Learner and Teacher Perspectives on Learning Rewards from Multiple Types of Human Feedback

Ali Larian
University of Utah
Salt Lake City, UT, USA
ali.larian@utah.edu

Atharv Belsare
University of Utah
Salt Lake City, UT, USA
atharv.belsare@utah.edu

Zifan Wu
University of Utah
Salt Lake City, UT, USA
zifan.wu@utah.edu

Daniel S. Brown
University of Utah
Salt Lake City, UT, USA
daniel.s.brown@utah.edu

***Abstract*—In safety-critical human-robot interaction domains, such as autonomous driving and assistive robotics, it is important for robots to autonomously assess whether they have received sufficient human feedback to learn reliable policies. Simultaneously, human teachers need to understand the informativeness of their feedback to provide effective guidance. This paper presents a unified framework that integrates both the robot's self-assessment and the human teacher's feedback informativeness, incorporating diverse feedback types, including demonstrations, comparisons, E-stops, and corrections. We extend demonstration sufficiency evaluation—determining whether enough feedback has been received to ensure reliable policy learning—to these modalities, empirically analyzing their impact on the robot's learning progress and policy confidence. Additionally, we compare entropy-based and regret-based stopping criteria for determining feedback sufficiency, finding that the regret-based approach offers more reliable performance across feedback types. Through empirical and theoretical analysis, we compare the theoretical representational power and empirical effectiveness of different feedback types. Our findings enable robots to better self-assess their performance and empower humans to deliver more informative feedback, enhancing mutual understanding in human-robot collaboration. Through theoretical analysis and experiments, our work provides a novel foundational understanding of feedback-driven reward learning.**

I. Introduction

In human-robot interaction, a robot's ability to autonomously assess its performance is crucial for safe and efficient collaboration, particularly in safety-critical domains like autonomous driving and assistive robotics. Such self-assessment enables robots to determine whether they have received sufficient human feedback to learn a reliable policy or if additional input is needed, reducing reliance on costly or infeasible human supervision. Simultaneously, from the human teacher's perspective, understanding the informativeness of their feedback—such as demonstrations, pairwise preferences, E-stops, or corrections—can enhance their ability to provide effective guidance and improve mutual. Despite the importance of these dual perspectives, prior work has rarely addressed both the learner's (robot's) self-assessment and the teacher's (human's) feedback informativeness within a unified framework that leverages diverse feedback types.

From the learner's perspective, prior work utilizes Bayesian Inverse Reinforcement Learning (BIRL) to tackle robot self-assessment by maintaining a belief distribution over the human's unobserved reward function and evaluating policy regret—the performance gap between learned and optimal policies [32]. However, prior work is limited to demonstrations and does not account for other feedback modalities, such as preferences, E-stops, or corrections, which humans naturally provide in real-world interactions [16, 25]. This limitation hinders comprehensive self-assessment, especially for complex tasks where demonstrations are challenging to provide due to task intricacy or equipment constraints [33]. From the learner's perspective, prior work in active learning has also explored how robots can actively elicit feedback to enhance learning [4]. Bıyık et al. [5] used information gain in active learning to select preference queries and halt querying. We adapt Bıyık et al. [5]'s active learning stopping criterion, which captures reward function uncertainty via information gain, for our passive learning self-assessment problem, and investigate whether monitoring reward function uncertainty reduction via information gain is a good indicator for feedback sufficiency.

From the teacher's perspective, understanding the informativeness of feedback modalities—such as E-stops, corrections, demonstrations, and pairwise preferences—is critical for providing AI systems with effective human guidance. Prior work has explored reward ambiguity through human preference modeling [19, 30] and demonstrated that ranked demonstrations reduce ambiguity more effectively than unranked ones [8]. However, analyzing diverse feedback modalities' effectiveness and theoretical representation power in reducing reward ambiguity remains unexplored. Our work addresses this gap by empirically and theoretically analyzing e-stops, corrections, demonstrations, and pairwise preferences, finding that pairwise preferences are the most effective in reducing reward ambiguity, followed by corrections, then demonstrations, with E-stops being the least effective. Surprisingly, demonstrations lead to low representational power, yet they are highly effective due to the number of implicit comparisons they produce. These findings enable human teachers to provide more informative feedback to AI systems, enhancing efficiency and mutual understanding.

In summary, we make the following contributions: (1) **Unified evaluation of diverse feedback modalities:** We are the first to study self-assessment sufficiency by extending demonstration sufficiency to incorporate diverse human feedback types, including demonstration, pairwise comparisons, E-stops, and corrections. (2) **Comparison of stopping criteria for**

feedback sufficiency: We compare two stopping criteria for feedback sufficiency: an entropy-based approach that leverages normalized information gain to assess reward function uncertainty and a regret-based approach. The regret-based criterion offers more reliable and practical stopping conditions across diverse feedback types. (3) **Empirical and theoretical analysis of reducing reward ambiguity:** We assess the effectiveness of pairwise preferences, corrections, demonstrations, and E-stops in reducing reward ambiguity, measured as the volume of the feasible reward region, and provide the first categorization of feedback types based on their theoretical reward representation and empirical performance. (4) **Venn Diagram representation:** a Venn Diagram visualizing the subset relationships among these feedback types as pairwise comparisons.

II. Related Work

Our work explores reward learning from diverse human feedback types—including demonstrations, pairwise preferences, E-stops, and corrections—considering both the learner's (robot's) and the teacher's (human's) perspectives. Notably, none of the related works below comprehensively addresses these viewpoints within the context of diverse feedback types.

A. Learner's Perspective: Diverse Feedback Learning and Autonomous Assessment

Prior work explores reward learning from diverse human feedback in reinforcement learning including rankings[6, 27], pairwise preferences[34, 11], and corrections [24, 33]. Mehta and Losey [25] combines demonstrations, corrections, and preferences. Ghosal et al. [13] models human rationality via a rationality coefficient. Ibarz et al. [15] and Bıyık et al. [5] learn rewards from pairwise preferences and demonstrations. Jeon et al. [16] and Metz et al. [26] propose unified approaches for reward learning from different human feedback types, but differ from our work which focuses on learner self-assessment and an analysis of teacher feedback informativeness.

There has been relatively little research on self-assessment when an AI system employs reward learning. Norton et al. [28], Burghouts et al. [9] focus on communicating uncertainty to humans, not integrating feedback for learning. Koenig et al. [20], Hayes et al. [14] emphasize learning from demonstrations, neglecting continuous self-assessment. Trinh et al. [32] use BIRL to assess demonstration sufficiency, omitting other feedback types. By contrast, our framework integrates pairwise preferences, corrections, demonstrations, and E-stops into autonomous self-assessment. A large body of research has focused on active learning, where the learner actively queries humans for specific feedback to obtain the most informative input [1, 3, 4, 2, 12]. These approaches often use information-theoretic objectives—such as maximizing expected information gain—to guide both query selection and stopping decisions. For example, Bıyık et al. [5] proposes selecting preference queries that maximize mutual information between human responses and the reward function, coupled with an optimal stopping rule based on information gain. In contrast, our work studies feedback sufficiency in a passive learning setting, where the agent does not actively generate queries. We provide the first comparison between information gain and regret as stopping criteria for determining feedback sufficiency across different feedback types.

B. Teacher's Perspective: Reward Ambiguity and Pedagogic Teaching

From the teacher's perspective, understanding reward ambiguity is critical for providing feedback that effectively guides the robot. Recent studies[19, 30] provide insights into reward ambiguity through human preference modeling and policy optimization invariance, respectively, but they do not examine how different feedback types vary in reducing this ambiguity and a comprehensive framework for analyzing various feedback modalities is lacking. This gap complicates assessing the informativeness of feedback types for reward learning. Our work addresses this by comparing demonstration, pairwise preferences, corrections, and E-stops. We find that pairwise preferences most effectively reduce reward ambiguity, constraining the feasible reward parameter space, followed by corrections, demonstrations, and E-stops. These findings extend the theoretical contributions of Knox et al. [19], Skalse et al. [30] to a broader set of feedback modalities. Beyond resolving reward ambiguity, the teacher's perspective also involves designing pedagogically informative feedback. Prior work has focused on the teacher's role in crafting maximally informative demonstrations, often assuming the teacher knows the learner's algorithm and target reward function [7, 10, 18, 35, 21]. However, none of these studies addresses both the learner's and teacher's perspectives in the context of diverse feedback types. Our work fills this gap by integrating both viewpoints, enabling robots to self-assess their performance and analyze reward ambiguity from diverse feedback.

