

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

Ali Larian, Qian Lin, Chang Zong Wu, Daniel S. Brown

Keywords: Learning from Human Feedback, 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 environment. Inverse reinforcement learning (IRL) provides a principled way to infer such objectives from human feedback. However, existing analyses 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 reward learning 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 machine teaching of reward functions 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 Inverse RL agents (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) only consider single-environment teaching settings.

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

Ali Larian¹, Qian Lin¹, Chang Zong Wu¹, Daniel S. Brown¹

{ali.larian, qian.lin, u1440478, daniel.s.brown}@utah.edu

¹Kahlert School of Computing, University of Utah.

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 analyses 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 reward learning 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 use heterogeneous feedback to teach an agent so that its behavior remains aligned with the human’s intent across multiple operating conditions?*

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 reward learning, 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 reward function learning (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 reward learning 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 reward learning.** We examine how different feedback types constrain reward identification in machine teaching for reward learning. In the unlimited-data regime, comparisons impose the strongest global constraints; however, under limited budgets, demonstrations are more constraint-efficient per query.
- **Environment-dependent reward identifiability.** We study how reward identifiability depends on environment dynamics and formally characterize how 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.

Our implementation and experiment code are available online.¹

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 reward learning across multiple environments with heterogeneous feedback. The most closely related work studies machine teaching for Inverse Reinforcement Learning (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

¹<https://github.com/Alilarian/multienv-reward-teaching>

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 (Tschitschek 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. Brown et al. (2021) studied the complementary problem of machine testing: how to check whether a learned reward is aligned across multiple environments. To our knowledge, we are the first to study machine teaching for reward learning 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). However, 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). An orthogonal line of work, meta-IRL (Yu et al., 2019; Chen et al., 2024), learns priors for fast task-specific reward learning, whereas we aim to teach a single reward that generalizes across environments without further fine-tuning. 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 either teaches within a single MDP or infers rewards from heterogeneous feedback without structuring the teaching itself. We bridge these directions with a teacher-centric, multi-MDP framework that unifies feedback modalities through generalized behavioral equivalence classes.

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, $T(s' | s, a)$ the transition function, and $\gamma \in (0, 1)$ the discount factor. We assume a linear reward function defined by a feature map $\phi : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}^d$, so that $R_w(s, a) = w^\top \phi(s, a)$, where $w \in \mathbb{R}^d$. For any policy π , we denote the discounted feature counts as $\mu^\pi(s) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t \phi(s_t, a_t) | s_0 = s]$ and $\mu^\pi(s, a) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t \phi(s_t, a_t) | s_0 = s, a_0 = a]$. 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 .

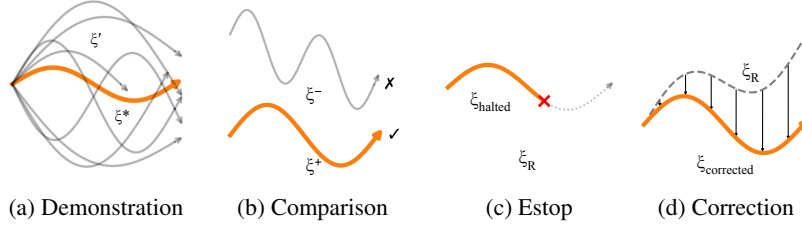


Figure 1: **Human feedback modalities as trajectory comparisons.** Orange trajectories denote the preferred outcome. (a) Demonstration: expert trajectory ξ^* preferred over alternatives. (b) Comparison: preference $\xi^+ \succ \xi^-$. (c) E-stop: halted trajectory $\xi_{\text{halted}} \succ \xi_R$. (d)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

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 $\psi : \mathcal{C} \rightarrow \Xi$. For any trajectory $\xi = (s_0, a_0, s_1, a_1, \dots)$, define its discounted feature vector $\Phi(\xi) = \sum_{t=0}^{\infty} \gamma^t \phi(s_t, a_t)$. Given reward weights w and rationality parameter $\beta \geq 0$, the likelihood of observing choice c is $P(c \mid w, \beta) \propto \exp(\beta r_w(\psi(c)))$, 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 $\beta \rightarrow \infty$, 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,

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

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

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

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 $\beta \rightarrow \infty$, 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 regime 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

$$\text{BEC}(\pi) = \{w \in \mathbb{R}^d \mid w^\top (\mu^\pi(s, \pi(s)) - \mu^\pi(s, b)) \geq 0, \forall s \in \mathcal{S}, \forall b \in \mathcal{A}\},$$

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

$$\text{gBEC}(D) = \{w \in \mathbb{R}^d \mid w^\top \Delta\Phi \geq 0, \forall \Delta\Phi \in \Delta\Phi(D)\}. \quad (1)$$

Connection to BECs. When D consists of optimal demonstrations from a policy π that visits all states, $\text{gBEC}(D)$ coincides with the classical $\text{BEC}(\pi)$, 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 $\text{gBEC}(D)$. We quantify reward ambiguity by its volume

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

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 $\text{gBEC}(D)$ 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 of the MDP in Figure 4a under tight budgets as heatmaps over reward weight space, where darker areas indicate weights that remain consistent across many randomly sampled feedback sets. Here every modality uses the same fixed budget of five queries. Demonstrations (Fig. 2c) produce the smallest $\text{gBEC}(D)$ and lowest ambiguity $G(D)$. Each demonstration implicitly enforces many optimality constraints, yielding strong shrinkage of the feasible region. 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 $\text{gBEC}(D)$ 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. A budget-matched sweep isolating per-query efficiency via feasibility heatmaps and feasible-region volume curves is provided in Appendix 10.2.

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 $\text{gBEC}(D)$ 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

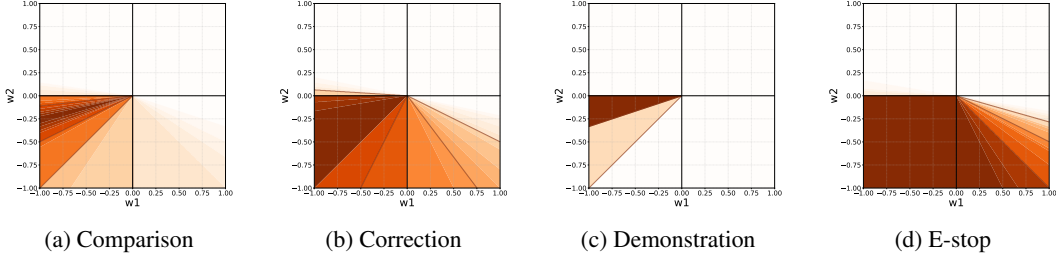


Figure 2: Finite-data feasible reward regions under different types of feedback. Heatmaps visualize the empirical feasibility probability over the reward space—darker colors indicate weight vectors that remain feasible across many randomly sampled feedback sets. The heatmap illustrates how different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

