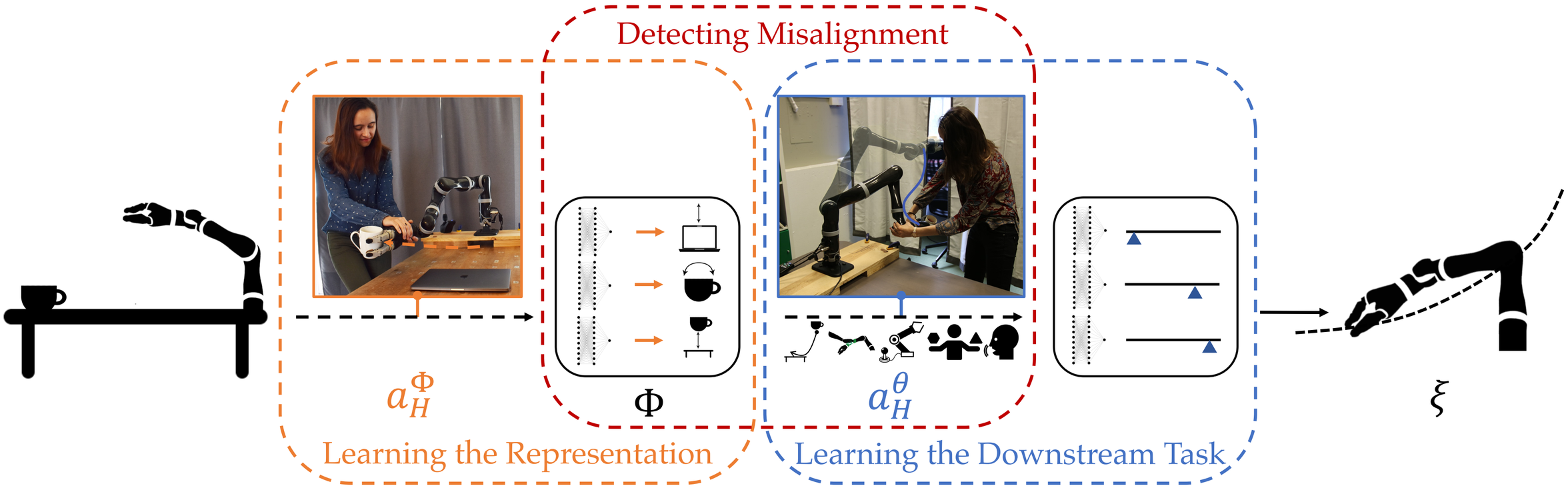
Aligning Robot and Human Representations

Andreea Bobu

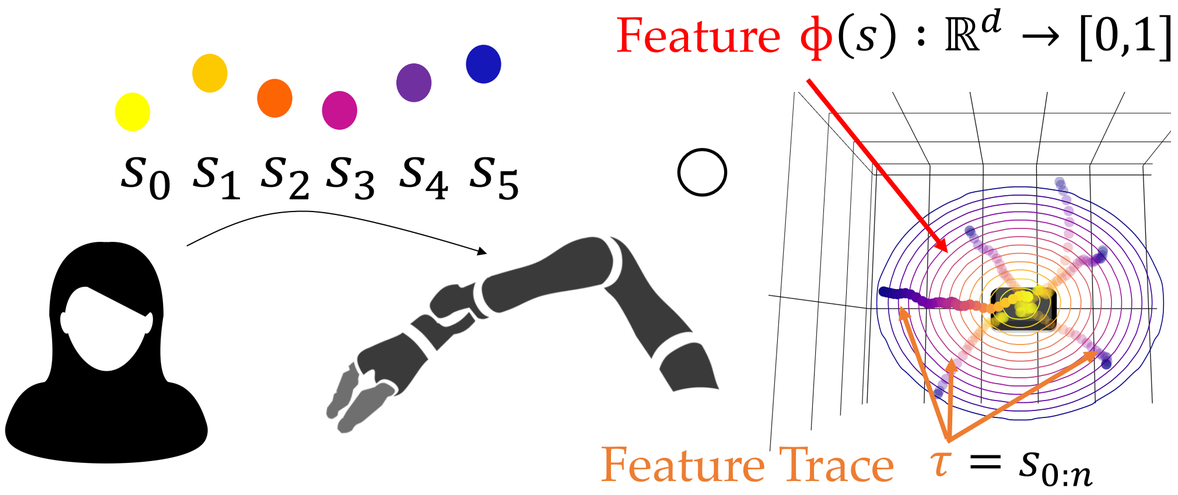
InterACT Laboratory · Berkeley UNIVERSITY OF CALIFORNIA

Problem Statement: How can the robot align its representation with the human in order to better interpret their task input?

