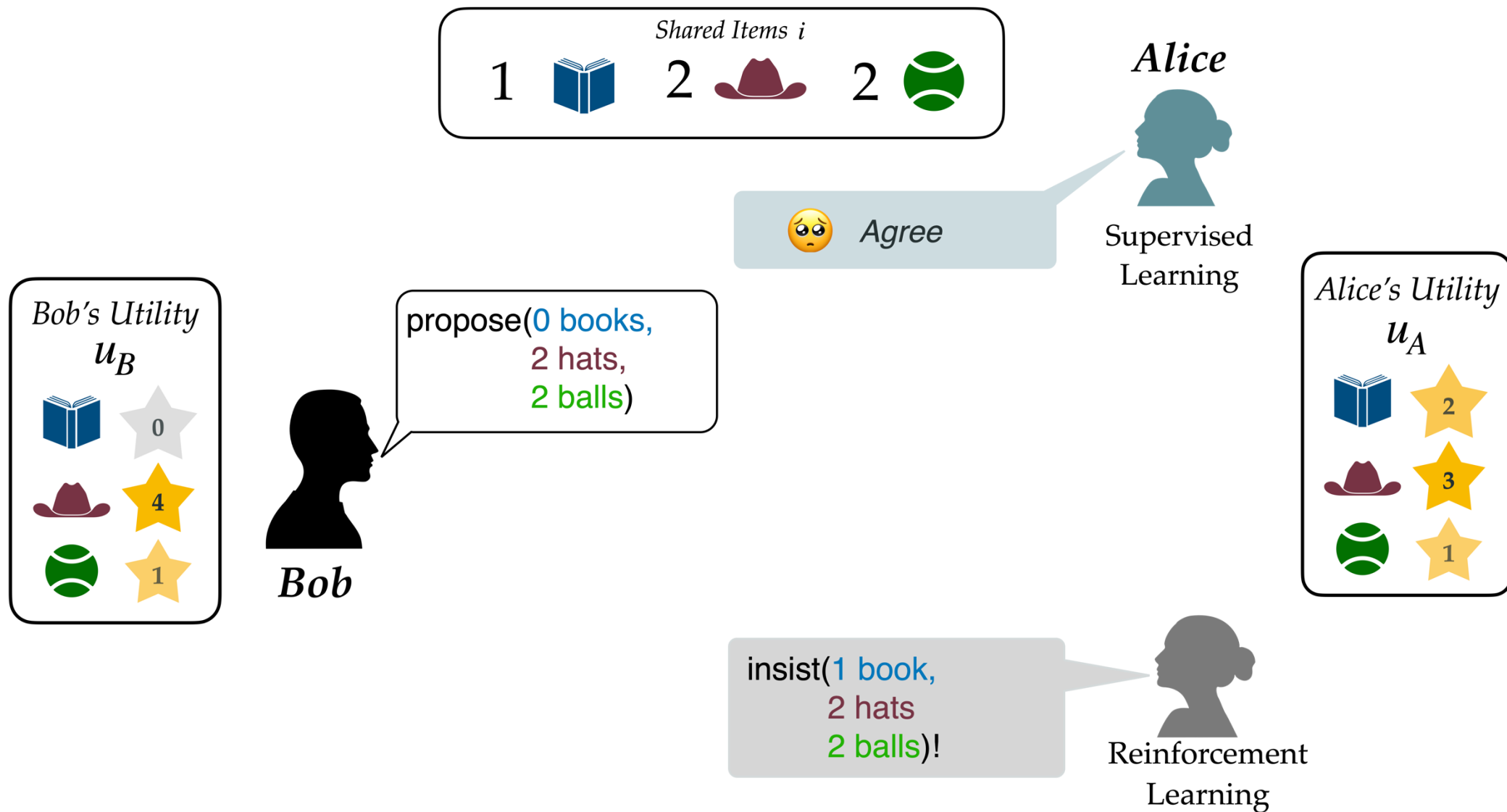
propose(0 books,
2 hats,
2 balls)