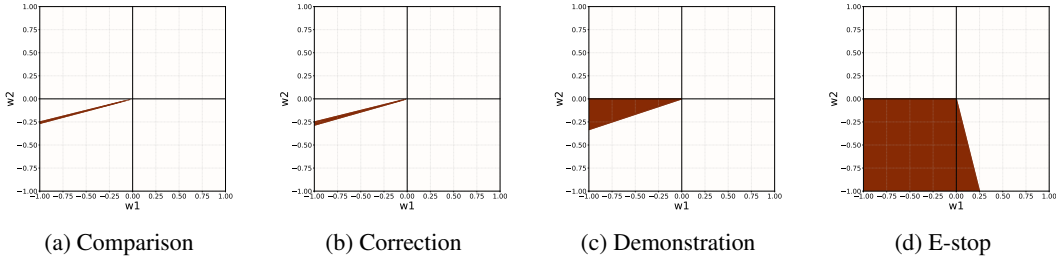


Figure 3: Feasible reward regions $\text{gBEC}(D)$ for different feedback modalities. Subfigures (a)–(d) show how infinite comparisons, corrections, demonstrations, and E-stops, respectively, shape the reward feasibility region. The black dot indicates the ground-truth reward parameter (corresponding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

key teaching trade-off. Different feedback modalities influence the geometry of $\text{gBEC}(D)$ 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.

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 comparisons as a reference modality. The comparison modality compares 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^* induces a full ordering over Ξ and yields the feasible region

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

We compare this region to those induced by demonstrations, corrections, 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 feedback, in contrast, can rank any two trajectories in Ξ , imposing ordering constraints across the entire trajectory space.

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

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

Then

$$H_{comparison} \subsetneq H_{demo}, \quad \text{and} \quad G(H_{comparison}) < G(H_{demo}).$$

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

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

Then

$$H_{comparison} \subsetneq H_{correction}, \quad \text{and} \quad G(H_{comparison}) < G(H_{correction}).$$

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*

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

In contrast, a pairwise comparison between ξ and any $\xi' \in \Xi$ that visits a state-action pair (s', a') with $\phi(s', a')$ linearly independent of this subspace produces a feature difference that lies outside it.

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

$$H_{E-stop} = \left\{ w \in \mathbb{R}^d : w^\top (\Phi(\xi^{0:t}) - \Phi(\xi)) \geq 0 \ \forall \xi \in \Xi, \ \forall t \right\}.$$

Then

$$H_{comparison} \subsetneq H_{E-stop}, \quad \text{and} \quad G(H_{comparison}) < G(H_{E-stop}).$$

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 common feature map $\phi : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}^d$ and a common ground-truth reward parameter $w^* \in \mathbb{R}^d$, where rewards take the linear form $r(s, a) = w^{*\top} \phi(s, a)$. 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*

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

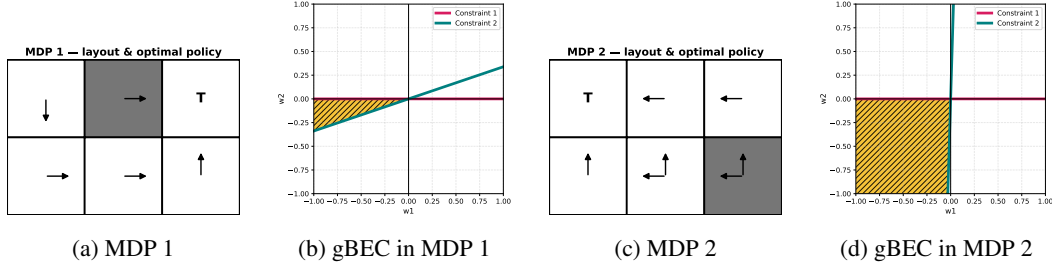


Figure 4: **Environment-dependent reward ambiguity.** Two gridworld MDPs share the same reward features but differ in layout; each MDP has two features, shown as white and gray cells, and **T** marks the terminal state. Arrows in (a) and (c) show an optimal policy. The feasible reward region in MDP 2 is strictly larger than in MDP 1 ($\text{gBEC}_1 \subsetneq \text{gBEC}_2$), illustrating how environment structure affects reward identifiability.

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 $w \neq w^*$ such that $w \in \mathcal{W}(M_1)$ but $w \notin \mathcal{W}(M_1) \cap \mathcal{W}(M_2)$.

The proof is provided in Appendix 10.1.

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^* , but differ in layout, cell arrangement, and terminal state position (see Figures 4a and 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: $\text{gBEC}_1 \subsetneq \text{gBEC}_2$. Every reward consistent with all optimal demonstrations in MDP 1 remains feasible in MDP 2 (but not vice versa). Teaching only in M_2 leaves reward ambiguity that can cause generalization failures to M_1 , while teaching in M_1 recovers a reward that transfers across both environments. This asymmetry stems purely from differences in the environment layout. Appendix 10.3 confirms this layout-dependent ambiguity across 200 randomly sampled layouts.

These observations reveal that single-MDP teaching is frequently insufficient for reward generalization across environments, motivating the use of diverse training environments. 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 $\mathcal{M} = \{M_1, \dots, M_K\}$ be a family of MDPs sharing a feature map $\phi : \bigcup_k (\mathcal{S}_k \times \mathcal{A}_k) \rightarrow \mathbb{R}^d$, 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 $D = \{x_1, \dots, x_N\}$ is a collection of such atoms. In the high-rationality limit ($\beta \rightarrow \infty$), 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^* across a given set of evaluation environments, $\mathcal{M}'$. Let $V_k^\pi(w) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r_w(s_t, a_t) \mid \pi, M_k]$ denote the expected discounted return of policy π in M_k under reward parameter w . We then define

$$\text{Loss}(w^*, w, \mathcal{M}') = \frac{1}{|\mathcal{M}'|} \sum_{k \in \mathcal{M}'} \left(V_k^{\pi_k^*(w)}(w^*) - V_k^{\pi_k(w)}(w^*) \right) \quad (2)$$

where $\pi_k^*(w)$ denotes the optimal policy in M_k under reward w . We assume an expert teacher who knows w^* and the dynamics of all $M_k \in \mathcal{M}_{\text{train}}$, but not the evaluation environments. A teaching instance is therefore specified by $\mathcal{T} = (\mathcal{M}, \mathcal{M}_{\text{train}}, \mathcal{M}_{\text{test}}, \phi, \mathbb{R}^d, w^*, \varepsilon)$, 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 in Eq. (2). 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 aims to find a minimal subset of environments $\mathcal{K} \subseteq \mathcal{M}_{\text{train}}$ and, within each selected environment, provide a minimal set of feedback instances such that the inferred reward generalizes across $\mathcal{M}_{\text{train}}$.