III. Preliminaries

A. Markov Decision Processes

We model the environment as an MDP with states $\mathcal{S}$, actions $\mathcal{A}$, transition function $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$, reward function $r : \mathcal{S} \to \mathbb{R}$, initial state distribution μ, and discount factor $\gamma \in [0, 1)$. A policy π maps states to action distributions. The expected return of π under r is $V_r^\pi = \mathbb{E}_{s \sim \mu} V_r^\pi(s)$, where $V_r^\pi(s) = \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r(s_t) \mid s_0 = s]$ and the Q-value function is $Q_R^\pi(s, a) = r(s) + \gamma \sum_{s'} T(s, a, s') V_r^\pi(s')$. Following prior work [5, 13, 32, 7, 19], we assume r is a linear function of features, $r(s) = w^T \phi(s)$, where $\phi(s) \in \mathbb{R}^k$ represents state features. We define Ξ as the trajectory class, consisting of all possible finite trajectories $\xi = (s_0, s_1, \ldots, s_T)$ with $s_0 \sim \mu$, where for each $t = 0, \ldots, T-1$, there exists an action $a_t \in \mathcal{A}$ such that $s_{t+1} \sim T(s_t, a_t, \cdot)$, and $T \leq T_{\max}$ for some maximum trajectory length $T_{\max}$.

B. Human Feedback Types

We model feedback using the reward-rational choice framework [16], where each instance is a choice $c \in \mathcal{C}$ mapped to trajectories in Ξ via a grounding function $\psi : \mathcal{C} \to \Xi$, constraining plausible reward functions.

Demonstrations: Choices of state-action pairs, with $\mathcal{C} = \mathcal{S} \times \mathcal{A}$, $\psi(s,a) = (s,a)$. Likelihood:

$$P(\xi \mid r, \beta) = \prod_{(s_t,a_t)\in\xi} \frac{\exp(\beta\, Q(s_t, a_t \mid r))}{\sum_{b\in\mathcal{A}} \exp(\beta\, Q(s_t, b \mid r))}. \quad (1)$$

Comparisons: Choices between trajectories $\xi_A, \xi_B \in \Xi$, with $\mathcal{C} = \{\xi_A, \xi_B\}$, $\psi(\xi_i) = \xi_i$. Preference probability:

$$P(\xi_A \mid r, \beta) = \frac{\exp(\beta\, r(\xi_A))}{\exp(\beta\, r(\xi_A)) + \exp(\beta\, r(\xi_B))}. \quad (2)$$

E-stop: Binary choice to halt (off) or continue a trajectory ξ_R at time t, with $\mathcal{C} = \{\text{off}, -\}$, $\psi(\text{off}) = \xi_{\text{halted}}$, $\psi(-) = \xi_R$, where $\xi_{\text{halted}} = \xi_R^{0:t}\xi_R^t \ldots \xi_R^t$. Likelihood:

$$P(\text{off} \mid r, \beta) = \frac{\exp(\beta\, r(\xi_{\text{halted}}))}{\exp(\beta\, r(\xi_{\text{halted}})) + \exp(\beta\, r(\xi_R))}. \quad (3)$$

Corrections: Comparison of a corrected trajectory $\xi_{\text{corrected}}$ to the robot's ξ_R, with $\mathcal{C} = \{\xi_R, \xi_{\text{corrected}}\}$, $\psi(\xi) = \xi$. Likelihood:

$$P(\xi_{\text{corrected}} \mid r, \beta) = \frac{\exp(\beta\, r(\xi_{\text{corrected}}))}{\exp(\beta\, r(\xi_{\text{corrected}})) + \exp(\beta\, r(\xi_R))}. \quad (4)$$

C. Value-at-Risk Bounds

Value-at-Risk is a probabilistic measure of worst-case performance [31, 17]. The α-Value-at-Risk (α-VaR) is the α-worst-case value of a random variable Z, defined as

$$\nu_\alpha(Z) = F_Z^{-1}(\alpha) = \inf\{z : F_Z(z) \geq \alpha\} \quad (5)$$

where $F_Z(z) = P(Z \leq z)$ is the cumulative distribution function of Z. The higher the α, the more risk-sensitive the measure. We use α-VaR to estimate regret under an unknown reward function, to provide risk-aware bound on policy performance with tunable risk sensitivity.

IV. Problem Definition

We seek to enable an agent to determine whether the user has received sufficient feedback to learn a policy that aligns with the expert's intended behavior, represented by the unobserved reward function with weights w^*. To address the problem of feedback sufficiency, we formulate this problem in two different ways: (1) Regret-based feedback sufficiency and (2) Entropy-based confidence feedback sufficiency.

A. Regret-Based Feedback Sufficiency

The regret-based approach assesses feedback sufficiency by measuring how close the robot's policy is to the expert's intended policy using normalized expected value difference (nEVD) [32]:

$$nEVD(\pi_{\text{robot}}, w^*) = \frac{V_{w^*}^* - V_{w^*}^{\pi_{\text{robot}}}}{V_{w^*}^* - V_{w^*}^{\pi_{\text{rand}}}}, \quad (6)$$

where π_{rand} is a random policy. Normalized regret allows interpretable thresholding by performance percentage. Since the true reward function weights w^* are unknown, we use Bayesian inference [29] to sample from the posterior $P(\theta \mid \mathcal{H}_{1:i})$, where $\mathcal{H}_{1:i}$ is the feedback history (e.g., demonstrations, preferences, E-stops). These samples estimate regret via α-VaR bound, declaring sufficiency when:

$$P(nEVD(\pi_{\text{robot}}, \theta) \leq \epsilon \mid \mathcal{H}_{1:i}) \geq \alpha, \quad \theta \sim P(\theta \mid \mathcal{H}_{1:i}). \quad (7)$$

B. Entropy-Based Confidence for Feedback Sufficiency

To compare with the regret-based stopping criterion, we assess whether information gain [5], typically used in active learning for preference query selection, can serve as an effective stopping criterion for passive learning self-assessment, as opposed to a regret-based approach [32]. We adopt a confidence-based stopping criterion using entropy to measure uncertainty in the reward posterior. Inspired by speech recognition [22, 23], we define confidence as normalized entropy reduction:

$$F(w \mid \mathcal{H}_{1:i}) = \frac{H_{\max} - H(w \mid \mathcal{H}_{1:i})}{H_{\max} - H_{\min}} \quad (8)$$

where $H(w \mid \mathcal{H}_{1:i})$ is the entropy of the reward posterior after i feedback instances, $H_{\max}$ is the maximum entropy, and $H_{\min}$ is the minimum entropy, estimated via calibration experiments (see Appendix F). Higher confidence (closer to 1) indicates greater certainty. As direct entropy computation is intractable, we approximate it using MCMC posterior samples, estimating marginal likelihood via importance sampling with the harmonic mean estimator:

$$\hat{H}(w \mid \mathcal{H}_{1:i}) \approx -\frac{1}{m_i}\sum_{k=1}^{m_i} \log P(\mathcal{H}_{1:i} \mid w^{(k)}) - \log A_d - \log\left(\frac{1}{m_i}\sum_{k=1}^{m_i} \frac{1}{P(\mathcal{H}_{1:i} \mid w^{(k)})}\right) \quad (9)$$

where $P(\mathcal{H}_{1:i} \mid w^{(k)})$ is the feedback likelihood under sampled reward $w^{(k)}$, m_i is the number of samples, and $A_d = 2\pi^{d/2}/\Gamma(d/2)$ is the d-dimensional unit sphere's surface area, required as reward weights satisfy $\|w\| = 1$, constraining them to the unit sphere. Details are in Appendix A. The agent halts querying when confidence $F(w \mid \mathcal{H}_{1:i})$ exceeds threshold $\tau \in [0, 1]$, indicating sufficient feedback. We evaluate threshold effectiveness across pairwise preferences, corrections, demonstrations, and E-stops.

V. Analysis of Feedback Types from the Teacher's Perspective

To understand the role of feedback type in learning efficiency, we first analyze how demonstrations, E-stops, pairwise comparisons, and correction feedback impact the feasible region of reward parameters. A key aspect of understanding feedback sufficiency is the *feasible region*, which defines the set of reward parameters w that rationalize a given set of human feedback. Formally, we represent this as:

$$H_F = \bigcap_{(\xi \succ \xi') \in F} \left\{ w \mid w^\top(\Phi_\xi - \Phi_{\xi'}) \geq 0 \right\}, \quad (10)$$

where $\Phi(\xi) = \sum_{t=0}^{T} \gamma^t \phi(s_t)$ is the expected discounted sum of state features from $\xi = (s_0, s_1, \ldots, s_T)$, and F is feedback type dependent. As noted previously, comparison, correction, and E-stop feedback all provide explicit preferences over trajectories. By contrast, demonstrations provide an infinite number of implicit preferences of the form $\xi^* \succeq \xi', \forall \xi'$.

We define the *reward ambiguity*, $G(H_F) = \text{Volume}(H_F)$, as the volume of the intersection of half-spaces. Without loss of generality, we assume $\|w\|_2 = 1$ to ensure this volume is bounded. A smaller $G(H_F)$ implies stronger constraints on the reward parameters, reducing ambiguity. Different feedback types influence this reduction to varying degrees: we find that pairwise comparisons constrain the feasible region more than E-stop, while correction feedback imposes even tighter constraints than pairwise comparisons. In the following sections, we present experimental results on reward ambiguity across feedback types, followed by a theoretical analysis comparing their relative informativeness.