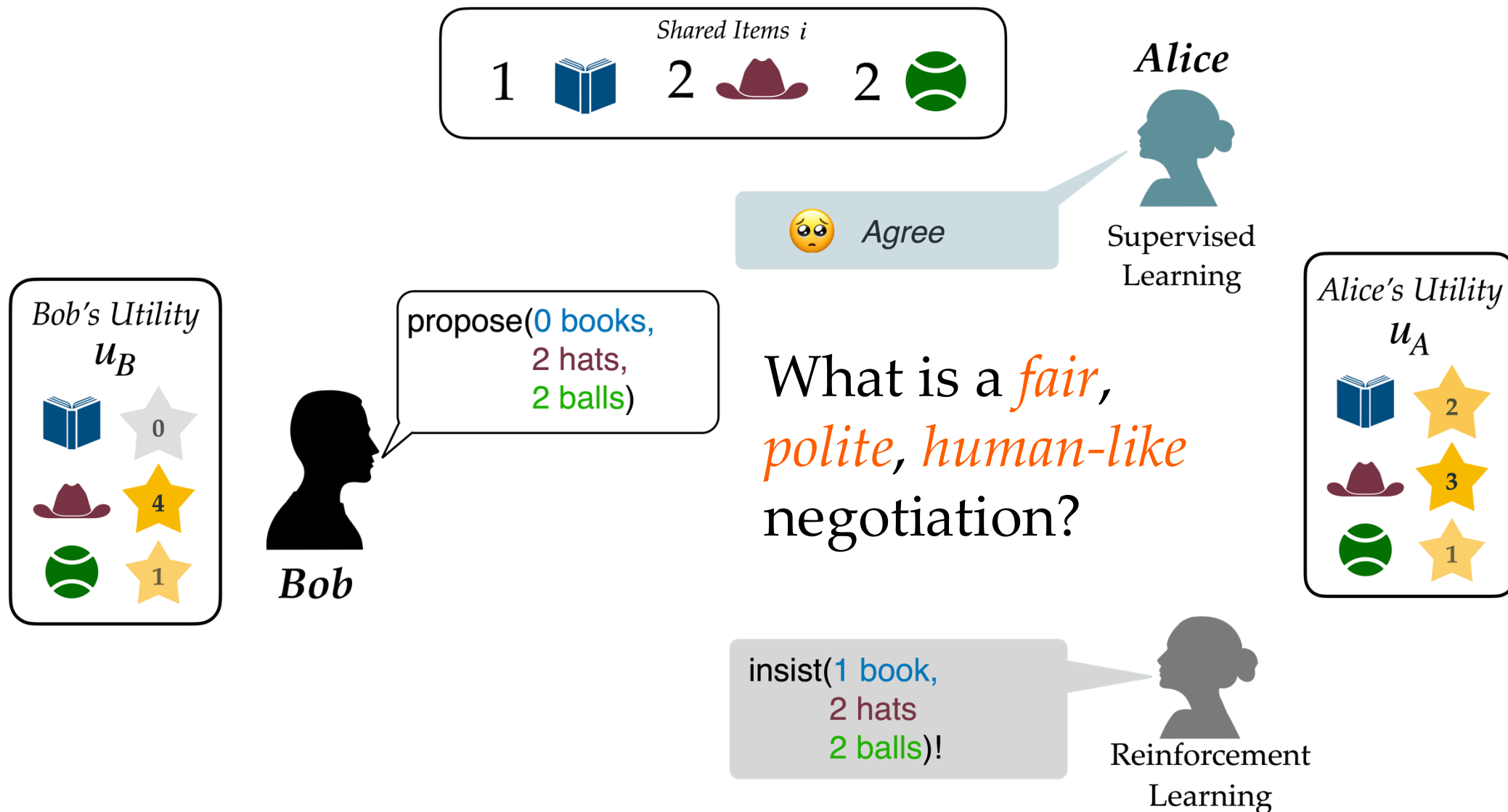
Negotiation Domain



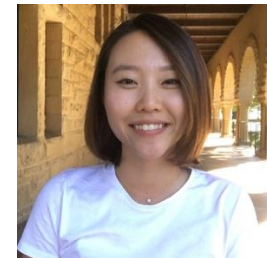
Negotiation Domain



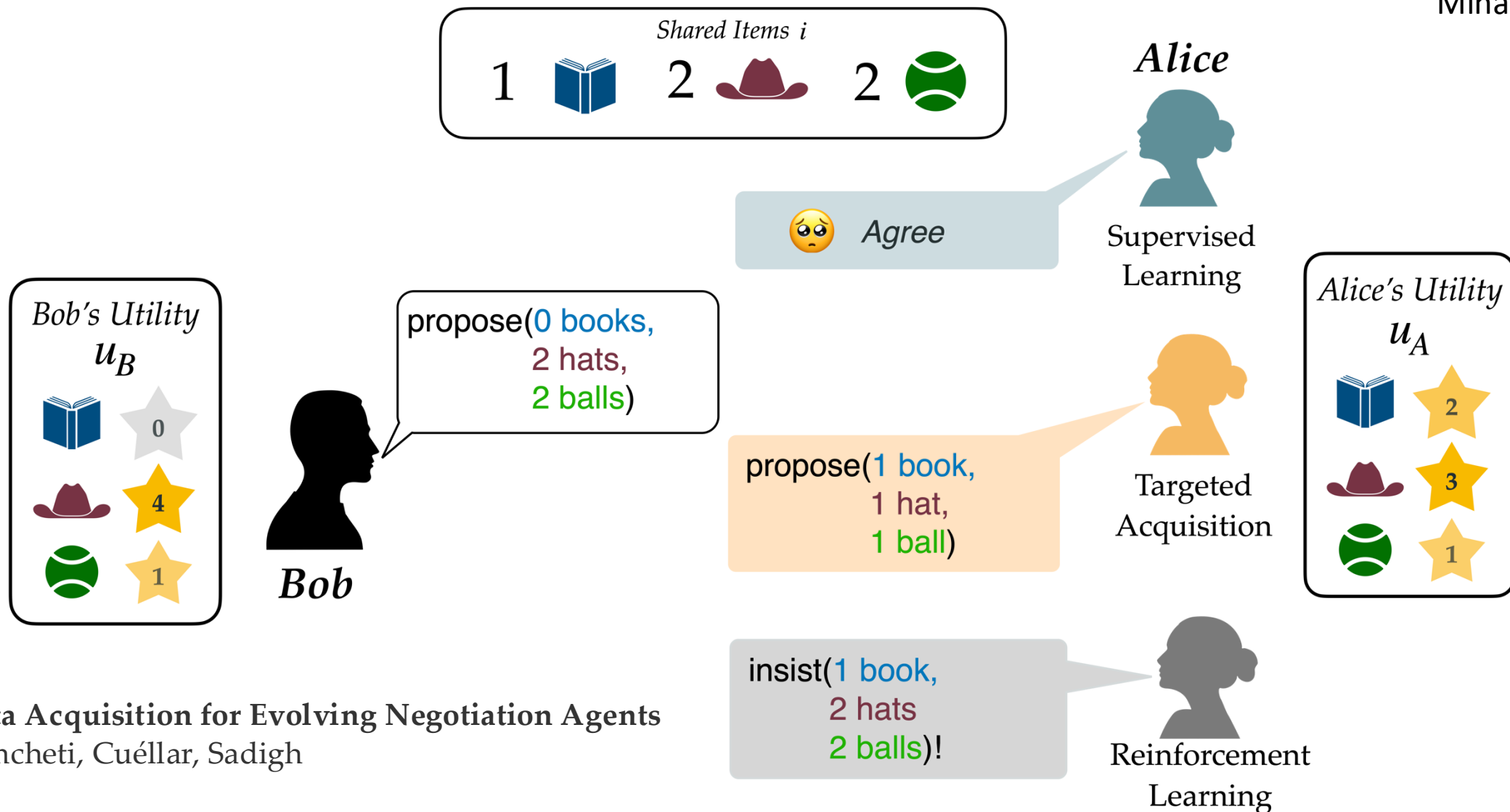
Negotiation Domain



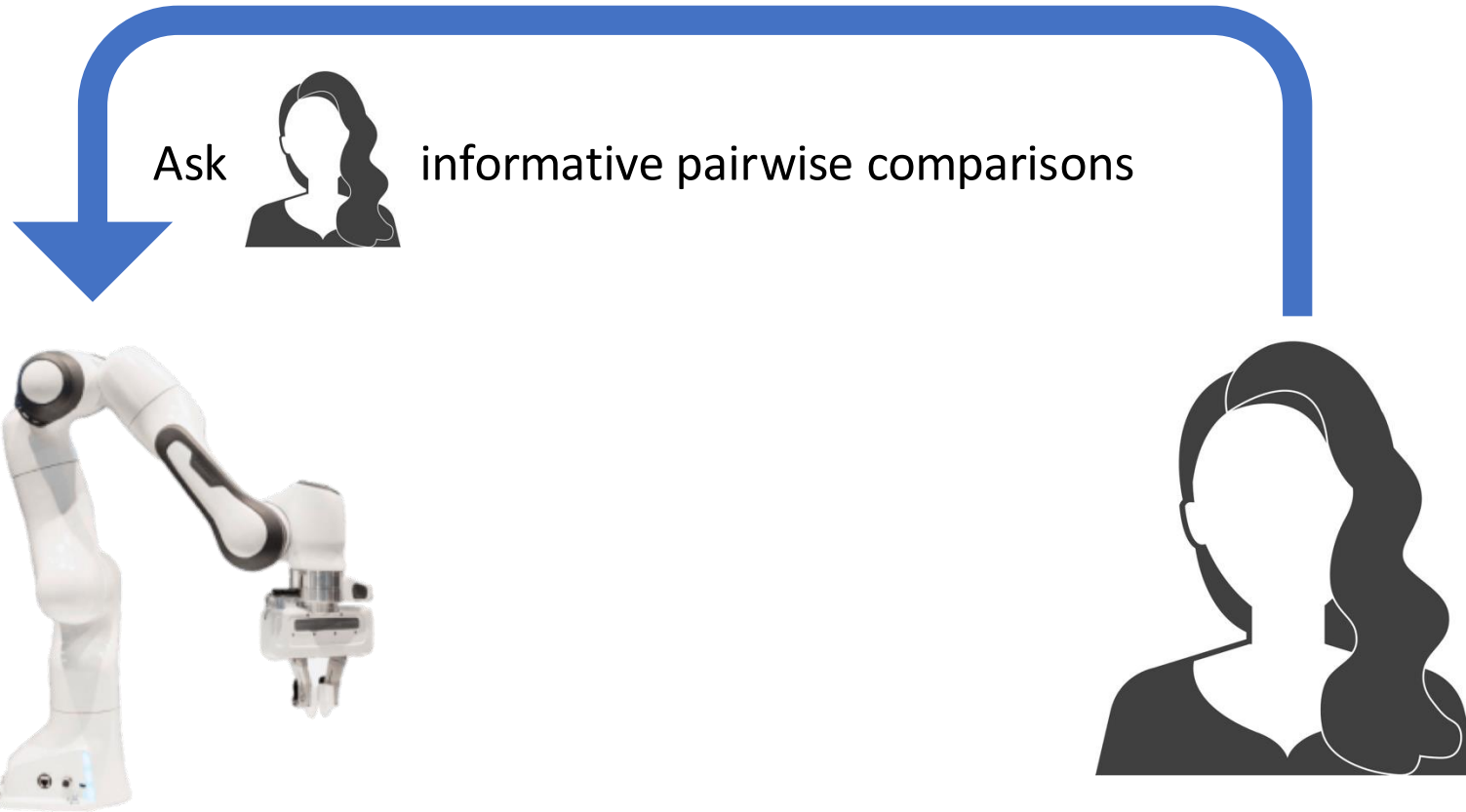
Negotiation Domain



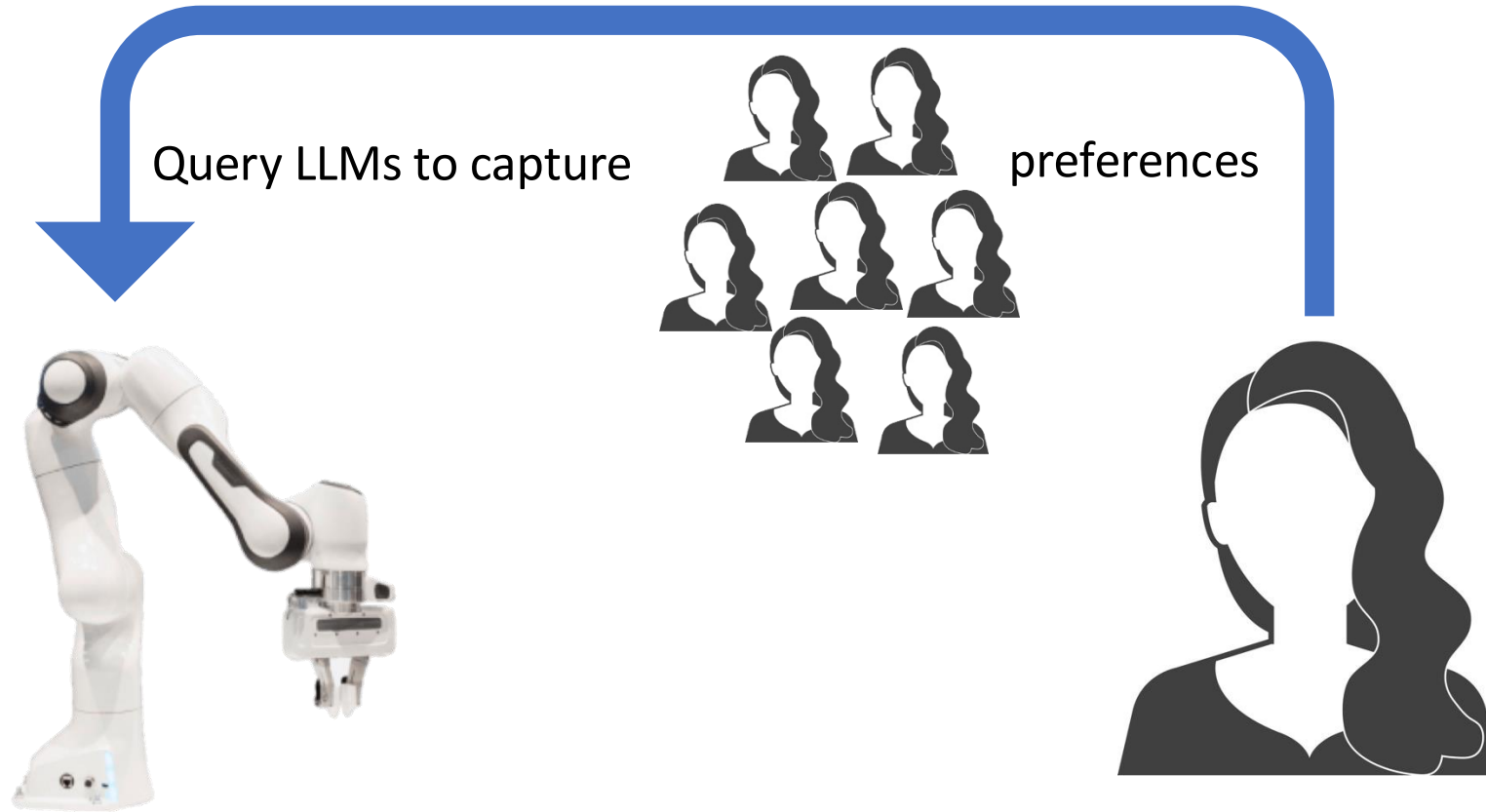
Minae Kwon



Learn Human Preferences



Learn Human Preferences



We use LLMs as a proxy reward function
to train RL agents from user inputs

Task description (ρ_1)

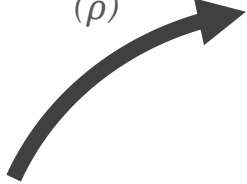
