We would like to thank the reviewers for appreciating our novel contributions on the algorithmic and theoretical front! We focus on clarifying our experimental results in this rebuttal.

[Why DM fails at ModelFail and SSD-IS achieve EXACTLY the same results as DM at ModelFail?].

ModelFail was first introduced by Thomas and Brunskill [2016] to show the failure of model-based approach in the MDPs with some partial observability. In ModelFail, the agent cannot tell the difference between any of the states except for s_1, but both DM and SSD-IS require full observability. From the point of view of both DM and SSD-IS, the actions have no impact on state transitions or rewards, so every policy has the same cumulative reward (equal to the the true cumulative reward of the behavior policy). A detailed discussion about why DM fails at ModelFail can be found in [Thomas and Brunskill, 2016, Section D.1]. MIS can handle partial observability by using observable states and the partial trajectories between them. Please refer Section 5.1 (line 258-262, there is a typo in Line 262, $\frac{\pi(a_{2\tau}^{(i)}|s_{2\tau}^{(i)})}{\mu(a_{2\tau}^{(i)}|s_{2\tau}^{(i)})}$ should be $\frac{\pi(a_{2\tau}^{(i)}|?)}{\mu(a_{2\tau}^{(i)}|?)}$, where symbol "?" stands for "unobserved", is an observed variable that the policy needs to react upon). Also see Section C (line 567-575) in the supplement for more details.

[Why MIS outperforms SSD-IS in time-invariant environments (including MountainCar) when n is large?].

The time-invariant ModelWin and MountainCar we used in the paper are finite-horizon *undiscounted* MDPs. Even though these environments have time-invariant transitions, the state marginal distributions at each t actually change with time and only converge to the stationary distribution as $t \to \infty$.

SSD-IS uses the stationary distribution ($t \to \infty$) to approximate that for all $t = 1, ..., H$ which is biased and not consistent even as the number of episodes $n \to \infty$. MIS, on the other hand, uses nearly unbiased and consistent estimators of the state marginals at every t. This allows MIS to outperform SSD-IS on Mountain Car when n gets large. We believe this is the reason and we will investigate it in details in our future work.

Reviewer #1

["A specific baseline I would really like to see the authors add is the PDIS (per-decision IS) and CWPDIS (consistent weighted per-decision IS)."]

The IS and WIS in the experiments are step-wise, which are essentially PDIS and CWPDIS. The detailed explanation is in Section 3 and Section C.

["Why does it (SSD-IS) achieve . . . perform as well as MIS for mountain car but eventually stops improving?"]

Please check the answers at the beginning.

["If p_t is sampled uniformly at each time step, isn't . . . setting equivalent to a time-invariant MDP with $p = 3.5$?"]

Sorry for the confusion. Note that each transition probability p_t is only sampled before the experiments and fixed during the experiments for all episodes. We will clarify it in the final version.

Reviewer #2

Thanks for supporting our paper. We are planning to extend our approach to large-scale environments with extensive function approximation.

Reviewer #3

["In Figure 2 and 3, why DM and SSD-IS method works well in ModelWin but perform very bad at ModelFail?"]
["For me it is surprised in time-invariant environment SSD-IS method perform worse than MIS method."]

Please check the answers at the beginning.

["In Figure 3 (b) and (d), why the curve is not smooth even after 128 repetition?"]

Note that the Y-axis is relative MSE, which is normalized by the true cumulative reward. In this time-varying MDP (Figure 3), the true cumulative reward is related to the transition probabilities p_t at each time step. We sample each p_t before the experiments and then fix them during the experiments, so the true cumulative reward is a non-smooth function of H and the figures with increasing H should not be smooth. In the time-invariant MDP (Figure 2), the true cumulative reward is a smooth function of H and the corresponding figures are smooth.

References

Thomas, P. and Brunskill, E. (2016). Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139–2148.