How the Teaching Style and Interpretation Type of State Interventions Shape Multi-Agent Coordination

Zhuolun Zhong (zhongzhuolun995@gmail.com) **Luc Caspar (luc.caspar@cross-compass.com)**
Cross Compass
Tokyo, Japan

Cross Compass
Tokyo, Japan

Joseph L. Austerweil (joseph.austerweil@gmail.com)
Chiba Institute of Technology
Tsudanuma, Japan

Abstract

How can an external teacher accelerate coordination among multiple learning agents? We investigate this question in a multi-agent foraging task where a simulated teacher can physically relocate agents—a direct but ambiguous pedagogical signal. We formalize a taxonomy of teaching styles (e.g., *Undoing*, *Correcting*) and agent interpretation rules. Our computational experiments reveal that the ***Undoing* style is a uniquely powerful catalyst for spatial coordination**, most effectively breaking behavioral symmetry between two agents by imposing spatial constraints that define "home regions." This catalytic effect scales, reinforcing niche specialization in three-agent systems. However, success is governed by a **pedagogical matching principle**: interventions fail when the agent's interpretation rule mismatches the teacher's intent (e.g., *Correcting* fails when interventions are interpreted as punitive). Our work provides a formal framework showing how structured guidance and agent interpretation jointly shape the emergence of collective organization.

Keywords: Multi-agent reinforcement learning; pedagogy; state intervention; coordination; spatial specialization.

Introduction

Teaching in complex, dynamic environments often extends beyond verbal instruction to include direct manipulation of the learner's state. In multi-agent settings, the challenge of such pedagogical intervention intensifies: the teacher must guide individual learning while fostering collective coordination under real-time resource competition. While previous research has explored how agents interpret physical corrections in single-agent tasks, the role of state interventions as a tool for shaping group-level behavior remains under-explored.

This paper investigates how algorithmic teaching styles—defined by when and where an agent is moved—catalyze spatial coordination. Unlike traditional RL, we treat state intervention as fast environmental shaping to break behavioral symmetry and foster *spatial specialization*. To study this, we use a grid-world where a teacher **teleports** agents to specific **tiles**. Agents must then **interpret** these relocations to update their reward representations.

The complexity of this teaching scenario is defined by delayed pedagogical outcomes and the teacher's need to manage multiple agents under spatial uncertainty. Rather than focusing on the agent's internal interpretive ambiguity, we adopt a **system-level modeling approach** to map the performance landscape across various teaching styles and agent interpretation rules.

To our knowledge, this is the first systematic investigation of state interventions in real-time, multi-agent, discrete environments where the teacher lacks access to the learner's feature representation. By bridging the gap between heuristic pedagogical strategies and multi-agent coordination, our work provides a computational foundation for designing adaptive agents that learn effectively from structured external guidance. **Our contributions are threefold:** (1) we formalize a **taxonomy of teaching styles** for multi-agent coordination; (2) we identify a **pedagogical matching principle** where effectiveness depends on teacher-agent alignment; and (3) we demonstrate that specific interventions act as **catalysts for stable spatial niche formation**.

Prior work

Our work sits at the intersection of computational models of teaching, reinforcement learning from human feedback, and multi-agent systems. This section provides an overview of the relevant literature along these axes, highlighting how our approach diverges from and extends prior research.

Computational Models of Teaching and Intervention

Formal models of teaching often build upon the framework of Partially Observable Markov Decision Processes (POMDPs) to capture the interactive dynamics between a teacher and a learner (Kaelbling, Littman, & Cassandra, 1998). This formalism has been productively applied to understand human pedagogical reasoning (Shafto, Goodman, & Griffiths, 2014) and to design algorithms for efficient instruction (Rafferty, Brunskill, Griffiths, & Shafto, 2016). A key insight from this line of work is that teaching is not merely reinforcement; it is a communicative act where signals are chosen to influence the learner's beliefs and future behavior (Ho, Cushman, Littman, & Austerweil, 2019, 2021).

