Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Observational Learning with Gated Information: A Lightweight RL Simulation Testbed and Latent Signal Models

**Chengyuan Zhu**
Nanchang Hangkong University
**Jialai Wu**
wujailai106@qq.com
Northeastern University USA

**Research Article**

**Keywords:**

**Posted Date:** March 12th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-9091145/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** The authors declare no competing interests.

# Abstract

Observational learning can be framed as active information sampling: an agent must decide when to spend time watching another agent to reduce uncertainty about when reward will be available. We present a compact simulation testbed for this social-timing problem in which (i) a demonstrator produces brief, stochastic cue bursts, (ii) a short reward window opens after each burst ends, and (iii) the observer can only access social timing features by choosing an explicit Observe action, with imperfect cue detection and a noisy window-remaining estimate. Building on the released multi-algorithm pipeline (dueling double DQN baseline, PPO actor-critic, and a tabular Q-learning control), we add controlled manipulations of social sensing and teacher dynamics: (1) a social-blind observer with the social channel disabled, (2) an impaired tracking observer (reduced detection reliability) with reduced detection reliability and noisier timing estimates, (3) a familiarity variable that gradually improves the observer's effective sensing with repeated exposure, and (4) teacher palatability conditions that change cue-burst (consumption) duration distributions. We compare learning speed and reward-related latent signals across these conditions by logging behavior and synthesizing candidate internal variables (TD error/RPE, observation-driven value updates, action salience, and information-gain impulses), optionally filtered into signal-like traces and analyzed with peri-event averages. The framework provides a lightweight, code-aligned template for diagnosing observational RL policies and for testing which latent variables best explain event-aligned "teaching signals" under different social-information constraints.


# Introduction

Many real-world agents learn faster by watching others. A robot can watch a peer interact with an object to infer affordances; a game agent can monitor another player's actions to predict upcoming opportunities; and in general, an observer can use another agent as a noisy sensor that reveals hidden state (e.g., when a resource becomes available). These problems are naturally framed as partially observable decision processes in which information is costly and must be actively sampled.

This paper focuses on a minimal but expressive "active observational learning" structure: a demonstrator emits a short cue burst that predicts a brief reward window for the observer. The observer must decide when to spend time observing to reduce timing uncertainty versus when to commit to acting. Importantly, the social channel is gated—the cue and the estimate of remaining window time are only available on Observe steps—so the task explicitly implements observation as an action with opportunity cost.

A second motivation is interpretability. Even when a policy succeeds, different internal learning variables can drive behavior: classical temporal-difference (TD) errors, value updates triggered by newly revealed information, action-initiation salience, or information-gain signals that scale with surprise. These alternatives make distinct, testable predictions about event-aligned transients and how they evolve from early exploration to late exploitation.

Contributions: (i) a fast observational-timing environment with explicit Observe-as-action gating and bout-locked reward windows; (ii) a multi-algorithm learner suite (dueling double DQN baseline, PPO actor-critic, and a tabular Q-learning control) sharing the same logging and evaluation code; (iii) controlled condition manipulations for social information access (social-blind), impaired social tracking (reduced detection reliability), familiarity-dependent sensing, and teacher palatability (short vs long cue bursts); and (iv) an algorithm-agnostic latent-signal synthesis pipeline (TD error/RPE, observation-driven value updates, salience, information gain) with optional filtering and peri-event analysis.

From an AI standpoint, the learner's decision problem is naturally framed as a partially observable Markov decision process (POMDP) with an action-dependent observation model and explicit information costs: the agent must trade off task performance against when and how often to acquire informative social cues. This links the task to classic POMDP ideas [3], value-of-information analyses [12], and rational inattention in decision-making [13]. Importantly, our focus is not on copying the demonstrator's actions, but on learning an attention policy that acquires just enough information to exploit time-locked social contingencies.

## Related work

Active observational learning sits at the intersection of (i) learning from demonstrations / social learning, where an agent improves by leveraging another agent's behavior [9-11,14-16], and (ii) active sensing / value-of-information settings, where observations are partial, optional, and costly [3,12,13]. Our formulation emphasizes when to attend and what to extract from brief social events, rather than directly imitating the demonstrator's actions.

From an RL perspective, this environment is a compact, interpretable benchmark for attention-gated decision making. It can be solved by value-based methods (e.g., DQN variants [4-6,8,9]) or actor-critic methods (e.g., PPO [7]). Unlike standard MDP benchmarks, the key uncertainty is not hidden state dynamics but the availability and reliability of social information, which the learner can sample via an explicit OBSERVE action (a POMDP-style control of the observation model [3]).

We intentionally avoid cooperation, explicit communication, or shared rewards: the demonstrator is an exogenous stochastic process that generates bouts, and the learner's reward is purely individual. This design isolates observational timing from coordination and multi-agent strategy learning, which are central but separate problems in multi-agent RL [26,27].

