We thank the Reviewers for their constructive comments that helped improve the paper. We address their comments and questions below, and modified the manuscript accordingly.

Reviewer 3 raised the question of the generalization to more complex agent morphologies, which was not addressed in the submission. Following their suggestion, we ran additional experiments with higher dimensional motor spaces (6D) and more complex forward models: i) involving non-linear functions and random linear combinations of the motor command (discrete world), and ii) a 4-segment arm with 4 hinge and 2 translational joints (arm in a room). The new results are qualitatively equivalent to the ones presented in the submission, which indicate that the capture of spatial invariants is robust to the complexity of the agent. Due to compute limitations, only preliminary results are currently available and have been added in the Appendix. They will be updated as soon as all simulations have finished running. (Remark: for the new arm, we collected more samples and increased the size of Net_{enc} to be able to approximate the more complex forward mapping.)

Following the suggestions of Reviewers 3 and 4, we extended the description of the problem setup to better ground the notations and concepts. We also added diagrams in the Appendix to illustrate i) the connection between our generic formalism and the simulations, and ii) the spatial invariants, that were previously only described mathematically.

Following Reviewer 4's recommendation, we added a simple and concise description of the paper's aim at the end of the Related work section: *This work is in line with the theoretical developments of [...], which address the fundamental problem of space perception in the framework of the SMCT, but frame them in an unsupervised machine learning framework. We show that the structure of space can get naturally captured as a by-product of sensorimotor prediction.*

Regarding the additional comments of Reviewers 1 and 4, it might be important to point out that our perspective significantly differs from the bulk of literature on spatial representation. Our aim is not to solve a task or to compete over performance measures but, more fundamentally, to study how the concept of *space* can emerge in a naive agent. Thus a direct performance comparison with other methods/tools (self-organizing maps, slow feature analysis...) is of limited interest. (Additionally, depth estimation is a problem of a different nature than the one addressed in this work). We can however say that most approaches to this problem build allocentric representations, based primarily on sensory information (with occasional auxiliary interactions with motor information), and using a priori defined constraints to shape the representation. These representations are build either to solve a task, in which case their structure is of secondary importance, or enforced to be spatial-like via prior constraints. As a consequence, their structure at most captures the topology of a sensory manifold (equivalent to the topology of space for a single environmental state only) or one imposed from the motor space via additional prior constraints. In contrast, we study the conditions for the emergence of an egocentric spatial representation, based on the fundamental study of how an external space induces invariants in an agent's sensorimotor experience (originating from the concept of *compensability* introduced by H.Poincaré). This means describing how the very structure of space shapes sensorimotor experiences, and how it impacts sensorimotor prediction, in the absence of any specific task or extraneous constraint. This way, the approach can be seen as in agreement with the SMCT and Predictive Coding. Moreover, the nature of the spatial invariants we identified means that our representation is grounded in the motor space, abstracted from the specific content of the environment, and captures the metric structure of space; a result that no other non-supervised method achieves.
Due to its originality, only few previous works align directly with this line of research. They are all mentioned in the introduction and related work of the original submission (refs 10, 23-27, 35, 36, 38). This work is the first to frame their theoretical considerations about spatial invariants into a self-supervised machine learning framework. We updated the end of the introduction to better emphasize the nature of this contribution, and significantly extended the related work section to better explain how our approach differs from the cited works.

We include below a (tiny) overview of the five diagrams and figures that were added to the paper:

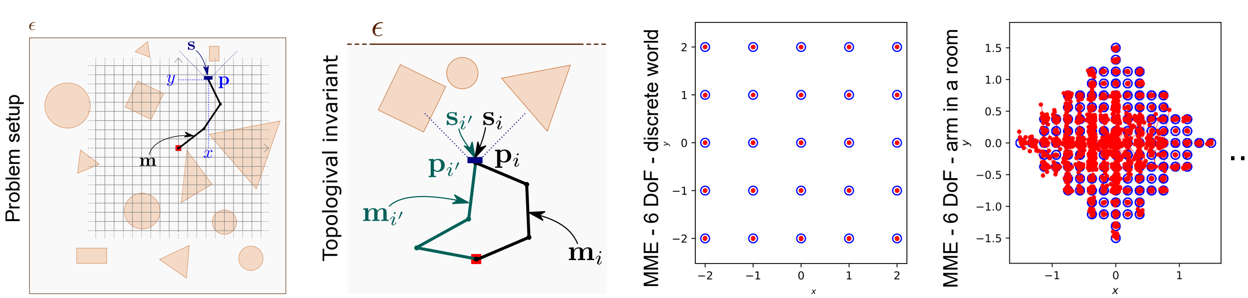

Figure 1: From left to right: Partial view of the problem setup diagram; Partial view of the invariants diagrams; MME results for the more complex discrete agent; MME results for the more complex arm in a room.