A. Comparing Ambiguity of the Learned Reward Leveraging Different Feedback Types

To visualize the informativeness of different feedback types in reducing reward ambiguity, we conducted experiments in a 2×3 grid environment with two features per cell and a single terminal state at the bottom-right cell. We compared four feedback types—demonstrations, pairwise preferences, corrections, and E-stops—by analyzing their impact on the feasible region of the learned reward function, visualized in Figure 1. Details of the environment layout and the methodology for generating these results are provided in Appendix E.

The results show that, as expected, all feedback types include the true reward value within their feasible regions, validating their ability to capture the correct reward function. However, the reward ambiguity $G(H_F)$, measured as the volume of the feasible region, varies significantly across feedback types. Pairwise preferences (Figure 1a) yield the smallest $G(H_F)$, indicating the highest reduction in reward ambiguity due to their total ranking of trajectories. Corrections (Figure 1b) produce a moderately constrained region with a larger $G(H_F)$ than pairwise preferences, as they compare trajectories with the same start state, unlike pairwise preferences' comparisons between any two trajectories, failing to capture the total ranking. Demonstrations (Figure 1c) yield a larger $G(H_F)$ than corrections but smaller than E-stops, as they compare optimal trajectories to others but miss additional trajectory comparisons, failing to capture the total ranking, as shown theoretically by Brown et al. [8]. E-stops (Figure 1d) yield the largest feasible region – and hence the highest $G(H_F)$ – as they compare a trajectory halted at some point to the full trajectory, indicating that early termination is preferable, without comparing the original trajectory to others as pairwise preferences do.

B. Teacher's Perspective on Pairwise Preference vs. E-stop Feedback

Motivated by our empirical findings in Section V-A, where pairwise preferences led to lower reward ambiguity than E-stop feedback, we now provide a theoretical explanation for this phenomenon. Specifically, we analyze the geometric constraints each feedback type imposes on the reward space and prove why pairwise comparisons reduce ambiguity more effectively than E-stops.

Consider an MDP with state space $\mathcal{S}$ and feature vectors $\phi(s) \in \mathbb{R}^k$. For a trajectory $\xi = (s_0, \ldots, s_T) \in \Xi$, the cumulative feature sum is $\Phi(\xi) = \sum_{t=0}^{T} \phi(s_t)$. An *E-stop comparison*, preferring stopping at time $t < T$, imposes a constraint $w^\top(\Phi(\xi_{0:t}) - \Phi(\xi_{0:T})) \geq 0$, where $w \in \mathbb{R}^k$, $\|w\|_2 = 1$. A trajectory $\xi' \in \Xi$ is *distinct* if it includes a state s' with $\phi(s') \notin \text{span}\{\phi(s_0), \ldots, \phi(s_T)\}$.

Lemma 1. *A constraint on w imposed by an E-stop comparison lies within the subspace* $\text{span}\{\phi(s_0), \ldots, \phi(s_T)\}$. *In contrast, a pairwise comparison between ξ and a distinct trajectory $\xi' = (s'_0, \ldots, s'_{T'})$ imposes a constraint that includes directions outside* $\text{span}\{\phi(s_0), \ldots, \phi(s_T)\}$.

The proof of Lemma 1 is provided in Appendix B

Proposition 1. *Let $H_{\textit{E-stop}}$ be the feasible region of reward parameters $w \in \mathbb{R}^k$ constrained by E-stop feedback, defined as the set of w that satisfy constraints derived from preferring halted trajectories $\xi_{\textit{halted}}$ over trajectories $\xi \in \Xi$. Let $H_{\textit{pairwise}}$ be the feasible region constrained by pairwise preferences, defined as the set of w that satisfy constraints from comparing a trajectory $\xi \in \Xi$ to all other trajectories $\xi' \in \Xi$. Then $H_{\textit{pairwise}} \subset H_{\textit{E-stop}}$. Consequently, the reward ambiguity, defined as the volume of the feasible region, satisfies: $G(H_{\textit{pairwise}}) < G(H_{\textit{E-stop}})$, indicating that pairwise preferences reduce reward ambiguity more than E-stop feedback.*

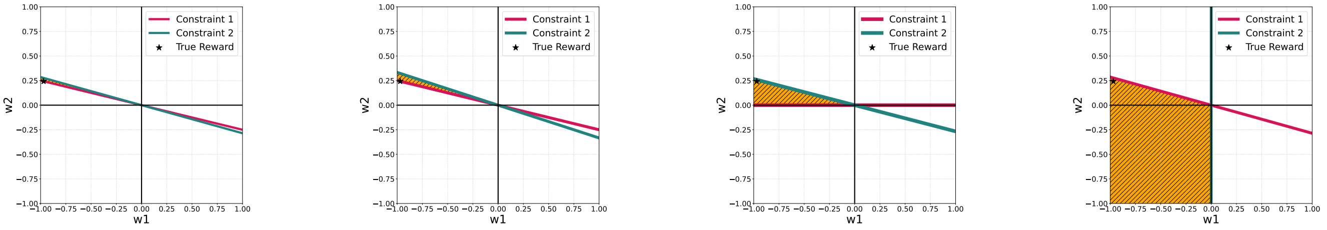
C. Teacher's Perspective on Pairwise Preference vs. Correction Feedback

To understand the impact of different feedback types on reward learning, we compare pairwise preferences and correction feedback in the context of a Markov Decision Process (MDP). Motivated by empirical observations that pairwise preferences often lead to lower reward ambiguity than other feedback types, we provide a theoretical explanation for why pairwise preferences are more effective than correction feedback at constraining the reward space.

Correction feedback involves comparing two trajectories that start from the same initial state $s_0 \sim \mu$. Given a trajectory $\xi = (s_0, s_1, \ldots, s_T)$ in the trajectory class Ξ, a corrected trajectory $\xi' = (s_0, s'_1, \ldots, s'_{T'})$ in Ξ starts at s_0 but diverges at some point (e.g., by taking a different action). If ξ' is preferred over ξ, this imposes a constraint on the reward parameters, $w^T(\Phi(\xi') - \Phi(\xi)) \geq 0$, where $\Phi(\xi) = \sum_{t=0}^{T} \gamma^t \phi(s_t)$.

In contrast, *pairwise preferences* involve comparing any two trajectories ξ and ξ' in the trajectory class Ξ, regardless of their starting states. This comparison imposes a constraint $w^T(\Phi(\xi) - \Phi(\xi')) \geq 0$, allowing preferences between trajectories from different parts of the state space.

Proposition 2. *Let $H_{\textit{correction}}$ be the feasible region of reward parameters $w \in \mathbb{R}^k$ constrained by correction feedback, defined as the set of w that satisfy constraints from comparing a trajectory $\xi \in \Xi$ to corrected trajectories $\xi_{corrected} \in \Xi$ starting from the same initial state. Let $H_{\textit{pairwise}}$ be the feasible*



(a) Pairwise preference (b) Correction (c) Demonstration (d) E-stop

Fig. 1: Feasible reward regions for different feedback types in a 2×3 grid environment with a terminal state at the bottom-right cell. Subfigures (a) to (d) depict the regions for pairwise preferences, corrections, demonstrations, and E-stops, respectively, ordered by increasing reward ambiguity $G(H_F)$, with the black dot indicating the ground-truth reward parameter.

region constrained by pairwise preferences, defined as the set of w that satisfy constraints from comparing $\xi \in \Xi$ to all other trajectories $\xi' \in \Xi$. Then: $H_{pairwise} \subset H_{correction}$. Consequently, the reward ambiguity, defined as the volume of the feasible region, satisfies: $G(H_{pairwise}) < G(H_{correction})$, indicating that pairwise preferences reduce reward ambiguity more effectively than correction feedback.

The proof of Preposition 1 and Preposition 2 is provided in Appendix C and Appendix D respectively.

VI. Experiments from the Learner's Perspective

We conduct two sets of experiments to support our claims. Each subsequent subsection presents a set of experiments and tests the relevant hypotheses. We conduct experiments to evaluate the learner's ability to determine feedback sufficiency in a passive learning setting, using diverse human feedback types (demonstrations, pairwise preferences, corrections, and E-stops) in a simulated gridworld environment. Our experiments assess the effectiveness of the normalized expected value difference (nEVD) stopping criterion and compare it with the normalized entropy criterion and a convergence baseline.

A. Experimental Setup

We evaluate feedback sufficiency in a passive learning setting using a gridworld environment with demonstrations, pairwise preferences, corrections, and E-stops. Two experiments test: (1) **Feedback effect on sufficiency declaration**: Using the nEVD stopping criterion, we measure how quickly each feedback type achieves low regret via α-VaR bounds. (2) **Comparison of stopping criteria**: We compare nEVD with entropy-based and convergence baseline criteria [32], where convergence declares sufficiency when the policy (π) stabilizes.