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

$$\min_D \text{TeachingCost}(D) \quad (3)$$

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

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

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 $\Delta\Phi(x) = \{\Delta\Phi_1, \dots, \Delta\Phi_m\}$.

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, decomposes teaching into two stages: an outer stage that selects environments by their marginal constraint coverage, and an inner stage that selects feedback atoms within those environments. 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 teaching, the learner estimates the reward $\hat{w} = \text{IRL}(D)$ and deploys the resulting policy; we then evaluate generalization performance 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^* (normalized so $\|w^*\|_2 \leq 1$), 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*) with corresponding ground-truth weights $w^* = [-1.0, -8.0, -2.0, -0.05]$, and generate 50 MDPs with the same ground-truth reward but different layouts (see Appendix 10.5 for visualizations of the environments). For generalization, 20% of MDPs (10 of 50) are held out and never seen during teaching. All results are averaged over 10 independent seeds, where each seed is a full replicate of the entire pipeline—environment generation, ground-truth reward, train/hold-out split, and feedback draws.

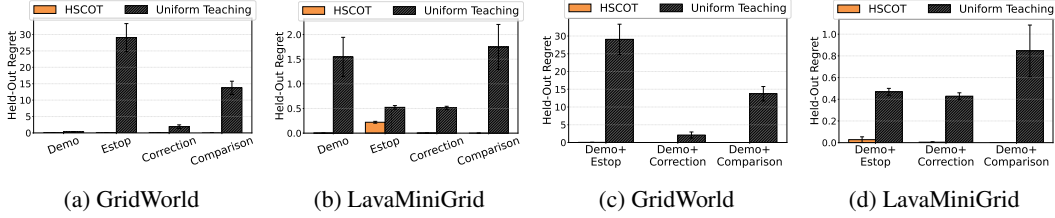


Figure 5: **Held-out regret** (lower is better) averaged over 10 random seeds on 20% held-out MDPs. Bars show mean regret and error bars denote standard error across seeds. Subfigures (a,b) show results for single-modality feedback. Subfigures (c,d) show settings where demonstrations are combined with 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.

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).

Held-out regret. Single-modality results are shown in Figures 5a and 5b. In GridWorld, HSCOT achieves zero held-out regret across all modalities, while uniform teaching incurs substantially higher regret in every case, confirming that strategic environment selection is critical even under identical budgets. In LavaMiniGrid, HSCOT again attains near-zero regret across modalities, whereas uniform teaching leaves residual regret throughout. Under uniform teaching, demonstrations remain comparatively effective since each one implicitly imposes many optimality constraints, and corrections benefit from steering rollouts toward the goal, yielding informative goal-directed constraints. Notably, E-stop outperforms comparisons under uniform querying in LavaMiniGrid: halting points concentrate on imminent lava entry—the dominant reward direction—whereas randomly sampled trajectory pairs rarely differ in lava contact, making many comparison queries uninformative. Mixed-feedback results (Figures 5c and 5d) show that combining demonstrations with another modality allows HSCOT to reach near-zero regret in both domains, as complementary constraint directions across environments tighten the feasible reward region and yield rewards that transfer reliably to unseen MDPs.

Constraint coverage. Figure 6 shows the fraction of the universal constraint set covered by the selected feedback atoms, computed over the training MDPs. HSCOT’s two-level greedy approach (environment selection followed by atom selection) consistently achieves complete coverage across both GridWorld and LavaMiniGrid and 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 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

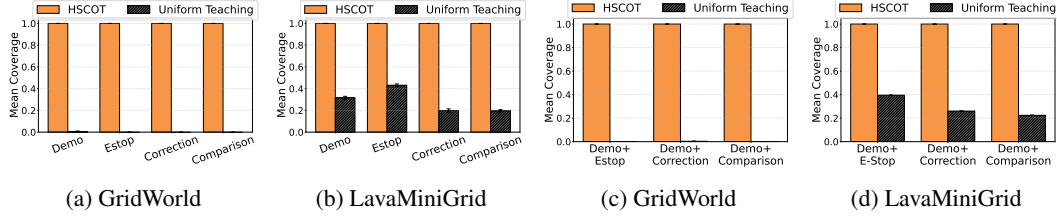


Figure 6: **Constraint coverage** (higher is better) averaged over 10 random seeds. Bars show the fraction of the universal constraint set covered, and error bars denote standard error across seeds. Subfigures (a,b) correspond to single-modality feedback, while (c,d) show mixed-feedback settings where demonstrations are combined with another modality. HSCOT (solid bars) achieves 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 environments activated (out of 50 training MDPs), mean $\pm$ std over 10 seeds. Aggregate averages the two domain means. Lower is better.

Feedback modality	GridWorld		LavaMiniGrid		Aggregate	
	HSCOT	Uniform	HSCOT	Uniform	HSCOT	Uniform
Demonstration	4.3 $\pm$ 0.4	5.2 $\pm$ 0.4	5.5 $\pm$ 0.4	7.1 $\pm$ 0.4	4.9	6.2
Correction	6.9 $\pm$ 0.4	7.7 $\pm$ 0.5	6.9 $\pm$ 0.6	7.0 $\pm$ 0.5	6.9	7.4
Comparison	6.3 $\pm$ 0.4	7.2 $\pm$ 0.4	7.5 $\pm$ 0.3	7.9 $\pm$ 0.4	6.9	7.6
E-stop	5.8 $\pm$ 0.4	6.9 $\pm$ 0.4	12.0 $\pm$ 0.4	15.0 $\pm$ 0.5	8.9	11.0

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 lowest for demonstration feedback, consistent with each demonstration inducing many optimality constraints at once, and highest for E-stop feedback, whose weaker and more trajectory-local constraints require activating more environments to achieve comparable coverage. Per-environment teaching maps that visualize, for each modality, the LavaMiniGrid layouts, the environments HSCOT activates, the held-out split, and the resulting per-environment regret are provided in Appendix 10.5.

9 Conclusion

We studied how to teach reward functions that remain robust across environment dynamics. Our analysis showed that feedback modalities differ sharply in informativeness: comparisons impose the strongest global constraints with unlimited data, while demonstrations are more constraint-efficient per query under tight budgets. We further proved that reward identifiability is environment-dependent—unlimited feedback in a single MDP can leave residual ambiguity that only additional environments resolve. Building on these insights, we introduced HSCOT, a two-stage greedy algorithm that selects informative environments and then queries feedback within them. Across GridWorld and LavaMiniGrid, HSCOT achieved lower held-out regret, complete constraint coverage, and fewer activated environments than uniform teaching under identical budgets, though extending HSCOT to continuous domains and noisy human feedback remains an important direction.