# 2.1 Learning from demonstrations and learning from observation

A large literature studies how agents can learn from others via behavior cloning, inverse reinforcement learning, and more general imitation learning (IL). In IL, the target is typically to match or generalize the demonstrator's actions and trajectories; methods such as dataset aggregation (DAgger) stabilize

learning when the learner's state distribution drifts from the demonstrator's distribution [14], and adversarial formulations such as GAIL can learn policies that reproduce expert occupancy measures [15]. More recently, learning from observation relaxes the assumption that demonstrator actions are labeled, enabling policy learning from state-only videos [16]. Our setting is complementary: the learner does not need to reproduce the teacher's actions (bouts are not a target behavior), but instead must infer when the environment becomes temporarily exploitable. This makes attention allocation—rather than action matching—the core computational problem.

## 2.2 Active sensing, rational inattention, and costly information

Explicit costs for observing or querying the environment appear in classic value-of-information theory [12] and in POMDP formulations where actions can control what is observed [3]. Economic and decision-theoretic accounts of rational inattention similarly trade reward against information processing [13]. Unlike intrinsic-motivation or curiosity frameworks (which use novelty/uncertainty bonuses to drive exploration) [17-20], our environment provides externally anchored information: the signal is generated by a social partner and has direct instrumental value because it gates access to a short reward window.

## 2.3 Links to computational neuroscience

Although our analysis uses AI terminology, the task was designed to map cleanly onto animal paradigms where neuromodulatory signals (e.g., dopamine) track learned expectations and guide action selection. Canonical reinforcement-learning accounts interpret phasic dopamine as a reward prediction error [22,23], and the vigor/latency of reward seeking is shaped by opportunity costs and effort [25]. We therefore treat "DA activity" as an algorithmic reward-prediction proxy, and ask how different assumptions about social cue reliability and internal costs (familiarity, empathic shaping) produce distinct trajectories in both behavior and simulated reward signals that can be compared to photometry.

## 3.1 Active observational timing environment

We simulate an episode as a 1D grid world with a discrete time step dt = 0.5 s. A demonstrator process generates periodic, short cue bursts (implemented in code as "licking bouts"; period 30 s with uniform jitter ±6 s). Each burst has a stochastic duration drawn from a gamma distribution with mean 1 s (shape = 2, scale = mean/shape). The end of a burst opens a brief reward-availability window of 3 s (6 steps) for the observer.

An observer agent occupies a grid position and has its own fixed resource site distinct from the demonstrator's position. At each step, the agent selects one of five actions: move left, stay, move right, Observe, or Consume. Consuming has a cooldown of 2 s (4 steps), and both Observe and Consume

carry small step costs. The core manipulation is information gating: the demonstrator cue and a noisy estimate of the remaining reward-window time are only revealed on Observe steps.

On an Observe step, cue detection is imperfect: during an active burst, the agent detects the cue with probability p = 0.7 per Observe step and thereby “registers” the current event. The agent also receives a normalized estimate of the remaining window time (true remaining steps plus small integer noise). On non-Observe steps, these social features are set to zero. This creates the intended tradeoff: observing reduces timing uncertainty but consumes time that could be used to reach the resource site and consume within the window.

Reward is delivered at most once per event. A reward of +1 is given when the agent (i) is at its own resource site, (ii) consumes during the open window after burst termination, and (iii) has registered that event by detecting the cue during the burst (registered event id matches the last-cue event id). Optional shaping adds a small bonus for observing during a burst before it is registered, encouraging attention early in training.

## 3.2 State representation

The agent observes a 6-dimensional state vector (matching the released code): (1) normalized agent position, (2) normalized resource-site position, (3) gated cue indicator (non-zero only on Observe steps), (4) gated (noisy) reward-window remaining estimate (non-zero only on Observe steps), (5) a phase-to-next-nominal-bout feature (periodic context clock), and (6) normalized consume cooldown. In familiarity-enabled variants, we optionally append the familiarity scalar $f \in [0,1]$ as a seventh feature to allow the learner to condition its Observe/act tradeoff on accumulated exposure; alternatively, f can be treated as a latent environment variable that only affects observation quality.

## 3.3 Deep Q-learning agent

We provide a strong value-based baseline using a dueling double DQN. The Q-network is a two-layer MLP (hidden size 128) producing a scalar value stream V(s) and an advantage stream A(s,a), combined as $Q(s,a) = V(s) + (A(s,a) - \text{mean}_a A(s,a))$.

Training uses an experience replay buffer (capacity 50,000 transitions), optimized with Adam (learning rate $3\times10-4$), discount factor $\gamma = 0.99$, batch size 256, a Huber loss, and gradient clipping (norm 10). We employ Double DQN updates [5] with a dueling value/advantage decomposition [6], and stabilize learning with a slowly updated target network (Polyak averaging, $\tau = 0.02$).

Alternative learner models (drop-in): Although our released implementation uses DQN for speed and stable value latents, the environment is compatible with other common RL learners. (1) Actor-critic methods such as A2C/PPO can learn stochastic Observe-Act policies and provide policy-gradient-based

latents (advantages, entropy) as candidate internal signals. (2) Recurrent agents (e.g., DRQN/R2D2-style) can integrate the gated cue stream over time and are a natural fit when the periodic phase feature is removed, making the task more strictly POMDP. (3) Model-based timing/belief baselines can learn a hazard model for event onset/offset and plan with an explicit belief over window remaining time. In Section 3.5 we define latent signal generators in a way that can use either value-based quantities (TD error, ΔV on observation) or actor-critic quantities (advantage-like errors).