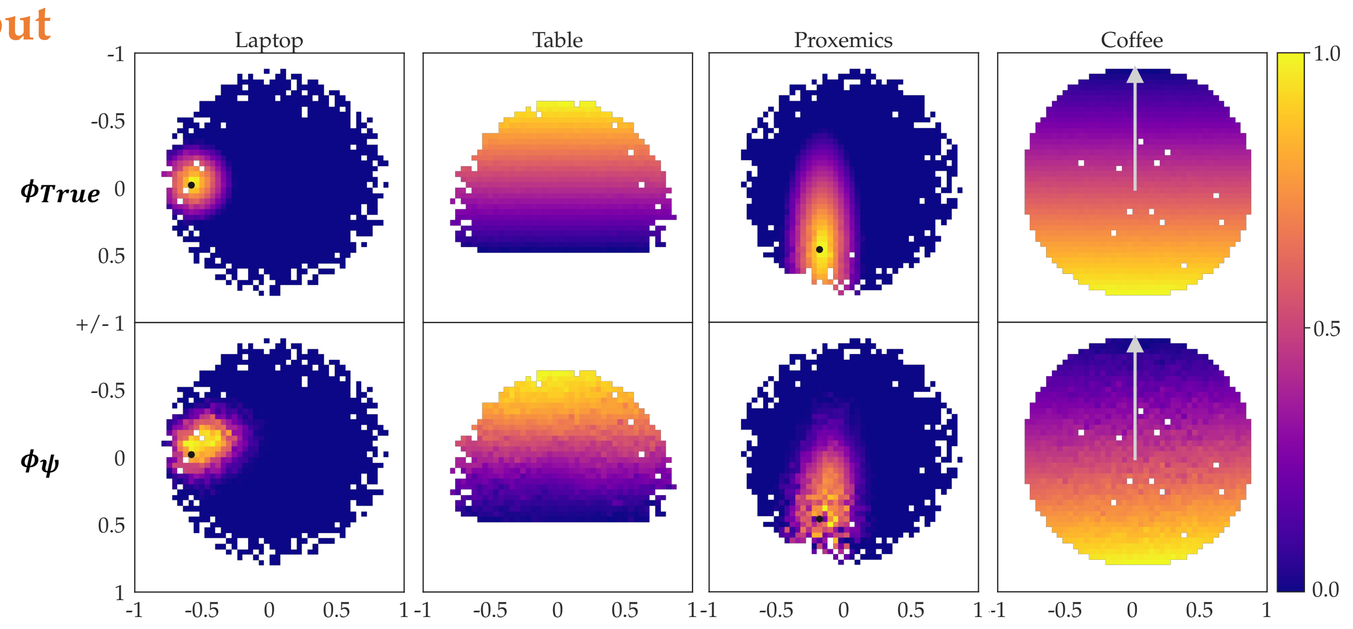


Key Idea: Focus explicitly on learning the representation from human data before using it for learning the downstream task.

