Thanks for the insightful feedback. Here we make some general clarification followed by individual responses.

(a) Probability of anomaly: In this paper, we report $1 - \eta_\pi^{S_t}([\bar{G}_t - \Delta, \bar{G}_t + \Delta])$ as the probability of anomaly, where $\eta_\pi^{S_t}$ is the probability density function of $\bar{G}_t$ and Δ is a hyperparameter. If we use a large Δ, the reported probability of anomaly before 10^4 steps will be close to 0, but then the algorithm becomes less sensitive to anomaly (e.g., if Δ is ∞, the output will always be 0). So Δ achieves a trade-off between reducing the false alarms (i.e., making the output as low as possible when no anomaly) and increasing sensitivity to the anomaly.

This is a simple but intuitive approach. A more formal approach requires properly defined priors over $\bar{G}_t$ and the occurrence of anomalies to make use of Bayes' rule. However, those priors depend heavily on the application and complicate the presentation of the central idea to conduct anomaly detection with reverse GVF. Therefore, we use the simple approach in our paper. We believe detecting anomaly using only a single observation based on a known p.d.f. itself is an interesting statistical problem that is out of the scope of this paper. Our contribution is to propose a (reverse) RL approach to efficiently learn the p.d.f. We will clarify all the above in the next revision.

(b) Figure 3c: In Figure 3c, the probability of anomaly is higher than 0.8. This is mainly due to the larger variance of the observed reward (compared with the toy MDP used in Figure 3b), resulting from the large stochasticity of the policy being followed. When the variance of a random variable is large, the probability mass is not concentrated. Consequently, the information that a single observation can provide is less. So our anomaly detection has a higher chance for a false alarm. If the variance of $\eta_\pi^{S_t}$ is large (intuitively, the curve of the p.d.f. is likely to be flatter), then $1 - \eta_\pi^{S_t}([\bar{G}_t - \Delta, \bar{G}_t + \Delta])$ will in general be large for all $\bar{G}_t$ due to the large randomness. We will clarify all the above in the next revision.

(R1) Legend: In Figure 3b, "2" means the reward becomes $R_t + 2$ from R_t after the anomaly occurs at 10^4 step. "0.9" means the probability of taking a_1 becomes 0.9 (from 0.1) after the anomaly occurs at 10^4 step. These are explained in Line 229 - 243 and we will clarify this further in the next revision. **Related work:** We will include more discussion about anomaly detection from the non-RL community in the next revision.

(R2) Baseline: Thanks for the insightful suggestion; we will include a comparison with IMPALA+GVF in the next revision. We will also move the comparison with IMPALA+PixelControl from the appendix to the main text. **Performance:** It is not entirely clear why some games are negative for Reverse GVF. Our initial conjecture is that it is related to the planning horizon. In the 10 tested games, Amidar seems to require the longest planning horizon. As Reverse GVF plans in the reverse direction, the representation for Reverse GVF may be less useful if we require longer planning horizon for the original problem. We will discuss this more in the next revision. **Assumption1:** As long as the problem we consider in the real world has a recurring structure, the ergodic assumption can usually be fulfilled, e.g., when we consider a bus commuting between two cities. As long as we set γ to 0 for any state, the inversion exists. E.g., if we are interested in the fuel of the bus, we could set $\gamma(\texttt{gas station}) = 0$. We will clarify this more in the next revision. **Metric:** Let A and B be two algorithms; we use R_A to denote the average undiscounted episodic return of the evaluation episodes at the end of training of the algorithm A. The improvement of A over B is computed as $\frac{R_A - R_B}{|R_B|}$. We will clarify this in the next revision.

(R3) Real-world data & other scenarios: Thanks for your insightful suggestion. This paper serves as the first work to introduce the reverse RL framework and establish its theoretical foundations. We, therefore, focus on providing motivating and easy-to-understand examples with synthetic data. Using real-world data may require extra engineering tricks that complicate the presentation of the central idea, which we therefore leave for future work. Reverse GVF can possibly be used for control as well if we learn $\bar{q}_\pi$ instead of $\bar{v}_\pi$, which we think deviates from the goal of this paper and we leave for future work. **Figure 3c:** Please see **(a) & (b)**. Moreover, we will plot the curves for different setups in Figure 3c separately in the next revision to improve readability. The overall message we want to convey is that our anomaly detection method is robust across all the tested setups.