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 sub-optimal 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>.
- Daniel S Brown, Jordan Schneider, Anca Dragan, and Scott Niekum. Value alignment verification. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1105–1115. PMLR, 18–24 Jul 2021. URL <https://proceedings.mlr.press/v139/brown21a.html>.
- 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.
- Yang Chen, Xiao Lin, Bo Yan, Libo Zhang, Jiamou Liu, Neset Özkan Tan, and Michael Witbrock. Meta-inverse reinforcement learning for mean field games via probabilistic context variables. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 11407–11415, 2024.
- 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 Daei, 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 Akrou, 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.
- Lantao Yu, Tianhe Yu, Chelsea Finn, and Stefano Ermon. Meta-inverse reinforcement learning with probabilistic context variables. *Advances in neural information processing systems*, 32, 2019.
- 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

Assumption 1 (Genericity / non-redundancy). For each modality $m \in \{\text{demo}, \text{corr}, \text{E-stop}\}$, there exists a w^* -consistent comparison ordering $w^{*\top}(\Phi(\xi_i) - \Phi(\xi_j)) \geq 0$ that is *not* implied by the constraints of modality m ; equivalently, $H_m \cap \{w : w^\top(\Phi(\xi_i) - \Phi(\xi_j)) < 0\}$ has positive measure. This rules out degenerate environments in which the modalities induce identical constraints, and is what makes the inclusions in Propositions 1–3 strict.

Remark 1 (Sufficient conditions for Assumption 1). Concretely, Assumption 1 holds when: (i) for Prop. 1, there exist suboptimal trajectories $\xi_a, \xi_b \in \Xi \setminus \Xi^*$ whose ordering is not implied by the demonstration (same-initial-state optimality) constraints; (ii) for Prop. 2, M admits ≥ 2 distinct initial states and some cross-initial-state ordering is not implied by the same-initial-state constraints; (iii) for Prop. 3, some w^* -consistent comparison ordering is not implied by the union of all E-stop (prefix-vs-continuation) constraints across Ξ .

All feasible regions below are intersected with the unit ball $\{w : \|w\|_2 \leq 1\}$.

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

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

Then

$$H_{\text{comparison}} \subsetneq H_{\text{demo}}, \quad \text{and} \quad G(H_{\text{comparison}}) < G(H_{\text{demo}}).$$

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

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

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

Each demonstration constraint therefore has the form

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

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 $\xi_a, \xi_b \in \Xi \setminus \Xi^*$ 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. $\square$

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

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

Then

$$H_{\text{comparison}} \subsetneq H_{\text{correction}}, \quad \text{and} \quad G(H_{\text{comparison}}) < G(H_{\text{correction}}).$$

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 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*

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

In contrast, a pairwise comparison between ξ and any $\xi' \in \Xi$ that visits a state-action pair (s', a') with $\phi(s', a')$ linearly independent of this subspace produces a feature difference that lies outside it.

Proof. Let

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

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

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

Then

$$H_{\text{comparison}} \subsetneq H_{E\text{-stop}}, \quad \text{and} \quad G(H_{\text{comparison}}) < G(H_{E\text{-stop}}).$$

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

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

Step 2: Strict inclusion.

By Lemma 1, 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

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

Step 3: Volume comparison.

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

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

□

Theorem 1 (Single-MDP ambiguity). *Let $\mathcal{M} = \{M_1, M_2\}$ be two finite MDPs that share a common feature map $\phi : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}^d$ and a common ground-truth reward parameter $w^* \in \mathbb{R}^d$, where rewards take the linear form $r(s, a) = w^{*\top} \phi(s, a)$. 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*

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

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 $w \neq w^*$ such that $w \in \mathcal{W}(M_1)$ but $w \notin \mathcal{W}(M_1) \cap \mathcal{W}(M_2)$.

Proof. **Step 1: Cone form of the gBEC.** For each MDP M_k , let

$$\mathcal{C}_k = \{\Delta\Phi(\xi^+, \xi^-) : \xi^+, \xi^- \text{ feasible in } M_k, (w^*)^\top \Delta\Phi(\xi^+, \xi^-) \geq 0\}$$

be the feature-difference directions that form valid constraints under w^* . Then

$$\mathcal{W}(M_k) = \{w \in \mathbb{R}^d : w^\top \Delta\Phi \geq 0 \ \forall \Delta\Phi \in \mathcal{C}_k\}.$$

For any pair (ξ_i, ξ_j) , one of $\Delta\Phi(\xi_i, \xi_j)$ or its negation $\Delta\Phi(\xi_j, \xi_i)$ satisfies the sign condition, so $\mathcal{C}_k$ contains a spanning vector for every line generated by a feature difference, giving $\text{span}(\mathcal{C}_k) = \mathcal{V}_k$.

Step 2: Perturbation direction u . Since $\mathcal{V}_1 \subsetneq \text{span}(\mathcal{V}_1 \cup \mathcal{V}_2)$, we have $\mathcal{V}_2 \not\subseteq \mathcal{V}_1$. Fix $z \in \mathcal{V}_2 \setminus \mathcal{V}_1$. Let $\{b_1, \dots, b_m\}$ be an orthonormal basis of $\mathcal{V}_1$ and define

$$u = z - \sum_{i=1}^m (z^\top b_i) b_i,$$

the component of z orthogonal to $\mathcal{V}_1$. Then $u \neq 0$ (as $z \notin \mathcal{V}_1$), and $u^\top \Delta\Phi = 0$ for all $\Delta\Phi \in \mathcal{V}_1$. Moreover $u^\top z = \|u\|_2^2 > 0$ with $z \in \mathcal{V}_2$, so $u \notin \mathcal{V}_2$.

Step 3: A violable constraint. Since $\text{span}(\mathcal{C}_2) = \mathcal{V}_2$ and $u \notin \mathcal{V}_2$, there exists $\Delta\Phi' \in \mathcal{C}_2$ with $u^\top \Delta\Phi' \neq 0$. Replacing z by $-z$ if its projection gives the wrong sign, take u so that

$$u^\top \Delta\Phi' < 0, \quad (w^*)^\top \Delta\Phi' \geq 0,$$

where the second inequality holds because $\Delta\Phi' \in \mathcal{C}_2$.

Step 4: Explicit perturbation magnitude. Choose

$$\varepsilon > \frac{(w^*)^\top \Delta\Phi'}{|u^\top \Delta\Phi'|} \ (\geq 0).$$

Then

$$(w^* + \varepsilon u)^\top \Delta\Phi' = (w^*)^\top \Delta\Phi' - \varepsilon |u^\top \Delta\Phi'| < 0,$$

for any margin $(w^*)^\top \Delta\Phi' \geq 0$.