## 3.4 Training and behavioral metrics

Training runs for a fixed number of episodes (default 1200), each lasting 120 s (240 steps). We log step-wise events and compute per-episode summary metrics: bout success rate (successful rewarded bouts / total bouts), mean episodic reward, Observe rate, Observe-detect rate (fraction of Observe steps that successfully detect the cue), and eat-in-window rate (fraction of steps with Consume taken while the window is open). For social-blind controls (or when comparing pure timing ability), we also compute an auxiliary asocial window-hit metric that counts Consume-at-site-in-window events regardless of whether the bout was registered. Learning speed is summarized by (i) time-to-threshold (episodes required to exceed a fixed success level) and (ii) the area under the success-curve over training.

## 3.5 Candidate reward-related latent signal hypotheses

To synthesize candidate latent “teaching signals”, we combine step rewards with learner-specific quantities such as state values, advantages, or belief updates, and we gate certain components on Observe steps to reflect information availability. In the current DQN implementation, we use a value proxy $V(s)=\max_a Q(s,a)$; for actor-critic learners, the same constructions can be expressed with a critic $V(s)$ and advantage estimates.

• Temporal-difference error (TD error): $\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t)$.

• Observation-driven value update (cue-update): on Observe steps, the value jump caused by revealing gated cue features, $\Delta V_t = V_{post} - V_{prior}$, where $V_{prior}$ is computed with cue features zeroed.

• Cue value (availability-gated): the same $V_{post} - V_{prior}$ term treated as a value-like signal that only exists when the cue channel is sampled (Observe).

• Action-initiation salience: impulses at Observe and Consume onsets, capturing initiation/vigor signals independent of outcome.

• Information gain/surprise: impulses when an Observe step successfully detects the cue, scaled by surprise under a timing prior (implemented using the phase-to-next-event feature, with the prior sharpening from early to late learning).

## 3.6 Signal-like filtering and event-aligned analysis

Impulse sequences can be mapped to smooth signal-like traces using a simple causal kernel (e.g., a bi-exponential rise/decay filter). This step is optional for pure-AI analysis, but is useful when comparing latent impulses to any measurement channel with slow dynamics (e.g., low-pass sensors or neural recordings). We compute peri-event averages aligned to (i) event termination (cue offset; estimated by the falling edge of the demonstrator cue), (ii) reward delivery (valid Consume), and (iii) Observe steps that successfully detect the cue. To visualize learning-stage effects, peri-event averages are compared between early episodes (first 20%) and late episodes (last 20%).

## 3.7 Condition manipulations for social access, familiarity, empathy, and teacher palatability

Observer phenotypes. We evaluate three observer variants that preserve the same action set and reward logic (Observe is an explicit, time-consuming action; reward is delivered at most once per bout when the agent consumes at its own site during the reward window and has registered the bout by detecting the cue during the burst). The variants differ only in the reliability of the gated social channel: (i) Typical observers use the default detection probability and window-estimate noise; (ii) impaired-tracking observers have reduced cue detection probability and increased timing-estimate noise; (iii) social-blind observers have the social channel disabled so Observe steps never reveal cue or window-remaining features. The social-blind condition serves as a negative control for “social” success; we additionally report an auxiliary asocial timing metric (eat-in-window regardless of registration) to quantify whether agents can exploit periodic structure without social sensing.

Familiarity and latent “bonding”. When enabled, familiarity is a latent scalar $f \in [0,1]$ that is updated across episodes when the learner chooses Observe during demonstrator bouts (with larger updates when a cue is successfully seen). Familiarity has two linked roles. First, it improves the social information channel by increasing the probability of cue detection and reducing sensory timing noise on the perceived cue. Second, it reshapes the agent’s cost/benefit trade-off for attention: the opportunity cost of Observe is modestly reduced and the internal value of successfully seeing a cue is increased. In the current implementation we additionally apply small shaping terms that (i) bias the agent to move toward its own food patch when a reward window is open and (ii) discourage over-observation during the reward window, helping to dissociate “attend-to-learn” from “act-to-consume”. All of these effects are controlled by the familiarity flag and can be ablated to isolate channel-quality vs. shaping contributions.

Empathy / intrinsic social valuation. To model a learner that assigns intrinsic value to social signals (i.e., a “social salience” drive that is separable from external food reward), we optionally add a small intrinsic reward when the learner successfully sees a demonstrator cue during a bout. This empathy bonus increases the incentive to Observe and can raise cue-sighting rates even when external reward is unchanged, allowing us to dissociate attention/engagement from reward harvesting. Because Observe

consumes time, strong empathy can also induce a trade-off with timely eating, potentially increasing response latency or slightly reducing bout success depending on weighting.

Teacher palatability manipulation. We manipulate “demonstrator palatability” by changing the temporal statistics of demonstrator lick bouts. In the current implementation, high palatability corresponds to longer mean bout durations (e.g., mean≈1.4 s), which provides a longer window for the learner to sample the cue stream during a bout. Low palatability corresponds to shorter bouts (e.g., mean≈0.6 s), making cues sparse and observation more time-critical. Critically, palatability here modulates the opportunity to observe (cue availability in time), not the magnitude of the learner’s external food reward.