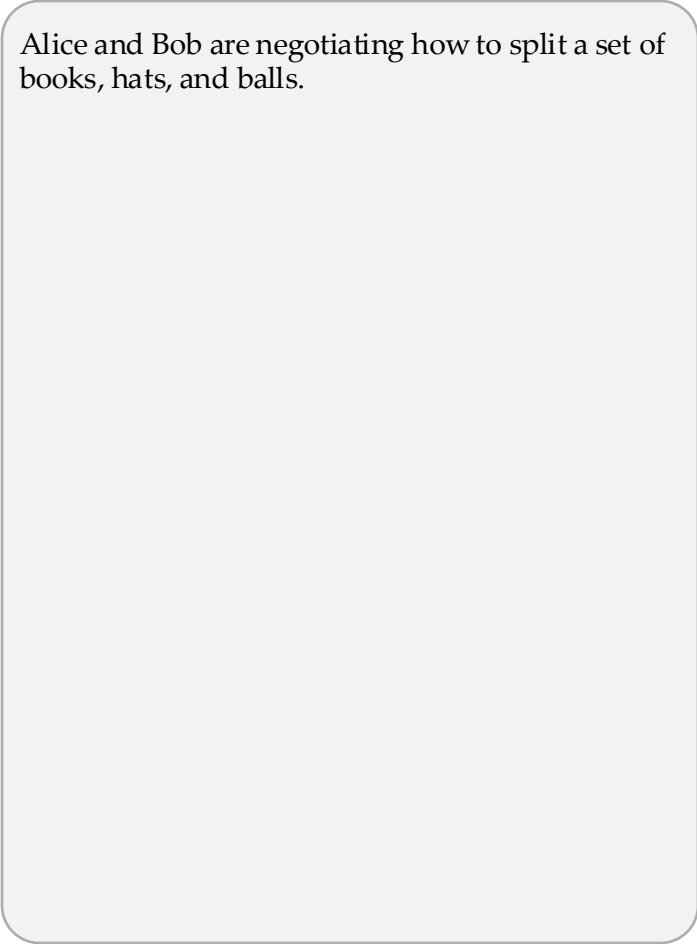
Prompt (ρ)

Alice and Bob are negotiating how to split a set of books, hats, and balls.

(1)
Feed prompt
(ρ)

LLM

Construct
prompt (ρ)



Task description (ρ_1)



Example from user describing
objective (versatile behavior)
(ρ_2)

Prompt (ρ)

Alice and Bob are negotiating how to split a set of books, hats, and balls.

Alice : propose: book=1 hat=1 ball=0

Bob : propose: book=0 hat=1 ball=0

Alice : propose: book=1 hat=0 ball=1

Agreement!

Alice : 4 points

Bob : 5 points

Is Alice a versatile negotiator?

Yes, because she suggested different proposals.

(1)
Feed prompt
(ρ)

LLM

Construct
prompt (ρ)

Task description (ρ_1)



Example from user describing objective (versatile behavior) (ρ_2)



Episode outcome described as string using parse f (ρ_3)

Prompt (ρ)

Alice and Bob are negotiating how to split a set of books, hats, and balls.

Alice : propose: book=1 hat=1 ball=0
Bob : propose: book=0 hat=1 ball=0
Alice : propose: book=1 hat=0 ball=1

Agreement!
Alice : 4 points
Bob : 5 points

Is Alice a versatile negotiator?
Yes, because she suggested different proposals.

Alice : propose: book=1 hat=1 ball=0
Bob : propose: book=0 hat=1 ball=0
Alice : propose: book=1 hat=1 ball=0

Agreement!
Alice : 5 points
Bob : 5 points

(1)
Feed prompt (ρ)

LLM

Construct prompt (ρ)

Task description (ρ_1)



Example from user describing objective (versatile behavior) (ρ_2)



Episode outcome described as string using parse f (ρ_3)

Question (ρ_4)

Prompt (ρ)

Alice and Bob are negotiating how to split a set of books, hats, and balls.

