Multi-Modal, Multi-Environment Machine Teaching for Robust Reward Learning

Anonymous authors
Paper under double-blind review

Keywords: Inverse RL, Machine Teaching, Reward Learning

Summary

As autonomous agents are increasingly deployed across diverse operational contexts, aligning their behavior with human intent demands reward functions that remain robust to such changes rather than overfitting to any single operating condition. Inverse reinforcement learning (IRL) provides a principled way to infer such objectives from human feedback. However, existing machine teaching approaches for IRL focus on single-environment, demonstration-only settings, leaving underexplored how heterogeneous feedback modalities and environment dynamics jointly constrain reward functions that generalize across multiple environments. Because demonstrations in one MDP entangle reward information with that environment's specific structure, the resulting rewards frequently fail to generalize when the agent is deployed in a new setting. We first analyze how different feedback modalities constrain rewards, showing that, in the unlimited-data regime, comparisons impose strictly stronger global constraints than other modalities. We further examine the geometry of feasible reward regions empirically, showing that demonstrations impose tighter reward constraints under finite budgets. Beyond this theoretical analysis, we introduce a hierarchical machine teaching algorithm for IRL that operates across multiple MDPs. The algorithm first greedily selects informative environments that expose complementary reward constraints, then strategically queries low-cost feedback within those environments. Empirically, our method achieves substantially lower regret and stronger generalization to held-out environments than uniform teaching baselines under identical feedback budgets, demonstrating the importance of multi-environment, multi-modal teaching for learning dynamics-robust reward functions.

Contribution(s)

1. Analysis and insights into different feedback types for teaching IRL in both unlimited-data and limited-budget regimes, showing that comparisons impose the strongest global constraints while demonstrations are more constraint-efficient per query under tight budgets.
   Context: Prior work on teaching IRL (Büning et al., 2022; Brown and Niekum, 2019) has primarily focused on demonstrations and did not systematically analyze how different feedback types constrain reward recovery.
2. A formal characterization of environment-dependent reward identifiability, showing that even unlimited feedback in a single MDP leaves residual reward ambiguity.
   Context: Existing machine teaching approaches for IRL (Büning et al., 2022; Brown and Niekum, 2019) typically assume a fixed environment and do not analyze how environment dynamics affect reward identifiability.
3. Hierarchical Set Cover Optimal Teaching (HSCOT), a framework that selects informative environments and feedback queries to efficiently constrain rewards across multiple MDPs.
   Context: Prior teaching methods (Büning et al., 2022; Brown and Niekum, 2019) operate within a single environment and cannot exploit variation in environment dynamics to reveal complementary reward constraints.
4. Empirical validation showing that HSCOT achieves higher constraint coverage and lower regret on held-out environments compared to uniform teaching under identical budgets.
   Context: Existing evaluations of machine teaching for IRL (Büning et al., 2022; Brown and Niekum, 2019) mainly consider single-environment teaching settings.

Multi-Modal, Multi-Environment Machine Teaching for Robust Reward Learning

Anonymous authors
Paper under double-blind review

Abstract

As autonomous agents are increasingly deployed across diverse operational contexts, aligning their behavior with human intent demands reward functions that remain robust to such changes rather than overfitting to any single environment. Inverse reinforcement learning (IRL) provides a principled way to infer such objectives from human feedback. However, existing analysis of optimal teaching approaches for IRL focus on single-environment, demonstration-only settings, leaving underexplored how heterogeneous feedback modalities and environment dynamics jointly constrain reward functions that generalize across multiple environments. Because demonstrations in one MDP entangle reward information with that environment's specific structure, the resulting rewards frequently fail to generalize when the agent is deployed in a new setting. We first analyze how different feedback modalities constrain rewards, showing that, in the unlimited-data regime, comparisons impose strictly stronger global constraints than other modalities. Beyond this theoretical analysis, we introduce a hierarchical machine teaching algorithm for IRL that operates across multiple MDPs. The algorithm first greedily selects informative environments that expose complementary reward constraints, then strategically queries low-cost feedback within those environments. Empirically, our method achieves substantially lower regret and stronger generalization to held-out environments than uniform teaching baselines under identical feedback budgets, demonstrating the importance of multi-environment, multi-modal teaching for learning dynamics-robust reward functions.

1 Introduction

As autonomous agents become increasingly common, a central challenge is enabling humans to convey their intent in sequential decision-making settings. Rather than merely imitating behavior, agents must infer underlying objectives that remain aligned across varied situations (Kaufmann et al., 2024). In realistic settings, humans provide heterogeneous feedback rather than demonstrations alone (Mehta and Losey, 2024). Moreover, agents are rarely confined to a single operating condition; instead, they are deployed across multiple operational contexts that differ in dynamics, constraints, or embodiment. A service robot, for instance, may be required to perform the same task across different physical layouts, action limitations, or safety conditions, while still adhering to a consistent human-defined objective — for example, reaching a target location efficiently while avoiding hazards. Understanding how to best teach an agent to act appropriately across heterogeneous settings is challenging. The teacher must convey an objective that is invariant to changes in dynamics or environment-specific details, while the agent must learn a representation that generalizes beyond the conditions under which feedback is provided. This raises the central question of this work: *How can a human teach an agent that operates across multiple settings using heterogeneous feedback such that the agent's behavior remains aligned with human intent in all of them?*

This question naturally motivates a machine teaching perspective on sequential decision-making, where a teacher strategically selects feedback to induce a desired learning outcome with minimal cost. A natural formalism for representing human intent is inverse reinforcement learning (IRL),

in which an agent infers a reward function from human-provided feedback (Abbeel and Ng, 2004). The teaching problem can therefore be cast as machine teaching for IRL (Brown and Niekum, 2019), where the teacher selects a minimal yet maximally informative set of feedback instances to recover a reward that induces low-regret behavior. However, existing machine teaching frameworks for IRL largely rely on demonstrations as the sole feedback modality (Kamalaruban et al., 2019; Büning et al., 2022). Besides the single-modality setting, prior work on machine teaching for IRL has largely restricted teaching to a single environment with fixed dynamics (Haug et al., 2018; Kamalaruban et al., 2019; Yengera et al., 2021; Büning et al., 2022). Rewards learned in a single MDP often overfit to environment-specific dynamics, leading to failures when deployed under different dynamics or learning assumptions (Booth et al., 2023; Fu et al., 2017; He and Dragan, 2021). We formally characterize this limitation by showing that reward identifiability is environment-dependent. Consequently, resolving reward ambiguity often requires teaching across multiple environments rather than collecting more feedback within one.

To address the limitations of single-modality and single-environment teaching, we introduce *Hierarchical Set Cover Optimal Teaching (HSCOT)*, a framework that jointly selects informative environments and heterogeneous feedback queries to constrain the reward space across multiple MDPs. Unlike prior approaches that operate within a single environment, HSCOT exploits variation in environment dynamics to expose complementary reward constraints, yielding a hierarchical strategy in which environments are selected first and feedback instances are chosen within them. Our contributions are as follows:

- **Analysis of feedback modalities for teaching IRL.** We examine how different feedback types constrain reward identification in machine teaching for IRL. In the unlimited-data regime, comparisons impose the strongest global constraints; however, under limited budgets, demonstrations more constraint-efficient per query.
- **Environment-dependent reward identifiability.** We show that reward identifiability depends on environment dynamics. We formally characterize this limitation and show that additional environments can expose complementary constraint directions for better reward identification.
- **Hierarchical multi-environment machine teaching.** We introduce Hierarchical Set Cover Optimal Teaching (HSCOT), a framework that jointly selects informative environments and feedback queries to efficiently constrain rewards across multiple MDPs.
- **Empirical validation.** Experiments show that HSCOT achieves higher constraint coverage and significantly lower regret on held-out environments compared to uniform teaching.

2 Related work

The problem of identifying minimal teaching sets has been extensively studied in Algorithmic Teaching (Balbach and Zeugmann, 2009) and Machine Teaching (Liu et al., 2017; Zhu et al., 2018; Zhu, 2015; Devidze et al., 2020; Wang et al., 2021). These works primarily consider supervised learning settings such as classification and regression, where teaching reduces to selecting informative examples for a learner with a fixed hypothesis class. In contrast, teaching in sequential decision-making requires reasoning about policies, trajectories, and reward functions, introducing additional structure absent from standard supervised settings (Cakmak and Lopes, 2012). Our work builds on this paradigm and studies machine teaching for inverse reinforcement learning (IRL) across multiple environments with heterogeneous feedback. The most closely related work studies machine teaching for IRL, introduced by Brown and Niekum (2019), which formulates teaching as selecting demonstrations that induce a target reward in a learner within a single Markov decision process (MDP). Subsequent work extends this framework to interactive settings where the teacher adaptively selects demonstrations based on the learner's current policy or posterior over rewards (Kamalaruban et al., 2019; Büning et al., 2022). Related work considers learner-specific preferences or internal constraints (Tschiatschek et al., 2019) or feature mismatch (Haug et al., 2018). However, these approaches assume a single environment and rely on demonstrations as the primary feedback modality.

