

Crowd-in-the-Loop Reinforcement Learning

Zhefan Ye

Computer Science and Engineering

University of Michigan

January 2026

Committee:

Professor Benjamin Kuipers (Chair)

Associate Professor Nikola Banovic

Professor Satinder Singh Baveja

Professor Vineet R. Kamat

The Core Challenges for RL

- Reinforcement learning (RL) requires two critical components to learn effectively:
 - Acquiring a reward function.
 - Understanding how environment dynamics work.

The Core Challenges for RL

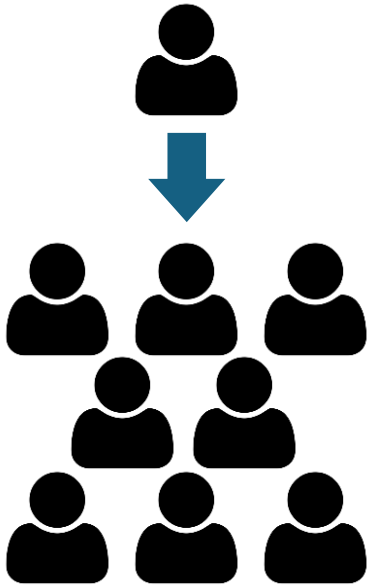
- Reinforcement learning (RL) requires two critical components to learn effectively:
 - Acquiring a reward function.
 - Understanding how environment dynamics work.
- Where does reward come from?
 - In synthetic environments (e.g., chess), reward functions are built-in.
 - In the real world, they ultimately come from *people*.

The Core Challenges for RL

- Reinforcement learning (RL) requires two critical components to learn effectively:
 - Acquiring a reward function.
 - Understanding how environment dynamics work.
- Where does reward come from?
 - In synthetic environments (e.g., chess), reward functions are built-in.
 - In the real world, they ultimately come from people.
- How is environment dynamics learned?
 - RL agents explore via trial and error (high sample complexity).
 - People possess prior knowledge of environment dynamics.

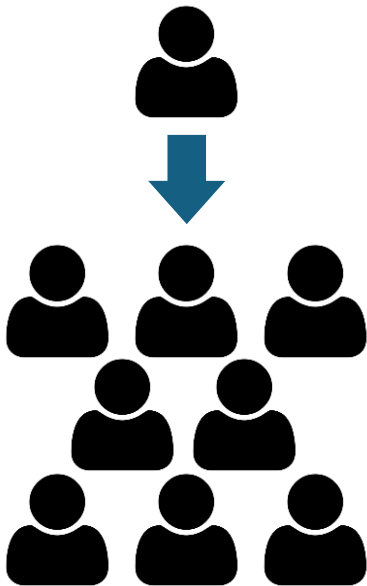
Why Crowds?

Scalability

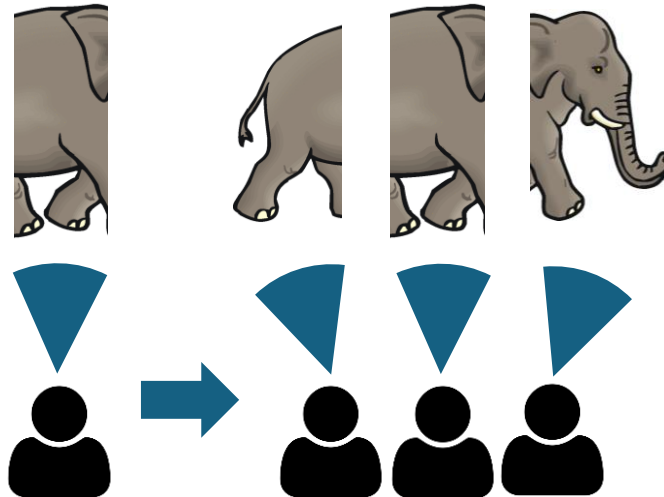


Why Crowds?

Scalability

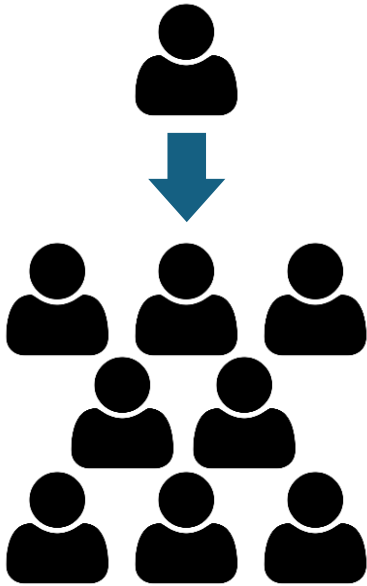


Diversity

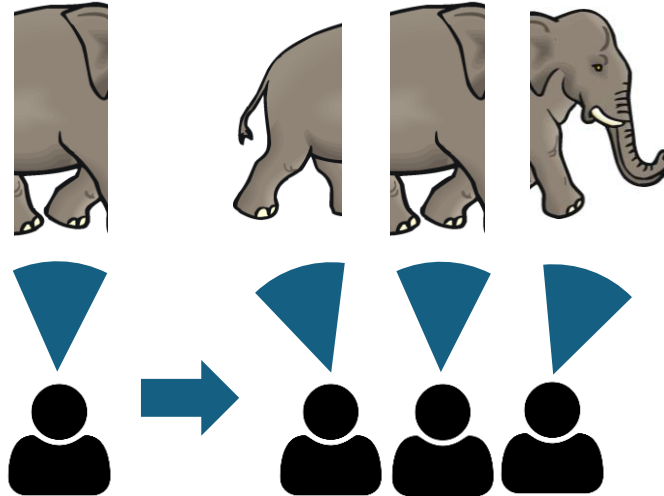


Why Crowds?

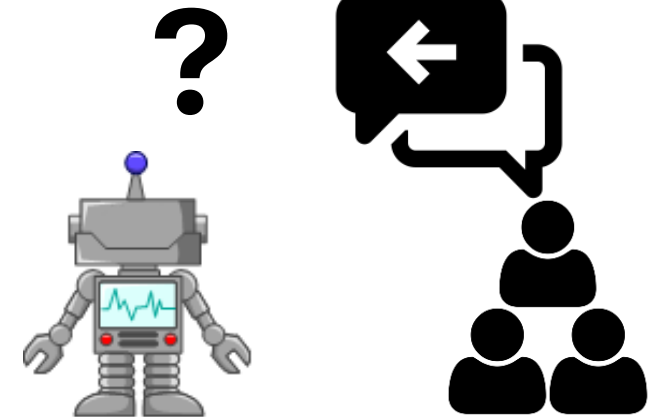
Scalability



Diversity

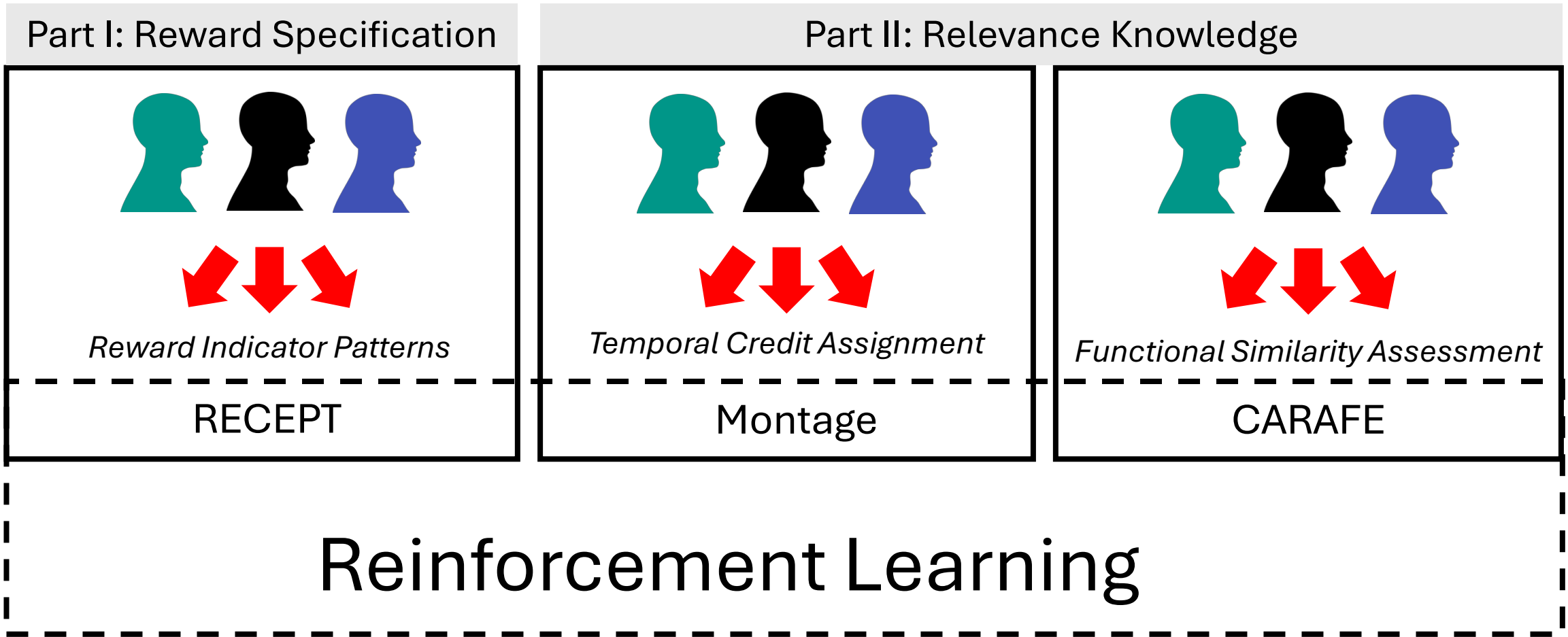


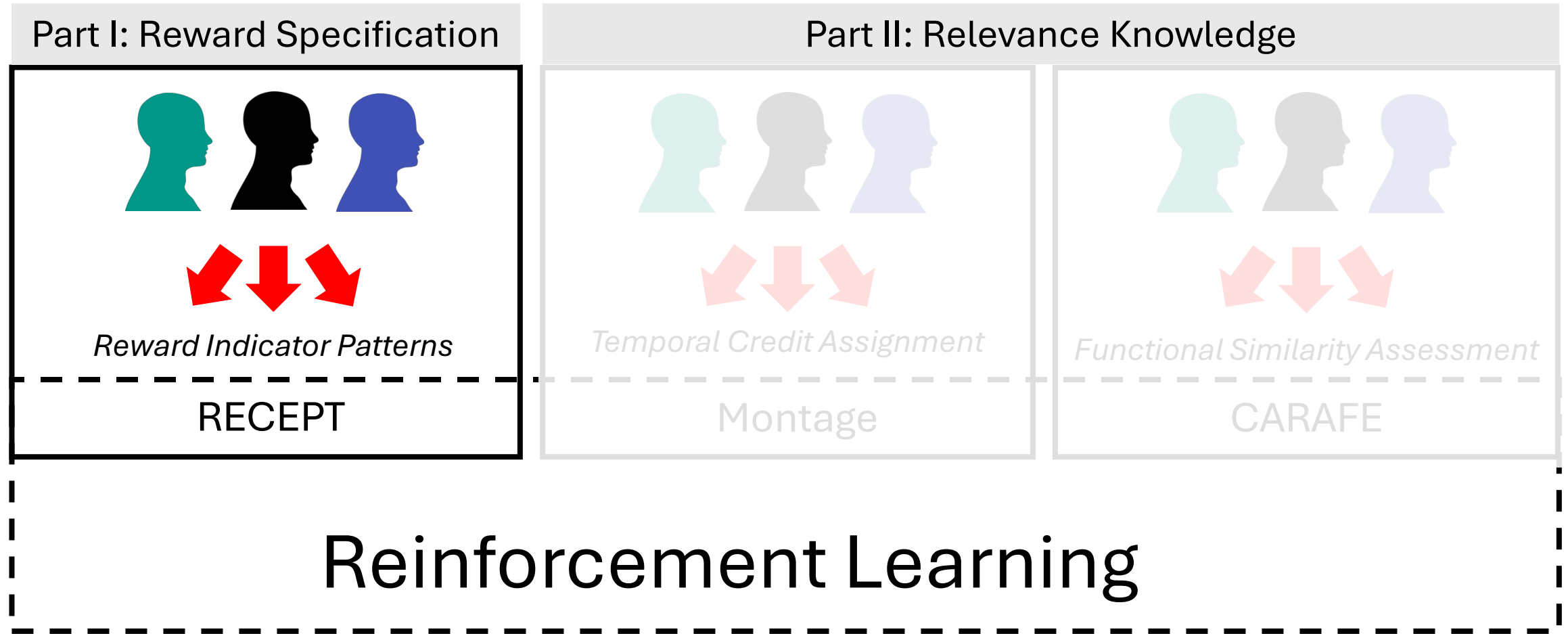
Flexibility



Thesis Statement

By leveraging collective intelligence regarding task objectives and environment dynamics, we can broaden the means to specify rewards and inject prior knowledge to improve reinforcement learning.





Motivation

- The real world has no scoreboard.



An autonomous vehicle passing a school bus with stop sign extended.

Motivation

- The real world has no scoreboard.
- Existing human-guided RL methods often ask for
 - Manual specification
 - Demonstration
 - Preference



An autonomous vehicle passing a school bus with stop sign extended.

Motivation

- The real world has no scoreboard.
- Existing human-guided RL methods often ask for
 - Manual specification
 - Demonstration
 - Preference
- The Missing Piece: **Patterns**



An autonomous vehicle passing a school bus with stop sign extended.

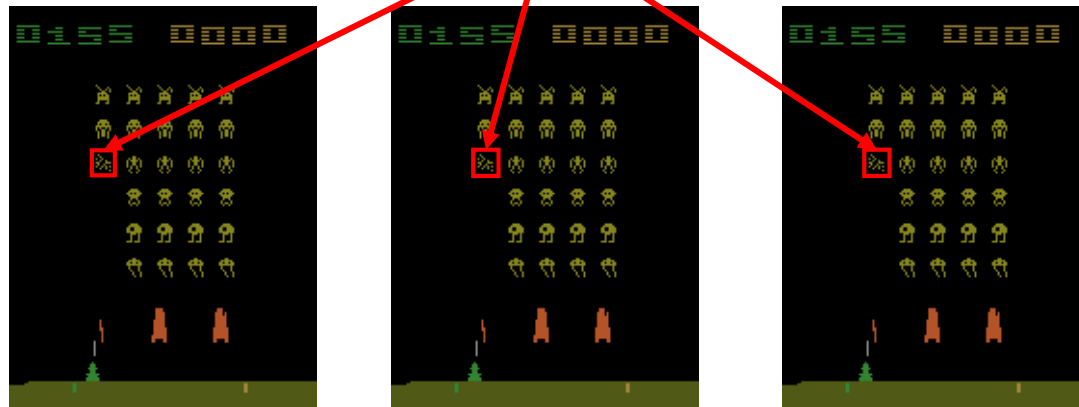
Patterns and Rewards

- People can identify patterns.



Patterns and Rewards

- People can identify patterns.
- Patterns can be scaled.



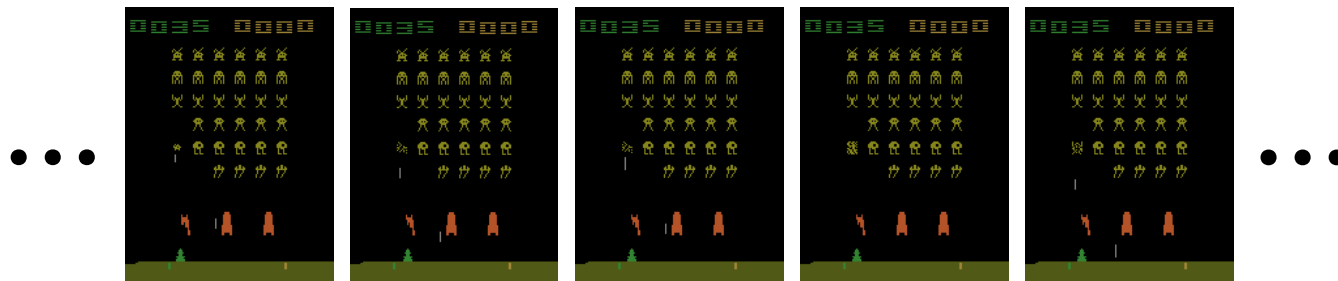
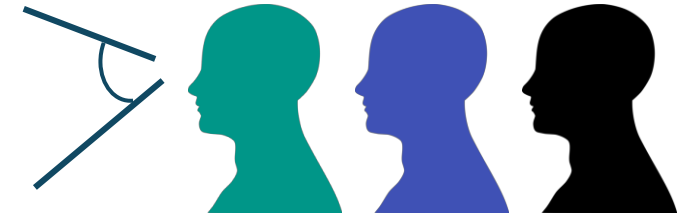
6/6/2026



15

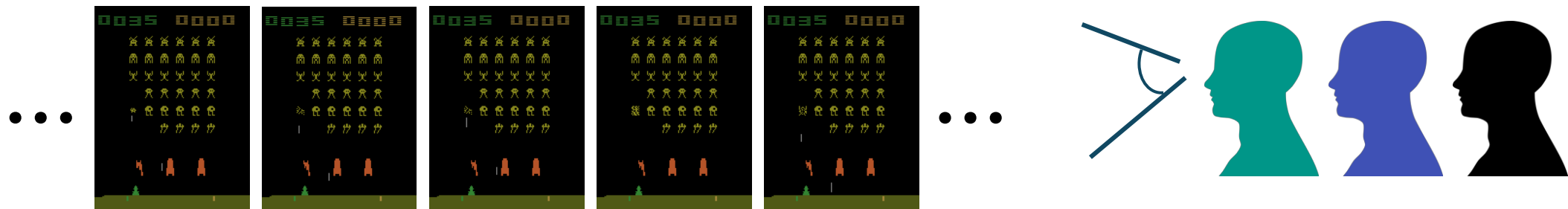
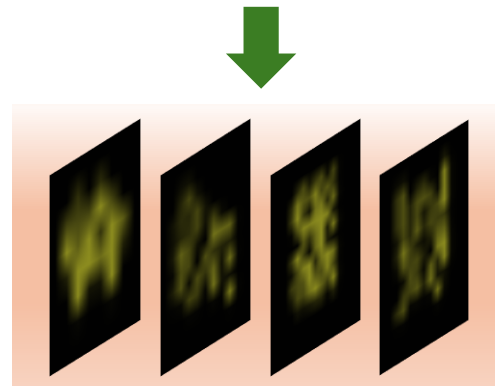
PRFG: Construct a Reward Function