Alice : propose: book=1 hat=1 ball=0
Bob : propose: book=0 hat=1 ball=0
Alice : propose: book=1 hat=0 ball=1

Agreement!
Alice : 4 points
Bob : 5 points

Is Alice a versatile negotiator?
Yes, because she suggested different proposals.

Alice : propose: book=1 hat=1 ball=0
Bob : propose: book=0 hat=1 ball=0
Alice : propose: book=1 hat=1 ball=0

Agreement!
Alice : 5 points
Bob : 5 points

Is Alice a versatile negotiator?

(1)
Feed prompt (ρ)

LLM

Construct prompt (ρ)

Task description (ρ_1)



Example from user describing objective (versatile behavior) (ρ_2)



Episode outcome described as string using parse f (ρ_3)

Question (ρ_4)

Prompt (ρ)

Alice and Bob are negotiating how to split a set of books, hats, and balls.

Alice : propose: book=1 hat=1 ball=0
Bob : propose: book=0 hat=1 ball=0
Alice : propose: book=1 hat=0 ball=1

Agreement!
Alice : 4 points
Bob : 5 points

Is Alice a versatile negotiator?
Yes, because she suggested different proposals.

Alice : propose: book=1 hat=1 ball=0
Bob : propose: book=0 hat=1 ball=0
Alice : propose: book=1 hat=1 ball=0

Agreement!
Alice : 5 points
Bob : 5 points

Is Alice a versatile negotiator?

(1)
Feed prompt (ρ)

LLM

Construct prompt (ρ)

Prompt (ρ)

Task description (ρ_1)



Example from user describing objective (versatile behavior) (ρ_2)



Episode outcome described as string using parse f (ρ_3)

Question (ρ_4)

Alice and Bob are negotiating how to split a set of books, hats, and balls.

Alice : propose: book=1 hat=1 ball=0
 Bob : propose: book=0 hat=1 ball=0
 Alice : propose: book=1 hat=0 ball=1

Agreement!
 Alice : 4 points
 Bob : 5 points

Is Alice a versatile negotiator?
 Yes, because she suggested different proposals.

Alice : propose: book=1 hat=1 ball=0
 Bob : propose: book=0 hat=1 ball=0
 Alice : propose: book=1 hat=1 ball=0

Agreement!
 Alice : 5 points
 Bob : 5 points

Is Alice a versatile negotiator?

Construct prompt (ρ)

(1) Feed prompt (ρ)

LLM

(2) LLM provides textual output

"No"

Prompt (ρ)

Task description (ρ_1)



Example from user describing objective (versatile behavior) (ρ_2)



Episode outcome described as string using parse f (ρ_3)

Question (ρ_4)

Alice and Bob are negotiating how to split a set of books, hats, and balls.

Alice : propose: book=1 hat=1 ball=0
 Bob : propose: book=0 hat=1 ball=0
 Alice : propose: book=1 hat=0 ball=1

Agreement!
 Alice : 4 points
 Bob : 5 points

Is Alice a versatile negotiator?
 Yes, because she suggested different proposals.

Alice : propose: book=1 hat=1 ball=0
 Bob : propose: book=0 hat=1 ball=0
 Alice : propose: book=1 hat=1 ball=0

Agreement!
 Alice : 5 points
 Bob : 5 points

Is Alice a versatile negotiator?

Construct prompt (ρ)

(1) Feed prompt (ρ)

LLM

(2) LLM provides textual output

"No"



(3) Convert to int "0" using parse g and use as reward signal

Prompt (ρ)

Task description (ρ_1)



Example from user describing objective (versatile behavior) (ρ_2)



Episode outcome described as string using parse f (ρ_3)

Question (ρ_4)

Alice and Bob are negotiating how to split a set of books, hats, and balls.

Alice : propose: book=1 hat=1 ball=0
Bob : propose: book=0 hat=1 ball=0
Alice : propose: book=1 hat=0 ball=1

Agreement!
Alice : 4 points
Bob : 5 points

Is Alice a versatile negotiator?
Yes, because she suggested different proposals.

Alice : propose: book=1 hat=1 ball=0
Bob : propose: book=0 hat=1 ball=0
Alice : propose: book=1 hat=1 ball=0

Agreement!
Alice : 5 points
Bob : 5 points

Is Alice a versatile negotiator?

(1) Feed prompt (ρ)

LLM

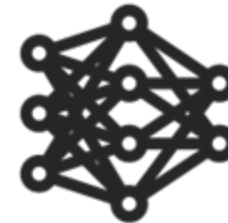
(2) LLM provides textual output

"No"

(3) Convert to int "0" using parse g and use as reward signal

(4) Update agent (Alice) weights and run an episode

Construct prompt (ρ)



Prompt (ρ)

Task description (ρ_1)



Example from user describing objective (versatile behavior) (ρ_2)



Episode outcome described as string using parse f (ρ_3)

Question (ρ_4)

Alice and Bob are negotiating how to split a set of books, hats, and balls.