To our knowledge, we are the first to study machine teaching for IRL across multiple environments with heterogeneous feedback modalities.

Outside classical IRL teaching, several works study teaching or guidance for sequential decision-making agents under alternative assumptions, including curriculum learning (Yengera et al., 2021), planning-based teaching MDPs (Peltola et al., 2019), and teaching-by-reinforcement for online RL agents (Zhang et al., 2021). These methods do not address reward inference or reward generalization across multiple environments.

A parallel line of work studies reward learning from human feedback, including demonstrations, rankings, preferences, and corrective signals (Wirth et al., 2017; Christiano et al., 2017; Brown et al., 2019; Myers et al., 2022; Losey et al., 2022; Wang et al., 2025). Several approaches develop unified probabilistic frameworks for combining heterogeneous feedback signals (Jeon et al., 2020; Mehta and Losey, 2024; Metz et al., 2025), while others study the effects of human rationality assumptions or feedback noise (Ghosal et al., 2023). These works primarily adopt a learner-centric perspective, focusing on how to infer rewards from feedback rather than how a teacher should select informative environments and queries. Recent studies also show that rewards learned in a single environment can become entangled with environment dynamics and fail to generalize across settings (Fu et al., 2017; He and Dragan, 2021; Booth et al., 2023). While prior machine teaching approaches implicitly assume that sufficient demonstrations in one MDP can identify the reward, some reward ambiguities are structural rather than data-driven. Our work addresses this limitation by teaching across multiple MDPs, where variation in dynamics reveals complementary reward constraints.

Conceptually, our approach extends the Behavioral Equivalence Class framework of Brown and Niekum (2019) beyond single-MDP demonstration teaching. We define *generalized behavioral equivalence classes*, which unify reward constraints induced by heterogeneous feedback across multiple environments. We further propose a hierarchical teaching structure that separates environment and feedback selection: the outer level selects environments whose dynamics expose complementary reward constraints, while the inner level chooses feedback that efficiently shrinks the feasible reward region. To our knowledge, this is the first framework to jointly address multi-environment teaching and heterogeneous feedback selection within a unified optimization formulation.

In summary, prior work has largely progressed along two separate directions: machine teaching methods that operate within a single MDP, and learner-centric approaches that infer rewards from heterogeneous feedback without addressing how teaching should be structured across environments. Our work bridges this gap by connecting the teaching problem with the fundamental challenges of reward learning in single-environment settings. We introduce a teacher-centric framework that operates across multiple MDPs, unifies diverse feedback modalities through generalized behavioral equivalence classes, and employs a hierarchical teaching strategy that improves reward identifiability and promotes generalization across varying dynamics.

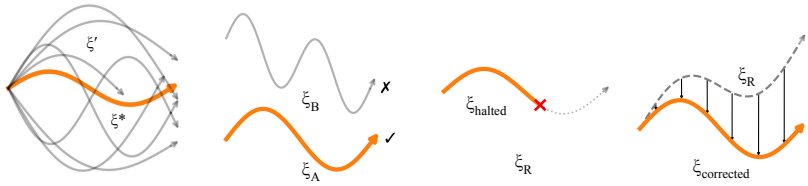
3 Preliminaries

3.1 Markov decision processes

We consider a finite Markov decision process (MDP) $M = (\mathcal{S}, \mathcal{A}, T, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ the action space, $T(s' \mid s, a)$ the transition function, and $\gamma \in (0, 1)$ the discount factor. We assume a linear reward function defined by a feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, so that $R_w(s, a) = w^\top \phi(s, a)$, where $w, w^\star \in \mathbb{R}^d$ For any policy π, we denote the discounted feature counts as $\mu^\pi(s) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t \phi(s_t, a_t)]$ and $\mu^\pi(s, a) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t \phi(s_t, a_t) \mid s_0 = s, a_0 = a]$. The corresponding value and Q-functions are $V_w^\pi(s) = w^\top \mu^\pi(s)$ and $Q_w^\pi(s, a) = w^\top \mu^\pi(s, a)$, and we let $\pi(w)$ denote an optimal policy under reward parameter w.

3.2 Human feedback models

We model human feedback under the reward-rational choice framework (Jeon et al., 2020). A feedback instance corresponds to a choice $c \in \mathcal{C}$ from a choice set, grounded to a trajectory through a mapping $\psi : \mathcal{C} \to \Xi$. For any trajectory $\xi = (s_0, a_0, s_1, a_1, \dots)$, define its discounted feature vector



(a) Demonstration (b) Comparison (c) Estop (d) Correction

Figure 1: **Human feedback modalities as trajectory comparisons.** Orange trajectories denote the preferred outcome. (a) Demonstration: expert trajectory ξ^* preferred over alternatives. (b) Comparison: preference $\xi_A \succ \xi_B$. (c) E-stop: halted trajectory $\xi_{\text{halted}} \succ \xi_R$. (d) Correction: corrected trajectory $\xi_{\text{corr}} \succ \xi_R$.

$\Phi(\xi) = \sum_{t=0}^{\infty} \gamma^t \phi(s_t, a_t)$. Given reward weights w and rationality parameter $\beta \geq 0$, the likelihood of observing choice c is $P(c \mid w, \beta) \propto \exp\big(\beta\, r_w(\psi(c))\big)$, where $r_w(\xi) = w^\top \Phi(\xi)$.

Demonstrations. Demonstrations correspond to reward-rational choices of state–action pairs with $\mathcal{C} = \mathcal{S} \times \mathcal{A}$ and $\psi(s, a) = (s, a)$; the trajectory likelihood factorizes as

$$P(\xi \mid w, \beta) = \prod_{(s_t, a_t) \in \xi} \frac{\exp(\beta Q_w(s_t, a_t))}{\sum_{b \in \mathcal{A}} \exp(\beta Q_w(s_t, b))}.$$

In the high-rationality limit $\beta \to \infty$, demonstrations implicitly induce preferences $\xi^+ \succeq \xi^-$, where ξ^+ is the demonstrated trajectory ξ and ξ^- denotes any alternative trajectory.

Comparisons. Comparisons select between trajectories $\xi^+, \xi^- \in \Xi$ with grounding $\psi(\xi_i) = \xi_i$. Under the Bradley–Terry model,

$$P(\xi_A \mid w, \beta) = \frac{\exp(\beta r_w(\xi^+))}{\exp(\beta r_w(\xi^+)) + \exp(\beta r_w(\xi^-))}.$$

Emergency stop (E-stop). Given a rollout ξ_R halted at time t, define $\xi_{\text{halted}} = \xi_R^{0:t} \xi_R^t \ldots \xi_R^t$, a trajectory of equal length that remains at state ξ_R^t after halting. The binary choice set $\mathcal{C} = \{\texttt{off}, -\}$ grounds to $\psi(\texttt{off}) = \xi_{\text{halted}}$ and $\psi(-) = \xi_R$, yielding

$$P(\texttt{off} \mid w, \beta) = \frac{\exp(\beta r_w(\xi_{\text{halted}}))}{\exp(\beta r_w(\xi_{\text{halted}})) + \exp(\beta r_w(\xi_R))}.$$

Corrections. Corrections provide an improved trajectory ξ_{corr} relative to a rollout ξ_R, interpreted as a localized comparison:

$$P(\xi_{\text{corr}} \mid w, \beta) = \frac{\exp(\beta r_w(\xi_{\text{corr}}))}{\exp(\beta r_w(\xi_{\text{corr}})) + \exp(\beta r_w(\xi_R))}.$$

As $\beta \to \infty$, all modalities reduce to linear preference constraints $w^\top(\Phi(\xi^+) - \Phi(\xi^-)) \geq 0$, providing the unified constraint interpretation used throughout. We adopt a teacher-centric view in which environments and feedback instances are selected across multiple MDPs to expose informative reward constraints. Illustrations of the feedback modalities are shown in Figure 1.