Physical state interventions encode a direct yet ambiguous teaching signal. Prior work explored how people use interventions to guide an agent to a goal while avoiding obstacles (Ho, Littman, & Austerweil, 2017). More recently, Losey, Bajcsy, O'Malley, and Dragan (2022) formalized physical human-robot interaction as a POMDP where the robot maintains a belief over the human's unknown reward function, that is updated through physical corrections interpreted as intentional signals. Their work demonstrates

that heuristic interpretations of interventions (e.g., updating one feature weight at a time) can improve learning efficiency. However, these approaches are designed for **single-agent, continuous-control tasks** (e.g., robotic manipulation) and assume a **shared, known feature space** between human and robot, which allows for a relatively straightforward mapping from physical corrections to reward function updates.

In contrast, our work investigates discrete grid-based environments where agents learn *tabula rasa* via model-free RL without prior knowledge of task features or reward structure. The teacher's interventions (e.g., dragging an agent to a new tile) must therefore be interpreted through heuristic rules defined at the level of state transitions and value updates, not feature-weight adjustments. We extend the idea of heuristic interpretation by proposing and systematically evaluating a broader taxonomy of such rules.

Reinforcement Learning from Human Feedback and Guidance

The paradigm of Reinforcement Learning from Human Feedback (RLHF) and related interactive techniques provide a rich foundation for learning from various forms of human input (Luo, Xu, Wu, & Levine, 2024). In such frameworks, human interventions are often treated as demonstrations or preference labels. For instance, Luo et al. (2024) treat human interventions during high-dimensional manipulation tasks as *suggestions*, where the agent learns directly from the state-action pair induced by the human.

A crucial distinction in our setting is the nature of the task and agent embodiment. Unlike continuous control tasks where a human can kinesthetically guide an end-effector through a physically plausible trajectory, our discrete environment introduces a fundamental asymmetry: the teacher can *teleport* an agent to any tile. An action the agent cannot replicate on its own. This breaks the direct *"suggestion as demonstratable action"* assumption common in RLHF for robotics. Consequently, we must consider interpretations beyond direct imitation, such as treating an intervention as a reset, an interruption, or even a punitive signal.

Multi-Agent Learning and External Shaping

In multi-agent reinforcement learning (MARL), coordination often emerges from self-interested learning (Axelrod, 1984; Austerweil et al., 2015) or is engineered through specific reward structures (Wolpert & Tumer, 2001) and communication protocols (Foerster, Assael, De Freitas, & Whiteson, 2016). External shaping by a teacher adds a new dimension to this problem. While there is extensive literature on centralized training for decentralized execution (Lowe et al., 2017; Rashid et al., 2020; Foerster, Farquhar, Afouras, Nardelli, & Whiteson, 2018) and on learning communication protocols (Foerster et al., 2016), the explicit study of how **an external teacher can use state interventions to shape multi-agent coordination** is nascent.

Classic game-theoretic analyses show that cooperation can emerge in repeated interactions without explicit social motivations (Axelrod, 1984), and self-interested agents can discover cooperative strategies in many grid-world games (Austerweil et al., 2015). However, analyzing two or more agents with continuous-time, interactive signals remains open. Our work investigates how targeted state interventions (a form of fast, asymmetric environmental shaping) can accelerate and stabilize the emergence of coordinated behaviors, such as division of labor, in a learning population. This connects to broader questions in social learning about how environmental structure mediates the value of external information (Wu et al., 2025). We contribute by providing a formal framework to study this specific shaping mechanism.

Method

We investigate a real-time foraging task in a 12×12 discrete grid world. A fixed set of tiles are designated as spawn tiles, where pellets appear stochastically over time (with a spawn rate of 100ms) following a Poisson process (Kingman, 1992). We focus on a *smooth* distribution, where spawn tiles are grouped into a small number of contiguous, spatially coherent regions (Wu et al., 2025). Agents occupy a single tile and move continuously toward adjacent tiles (including diagonals) over a fixed duration. When an agent overlaps with a pellet, it receives a reward R_{pellet} and the pellet is consumed. We compare two spatial configurations: (1) a two-agent scenario where spawn tiles are divided into two contiguous regions (9 tiles each), positioned in **opposite corners (top-left and bottom-right)** of the grid; and (2) a three-agent scenario with three contiguous regions (6 tiles each), **arranged in an inverted triangular layout (top-left, top-right, and bottom-center)**. Regions are separated by non-spawn tiles, creating spatial gaps that make division of labor a viable coordination strategy. In both scenarios, agents' initial states are randomly selected at the start of each episode.

