We would like to thank all reviewers for their insightful comments.

Reviewer 1

• **"I wonder if there is really novel technique or idea in the paper. (...) It is natural to choose optimistic representation model (...)"**

We agree that our solution is simple and natural. The fact that previous approaches have used more complicated algorithms to obtain worse bounds should be an indication that algorithm design and analysis are not obvious, however.

• **"(...) I am not sure the significance of the result. For example, when solving RL problems such as Atari games, we may test different representation methods. If one of them approximately satisfy Markov property and achieve good performance, then we use this representation and the problem is solved."**

In the mentioned scenario it has to be specified how to efficiently test (a) the (approximate) Markov property and (b) good performance for each representation. (Note that for (b) you still have to find a good/optimal policy under each representation.) If you have a simulator and you don't care about efficiency and cost of exploration (i.e., reward), you can test (i.e., try to learn the optimal policy) using multiple models and then cross-validate the best. But if you care about efficiency and overall performance, this is just not possible. Our approach shows that (a) is not necessary, while for (b) we offer an explicit solution for the arising exploration-exploitation problem (for simultaneously choosing a model and a policy) that is also efficient as the regret bound shows.

• **"Are there cases where we need to (...) learn the best representation methods in a finite representation set?"**

In many applications several sensory measures are available but it is not clear what is the most suited representation of the system. Experts often have ideas about "reasonable" models and/or features combinations. One can also think of using representations that worked well in other similar problem settings. The proposed algorithm may leverage this information and quickly discard models that are not suited for the specific problem. Clearly, this is not the most generic case but it is a first step toward having an efficient algorithm for selecting the representation.

• **"Besides, the regret can scale linearly in $\sqrt{|\Phi|}$ since S_Σ scales linearly in $\sqrt{|\Phi|}$ in worst case, which means the algorithm cannot perform well for large or even infinite representation set."**

While performance bounds will generally depend on the model space Φ, it is an open problem whether the $\sqrt{|\Phi|}$-dependence can be improved in the considered setting, cf. also our response to Reviewer 2. Still, there are techniques that still work for infinite representation sets (see e.g. reference [6]) that will work for our approach as well.

• **"(...) there are many algorithms such as UCBVI which enjoys square root dependence on S. Is it possible to use the framework and techniques of UCBVI to improve the dependence of the number of states?**

Please note that the regret bound for UCBVI has been derived for the simpler *episodic* setting. For the average reward setting, it is still an open question whether $\sqrt{S}$-bounds are achievable. Our approach can be adapted to the episodic case when the regret bounds would benefit from the improved bounds available in this setting. As discussed in the paper, any optimistic algorithm with improved bounds could be used within our framework to obtain better bounds.

• **"The paper could be more clear if lemma 3 was proved in appendix."**

An explicit derivation of Lemma 3 would have taken a few pages mostly repeating material from [9], so we decided not to include it in the paper. However, we will think about whether it is possible to give more details about the proof without the need to copy content from [9].

Reviewer 2

• **"Can UCB-MS ensure the final model to be a Markov model? If not, does the result still make sense?"**

An important point of the paper is that for obtaining regret bounds in the online setting it is actually not necessary (and in some cases not even possible) to identify the true (Markov) model. As long as a non-Markov model gives at least the same reward that would be expected from a Markov model there is no need to discard it. Such a model could be e.g. a good (non-Markovian) approximation. The regret is always measured with respect to the true Markov model however.

• **"Is there any discussion about the lower bound for this task? (...) I think the regret bound of UCB-MS should match lower bounds with respect to T. How about other parameters like S, A or Φ?"**

The $\sqrt{A}$-dependence is optimal as for UCRL2, while the optimal dependence on S is still an open question (also for the MDP case). The optimal dependence on $|\Phi|$ in our setting is also open. The closest result we know is for aggregation techniques with *full* information where it is possible to obtain $\log(|\Phi|)$. Obviously, in our setting we have less information and it is not clear if it is possible to obtain logarithmic dependence.

Reviewer 3

Thank you very much for the positive feedback!