Adaptive Querying for Reward Learning from Human Feedback

APPENDIX

This appendix includes additional evaluations conducted in simulation, as well as provides a detailed description on the user study.

A. Additional Details on Simulated Environments

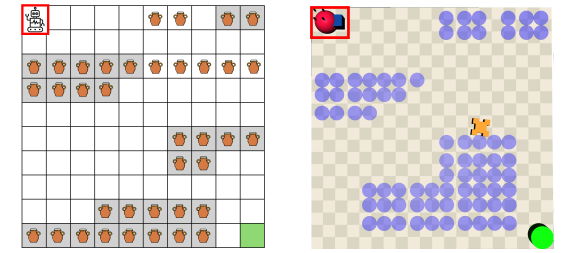
In addition to the Vase and Push domains, we evaluated our approach in a Navigation domain and on Atari Freeway domain. Fig. 1 shows an illustration of the four domains in simulation. The Navigation and Freeway domains are discussed in detail below.

a) Navigation: In this ROS-based city environment, the robot optimizes the shortest path to the goal location. A state is represented as $\langle x, y, f, p \rangle$, where, x and y are robot coordinates, f is the surface type (concrete or grass), and p indicates the presence of a puddle. The robot can move in all four directions and each costs $+1$. Navigating over grass damages the grass and is a mild NSE. Navigating over grass with puddles is a severe NSE. Features used for training are $\langle f, p \rangle$. Here, NSEs are unavoidable.

b) Atari Freeway: In the Atari game, the robot (a chicken) navigates ten cars moving at varying speeds to reach the destination quickly while avoiding being hit. Being hit by a car moves the robot back to its previous position, and is a severe NSE. A game state is defined by coordinates (x_1, y_1) and (x_2, y_2), i.e., the top left and bottom right corners of the robot and cars, extracted from the Atari-HEAD dataset [1]. Similar to [2], only car coordinates within a specific range of the robot are considered. The robot can move up, down or stay in place, with unit cost and deterministic transitions.

c) Feedback Formats: In the Vase, Push, and Navigation domains, we use the following feedback formats: Approval, Correction, Rank, Demo-Action Mismatch, Annotated Approval, and Annotated Correction. Each format provides different levels of detail about NSEs, influencing how the agent learns reward models and shaping its belief of NSEs accordingly. Fig. 3 shows the interaction for each format and learned reward values, using different feedback formats in isolation on the vase domain. For the Freeway domain, we use human demonstrations (explicit feedback) and gaze (implicit feedback) data. Note that we do not collect the human feedback data for the Freeway domain but instead use the data that is already available in [1].

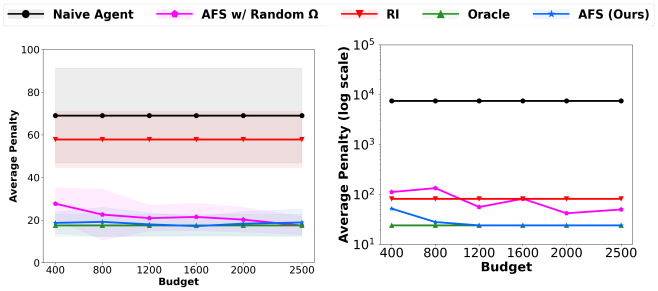
d) Results and Discussion: Fig. 2 shows the average NSE penalties when operating based on an NSE model learned using different querying approaches. Clusters for critical state selection were generated using KMeans clustering algorithm with $K = 3$ for the navigation domain and $K = 5$ for the Atari Freeway domain. Table I shows the average cost



(a) Vase: unavoidable NSE (b) Push: avoidable NSE

(c) Navigation: unavoidable NSE (d) Freeway: avoidable NSE

Fig. 1. Illustrations of the domains used for evaluation. The red box denotes the agent, and the goal location is marked in green.



(a) Navigation: Average Penalty (b) Freeway: Average Penalty

Fig. 2. Average penalty incurred.

for task completion. While the Naive Agent has a lower cost for task completion, it incurs the highest NSE penalty as it has no knowledge of R_N. RI causes more NSEs, especially when they are unavoidable, as its reward function does not fully model the penalties for mild and severe NSEs. Overall, the results show that our approach consistently mitigates avoidable and unavoidable NSEs, without affecting the task performance substantially.