State Samples



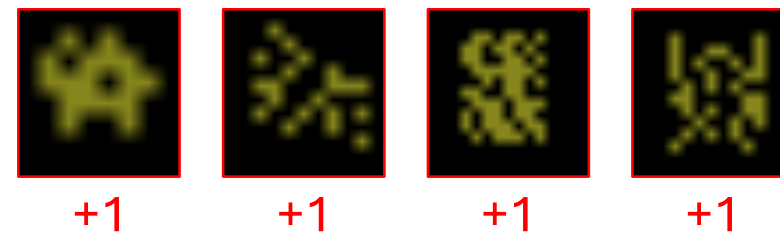
PRFG: Construct a Reward Function

State Samples

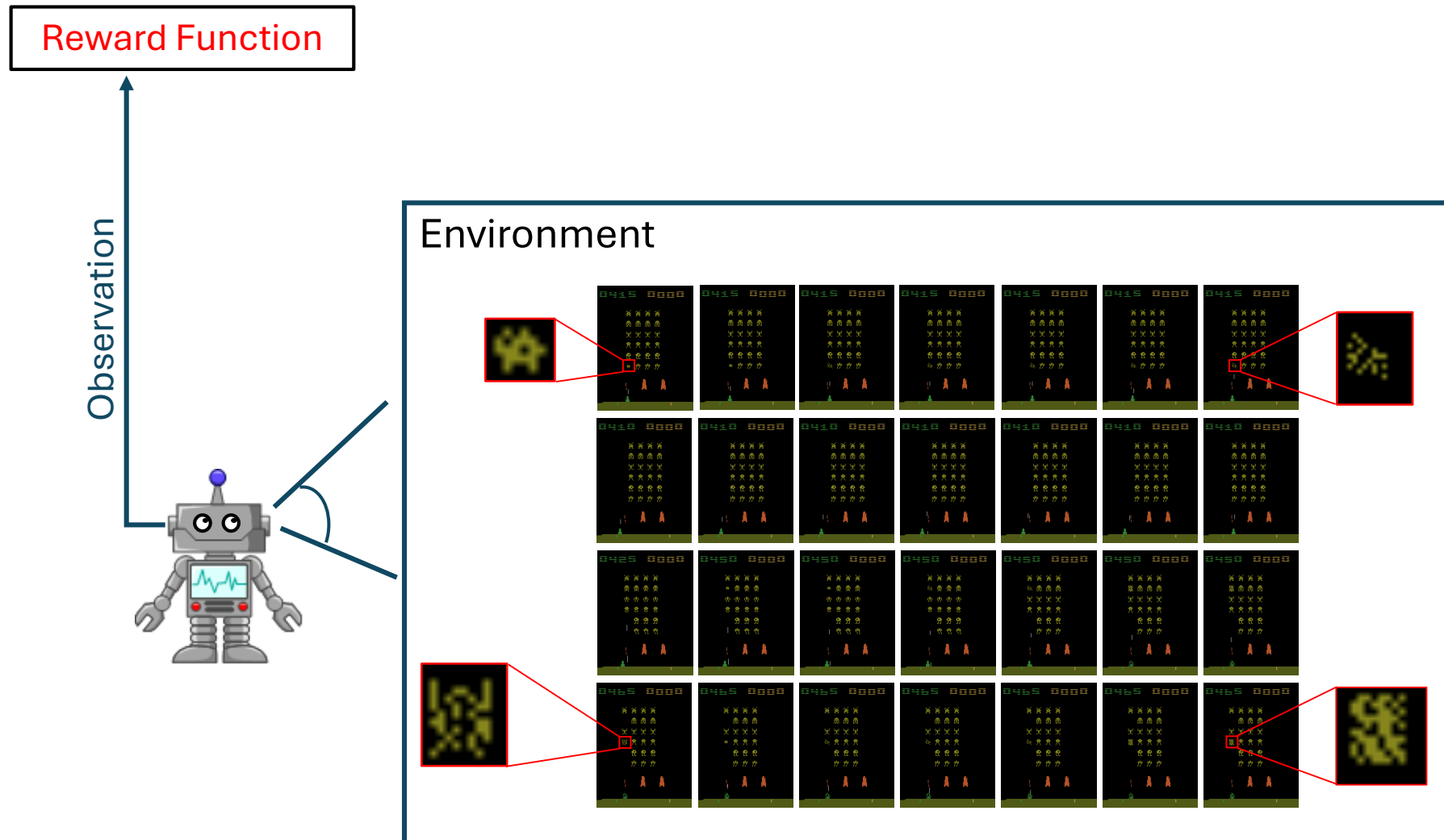


Patterns

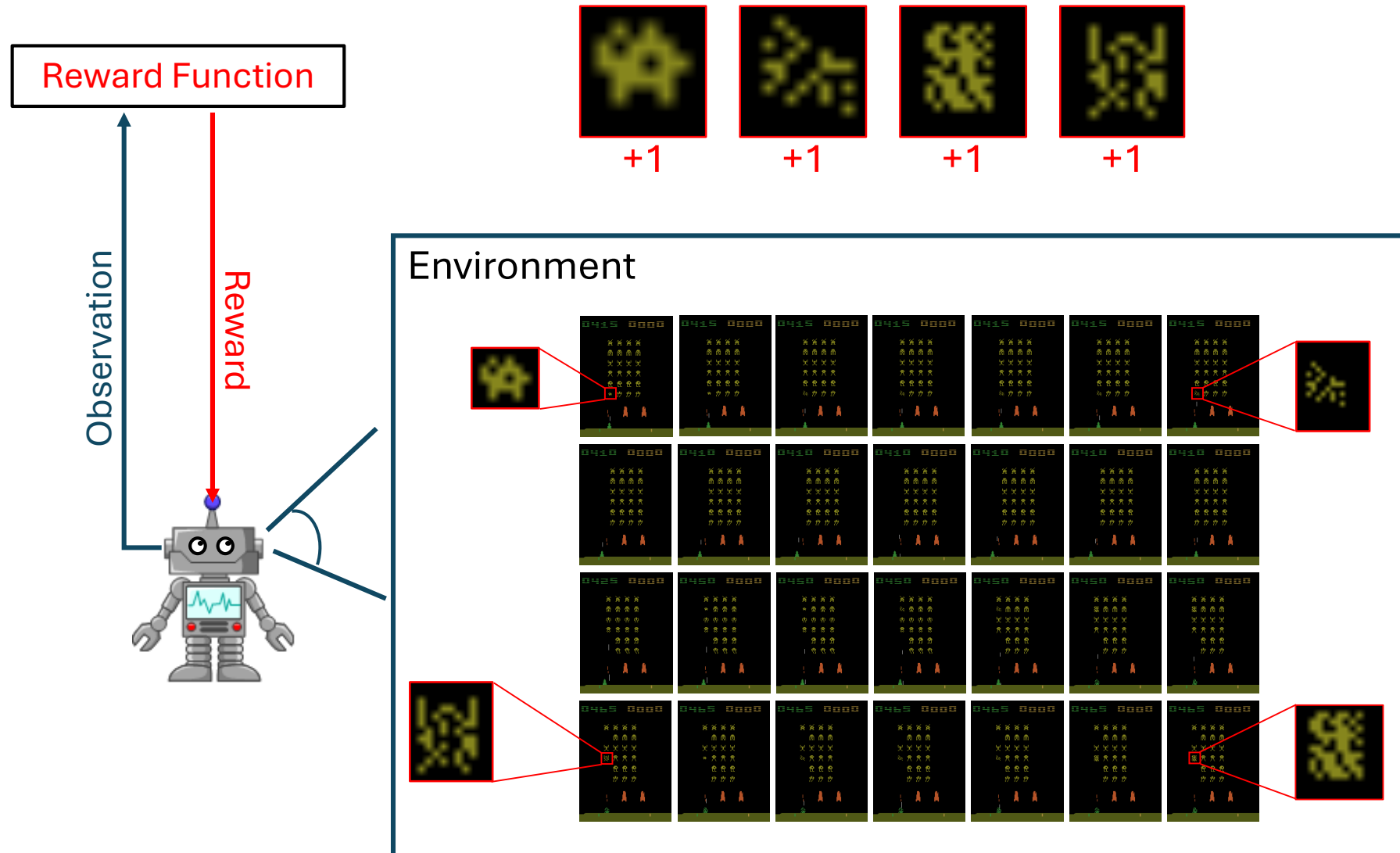
Reward Function



PRFG: Deploy a Reward Function

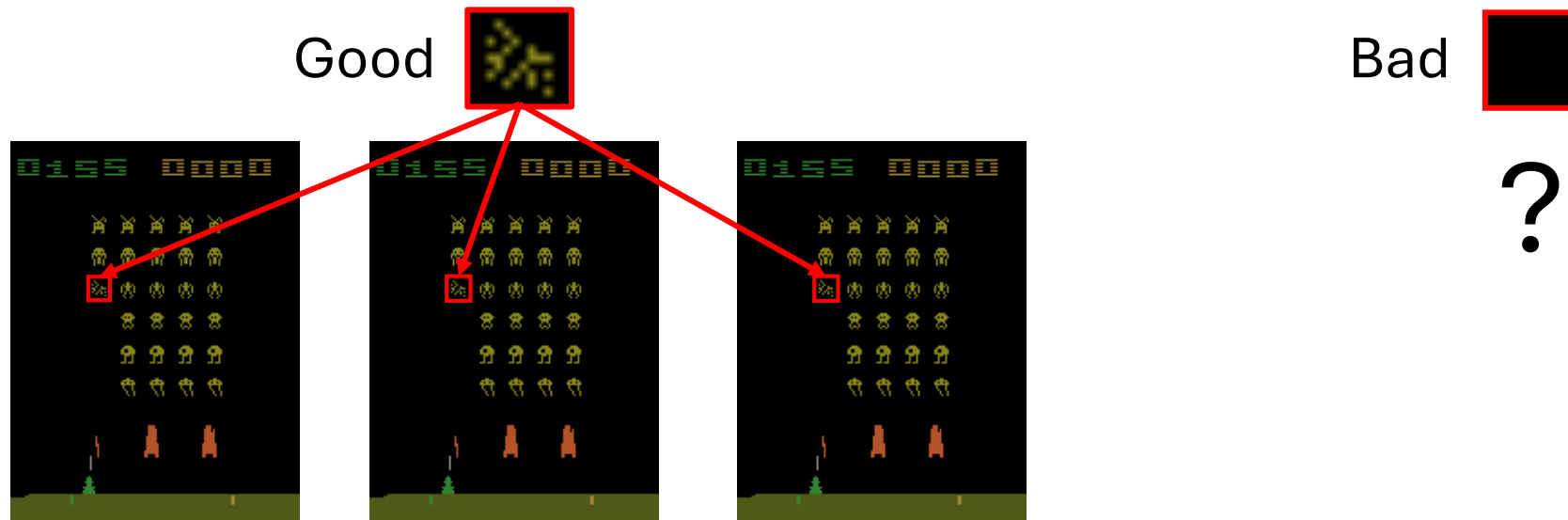


PRFG: Deploy a Reward Function



Potential Errors of PRFG

- PRFG can scale human feedback.
- But crowd workers may bring erroneous patterns.



Failure Modes

- Missing Reward
- False Reward
- Duplicate Reward
- Delayed Reward



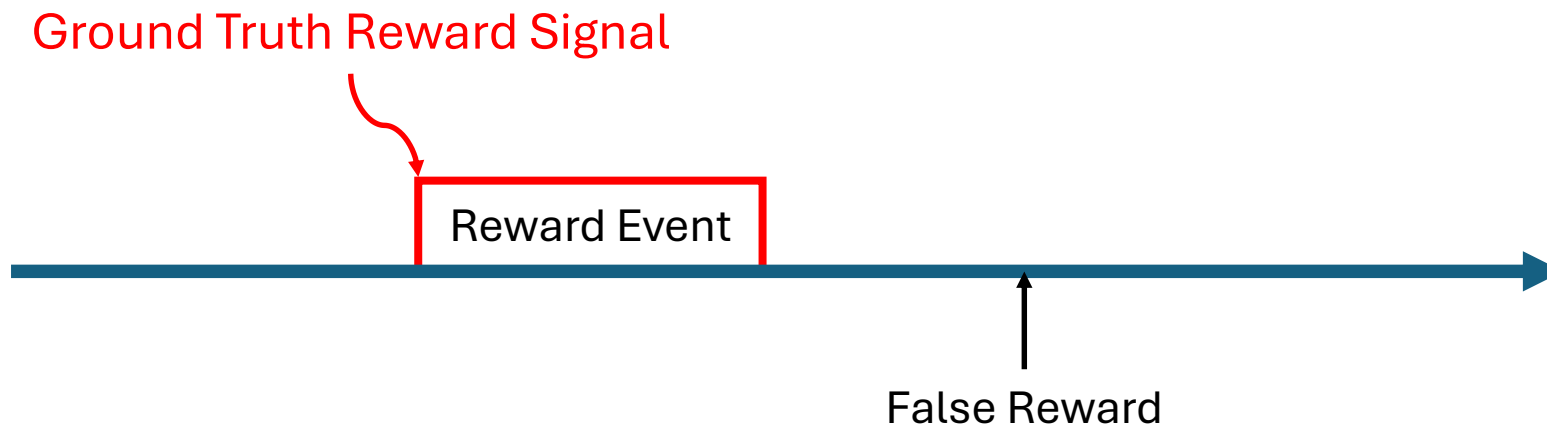
Missing Reward

- Missing Reward
- False Reward
- Duplicate Reward
- Delayed Reward



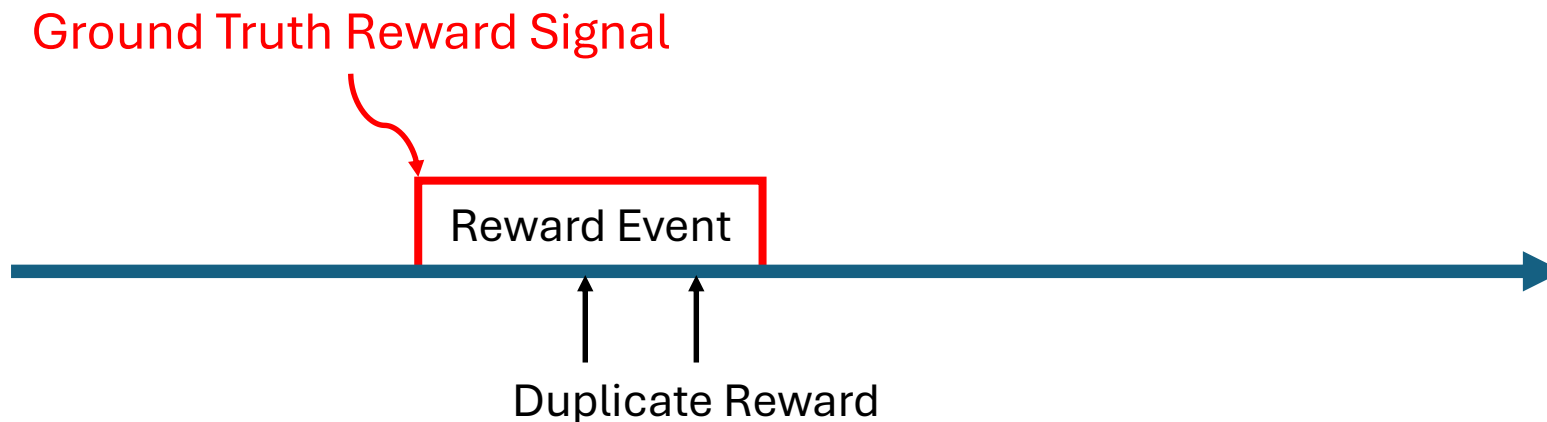
False Reward

- Missing Reward
- **False Reward**
- Duplicate Reward
- Delayed Reward



Duplicate Reward

- Missing Reward
- False Reward
- Duplicate Reward
- Delayed Reward



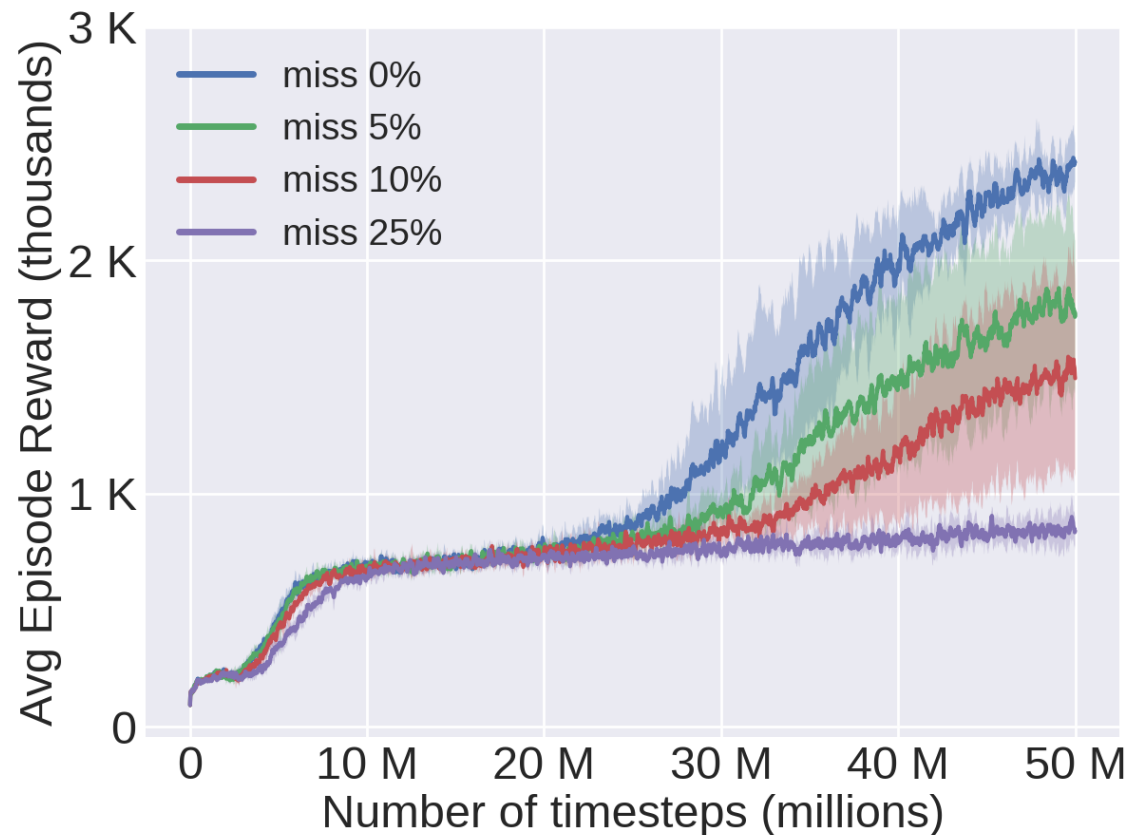
Delayed Reward

- Missing Reward
- False Reward
- Duplicate Reward
- Delayed Reward

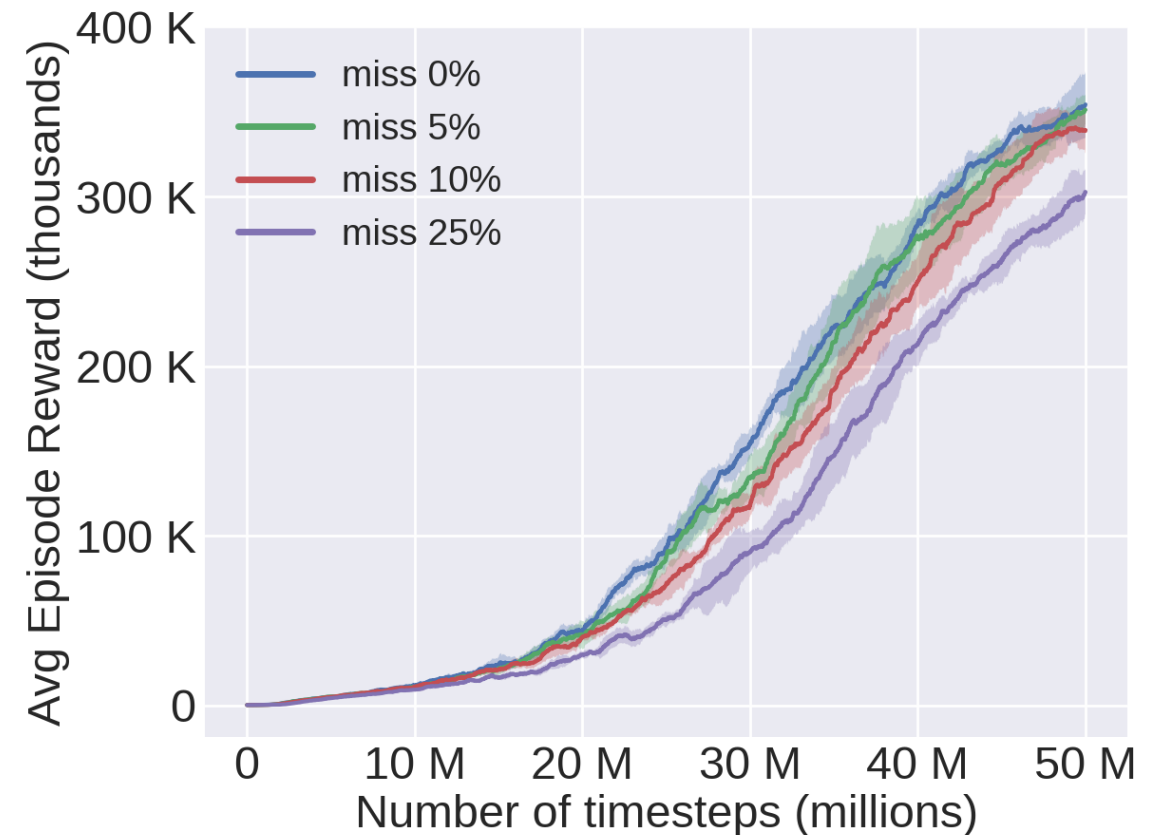


Missing Reward

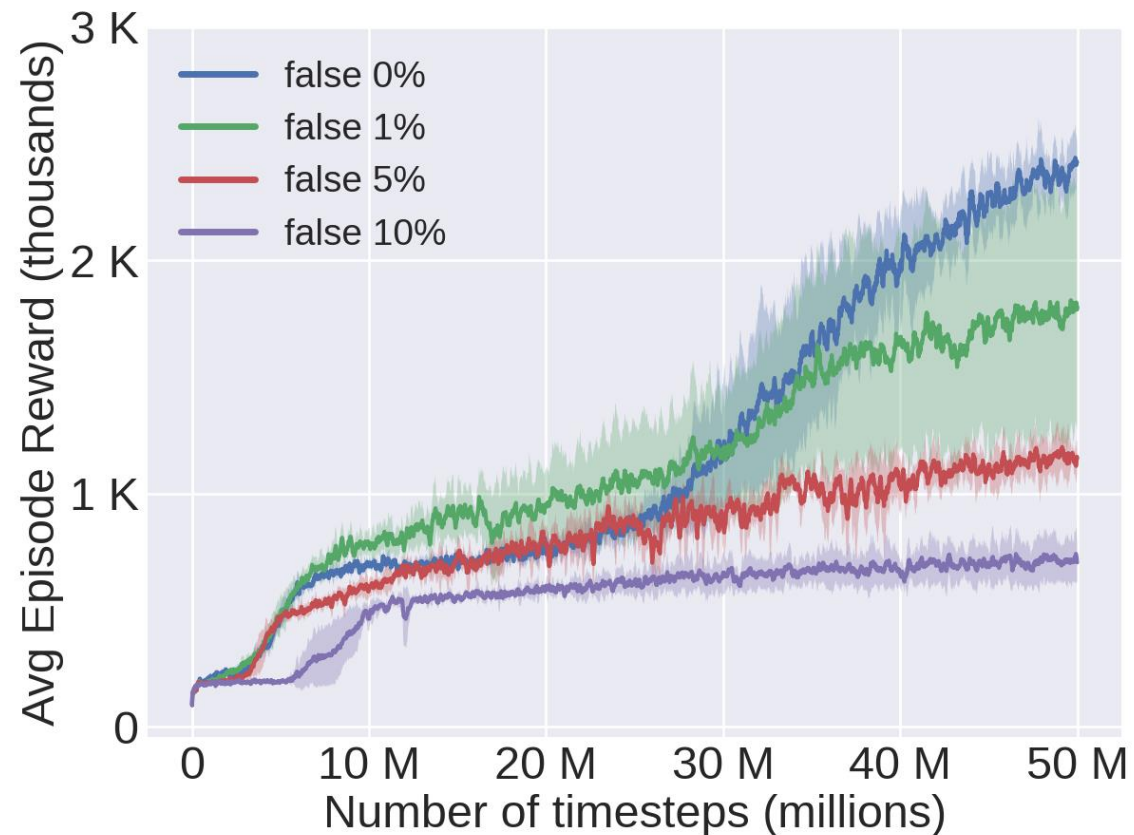
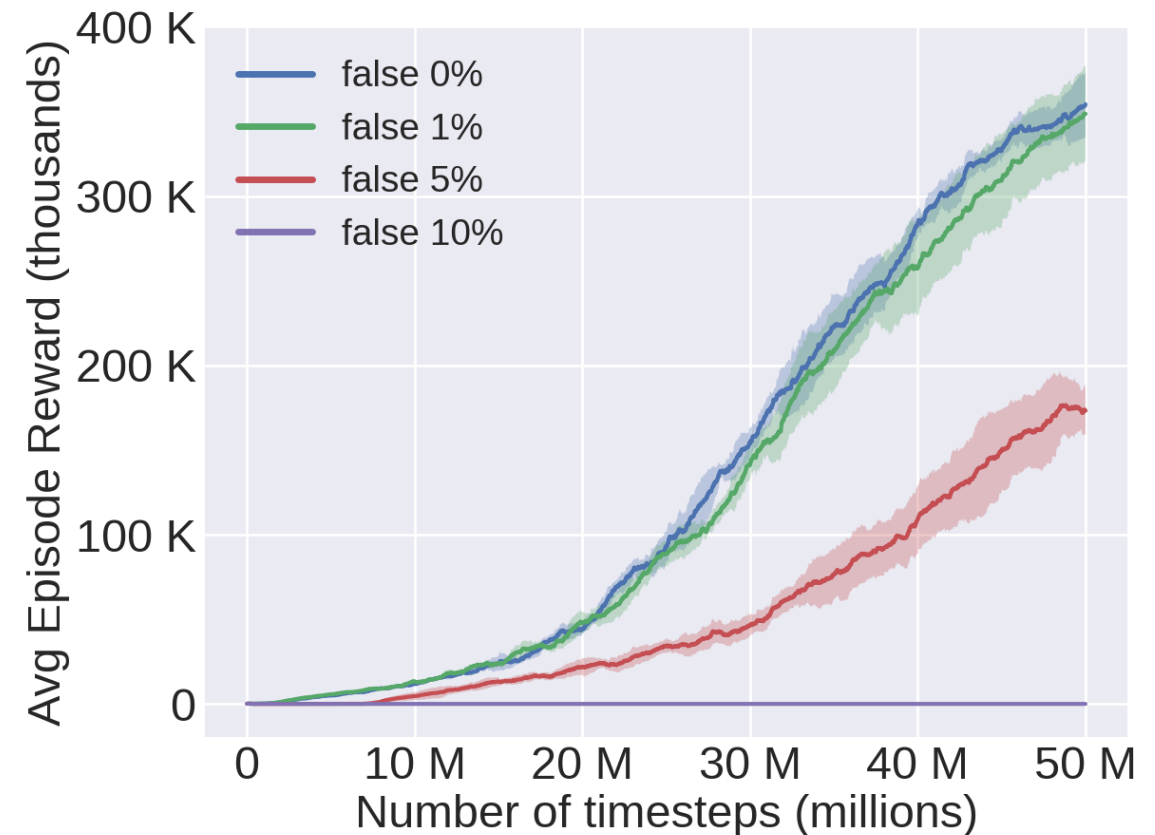
Space Invaders