B. Feedback Effect on Sufficiency Declaration

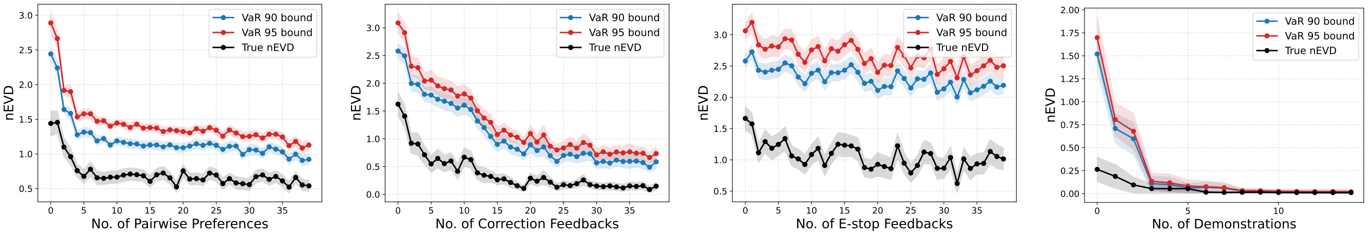
We investigate how different feedback modalities affect the learner's ability to declare feedback sufficiency, i.e., when the robot determines it has received enough feedback to learn a low-regret policy. We use the normalized nEVD α-VaR bound, which measures the performance gap between the learned and optimal policies under uncertainty, as the criterion for sufficiency. A lower nEVD α-VaR bound indicates higher confidence in the learned policy. We test the following hypotheses: **H1:** *The learner relies on a greater number of pairwise comparisons than demonstrations when the agent declares feedback sufficiency;* **H2:** *The learner depends on more E-stop feedback instances than pairwise comparisons when the agent declares feedback sufficiency;* **H3:** *The learner requires more pairwise preferences than corrections when the agent declares feedback sufficiency.*

Figure 2 shows nEVD α-VaR bounds, averaged over 20 seeds, with 15 demonstrations and 40 instances each of pairwise preferences, corrections, and E-stops. Demonstrations (Figure 2d) drop to near zero after 4 instances, reflecting optimal trajectory information, outperforming pairwise preferences (Figure 2a), which drop initially but stabilize above zero due to partial trajectory comparisons, supporting **H1**. Partial preferences, unlike total rankings [8], reduce informativeness, diverging from the teacher's perspective. Pairwise preferences decline steadily, while E-stops (Figure 2c) fluctuate without convergence, as their partial constraints limit policy improvement, supporting **H2**. Proposition 1 confirms E-stops' greater ambiguity, aligning perspectives. Corrections outperform pairwise preferences (Figure 2b), achieving lower regret faster, supporting **H3**. This contrasts the teacher's perspective, where pairwise preferences reduce ambiguity more (Proposition 2), as few comparisons prevent full trajectory ranking and corrections excel in goal-reaching tasks.

C. Comparison of Stopping Criteria

We examine how feedback modalities (demonstrations, pairwise preferences, corrections, E-stops) affect posterior convergence, measured by normalized entropy, and evaluate nEVD, entropy-based, and convergence-based stopping criteria for feedback sufficiency. The hypotheses we test are: **H4:** *Normalized entropy increases and converges to 1 fastest with demonstrations, followed by pairwise preferences, corrections, and E-stops;* **H5:** *The nEVD stopping criterion outperforms entropy-based and convergence-based criteria in F1 scores in the majority of feedback modalities.*

Figure 3 shows normalized entropy of the reward distribution averaged over twenty seeds, with informativeness order (demonstrations > pairwise preferences > corrections > E-stops) supporting **H4**. Demonstrations increase normalized entropy faster than pairwise preferences, which fail to reach demonstration levels despite double instances. This deviates from the teacher's



(a) Pairwise preference (b) Correction nEVD (c) E-stop nEVD (d) Demonstration nEVD

Fig. 2: Effect of pairwise preferences, correction, E-stop, and demonstration feedback on nEVD in the grid world environment, with results averaged over twenty seeds. The subfigures show: Pairwise Preference nEVD (2a), Correction nEVD (2b), E-stop nEVD (2c), and Demonstration nEVD (2d), illustrating how nEVD evolves as more feedback is provided.

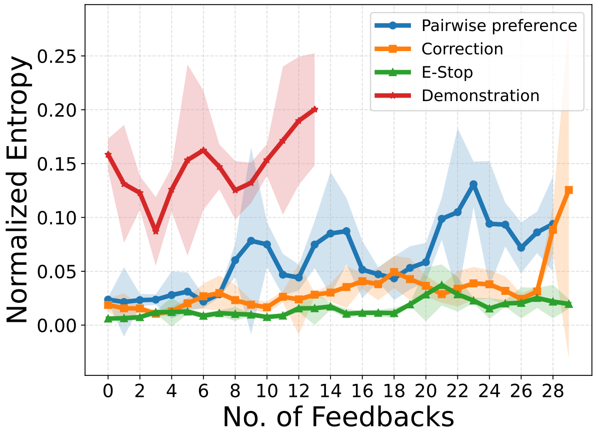


Fig. 3: Normalized entropy over different feedback modalities.

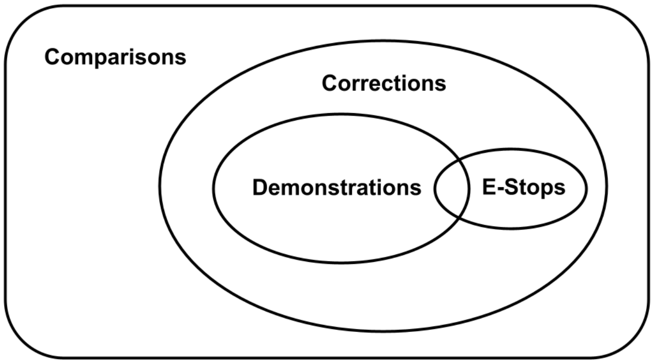


Fig. 4: The relationships between the sets of all implicit or explicit pairwise comparisons that result from different forms of human feedback.

perspective due to partial preferences' lack of total ordering [8]. Pairwise preferences outpace corrections, which cannot capture total ordering by comparing only same-start-state trajectories, consistent with Proposition 2. Pairwise preferences surpass E-stops, whose entropy fluctuates minimally, aligning with Proposition 1. For **H5**, Appendix Table I in Section G shows nEVD's higher F1 scores across all modalities compared to entropy-based criteria. Unlike entropy-based criteria, nEVD excels because low-regret policies can be learned despite high reward uncertainty (small normalized entropy). So we can see, from our comparison, that nEVD enabled low-regret policies without requiring reward distribution convergence, prioritizing policy performance over reward certainty. This makes entropy-based criteria less effective for self-assessment, as high uncertainty is tolerable with strong policies, favoring nEVD's effectiveness. However, convergence outperforms nEVD for E-stops (F1: 0.07 vs. nEVD's 0.05) by detecting sufficiency when normalized entropy stabilizes for p consecutive steps. E-stops' fluctuating, minimally changing entropy allows convergence to trigger, despite suboptimal policies with high nEVD. See Appendix H for F1 calculation, implementation, and criterion thresholds.

As discussed earlier, we propose to view all feedback types as special types of pairwise comparisons of the form $\xi_A \succ \xi_B$ as shown in Appendix Figure 6. One of our contributions is to characterize the relationships between the explicit and implicit pairwise comparisons induced by different feedback types. These relationships are shown in Figure 4. We describe the subset and superset nature of the different feedback types as follows: **Demonstrations** induce an infinite set of pairwise comparisons: $\xi^* \succeq \xi'$, $\forall \xi'$. If ξ^* is a prefix of ξ', then the corresponding pairwise comparison is also in the set of implicit pairwise comparisons that result from an e-stop. Otherwise, the pairwise comparison is not in the set of implicit comparisons that result from e-stops. **E-Stops** induce an implicit pairwise comparison: $\xi_{\text{halted}} \succ \xi_{\text{R}}$. If ξ_{halted} is an optimal trajectory, then this pairwise comparison is also contained in the set of implicit pairwise comparisons that result from demonstrations. Otherwise, the pairwise comparison is not contained in the demonstration set since the preferred trajectory is not optimal. **Corrections** are of the form $\xi_{\text{corrected}} \succ \xi_R$ where both trajectories start at the same state. Thus the implicit pairwise comparisons that result from a demonstration, $\xi^* \succeq \xi'$, $\forall \xi'$ are all special cases of correction feedback where the more preferred trajectory is optimal. The pairwise comparison implicit in an E-Stops, $\xi_{\text{halted}} \succ \xi_R$, is also a special case of a correction feedback since both ξ_R and ξ_{halted} start at the same state. **Comparisons** are of the form $\xi_A \succ \xi_B$, with no constraints on the trajectories. Thus, all other feedback types can be interpreted as special cases of comparisons and thus comparisons subsume the other feedback types.

