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. Evaluations on Domains in Simulation

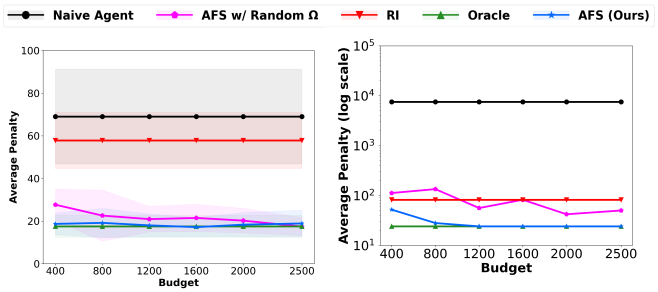
In addition to the Vase and Push domains, we evaluated our approach in a Navigation domain and on Atari Freeway domain. Fig. 2 shows an illustration of the four domains in simulation. The Navigation and Freeway domains are discussed in detail below.

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.

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.

While we use the six *explicit* feedback formats described in the paper, for the Freeway domain, we use human demonstrations (explicit feedback) and gaze (implicit feedback), available in the Atari-HEAD dataset [1].

Results and Discussion Fig. 1 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 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.



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

Fig. 1. Average penalty incurred.

TABLE I

AVG. COST AND STANDARD ERROR AT TASK COMPLETION.

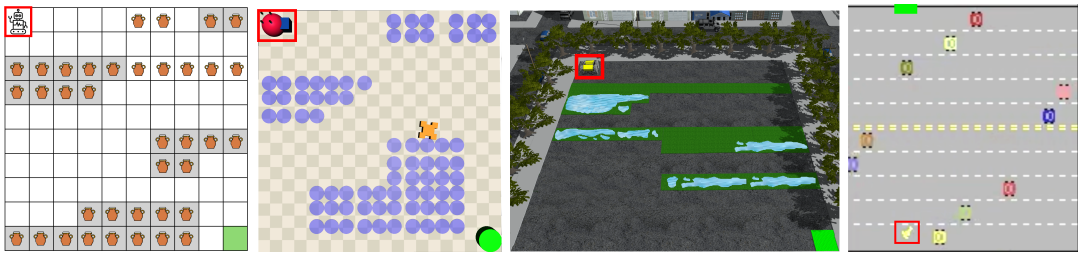
Method	Navigation: unavoidable NSE	Freeway: avoidable NSE
Oracle	51.37 ± 2.69	3759.8 ± 0.00
Naive	36.11 ± 1.39	61661.0 ± 0.00
RI	40.10 ± 0.69	71716.6 ± 0.00
AFS (Ours)	49.18 ± 13.67	1726.5 ± 0.00

B. User Study Details

The user study was conducted in accordance with IRB approval. We recruited 12 graduate students who had completed at least one course in Reinforcement 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 study was designed to take approximately 40 minutes to complete, and participants were compensated $10 for their time.

Study Design: After describing the domain and the objective of the agent in the domain, the users were required to take a tutorial, in which they interacted with the system by providing feedback in each of the six formats. Fig. 3 illustrates the interface used during the study. The available feedback buttons varied based on the feedback format. The calibration phase followed this, where the users' preference model was learned. Each user is prompted five times per format to provide feedback, with the option to respond or ignore, reflecting their interaction preference. The probability of receiving feedback in a given format is determined by the fraction of prompts the user responds to, while the cost is based on their self-reported effort.

The study took place in three phases. In each phase, the users were queried for feedback in one of the three approaches – RI, AFS with Random Ω and our approach.



(a) Vase: unavoidable NSE (b) Push: avoidable NSE (c) Navigation: unavoidable NSE (d) Freeway: avoidable NSE

Fig. 2. Illustrations of the domains used for evaluation. The red box denotes the agent, and the goal location is marked in green.

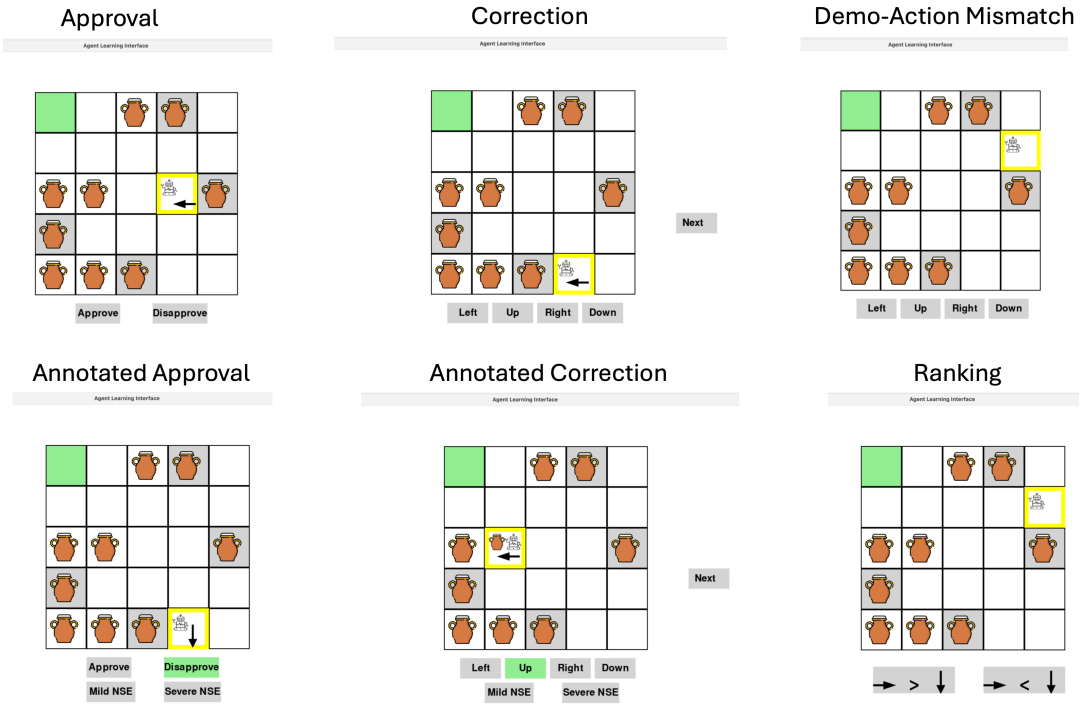


Fig. 3. User Study Interface: Participants provide feedback via button clicks, with available options varying by format, as shown in the figure.

To avoid bias towards any one approach, the users were not informed which approach was being used. After completing each phase, the users were shown a trajectory of the agent's learned policy. They also evaluated each phase by answering a questionnaire with the following questions,

1) Were the states in which the agent requested for feedback critical to its learning?
2) On a scale of 1 to 5, how intelligent do you think the agent's choice of feedback formats are?
3) Did the agent's performance improve at the end of the learning 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.