Demon Attack

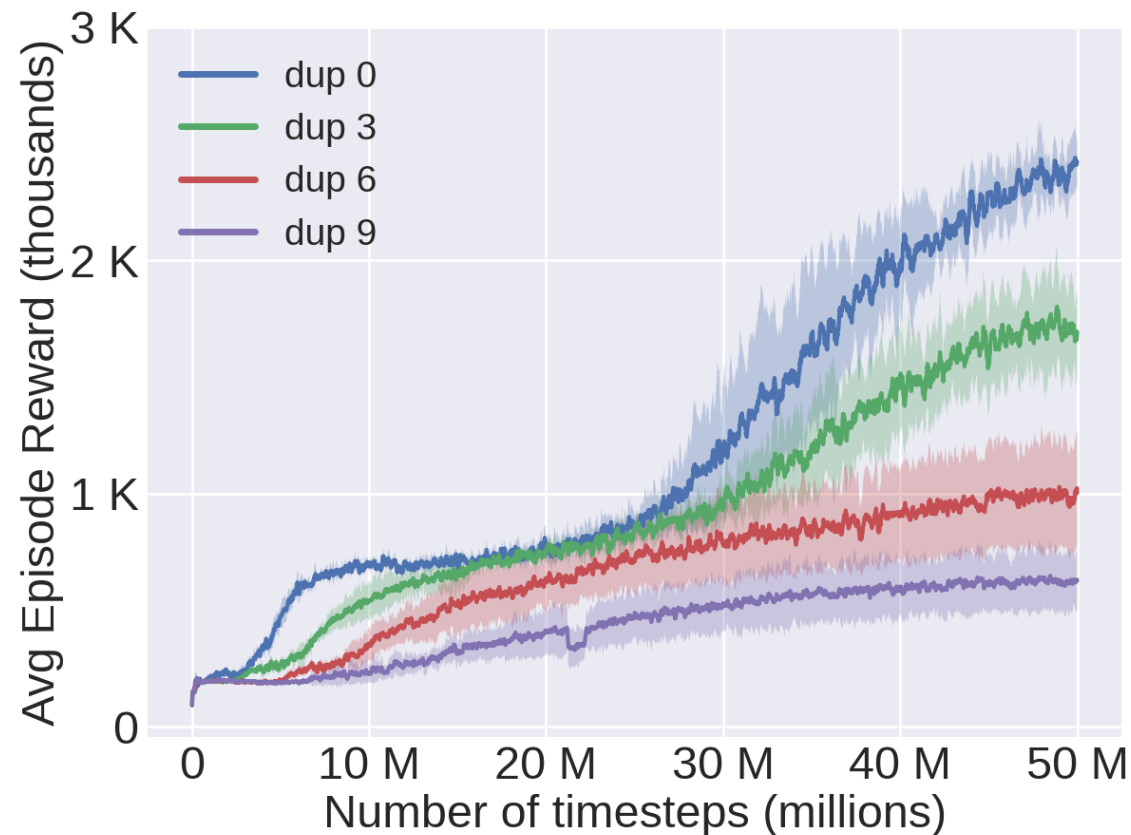


False Reward

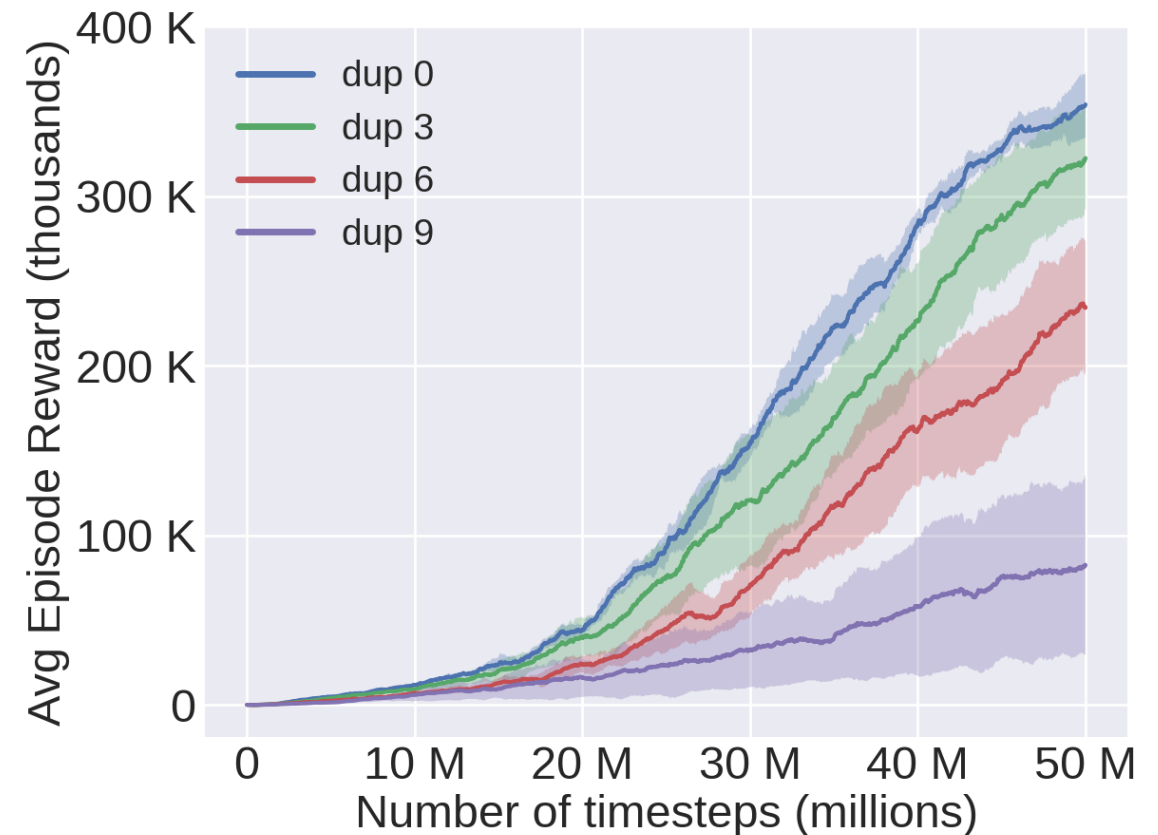
Space Invaders**Demon Attack**

Duplicate Reward

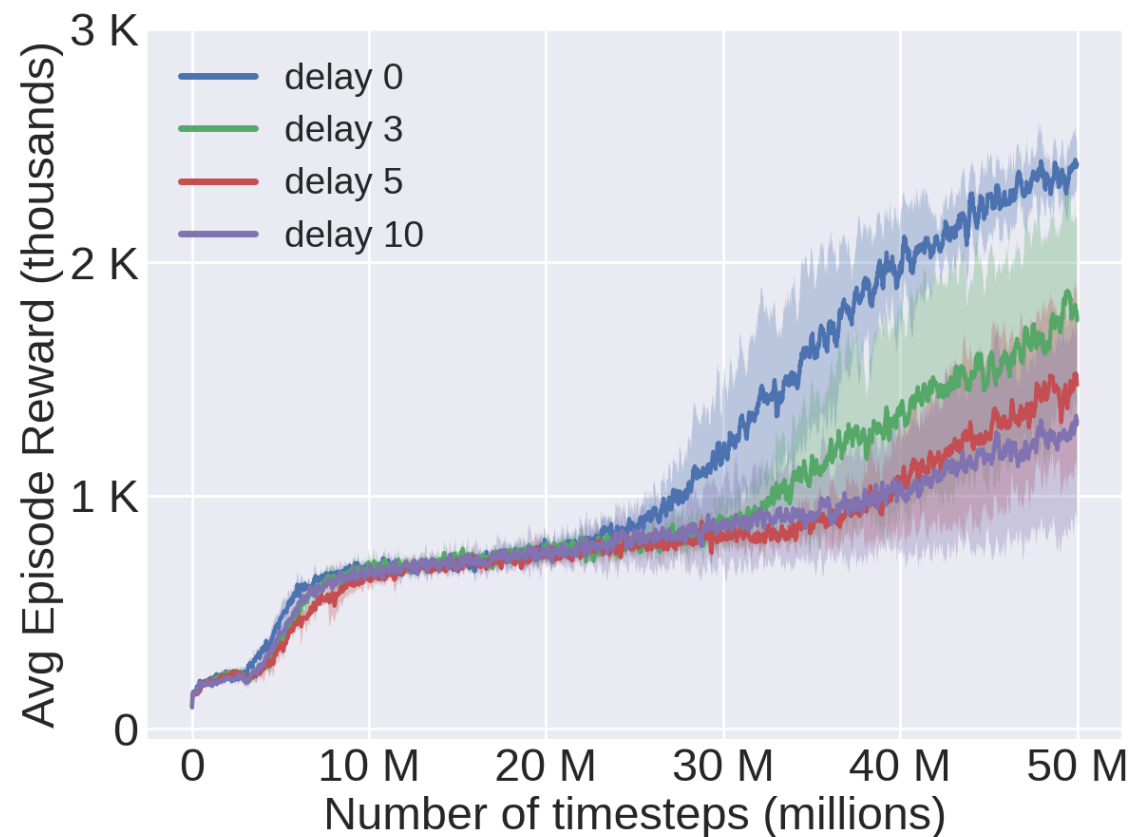
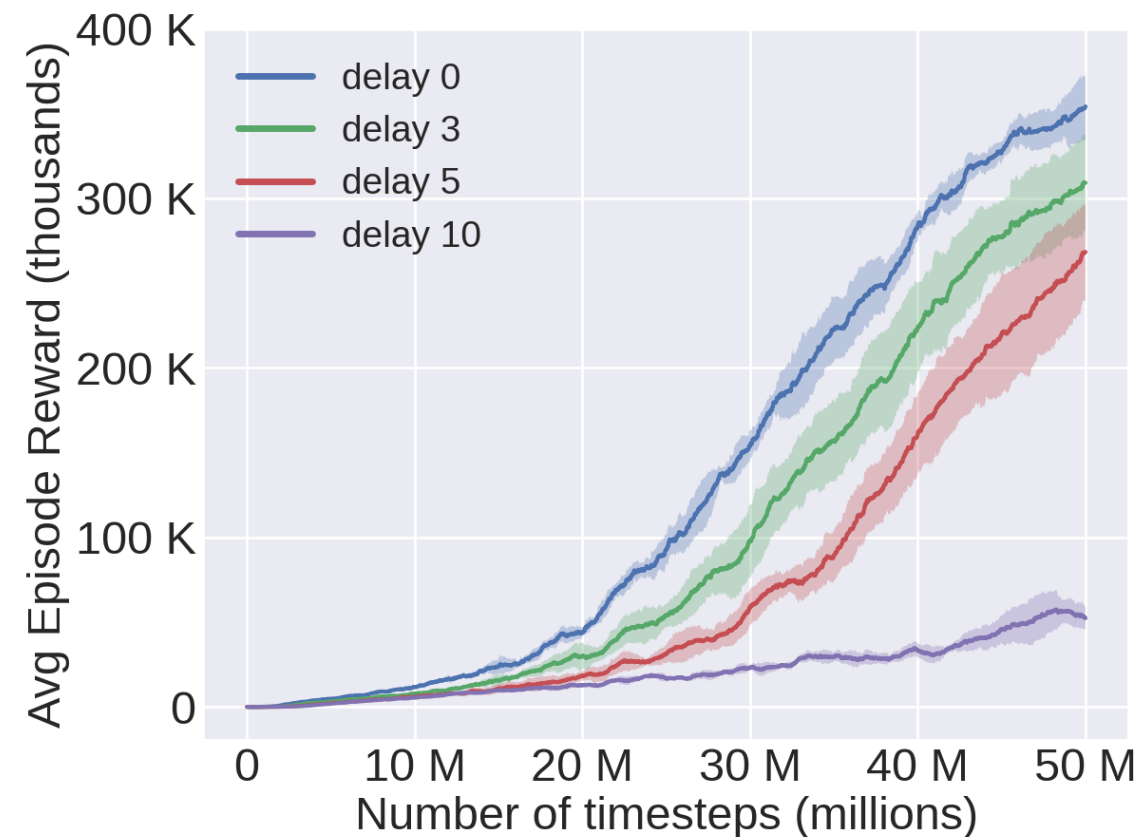
Space Invaders



Demon Attack



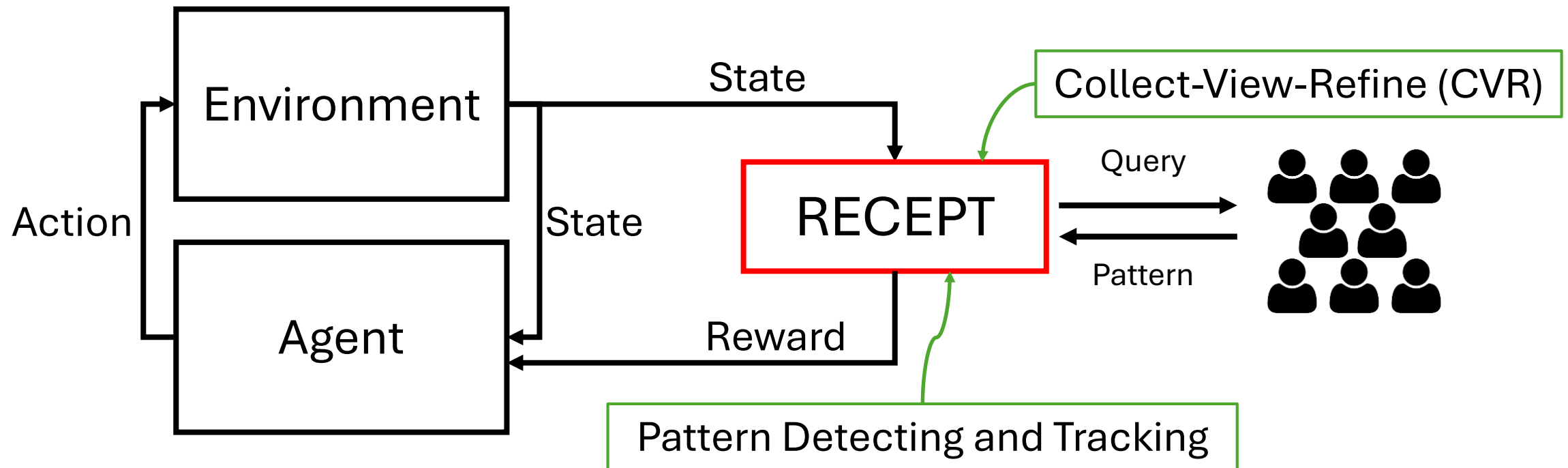
Delayed Reward

Space Invaders**Demon Attack**

Takeaway

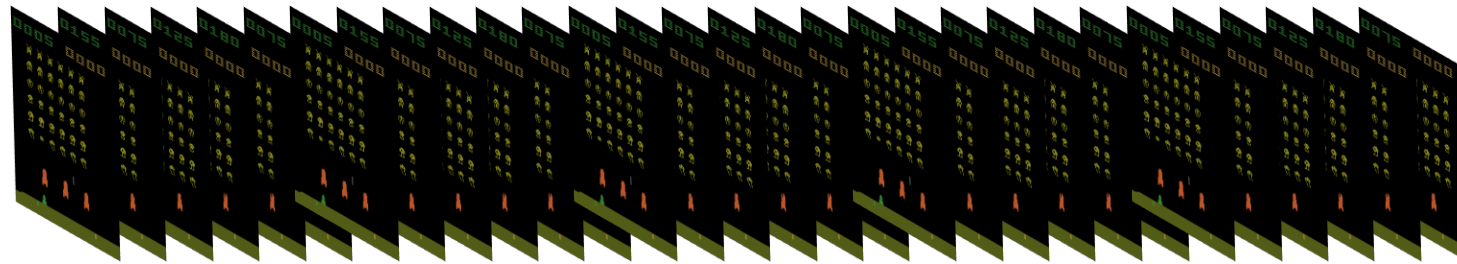
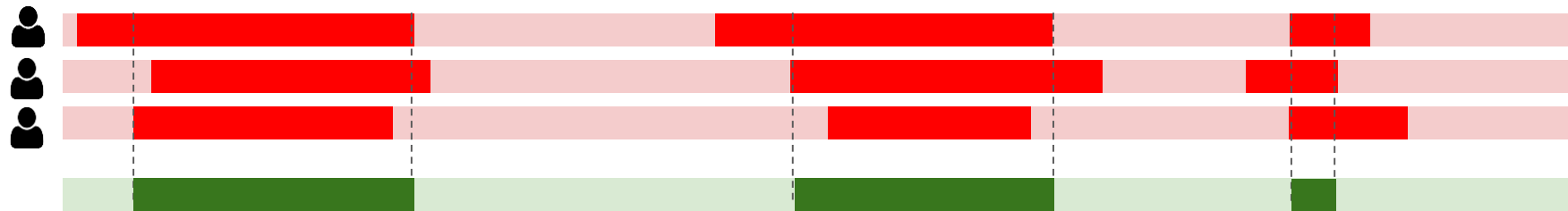
- Missing Reward (Reduce)
- **False Reward** (Minimize)
- **Duplicate Reward** (Minimize)
- Delayed Reward (Reduce)

REwards from Crowd-Elicited PaTterns (RECEPT)



