We thank all reviewers for their comments and for their time in reviewing our work. We appreciate the reviewers' assessment of the strengths of the paper, with all reviewers highlighting our contribution towards putting into practice the symmetry-based notion of disentanglement. We address their questions and issues in the following.

Limitations and scalability *(we will amend our paper to discuss the following point)*

Reviewers 2, 3, and 5 note that the discussion of our work's limitations could be improved. Although $\mathrm{SO}(n)$ readily occurs in many systems of interest, there are indeed environments where this will not be the case — for example, if objects appear and disappear, or split apart and merge together. With regards to scalability, there is a polynomial scaling associated with the $n(n-1)/2$ rotations that describe a transformation in an n-dimensional spherical latent space. Whilst each of these $n \times n$ rotation matrices is relatively sparse (with 4 parameterised elements as per eq. (2) of the manuscript), ultimately it is unavoidable that computational cost increases with system size.

Related Work *(the following discussions, and the papers mentioned by reviewers 2 and 3, will be added to our work)*

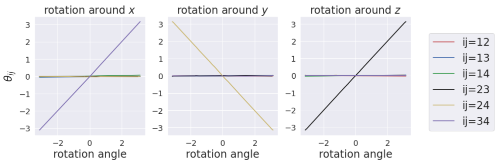
Reviewer 4 enquires about the relation between our work and Connor and Rozell (AAAI Feb. 2020), which is indeed insufficiently described in our paper. Their work requires first learning a structured latent space with an auto-encoder, which, as we show in our figure 2, requires neighbouring states to overlap. In their experiments, to satisfy this condition, differences between original states and transformed states had to be made quite small. We speculate that their method requires this first step because their use of general matrices (instead of $\mathrm{SO}(n)$) to represent transformations may not enforce a sufficiently strong prior on the type of transformations operating on their system. Another important difference between their work and ours is that their sparsity regularisation aims to reduce the number of transport operators that describe the system and not enforce disentanglement within each operator as in our method.

Reviewer 2 has some questions concerning the connections between our work and both static and dynamical VAEs, which we indeed insufficiently explain. Our work and SSMs use some similar tools, and interesting representations can indeed be found in both. However, there are fundamental differences that lead to us treating them separately and not using SSMs as experimental baselines. First and foremost, it is not straightforward to reconcile the SSM objective of modelling probabilistic generative processes with the inherently deterministic framework of Higgins et al (which is why we do not use a VAE). Our method is the first to propose a path to disentanglement (in the sense of Higgins) in either deterministic or probabilistic frameworks. As such, we feel that attempting to bridge the gap between symmetry-based representation learning and SSMs is certainly an exciting research direction but is outside the scope of our current work.

Reviewer 3 asks a question about the limitations of the work of Caselles-Dupré, which we indeed insufficiently describe. Their method only applies to transformations that are identity except for a single unknown 2x2 block on the diagonal (Sec 6.2 of their paper); their formalism therefore applies to the gridworld experiment, where they do find similar representations to ours, but not to the other environments considered in our work. Importantly, their framework can only work when the type of symmetries are 1) a priori known and 2) described by this type of matrix.

Experiments *(results to be added to the appendix)*

Reviewers 3 and 5 have questions about experimental results when we increase the number of latent space dimensions. For the Lie group experiment with a latent space dimension of 4, we do find that the additional dimension is ignored (see figure). Similar results are found for a latent space dimension of 6 for lines 239-241.



Reviewer 3 is correct that the "direct prediction" baseline in Sec 5.5 could be replaced by a stronger baseline. However, the intent of this experiment is simply to demonstrate that the representations learnt by our framework do indeed exhibit desirable properties (beyond interpretability) that are often associated with disentanglement. The chosen baselines are then intended as direct ablations of our regularisation, and (in the case of "direct prediction") our prior assumptions on the environmental structure. With regards to reviewer 3's suggestion that we present a more complex disentanglement task, we did apply our framework to the established datasets of 3D Cars (Kim and Mnih, "Disentangling by factorising", 2019) and 3D Shapes (Reed et al. "Deep visual analogy-making", 2015). We ultimately decided that the experiments in the main text sufficiently presented our contribution, but would be happy to include these results in the appendix.

Clarifications *(the following clarifications will be added to our paper)*

To address Reviewer 4's specific queries: 1) The encoder, decoder and thetas are all learned within the same loop. 2) We will fully describe all network architectures in supplementary Sec 3.3. 3) Each symbol/colour pair corresponds to a unique state of the environment. This mapping is consistent for all latent spaces. 4) In Sec 5.2 and 5.3, independent sets of parameters (thetas) are learnt for each discrete environmental actions. In Sec 5.4, this look-up-table is replaced by a network, ρ_σ, that maps (continuous) environmental actions to rotations in the latent space. 5) The same training protocol as for the other experiments was used here too; further information will be added to the appendix.