4 Feedback informativeness

We analyze how different feedback modalities constrain the feasible reward space. We provide intuition for the finite-data and analyze idealized unlimited-data regimes to understand how each modality reduces reward ambiguity. A central tool for reasoning about reward ambiguity is the Behavioral Equivalence Class (BEC) (Brown and Niekum, 2019).

Definition 1 (Behavioral equivalence class (BEC))**.** Let π be a policy in a single MDP M with linear reward $R_w(s,a) = w^\top \phi(s,a)$. The behavioral equivalence class of π is

$$\mathrm{BEC}(\pi) = \left\{ w \in \mathbb{R}^d \;\middle|\; w^\top \big(\mu^\pi(s,\pi(s)) - \mu^\pi(s,b)\big) \geq 0,\ \forall s \in \mathcal{S},\ \forall b \in \mathcal{A} \right\},$$

i.e., the intersection of halfspaces defined by the optimality constraints of π. This constraint-based view extends naturally to arbitrary feedback datasets.

Definition 2 (Generalized behavioral equivalence class (gBEC))**.** Let D be a feedback dataset collected across one or more environments. Each feedback instance induces one or more linear preference constraints of the form $w^\top \Delta\Phi \geq 0$, where $\Delta\Phi = \Phi(\xi^+) - \Phi(\xi^-)$.

Let $\Delta\Phi(D)$ denote the set of all feature-difference vectors induced by the feedback dataset D. The generalized behavioral equivalence class induced by D is

$$\mathrm{gBEC}(D) = \left\{ w \in \mathbb{R}^d \;\middle|\; w^\top \Delta\Phi \geq 0,\ \forall \Delta\Phi \in \Delta\Phi(D) \right\}. \tag{1}$$

Connection to BECs. When D consists of optimal demonstrations from a policy π that visit all states, $\mathrm{gBEC}(D)$ coincides with the classical $\mathrm{BEC}(\pi)$, since each demonstrated state s induces optimality constraints of the form $w^\top(\mu^\pi(s,\pi(s)) - \mu^\pi(s,b)) \geq 0$ for all $b \in \mathcal{A}$.

Feasible reward region and ambiguity. For any dataset D, the feasible reward region is $\mathrm{gBEC}(D)$. We quantify reward ambiguity by its volume

$$G(D) = \mathrm{Volume}\big(\mathrm{gBEC}(D)\big),$$

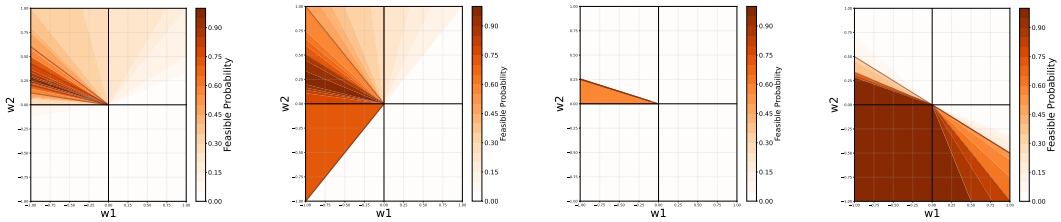
assuming $\|w\|_2 \leq 1$ for boundedness. Smaller $G(D)$ indicates tighter reward identification.

4.1 Constraint geometry under limited and unlimited feedback

To build intuition about how feedback modalities shape the geometry of the feasible reward region, we first analyze how different forms of feedback constrain $\mathrm{gBEC}(D)$ and reduce reward ambiguity. We consider two regimes: (i) finite-data teaching with a limited feedback budget, and (ii) unlimited-data teaching, which isolates intrinsic informativeness by examining the feasible region induced by arbitrarily many feedback instances of a single type.

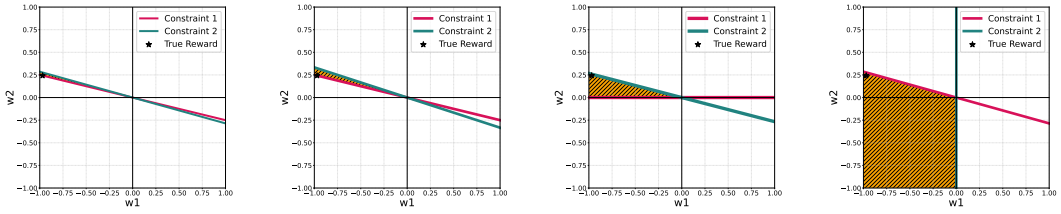
Limited-data teaching regime. Figure 2 visualizes feasible reward regions under tight budgets as heatmaps over reward weight space, where brighter areas indicate weights that remain consistent across many randomly sampled feedback sets. We compare three demonstrations against five instances of each other modality. Demonstrations (Fig. 2c) produce the smallest $\mathrm{gBEC}(D)$ and lowest ambiguity $G(D)$. Each demonstration implicitly enforces many optimality constraints, yielding strong shrinkage of the feasible region even with fewer queries. In contrast, comparisons (Fig. 2a), corrections (Fig. 2b), and E-stop feedback (Fig. 2d) each contribute only one linear constraint per instance, resulting in progressively larger feasible regions. Corrections stay more localized thanks to their shared start-state structure, while E-stop feedback yields the largest $\mathrm{gBEC}(D)$ because its constraints are highly trajectory-specific and local. Demonstrations are especially powerful under limited budgets, as each piece of feedback delivers rich, multi-faceted behavioral information.

Unlimited-data teaching regime. We next examine the idealized limit in which a teacher can provide arbitrarily many feedback instances of a single type. Figure 3 shows the corresponding feasible reward regions. In contrast to the finite-data setting, comparisons induce the smallest $\mathrm{gBEC}(D)$ and lowest ambiguity $G(D)$, followed by corrections, demonstrations, and E-stops. This reversal highlights an important distinction: while demonstrations provide strong information per example, comparison feedback becomes more informative in the unlimited-data setting because it can enforce global ordering constraints across the trajectory space. Taken together, these observations reveal a key teaching trade-off. Different feedback modalities influence the geometry of $\mathrm{gBEC}(D)$ in fundamentally different ways, and their relative effectiveness depends on the available teaching budget. This motivates the theoretical analysis in the next subsections, where we formally characterize how feedback modalities constrain reward ambiguity in the unlimited-data regime.



(a) Comparison (b) Correction (c) Demonstration (d) E-stop

Figure 2: Finite-data feasible reward regions under different types of feedback. Heatmaps visualize the empirical feasibility probability over the reward space—brighter colors indicate weight vectors that remain feasible across many randomly sampled feedback sets. The heatmap illustrates how different feedback modalities constrain the generalized behavioral equivalence class $\mathrm{gBEC}(D)$ and affect reward ambiguity $G(D)$ in the data-limited regime (corresponding to the layout in Figure 4a).



(a) Comparison (b) Correction (c) Demonstration (d) E-stop

Figure 3: Feasible reward regions $\mathrm{gBEC}(D)$ for different feedback modalities. Subfigures (a)–(d) show how infinite comparisons, corrections, demonstrations, and E-stops, respectively, shape the reward feasibility region. The black dot indicates the ground-truth reward parameter (corresponding to the layout in Figure 4a).

5 Theoretical analysis of feedback modalities

The geometric analysis above showed that different feedback modalities reduce reward ambiguity in qualitatively different ways. We now formalize these differences by analyzing how each modality constrains the feasible reward region in the unlimited-data regime.

Let Ξ denote the trajectory space of environment M, i.e., the set of all trajectories that can arise in M. Throughout this section we assume the teacher can obtain arbitrarily many feedback instances of a given modality. We use comparison as a reference modality. comparison compare any two trajectories $\xi_i, \xi_j \in \Xi$, allowing ordering constraints across the entire trajectory space. Therefore, in the unlimited-data regime, a complete set of preferences consistent with the true reward $w^\star$ induces a full ordering over Ξ and yields the feasible region

$$H_{\text{comparison}} = \Big\{ w \in \mathbb{R}^d : \forall \xi_i, \xi_j \in \Xi \text{ with } w^{\star\top}\Phi(\xi_i) > w^{\star\top}\Phi(\xi_j),\ w^\top(\Phi(\xi_i) - \Phi(\xi_j)) \geq 0 \Big\}.$$

We compare this region to those induced by demonstrations, corrective, and E-stop feedback.

5.1 Comparison vs. demonstrative feedback

As mentioned previously, demonstrations induce implicit constraints involving trajectories from the same initial state. comparison, in contrast, compare any two trajectories in Ξ, imposing ordering constraints across the entire trajectory space.