Learning a Feature with Representation-Specific Input



Train ϕ_ψ $s.t. \forall i < j, k: \phi_\psi(s_i) \geq \phi_\psi(s_j)$



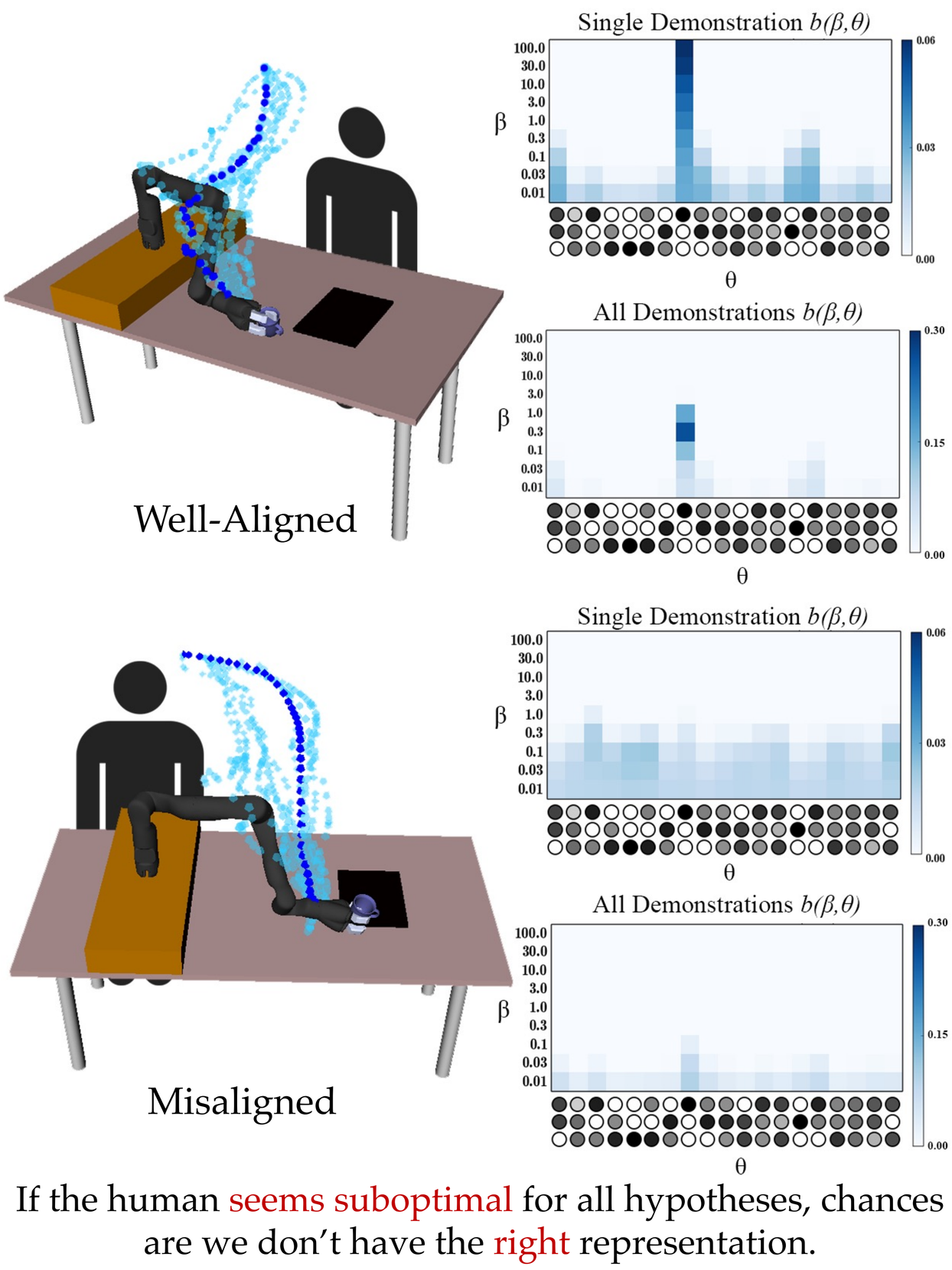
Detecting Misalignment with Situational Confidence

Demonstration

$$P(\xi_H \mid \beta, \theta) = \frac{e^{-\beta C_\theta(\xi_H)}}{\int e^{-\beta C_\theta(\bar{\xi}_H)} d\bar{\xi}_H}$$