Evaluation protocol. We evaluate learner performance under controlled factorial manipulations of demonstrator palatability ($T \in \{low, high\}$), observer phenotype ($L \in \{normal, blind, impaired\}$), and familiarity ($F \in \{off, on\}$); we also evaluate an empathy manipulation ($E \in \{off, on\}$) that adds intrinsic value to successfully seeing cues. For efficiency, the main manuscript focuses on the core 3-way factorial ($T \times L \times F$), while empathy effects are reported as targeted comparisons (Supplementary Fig. S3). For each condition we train multiple random seeds with identical hyperparameters and summarize learning curves using exponentially weighted moving averages (EMA), reporting mean±SEM across seeds and simple permutation tests for between-condition differences in final performance.

Note: the impaired-tracking condition is an algorithmic sensory/attention ablation (reduced and noisier social observations) and is not intended as a clinical model.

# Results

## 4.1 Learned observe–act policies in an observational timing task

The gated-observation contingency produces the intended exploration–exploitation tradeoff: Observe consumes a time step and is the only way to access social timing information, but acting too often (moving/eating) without observing risks missing the correct window. In the baseline condition (typical observer; default teacher bout statistics), the DQN learner increases bout success steadily over training (Figure 2), indicating that it learns to coordinate brief information sampling with rapid relocation to the resource site and timely Consume actions.

Behaviorally, the learned strategy increases its overall Observe rate early in training (reaching ~0.4 Observe/step in this run) while the Observe-detect rate remains low because cue detection is probabilistic and cue bursts are brief. As the policy improves, the eat-in-window rate also rises (peaking around ~0.02 Eat-in-window/step), and bout success reaches ~0.45 before partially degrading late in training in this particular seed (Figure 2). These diagnostics help separate “watching a lot” from “watching at the right times,” and motivate the condition manipulations below (Figure 5) that directly change observation reliability and teacher bout duration.

Viewed through the lens of active sensing, the learned policy implements sparse sampling: it concentrates OBSERVE actions during periods of high expected information value (teacher bouts) and suppresses observing elsewhere, where the same action cost yields little utility. This pattern is consistent with value-of-information and rational-inattention accounts of decision making [12,13].

## 4.2 Divergent predictions for latent teaching-signal hypotheses

The candidate latent-signal models make distinct predictions about when reward-related transients should occur relative to task events. TD-error-based signals produce strong transients near reward delivery and can shift earlier with learning as predictive information is incorporated into V(s). In contrast, cue-update and cue-value signals produce transients on Observe steps when gated cue features become available—often preceding reward and appearing even on unrewarded attempts when observation resolves timing uncertainty.

The action-initiation salience model predicts transients at Observe/Consume onsets largely independent of outcome, whereas the information-gain model predicts large transients specifically when observing yields an unexpectedly informative cue (high surprise under the timing prior). These contrasts are clearest in peri-event averages aligned to successful Observe-detect events and to cue-burst termination.

Mechanistically, familiarity bundles two factors that are often separated in active sensing models: (i) a reduced price of information (lower observation cost) and (ii) an improved quality of information (higher detection and less timing noise). Both factors increase the expected value of sampling in a way that mirrors value-of-information formulations [12] and rational-inattention accounts that treat information as a limited, costly resource [13].

## 4.3 Practical use for model comparison and interpretability

Because each hypothesis yields a full predicted latent impulse train (and an optional smoothed trace), the framework can be used as a forward model for algorithm diagnostics and, when available, comparing against recorded signals with slow dynamics. For example, one can regress an observed trace onto model-derived predictors (TD error, $\Delta V$ on observation, salience, surprise) with cross-validation or compare hypotheses via their ability to reproduce learning-stage shifts in event-aligned responses.

The empathy manipulation can be interpreted as an explicit social utility term or, algorithmically, as reward shaping that adds value to cue-contingent social events. Reward shaping is a standard technique

for accelerating learning without changing the optimal policy when implemented in a potential-based form [21], and related ideas appear in multi-agent RL where auxiliary social rewards encourage prosocial or information-seeking behaviors [26,27]. Here, empathic shaping selectively increases the marginal value of attending to another's consummatory bout, thereby shifting the learned attention budget toward social moments.

## 4.4 Condition comparisons: social blindness, impaired tracking, familiarity, empathy, and palatability

We evaluate how learning and reward-related latent signatures change when access to the social channel is degraded (social-blind and impaired-tracking observers), when repeated exposure improves the information channel (familiarity), when intrinsic social salience biases attention (empathy), and when the demonstrator's cue bursts are shorter or longer (teacher palatability). The supplementary simulations included with this revision illustrate these effects: familiarity increases asymptotic bout success and cue sightings; empathy increases observation/cue sightings but can slightly reduce bout success by promoting over-attendance; and low palatability (short bursts) can collapse learning by making observation time-critical. Figure 5 summarizes the design and parameterization; Supplementary Figures S1–S4 show the full curves.

Prediction 1 (social blindness): disabling the social channel should collapse social bout success (because bout registration becomes impossible), while leaving an asocial timing metric (eat-in-window) largely governed by periodic structure and exploration. This isolates which parts of performance truly depend on social sensing.