Step 5: Witness reward. Set $\tilde{w} = w^* + \varepsilon u \neq w^*$. For every $\Delta\Phi \in \mathcal{C}_1$, $u^\top \Delta\Phi = 0$, so

$$\tilde{w}^\top \Delta\Phi = (w^*)^\top \Delta\Phi \geq 0,$$

giving $\tilde{w} \in \mathcal{W}(M_1)$. By Step 4, $\tilde{w}^\top \Delta\Phi' < 0$, so $\tilde{w} \notin \mathcal{W}(M_2)$ and hence $\tilde{w} \notin \mathcal{W}(M_1) \cap \mathcal{W}(M_2)$. Therefore no feedback obtained solely from M_1 identifies w^* . $\square$

10.2 Per-budget feasible region analysis

To analyze the limited-budget regime more rigorously, we go beyond the single fixed-budget heatmaps and study how each feedback modality behaves as the budget varies. We proceed in two steps. First, we visualize the feasible reward region as a heatmap across a range of feedback budgets $B \in \{1, 2, 3, 5, 7, 10, 15, 20\}$, showing how the region shrinks for each modality as more queries become available. Second, to obtain a quantitative, budget-matched comparison, we approximate the *volume* of the feasible region at each budget and plot it as a function of B . All figures in this section are computed for the same single MDP in Figure 4a, matching the finite-data regions in Figure 2.

Per-budget heatmap results. The per-budget feasibility heatmaps (Figures 8 and 8) show how each modality contracts the feasible region as the budget grows. Demonstrations are tight from the start: already at $B = 1$ (Fig. 8a) the demonstration region is a narrow wedge around w^* and it changes little thereafter, since a single demonstration contrasts the optimal action at each visited state against its alternatives and thus instantiates many optimality constraints at once. Comparisons narrow steadily: the comparison region starts as a wide wedge at $B = 1$ but contracts at each successive budget, becoming a thin sliver comparable to demonstrations by $B = 15$ – 20 (Fig. 8g–8h), because comparisons can contrast arbitrary trajectory pairs and each query may cut along a new feature-difference direction. Corrections and E-stops, by contrast, saturate: they reduce the feasible region far more slowly and retain wide wedges even at $B = 20$, since by Lemma 1 E-stop constraints lie in the span of features along a single trajectory and corrections are limited by their shared start-state structure, so repeated queries re-cut along similar directions and quickly stop adding information. Taken together, the heatmaps make a regime crossover visible: demonstrations dominate under tight budgets, while comparisons overtake them as the budget grows. This trend is consistent with our unlimited-data analysis (Figure 3): as the budget increases, the comparison wedge approaches and begins to match the demonstration wedge, moving toward the unlimited-data regime in which comparisons impose the strongest global ordering constraints and yield the smallest feasible region (Proposition 1). The crossover is only partial at the budgets shown—corrections and E-stops remain far from their idealized regions because their constraints saturate—but the comparison trend clearly anticipates the unlimited-data ordering.

Feasible-region volume results. Aggregating these regions into a single scalar, Figure 7 plots the feasible-region volume $G(D)$ against the budget B , providing the budget-matched comparison across modalities. Demonstrations attain the smallest volume at every budget, collapsing the feasible region to a thin sliver already at $B = 1$ and leaving little room to shrink further. Comparisons begin with a much larger volume but contract steadily, approaching the demonstration curve only at large budgets ($B \geq 15$), consistent with comparisons needing many queries to accumulate the global ordering constraints that make them dominant in the unlimited-data regime. Corrections and E-stops reduce ambiguity far more slowly and plateau at large volumes—E-stop the largest—because their constraints are confined to same-start-state or trajectory-local feature-difference subspaces (Lemma 1) and saturate quickly. The ordering $G(D)_{\text{demo}} \leq G(D)_{\text{comp}} \leq G(D)_{\text{corr}} \leq G(D)_{\text{E-stop}}$ holds across the entire budget range with tight variance bands, confirming that demonstrations are the most constraint-efficient modality under limited budgets. This steady contraction of the comparison curve toward the demonstration curve mirrors the crossover anticipated by the unlimited-data analysis (Figure 3): with enough queries, comparisons accumulate the global ordering constraints that make them the most constraining modality in the limit, so increasing the budget moves the empirical ordering toward the idealized one in which comparisons yield the smallest feasible region. The convergence is clearest for comparisons; corrections and E-stops plateau well above their idealized regions, indicating that the budgets considered here approach but do not fully reach the unlimited-data regime for those modalities.

How the feasible-region volume is approximated. For each budget B and each modality, we draw B random feedback queries and extract the induced homogeneous linear constraints $v^\top w \geq 0$,

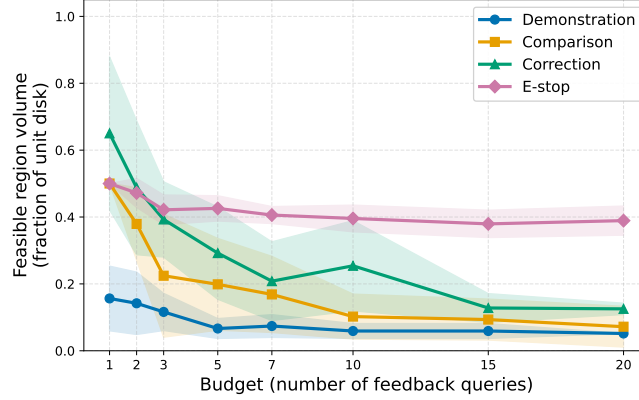


Figure 7: **Budget-matched reward ambiguity across modalities.** Feasible-region volume $G(D)$ (lower is better) versus budget B , averaged over 100 random samples per setting; shaded bands denote ± 1 standard deviation. Demonstrations attain the smallest volume at every budget; comparisons shrink steadily and approach demonstrations only for large B ; corrections and E-stops reduce ambiguity slowly and plateau at larger volumes. The ordering $G(D)_{\text{demo}} \leq G(D)_{\text{comp}} \leq G(D)_{\text{corr}} \leq G(D)_{\text{E-stop}}$ holds across the entire range with tight variance bands.

where $v = \Phi(\xi^+) - \Phi(\xi^-)$ is the discounted feature-difference between the preferred and dispreferred trajectory. The feasible reward region is the set of weights satisfying all of these constraints simultaneously. Because every constraint is homogeneous and we restrict to $\|w\|_2 \leq 1$, this region is a convex cone intersected with the unit disk, i.e., an angular wedge. We approximate its volume $G(D)$ as the *fraction of the unit reward disk that remains feasible*: we evaluate all constraints on a dense grid over the disk and report the proportion of points satisfying every constraint. Smaller $G(D)$ thus indicates a tighter, less ambiguous reward region. We repeat the procedure 100 times per $(B, \text{modality})$ configuration and report the mean and standard deviation, yielding the curves in Figure 7.