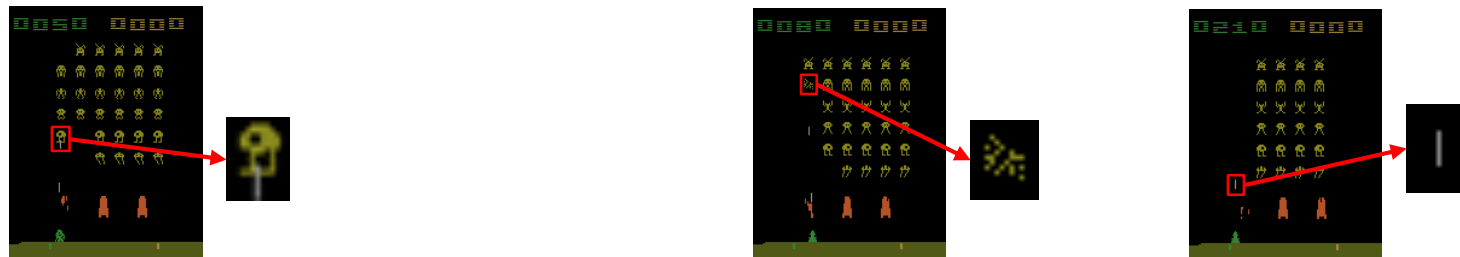
CVR Workflow --- Collect

Input:
Gameplay Images

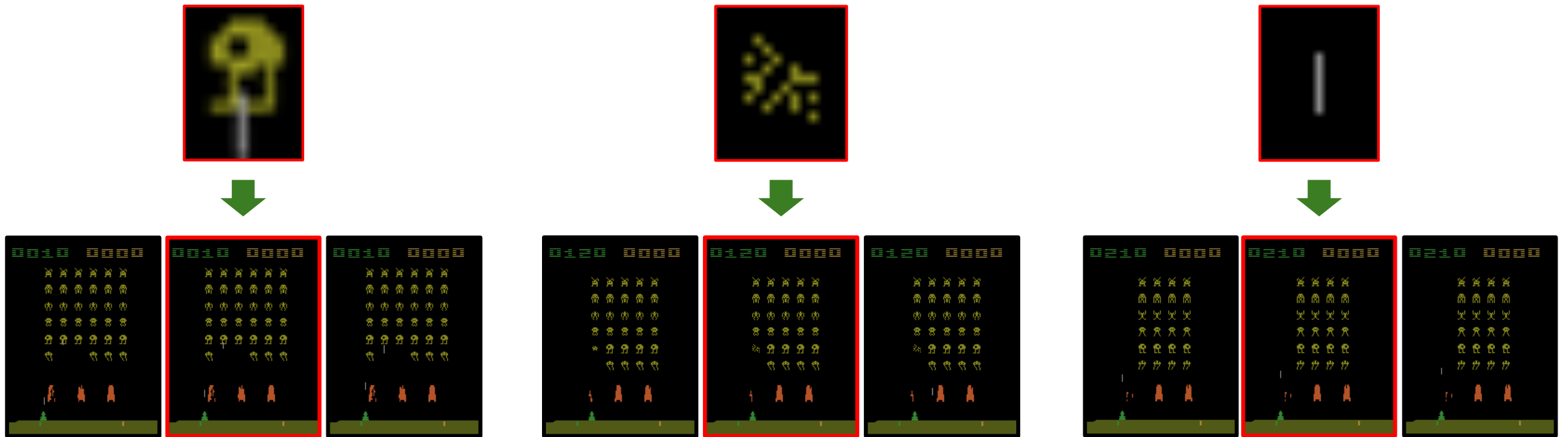


Aggregate

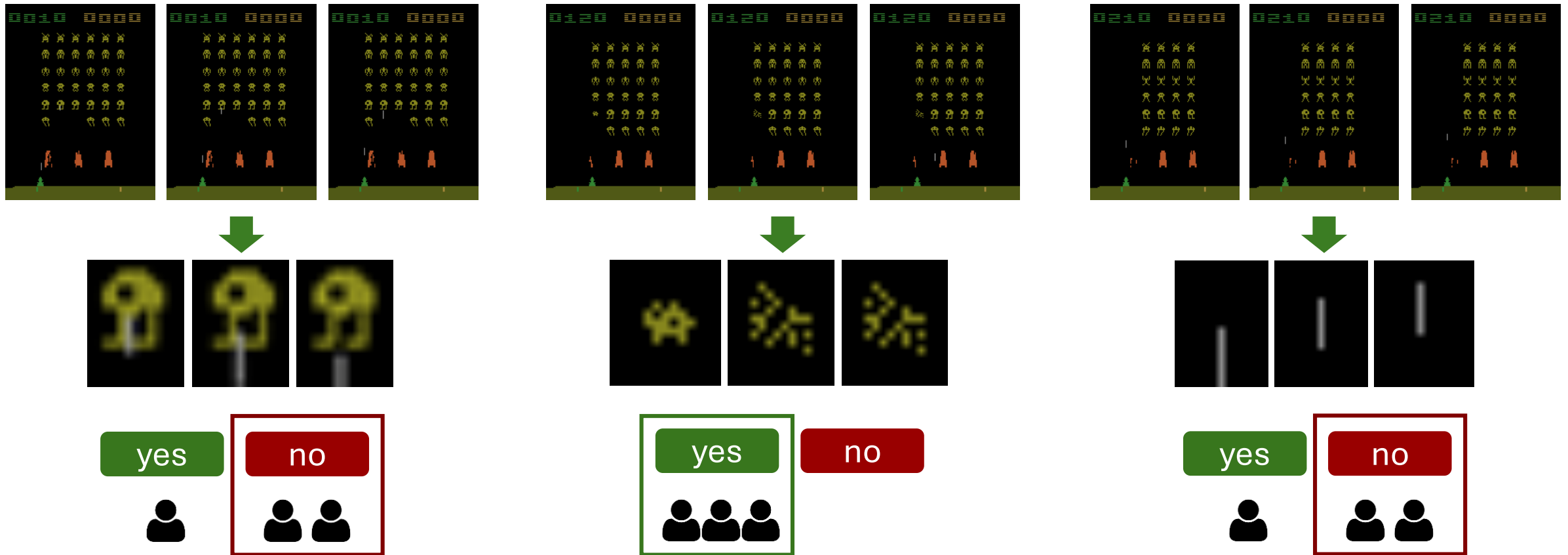
Crop



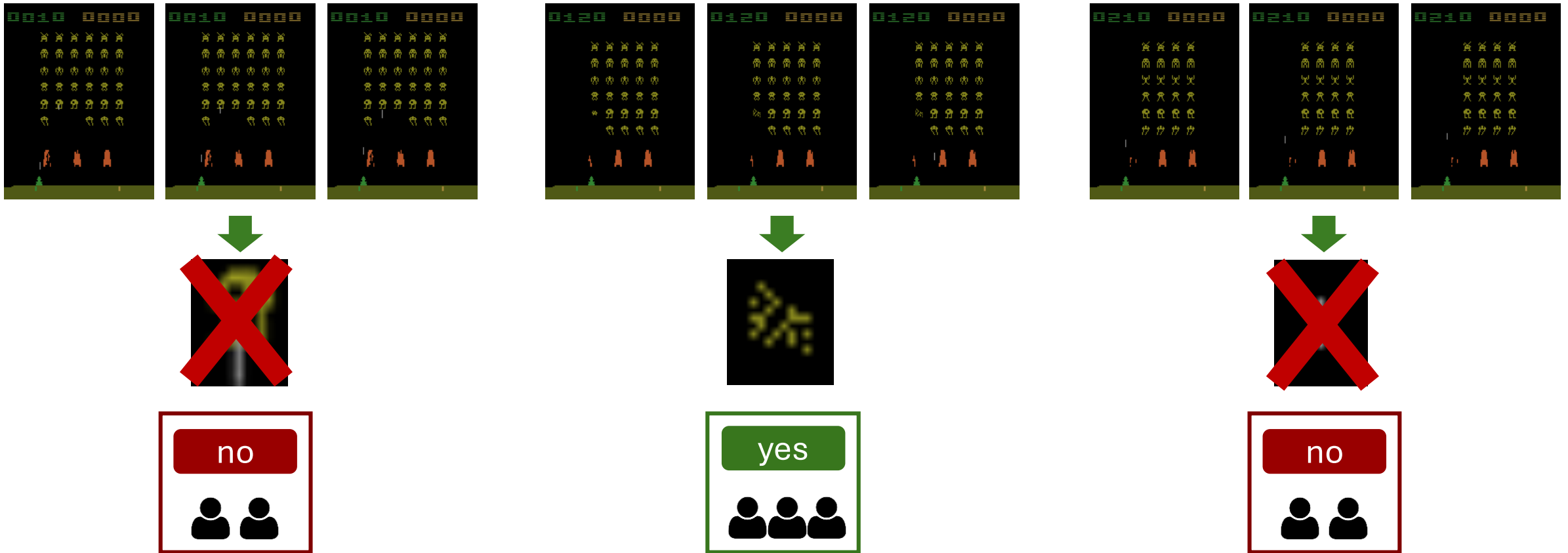
CVR Workflow --- View



CVR Workflow --- Refine

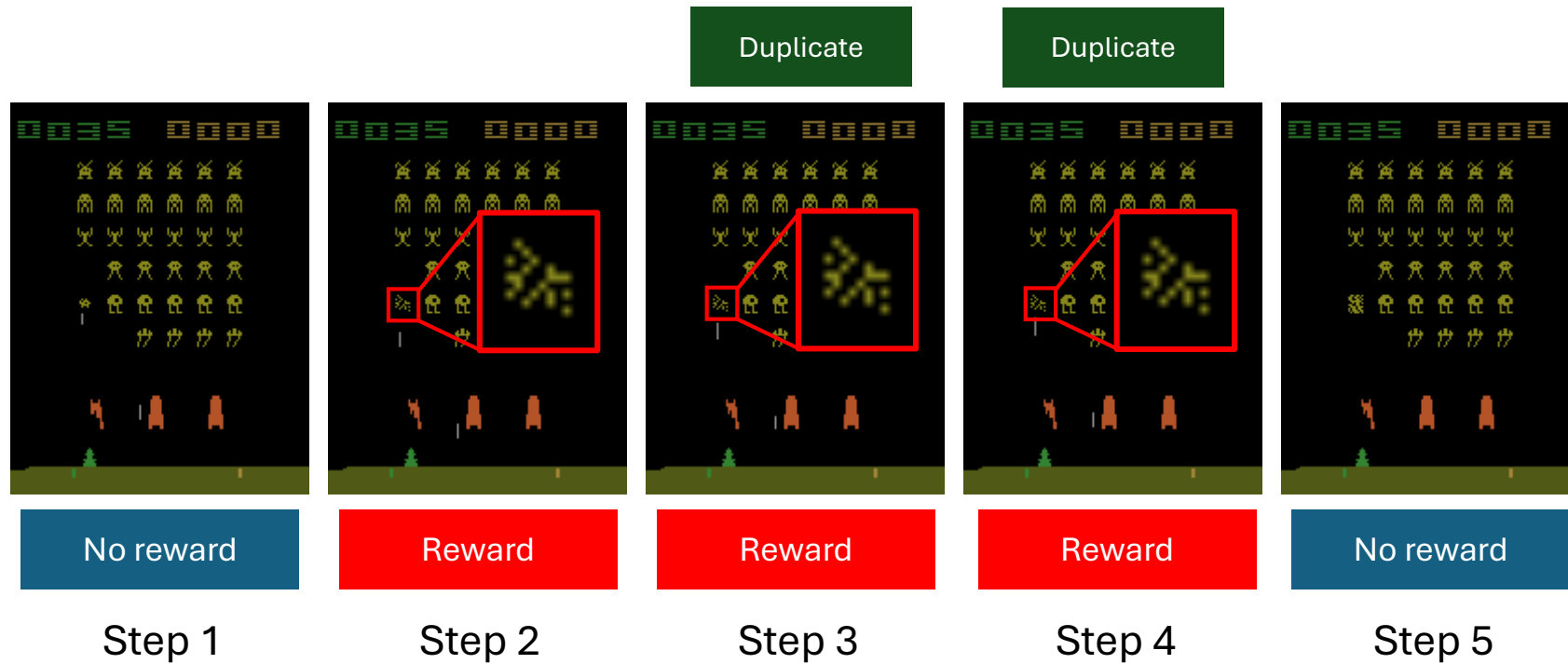


CVR Workflow --- Refine

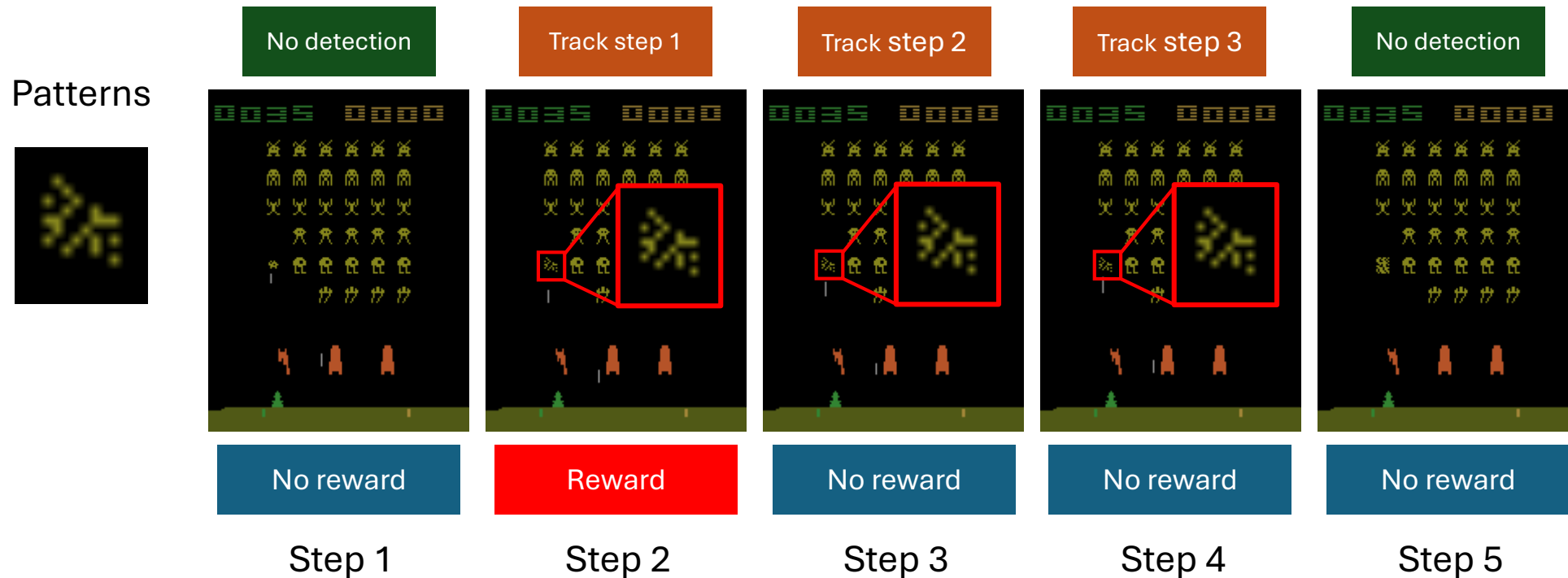


Pattern Detection

Patterns

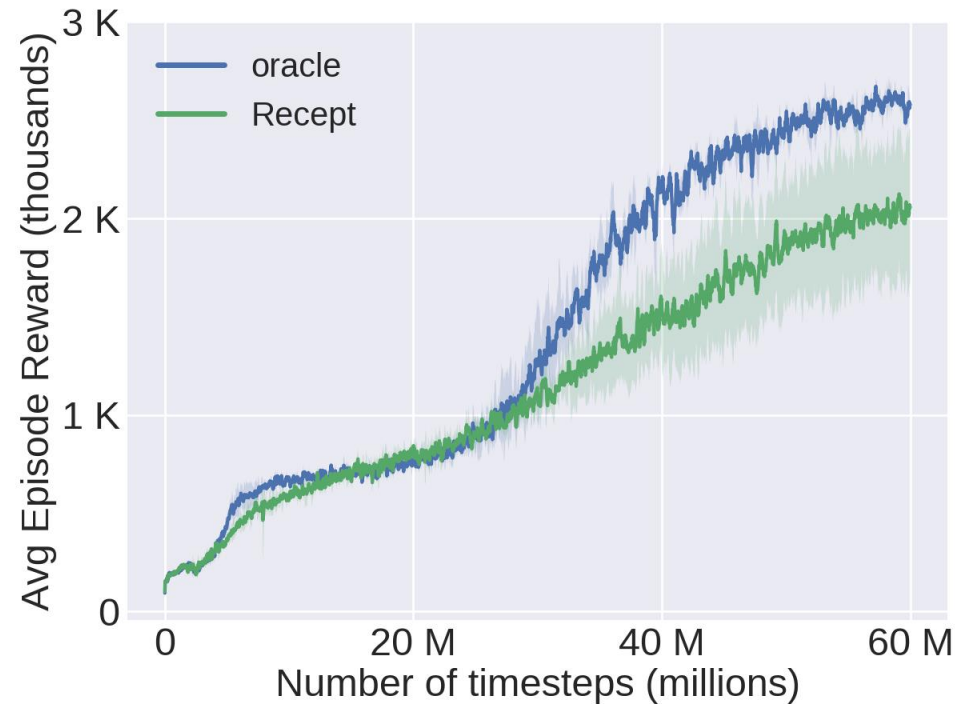


Pattern Detection and Tracking



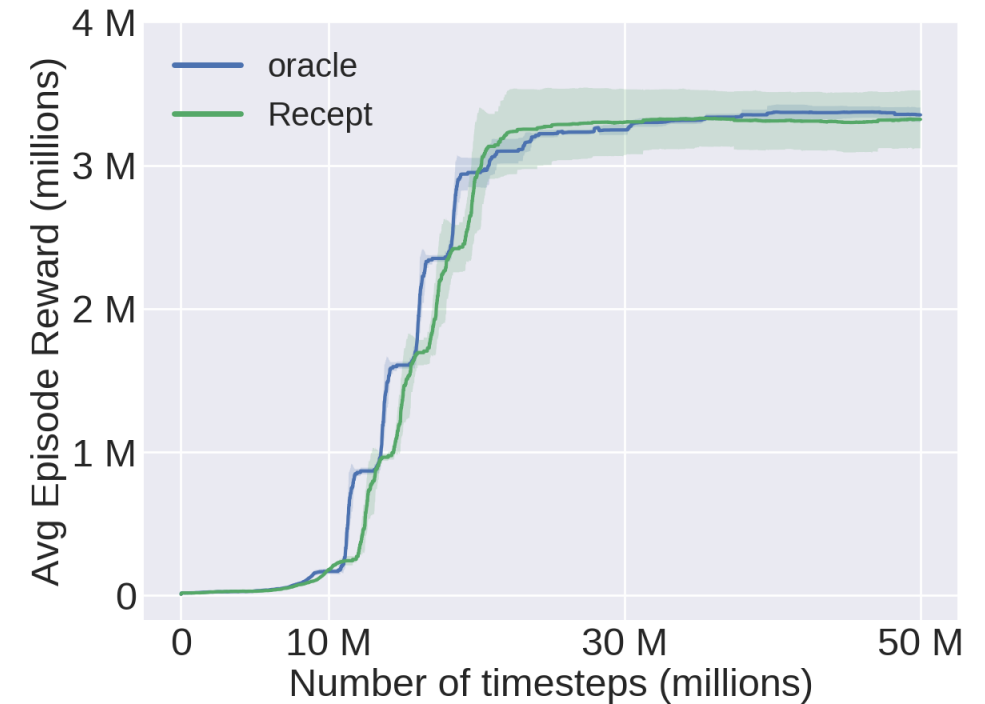
Comparison with Ground Truth Reward

Space Invaders



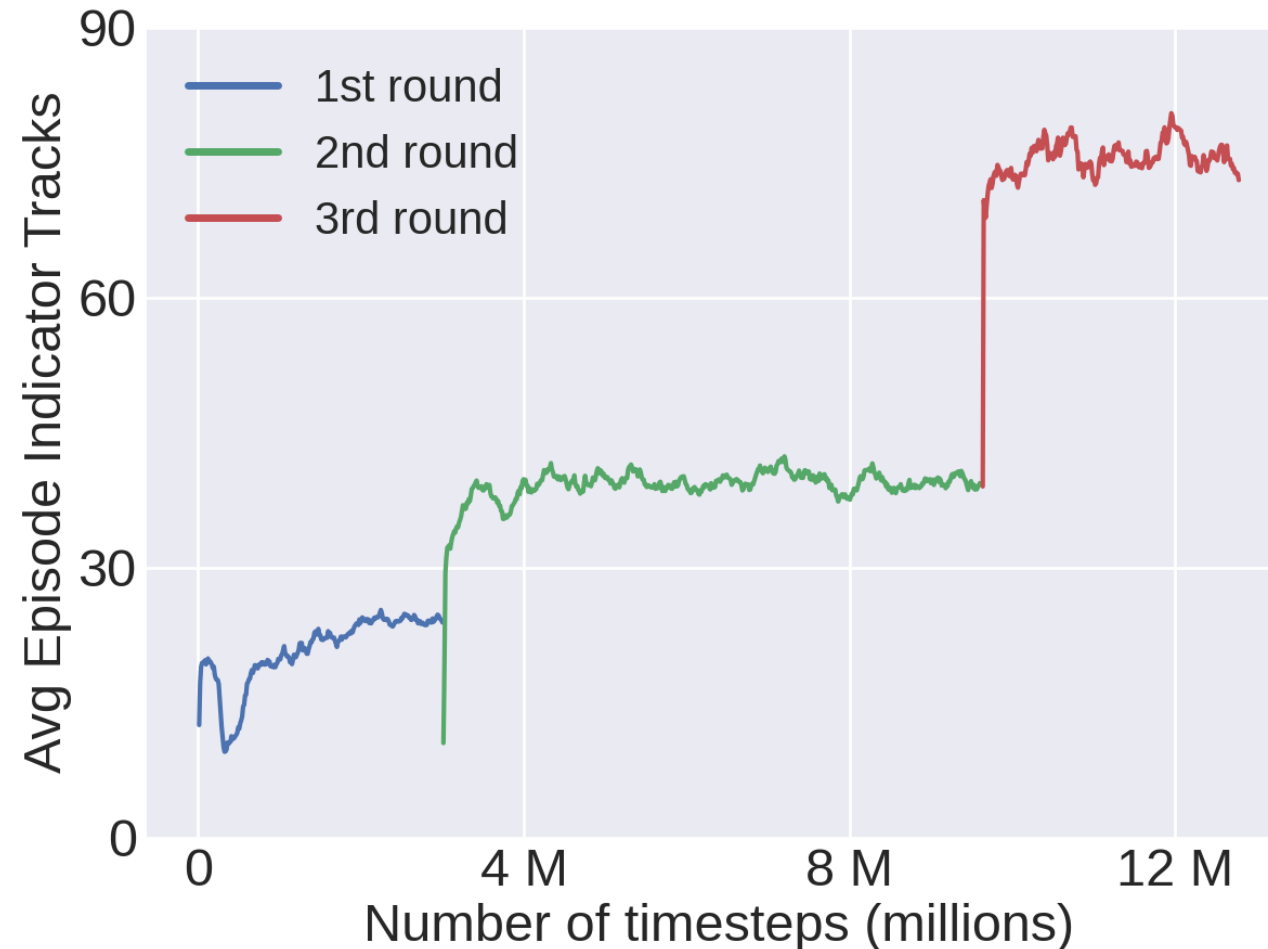
Space Invaders: 80% of oracle AULC

Atlantis

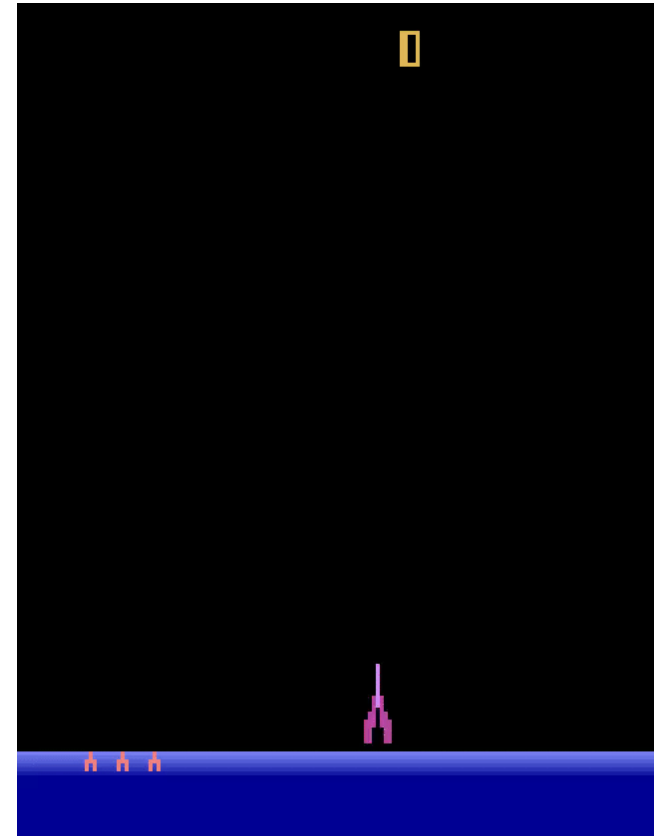


Atlantis: 98% of oracle AULC

Progressive RECEPT in Demon Attack



RECEPT-Trained RL Agents



Takeaway

- Pattern-based reward function generation (PRFG) method generates rewards via pattern detection and tracking.

Takeaway

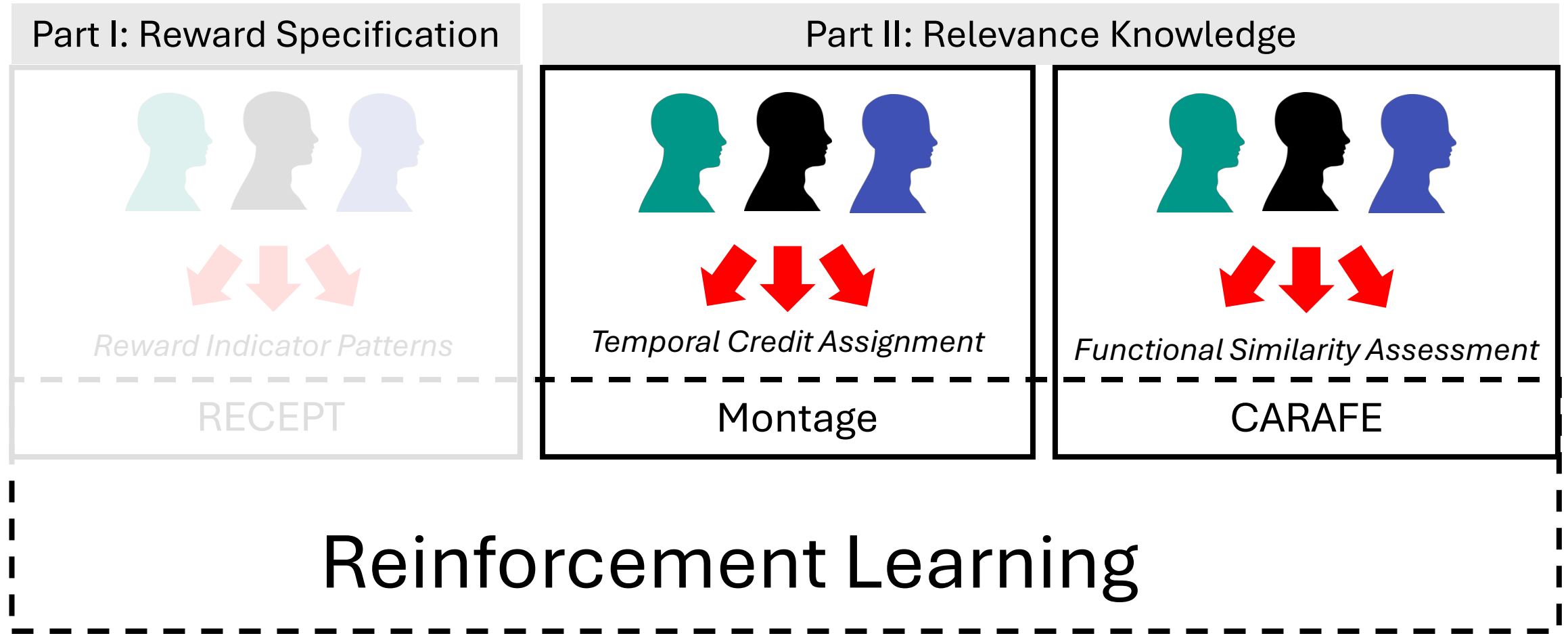
- Pattern-based reward function generation (PRFG) method generates rewards via pattern detection and tracking.
- Collect-View-Refine (CVR) workflow elicits accurate patterns from crowd workers.

Takeaway

- Pattern-based reward function generation (PRFG) method generates rewards via pattern detection and tracking.
- Collect-View-Refine (CVR) workflow elicits accurate patterns from crowd workers.
- RECEPT can produce scalable reward signals and train RL agents to achieve comparable performance to the oracle agents.

Takeaway

- Pattern-based reward function generation (PRFG) method generates rewards via pattern detection and tracking.
- Collect-View-Refine (CVR) workflow elicits accurate patterns from crowd workers.
- RECEPT can produce scalable reward signals and train RL agents to achieve comparable performance to the oracle agents.
- Modern foundation models, e.g., VLMs, can provide more advanced pattern-matching capabilities.
 - E.g., “Penalize the autonomous vehicle if it passes a school bus with its stop sign extended.”



From Rewards to Relevance

- Rewards define the **goal**, but not the **path**. This forces agents into inefficient, exhaustive exploration.
- **Relevance** — a concept that leverages human intuition to identify states that can significantly influence outcomes.

From Rewards to Relevance

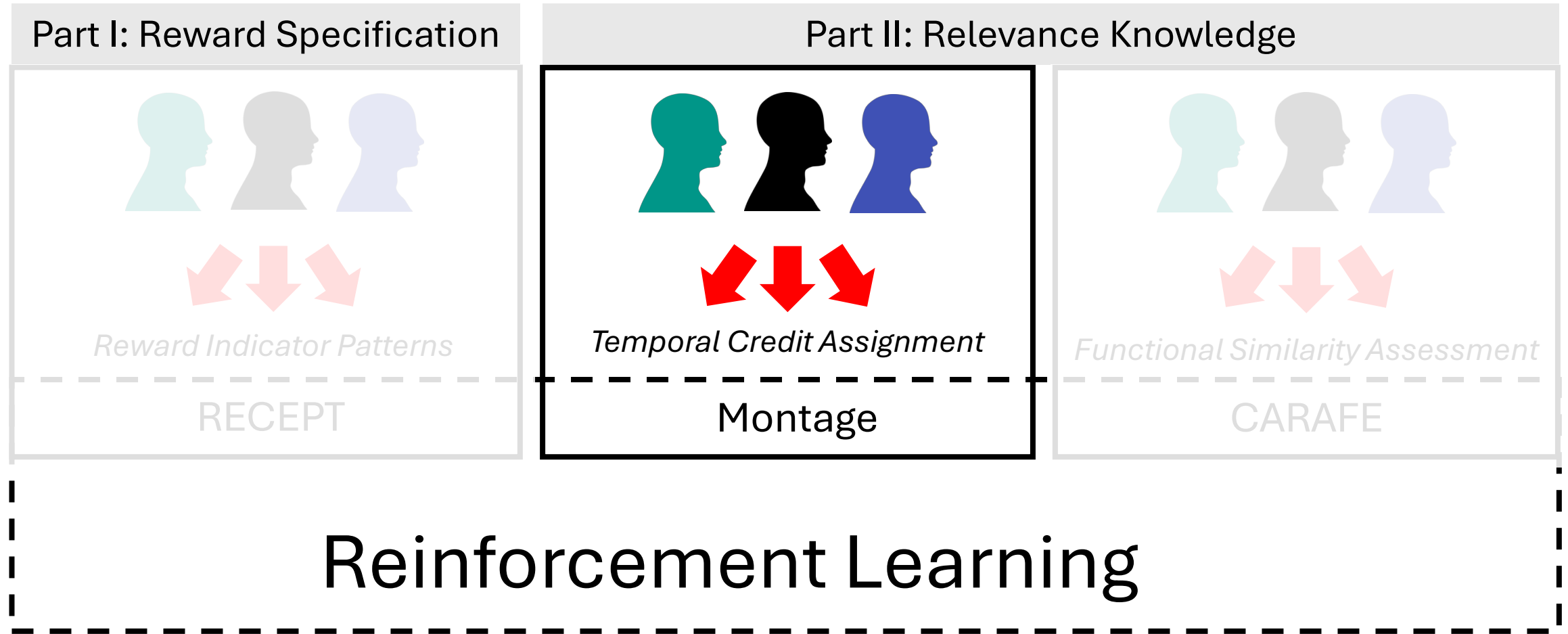
- Rewards define the **goal**, but not the **path**. This forces agents into inefficient, exhaustive exploration.
- **Relevance** — a concept that leverages human intuition to identify states that can significantly influence outcomes.
- Two Proposed Frameworks:
 - **Retrospective:** Improves **temporal credit assignment** by identifying critical past decisions.



From Rewards to Relevance

- Rewards define the **goal**, but not the **path**. This forces agents into inefficient, exhaustive exploration.
- **Relevance** — a concept that leverages human intuition to identify states that can significantly influence outcomes.
- Two Proposed Frameworks:
 - **Retrospective**: Improves **temporal credit assignment** by identifying critical past decisions.
 - **Prospective**: Inject **relevance prior** by comparing states of their potential future outcomes.





Motivation

- What leads to the undesired behavior?

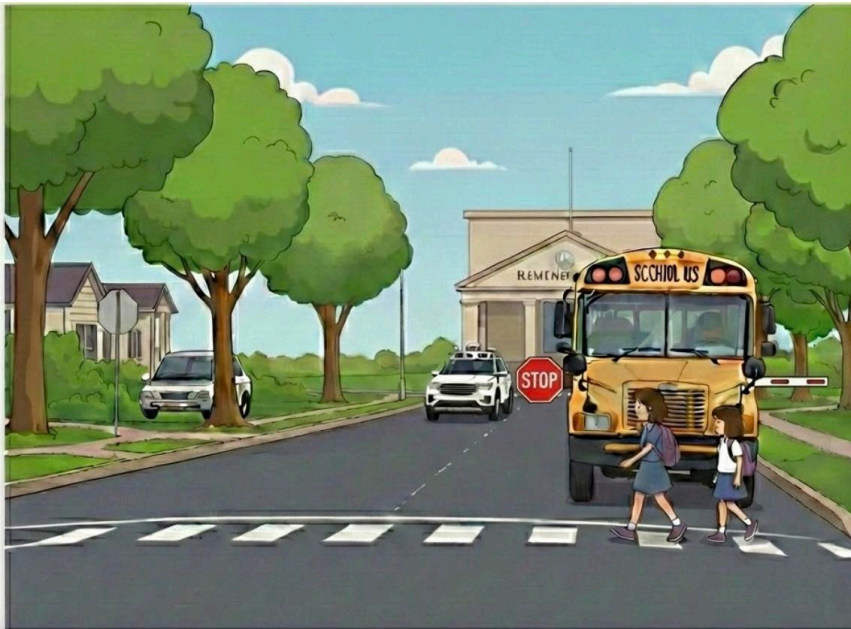
Observed Failure



Retrospective Relevance

- What leads to the undesired behavior?