Proposition 1 (comparison reduce reward ambiguity compared to demonstrations)**.** *Let $\Xi^\star \subseteq \Xi$ denote the trajectories that are optimal under $w^\star$. Demonstrations induce the feasible region*

$$H_{demo} = \Big\{ w \in \mathbb{R}^d : \forall \xi^\star \in \Xi^\star,\ \forall \xi' \in \Xi \textit{ with the same initial state as } \xi^\star,\ w^\top(\Phi(\xi^\star) - \Phi(\xi')) \geq 0 \Big\}.$$

Then

$$H_{comparison} \subsetneq H_{demo}, \qquad G(H_{comparison}) < G(H_{demo}).$$

5.2 Comparison vs. corrective feedback

Corrective feedback compares a corrected trajectory ξ_{corr} with the agent's rollout ξ_R, where both trajectories share the same initial state.

Proposition 2 (Comparison strictly reduce reward ambiguity compared to corrective feedback)**.** *Corrective feedback induces the feasible region*

$$H_{\textit{correction}} = \Big\{ w \in \mathbb{R}^d : \ \forall \xi_{\text{corr}}, \xi_R \in \Xi \textit{ with the same initial state}, \ w^\top(\Phi(\xi_{\text{corr}}) - \Phi(\xi_R)) \geq 0 \Big\}.$$

Then

$$H_{\textit{comparison}} \subsetneq H_{\textit{correction}}, \qquad G(H_{\textit{comparison}}) < G(H_{\textit{correction}}).$$

5.3 Comparison vs. E-stop feedback

E-stop feedback compares a halted prefix $\xi^{0:t}$ with the continuation of the same trajectory ξ, penalizing undesirable future behavior. Because both share the same prefix, the feature difference $\Phi(\xi^{0:t}) - \Phi(\xi)$ depends only on states visited after time t along the same trajectory. Consequently, E-stop constraints are confined to the feature subspace induced by that trajectory.

Lemma 1 (E-stop constraints are trajectory-local)**.** *Fix a trajectory $\xi \in \Xi$. Any E-stop constraint comparing $\xi^{0:t}$ with ξ lies in the subspace*

$$\text{span}\{\phi(s,a) \mid (s,a) \textit{ appears in } \xi\}.$$

Comparison between trajectories ξ and $\xi' \in \Xi$ producex feature differences outside this trajectory-specific subspace.

Proposition 3 (Comparison strictly reduce reward ambiguity compared to E-stop feedback)**.** *E-stop feedback induces the feasible region*

$$H_{\textit{E-stop}} = \Big\{ w \in \mathbb{R}^d : \|w\|_2 \leq 1, \ \forall \xi \in \Xi, \ \forall t, \ w^\top(\Phi(\xi^{0:t}) - \Phi(\xi)) \geq 0 \Big\}.$$

Then

$$H_{\textit{comparison}} \subsetneq H_{\textit{E-stop}}, \qquad G(H_{\textit{comparison}}) < G(H_{\textit{E-stop}}).$$

6 Single-MDP limitations: environment-dependent reward identifiability

The previous section characterized the intrinsic informativeness of different feedback modalities in the unlimited-data regime within a single MDP. In realistic settings, however, teaching must operate under finite feedback budgets and produce rewards that remain effective across different environments. Even with unlimited feedback available in one MDP, the constraints induced by that particular environment may fail to uniquely identify the ground-truth reward when the agent is deployed in other environments. We now formalize a key limitation of single-MDP teaching: reward identifiability depends on the environment.

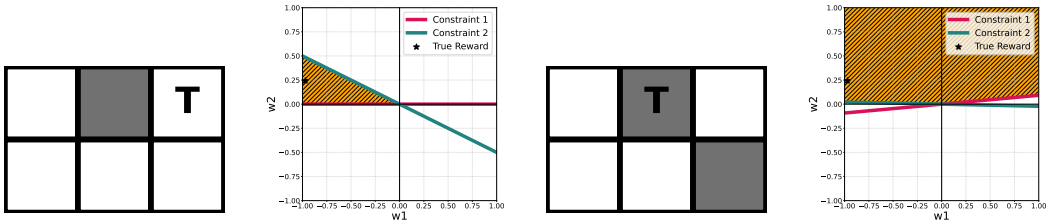
Theorem 1 (Single-MDP ambiguity)**.** *Let $\mathcal{M} = \{M_1, M_2\}$ be two finite MDPs that share a feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ and linear rewards $r_w(s,a) = w^\top \phi(s,a)$ with $w \in \mathbb{R}^d \subset \mathbb{R}^d$. Let $w^\star$ denote the ground-truth reward parameter. For any MDP M_k, let $\mathcal{W}(M_k)$ denote the generalized behavioral equivalence class (gBEC), i.e., the set of reward parameters consistent with* all possible *idealized feedback constraints obtainable within M_k. Define the feature-difference span induced by M_k as*

$$\mathcal{V}_k = \text{span}\{\Delta\Phi(\xi^+, \xi^-) \ : \ \xi^+, \xi^- \textit{ are feasible trajectories in } M_k\},$$

where $\Delta\Phi(\xi^+, \xi^-) = \Phi(\xi^+) - \Phi(\xi^-)$ and $\Phi(\xi)$ denotes the feature expectation of trajectory ξ.

Assume that the induced spans satisfy $\mathcal{V}_1 \subsetneq \text{span}(\mathcal{V}_1 \cup \mathcal{V}_2)$. Then there exists a reward parameter $w \neq w^\star$ such that $w \in \mathcal{W}(M_1)$ but $w \notin \mathcal{W}(M_1) \cap \mathcal{W}(M_2)$. In particular, no amount of feedback obtained solely from M_1 is sufficient to uniquely identify $w^\star$.

The proof is provided in Appendix 10.1.



(a) MDP 1 (b) gBEC in MDP 1 (c) MDP 2 (d) gBEC in MDP 2

Figure 4: **Environment-dependent reward ambiguity.** Two gridworld MDPs share the same reward features but differ in layout. The feasible reward region in MDP 2 is strictly larger than in MDP 1 ($\mathrm{gBEC}_1 \subsetneq \mathrm{gBEC}_2$), illustrating how environment structure affects reward identifiability.

The theorem shows that reward identifiability depends critically on the span of feature differences achievable by feasible trajectories in each MDP. When a single MDP does not span all reward-relevant directions, residual ambiguity persists even with unlimited idealized feedback. For a concrete example, consider two gridworld MDPs that share the same 2D feature representation (white/gray cells) and ground-truth linear reward $w^\star$, but differ in layout, cell arrangement, and terminal state position (see Figures 4a–4c). These layout differences produce distinct sets of feasible trajectories and thus distinct spans of achievable feature differences, leading to different generalized behavioral equivalence classes. The resulting feasible reward regions appear in Figures 4b and 4d: $\mathrm{gBEC}_1 \subsetneq \mathrm{gBEC}_2$. Every reward consistent with all optimal demonstrations in MDP 1 remains feasible in MDP 2 (but not vice versa). Teaching only in the "larger" MDP 2 therefore leaves reward ambiguity that can cause generalization failures to MDP 1, while teaching in MDP 1 excludes more inconsistent rewards. This asymmetry stems purely from differences in environment layout.

These observations reveal that single-MDP teaching is frequently insufficient to achieve unambiguous reward identification, particularly when generalization across environments is required. Under realistic finite feedback budgets, the situation becomes even more challenging: the teacher faces two tightly coupled problems—(i) identifying environments whose combined constraint spans resolve the remaining ambiguity, and (ii) selecting the most informative feedback instances within those environments under a limited total query budget. This naturally motivates a hierarchical approach, in which environment selection activates complementary constraint subspaces and feedback selection efficiently covers those subspaces. We next formalize this two-stage strategy as a hierarchical set-cover optimization problem and introduce hierarchical set cover optimal teaching (HSCOT), an algorithm for multi-environment teaching with heterogeneous feedback.

7 Hierarchical Set Cover Optimal Teaching (HSCOT)

We introduce *Hierarchical Set Cover Optimal Teaching (HSCOT)*, a machine teaching framework that learns reward functions robust to changes in environment dynamics by strategically selecting both environments and feedback queries. The key idea is that different environments expose different reward constraints through their feasible trajectory spaces. HSCOT exploits this structure by first selecting environments whose dynamics reveal complementary reward constraints and then selecting feedback instances within those environments that efficiently cover those constraints.