Alice : propose: book=1 hat=1 ball=0
Bob : propose: book=0 hat=1 ball=0
Alice : propose: book=1 hat=0 ball=1

Agreement!
Alice : 4 points
Bob : 5 points

Is Alice a versatile negotiator?
Yes, because she suggested different proposals.

Alice : propose: book=1 hat=1 ball=0
Bob : propose: book=0 hat=1 ball=0
Alice : propose: book=1 hat=1 ball=0

Agreement!
Alice : 5 points
Bob : 5 points

Is Alice a versatile negotiator?

(1) Feed prompt (ρ)

LLM

(2) LLM provides textual output

"No"

(3) Convert to int "0" using parse g and use as reward signal

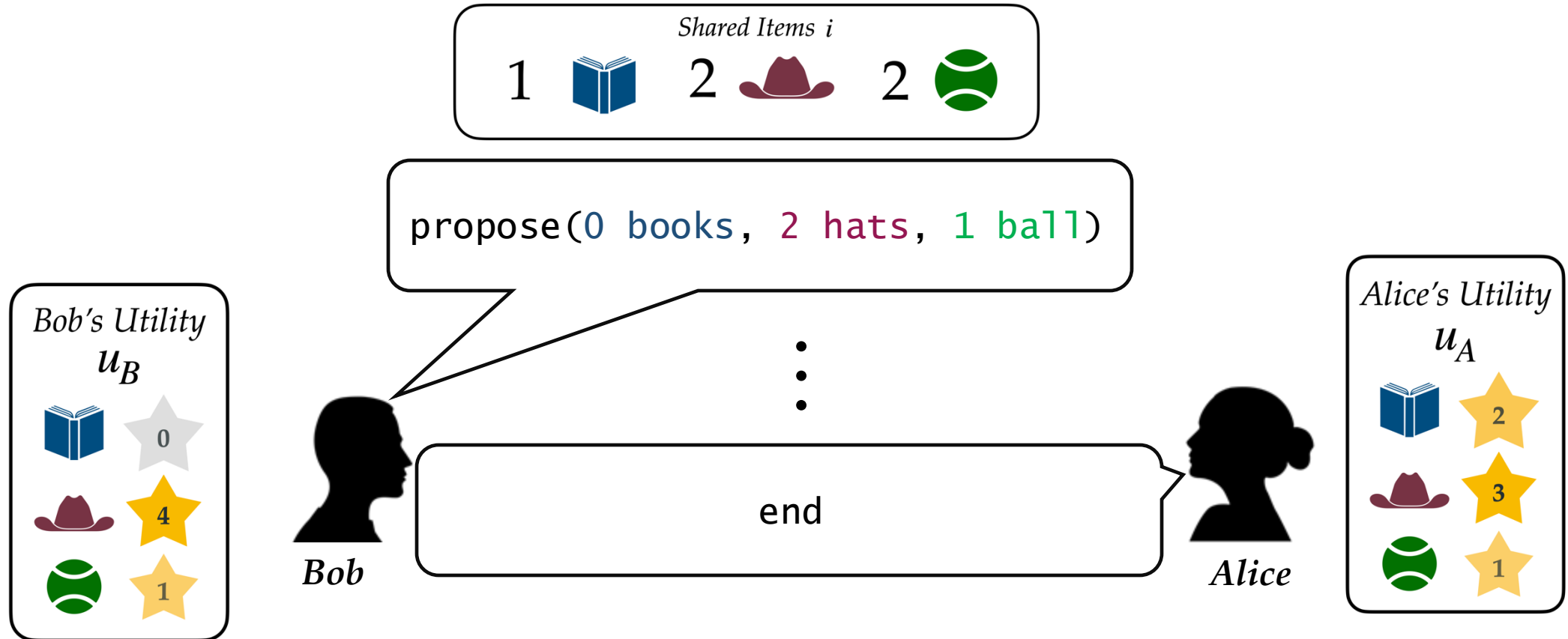


(4) Update agent (Alice) weights and run an episode

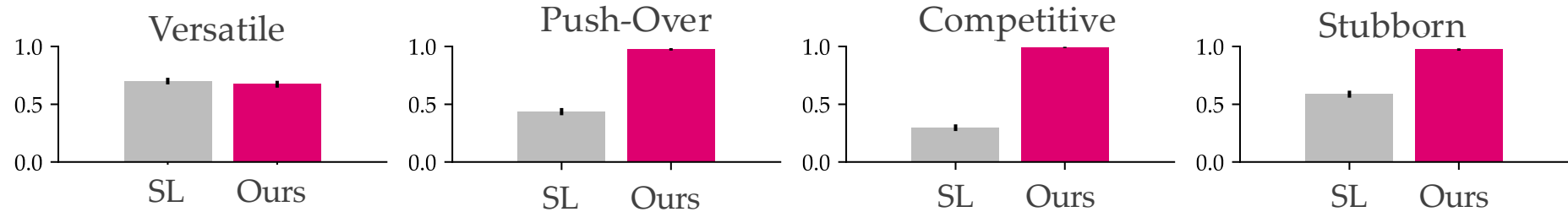
Construct prompt (ρ)