Critical Decision Point



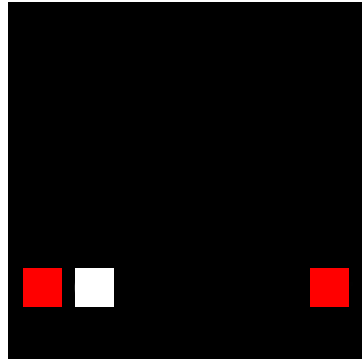
Observed Failure



←
Backtracking

DotShooter

Observation:



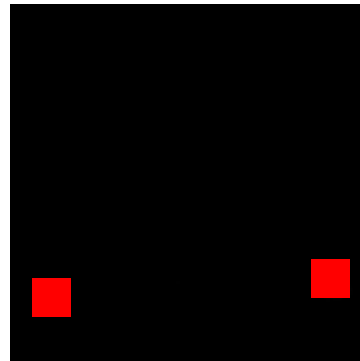
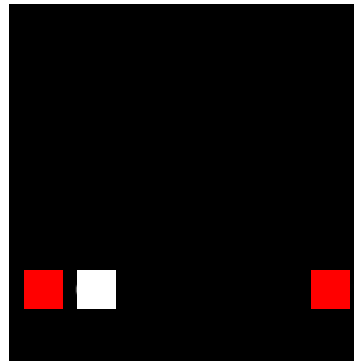
Action:

Up / Down / Shoot / Noop

- Goal: Emit pellets (white block) to hit the target (red block on the right)
 - One episode allows one hit.

DotShooter

Observation:



Action:

Up

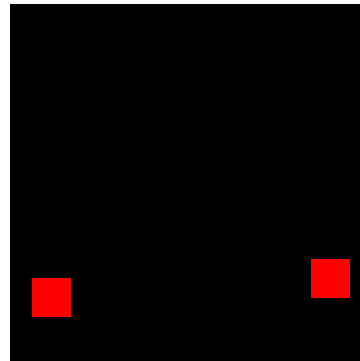
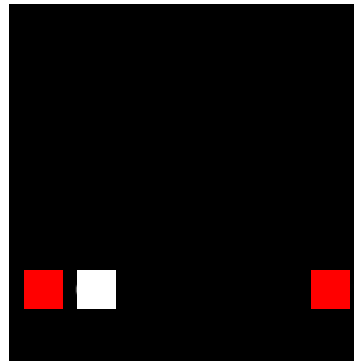
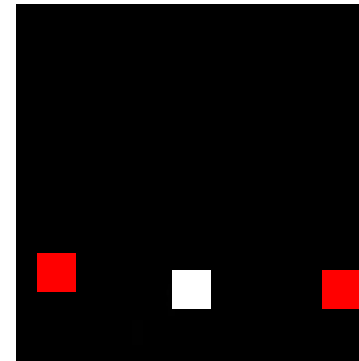
Shoot

Step 1

Step 2

DotShooter

Observation:



Action:

Up

Shoot

Up

Step 1

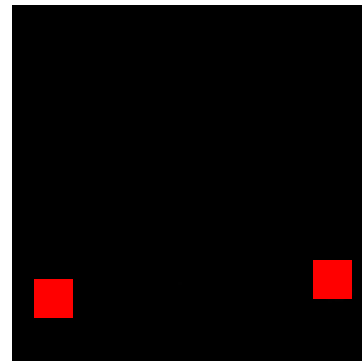
Step 2

Step 3

DotShooter

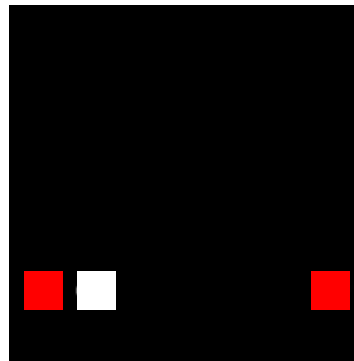
Observation:

Action:



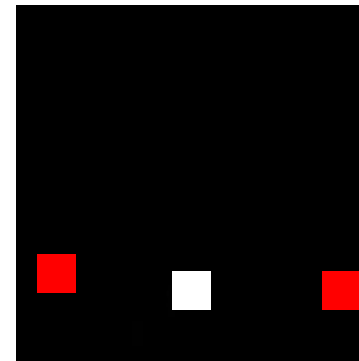
Up

Step 1



Shoot

Step 2



Up

Step 3

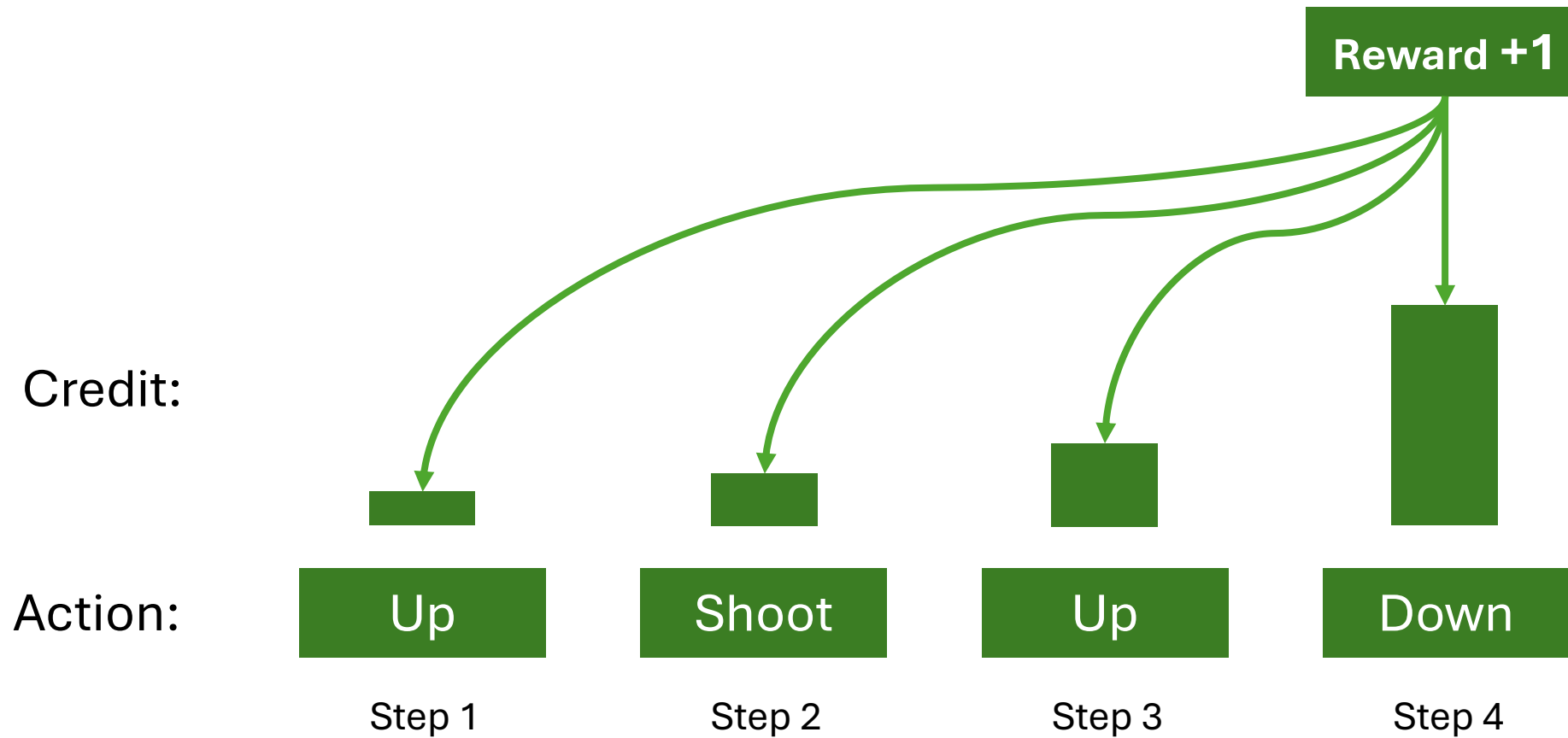


Down

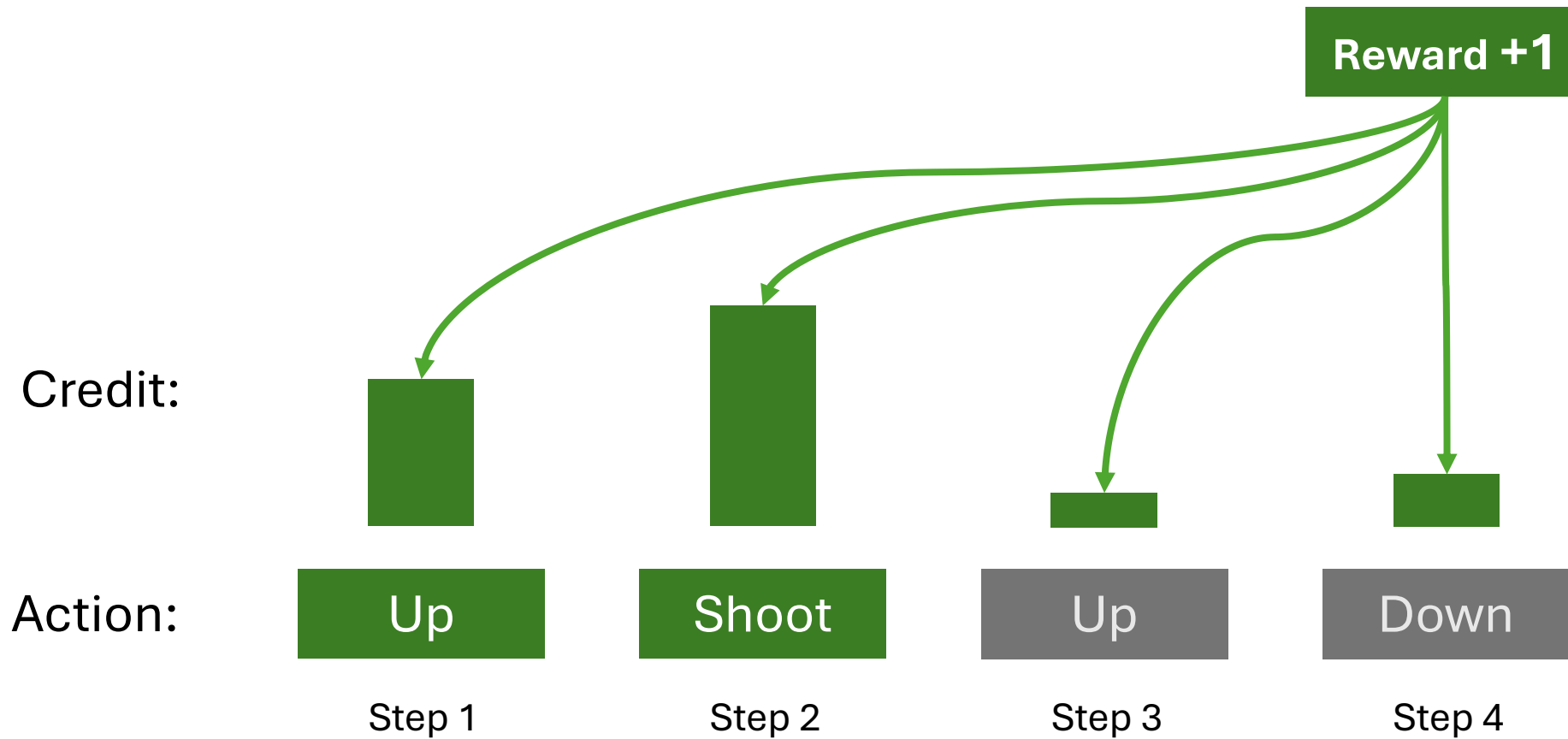
Step 4

Reward +1

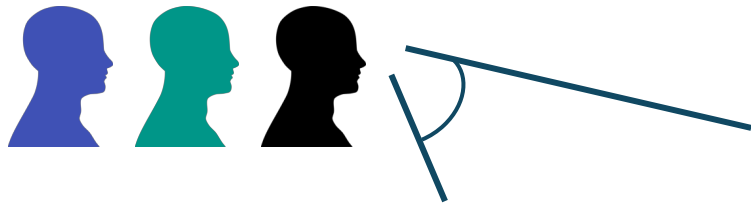
Recency Weighting



After Many Updates

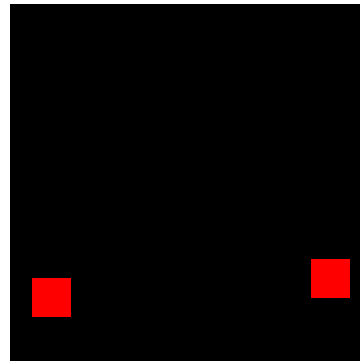
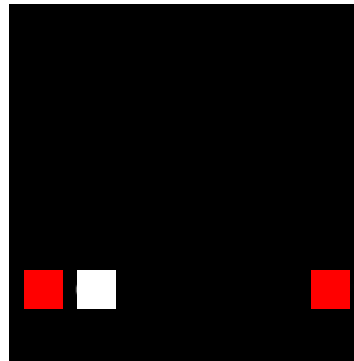
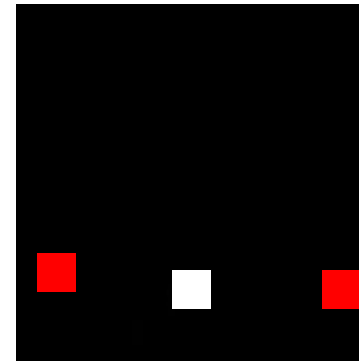


Bring a Crowd to the Loop



Reward +1

Observation:



Action:

Up

Shoot

Up

Down

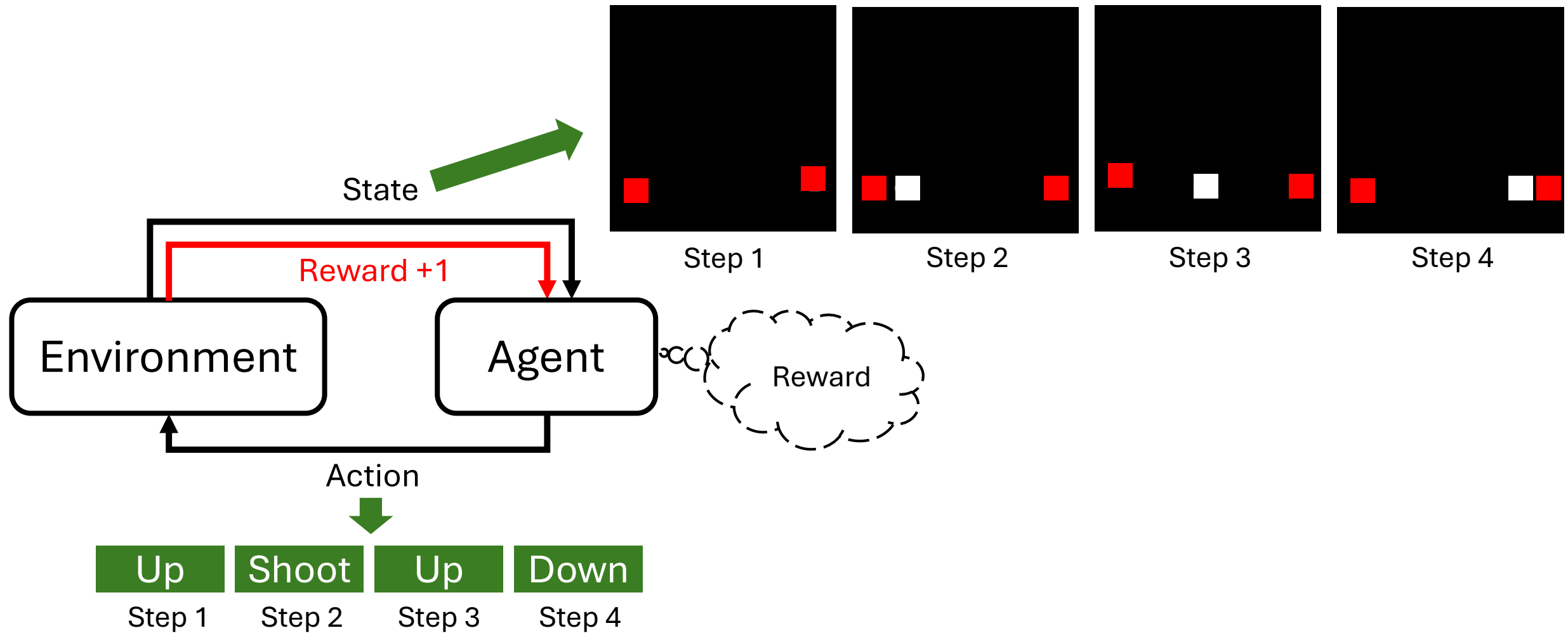
Step 1

Step 2

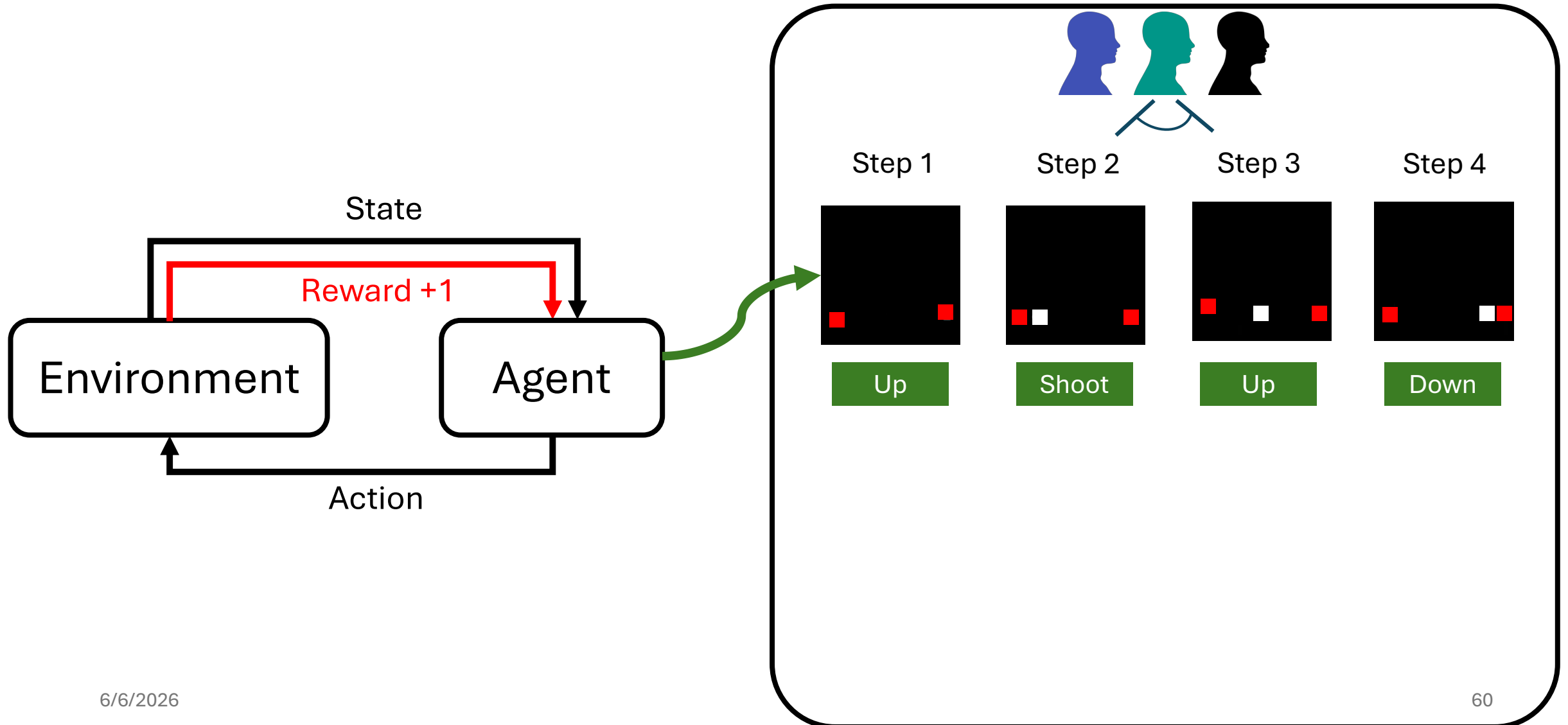
Step 3

Step 4

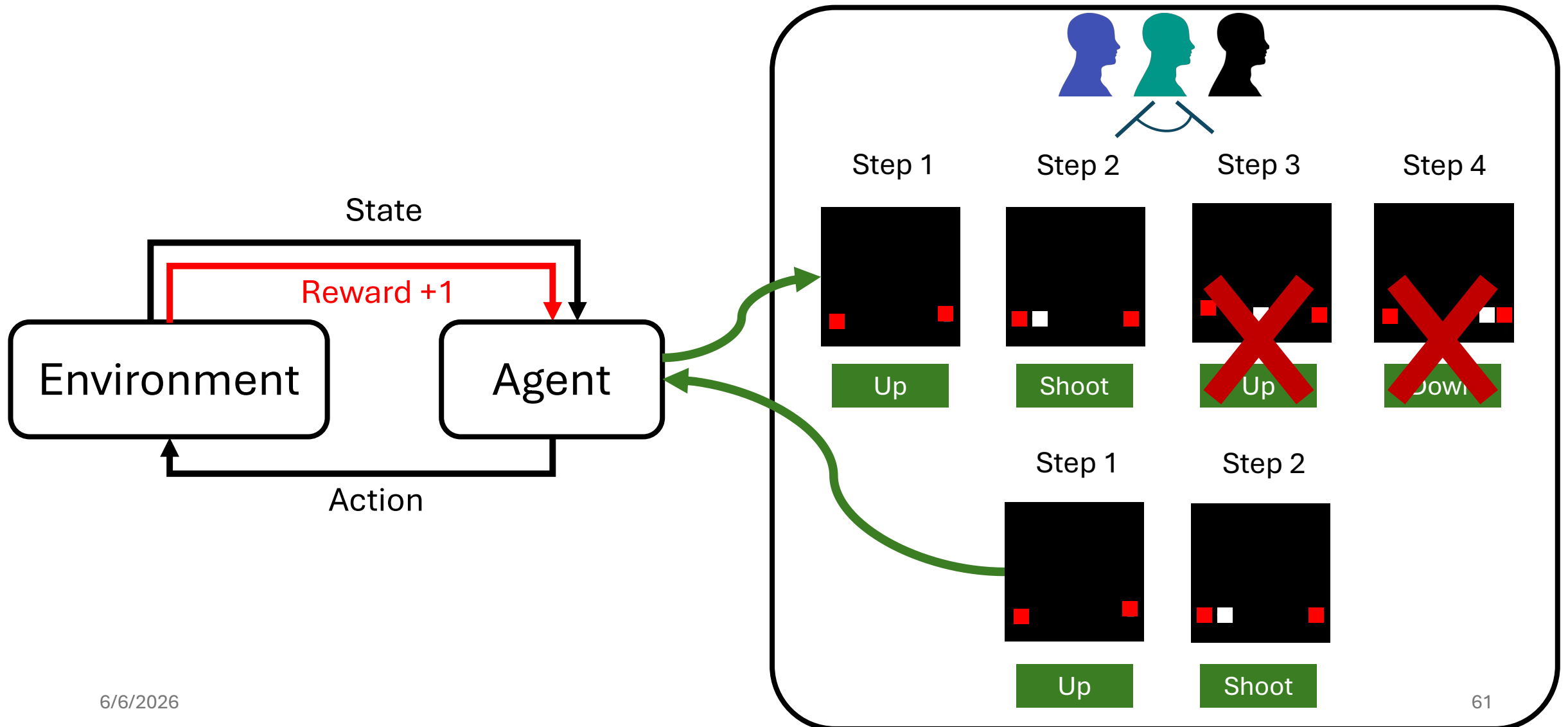
Selective Memory Learning



Selective Memory Learning

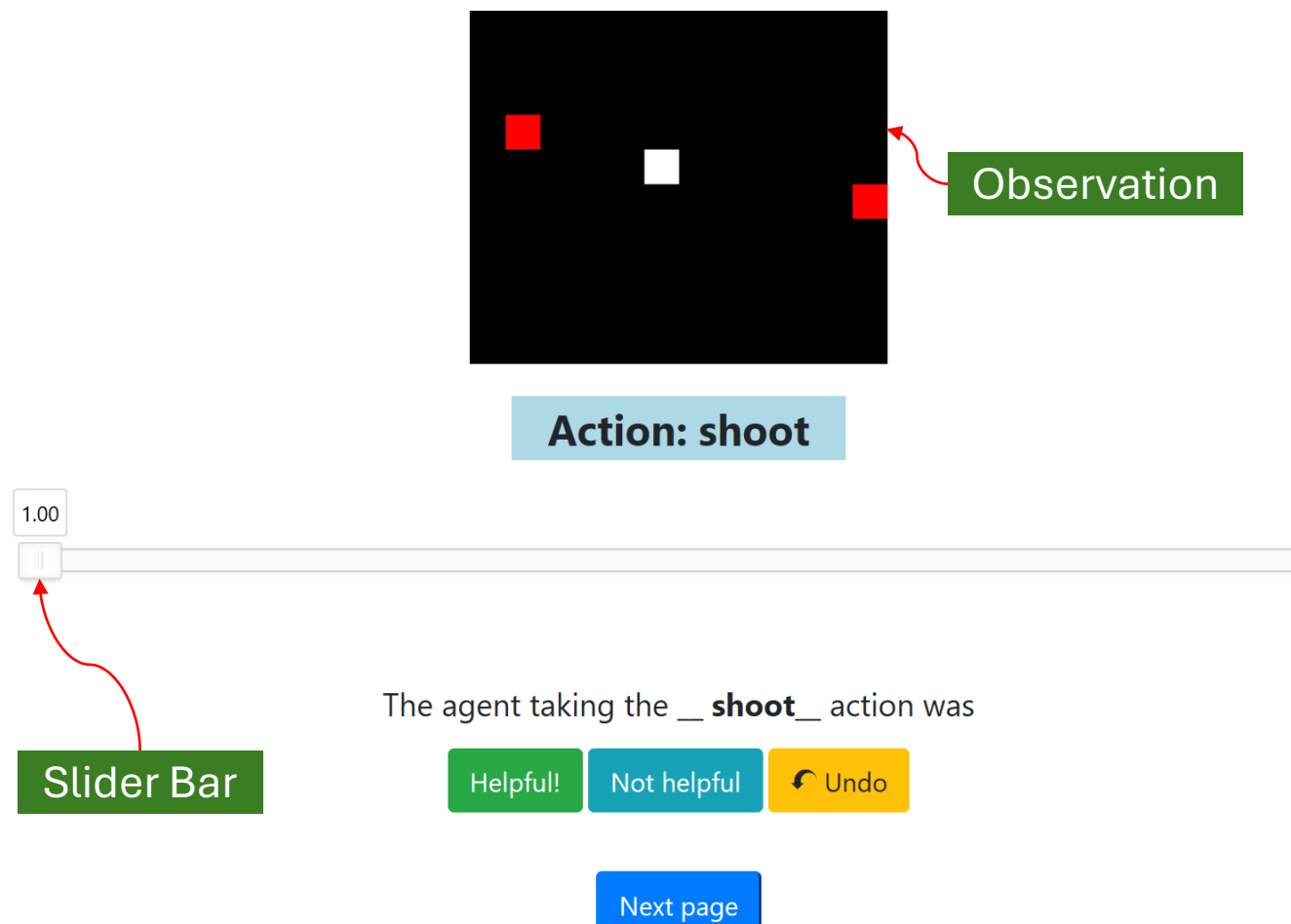


Selective Memory Learning



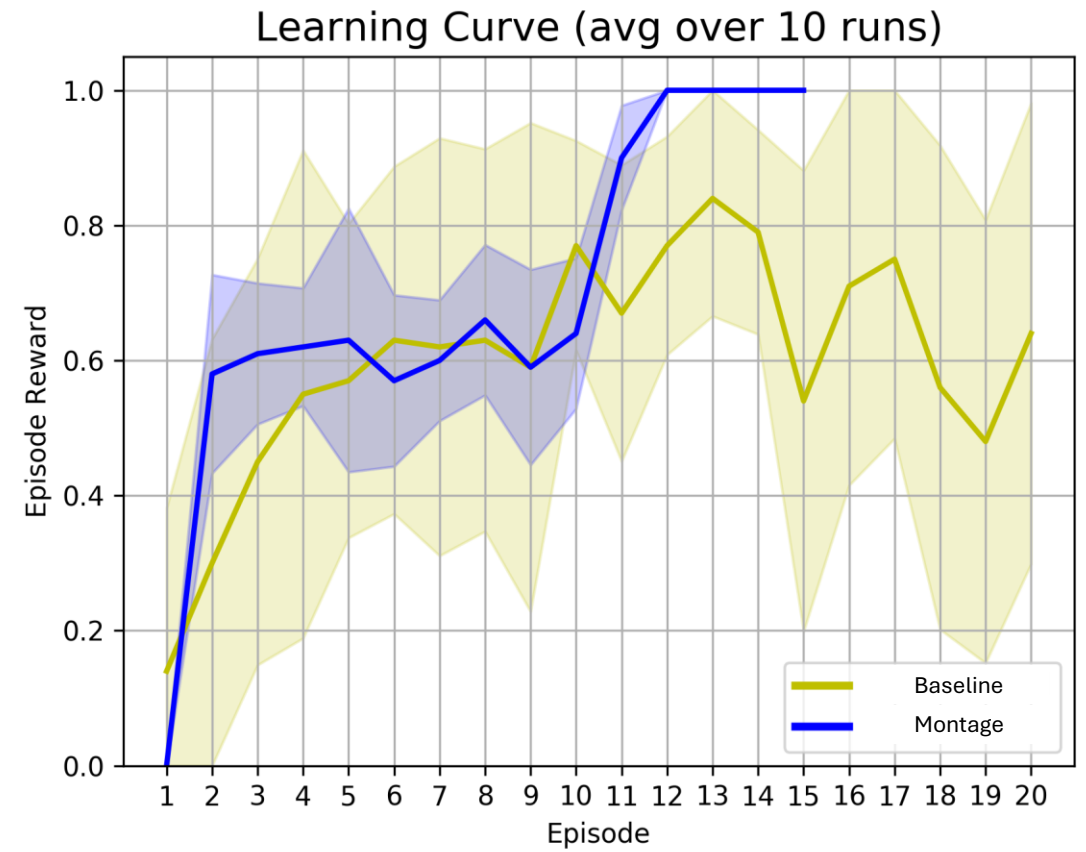
Montage User Interface

- Workers can click on the slider bar to view each step and choose **Helpful!** or **Not helpful** to determine whether the state and the action contributes to the reward.



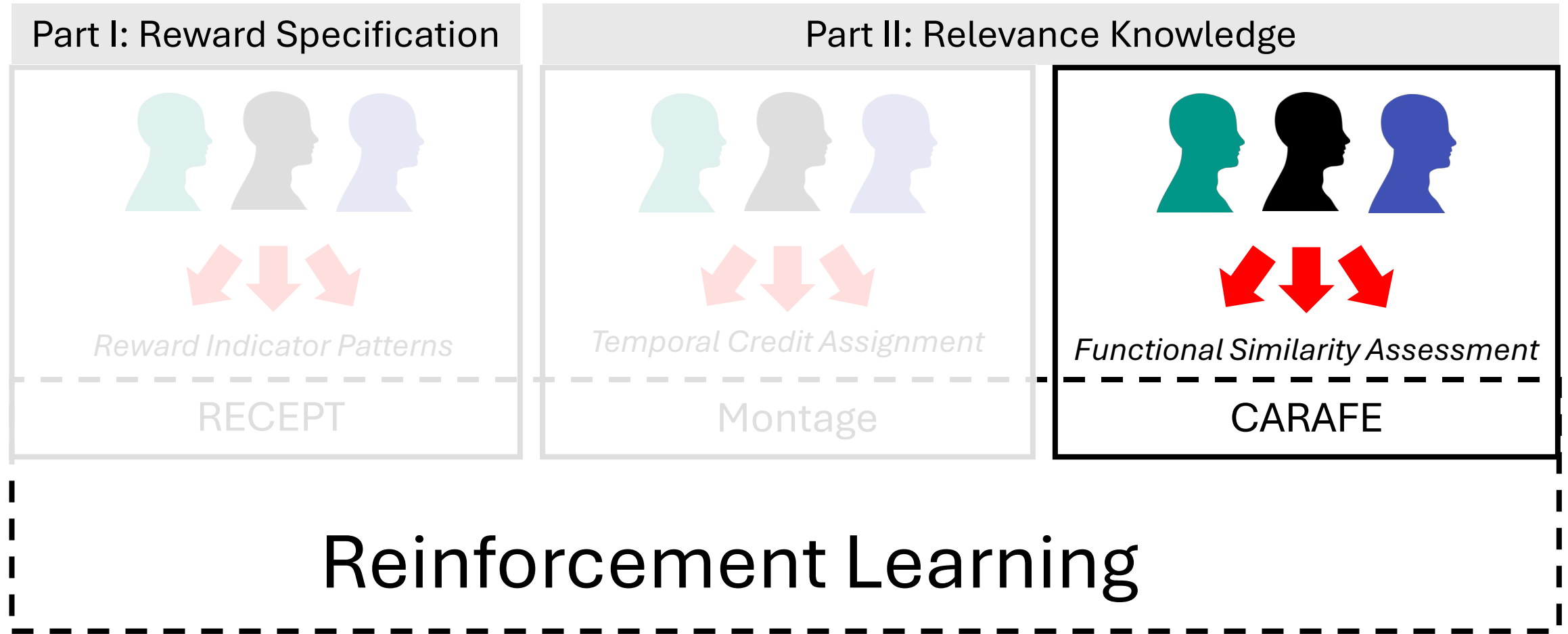
Crowd Study Result

- **Yellow curve** -> Monte Carlo Control (baseline)
- **Blue curve** -> Montage
- Montage achieved **85.6% accuracy** and **1.0 Recall** against the optimal policy.
 - We collected 10 worker annotations per episode and analyzed every possible 5-person team combination (252 total teams).



From Retrospective to Prospective

- Montage's Limitations
 - Retrospective: Requires collected trajectories before learning
 - Gated learning loop: Cannot update policy until human review
 - Lack scalability: Continuous human intervention needed



Motivation

- Similarity comparison:
 - Passing a school bus vs. delivery truck vs. ambulance
- Relevance prior:
 - Learn a metric to quantify this similarity, effectively internalizing human intuition
- The agent generalizes to new situations without needing constant human feedback.



Similarity Elicitation



Crowd Annotation:
“which two states are the most similar?”



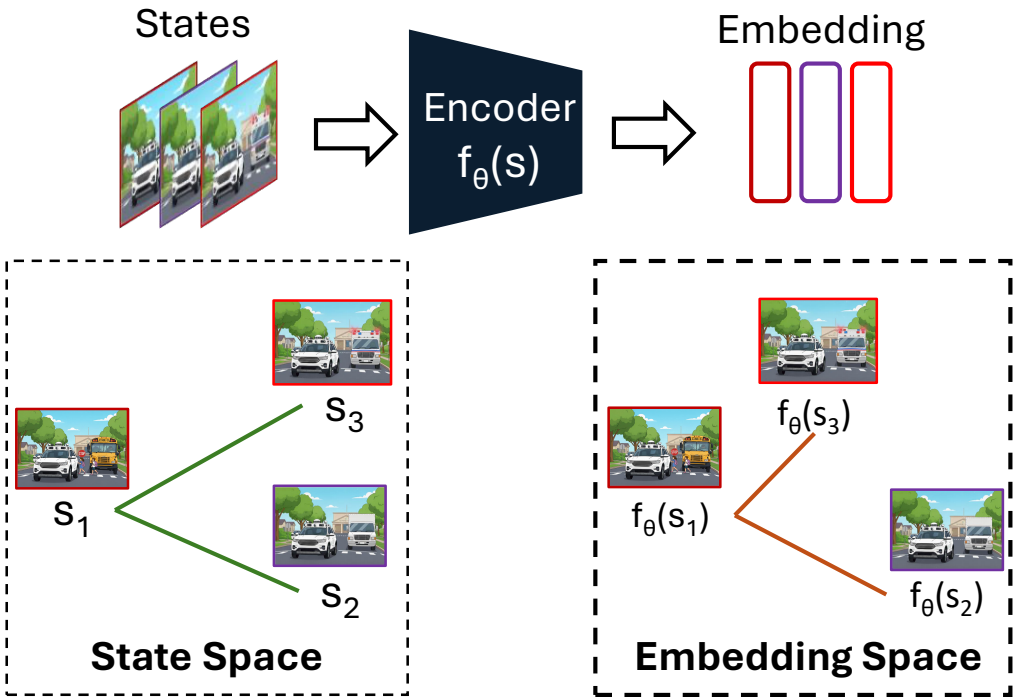
s₁ s₂ s₃

Functional Similarity Judgement

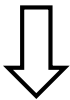
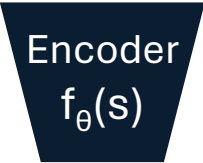
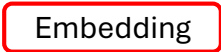
Triplet-based Representation Learning

Triplet Loss Objective:

$$\left\| \boxed{f_{\theta}(s_1)} - \boxed{f_{\theta}(s_3)} \right\| < \left\| \boxed{f_{\theta}(s_1)} - \boxed{f_{\theta}(s_2)} \right\|$$



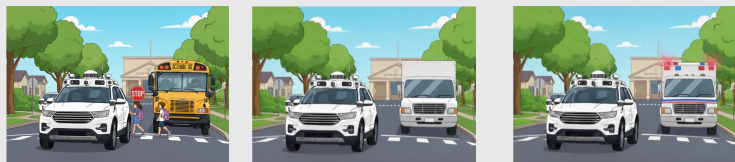
RL in Target Env



Similarity Elicitation



Crowd Annotation:
“which two states are the most similar?”



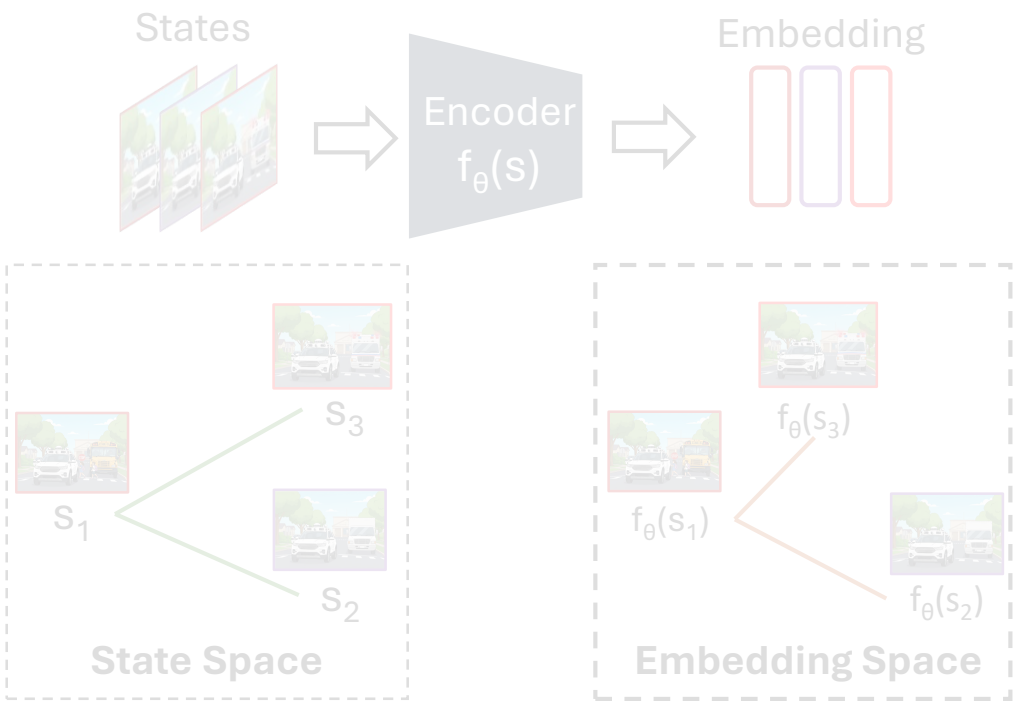
S₁ S₂ S₃

Functional Similarity Judgement

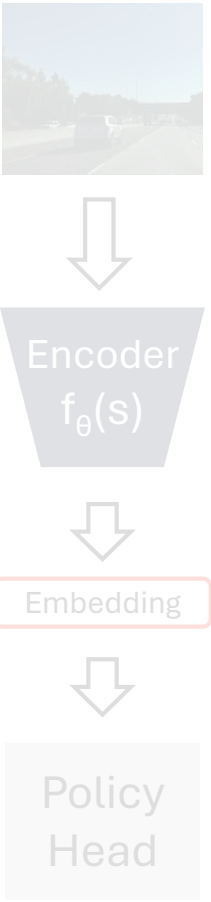
Triplet-based Representation Learning

Triplet Loss Objective:

$$\left\| f_{\theta}(s_1) - f_{\theta}(s_3) \right\| < \left\| f_{\theta}(s_1) - f_{\theta}(s_2) \right\|$$

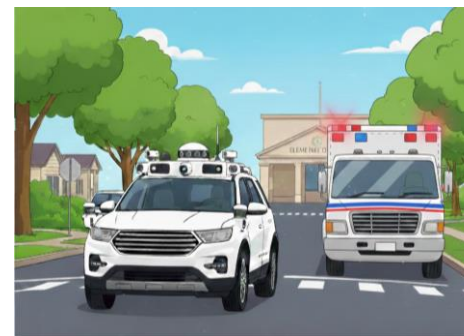


RL in Target Env



Eliciting Similarity via *Comparison*

- Asking humans directly to provide relevance prior could be demanding for non-experts.
- However, when presented as a similarity comparison task, humans can provide more robust responses via side-by-side comparison.

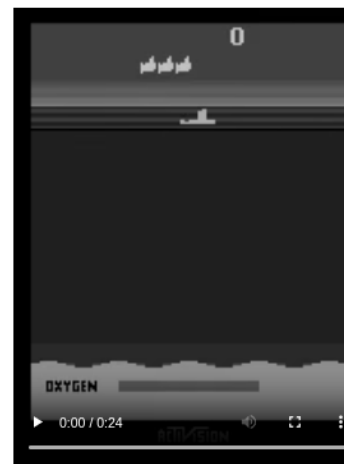


CARAFE User Interface

- Design Features
 - Game-specific tutorials and instructions

Do not open this page on multiple tabs.

Do not use the back button.



The video to the left shows the ATARI game **seaquest** being played. Your task is to identify when different scenes from the game capture similar situations.

[Click here to read more about how the game is played](#)

The player uses a submarine to shoot at enemies and rescue divers. Enemies include sharks and submarines, which shoot missiles at the player's submarine. The player must ward off the enemies by firing an unlimited supply of missiles while trying to rescue divers swimming through the water. The points awarded to the player for shooting an enemy starts at 10 points each, and increases as the game advances. The sub can hold up to six divers at a time. Each time the player resurfaces prior to having a full load of six divers, one of the divers is removed.

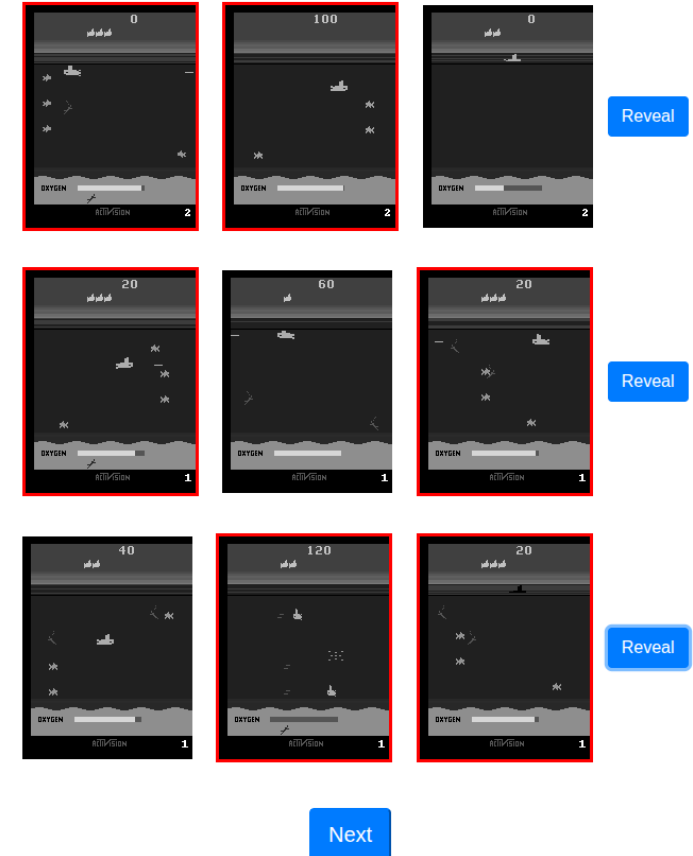
The submarine has a limited amount of oxygen. The player must surface often in order to replenish the oxygen, but if the player resurfaces without any rescued divers, they will lose a life. If the player resurfaces with the maximum amount of divers, they will gain bonus points for the sub's remaining oxygen. Each time the player surfaces, the game's difficulty increases; enemies increase in number and speed. Eventually an enemy sub begins patrolling the surface, leaving the player without a safe haven.

[Next](#)

CARAFE Interface

- Design Features
 - Game-specific tutorials and instructions
 - Pre-selection Priming

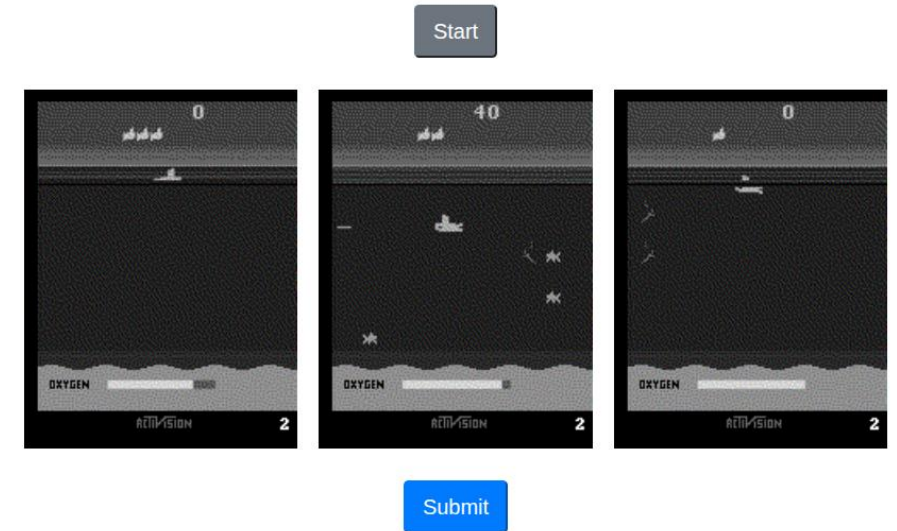
Click **Reveal** to show which two scenes are most similar. Then, click **Next** to start the tasks.



CARAFE Interface

- Design Features
 - Game-specific tutorials and instructions
 - Pre-selection Priming
 - Selection Page

Click [Start](#) button to start the task.
Which two of the three scenes below are the most similar to one another?
Select two by clicking them, and then click [Submit](#).



Number of tasks left: 4

[View Instructions](#)

Similarity Elicitation



Crowd Annotation:
“which two states are the most similar?”



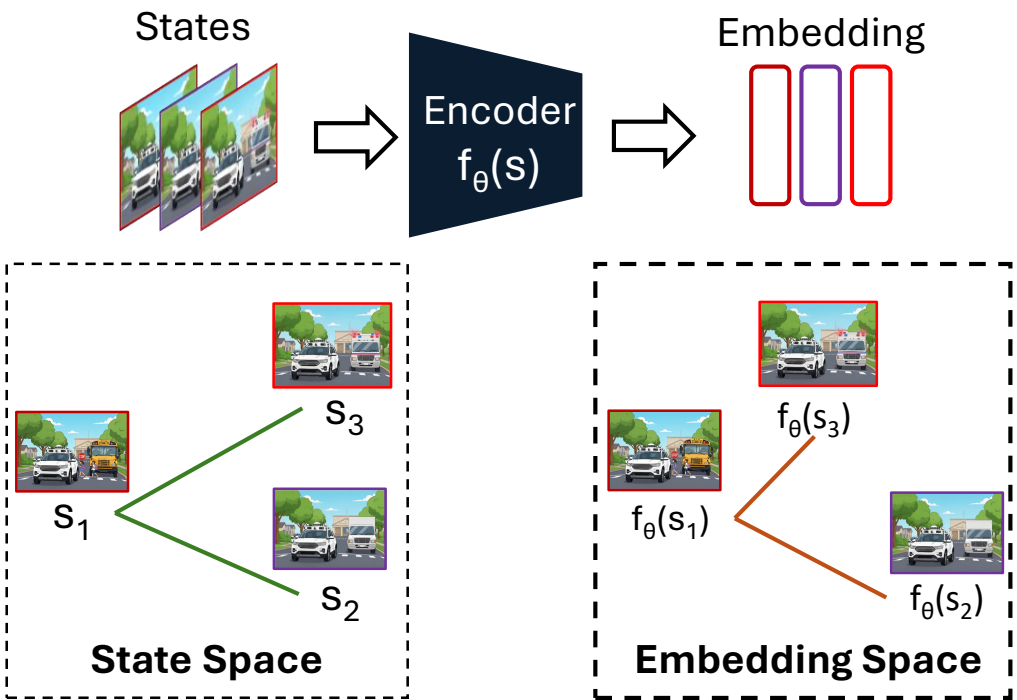
s_1 s_2 s_3

Functional Similarity Judgement

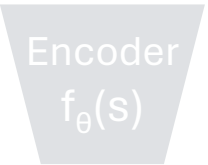
Triplet-based Representation Learning

Triplet Loss Objective:

$$\left\| f_{\theta}(s_1) - f_{\theta}(s_3) \right\| < \left\| f_{\theta}(s_1) - f_{\theta}(s_2) \right\|$$



RL in Target Env



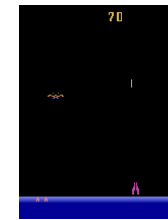
From Annotation to Learning Objective

- Once a triplet is annotated, we use *triplet loss* to train a state encoder.