7.1 Generalized teaching formulation

We formulate teaching a reward learning agent across multiple MDPs with heterogeneous feedback as a machine teaching problem. This generalizes the single-MDP, demonstration-only setting of Brown and Niekum (2019) to a multi-environment setting where the teacher selects both environments and feedback instances to efficiently constrain the reward and promote generalization.

Let $\mathcal{M} = \{M_1, \ldots, M_K\}$ be a family of MDPs sharing a feature map $\phi : \bigcup_k (\mathcal{S}_k \times \mathcal{A}_k) \to \mathbb{R}^d$, where $\mathcal{S}_k$ and $\mathcal{A}_k$ denote the state and action spaces of M_k. Let $\mathcal{M}_{\text{train}}, \mathcal{M}_{\text{test}} \subseteq \mathcal{M}$ denote

training and evaluation environments. A feedback atom is $x = (k, f, c)$, where $M_k \in \mathcal{M}_{\text{train}}$, $f \in \{\text{demo}, \text{comp}, \text{estop}, \text{corr}\}$ denotes the modality, and c is the queried choice. A dataset $D = \{x_1, \ldots, x_N\}$ is a collection of such atoms. In the high-rationality limit ($\beta \to \infty$), corresponding to an expert teacher, each atom induces linear preference constraints $w^\top \Delta\Phi \geq 0$, where $\Delta\Phi$ is the discounted feature-count difference between preferred and dispreferred trajectories. These constraints define the feasible reward region

$$\mathcal{W}(D) = \{w \in \mathbb{R}^d \mid w^\top \Delta\Phi \geq 0,\ \forall \Delta\Phi \text{ induced by } x \in D\}.$$

We evaluate a learned reward w using the average performance gap relative to the ground-truth reward $w^\star$ across a given set of evaluation environments, $\mathcal{M}'$,

$$\text{Loss}(w^\star, w, \mathcal{M}') = \frac{1}{|\mathcal{M}'|} \sum_{k \in \mathcal{M}'} \left(V^{\pi_k(w^\star)}_{w^\star,k} - V^{\pi_k(w)}_{w^\star,k} \right) \tag{2}$$

where $\pi_k(w)$ denotes the optimal policy in M_k under reward w.

We assume an expert teacher who knows $w^\star$ and the dynamics of all $M_k \in \mathcal{M}_{\text{train}}$, but not the evaluation environments. A teaching instance is therefore specified by $\mathcal{T} = (\mathcal{M}, \mathcal{M}_{\text{train}}, \mathcal{M}_{\text{test}}, \phi, \mathbb{R}^d, w^\star, \varepsilon)$, where ε is a target loss tolerance.

We want to solve for a teaching set D that minimizes some notion of teaching cost and also minimizes loss. However, as is typical of machine teaching problems in general (Zhu et al., 2018), this is a difficult multi-objective optimization problem. Minimizing loss on an unseen set of MDPs is also very difficult without making any assumptions on these unseen test environments. To make our approach tractable, we focus on a constrained form of the machine teaching problem (Zhu et al., 2018), where we focus on minimizing loss on the known train set of MDPs. This mirrors the standard empirical risk minimization paradigm in machine learning, where performance is optimized on a training set to minimize error on unseen test instances. Under the assumption that training and evaluation environments are drawn i.i.d. from the same distribution, minimizing loss on the training MDPs provides a principled surrogate for minimizing loss on unseen environments.

The teacher selects a subset of environments $\mathcal{K} \subseteq \mathcal{M}_{\text{train}}$ and queries a dataset D whose atoms originate only from those environments. Let $\hat{w} = \text{IRL}(D)$ denote the inferred reward. We define the hierarchical teaching cost as $\text{TeachingCost}(D) = (|\text{MDPs}(D)|, |D|)$ and solve

$$\min_{D} \quad \text{TeachingCost}(D) \tag{3}$$

$$\text{s.t.} \quad \text{Loss}(w^\star, \hat{w}, \mathcal{M}_{\text{train}}) \leq \varepsilon, \tag{4}$$

$$\hat{w} = \text{IRL}(D). \tag{5}$$

The minimization is lexicographic: first minimizing the number of environments used and then the number of feedback queries.

7.2 Constraint universe

To operationalize the teaching objective, we represent feedback in terms of the reward constraints it induces. For each training MDP $M_k \in \mathcal{M}_{\text{train}}$, let $\mathcal{X}_k$ denote the finite set of candidate feedback atoms available in that environment. Each atom $x \in \mathcal{X}_k$ induces a set of discounted feature-difference vectors $\Delta\Phi(x) = \{\Delta\Phi_1, \ldots, \Delta\Phi_m\}$.

We define the *constraint universe* $\mathcal{U}$ as the set of all distinct linear preference constraints inducible from the training environments, $\mathcal{U} = \bigcup_{k \in \mathcal{M}_{\text{train}}} \bigcup_{x \in \mathcal{X}_k} \Delta\Phi(x)$, where duplicates are removed so each feature-difference direction appears only once. For each environment M_k, we similarly define $\mathcal{U}_k = \bigcup_{x \in \mathcal{X}_k} \Delta\Phi(x) \subseteq \mathcal{U}$, representing the constraint directions accessible within that environment.

This representation highlights a key structural property of the teaching problem: environment dynamics determine which reward constraints are reachable, while individual feedback atoms instantiate those constraints. This separation provides the basis for the hierarchical teaching strategy introduced next.

7.3 Hierarchical teaching strategy

The constraint representation above exposes a natural hierarchical structure in the teaching problem. This observation allows the teaching objective to be decomposed into two coupled decisions: selecting informative environments and selecting feedback instances within those environments.

Outer level (environment selection). The first stage selects a minimal subset of environments $\mathcal{K} \subseteq \mathcal{M}_{\text{train}}$ such that $\bigcup_{k \in \mathcal{K}} \mathcal{U}_k = \mathcal{U}$, identifying environments whose trajectory spaces collectively expose all reward-relevant constraint directions.

Inner level (atom selection). Given $\mathcal{K}$, the second stage selects a minimal dataset D satisfying $\bigcup_{x \in D} \Delta\Phi(x) = \mathcal{U}$ with $\text{MDPs}(D) \subseteq \mathcal{K}$, determining the feedback instances required to realize those constraints.

Greedy hierarchical approximation (HSCOT). Both stages correspond to set-cover problems over the finite constraint universe $\mathcal{U}$. We therefore take advantage of submodularity (Nemhauser et al., 1978) and employ greedy selection procedures that iteratively choose the environment or feedback atom providing the largest marginal increase in uncovered constraint directions. The resulting algorithm, HSCOT, separates environment informativeness from feedback efficiency while maintaining a tractable approximation to the original teaching objective. The full pseudocode for the outer-stage environment selection, the inner-stage atom selection, and the combined hierarchical procedure appears in Algorithms 1–3.

After the teaching algorithm constructing D, the learner estimates the reward $\hat{w} = \text{IRL}(D)$ and evaluates the resulting policy using the loss in Equation (2).

8 Experimental results

We evaluate whether hierarchical multi-environment teaching (HSCOT) improves reward identifiability and policy generalization under limited feedback budgets.

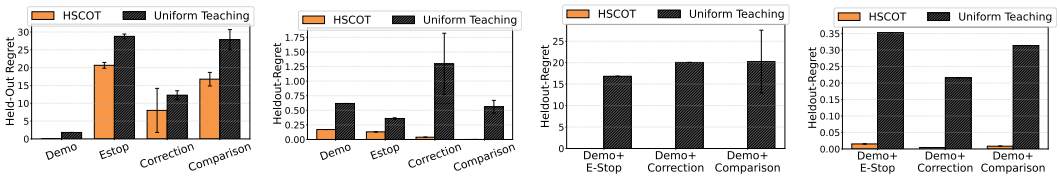
Domains and setup. We evaluate on two domains with deterministic transitions and fixed discount factor. In a 6×6 GridWorld, we generate 50 MDPs sharing the same linear reward $w^\star$ (normalized so $\|w^\star\|_2 \leq 1$), but differing in layout and transition structure; feature vectors are sampled per environment.In LavaMiniGrid (Chevalier-Boisvert et al., 2023), we use four hand-designed features (*distance-to-goal*, *on-lava*, *adjacent-to-lava*, *per-step cost*) and generate 50 MDPs with the same ground-truth reward but different layouts. For generalization, 20% of MDPs are held out and never seen during teaching.

Baseline and Reward Learning. We compare HSCOT against a *Uniform Teaching* baseline that samples feedback atoms uniformly across environments under the same global query budget as HSCOT. Reward recovery uses max-margin inverse reinforcement learning (Abbeel and Ng, 2004).