10.3 Feasible reward regions across sampled layouts

Figures 9 and 10 extend the single MDP-pair illustration of Figure 4 to a larger population of layouts, confirming that the environment-dependent ambiguity of Theorem 1 is typical rather than an artifact of one hand-picked example. We use 2×3 gridworlds with two cell features (drawn gray and white) and a randomly placed terminal cell $\mathbf{T}$ that may occupy either feature. For each $k \in \{1, \dots, 5\}$, where k is the number of gray-feature cells, we sample 40 layouts, for 200 layouts in total. Figure 9 shows the sampled layouts with their optimal policies (one arrow per tied optimal action), and Figure 10 shows the corresponding feasible reward regions $\text{gBEC}(D)$. The feasible region varies markedly across layouts: some pin the reward down to a thin wedge, while others—typically those with more gray cells, whose optimal trajectories are less distinguishable in feature space—leave a much larger region. This reflects that reward identifiability is governed by the layout-induced feature-difference directions rather than by the amount of feedback.

How the feasible regions are computed. Each region is obtained from the constraints induced by *all* optimal demonstrations in a layout, expressed through action-successor features. Recall the discounted feature counts (successor features) of a policy π introduced in the preliminaries, which we write here as $\psi^\pi(s) \equiv \mu^\pi(s)$ and $\psi^\pi(s, a) \equiv \mu^\pi(s, a)$:

$$\psi^\pi(s) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \phi(s_t, a_t) \mid s_0 = s \right],$$

$$\psi^\pi(s, a) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \phi(s_t, a_t) \mid s_0 = s, a_0 = a, a_{t \geq 1} \sim \pi \right].$$

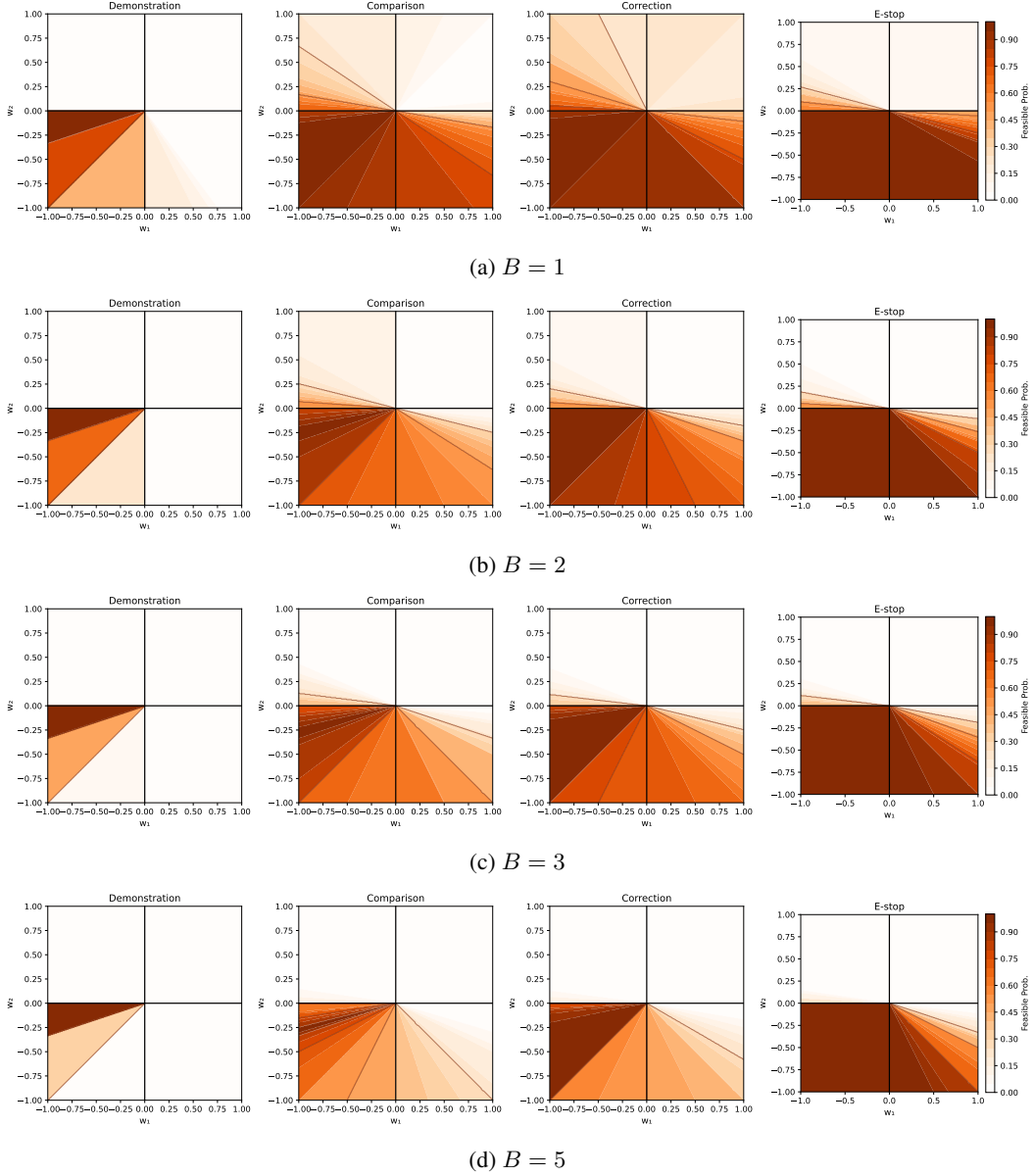


Figure 8: **Per-budget feasible reward regions across feedback modalities.** Each row shows the feasible reward region for a single feedback budget B , with the four columns corresponding to demonstration, comparison, correction, and E-stop feedback. Colors encode the empirical feasibility probability over 100 independent random feedback samples—darker regions are feasible across more samples. All panels are computed for the same MDP whose layout is shown in Figure 4a. (Continued in Figure 8.)