A *simulated teacher* with perfect knowledge of the spawn distribution (but not real-time pellet presence) can perform *state interventions*: forcibly relocating an agent to a new tile at any moment. The agent then resumes its policy from this new state. The teacher decides when to intervene by estimating, for each possible action an agent could take from its current position, the long-term value of that action based solely on the known spawn distribution. If the agent's chosen action is estimated to be suboptimal (i.e., less valuable than the best available alternative from that position), the teacher triggers an intervention with a probability specified by the *intervention rate* parameter. The intervention itself is determined by the *teaching style*.

This setup introduces several learning challenges: (1) rewards are dynamic and stochastic; (2) agents learn online via reinforcement learning with no prior knowledge; (3) we analyze systems involving **two or three agents** operating concurrently, which creates competitive pressure and requires spatial coordination; (4) interventions provide an ambiguous, real-time pedagogical signal that agents must interpret within these multi-agent social dilemmas. This forms a complex

testbed for evaluating how state interventions affect collective coordination and competition in dynamic environments.

Intervention styles

Our environment is modeled as a Markov Decision Process (MDP) (S, A, T, R, γ), where the state space S corresponds to the agent's discrete grid position; the action space A includes movements to the eight adjacent tiles; the transition function T is deterministic; the immediate reward $R(s_t, a_t, s_{t+1})$ is stochastic, depending on whether a pellet is collected during the movement from s_t to s_{t+1}; and γ is the discount factor.

Agents learn via Q-learning (Sutton & Barto, 2018) with an ε-greedy policy. The Q-update for (s_t, a_t, s_{t+1}, r_t) is:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[r_t + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t) \right] \tag{1}$$

Based on empirical tuning, we set the reward $R_{\text{pellet}} = 6$, the learning rate $\alpha = 0.1$, the discount factor $\gamma = 0.9$, and a fixed exploration rate $\varepsilon = 0.3$. Each movement action occupies a fixed duration of 200ms. These parameters remain consistent across all experimental conditions to ensure comparability.

A core challenge is how an agent should update its value function when its movement is interrupted by a state intervention. We formalize this by distinguishing between:

- s^A_{t+1}: the state the agent would have reached autonomously,
- s^I_{t+1}: the agent's state post-intervention,
- a_t: the original action attempted by the agent,
- r_t: the environmental reward received prior to interruption.

We also define a^I as the action from s_t that best reduces distance to s^I_{t+1}, and r^I as the reward obtained at s^I_{t+1}. To systematically evaluate these interpretation types, we define four algorithmic **teaching styles** that determine *when* and *where* to intervene:

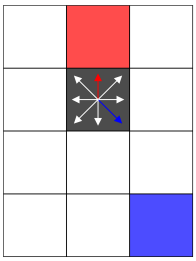


Figure 1: As an agent moves along the red arrow (action a_t) moving from the black (s_t) to red tile (s^A_{t+1}), it is intervened on and is placed on the blue tile (s^I_{t+1}). The blue arrow a^I is the closest action from the starting black tile s_t the post-intervention blue tile s^I_{t+1}.

- **Undoing**: Returns the agent to its previous state s_t.
- **Correcting**: Moves the agent to a randomly selected pellet spawn tile within a 5×5 neighborhood of s_t.
- **Exploration-encouraging Correcting**: Moves the agent to a non-spawn tile adjacent (within a 3×3 area) to a spawn tile in the 5×5 neighborhood of s_t.
- **Restart**: Moves the agent to a non-spawn tile in the 5×5 neighborhood of s_t, ensuring the relocation tile is outside the current contiguous spawn region to reset the agent's trajectory.

An intervention is triggered probabilistically (via an *intervention rate* parameter) when the agent's current action is suboptimal relative to the known spawn tile distribution.