Evaluation metrics. We report *held-out regret* using the average performance gap (Eq. (2)) between the policy induced by the learned reward $\hat{w}$ and the ground-truth optimal policy, averaged over held-out MDPs. We also report *constraint coverage*, the fraction of the universal constraint set $\mathcal{U}$ covered by the selected feedback atoms D:

$$\text{coverage}(D) = \frac{|\mathcal{U}(D)|}{|\mathcal{U}|},$$

where $\mathcal{U}$ is the set of all distinct linear preference constraints $w^\top(\Phi(\xi^+) - \Phi(\xi^-)) \geq 0$ inducible by any feedback atom and modality across all training MDPs, and $\mathcal{U}(D)$ is the subset induced by D. Coverage serves as a proxy for reduction in reward ambiguity (shrinkage of the generalized behavioral equivalence class).



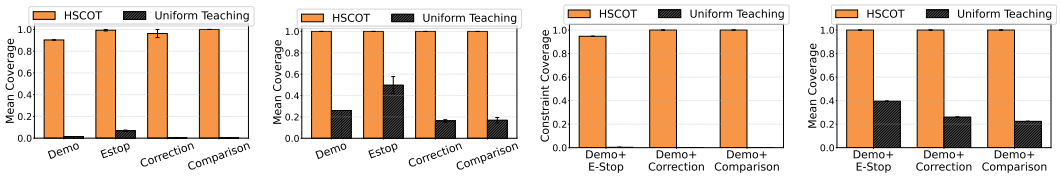
(a) GridWorld (b) LavaMiniGrid (c) GridWorld (d) LavaMiniGrid

Figure 5: **Held-out regret** (lower is better) averaged over 10 random seeds on 20% held-out MDPs. Bars show mean regret and error bars denote standard error across seeds. Subfigures (a,b) show results for single-modality feedback. Subfigures (c,d) show settings where demonstrations are combined with an additional modality. HSCOT (solid bars) consistently outperforms uniform teaching (dashed bars) across both domains and feedback settings. Overall, rewards learned by HSCOT generalize better to unseen environments.

Held-out regret. Single-modality results are shown in Figures 5a and 5b. The regret is evaluated on previously unseen environments whose dynamics were not observed during teaching, so even rewards that perform optimally on the training MDPs may incur small performance gaps when transferred to new layouts. In GridWorld (Figure 5a), demonstration feedback alone is sometimes sufficient to recover a reward that generalizes perfectly, resulting in near-zero regret, whereas the other modalities still produce non-zero regret on unseen environments. In LavaMiniGrid (Figure 5b), environment variation is stronger and even demonstrations leave residual regret, indicating the increased difficulty of identifying rewards that generalize to unseen environment layouts. Mixed-feedback results are shown in Figures 5c and 5d. When demonstrations are combined with another modality, regret improves substantially and approaches zero in both domains. This improvement is consistent with the limited-budget analysis: demonstrations implicitly introduce multiple optimality constraints, and combining them with other feedback types exposes complementary constraint directions across environments, tightening the feasible reward region and enabling HSCOT to recover rewards that transfer almost perfectly to unseen MDPs.

Constraint coverage. Figure 6 shows the fraction of the universal constraint set covered by the selected feedback atoms under a fixed global query budget. HSCOT's two-level greedy approach (environment selection followed by atom selection) consistently achieves near-complete coverage in both GridWorld and LavaMiniGrid across almost all feedback modalities. In contrast, uniform teaching recovers only a small fraction of the universal set—even when using demonstrations, where each individual demonstration implicitly imposes many constraints. This large and consistent gap demonstrates two important points: (1) HSCOT can recover nearly all constraints that can be generated across the training MDPs, and (2) intelligent selection of environments remains critical for constraint coverage, even when each piece of feedback is highly informative on its own. Uniform sampling, by spreading queries thinly across uninformative MDPs, misses most of the complementary constraints needed for tight reward identification.

Selective environment activation. Table 1 reports the average number of distinct environments activated during teaching. HSCOT consistently uses fewer environments than uniform teaching under the same global budget, reflecting its outer-stage greedy selection of MDPs that contribute maximal marginal constraint coverage. The results also highlight the necessity of multi-environment teaching under limited feedback. Each MDP exposes only a subset of the global constraint universe due to its specific transition dynamics, so no single environment suffices to eliminate reward ambiguity. HSCOT selects a small but informative subset whose joint constraints approximate the full intersection over all training MDPs. Environment counts are highest for E-stop feedback, consistent with its weaker and more local constraints, and lowest for comparison feedback, which imposes stronger global ordering constraints.



(a) GridWorld (b) LavaMiniGrid (c) GridWorld (d) LavaMiniGrid

Figure 6: **Constraint coverage** (higher is better) averaged over 10 random seeds. Bars show the fraction of the universal constraint set covered, and error bars denote standard error across seeds. Subfigures (a,b) correspond to single-modality feedback, while (c,d) show mixed-feedback settings where demonstrations are combined with another modality. HSCOT (solid bars) achieves near-complete coverage in both domains across all settings, whereas uniform teaching (dashed bars) covers only a small fraction of the constraint universe.

Table 1: Average number of environments activated (out of 50 training MDPs) over 10 random seeds. Lower is better.

Feedback modality	GridWorld		LavaMiniGrid		Aggregate	
	HSCOT	Uniform	HSCOT	Uniform	HSCOT	Uniform
Demonstration	4.60	5.05	5.00	5.99	**4.80**	5.52
Correction	4.95	5.21	6.55	6.68	**5.75**	5.95
Comparison	2.55	2.66	6.91	6.87	**4.73**	4.77
E-stop	8.44	10.83	13.77	17.50	**11.11**	14.17

9 Conclusion

We studied how to teach reward functions that remain robust across diverse environment dynamics. Single-MDP teaching frequently entangles reward information with transition structure, limiting generalization when dynamics change. We first showed, through constraint geometry and formal comparison, that feedback modalities differ markedly in informativeness. In the unlimited-data regime, comparison impose the strictest global ordering constraints and yield the smallest feasible reward region, strictly outperforming demonstrations, corrections, and E-stops. Under tight budgets, however, demonstrations become more efficient per query due to their implicit multi-constraint nature. We further proved that reward identifiability is inherently environment-dependent: even idealized unlimited feedback in one MDP generally leaves residual ambiguity in others unless their combined dynamics expose complementary feature-difference directions. To address these limitations, we introduced HSCOT, a greedy two-stage algorithm that first selects informative environments and then strategically queries heterogeneous feedback within them. In GridWorld and LavaMiniGrid experiments with held-out evaluation MDPs, HSCOT consistently achieved substantially lower regret, near-complete constraint coverage, and more efficient environment utilization than uniform sampling under identical global query budgets. These results underscore that robust reward learning depends less on collecting more feedback and more on deliberately choosing *which environments* and *which queries* expose the most informative constraints on the true objective.

References

Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1, 2004.

Frank J. Balbach and Thomas Zeugmann. Recent developments in algorithmic teaching. In Adrian Horia Dediu, Armand Mihai Ionescu, and Carlos Martín-Vide, editors, *Language and Automata Theory and Applications*, volume 5457 of *Lecture Notes in Computer Science*, pages

1–18. Springer, Berlin, Heidelberg, 2009. doi: 10.1007/978-3-642-00982-2_1. URL https://doi.org/10.1007/978-3-642-00982-2_1.

Serena Booth, W Bradley Knox, Julie Shah, Scott Niekum, Peter Stone, and Alessandro Allievi. The perils of trial-and-error reward design: misdesign through overfitting and invalid task specifications. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5920–5929, 2023.

Daniel Brown, Wonjoon Goo, Prabhat Nagarajan, and Scott Niekum. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. In *International conference on machine learning*, pages 783–792. PMLR, 2019.

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 7749–7756. AAAI Press, 2019. URL https://ojs.aaai.org/index.php/AAAI/article/view/4771.

Thomas Kleine Büning, Anne-Marie George, and Christos Dimitrakakis. Interactive inverse reinforcement learning for cooperative games. In *International Conference on Machine Learning*, pages 2393–2413. PMLR, 2022.

Maya Cakmak and Manuel Lopes. Algorithmic and human teaching of sequential decision tasks. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, AAAI'12, page 1536–1542. AAAI Press, 2012.

Maxime Chevalier-Boisvert, Bolun Dai, Mark Towers, Rodrigo Perez-Vicente, Lucas Willems, Salem Lahlou, Suman Pal, Pablo Samuel Castro, and Jordan Terry. Minigrid & miniworld: Modular & customizable reinforcement learning environments for goal-oriented tasks. *Advances in Neural Information Processing Systems*, 36:73383–73394, 2023.