VII. Conclusion and Future Work

We present a unified framework integrating robot self-assessment and human feedback informativeness across demonstrations, pairwise preferences, E-stops, and corrections. Our analyses show pairwise preferences best reduce reward ambiguity, followed by corrections, demonstrations, and E-stops, with regret-based stopping criteria outperforming entropy-based ones. These findings enhance human-robot collaboration. Future work will explore feedback burden vs. informativeness, considering human effort (e.g., low-effort E-stops vs. high-effort demonstrations).

References

[1] Chandrayee Basu, Erdem Bıyık, Zhixun He, Mukesh Singhal, and Dorsa Sadigh. Active learning of reward dynamics from hierarchical queries. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 120–127, 2019. doi: 10.1109/IROS40897.2019.8968522.

[2] Erdem Biyik and Dorsa Sadigh. Batch active preference-based learning of reward functions. In *Conference on robot learning*, pages 519–528. PMLR, 2018.

[3] Erdem Bıyık, Kenneth Wang, Nima Anari, and Dorsa Sadigh. Batch active learning using determinantal point processes. *arXiv preprint arXiv:1906.07975*, 2019.

[4] Erdem Bıyık, Nicolas Huynh, Mykel J Kochenderfer, and Dorsa Sadigh. Active preference-based gaussian process regression for reward learning. *arXiv preprint arXiv:2005.02575*, 2020.

[5] Erdem Bıyık, Dylan P. Losey, Malayandi Palan, Nicholas C. Landolfi, Gleb Shevchuk, and Dorsa Sadigh. Learning reward functions from diverse sources of human feedback: Optimally integrating demonstrations and preferences. *Int. J. Rob. Res.*, 41(1):45–67, January 2022. ISSN 0278-3649. doi: 10.1177/02783649211041652. URL https://doi.org/10.1177/02783649211041652.

[6] Daniel Brown, Wonjoon Goo, Prabhat Nagarajan, and Scott Niekum. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 783–792. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/brown19a.html.

[7] Daniel S Brown and Scott Niekum. Machine teaching for inverse reinforcement learning: Algorithms and applications. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7749–7758, 2019.

[8] Daniel S Brown, Wonjoon Goo, and Scott Niekum. Better-than-demonstrator imitation learning via automatically-ranked demonstrations. In *Conference on robot learning*, pages 330–359. PMLR, 2020.

[9] Gertjan J Burghouts, Albert Huizing, and Mark A Neerincx. Robotic self-assessment of competence. *arXiv preprint arXiv:2005.01546*, 2020.

[10] Maya Cakmak and Manuel Lopes. Algorithmic and human teaching of sequential decision tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, pages 1536–1542, 2012.

[11] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[12] Tesca Fitzgerald, Pallavi Koppol, Patrick Callaghan, Russell Quinlan Jun Hei Wong, Reid Simmons, Oliver Kroemer, and Henny Admoni. Inquire: Interactive querying for user-aware informative reasoning. In Karen Liu, Dana Kulic, and Jeff Ichnowski, editors, *Proceedings of The 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pages 2241–2250. PMLR, 14–18 Dec 2023. URL https://proceedings.mlr.press/v205/fitzgerald23a.html.

[13] Gaurav R Ghosal, Matthew Zurek, Daniel S Brown, and Anca D Dragan. The effect of modeling human rationality level on learning rewards from multiple feedback types. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5983–5992, 2023.

[14] Cory J. Hayes, Maryam Moosaei, and Laurel D. Riek. Exploring implicit human responses to robot mistakes in a learning from demonstration task. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, page 246–252. IEEE Press, 2016. doi: 10.1109/ROMAN.2016.7745138. URL https://doi.org/10.1109/ROMAN.2016.7745138.

[15] Borja Ibarz, Jan Leike, Tobias Pohlen, Geoffrey Irving, Shane Legg, and Dario Amodei. Reward learning from human preferences and demonstrations in atari. *Advances in neural information processing systems*, 31, 2018.

[16] Hong Jun Jeon, Smitha Milli, and Anca Dragan. Reward-rational (implicit) choice: A unifying formalism for reward learning. *Advances in Neural Information Processing Systems*, 33:4415–4426, 2020.

[17] Philippe Jorion. Value at risk: The new benchmark for managing financial risk. 01 2000.

[18] Parameswaran Kamalaruban, Rati Devidze, Volkan Cevher, and Adish Singla. Interactive teaching algorithms for inverse reinforcement learning. *arXiv preprint arXiv:1905.11867*, 2019.

[19] W. Bradley Knox, Stephane Hatgis-Kessell, Serena Booth, Scott Niekum, Peter Stone, and Alessandro G Allievi. Models of human preference for learning reward functions. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=hpKJkVoThY.

[20] Nathan Koenig, Leila Takayama, and Maja Matarić. 2010 special issue: Communication and knowledge sharing in human-robot interaction and learning from demonstration. *Neural Netw.*, 23(8–9):1104–1112, October 2010. ISSN 0893-6080. doi: 10.1016/j.neunet.2010.06.005. URL https://doi.org/10.1016/j.neunet.2010.06.005.

[21] Pallavi Koppol, Henny Admoni, and Reid Simmons. Interaction considerations in learning from humans. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 283–291. International Joint Conferences on

Artificial Intelligence Organization, 8 2021. doi: 10.24963/ijcai.2021/40. URL https://doi.org/10.24963/ijcai.2021/40. Main Track.

[22] Aleksandr Laptev and Boris Ginsburg. Fast entropy-based methods of word-level confidence estimation for end-to-end automatic speech recognition. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 152–159. IEEE, 2023.

[23] Qiujia Li, David Qiu, Yu Zhang, Bo Li, Yanzhang He, Philip C Woodland, Liangliang Cao, and Trevor Strohman. Confidence estimation for attention-based sequence-to-sequence models for speech recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6388–6392. IEEE, 2021.

[24] Dylan P Losey, Andrea Bajcsy, Marcia K O'Malley, and Anca D Dragan. Physical interaction as communication: Learning robot objectives online from human corrections. *The International Journal of Robotics Research*, 41(1): 20–44, 2022.

[25] Shaunak A. Mehta and Dylan P. Losey. Unified learning from demonstrations, corrections, and preferences during physical human–robot interaction. 13(3), 2024. doi: 10.1145/3623384. URL https://doi.org/10.1145/3623384.

[26] Yannick Metz, Andras Geiszl, Raphaël Baur, and Mennatallah El-Assady. Reward learning from multiple feedback types. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=9Ieq8jQNAl.

[27] Vivek Myers, Erdem Biyik, Nima Anari, and Dorsa Sadigh. Learning multimodal rewards from rankings. In Aleksandra Faust, David Hsu, and Gerhard Neumann, editors, *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pages 342–352. PMLR, 08–11 Nov 2022. URL https://proceedings.mlr.press/v164/myers22a.html.

[28] Adam Norton, Henny Admoni, Jacob Crandall, Tesca Fitzgerald, Alvika Gautam, Michael Goodrich, Amy Saretsky, Matthias Scheutz, Reid Simmons, Aaron Steinfeld, and Holly Yanco. Metrics for robot proficiency self-assessment and communication of proficiency in human-robot teams. *J. Hum.-Robot Interact.*, 11(3), July 2022. doi: 10.1145/3522579. URL https://doi.org/10.1145/3522579.

[29] Deepak Ramachandran and Eyal Amir. Bayesian inverse reinforcement learning. In *IJCAI*, volume 7, pages 2586–2591, 2007.

[30] Joar Max Viktor Skalse, Matthew Farrugia-Roberts, Stuart Russell, Alessandro Abate, and Adam Gleave. Invariance in policy optimisation and partial identifiability in reward learning. In *International Conference on Machine Learning*, pages 32033–32058. PMLR, 2023.

[31] Aviv Tamar, Yonatan Glassner, and Shie Mannor. Optimizing the cvar via sampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

[32] Tu Trinh, Haoyu Chen, and Daniel S. Brown. Autonomous assessment of demonstration sufficiency via bayesian inverse reinforcement learning. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '24, page 725–733. Association for Computing Machinery, 2024.

[33] Shuangge Wang, Anjiabei Wang, Sofiya Goncharova, Brian Scassellati, and Tesca Fitzgerald. Effects of robot competency and motion legibility on human correction feedback. *arXiv preprint arXiv:2501.03515*, 2025.

[34] Christian Wirth, Riad Akrour, Gerhard Neumann, and Johannes Fürnkranz. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research*, 18(136):1–46, 2017.

[35] Gaurav Yengera, Rati Devidze, Parameswaran Kamalaruban, and Adish Singla. Curriculum design for teaching via demonstrations: theory and applications. *Advances in Neural Information Processing Systems*, 34: 10496–10509, 2021.

