We thank all three reviewers for their helpful comments, which we attempt to answer hereafter.

Theory. About Reviewer #1's first comment on actions and symmetries, we are interested in practical ways to learn SBD representations, and make an analogy/parallel between the effect of a symmetry g (by the group action $\cdot_{\mathcal{W}}$) on the environment $(o_1, g, g \cdot_{\mathcal{W}} o_1 = o_2)$, and a transition that uses the dynamics f of the environment $(o_t, a_t, f(o_t, a_t) = o_{t+1})$. It allows us to consider an embodied scenario, where symmetries are applied via group actions to an agent evolving in an environment. In our analogy we simply state that $o_1 = o_t$, $o_2 = o_{t+1}$ and $a_t = g$ and $\cdot_{\mathcal{W}} = f$. However, we do not consider that symmetries and actions are always the same. A symmetry is an element of a group (in the mathematical sense) of functions $g : W \rightarrow W$, and the binary operation of the group is composition. In that sense, these functions can effectively be considered as actions, because actions take the environment from one state to another through the dynamics f, and symmetries take the environment from one state to another through the group action $\cdot_{\mathcal{W}}$. However it is important not to say that all actions are symmetries, for instance the action of eating a collectible item in the environment is not part of any group of symmetries of the environment because it might be irreversible for instance. This point is important to clarify in the paper, and we updated the paper in the beginning of Sec.3 with this explanation.

We thank Reviewer #1 for noticing the erroneous notation in claim 1 of Theorem 1. We clarify the notation: $\mathcal{W} = (W, \cdot_{\mathcal{W}})$ is a world, which comprises a set of states $W = (w_1, .., w_m)$, where each state w_i is a d-dimensional vector, and a group action $\cdot_{\mathcal{W}}$ w.r.t a group G. In our example, w_i is the position of the agent $(x, y) \in \mathbb{R}^2$. **In claim 1 of Theorem 1 we consider W_k to be the set of possible values for the k^{th} dimension of states** $w \in W$, e.g. all the possible values of x would be W_1 for the aforementioned example. We hope this clarifies Theorem 1, which in claim 1 establishes a lower bound, as a function of the cardinalities of $(W_1, .., W_m)$, of the number of possibilities of how the group action $\cdot_{\mathcal{W}}$ can be applied to W. All these possibilities form a set of possible worlds $(\mathcal{W}_1, .., \mathcal{W}_{k_{W,G}})$ that have the same W but different group actions. We added the definition of W_k in the formulation of Theorem 1, and fixed claim 1. We also added a notation clarification in Sec.2, in order to introduce the useful notations earlier in the paper.

"Using a training set T of still images" refers to using only unordered observations for training, i.e. for learning a representation model. We clarified this in Theorem 1.

As suggested by Reviewers #1 and #3, Theorem 2 has been moved to the appendix. In the main text we still keep Sec.5 and mention the result in order to motivate the experiments that follows. The additional freed space has been used by the clarifications and additional experiments of the rebuttal.

Experiments. Following Reviewer's #3 suggestion, we added hyperparameter and architecture details for our experiments. Note that we did not include them at first because our experiments did not depend on hyperparameter and architecture search, as we used standard choices. Our code is also provided in the Google colab.

Regarding a more general approach for the Forward-VAE architecture (Reviewer #3), we indeed explicitly design the model such that the resulting representation is Linear SB-disentangled, because we enforce linearity, force the representation to be SB (see points 1 and 2 in the definition in Sec.3 and by design have two separate subspaces for each symmetry. A more general approach would have been not to have those two separated subspaces and learn the entire action matrices, and thus we won't have the guarantee that the representation will satisfy the disentangled property. We ran this additional experiment and obtained the expected result: the learned representation is Linear-SB but not disentangled. This means that the x and y coordinates are not properly disentangled w.r.t to the considered group decomposition (i.e. a latent traversal over each dimension would not result in only a movement of the agent along the x or y coordinate). Still, the learned action matrices are able to describe how the symmetries affect the representation and in a linear way. Enforcing disentanglement is the only viable option we found for LSB-disentanglement with this architecture. We added this additional experiment and conclusion in the remarks section (Sec.6.3) about this experiment.

The instability (line 244) mentioned by Reviewer #3 can be expected during training due to the different contributions of the loss: at each training steps the goal of the forward part of the loss is to have a latent space that is suited for predicting z_{t+1} using z_t. The rest of the loss is the VAE, which tries to learn a latent space that allows reconstruction. Hence we considered the balance between these two seemingly unrelated objectives as a source of instability. However it worked in practice, without any reweighting of the objectives, which was a surprise. We rephrased the sentence.

Related works. It is indeed true that the paper is light on related work. Our initial intent was only to extend the work of Higgins et al., and so we omitted most of the related work in the paper, and pointed to the work of Higgins et al. for context. However, as suggested by Reviewer #3, better positioning the paper in the disentangled representation learning field might prove helpful for readers. The works mentioned by Reviewer #3, which we are aware of and agree are relevant in our paper, are now included in a paragraph in the Discussion section. This way, it is easier to map our paper in the current state of the art of the active field of disentangled representation learning. Finally, we updated all references that pointed to the arxiv version of the paper rather than the peer-reviewed one.