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.

Rati Devidze, Farnam Mansouri, Luis Haug, Yuxin Chen, and Adish Singla. Understanding the power and limitations of teaching with imperfect knowledge. *arXiv preprint arXiv:2003.09712*, 2020.

Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adversarial inverse reinforcement learning. *arXiv preprint arXiv:1710.11248*, 2017.

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.

Luis Haug, Sebastian Tschiatschek, and Adish Singla. Teaching inverse reinforcement learners via features and demonstrations. *Advances in Neural Information Processing Systems*, 31, 2018.

Jerry Zhi-Yang He and Anca D Dragan. Assisted robust reward design. *arXiv preprint arXiv:2111.09884*, 2021.

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.

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

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. A survey of reinforcement learning from human feedback. *Transactions on Machine Learning Research*, 2024.

Weiyang Liu, Bo Dai, Ahmad Humayun, Charlene Tay, Chen Yu, Linda B Smith, James M Rehg, and Le Song. Iterative machine teaching. In *International Conference on Machine Learning*, pages 2149–2158. PMLR, 2017.

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.

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.

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.

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.

George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical programming*, 14(1):265–294, 1978.

Tomi Peltola, Mustafa Mert Çelikok, Pedram Daee, and Samuel Kaski. Machine teaching of active sequential learners. *Advances in neural information processing systems*, 32, 2019.

Sebastian Tschiatschek, Ahana Ghosh, Luis Haug, Rati Devidze, and Adish Singla. Learner-aware teaching: Inverse reinforcement learning with preferences and constraints. *Advances in neural information processing systems*, 32, 2019.

Chaoqi Wang, Adish Singla, and Yuxin Chen. Teaching an active learner with contrastive examples. *Advances in Neural Information Processing Systems*, 34:17968–17980, 2021.

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.

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.

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.

Xuezhou Zhang, Shubham Bharti, Yuzhe Ma, Adish Singla, and Xiaojin Zhu. The sample complexity of teaching by reinforcement on q-learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10939–10947, 2021.

Xiaojin Zhu. Machine teaching: an inverse problem to machine learning and an approach toward optimal education. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 4083–4087. AAAI Press, 2015. ISBN 0262511290.

Xiaojin Zhu, Adish Singla, Sandra Zilles, and Anna N Rafferty. An overview of machine teaching. *arXiv preprint arXiv:1801.05927*, 2018.

Supplementary Materials

The following content was not necessarily subject to peer review.

10 Appendix

10.1 Proofs

Proposition 4 (Comparison strictly reduce reward ambiguity compared to demonstrations)**.** *Let $\Xi^\star \subseteq \Xi$ denote the trajectories that are optimal under $w^\star$. Demonstrations induce the feasible region*

$$H_{demo} = \Big\{ w \in \mathbb{R}^d : \forall \xi^\star \in \Xi^\star, \forall \xi' \in \Xi \text{ with the same initial state as } \xi^\star,\ w^\top(\Phi(\xi^\star) - \Phi(\xi')) \geq 0 \Big\}.$$

Then

$$H_{comparison} \subsetneq H_{demo}, \qquad G(H_{comparison}) < G(H_{demo}).$$

Proof. **Step 1: Inclusion $H_{\text{comparison}} \subseteq H_{\text{demo}}$.**

Demonstrations reveal trajectories that are optimal under the true reward $w^\star$. Thus for any $\xi^\star \in \Xi^\star$ and any trajectory ξ' with the same initial state,

$$w^{\star\top}\Phi(\xi^\star) \geq w^{\star\top}\Phi(\xi').$$

Each demonstration constraint therefore has the form

$$w^\top(\Phi(\xi^\star) - \Phi(\xi')) \geq 0.$$

This is a special case of the comparison ordering constraints defining $H_{\text{comparison}}$, since comparisons compare arbitrary trajectories $\xi_i, \xi_j \in \Xi$.

Therefore any w satisfying all comparison constraints automatically satisfies all demonstration constraints, implying

$$H_{\text{comparison}} \subseteq H_{\text{demo}}.$$

Step 2: Strict inclusion.

To show the inclusion is strict, it suffices to exhibit a reward vector $w \in H_{\text{demo}} \setminus H_{\text{comparison}}$.

Consider two suboptimal trajectories $\xi_a, \xi_b \in \Xi \setminus \Xi^\star$ such that

$$w^{\star\top}\Phi(\xi_a) > w^{\star\top}\Phi(\xi_b).$$

Let

$$v := \Phi(\xi_a) - \Phi(\xi_b).$$

Demonstration constraints compare optimal trajectories with alternatives from the same initial state, but do not constrain the relative ordering between suboptimal trajectories such as ξ_a and ξ_b. Thus the sign of $w^\top v$ is not determined by the demonstration constraints.

Consequently there exists a reward vector $w \in H_{\text{demo}}$ such that

$$w^\top v < 0.$$

Such a vector satisfies all demonstration constraints but violates the comparison ordering $\xi_a \succ \xi_b$, implying $w \notin H_{\text{comparison}}$.

Therefore

$$H_{\text{comparison}} \subsetneq H_{\text{demo}}.$$

Step 3: Volume comparison.

Since $H_{\text{comparison}}$ is obtained from H_{demo} by adding additional linear inequality constraints, and both sets are compact convex subsets of the unit sphere (after $\|w\|_2 = 1$ normalization), their volumes satisfy

$$G(H_{\text{comparison}}) < G(H_{\text{demo}}).$$

This completes the proof. □

Proposition 5 (Comparison strictly reduce reward ambiguity compared to corrective feedback)**.** *Corrective feedback induces the feasible region*

$$H_{\textit{correction}} = \Big\{ w \in \mathbb{R}^d : \forall \xi_{\text{corr}}, \xi_R \in \Xi \textit{ with the same initial state},\ w^\top (\Phi(\xi_{\text{corr}}) - \Phi(\xi_R)) \geq 0 \Big\}.$$

Then

$$H_{\textit{comparison}} \subsetneq H_{\textit{correction}}, \qquad G(H_{\textit{comparison}}) < G(H_{\textit{correction}}).$$

Proof. **Step 1: Inclusion $H_{\text{comparison}} \subseteq H_{\text{correction}}$.**

Each corrective feedback instance compares two trajectories ξ_{corr} and ξ_R that share the same initial state, imposing

$$w^\top (\Phi(\xi_{\text{corr}}) - \Phi(\xi_R)) \geq 0.$$

This is a special case of a comparison constraint. Hence every correction constraint appears among the comparison constraints, implying

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

Step 2: Strict inclusion.

Consider trajectories $\xi_a, \xi_b \in \Xi$ with different initial states such that

$$w^{\star\top} \Phi(\xi_a) > w^{\star\top} \Phi(\xi_b).$$

The corresponding comparison constraint is

$$w^\top (\Phi(\xi_a) - \Phi(\xi_b)) \geq 0.$$

However, correction constraints only compare trajectories that start from the same state. Therefore the ordering between ξ_a and ξ_b is not enforced by correction constraints.

Thus there exists $w \in H_{\text{correction}}$ such that

$$w^\top(\Phi(\xi_a) - \Phi(\xi_b)) < 0.$$

Hence $w \notin H_{\text{comparison}}$ and

$$H_{\text{comparison}} \subsetneq H_{\text{correction}}.$$

Step 3: Volume comparison.

Since $H_{\text{comparison}}$ adds additional linear constraints to $H_{\text{correction}}$, their volumes satisfy

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

□

Lemma 2 (E-stop constraints are trajectory-local)**.** *Fix a trajectory $\xi \in \Xi$. Any E-stop constraint comparing $\xi^{0:t}$ with ξ lies in the subspace*

$$\text{span}\{\phi(s,a) \mid (s,a) \textit{ appears in } \xi\}.$$

Comparison between trajectories ξ and ξ' may produce feature differences outside this trajectory-specific subspace.

Proof. Let

$$\xi = (s_0, a_0, s_1, a_1, \ldots, s_T, a_T)$$

be a trajectory.

For any stopping time $t < T$, the halted trajectory is $\xi^{0:t}$.