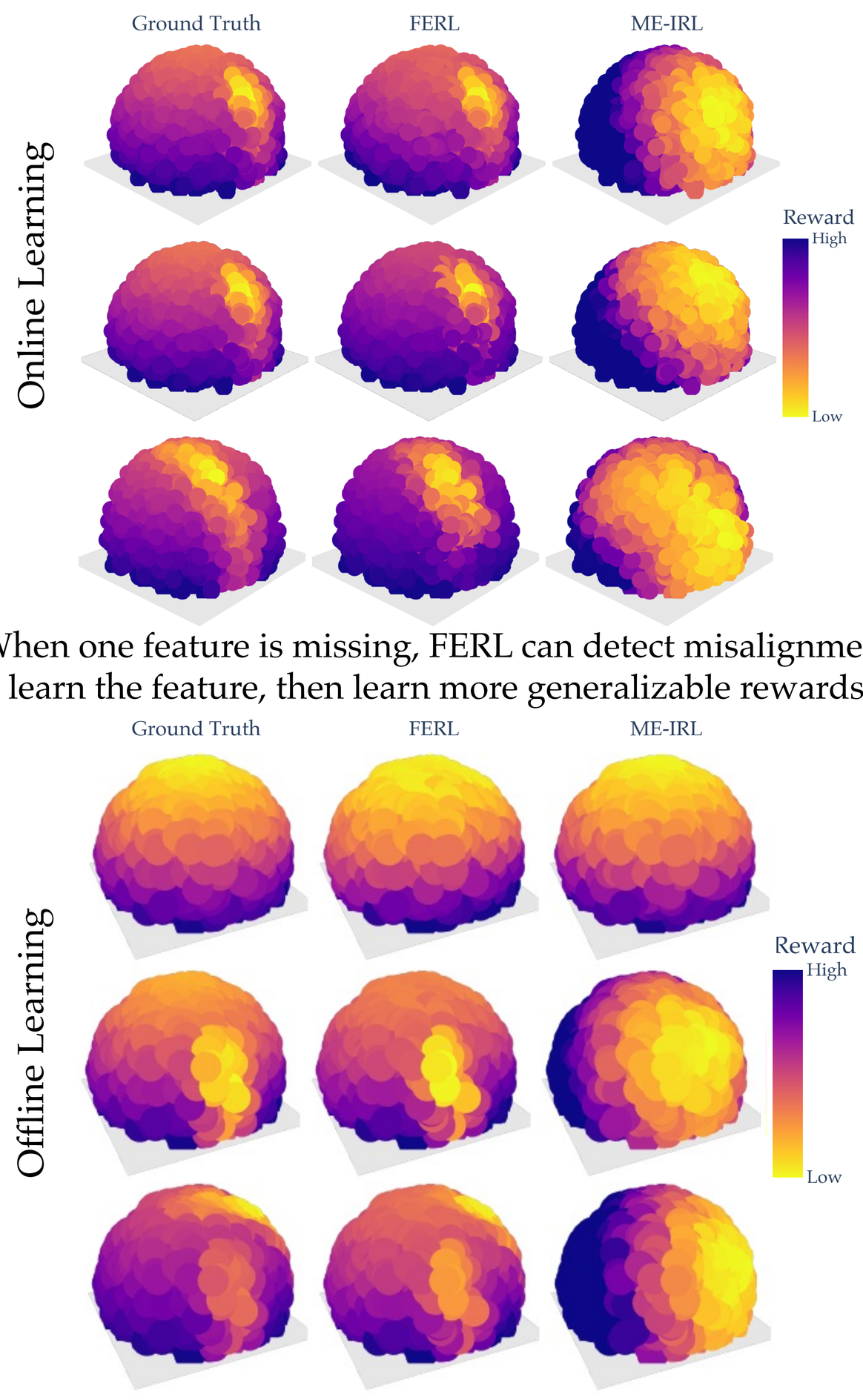
Confidence

$$b'(\beta, \theta) = \frac{P(\xi_H \mid \beta, \theta) b(\beta, \theta)}{\int P(\xi_H \mid \bar{\beta}, \bar{\theta}) b(\bar{\beta}, \bar{\theta}) d\bar{\theta} d\bar{\beta}}$$



If the human seems suboptimal for all hypotheses, chances are we don't have the right representation.

Feature Expansive Reward Learning (FERL) Generalizes Better than Deep End-to-End IRL



When one feature is missing, FERL can detect misalignment, learn the feature, then learn more generalizable rewards.

FERL learns the representation one feature at a time, leading to more generalizable rewards than deep end-to-end IRL.