Intervention Interpretation

How should an agent update its value function when its movement is interrupted by a state intervention? When externally displaced, an agent must infer the pedagogical intent behind the intervention. We propose six distinct **interpretation types** for mapping an intervention event to a Q-update:

- **Suggestion**: Treats s^I_{t+1} as a recommended goal. *Update:* $(s_t, a^I, s^I_{t+1}, r^I)$.
- **Reset**: Treats intervention as a neutral relocation adapted from (Ho et al., 2017). *Update:* $(s_t, a_t, s^A_{t+1}, r_t)$.
- **Interrupt**: Treats intervention as a mere interruption adapted from (Ho et al., 2017). No Q-update is performed.
- **Transition**: Treats intervention as a positive guided transition. *Update:* $(s_t, a^I, s^I_{t+1}, +2R_{\text{pellet}})$. If $s^I_{t+1} = s_t$, it is treated as negative: $(s_t, a_t, s^A_{t+1}, -2R_{\text{pellet}})$.
- **Disrupt**: Treats any intervention as inherently negative. *Update:* $(s_t, a^I, s^I_{t+1}, -2R_{\text{pellet}})$.
- **Impede**: Treats intervention as a punishment for the original action. *Update:* $(s_t, a_t, s^A_{t+1}, -2R_{\text{pellet}})$.

These interpretation types represent distinct computational hypotheses for how physical interventions function as communicative acts within the learning process.

Policy Evaluation and Coordination Metrics

The stochastic and dynamic nature of pellet generation, and multi-agent interaction violates the standard Markov assumption. Explicitly modeling pellet counts would lead to a combinatorial state explosion. Therefore, to evaluate and compare policies, we approximate the **expected reward** of a policy π using a **Monte Carlo simulation** approach:

$$\hat{V}(\pi) = \frac{1}{100} \sum_{k=1}^{100} \sum_{\tau=0}^{30} \gamma^\tau r_k^{(\tau)}, \tag{2}$$

where $r_k^{(\tau)}$ is the reward obtained at step τ in the k-th rollout. Each rollout starts the agent from a random position and executes π for 30 steps in a new environment instance (with other

agents). This metric estimates the agent's average foraging efficiency under environmental and interactive stochasticity.

We analyzed 4 *teaching styles*, 6 *interpretation types*, and 4 intervention rates with 50 replications ($96 \times 50 = 4800$). This approach allows us to investigate how specific combinations of pedagogical signals and agent-level heuristics interact with environmental topology. By evaluating these configurations in spatially continuous "smooth" environments, we aim to identify the computational conditions that yield the most stable policy acquisition and group-level performance. Coordination emerges as a critical challenge, when multi-agent systems scale. Ideally, agents should minimize competitive overlap by partitioning the environment—a phenomenon we term *spatial specialization*. The goal is to investigate how intervention styles influence coordination.

To evaluate multi-agent coordination, we define two key metrics: (1) **Policy Divergence** ($\Delta\pi$), which measures the degree of spatial specialization by the difference in agents' regional preferences (Expected Reward in Region 1 vs. 2); and (2) **System Equilibrium** (σ), defined as the absolute difference in agents' scores to capture resource fairness. In three-agent settings, we use **Region Attachment** to index niche specialization.

Results and Analysis

Table 1: Factorial ANOVA Results for Coordination Metrics. Degrees of freedom (df) are reported for each factor and interaction.

Factor (df_{factor}, df_{error})	Policy Div. (F)	Sys. Equil. (F)	Reg. Attach. (F)
Teaching Style $(3, 4704)$	418.96^{***}	267.02^{***}	8517.27^{***}
Interpretation $(5, 4704)$	9.03^{***}	1.56	55.92^{***}
Rate $(3, 4704)$	8.88^{***}	82.09^{***}	455.05^{***}
Style $\times$ Interp. $(15, 4704)$	2.86^{***}	2.99^{***}	22.19^{***}

Note: *** *denotes* $p < .001$.

Emergence of Coordination in Two-Agent Systems Factorial ANOVA results (Table 1) reveal that coordination is highly sensitive to teaching styles and interpretation rules. As shown in Fig. 2, the **Restart** style yielded the highest divergence ($M = 165.7$) but severely degraded equilibrium ($M = 2626$, $d = 0.87$). In contrast, **Undoing** emerged as the most balanced strategy, fostering moderate specialization ($M = 74.0$) while maintaining superior fairness ($M = 949.8$).