The discounted feature sums are

$$\Phi(\xi) = \sum_{\tau=0}^{T} \gamma^\tau \phi(s_\tau, a_\tau)$$

and

$$\Phi(\xi^{0:t}) = \sum_{\tau=0}^{t} \gamma^\tau \phi(s_\tau, a_\tau) + \sum_{\tau=t+1}^{\infty} \gamma^\tau \phi(s_t, a_t).$$

The difference vector is

$$\Phi(\xi^{0:t}) - \Phi(\xi) = \frac{\gamma^{t+1}}{1-\gamma}\phi(s_t, a_t) - \sum_{\tau=t+1}^{T} \gamma^\tau \phi(s_\tau, a_\tau).$$

All terms involve only feature vectors $\phi(s_\tau, a_\tau)$ for state-action pairs appearing in ξ.

Therefore

$$\Phi(\xi^{0:t}) - \Phi(\xi) \in \text{span}\{\phi(s,a) \mid (s,a) \text{ appears in } \xi\}.$$

Now consider another trajectory ξ' that visits a state-action pair (s', a') whose feature vector is linearly independent of this span. Then the comparison difference

$$\Phi(\xi) - \Phi(\xi')$$

contains a component outside the trajectory-specific subspace, which cannot be generated by E-stop constraints along ξ alone.

□

Proposition 6 (Comparison strictly reduce reward ambiguity compared to E-stop feedback)**.** *E-stop feedback induces the feasible region*

$$H_{\textit{E-stop}} = \Big\{ w \in \mathbb{R}^d : \|w\|_2 \le 1,\ \forall \xi \in \Xi,\ \forall t,\ w^\top (\Phi(\xi^{0:t}) - \Phi(\xi)) \ge 0 \Big\}.$$

Then

$$H_{\textit{comparison}} \subsetneq H_{\textit{E-stop}}, \qquad G(H_{\textit{comparison}}) < G(H_{\textit{E-stop}}).$$

Proof. **Step 1: Inclusion.**

Each E-stop constraint compares two trajectories $\xi^{0:t}$ and ξ. This is a special case of a comparison Thus every E-stop constraint appears among the comparison constraints, implying

$$H_{\text{comparison}} \subseteq H_{\text{E-stop}}.$$

Step 2: Strict inclusion.

By Lemma 2, E-stop constraints along a trajectory ξ only constrain reward vectors in the subspace spanned by features appearing in ξ.

Comparisons between trajectories ξ and ξ' can introduce feature differences outside this subspace.

Hence there exist reward vectors satisfying all E-stop constraints but violating a comparison ordering.

Therefore

$$H_{\text{comparison}} \subsetneq H_{\text{E-stop}}.$$

Step 3: Volume comparison.

Since $H_{\text{comparison}}$ adds additional constraints to $H_{\text{E-stop}}$, their volumes satisfy

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

□

Theorem 2 (Single-MDP ambiguity and necessity of multi-MDP teaching)**.** *Let $\mathcal{M} = \{M_1, M_2\}$ be two finite MDPs that share a feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ and linear rewards $r_w(s,a) = w^\top \phi(s,a)$ with $w \in \mathbb{R}^d \subset \mathbb{R}^d$. Let $w^\star$ denote the ground-truth reward parameter.*

For any MDP M_k, let $\mathcal{W}(M_k)$ denote the generalized behavioral equivalence class (gBEC), i.e., the set of reward parameters consistent with all possible *idealized feedback constraints obtainable within M_k.*

Define the feature-difference span induced by M_k as

$$\mathcal{V}_k = \mathrm{span}\{\Delta\Phi(\xi^+, \xi^-) \ : \ \xi^+, \xi^- \textit{ are feasible trajectories in } M_k\},$$

where $\Delta\Phi(\xi^+, \xi^-) = \Phi(\xi^+) - \Phi(\xi^-)$ and $\Phi(\xi)$ denotes the feature expectation of trajectory ξ.

Assume that the induced spans satisfy

$$\mathcal{V}_1 \subsetneq \mathrm{span}(\mathcal{V}_1 \cup \mathcal{V}_2).$$

Then there exists a reward parameter $w \neq w^\star$ such that

$$w \in \mathcal{W}(M_1) \quad \textit{but} \quad w \notin \mathcal{W}(M_1) \cap \mathcal{W}(M_2).$$

In particular, no amount of feedback obtained solely from M_1 is sufficient to uniquely identify $w^\star$.

Proof. Under idealized feedback, every constraint obtained from M_1 can be written as a linear inequality

$$w^\top \Delta\Phi \geq 0, \qquad \Delta\Phi \in \mathcal{V}_1,$$

which defines a closed convex cone in parameter space. Hence the gBEC induced by M_1 is

$$\mathcal{W}(M_1) = \left\{w \in \mathbb{R}^d : \ w^\top \Delta\Phi \geq 0 \ \ \forall \Delta\Phi \in \mathcal{V}_1\right\}.$$

Because $\mathcal{V}_1 \subsetneq \mathrm{span}(\mathcal{V}_1 \cup \mathcal{V}_2)$, there exists a direction $u \in \mathbb{R}^d$ such that

$$u \perp \mathcal{V}_1 \qquad \text{but} \qquad u \not\perp \mathcal{V}_2.$$

Define

$$w = \frac{w^\star + \epsilon u}{\|w^\star + \epsilon u\|_2}$$

for sufficiently small $\epsilon > 0$. By orthogonality, for every $\Delta\Phi \in \mathcal{V}_1$ we have

$$w^\top \Delta\Phi = (w^\star)^\top \Delta\Phi,$$

so w satisfies all constraints induced by M_1. Therefore $w \in \mathcal{W}(M_1)$.

However, since u has nonzero projection onto $\mathcal{V}_2$, there exists some $\Delta\Phi' \in \mathcal{V}_2$ for which

$$w^\top \Delta\Phi' \neq (w^\star)^\top \Delta\Phi'.$$

For sufficiently small but nonzero ϵ, the sign of the inner product can be reversed relative to $w^\star$, violating at least one constraint obtainable in M_2. Thus $w \notin \mathcal{W}(M_2)$, implying

$$w \notin \mathcal{W}(M_1) \cap \mathcal{W}(M_2).$$

Hence the reward parameter cannot be uniquely identified using feedback from M_1 alone, completing the proof. □

10.2 Supplementary Algorithms

This section provides the full pseudocode for the hierarchical set-cover teaching procedure used in the main text. The algorithms correspond to the two-stage greedy approximation described in Section 7.3.

Algorithm S1: Greedy Environment Selection (Outer Stage)

Algorithm 1 Greedy Environment Selection

Require: Per-MDP constraint coverage sets $\{\mathcal{U}_k\}_{k\in\mathcal{M}_{\text{train}}}$, universe $\mathcal{U}$
Ensure: Ordered list of selected MDP indices $\mathcal{K}$
covered $\leftarrow \emptyset$
$\mathcal{K} \leftarrow [\,]$
while covered $\neq \mathcal{U}$ **do**
    Select k maximizing $|\mathcal{U}_k \setminus \text{covered}|$
    $\mathcal{K} \leftarrow \mathcal{K} \cup \{k\}$
    covered $\leftarrow$ covered $\cup\, (\mathcal{U}_k \setminus \text{covered})$
end while
return $\mathcal{K}$

Algorithm S2: Greedy Atom Selection (Inner Stage)

Algorithm 2 Greedy Atom Selection

Require: Selected MDPs $\mathcal{K}$, candidate atoms and coverage sets within $\mathcal{K}$, universe $\mathcal{U}$
Ensure: Ordered list of chosen atoms D
covered $\leftarrow \emptyset$
$D \leftarrow [\,]$
while covered $\neq \mathcal{U}$ **do**
    Select atom x from any $k \in \mathcal{K}$ maximizing $|\text{coverage}(x) \setminus \text{covered}|$
    $D \leftarrow D \cup \{x\}$
    covered $\leftarrow$ covered $\cup\, (\text{coverage}(x) \setminus \text{covered})$
end while
return D

Algorithm S3: Hierarchical SCOT (HSCOT)

Algorithm 3 Hierarchical SCOT (HSCOT)

Require: Constraint universe $\mathcal{U}$, per-MDP inducible constraints $\{\mathcal{U}_k\}$, candidate atoms per MDP
Ensure: Selected environments $\mathcal{K}$ and feedback dataset D
$\mathcal{K} \leftarrow$ Greedy Environment Selection$(\{\mathcal{U}_k\}, \mathcal{U})$ (Alg. S1)
Restrict candidate atoms to those in environments $\mathcal{K}$
$D \leftarrow$ Greedy Atom Selection$(\mathcal{K}, \mathcal{U})$ (Alg. S2)
return $(\mathcal{K}, D)$