Appendix

A. Detailed Derivation of Entropy

The posterior entropy is defined as:

$$H(w|\mathcal{H}_{1:i}) = -\mathbb{E}_{P(w|\mathcal{H}_{1:i})}\left[\log P(w|\mathcal{H}_{1:i})\right]. \tag{11}$$

Using Bayes' theorem:

$$P(w|\mathcal{H}_{1:i}) = \frac{P(\mathcal{H}_{1:i}|w)P(w)}{P(\mathcal{H}_{1:i})}, \tag{12}$$

we obtain:

$$\begin{aligned}\log P(w|\mathcal{H}_{1:i}) = \log P(\mathcal{H}_{1:i}|w) + \log P(w) \\ - \log P(\mathcal{H}_{1:i}).\end{aligned} \tag{13}$$

We assume a uniform prior over the unit sphere, where the probability density $P(w)$ is:

$$P(w) = \frac{1}{A_d}, \quad A_d = \frac{2\pi^{d/2}}{\Gamma(d/2)}, \tag{14}$$

and A_d is the surface area of the d-dimensional unit sphere. Here, $\Gamma(d/2)$ is the Gamma function, defined as $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\,dt$ for $z > 0$, ensuring normalization of the uniform prior.

Given a uniform prior, $\log P(w) = -\log A_d$, so:

$$\begin{aligned}H(w|\mathcal{H}_{1:i}) = -\mathbb{E}_{P(w|\mathcal{H}_{1:i})}\left[\log P(\mathcal{H}_{1:i}|w)\right] \\ - \log A_d + \log P(\mathcal{H}_{1:i}).\end{aligned} \tag{15}$$

To approximate the expectation, we use m_i MCMC posterior samples $\{w^{(k)}\}_{k=1}^{m_i} \sim P(w|\mathcal{H}_{1:i})$:

$$\begin{aligned}&\mathbb{E}_{P(w|\mathcal{H}_{1:i})}\left[\log P(\mathcal{H}_{1:i}|w)\right] \\ &\qquad \approx \frac{1}{m_i}\sum_{k=1}^{m_i} \log P(\mathcal{H}_{1:i}|w^{(k)}).\end{aligned} \tag{16}$$

Direct computation of the marginal likelihood:

$$P(\mathcal{H}_{1:i}) = \int P(\mathcal{H}_{1:i}|w)P(w)\,dw, \tag{17}$$

is challenging due to the high-dimensional parameter space $w \in \mathbb{R}^d$ with $\|w\| = 1$. Importance Sampling addresses this by using posterior samples from MCMC, avoiding additional prior sampling.

Using the Harmonic Mean Estimator (HME), we exploit the identity:

$$\mathbb{E}_{P(w|\mathcal{H}_{1:i})}\left[\frac{1}{P(\mathcal{H}_{1:i}|w)}\right] = \frac{1}{P(\mathcal{H}_{1:i})}, \tag{18}$$

yielding:

$$\hat{P}(\mathcal{H}_{1:i}) = \left(\frac{1}{m_i}\sum_{k=1}^{m_i}\frac{1}{P(\mathcal{H}_{1:i}|w^{(k)})}\right)^{-1}. \tag{19}$$

Its logarithm is:

$$\log \hat{P}(\mathcal{H}_{1:i}) = -\log\left(\frac{1}{m_i}\sum_{k=1}^{m_i}\frac{1}{P(\mathcal{H}_{1:i}|w^{(k)})}\right). \tag{20}$$

The posterior entropy estimate is:

$$\begin{aligned}\hat{H}(w|\mathcal{H}_{1:i}) \approx & -\frac{1}{m_i}\sum_{k=1}^{m_i}\log P(\mathcal{H}_{1:i}|w^{(k)}) \\ & - \log A_d - \log\left(\frac{1}{m_i}\sum_{k=1}^{m_i}\frac{1}{P(\mathcal{H}_{1:i}|w^{(k)})}\right).\end{aligned} \tag{21}$$

B. Proof of Lemma 1

Lemma. *(Restatement of Lemma 1) A constraint on w imposed by an E-stop comparison lies within the subspace* $\operatorname{span}\{\phi(s_0), \dots, \phi(s_T)\}$*. In contrast, a pairwise comparison between ξ and a distinct trajectory $\xi' = (s'_0, \dots, s'_{T'})$ imposes a constraint that includes directions outside* $\operatorname{span}\{\phi(s_0), \dots, \phi(s_T)\}$.

Proof: **Step 1: E-stop constraints are confined to a single subspace.**

The feature sum of a trajectory halted at t is $\Phi(\xi_{0:t}) = \sum_{i=0}^{t}\phi(s_i)$, while the feature sum of the full trajectory is $\Phi(\xi_{0:T}) = \sum_{i=0}^{T}\phi(s_i)$. Apparently, they are both linear combinations of $\{\phi(s_0), \dots, \phi(s_t)\}$, and therefore we have $\Phi(\xi_{0:t}), \Phi(\xi_{0:T}) \in \operatorname{span}\{\phi(s_0), \dots, \phi(s_T)\}$. As a result, we have

$$\Phi(\xi_{0:t}) - \Phi(\xi_{0:T}) \in \operatorname{span}\{\phi(s_0), \dots, \phi(s_T)\}. \tag{22}$$

In other words, the constraint $w^\top(\Phi(\xi_{0:t}) - \Phi(\xi_{0:T})) \geq 0$ lies within this subspace.

Step 2: Pairwise comparisons introduce new directions.

Consider a distinct trajectory $\xi' = (s'_0, \dots, s'_{T'})$ containing state s' such that $\phi(s') \notin \operatorname{span}\{\phi(s_0), \dots, \phi(s_T)\}$. Since $\phi(s')$ has a component not in this span, $\Phi(\xi') = \sum_{t=0}^{T'}\phi(s'_t)$, which includes $\phi(s')$, does too. Thus, the difference

$$\Phi(\xi_{0:T}) - \Phi(\xi') \notin \operatorname{span}\{\phi(s_0), \dots, \phi(s_T)\}. \tag{23}$$

In other words, the constraint $w^\top(\Phi(\xi_{0:T}) - \Phi(\xi')) \geq 0$ imposes a condition in a new direction, unachievable by E-stop constraints confined to $\operatorname{span}\{\phi(s_0), \dots, \phi(s_T)\}$.

■

C. Proof of Proposition 1

Proposition. *(Restatement of Proposition 1) Let $H_{\text{E-stop}}$ be the feasible region of reward parameters $w \in \mathbb{R}^k$ constrained by E-stop feedback, defined as the set of w that satisfy constraints derived from preferring halted trajectories ξ_{halted} over trajectories $\xi \in \Xi$. Let H_{pairwise} be the feasible region constrained by pairwise preferences, defined as the set of w that satisfy constraints from comparing a trajectory $\xi \in \Xi$ to all other trajectories $\xi' \in \Xi$. Then $H_{\text{pairwise}} \subset H_{\text{E-stop}}$. Consequently, the reward ambiguity, defined as the volume of the feasible region, satisfies: $G(H_{\text{pairwise}}) < G(H_{\text{E-stop}})$, indicating that pairwise preferences reduce reward ambiguity more than E-stop feedback.*

Proof: **Step 1: Define feasible regions** Consider an MDP with state space $\mathcal{S}$, action space $\mathcal{A}$, initial state distribution μ, and feature mapping $\phi : \mathcal{S} \to \mathbb{R}^k$. Trajectories from the trajectory class Ξ have different lengths (i.e., they may consist of varying numbers of states) and start from the initial state distribution μ.

- *E-stop Feedback*: If we have E-stop trajectories for all trajectories in Ξ, the feasible region is the intersection of half-spaces over all trajectories and their early-stopping variants:

$$H_{\text{E-stop}} = \bigcap_{\substack{\xi_i \in \Xi \\ \xi_i^{\text{stop}} \succ \xi_i}} \left\{ w \mid w^\top(\Phi_{\xi_i^{\text{stop}}} - \Phi_{\xi_i}) \geq 0 \right\}, \tag{24}$$

where ξ_i^{stop} denotes a trajectory that follows ξ_i but stops early, and $\xi_i^{\text{stop}} \succ \xi_i$ indicates the preference under E-stop feedback.

- *Pairwise Comparisons*: If we have all possible pairwise comparisons for trajectories in Ξ, resulting in a total ordering over Ξ, the feasible region is:

$$H_{\text{pairwise}} = \bigcap_{\substack{\xi_i \succ \xi_j \\ \xi_i, \xi_j \in \Xi}} \left\{ w \mid w^\top(\Phi_{\xi_i} - \Phi_{\xi_j}) \geq 0 \right\}, \tag{25}$$

where $\xi_i \succ \xi_j$ denotes the preference in the total ordering over all trajectories in Ξ.

