We thank the reviewers for their positive comments and helpful feedback. Following these suggestions, we have made improvements to the clarity of the methods and experiments. We will also include an updated Figure 1 illustrating the full VAE framework in the next draft of the manuscript. We respond to the specific comments of the reviewers below.

Reviewer #1 We thank the reviewer for their positive comments and interest in our work.

Reviewer #2 We appreciate the reviewer's feedback and have made several changes to the manuscript to address the reviewer's concerns. We have improved the methods to provide additional intuition and implementation details.

1. Any imaging domain in which rotation and translation occur as nuisance variables will benefit from this approach. Single particle electron microscopy is a high impact application domain with exactly this problem. Understanding continuous variability in proteins imaged with EM is a pressing problem for which our method provides the first solution framework.

2. The MLP in figure 1 is specifically the generative network component of the VAE. The inference network is structured in the standard manner for fully connected inference networks. We will change Figure 1 to illustrate the full VAE framework and better illustrate the generative network.

3. We have updated the methods section to give more intuition and improve the description of the method. We will also release the source code with the camera-ready version of the manuscript, which will provide all implementation details.

4. Empirically, the Gaussian approximation works well, but we plan to explore other distributions in the future.

5. One network was trained for each dataset. We now clarify this in the text. These results are robust to the choice of prior values. We show in Appendix Figures 1 & 2 that spatial-VAEs trained with wide priors on these parameters learn the same manifold over digits and reach the same reconstruction error as the models with correctly matched priors. For the dimension of the unstructured latent variables, these settings represent a reasonable trade off between interpretability/compression and representation power. It is not surprising that with large z dimension the ELBOs become similar, as eventually there is enough capacity in z to represent both the content and the rotation and translation. However, the standard VAEs do not disentangle pose from content.

6. The purpose of these figures is to illustrate that the spatial-VAE successfully learns disentangled representations on real datasets. In Figure 4, the spatial-VAE, but not the standard VAE, recovers the ground truth variability in the dataset. As additional quantitative support for this claim, we now report the ELBOs for each model in the main text (was Appendix Figure 3) and also include a quantitative assessment of the ability of the spatial-VAE to recover the ground truth variability. We report the correlation coefficient of the mean of the approximate posterior for each latent variable with the known conformation and rotations of each image (Table 1). The latent variables learned by the standard VAEs do not separate into the ground truth conformation and rotation variables whereas the spatial-VAE latent variables correlated well with these features.

Model	Variable	Conformation	Rotation
vanilla-VAE [Z-D=1]	z_1	0.00	0.18
vanilla-VAE [Z-D=2]	z_1	0.09	0.02
vanilla-VAE [Z-D=2]	z_2	0.07	0.04
spatial-VAE	z_1	**0.95**	0.01
spatial-VAE	θ	0.01	**0.92**

Table 1: Correlation coefficients of the inferred latent variables with the ground truth factors in the 5HDB dataset.

Reviewer #3 We thank the reviewer for their helpful comments. We will clarify the method description in the final draft and will provide a comparison with the same effective total number of latent variables in the fixed/vanilla VAEs.

1. Section 2.1 and Figure 1 will be revised as suggested.

2. In Figure 2, the solid lines are training set ELBOs and the dashed lines are test set ELBOs. We now include this in the caption. Furthermore, we will include a comparison with the fixed/vanilla VAEs with the same effective total number of latent variables. For the transformed MNIST datasets, the spatial-VAEs with rotation/translation inference still outperform the standard VAEs even with the additional latent variables. We will update the discussion accordingly.

3. The reviewer is correct. Only the prior is defined to have mean zero. We have corrected this error in the text.

4. It is true that this work is limited to modeling global transformations and thus single objects. We think that extending this idea to handle multiple objects is an exciting future direction, which we now mention in the conclusion. Regarding object vs. camera transformations, generally speaking, object transformations and camera transformations are exact inverses. However, there are some interesting effects that can occur with light photography that are related to the angle of view and distance from the camera to the object (e.g. foreshortening, depth of field, etc.). Adaptation of this framework to explicitly handle these kinds of effects would also be an interesting future direction.