Prediction 2 (impaired tracking and familiarity): reducing cue detection and increasing timing noise slows learning and reduces observation efficiency. Introducing familiarity partially rescues performance via two mechanisms: (i) an improving sensory channel (higher effective p_detect, lower timing noise) and (ii) modest shaping that lowers the opportunity cost of Observe and discourages over-observation during the reward window, biasing the agent to transition from attending to eating. Together, these changes should produce earlier rises in bout success and cleaner peri-event signatures.

Prediction 3 (empathy / intrinsic salience): adding a small intrinsic bonus for successfully seeing the cue should increase Observe rate and cue-sighting rate even when extrinsic reward has not yet been mastered. Depending on its weight, this can either speed early learning (by encouraging information sampling) or slightly reduce bout success late in training (if attention competes with timely eating), providing a principled dissociation between social engagement and reward harvesting.

Prediction 4 (teacher palatability): in our implementation, high palatability corresponds to longer cue bursts (longer mean bout duration), providing more opportunities to Observe during the bout and thus higher bout success and lower latency. Low palatability corresponds to short bursts, making observation

time-critical and amplifying differences between typical and impaired observers; under this regime, familiarity should provide a larger relative benefit.

Across all conditions, we compare not only behavior but also latent reward-related signatures. For example, Figure 4 shows a representative TD-error/RPE peri-event average aligned to cue-burst termination (lick_end): the late-training RPE transient is larger and sharper than early training, consistent with improved prediction of imminent reward availability. Analogous peri-event comparisons can be computed for reward delivery and Observe-detect events to test which latent hypotheses best match recorded signals.

# Discussion

## 5.1 Active sensing interpretation

Our results can be summarized as learning a policy over information acquisition. Because the reward window is brief and cannot be exploited without observing, the agent faces a classic trade-off: observing too rarely misses windows, while observing too often wastes time and incurs explicit costs. In value-of-information terms [12], the learned strategy approximates a context-dependent threshold rule in which the marginal value of sampling rises sharply during bouts and during the final seconds when the window is expected to open; outside these moments, the rational action is to conserve attention. This perspective aligns with rational-inattention accounts that treat information as an economic resource with a “price” [13], and it suggests analytic extensions where the learner maintains an explicit belief state (bouts/no-bouts) and chooses observation rates accordingly [3].

## 5.2 Relation to imitation learning and multi-agent RL

While our demonstrator is another agent, the setting is intentionally not imitation learning. The teacher’s policy is exogenous and the learner’s goal is not behavioral matching; instead, the teacher provides intermittent, instrumentally useful information. This complements learning-from-observation methods [11], which target action policies from state-only trajectories, and suggests hybrid extensions: one can couple our “attention-gated” objective with imitation losses (e.g., behavior cloning or GAIL) to study how agents balance copying, inference, and exploitation in social settings [14, 15]. Conversely, if rewards were shared or the teacher adapted to the learner, the problem would transition toward genuine multi-agent RL with strategic coupling [26, 27].

## 5.3 Neuroscience-facing predictions

Because the environment exposes a clean latent variable (expected reward-window timing), it yields concrete predictions for reward-like signals under different hypotheses. In a TD/RPE interpretation [22, 23], “DA activity” should ramp or peak around the inferred bout end, reflecting a rise in state value as the reward window becomes imminent; missed windows should produce negative prediction errors at window close (consistent with phasic decreases when expected rewards fail to occur) [24]. Familiarity should compress observation-to-action latency by lowering the effective cost of information and

reducing temporal uncertainty, shifting reward prediction earlier and reducing trial-to-trial variability. Empathic shaping adds a social utility term that predicts additional positive transients during observed bouts even when no direct reward follows, analogous to auxiliary value signals that can bias vigor and engagement [25].

## 5.4 Extensions

This work provides a deliberately minimal simulation testbed for active observational learning: an exogenous demonstrator emits brief, time-localized cue bursts that gate a short reward-availability window for an observer. By making observation an explicit, costly action, the task isolates the central tradeoff between sampling information and acting under time pressure.

A key emphasis is algorithm-level interpretability. The same successful behavior can be supported by different latent quantities—TD errors, value jumps triggered by newly revealed information, initiation salience, or information-gain/surprise. Embedding these hypotheses in an environment with explicit information gating makes their event-aligned predictions sharply separable, and enables principled model comparison across learners and across learning stages.

Limitations are straightforward: the demonstrator is non-adaptive, the environment is 1D, and the latent signal models are intentionally stylized. However, each component is extensible: richer observation channels, multi-dimensional navigation, removal of the periodic phase feature (to enforce stronger partial observability), adaptive demonstrators, or multi-agent extensions can be added without changing the core logging and hypothesis-testing workflow.

The current benchmark is intentionally minimalist; several extensions would connect it more tightly to contemporary AI and behavioral neuroscience. First, richer partial observability (occlusions, noisy spatial perception, multiple concurrent events) would motivate explicit belief-state learning and representation learning. Second, scaling from 1D to 2D arenas and adding competing opportunities would test whether agents learn allocative attention policies that resemble active vision. Third, adding intrinsic-motivation terms [17–20] would let one disentangle social-instrumental information seeking from novelty-driven exploration. Finally, making the demonstrator adaptive would open the door to studying emergent communication and coordination in social foraging tasks [26, 27].