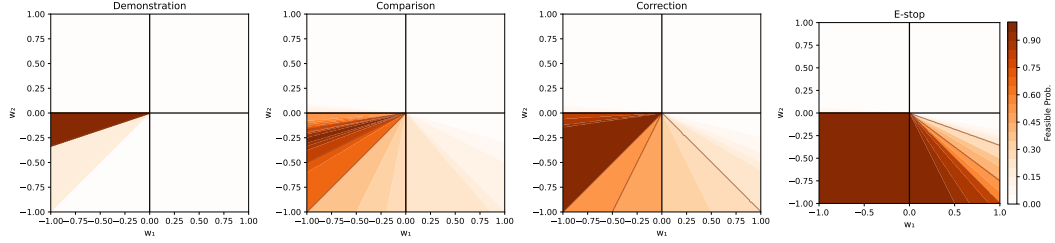
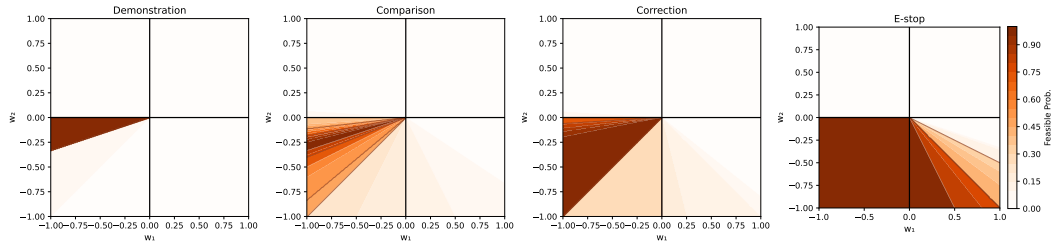
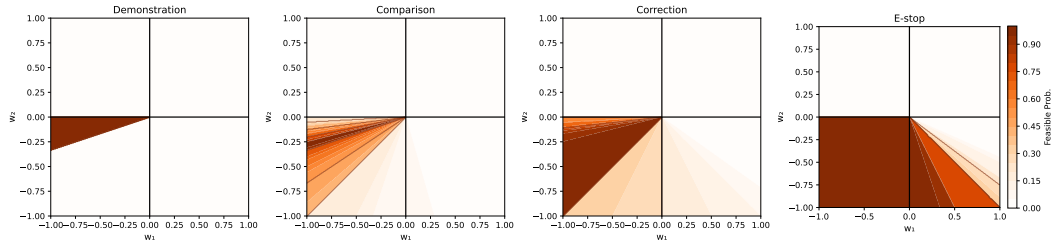
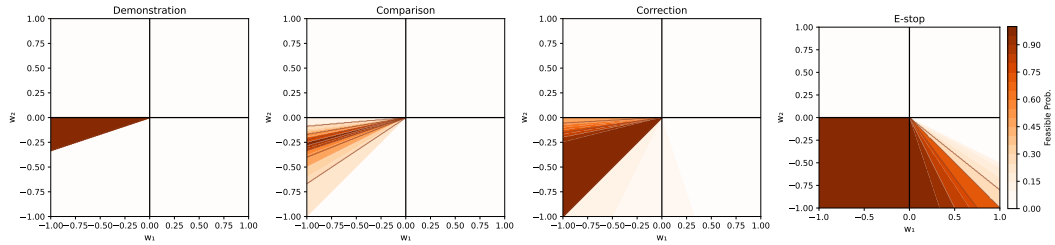
(e) $B = 7$ (f) $B = 10$ (g) $B = 15$ (h) $B = 20$

Figure 8: **Per-budget feasible reward regions (continued).** Budgets $B \in \{7, 10, 15, 20\}$, continuing Figure 8. As the budget grows, the comparison wedge tightens toward the demonstration wedge, while the correction and E-stop wedges shrink slowly and stay comparatively wide, reflecting their same-start-state and trajectory-local constraint structure.

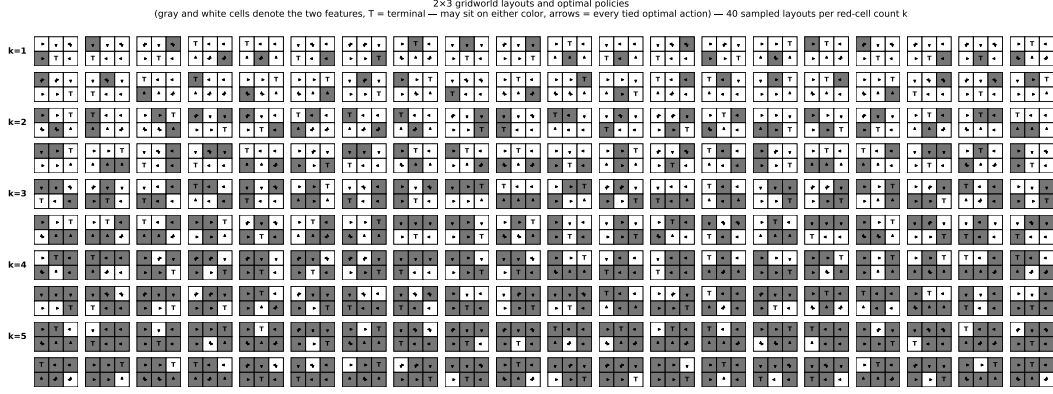


Figure 9: **Sampled 2×3 layouts and optimal policies.** Gray and white cells denote the two features and **T** the (randomly placed) terminal cell, which may sit on either feature; arrows show every tied optimal action. Rows are grouped by k , the number of gray-feature cells (40 layouts per k , 200 total).

The action-successor feature $\psi^\pi(s, a)$ takes action a in state s and then follows π ; it captures the full long-run feature consequence of that action choice and satisfies $Q_w^\pi(s, a) = w^\top \psi^\pi(s, a)$ (in these gridworlds the feature is a property of the cell, so $\phi(s, a) = \phi(s)$). We obtain the optimal policy π^* by value iteration and compute its action-successor features ψ^{π^*} . Optimality of any $a^* \in \arg \max_a Q_w^{\pi^*}(s, a)$ over an alternative action b then gives, for every non-terminal state s ,

$$w^\top (\psi^{\pi^*}(s, a^*) - \psi^{\pi^*}(s, b)) \geq 0,$$

which is exactly a behavioral-equivalence-class optimality constraint and a special case of the generalized constraint $w^\top \Delta \Phi \geq 0$ in Eq. (1), with $\Delta \Phi = \psi^{\pi^*}(s, a^*) - \psi^{\pi^*}(s, b)$. Collecting these feature-difference vectors over all states and all optimal/non-optimal action pairs, and discarding constraints implied by the rest, yields the half-spaces whose intersection is $\text{gBEC}(D)$. Restricting to $\|w\|_2 \leq 1$ makes this a bounded convex cone (an angular wedge for $d = 2$): in Figure 10 the red lines are the retained constraint boundaries and the gold hatch is their feasible intersection.

