We thank the reviewers for thoroughly commenting on our article; their comments give us the opportunity to improve the manuscript and clarify all the unclear points. We have performed the additional requested experiments: 1) we have replaced the W-MSE representation with Random Features, keeping the RNN model; 2) we have replaced RNN with a feed-forward network, keeping the W-MSE representation. For Montezuma's Revenge, the average prediction error is significantly higher in both cases and the final score does not exceed 400, which matches the RND results from Table 2. We noticed that both components work collaboratively: while the representation is optimised to preserve temporal consistency, the sequential prediction approximates well the trajectory of the agent. These results will be included in the revised version.

We did not include the recent methods as baselines for the Partially Observable Labyrinth experiment because they were mostly designed for high-dimensional observations and do not provide benefits for this environment compared to the theoretically justified methods for the tabular setting. At the same time, to the best of our knowledge, popular classical methods for exploration do not assume partial observability and are not applicable to this environment. We experimented with the MBIE-EB method [30 in manuscript] and we observed a degradation in performance (-55.7 for size 3x3, -1000 for 4x4). Replacing the RNN world model with a feed-forward network produces an even more dramatic decline to -969 for size 3x3. In this case, the irrelevant intrinsic reward completely obscures the target goal.

The world model prediction error can be a proxy on how much information the belief state is missing about the predicted future step. The less information is available about this step, the more uncertain the model and the higher the error. During experiments with Partially Observable Labyrinth, we noticed that the world model tends to output mean values of the state when the observation cannot be predicted; this can be an indicator of uncertainty. However, as noted by R4, in general we cannot guarantee that the prediction error is a measure of uncertainty. More advanced methods, e.g. based on ensembles [1], can provide the estimation and can be a good extension of our method. To be precise, we will replace "uncertainty" with "missing information" in lines 37-40 and all other places.

For an intuition about W-MSE representation and stochasticity, let's consider the noisy TV experiment: there is a TV in the environment, an agent can switch channels, but it always shows random images or noise. Observing it, most of the curiosity methods will produce an harmful high intrinsic reward, this effect being known as the "couch-potato" issue [26 in manuscript]. In our case, the W-MSE loss pushes the representations of neighbour frames to be as similar as possible, thus the representations of random images of the TV will converge to the mean and will be easily predictable by the world model, avoiding the described issue.

As suggested, for future work we plan to check the scalability of the method and the asymptotic performance in Atari and compare it with the best-performing methods such as NGU. Furthermore, we plan to run experiments in 3D maze-like environments (e.g. as in [26 in manuscript]), which should be suitable for the method. Additionally, we plan to run the described noisy TV experiment. Predictive State Representations is an interesting, theoretically grounded alternative to the world models; we will include this missing citation. Our approach has conceptual similarities with Slow Feature Analysis; in fact, the recently presented gradient-based version of the method [2] can be a good alternative to the W-MSE loss; we will include this citation as well.

To show how the seed affects the performance we included Fig. 1 with training dynamics in the supplementary. However, as was accurately spotted out by R3, there is a mistake in the Montezuma's Revenge plot, i.e., it's a bug of the plotting script. We examined each run separately, the 2500 score was reached first time at 24.5M, 43.2M, 33.6M, 28.2M, 41M frame for each seed respectively; the plot will be updated in the final version. The encoder is updated during the whole training; empirically, the representation adapts to the changing distribution of observations. Pretraining steps are included in the training budget, these observations are used to train all components including DQN; we will add these details in the Method section. We will remove the "information bottleneck" phrase on line 159, as it can be misleading. Figure 3 shows the trajectory created from the user input ("emb_vis" script); it is used only for demonstration, while the representation is trained only on random transitions, it will be noted in the final version. We will elaborate on line 115. We will rephrase the line 99, as we refer to the specific model suitable for the VIME method [16 in manuscript]. We will refer to the agent as "it" on line 77 and other places.

Additionally, we would like to update the "Broader Impact" section in the revised version of the manuscript with: "The presented work is a research in the field of reinforcement learning, focusing on the problem of exploration in real-world conditions (image-based observations, partial observability). Such algorithms can help searching for new important information, for non-trivial solutions. These algorithms can be the crucial component for the development of autonomous intelligent systems for solving complex tasks. Such systems can be used in many different fields, having both strong positive and negative impacts on society, and should be treated with care."

[1] Balaji Lakshminarayanan, Alexander Pritzel, Charles Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles, 2016. [2] Merlin Schüler, Hlynur Davíð Hlynsson, Laurenz Wiskott. Gradient-based Training of Slow Feature Analysis by Differentiable Approximate Whitening, 2018.