# Conclusion

We introduced a compact, code-aligned simulation pipeline for active observational learning with gated information. The framework supports interchangeable RL learners and a family of reward-related latent signal hypotheses, yielding event-aligned predictions that can be used for algorithm diagnostics, interpretability, and (optionally) comparison to smoothed measurement traces.

# Declarations

# Acknowledgments

No external funding was used. Simulations were run on standard compute resources.

# References

1. Sutton RS, Barto AG. Reinforcement Learning: An Introduction. 2nd ed. MIT Press. 2018.
2. Sutton RS. Learning to predict by the methods of temporal differences. Machine Learning. 1988;3:9–44.
3. Kaelbling LP, Littman ML, Cassandra AR. Planning and acting in partially observable stochastic domains. Artificial Intelligence. 1998;101(1–2):99–134.
4. Mnih V, et al. Human-level control through deep reinforcement learning. Nature. 2015;518:529–533.
5. Van Hasselt H, Guez A, Silver D. Deep reinforcement learning with Double Q-learning. AAAI. 2016.
6. Wang Z, Schaul T, Hessel M, van Hasselt H, Lanctot M, de Freitas N. Dueling network architectures for deep reinforcement learning. ICML. 2016.
7. Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal Policy Optimization Algorithms. arXiv:1707.06347. 2017.
8. Hausknecht M, Stone P. Deep recurrent Q-learning for partially observable MDPs. AAAI Fall Symposium. 2015.
9. Kapturowski S, et al. Recurrent experience replay in distributed reinforcement learning (R2D2). ICLR. 2019.
10. Argall BD, Chernova S, Veloso M, Browning B. A survey of robot learning from demonstration. Robotics and Autonomous Systems. 2009;57(5):469–483.
11. Bandura A. Social Learning Theory. Prentice Hall. 1977.
12. Howard, R. A. Information Value Theory. IEEE Transactions on Systems Science and Cybernetics, 1966.
13. Sims, C. A. Implications of rational inattention. Journal of Monetary Economics, 2003.
14. Ross, S., Gordon, G., and Bagnell, J. A. A reduction of imitation learning and structured prediction to no-regret online learning. Proceedings of AISTATS, 2011.
15. Ho, J. and Ermon, S. Generative Adversarial Imitation Learning. Advances in Neural Information Processing Systems (NeurIPS), 2016.
16. Hussein, A., Gaber, M. M., Elyan, E., and Jayne, C. Imitation Learning: A Survey. arXiv:1707.02748, 2017.
17. Schmidhuber, J. A possibility for implementing curiosity and boredom in model-building neural controllers. Proceedings of the International Conference on Simulation of Adaptive Behavior (SAB), 1991.

18. Oudeyer, P.-Y. and Kaplan, F. What is intrinsic motivation? A typology of computational approaches. Frontiers in Neurorobotics, 2007.
19. Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T. Curiosity-driven exploration by self-supervised prediction. Proceedings of ICML, 2017.
20. Burda, Y., Edwards, H., Storkey, A., and Klimov, O. Exploration by Random Network Distillation. arXiv:1810.12894, 2018.
21. Ng, A. Y., Harada, D., and Russell, S. Policy invariance under reward transformations: Theory and application to reward shaping. Proceedings of ICML, 1999.
22. Schultz, W., Dayan, P., and Montague, P. R. A neural substrate of prediction and reward. Science, 1997.
23. Montague, P. R., Dayan, P., and Sejnowski, T. J. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. Journal of Neuroscience, 1996.
24. Fiorillo, C. D., Tobler, P. N., and Schultz, W. Discrete coding of reward probability and uncertainty by dopamine neurons. Science, 2003.
25. Niv, Y., Daw, N. D., Joel, D., and Dayan, P. Tonic dopamine: opportunity costs and the control of response vigor. Psychopharmacology, 2007.
26. Buşoniu, L., Babuška, R., and De Schutter, B. A comprehensive survey of multiagent reinforcement learning. IEEE Transactions on Systems, Man, and Cybernetics, Part C, 2008.
27. Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, P., and Mordatch, I. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. Advances in Neural Information Processing Systems (NeurIPS), 2017.