(5) Summarize episode outcome as string (ρ_3) using parser f

DEALORNO DEAL Negotiation Task



Labeling Accuracy



RL Agent Accuracy

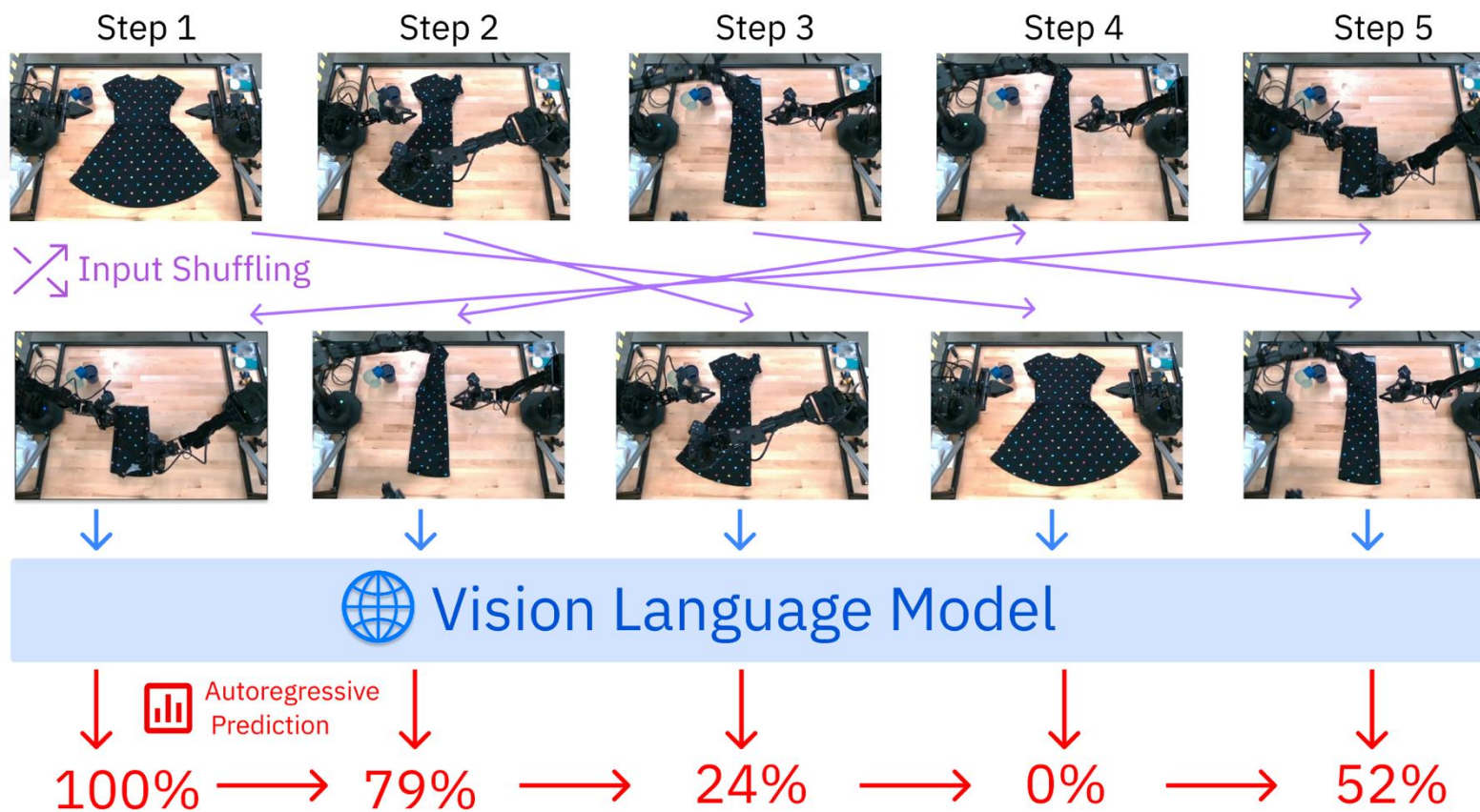


We outperform SL by avg. of 46%

We underperform True Reward by avg. of 4%

We can use an LLM as a proxy reward to train objective-aligned agents

Generative Value Learning



Key Takeaways

We can learn representations (**reward functions**) by

- 1) pretraining and **actively** querying for informative human feedback
- 2) leveraging the knowledge of **large language models**.