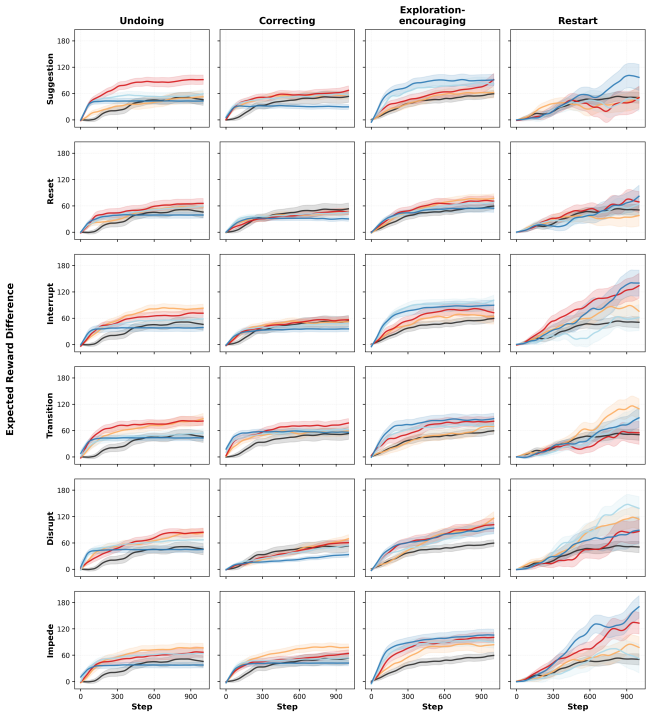
A critical **pedagogical mismatch** was observed between **Correcting** and **Disrupt**: when agents interpreted beneficial guidance as punitive, divergence dropped significantly ($d = -0.77$). This supports our **matching principle**: effectiveness hinges on alignment between teacher intent and learner heuristic.

Niche Reinforcement in Three-Agent Systems In complex three-agent settings, the superiority of **Undoing** and **Correcting** became even more pronounced, yielding dramatically higher spatial attachment than more aggressive styles ($d = 2.90$, Fig. 3). We found that higher intervention rates amplified coordination for Undoing ($r = .65$), yet actively dismantled it for **Restart** ($r = -.24$). This divergent correlation proves that more intervention is not always better; rather, the effect depends on whether the pedagogical signal reinforces or erases the emerging spatial structure. These results suggest that state interventions can scale to larger collectives, provided they function as a positive feedback loop for niche formation.