# Figures

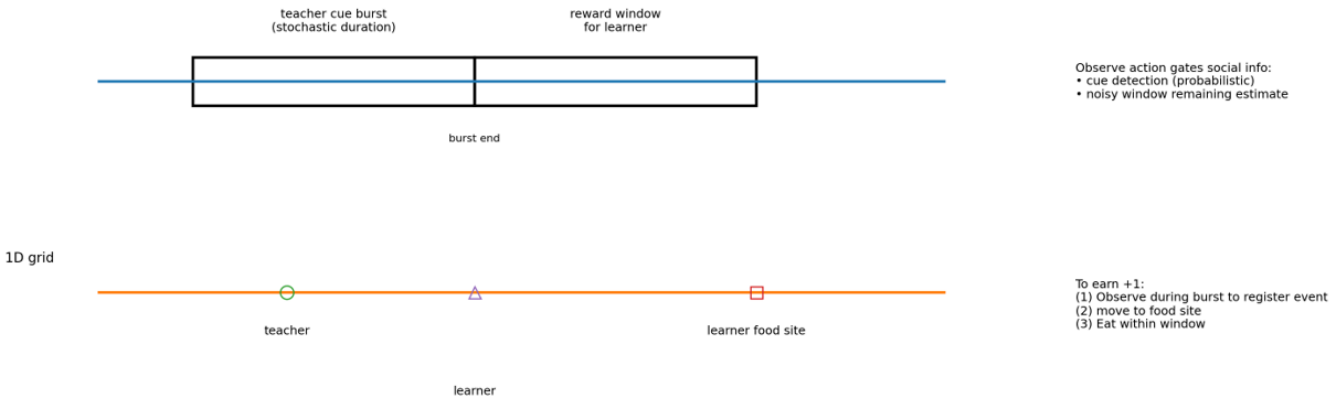
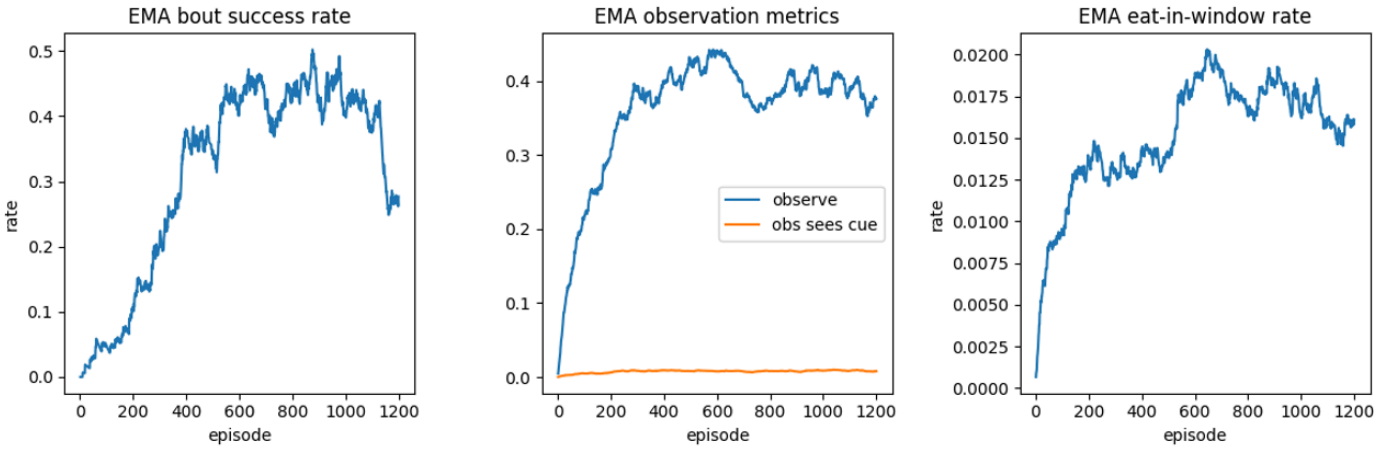



**Figure 1**

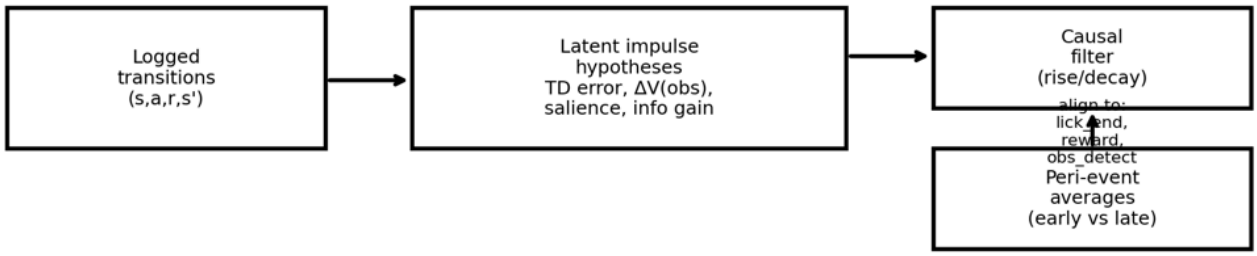
**Task schematic for active observational timing.** The demonstrator emits stochastic cue bursts (teacher "consumption" bouts). A short reward window opens after each burst ends. The observer must choose

Observe actions to access noisy social timing features, then move to its own resource site and Consume within the window to obtain +1 (at most once per bout).




## Figure 2

**Learning curves (baseline DQN).** EMA-smoothed per-episode bout success rate, Observe metrics (Observe rate and Observe-detect rate), and in-window consumption rate (eat-in-window) from the run shown in this paper.




## Figure 3

**Latent-signal synthesis pipeline.** Logged transitions are converted into candidate latent impulses (e.g., TD error/RPE, observation-driven value updates, action salience, information gain), optionally convolved with a causal rise/decay kernel to mimic signal dynamics, and summarized via peri-event averages.