Step 2: Show inclusion of feasible regions

To show that $H_{\text{pairwise}} \subseteq H_{\text{E-stop}}$, observe that every E-stop constraint $w^\top(\Phi_{\xi_i^{\text{stop}}} - \Phi_{\xi_i}) \geq 0$ for $\xi_i^{\text{stop}} \succ \xi_i$ is also a pairwise comparison constraint since ξ_i^{stop} can be considered a trajectory in Ξ. Therefore, the set of constraints defining H_{pairwise} includes all E-stop constraints, implying that any w in H_{pairwise} also satisfies the E-stop constraints, so $H_{\text{pairwise}} \subseteq H_{\text{E-stop}}$.

Step 3: Demonstrate strict inclusion Consider the trajectory class Ξ, where each trajectory $\xi_i \in \Xi$ is a sequence of states $(s_{i,0}, s_{i,1}, \dots, s_{i,T_i})$. By Lemma 1, an E-stop constraint for ξ_i, given by $w^\top(\Phi_{\xi_i^{\text{stop}}} - \Phi_{\xi_i}) \geq 0$, where ξ_i^{stop} is an early-stopping variant, lies within the subspace $\operatorname{span}\{\phi(s) \mid s \in \xi_i\}$. Thus, $H_{\text{E-stop}} = \bigcap_{\xi_i \in \Xi, \xi_i^{\text{stop}} \succ \xi_i}\{w \mid w^\top(\Phi_{\xi_i^{\text{stop}}} - \Phi_{\xi_i}) \geq 0\}$ is defined by constraining each local to a trajectory's subspace.

In contrast, pairwise preferences involve comparisons $\xi_i \succ \xi_j$ for any $\xi_i, \xi_j \in \Xi$, yielding constraints $w^\top(\Phi_{\xi_i} - \Phi_{\xi_j}) \geq 0$. Suppose ξ_j is a distinct trajectory containing a state s' such that $\phi(s') \notin \operatorname{span}\{\phi(s) \mid s \in \xi_i\}$. By Lemma 1, the vector

$\Phi_{\xi_i} - \Phi_{\xi_j}$ includes components outside $\text{span}\{\phi(s) \mid s \in \xi_i\}$, imposing a constraint in a direction not covered by any E-stop constraint on ξ_i. Across the trajectory class Ξ, which contains diverse trajectories potentially visiting distinct sets of states, each pairwise comparison between trajectories with non-overlapping or partially overlapping states introduces such unique constraints.

Since H_{pairwise} aggregates all such pairwise constraints, it enforces relationships across the feature spaces of different trajectories, unlike $H_{\text{E-stop}}$, which remains confined to trajectory-specific subspaces. This additional network of constraints ensures that there exists a $w \in H_{\text{E-stop}}$ satisfying all E-stop constraints (i.e., $w^\top(\Phi_{\xi_i^{\text{stop}}} - \Phi_{\xi_i}) \geq 0$ for all ξ_i) but violating at least one pairwise constraint, such as $w^\top(\Phi_{\xi_i} - \Phi_{\xi_j}) < 0$ for some $\xi_i \succ \xi_j$, where ξ_j introduces a new feature direction. Thus, $w \notin H_{\text{pairwise}}$, proving that $H_{\text{pairwise}} \subset H_{\text{E-stop}}$.

Step 4: Analyze reward ambiguity

The reward ambiguity $G(H_F)$ is defined as the volume of the feasible region H_F. Since $H_{\text{pairwise}} \subset H_{\text{E-stop}}$, and the additional constraints in H_{pairwise} reduce the volume of the feasible region, we have:

$$G(H_{\text{pairwise}}) < G(H_{\text{E-stop}}).$$

This indicates that pairwise comparisons provide stricter constraints on the reward parameters, thereby reducing the ambiguity in the reward function more effectively than E-stop feedback. We have shown that the feasible region under pairwise comparisons is a strict subset of the feasible region under E-stop feedback, leading to a smaller reward ambiguity. This completes the proof of Proposition 1. ■

D. Proof of Proposition 2

Proposition. *(Restatement of Proposition 2) Let $H_{\text{correction}}$ be the feasible region of reward parameters $w \in \mathbb{R}^k$ constrained by correction feedback, defined as the set of w that satisfy constraints from comparing a trajectory $\xi \in \Xi$ to corrected trajectories $\xi_{\text{corrected}} \in \Xi$ starting from the same initial state. Let H_{pairwise} be the feasible region constrained by pairwise preferences, defined as the set of w that satisfy constraints from comparing $\xi \in \Xi$ to all other trajectories $\xi' \in \Xi$. Then: $H_{\text{pairwise}} \subset H_{\text{correction}}$. Consequently, the reward ambiguity, defined as the volume of the feasible region, satisfies: $G(H_{\text{pairwise}}) < G(H_{\text{correction}})$, indicating that pairwise preferences reduce reward ambiguity more effectively than correction feedback.*

Proof: **Step 1: Define feasible regions**

Consider an MDP with state space $\mathcal{S}$, action space $\mathcal{A}$, initial state distribution μ, and feature mapping $\phi : \mathcal{S} \to \mathbb{R}^k$. Trajectories from the trajectory class Ξ have different lengths (i.e., they may consist of varying numbers of states) and start from the initial state distribution μ.

- *Correction Feedback*: The feasible region is defined by constraints from comparing a trajectory $\xi_i \in \Xi$ to a corrected trajectory $\xi_{\text{corr}} \in \Xi$ that starts from the same initial state s_0, with $\xi_{\text{corr}} \succ \xi_i$:

$$H_{\text{correction}} = \bigcap_{\substack{\xi_i, \xi_{\text{corr}} \in \Xi \\ s_0 = s_0' \\ \xi_{\text{corr}} \succ \xi_i}} \left\{ w \mid w^\top(\Phi_{\xi_{\text{corr}}} - \Phi_{\xi_i}) \geq 0 \right\},$$

where s_0 and s_0' are the starting states of ξ_i and ξ_{corr}.

- *Pairwise Preferences*: The feasible region is defined by constraints from comparing all pairs of trajectories in Ξ, assuming a total ordering:

$$H_{\text{pairwise}} = \bigcap_{\substack{\xi_i \succ \xi_j \\ \xi_i, \xi_j \in \Xi}} \left\{ w \mid w^\top(\Phi_{\xi_i} - \Phi_{\xi_j}) \geq 0 \right\}.$$

Step 2: Show inclusion of feasible regions

Every correction feedback constraint $w^\top(\Phi_{\xi_{\text{corr}}} - \Phi_{\xi_i}) \geq 0$, where $\xi_{\text{corr}} \succ \xi_i$ and both start at s_0, is also a pairwise preference constraint, since $\xi_{\text{corr}}, \xi_i \in \Xi$. Thus, the set of constraints defining H_{pairwise} includes all correction feedback constraints. Therefore, any $w \in H_{\text{pairwise}}$ satisfies all correction constraints, implying:

$$H_{\text{pairwise}} \subseteq H_{\text{correction}}.$$

Step 3: Demonstrate strict inclusion

To show that the inclusion $H_{\text{pairwise}} \subseteq H_{\text{correction}}$ is strict, we need to demonstrate that there exists a weight vector $w \in H_{\text{correction}}$ that is not in H_{pairwise}, thus proving $H_{\text{pairwise}} \subset H_{\text{correction}}$.

Let $\xi_a = (s_a, s_{a_2}, \ldots, s_{a_m})$ be a trajectory starting at state s_a, $\xi_c = (s_a, s_{c_2}, \ldots, s_{c_n})$ be another trajectory starting at s_a with subsequent states distinct from ξ_a's (i.e., $s_{c_i} \neq s_{a_j}$ for all $i, j \geq 2$), sharing only s_a, and $\xi_b = (s_b, s_{b_2}, \ldots, s_{b_p})$ be a trajectory starting at $s_b \neq s_a$, with no states shared with ξ_a (i.e., $s \notin \xi_a$ for all $s \in \xi_b$, and vice versa).

For correction feedback, consider the preference $\xi_c \succ \xi_a$, which is valid since both trajectories start at s_a. This imposes the constraint:

$$w^\top(\Phi_{\xi_c} - \Phi_{\xi_a}) \geq 0,$$

where $\Phi_\xi = \sum_{s \in \xi} \phi(s)$ represents the sum of feature vectors over the states in trajectory ξ. Define $v = \Phi_{\xi_c} - \Phi_{\xi_a}$. Since ξ_c and ξ_a share the starting state s_a, the terms $\phi(s_a)$ cancel out, so:

$$v = \sum_{s \in \xi_c \setminus \{s_a\}} \phi(s) - \sum_{s \in \xi_a \setminus \{s_a\}} \phi(s).$$

The correction feasible region includes all w satisfying $w^\top v \geq 0$, with $\|w\|_2 = 1$.