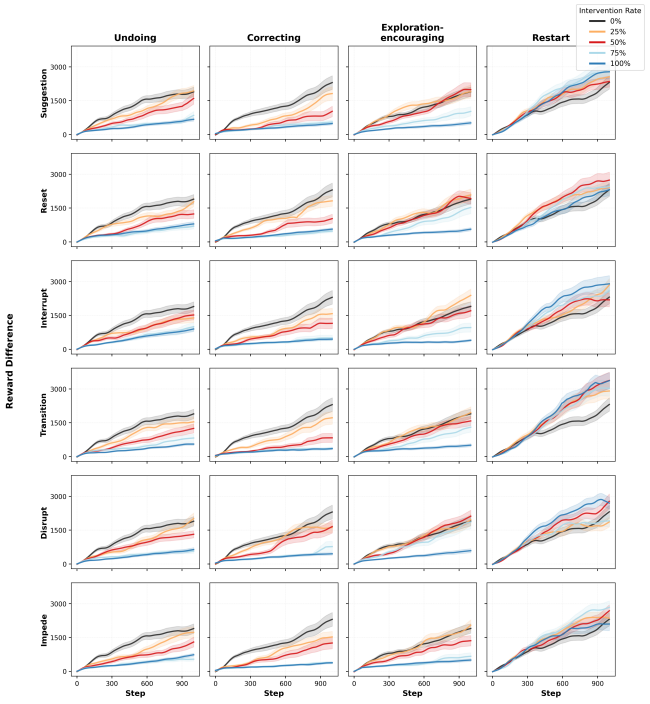
Discussion

This study systematically explores how algorithmic teaching styles, mediated by agent interpretation rules, can catalyze spatial coordination in multi-agent reinforcement learning. Our central finding is that the ***Undoing*** **intervention style serves as a uniquely powerful catalyst for multi-agent coordination**, most effectively breaking behavioral symmetry between two agents. The overall effectiveness of pedagogical intervention, however, is governed by a **pedagogical matching principle**: successful coordination requires coherence between the teacher's external action and the learner's internal heuristic. Our results provide concrete evidence for this principle by revealing critical *mismatches*. The *Reset* interpretation consistently led to poor performance across all teaching styles, indicating a universal failure to align with pedagogical intent. More specifically, a severe mismatch occurred when agents using the *Disrupt* interpretation (viewing all interventions as negative) were subjected to the *Correcting* style, which actively undermined the very coordination it aimed to promote. This pedagogical matching principle shares conceptual roots with information-theoretic accounts of communication, which posit that efficiency is maximized when the encoder and decoder utilize a shared codebook. However, our findings highlight a critical distinction: unlike digital signals with fixed mappings, **state interventions are inherently multivalent and socially ambiguous**. A physical relocation does not carry an intrinsic semantic label; its meaning—whether as a suggestion, a penalty, or a neutral reset—is socially constructed through the learner's heuristic interpretation of the teacher's latent intent. Our discovery of specific failure modes (e.g., the *Disrupt-Correcting* mismatch) demonstrates that even "accurate" physical guidance can impede coordination if the agent assigns a conflicting valence to the teacher's action. This suggests that pedagogical success in multi-agent systems is not merely a matter of signal alignment, but of managing the *intentionality* ascribed to physical interventions.

Crucially, the coordination-promoting effect of appropriate teaching scales to more complex multi-agent societies. In three-agent systems, both *Undoing* and *Correcting* styles robustly reinforced the emerging spatial niches, significantly increasing agents' attachment to their primary regions. This demonstrates that targeted state interventions can transcend simple pairwise coordination and act as a *positive feedback loop* that solidifies collective organization. The failure of the *Restart* style across both two- and three-agent scenarios further underscores that not all interventions are help-



(a) Policy Divergence. Larger values reflect better coordination.



(b) System Equilibrium. Smaller values reflect greater fairness.

Figure 2: Policy Divergence and System Equilibrium for Two-Agent Experiments. Rows and columns vary in intervention interpretations and teaching styles, respectively. Error bands encode standard error. An appropriate teaching style accelerates coordination among agents, allowing each agent to focus on a specific region with balanced scoring.

ful; pedagogical success requires the external guidance to be structured in a way that is computationally interpretable and aligned with the goal of reducing competitive overlap.

These requirements for interpretability and goal alignment are met by the distinct pedagogical policies that our teaching styles operationalize. The *Undoing* style, which forcibly returns the agent to its previous state, emerged as a uniquely powerful catalyst for multi-agent coordination. Its efficacy does not stem from conveying a specific reward or penalty, but from imposing a strong *spatial constraint*. By repeatedly resetting an agent's position, it effectively defines a "home region" and increases the cost of venturing into areas dominated by others. This mechanism rapidly breaks behavioral symmetry and enforces spatial niche formation—a crucial step in solving the "clash of interests" in collective learning. This finding demonstrates how structured external intervention can *accelerate and stabilize* the emergence of cooperative equilibria that self-interested learners might otherwise discover slowly or unreliably (Austerweil et al., 2015).

Conversely, the *Correcting* style functions more as a **stabilizing force**. Once a basic spatial division has emerged, *Correcting* is sufficient to reinforce these existing niches by guiding agents back to their respective optimal regions without the high cost of total displacement.

However, its effectiveness as a stabilizer is contingent upon the agent's interpretation, as evidenced by its severe incompatibility with the *Disrupt* rule. This observation refines the pedagogical matching principle: a well-intentioned guiding correction can fail if the learner misinterprets it as a cautionary demonstration.