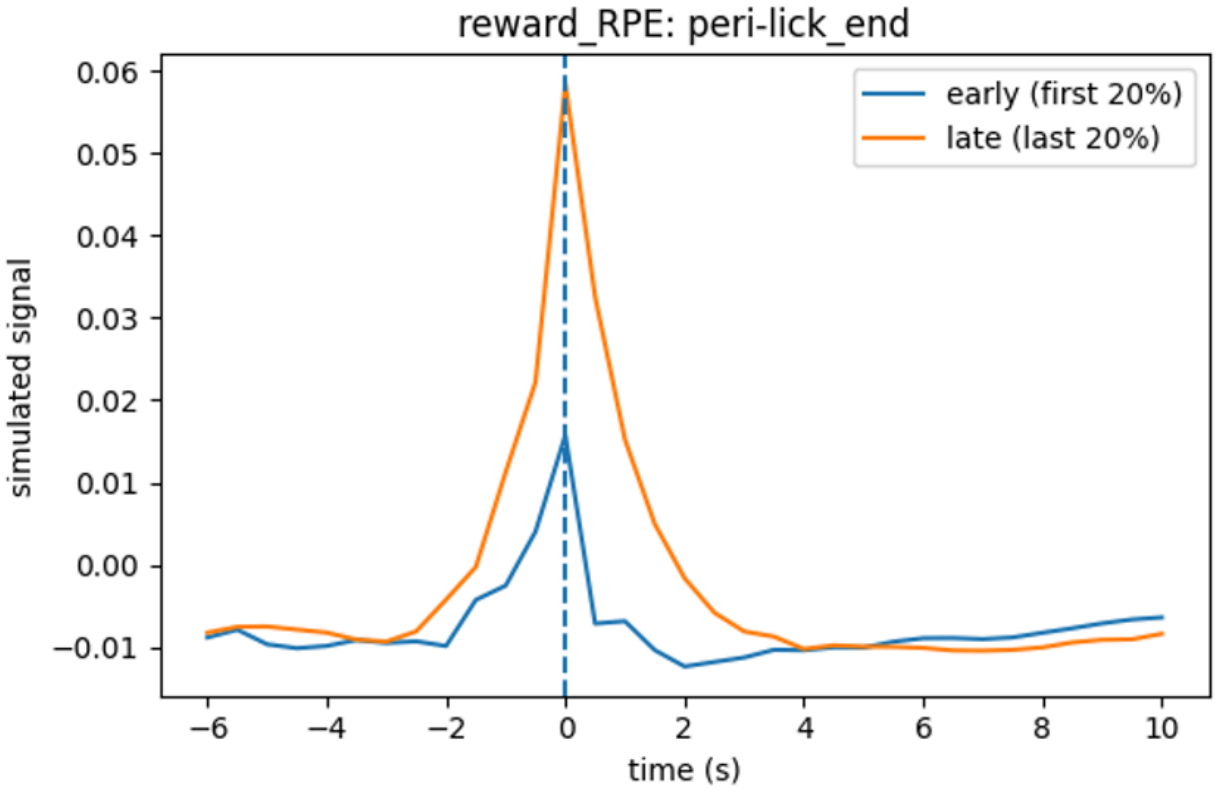



**Figure 4**

**Example peri-event TD-error/RPE prediction aligned to cue-burst termination (lick_end).** Early vs late training peri-event averages illustrate how reward-prediction transients sharpen and increase in amplitude as the policy improves.

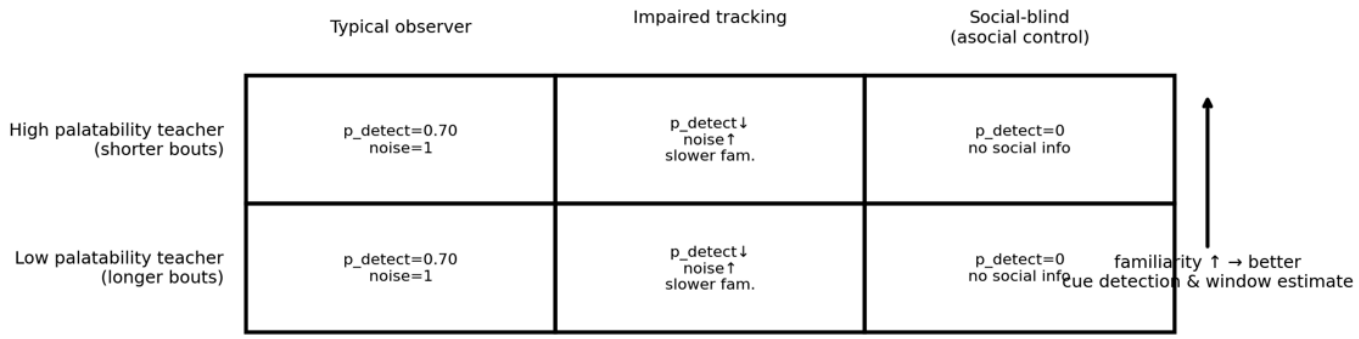



**Figure 5**

**Condition matrix used for controlled comparisons.** We vary teacher palatability by changing the mean duration of demonstrator cue bursts (T=high: longer bursts; T=low: shorter bursts), observer phenotype

(typical, impaired tracking, social-blind), and familiarity (F=on/off). Empathy (E=on/off) is evaluated as an additional manipulation in Supplementary Figure S3.

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFigures.docx