For pairwise preferences, consider the preference $\xi_a \succ \xi_b$, which imposes the constraint:

$$w^\top(\Phi_{\xi_a} - \Phi_{\xi_b}) \geq 0.$$

Define $u = \Phi_{\xi_a} - \Phi_{\xi_b}$. Since ξ_a and ξ_b do not share any states, we have:

$$u = \sum_{s \in \xi_a} \phi(s) - \sum_{s \in \xi_b} \phi(s).$$

The feature vectors in ξ_b introduce components that are distinct from those in ξ_a due to the absence of shared states.

Suppose ξ_b visits a state $s' \in \xi_b$ such that $\phi(s') \notin \text{span}\{\phi(s) \mid s \text{ visited by } \xi_a\}$. Because ξ_a and ξ_b share no states, and $\phi(s')$ is linearly independent of the feature vectors in ξ_a, the vector u includes a component outside the span of v, which depends only on states in $\xi_a \setminus \{s_a\}$ and $\xi_c \setminus \{s_a\}$. This introduces a constraint in a new direction not present in $H_{\text{correction}}$. There exists a $w \in H_{\text{correction}}$ that satisfies all same-start-state constraints (e.g., $w^\top v \geq 0$) but violates the pairwise constraint $w^\top u \geq 0$ (e.g., $w^\top u < 0$), so $w \notin H_{\text{pairwise}}$. Thus:

$$H_{\text{pairwise}} \subset H_{\text{correction}}.$$

Step 4: Analyze reward ambiguity

The reward ambiguity $G(H_F)$ is the volume of the feasible region. Since $H_{\text{pairwise}} \subset H_{\text{correction}}$, and pairwise preferences impose additional constraints (e.g., from different starting states), the volume of H_{pairwise} is strictly smaller:

$$G(H_{\text{pairwise}}) < G(H_{\text{correction}}).$$

This shows that pairwise preferences constrain the reward parameters more tightly, reducing ambiguity more effectively than correction feedback. This completes the proof. ■

E. Feasible Region Construction

Our experimental setup simulates a human teacher providing feedback to a robot navigating a 2 × 3 grid environment, as illustrated in Figure 5.

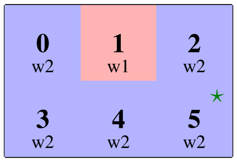


Fig. 5: Layout of the 2×3 grid environment. States are numbered 0 to 5 from the top-left corner, with colors indicating features ("Blue" or "Red"). The bottom-right cell (state 5) is the terminal state, marked with a green star.

The robot learns a reward function based on feedback, with the feasible region representing the set of possible reward parameters (w_1, w_2) consistent with the input. To ensure consistency across feedback types, we first generated 3,000,000 random trajectories with a maximum length of 10, which were then used to produce E-stops, corrections, and pairwise preferences, while demonstrations were generated separately.

Feedback was generated as follows: Demonstrations (Figure 1c) were created by determining optimal trajectories from each non-terminal state, selecting actions that maximize expected rewards until the terminal state or a maximum step limit was reached. Pairwise preferences (Figure 1a) were obtained by comparing pairs based on their total rewards and ranking the higher-reward trajectory as preferred, producing 3,000,000 comparisons. Corrections (Figure 1b) were simulated by taking random trajectories and generating alternative trajectories from the same starting state, selecting the one with the highest reward as the improved path. E-stops (Figure 1d) were simulated on 3,000,000 random trajectories by identifying unsafe states where the robot should stop, based on a human rationality model with a high confidence parameter and a discount factor of 0.99.

The feasible regions were visualized by translating feedback into constraints on the reward parameters. For demonstrations, we generated 100 random trajectories from each non-terminal state (states 0 to 4) and compared them to the optimal trajectory from that state. These comparisons created constraints ensuring the optimal trajectory had the highest reward.

Pairwise preferences imposed constraints that preferred trajectories outscored others. Corrections required the improved trajectories to have higher rewards, and E-stops excluded reward parameters giving positive rewards to unsafe states. These constraints formed a set of linear inequalities, and the feasible region was the intersection of these inequalities, plotted in the w_1-w_2 space. The regions' sizes reflect the informativeness order: pairwise preferences (narrowest), corrections (moderate), demonstration (larger), and E-stops (largest), as shown in Figures 1a, 1b, 1c, and 1d.

F. Normalizing Entropy

To estimate $H_{\max}$ and $H_{\min}$, we conducted calibration experiments for each feedback type (pairwise preferences, corrections, demonstrations, and E-stops). For each feedback type, we performed three independent experiments, tracking the entropy $H(w \mid \mathcal{H}_{1:i})$ of the reward posterior after each feedback instance. We then determined $H_{\max}$ as the maximum entropy observed across all experiments and all feedback types, and $H_{\min}$ as the minimum entropy observed. These values were used to normalize the confidence metric as described in Equation (8).

G. F1 Score Results

TABLE I: F1 scores for stopping criteria (5×5 gridworld)

Metric	**E-stop**	**Corr.**	**Demo**	**Pref.**
nEVD	0.05	**0.14**	**0.89**	**0.20**
Entropy	0.02	0.05	0.17	0.08
Conv.	**0.07**	0.09	0.75	0.07

H. Comparison of Stopping Criteria

We compare three stopping criteria in the feedback sufficiency problem and demonstrate that the normalized expected value difference (nEVD), as introduced in Equation 6, outperforms the normalized entropy and convergence criteria. The normalized entropy criterion, defined in Equation 8, captures the convergence of the posterior distribution over rewards, while the convergence criterion focuses on policy stabilization. We evaluate the performance of these criteria using the F1 score, with results summarized in Table I.

1) Identification Accuracy: To assess the accuracy of each stopping criterion in declaring feedback sufficiency, we compute the F1 score, which is the harmonic mean of precision and recall. This metric balances the ability of each criterion to correctly identify when sufficient feedback has been received. For each criterion, we define true positives (TP), false positives (FP), and false negatives (FN) as follows:

- **True Positive (TP)**: The criterion correctly declares feedback sufficiency when the condition for sufficiency is satisfied.
- **False Positive (FP)**: The criterion incorrectly declares feedback sufficiency when the condition is not met.
- **False Negative (FN)**: The criterion fails to declare feedback sufficiency when the condition is satisfied.

For the nEVD criterion, sufficiency is declared when $P(nEVD(\pi_{\text{robot}}, \theta) \leq \epsilon \mid \mathcal{H}_{1:i}) \geq \alpha$, where $\theta \sim P(\theta \mid \mathcal{H}_{1:i})$, and the F1 score is calculated as:

$$F1_{\text{nEVD}} = \frac{TP_{\text{nEVD}}}{TP_{\text{nEVD}} + \frac{1}{2}(FP_{\text{nEVD}} + FN_{\text{nEVD}})}.$$

For the normalized entropy criterion, sufficiency is declared when the confidence $F(w \mid \mathcal{H}_{1:i}) \geq \tau$, as defined in Equation 8, with the F1 score given by:

$$F1_{\text{entropy}} = \frac{TP_{\text{entropy}}}{TP_{\text{entropy}} + \frac{1}{2}(FP_{\text{entropy}} + FN_{\text{entropy}})}.$$

For the convergence criterion, sufficiency is declared when the policy remains unchanged over p consecutive steps, and the F1 score is computed similarly based on policy stabilization.

2) Threshold Values: To ensure a fair comparison across the stopping criteria, we evaluated each over a range of threshold values and computed the average F1 score over these ranges. This approach accounts for variability in performance depending on the threshold settings.

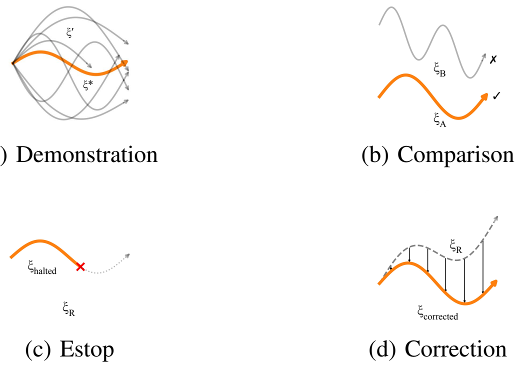
For the nEVD criterion, we tested a range of threshold values ϵ from 0.1 to 1.5 in increments of 0.1. The confidence level α was fixed at 0.95.

For the normalized entropy criterion, we evaluated a range of threshold values τ from 0.05 to 0.6 in increments of 0.05, corresponding to confidence levels of 5% to 60% in the reward posterior.

For the convergence criterion, we varied the number of consecutive steps p from 1 to 5.

The F1 scores reported in Table I represent the averages over these threshold ranges for each criterion, providing a robust comparison of their effectiveness in declaring feedback sufficiency.

I. Supplementary Figures



(a) Demonstration

(b) Comparison

(c) Estop

(d) Correction

Fig. 6: Feedback types: (a) Demonstration: $\xi^* \succeq \xi'$; (b) Comparison: $\xi_A \succ \xi_B$; (c) E-stop: $\xi_{\text{halted}} \succ \xi_R$; (d) Correction: $\xi_{\text{corrected}} \succ \xi_R$.