B. User Study Details

The user study was conducted with approval from Oregon State University IRB. We recruited 12 graduate students who had completed at least one course in Reinforcement

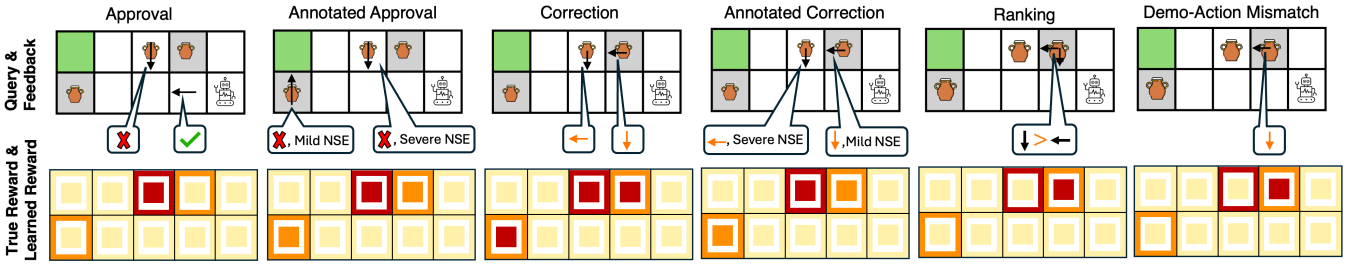


Fig. 3. Visualization of reward learned using different feedback types. **(Row 1)** Black arrows indicate queries, and feedback is in speech bubbles. **(Row 2)** ■ denotes high, ■ mild, and ■ zero penalty. Outer box is the true reward, and inner box shows the learned reward. Mismatches between the outer and inner box colors indicate incorrect learned model.

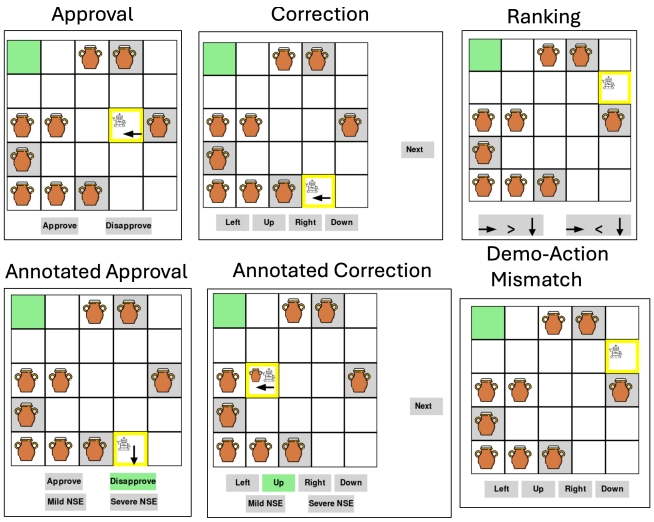


Fig. 4. User Study Interface: Participants provide feedback via button clicks, with available options varying by format, as shown in the figure.

TABLE I

AVG. COST AND STANDARD ERROR AT TASK COMPLETION.

Method	Navigation: unavoidable NSE	Freeway: avoidable NSE
Oracle	51.37 ± 2.69	3759.8 ± 0.00
Naive	36.11 ± 1.39	61661.0 ± 0.00
RI	40.10 ± 0.69	71716.6 ± 0.00
AFS (Ours)	49.18 ± 13.67	1726.5 ± 0.00

Learning. Participants were recruited through voluntary responses to a recruitment message sent to student groups. They were provided with an overview of the study and given the opportunity to review and ask questions about the consent form before providing their consent. The participants were compensated with $10 for their time.

a) Study Design: After introducing the domain and the agent's objective, users completed a tutorial where they interacted with the system by providing feedback in each of the six formats. Fig. 4 illustrates the study interface, with feedback buttons varying based on the format. This was followed by a calibration phase, during which the users' preference model was learned. Each user was prompted five times per format to provide feedback, with the option to respond or ignore, allowing them to express their interaction preferences. The probability of receiving feedback in a given format was determined by the fraction of prompts the user responded to, while the cost was based on their self-reported effort.

The study comprised three phases, each evaluating a different feedback selection approach—RI, AFS with Random Ω and AFS with our proposed method for critical state selection. To prevent bias, users were unaware of the approach used in each phase. After completing a phase, they were shown a trajectory of the agent's learned policy and asked to evaluate the approach used in that phase.

REFERENCES

[1] R. Zhang, C. Walshe, Z. Liu, L. Guan, K. Muller, J. Whritner, L. Zhang, M. Hayhoe, and D. Ballard, "Atari-head: Atari human eye-tracking and demonstration dataset," in *AAAI*, 2020.

[2] A. Saran, R. Zhang, E. S. Short, and S. Niekum, "Efficiently guiding imitation learning agents with human gaze," in *AAMAS*, 2021.