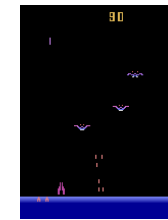
$$L(s_1, s_2, s_3) = \max(\|f_\theta(s_1) - f_\theta(s_3)\|_2^2 - \|f_\theta(s_1) - f_\theta(s_2)\|_2^2 + \alpha, 0)$$

Similarity Annotation

State 1



State 2



State 3



From Annotation to Learning Objective

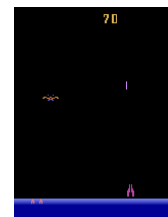
- Once a triplet is annotated, we use *triplet loss* to train a state encoder.

$$L(s_1, s_2, s_3) = \max(\|f_\theta(s_1) - f_\theta(s_3)\|_2^2 - \|f_\theta(s_1) - f_\theta(s_2)\|_2^2 + \alpha, 0)$$

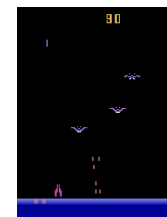
- If state 1 and 3 are selected, they ought to be closer in the embedding space.

Similarity Annotation

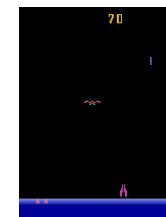
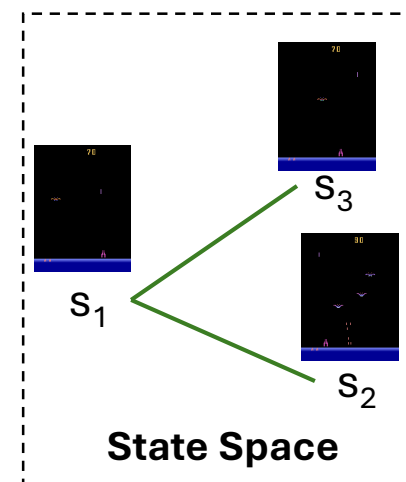
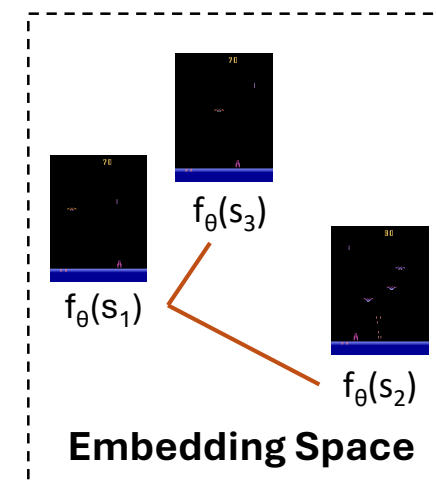
State 1



State 2



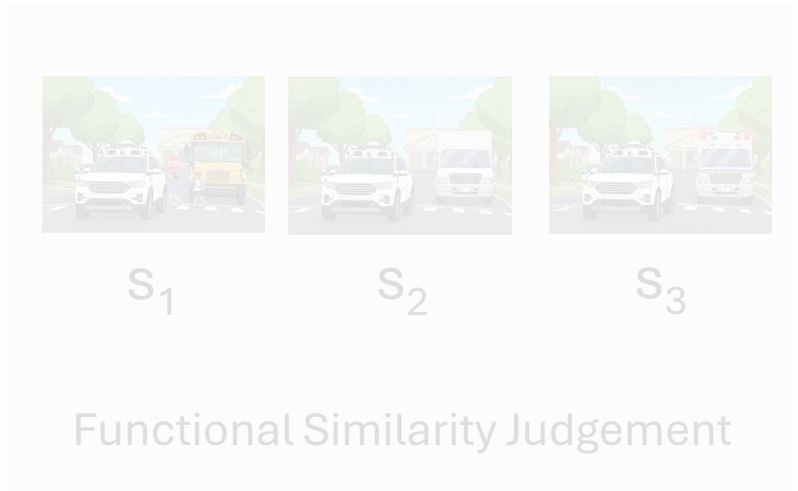
State 3



Similarity Elicitation



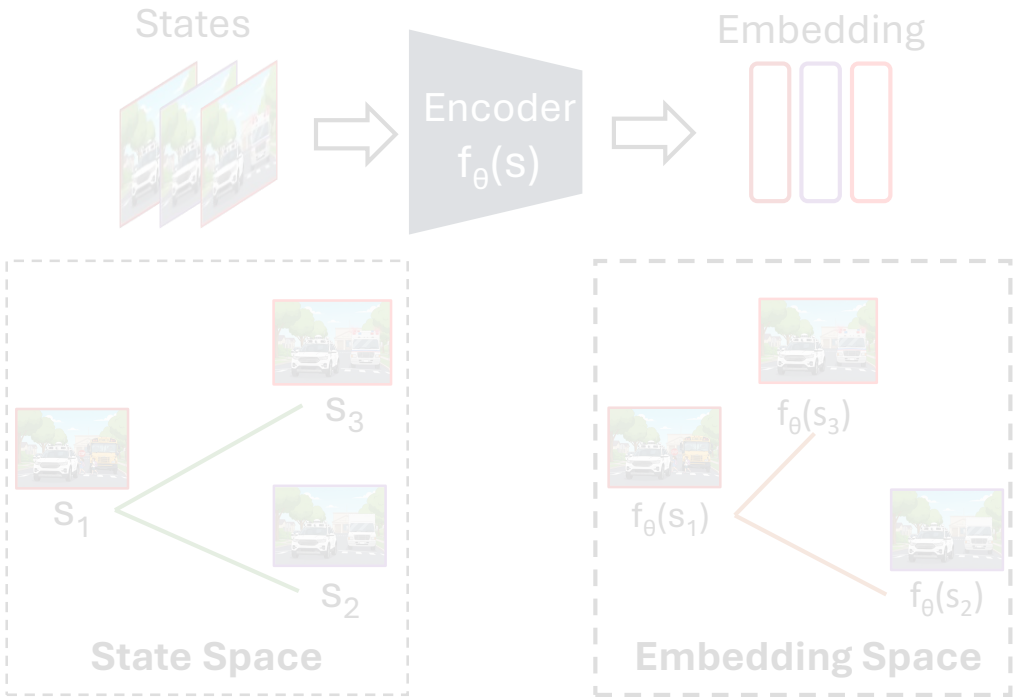
Crowd Annotation:
“which two states are the most similar?”



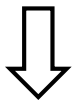
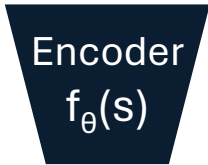
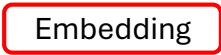
Triplet-based Representation Learning

Triplet Loss Objective:

$$\left\| f_{\theta}(s_1) - f_{\theta}(s_3) \right\| < \left\| f_{\theta}(s_1) - f_{\theta}(s_2) \right\|$$

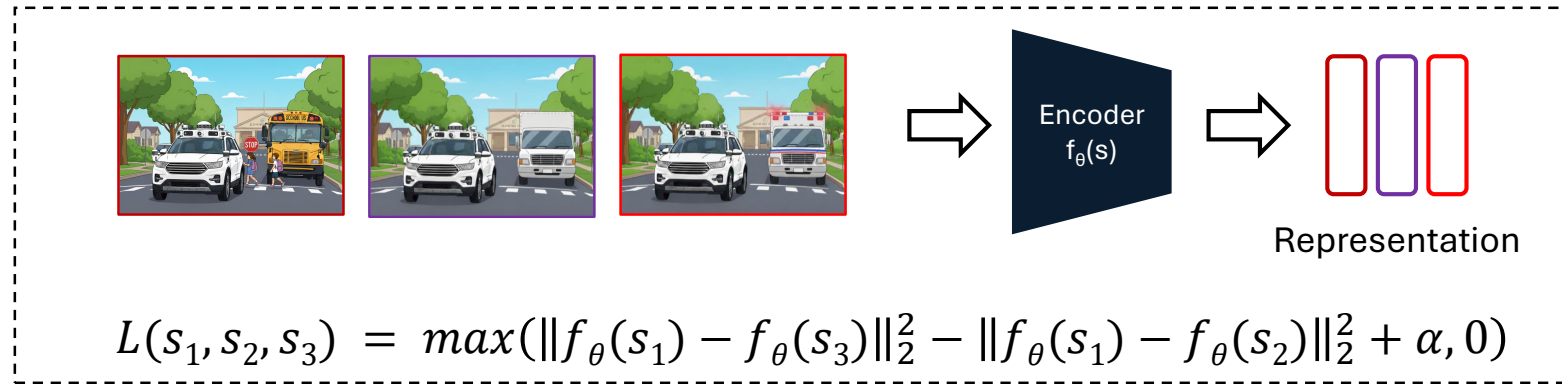


RL in Target Env



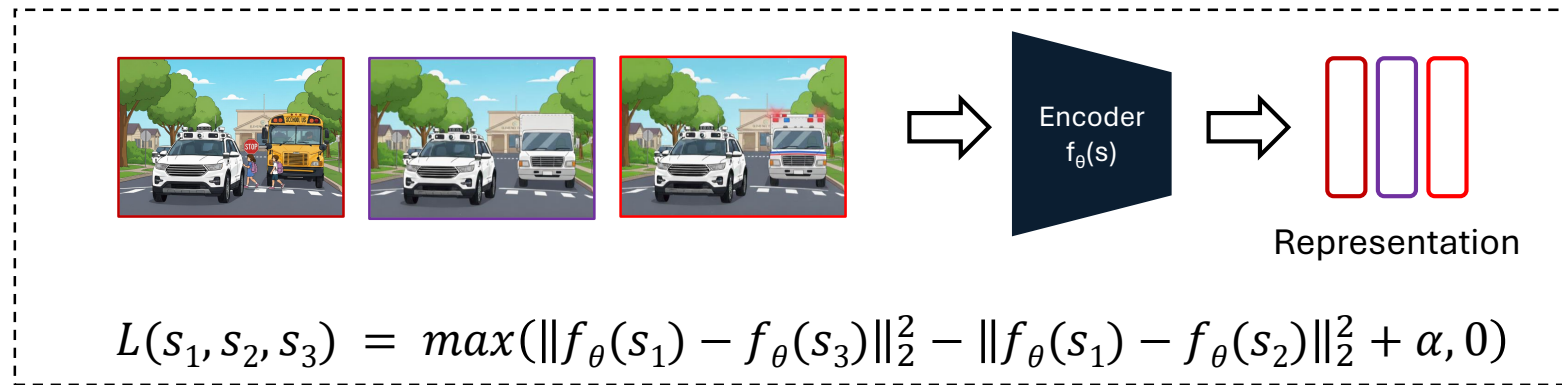
From Pretraining to RL

Pretraining

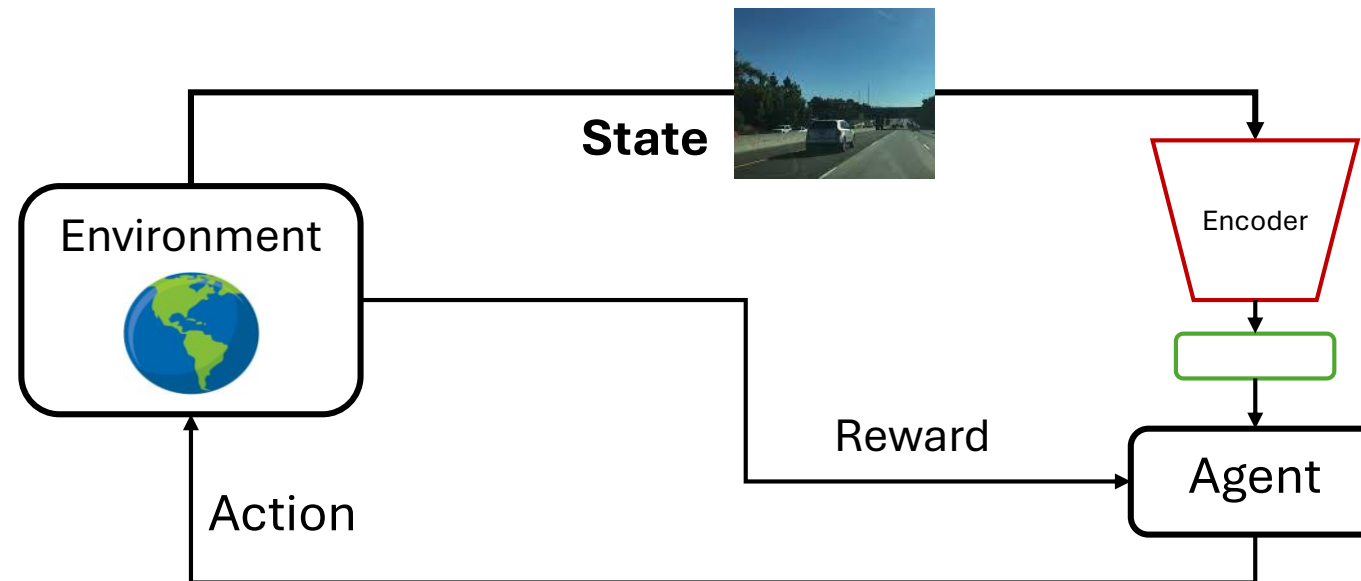


From Pretraining to RL

Pretraining

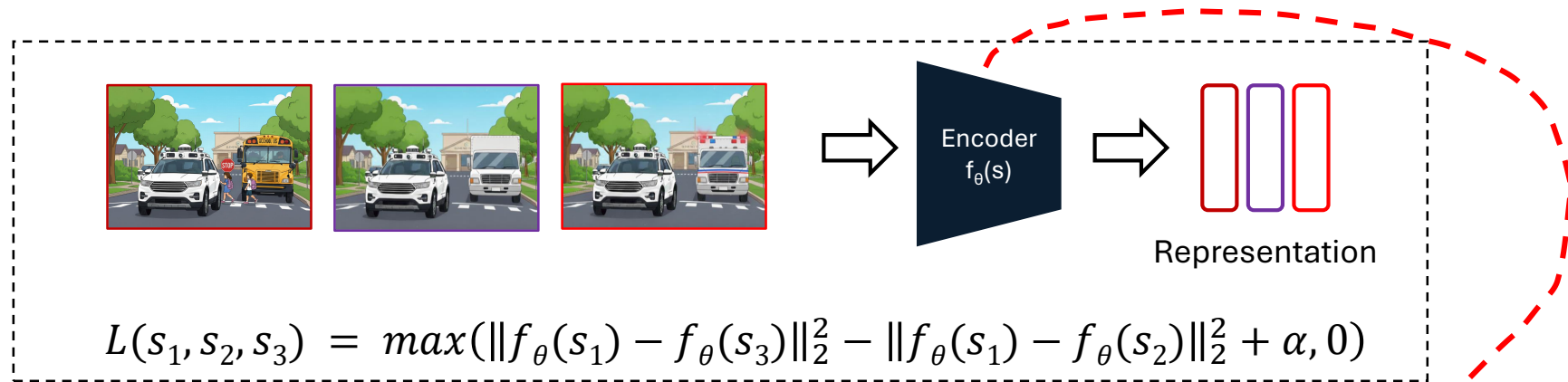


RL Training

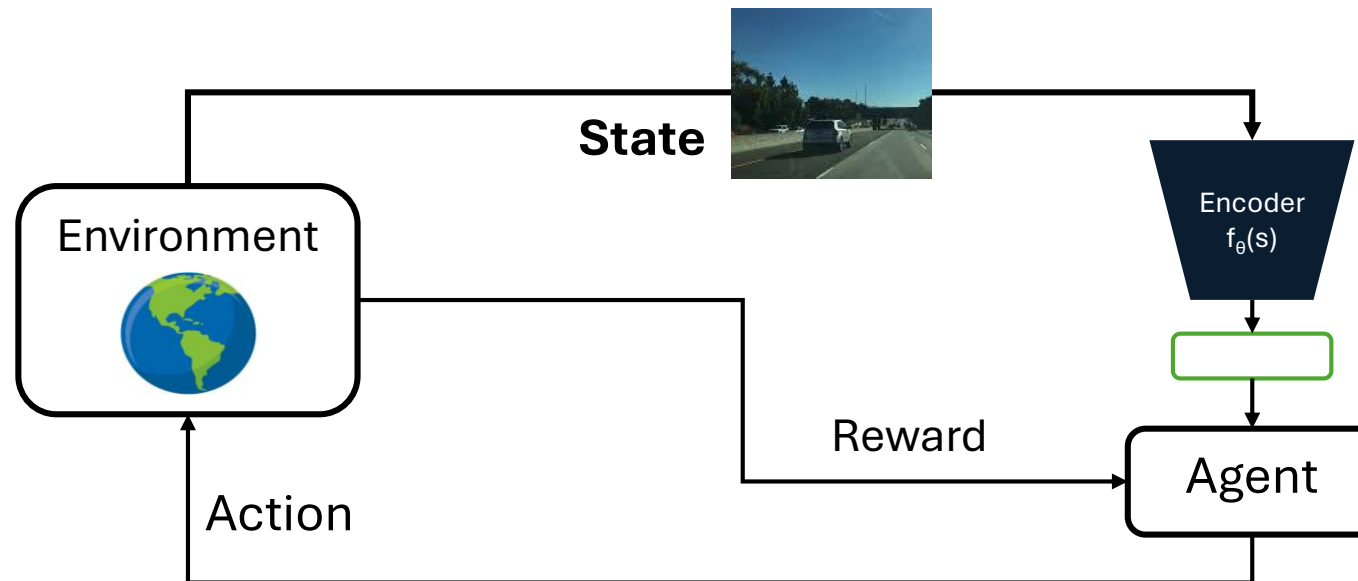


From Pretraining to RL

Pretraining



RL Training



Can Triplet-Induced Encoders Accelerate RL?

- **Premise:** Confirm relevance-based representations work in principle before using noisy crowd data.

Can Triplet-Induced Encoders Accelerate RL?

- **Premise:** Confirm relevance-based representations work in principle before using noisy crowd data.
- **Methodology:** Use an extensively-trained agent (*oracle*) to generate "perfect" similarity triplets.
 - *Assumption:* Expert embeddings naturally group functionally similar states.



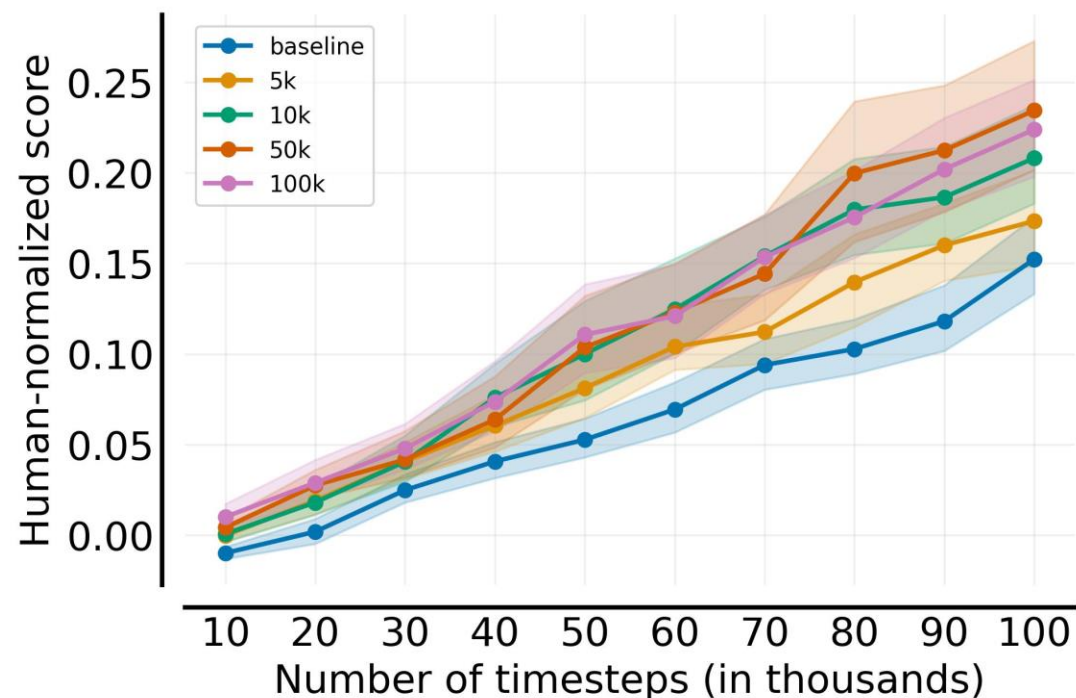
Can Triplet-Induced Encoders Accelerate RL?

- **Premise:** Confirm relevance-based representations work in principle before using noisy crowd data.
- **Methodology:** Use an extensively-trained agent (*oracle*) to generate "perfect" similarity triplets.
 - *Assumption:* Expert embeddings naturally group functionally similar states.



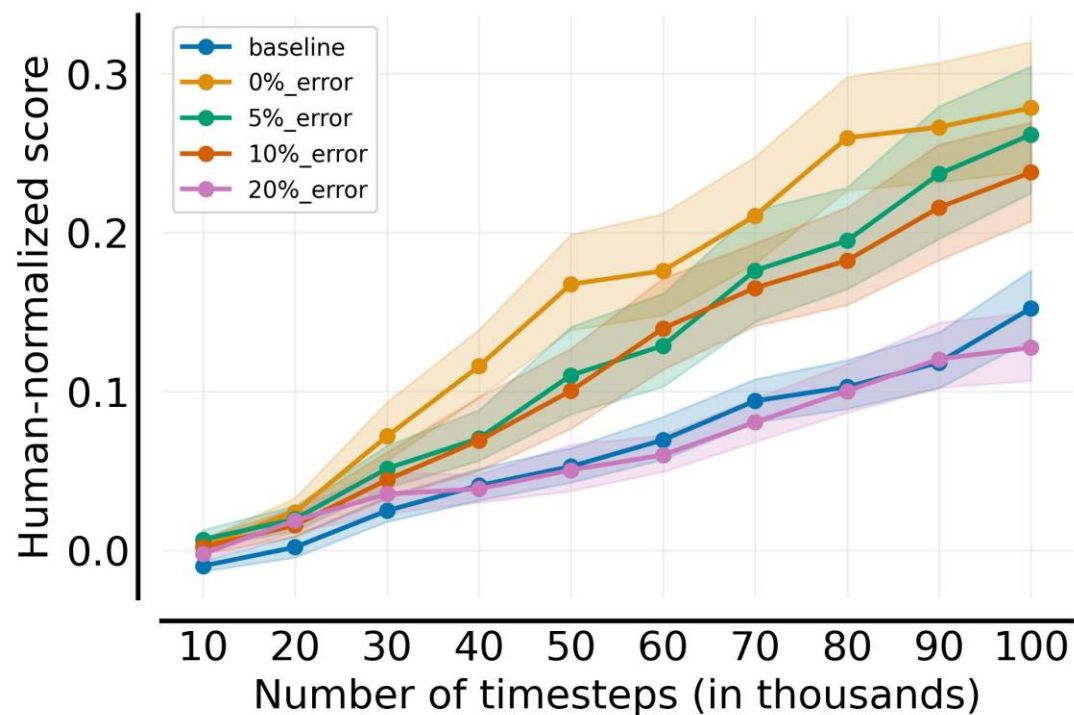
- **Robustness:** Stress-tested against data scarcity and synthetic noise

Dataset Size



- **Setup:** We varied the number of states in the dataset used to train the triplet encoder to test performance under data scarcity.
- **Result:** Encoders trained with as few as **10,000 states** enabled the agent to outperform the baseline, proving the method works with limited data.