Beyond the computational domain, these dynamics have direct implications for human-robot interaction and shared autonomy. For instance, state interventions are analogous to a human teacher physically guiding a robotic manipulator or a semi-autonomous vehicle "snapping" a user back to a safe trajectory. Our results suggest that the efficacy of such haptic or spatial guidance depends heavily on the user's internal model: if a corrective nudge is misinterpreted as a system failure or a restrictive penalty (the *Disrupt* heuristic), it may trigger counter-productive behavioral responses. Furthermore, our findings on spatial specialization suggest that external intervention could be a vital tool for organizing large-scale multi-robot swarms, where breaking initial symmetry is often a bottleneck for collective efficiency.

Together, these insights suggest a potential trajectory for adaptive pedagogical systems: transitioning from symmetry-breaking interventions (*Undoing*) to refinement-oriented guidance (*Correcting*) as the multi-agent system matures, while remaining mindful of the agents' evolving interpretative frameworks.

Future Work

The current study establishes that state interventions, particularly the *Undoing* style, can act as a powerful catalyst for spatial coordination under specific conditions (smooth resource distributions, tabula rasa learners). A critical trajectory for fu-

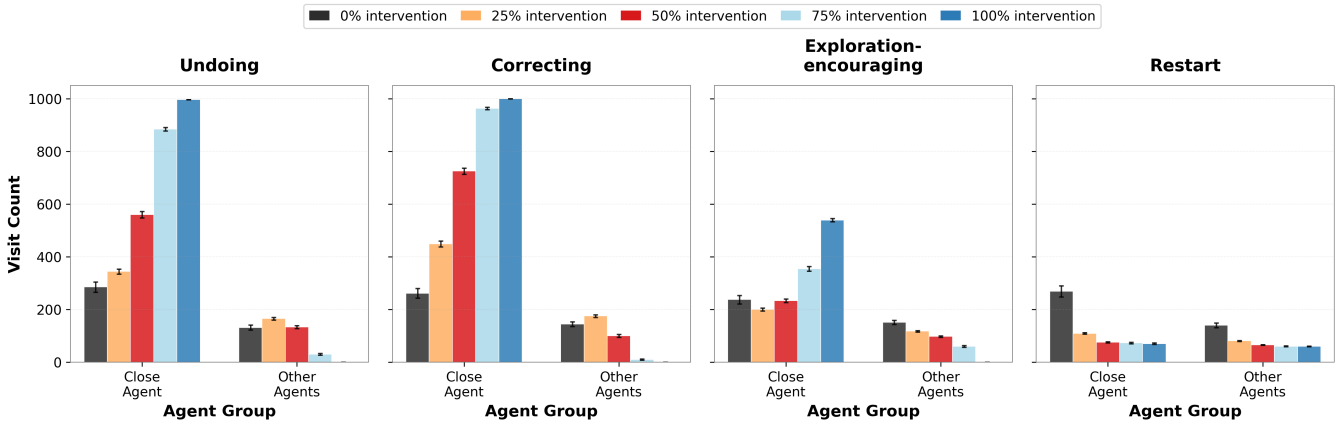


Figure 3: Three-agent simulation results. Visit count of agents for different teaching styles. An appropriate teaching style can enhance the coordination among multiple agents.

ture research lies in testing the boundary conditions and generalizability of these effects by introducing greater ecological complexity.

First, transitioning from fully observable grid worlds to Partially Observable Markov Decision Processes (POMDPs) would introduce realistic perceptual uncertainty. In such settings, state interventions could function not only as spatial corrections but also as *information-shaping signals* that reduce a learner's belief entropy about the reward landscape. The teacher's role would thus evolve from trajectory correction to active epistemic guidance, where interventions like *Undoing* might serve to prune sub-optimal hypotheses or direct attention. Exploring whether the catalytic effect of *Undoing* persists, or how the required **pedagogical matching** between style and interpretation changes under partial observability, is a key next step.

Second, the robustness of our identified mechanisms across different environmental structures remains to be explored. For instance, would *Undoing* remain the dominant catalyst in environments with fragmented or dynamically shifting resources? Investigating how the effectiveness of different teaching styles varies with environmental topology will clarify the scope of our findings and inform the design of context-aware pedagogical agents.

