Thank you very much for the thoughtful reviews!

All reviewers: All reviewers suggested ways to improve our treatment of the related work. If the paper is accepted we will use the additional content page to address this. We will move (a revised version of) Sec. A.3 from the appendix to the main paper and make sure the connections with the literature are explored in more depth (in particular with respect to VAML, as suggested by **R3** and **R4**, and several notions of MDP homomorphism, as suggested by **R1** and **R2**).

[**R1**]: We want to emphasize that our work is about model learning, *not* about state representation. In particular, we assume (except in the discussion of related work) the existence of a state signal s that is neither learned nor modified by VE. The VE loss is used exclusively to learn a model from state s to state s', but the notion of state itself never changes. It seems plausible that the model learned by VE induces a space of "compatible" state representations, but this is yet to be analyzed. We made a deliberate choice to isolate the representation learning aspect, both in the theory and experiments, to have as clean an analysis as possible. We believe that this misunderstanding may be the source of **R1**'s concerns. In light of this clarification, we kindly ask the reviewer to reconsider their assessment of our paper.

We argue that a tabular representation is the most appropriate setup to understand a new approach to model-learning. In particular, tabular experiments help to illustrate the differences between VE and MLE, since in this case it is trivial to define a distribution model with the appropriate capacity (a $n \times n$ transition matrix) and to limit its capacity in a meaningful way (through the matrix's rank). We also scale up to non-tabular experiments (Figs. 4c and 4d) and consider the relationship to existing work that uses VE in combination with deep learning (Sec. 6). The choice of MLE as our baseline also arises naturally, as this is by far the criterion most commonly used in conventional model learning (see supplement of [1]). Also, the "trivial-embedding" problem cannot be the explanation for the poor performance of MLE, since there is no embedding being learned (states s are fixed; only the transition function $f(s, a) = s'$ is learned). We believe other concerns raised by **R1** can be similarly resolved by noting that VE is not about representation learning.

Specific points: (**Q^π-irrelevance):** see response to **R2** on homomorphisms. **(L131/138):** We will use "functions" throughout, thanks for pointing that out! **(L197 & L29):** We will include references to standard RL texts that corroborate these claims. **(L201):** MLE generates probability distributions under which the observed data is most probable [2]; the fact that this is desirable does not mean MLE will always be the best strategy. **(L274):** The statement does not refer to the linear approximators themselves, but rather to models (linear or otherwise) that are VE with respect to them (see L277). **(Eq 7):** The loss in Eq. 7 is over (possibly very small) subsets of all possible functions and policies; this is exactly how VE is used to restrict the space of models and one of the main points of the paper. **(Policies):** We consider stochastic policies throughout, but describe after Proposition 1 how $|\mathcal{A}|$ deterministic policies can cover this space. **(L12/42):** It refers to adding functions / policies to the sets we are defining VE wrt. **(Property 3):** Adding more policies and functions to Π and $\mathcal{V}$ further constrains the set of models that are VE to the true model. **(Hamel):** see comment to **R3**.

[**R2**]: If the paper is accepted, we will use the extra space to improve the discussion on related work. We now describe 3 concrete modifications in this direction resulting from **R2**'s comments. **(Sec. 6)**: We will revise Sec. 6 prioritising clarity over breath, to make sure the main text is self-contained. **(Homomorphism)**: We will elaborate on the connection between VE and MDP homomorphisms, which we briefly touched upon in Sec. A.3. Note that any notion of equivalence over states (*e.g.*, Q^π-irrelevance, suggested by **R1**) can be recast as a form of state aggregation; in this case the functions mapping states to clusters can (and probably should) be used to enforce VE. But VE is more general than that: it applies to any representation (not only aggregation) and can be used to explore structure in the problem even when there is no clear notion of state abstraction (Sec. A.1.2). We will add this discussion. **(Linear models)**: The relation of VE with Sutton *et al.*'s and Parr *et al.*'s results is another interesting connection. As **R2** notes, VE is more general, since it also applies to nonlinear models (even when a linear approximator is used—L277). Re-deriving these results from VE's perspective is an intriguing idea; we will try to do so and add any eventual insights to the appendix.

[**R3**]: We will extend the discussion on VAML and move it to the main paper. (**Clarification of $\mathcal{P}(\Pi, \mathcal{V})$**): We use $\mathcal{P}$ to refer to the set of all transition kernels, and define $\mathcal{P}(\Pi, \mathcal{V})$ as the set of all such kernels that are value equivalent to the environment wrt Π an $\mathcal{V}$ when the reward is assumed to be known. Upon re-examination, it appears that we do not explicitly state this until the proof of Proposition 2 in the appendix. We will spell out this definition clearly in the main text in the subsequent version of the paper. (**Hamel**): Hamel dimension, matching the intuitive notion of dimensionality, describes the number of coordinates necessary to specify every point in a given vector-space. We will include a definition of this term in the text to improve readability.

[**R4**]: (**Practicality**): This is a valid point. However, note that we empirically observed that high quality value equivalent models can be produced without requiring prohibitively many value functions. We intend to provide theoretical support for these observations in future work. (**Property 2**): Good question! $\mathcal{M}(\Pi, \mathcal{V})$ is the set of models in class $\mathcal{M}$ that are value-equivalent to m^*. When we consider all policies and values, there is only one possible value equivalent model: m^* itself. Thus, if $m^* \notin \mathcal{M}$ then $m^* \notin \mathcal{M}(\mathbb{\Pi}, \mathbb{V})$. (**Related Work**): Thanks for pointing these out! We will include both of these papers in the next version of our related work section.

[1] Farahmand et al., 2017. Value-Aware Loss Function for Model-based Reinforcement Learning. [2] Millar, 2011. Maximum Likelihood Estimation and Inference.