Error Tolerance



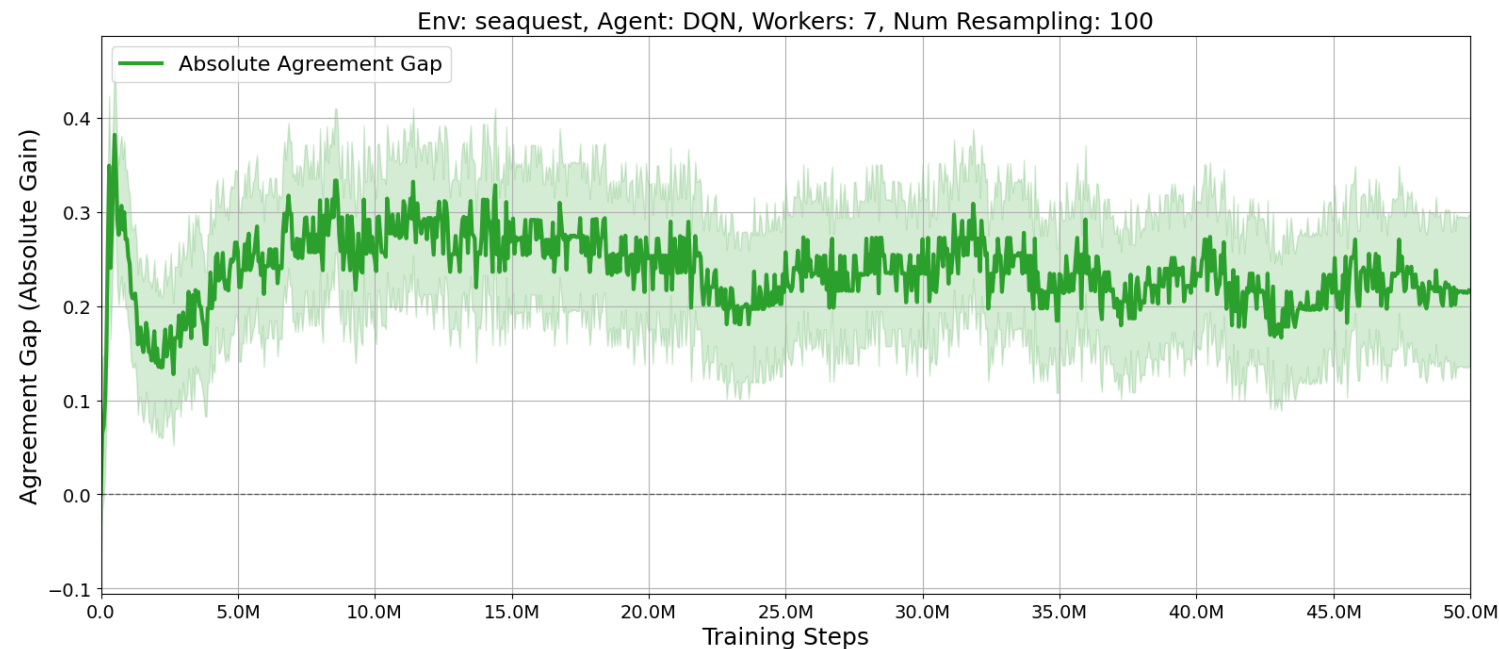
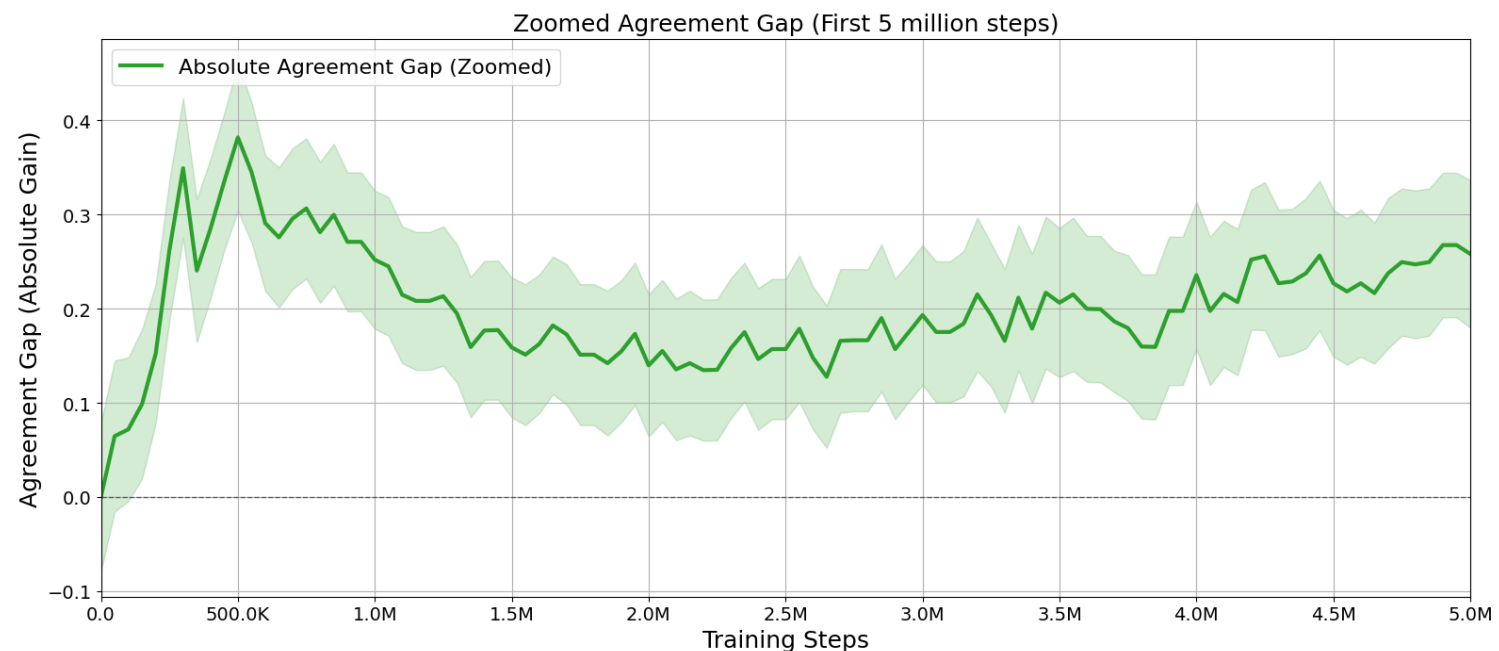
- **Setup:** We intentionally injected **5%**, **10%**, and **20%** erroneous labels into the dataset to mimic human error.
- **Result:** The method is robust and tolerates up to **10% error** without significant performance loss.

Can Humans Provide Coherent Signals?

- **The Challenge:** Can noisy human intuition approximate the "Oracle" signal established in RQ1?
- Constraint
 - End-to-end training with human data is expensive for every test case.
- Metric
 - Agreement
 - The fraction of triplets for which the majority vote of human annotators matches the similarity judgment produced by the agent's encoder.
 - Agreement Gap
 - Human – Latest vs. Human – Untrained

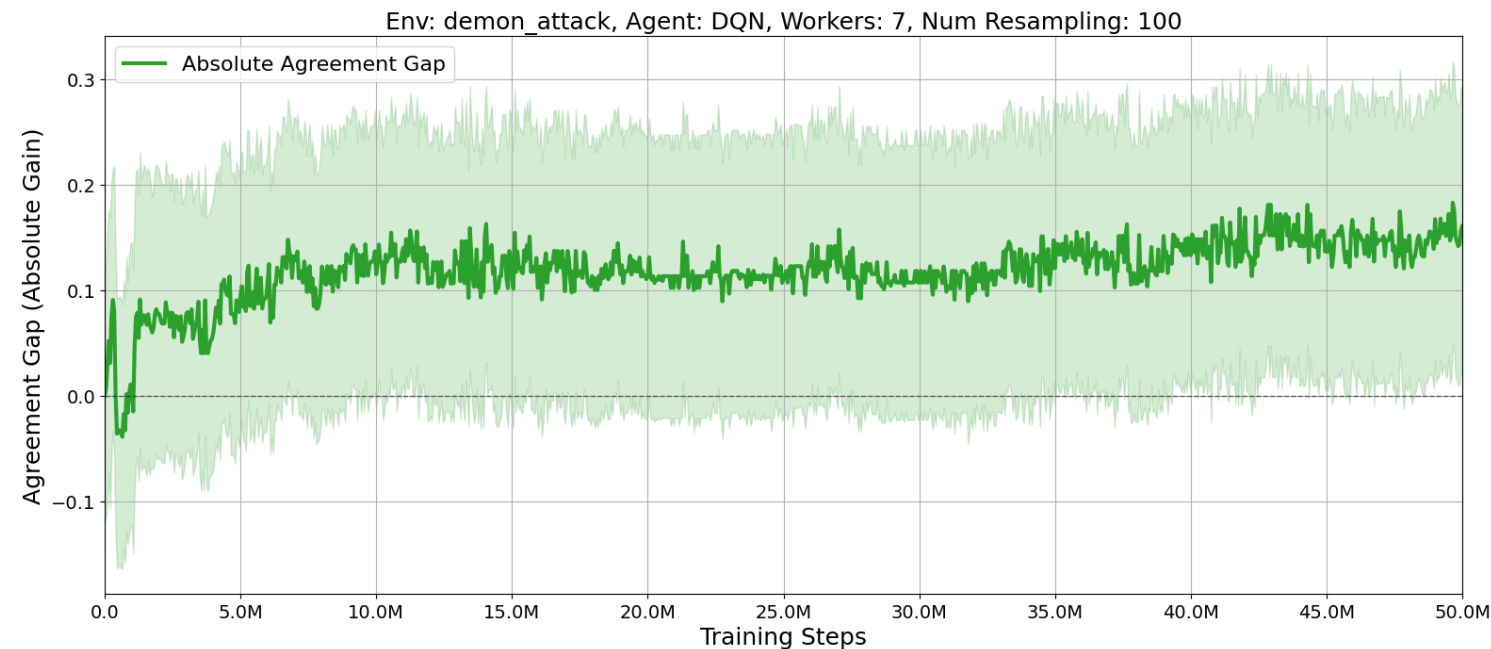
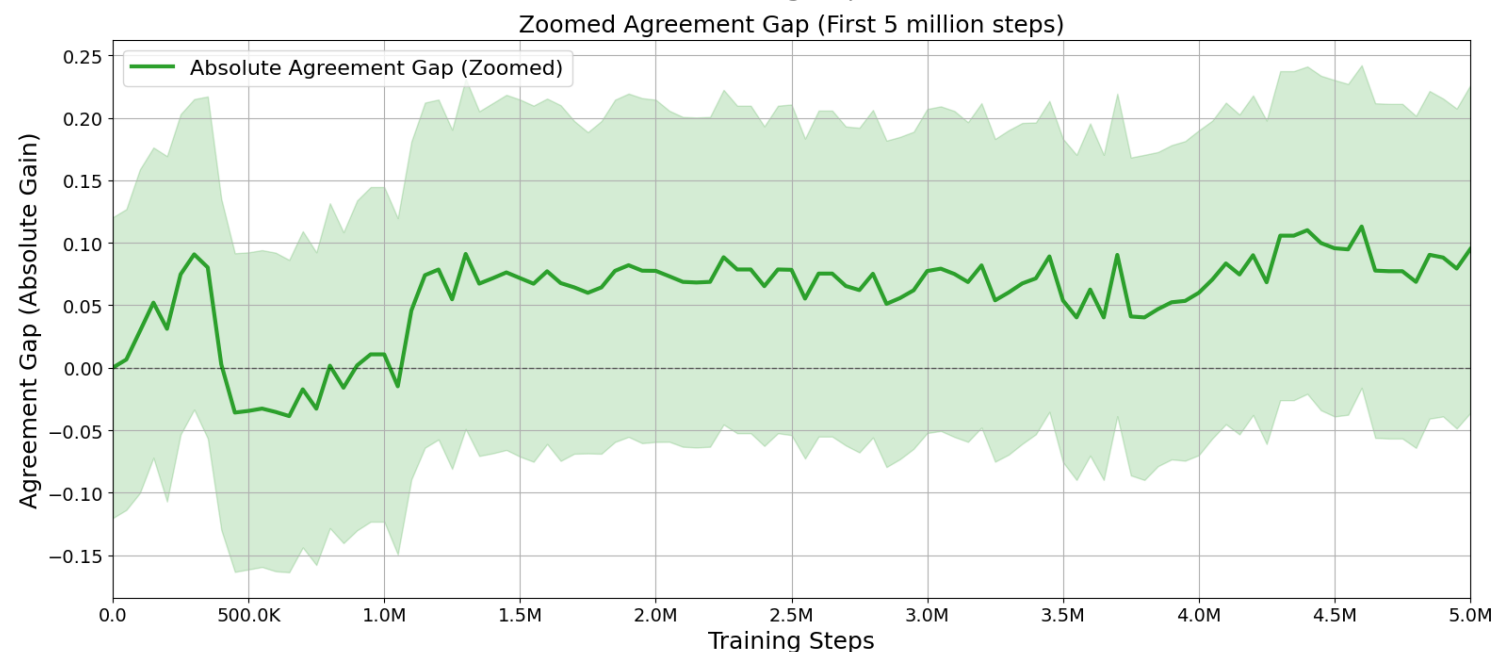
Seaquest

- The top plot shows the agreement gap throughout the entire training process (0~50M steps)
- The zoomed view (bottom, 0~5M steps) reveals a sharp peak in agreement (~0.4) within the first **500k steps**



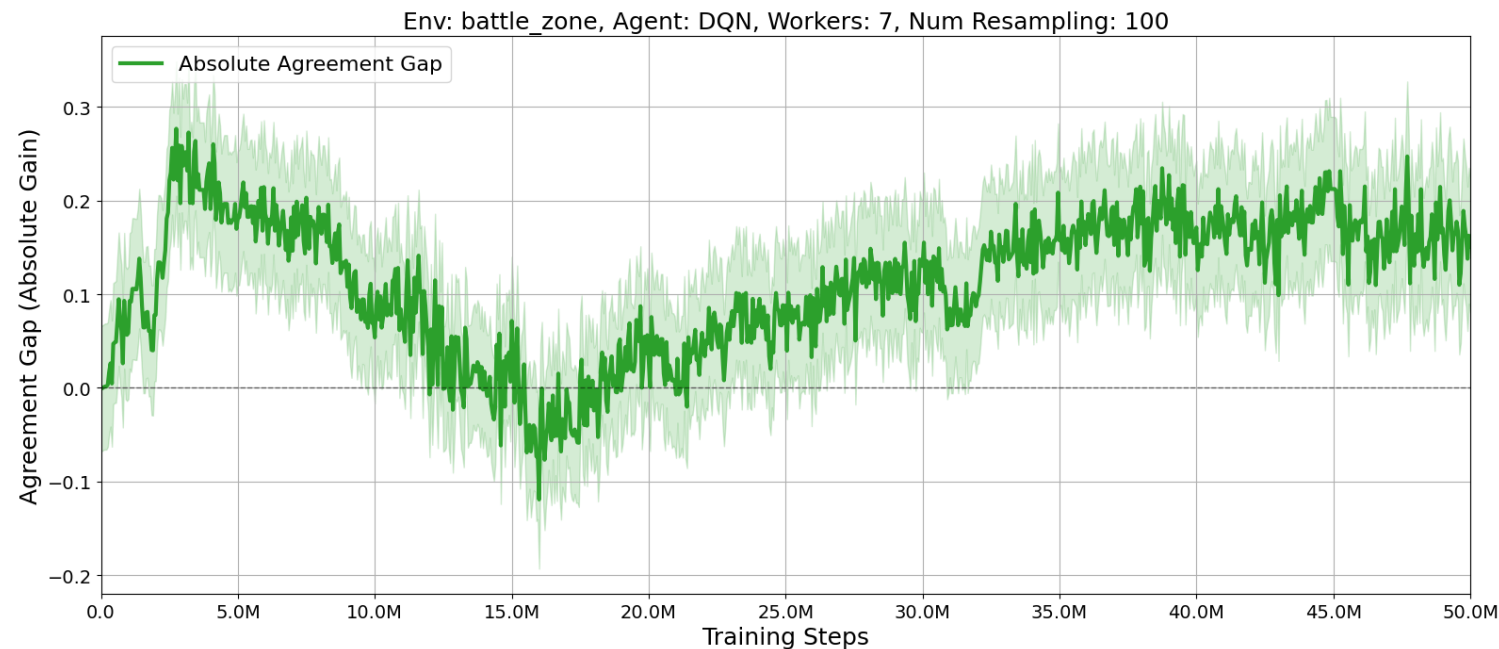
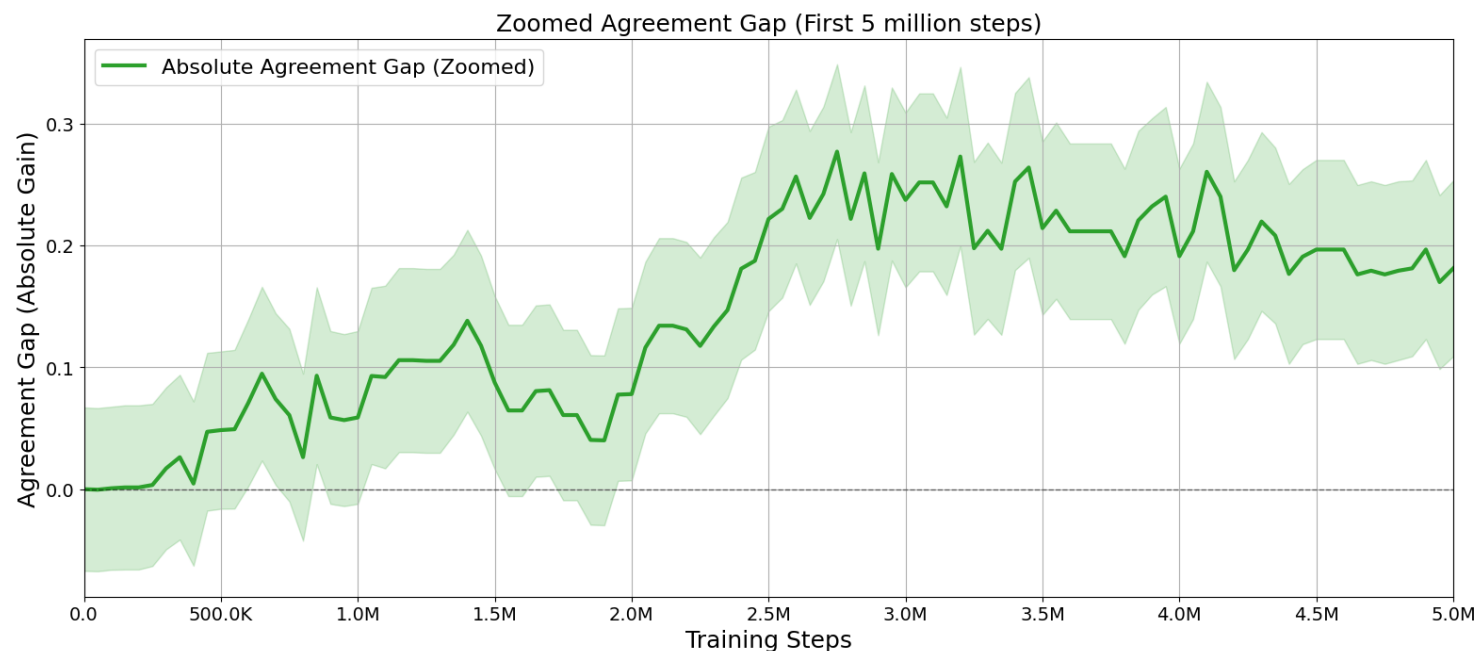
Demon Attack

- Similar to Seaquest, we see a rapid rise in agreement, peaking at **~0.1** within the first **300k** steps



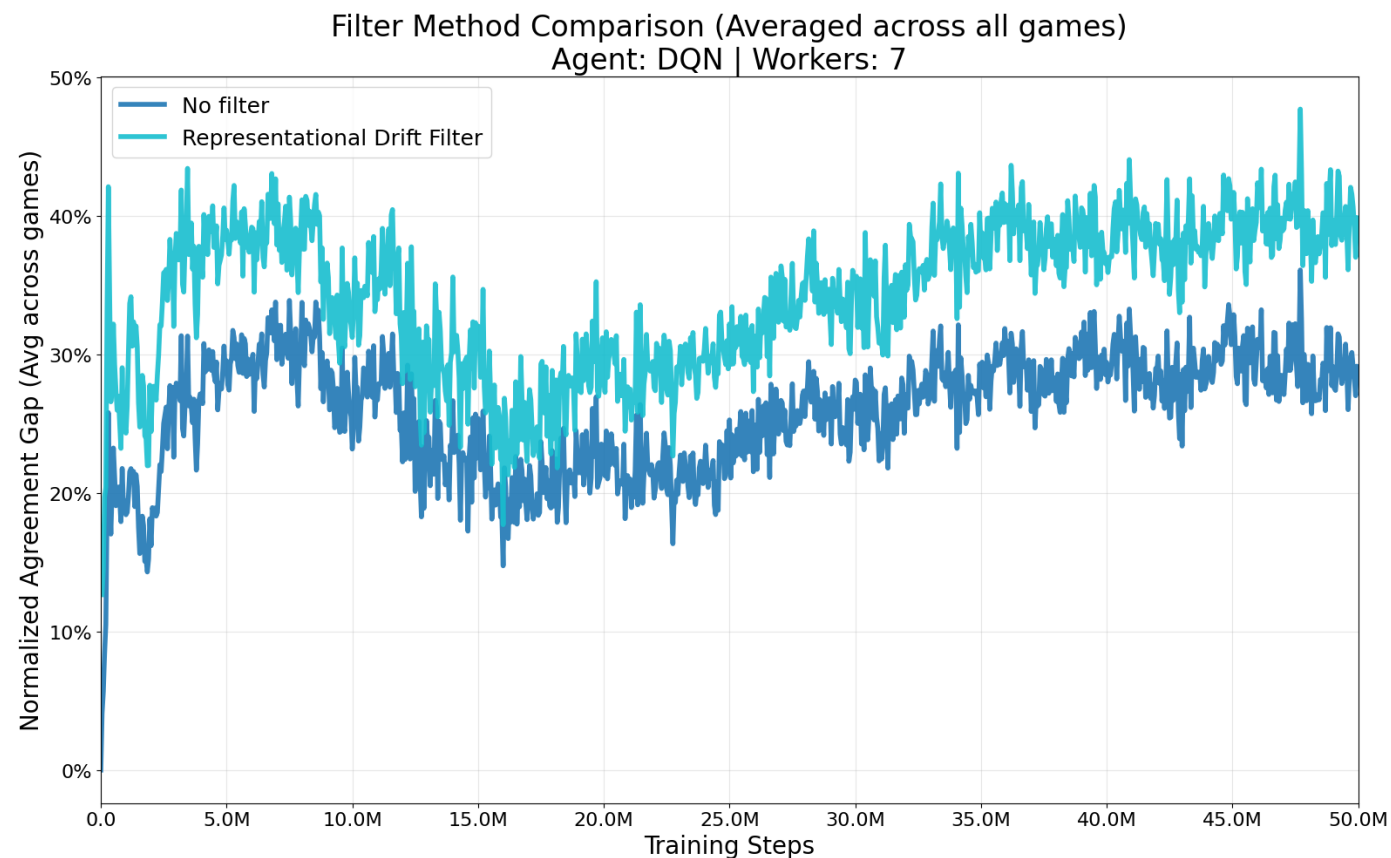
Battle Zone

- Unlike 2D games, *Battlezone* shows a significant "dip" in agreement around **15M steps**.
- The gap recovers in later stages, suggesting the agent eventually re-aligns with human judgement.



Representational Drift Filtering

- Human intuition is most valuable when focused on states where the agent is actively "changing its mind".
- This serves as an **existence proof**: while the baseline human signal is good, targeting consequential representational shifts significantly amplifies its value.



Takeaway

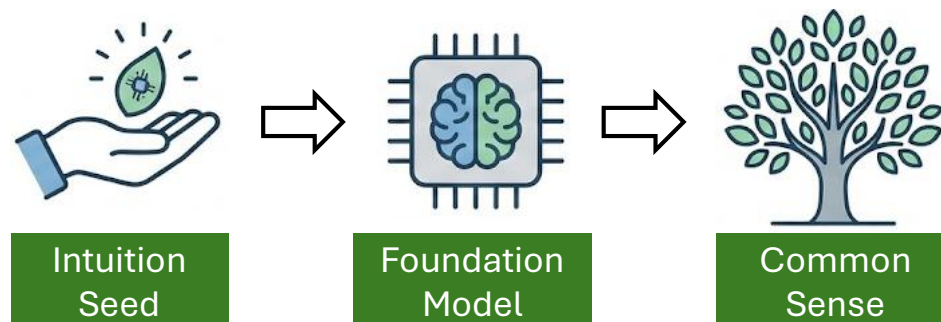
- Theoretical feasibility is established.
 - Simulation studies prove that **similarity-induced representations** accelerate RL training.
 - The method is robust to data scarcity and annotation noise.
- Human assessment is a valid signal.
 - Non-expert crowds provide a **~30% "head start"** in representational alignment compared to random initialization.
 - This signal is most effective for **early bootstrapping** (cold start) as well as states with **consequential stake**.

Contributions

- Part I: Human-aided Reward Design
 - **RECEPT**
 - Creates a Collective-View-Refine (CVR) workflow to elicit visual patterns from crowd workers.
 - Pattern-based Reward Generation (PRFG) produces scalable reward signals via pattern detection and tracking.
- Part II: Relevance-driven Efficient Learning
 - **Montage**
 - Retrospective Relevance
 - Uses Selective Memory Learning (SML) to filter policy updates to improve sample efficiency.
 - **CARAFE**
 - Prospective Relevance
 - Encodes functional similarity priors via crowdsourced triplet comparisons to accelerate downstream RL training.

Intuition Frontier

- Methods like RLVR prioritize automated verification for objective tasks (e.g., code, math) to minimize human bottlenecks.
 - This shift neglects subjective or ambiguous domains (e.g., social norms, trade-offs) where human "lived experience" is irreplaceable.
- **Intuition Scalers:** Foundation models should not replace human insight but scale it.
 - Humans provide the semantic "seed" (e.g., pattern), and an FM generalizes this prior across millions of unlabeled instances to achieve *common sense*.



Intuition Scaler



- Advanced pattern matching
 - Basic visual pattern -> “passing a school bus with stop sign extended is bad”
- Zero-shot Generalization
 - School bus -> Ambulance -> All emergency vehicles

Thank you!

Questions?