Furthermore, the scalability and inherent heterogeneity of multi-agent systems present a vital frontier. Beyond the small, homogeneous groups explored here, larger populations introduce complex collective dynamics with diverse learning rates, sensory capabilities, or intrinsic motivations. Future work should focus on how a teacher (centralized or distributed) can use state interventions to manage such a "heterogeneous classroom"—enforcing specialization, preventing maladaptive cascades, and maintaining a robust socio-technical equilibrium. This entails understanding how to balance individual learning efficiency with global stability as the system scales.

Ultimately, such research is essential for designing adaptive autonomous systems capable of learning and cooperating effectively from structured guidance in high-stakes, real-world environments.

Acknowledgments

We thank the anonymous reviewers for their insightful feedback on earlier drafts of this manuscript. JLA was funded by the Japanese Probabilistic Computing Consortium Association (JPCCA). **All experimental code and data analysis scripts are publicly available at: `https://github.com/ZhuolunZhong/Physical-Intervention-with-multi-agent-system-computation.git`.**

References

Austerweil, J. L., Brawner, S., Greenwald, A., Hilliard, E., Ho, M., Littman, M. L., ... Trimbach, C. (2015). The impact of other-regarding preferences in a collection of non-zero-sum grid games. In *Aaai spring symposium 2016 on challenges and opportunities in multiagent learning for the real world.*

Axelrod, R. (1984). *Cooperation*. New York: basic books.

Foerster, J., Assael, I. A., De Freitas, N., & Whiteson, S. (2016). Learning to communicate with deep multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems* (Vol. 29).

Foerster, J., Farquhar, G., Afouras, T., Nardelli, N., & Whiteson, S. (2018). Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 32).

Ho, M. K., Cushman, F., Littman, M. L., & Austerweil, J. L. (2019). People teach with rewards and punishments as communication not reinforcements. *Journal of Experimental Psychology: General*, *148*(3), 520–549.

Ho, M. K., Cushman, F., Littman, M. L., & Austerweil, J. L. (2021). Communication in action: Planning and interpreting communicative demonstrations. *Journal of Experimental Psychology: General*, *150*(11), 2246–2272.

Ho, M. K., Littman, M. L., & Austerweil, J. L. (2017). Teaching by intervention: Working backwards, undoing mistakes, or correcting mistakes? In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* (pp. 526–531). Austin, TX: Cognitive Science Society.

Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, *101*(1-2), 99–134.

Kingman, J. F. C. (1992). *Poisson processes*. Clarendon Press.

Losey, D. P., Bajcsy, A., O'Malley, M. K., & Dragan, A. D. (2022). Physical interaction as communication: Learning robot objectives online from human corrections. *The International Journal of Robotics Research*, *41*(1), 20–44.

Lowe, R., WU, Y., Tamar, A., Harb, J., Abbeel, P., OpenAI, & Mordatch, I. (2017). Multi-agent actor-critic for mixed cooperative-competitive environments. In I. Guyon et al. (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc.

Luo, J., Xu, C., Wu, J., & Levine, S. (2024). Precise and dexterous robotic manipulation via human-inthe-loop reinforcement learning. *arXiv preprint arXiv:2410.21845*, *2*(3).

Rafferty, A. N., Brunskill, E., Griffiths, T. L., & Shafto, P. (2016). Faster teaching via POMDP planning. *Cognitive Science*, *40*(6), 1290–1332.

Rashid, T., Samvelyan, M., De Witt, C. S., Farquhar, G., Foerster, J., & Whiteson, S. (2020). Monotonic value function factorisation for deep multi-agent reinforcement learning. *Journal of Machine Learning Research*, *21*(178), 1–51.

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*, *71*, 55–89.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Wolpert, D. H., & Tumer, K. (2001). Optimal payoff functions for members of collectives. *Advances in Complex Systems*, *4*(02n03), 265–279.

Wu, C. M., Deffner, D., Kahl, B., Meder, B., Ho, M. H., & Kurvers, R. H. (2025). Adaptive mechanisms of social and asocial learning in immersive collective foraging. *Nature communications*, *16*(1), 3539.