10.4 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 $\mathcal{K}$

```

1: covered  $\leftarrow \emptyset$ 
2:  $\mathcal{K} \leftarrow []$ 
3: while covered  $\neq \mathcal{U}$  do
4:   Select  $k$  maximizing  $|\mathcal{U}_k \setminus \text{covered}|$ 
5:    $\mathcal{K} \leftarrow \mathcal{K} \cup \{k\}$ 
6:   covered  $\leftarrow \text{covered} \cup (\mathcal{U}_k \setminus \text{covered})$ 
7: end while
8: return  $\mathcal{K}$ 
    
```

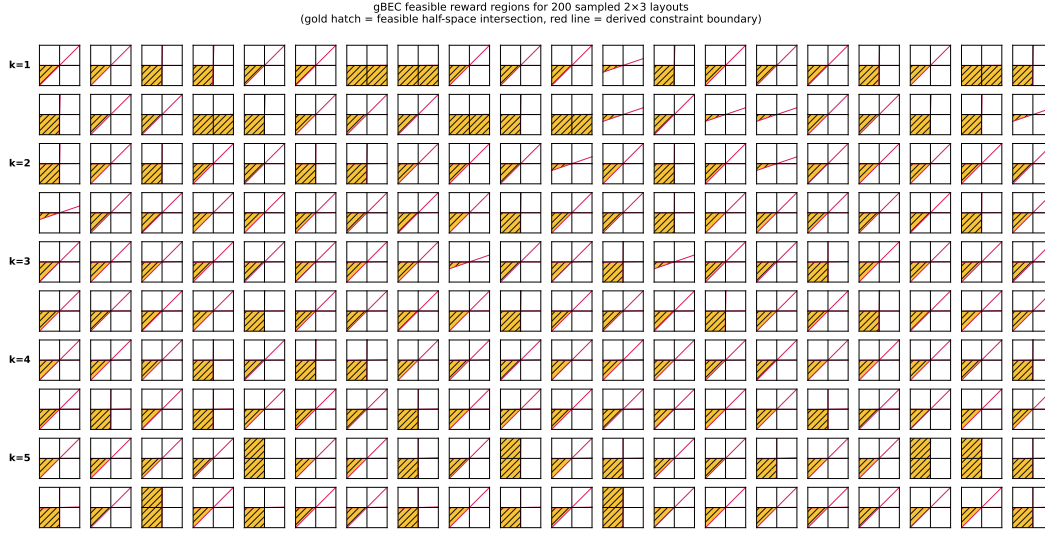


Figure 10: **Feasible reward regions gBEC(D) for the sampled layouts.** For each layout in Figure 9, the gold hatch is the feasible half-space intersection and the red lines are the derived constraint boundaries, computed from the action-successor-feature constraints described above. The region shape—and hence reward identifiability—varies strongly with layout, mirroring Figure 4.

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

- 1: $\text{covered} \leftarrow \emptyset$
 - 2: $D \leftarrow []$
 - 3: **while** $\text{covered} \neq \mathcal{U}$ **do**
 - 4: Select atom x from any $k \in \mathcal{K}$ maximizing $|\text{coverage}(x) \setminus \text{covered}|$
 - 5: $D \leftarrow D \cup \{x\}$
 - 6: $\text{covered} \leftarrow \text{covered} \cup (\text{coverage}(x) \setminus \text{covered})$
 - 7: **end while**
 - 8: **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

- 1: $\mathcal{K} \leftarrow \text{Greedy Environment Selection}(\{\mathcal{U}_k\}, \mathcal{U})$ (Alg. S1)
 - 2: Restrict candidate atoms to those in environments $\mathcal{K}$
 - 3: $D \leftarrow \text{Greedy Atom Selection}(\mathcal{K}, \mathcal{U})$ (Alg. S2)
 - 4: **return** $(\mathcal{K}, D)$
-

10.5 Per-environment teaching maps

To complement the aggregate metrics in Section 8, we visualize the full set of 50 LavaMiniGrid environments for each feedback modality. In every panel, red cells denote lava, green the goal, and white free space. An orange border marks environments selected by HSCOT, a cyan border marks held-out environments excluded from teaching, and a yellow border marks training environments with high residual regret. Each panel title reports its per-environment regret r , and each figure title reports the mean regret over all environments. These maps make HSCOT’s selective environment activation concrete. Comparison (Figure 11a), correction (Figure 11b), and demonstration (Figure 12a) all achieve zero mean regret, whereas E-stop (Figure 12b) leaves residual regret across many environments (mean regret 0.188).



(a) Comparison feedback (mean regret = 0.000).



(b) Correction feedback (mean regret = 0.000).

Figure 11: Per-environment teaching maps (comparison and correction). Each panel shows one LavaMiniGrid environment: red is lava, green the goal, white free space. Orange borders mark HSCOT-selected environments, cyan borders mark held-out environments, and yellow borders mark high-regret training environments. Panel titles report per-environment regret r .

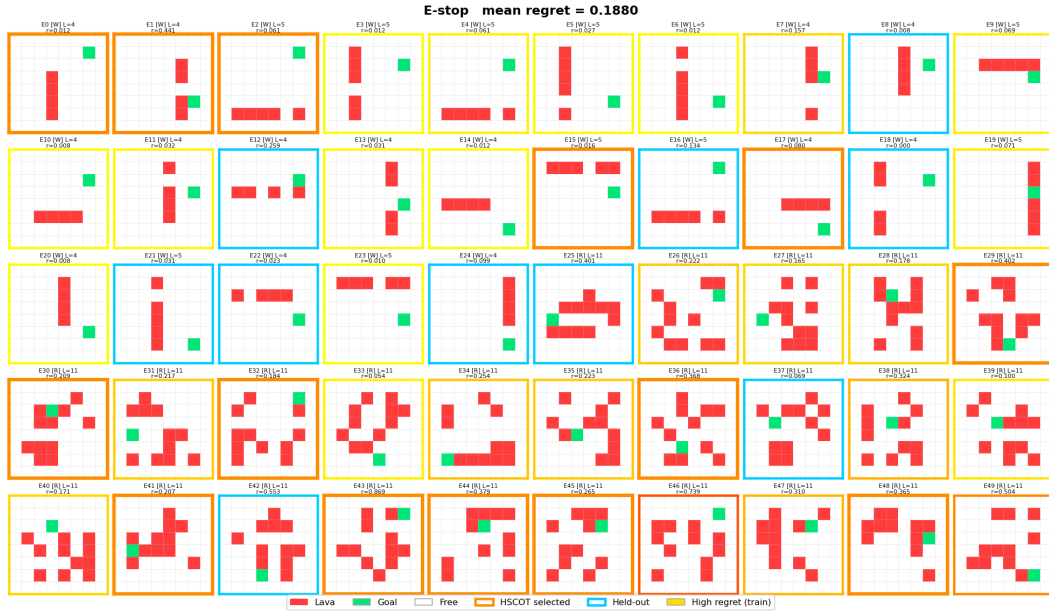


Figure 12: **Per-environment teaching maps (demonstration and E-stop).** Same color and border conventions as Figure 11. E-stop leaves substantial residual regret across many environments and triggers several high-regret (yellow) panels, unlike the other modalities.