We thank the reviewers for their thoughtful feedback. We are grateful that all the reviewers recommended the paper for acceptance, and noted clear presentation, strong empirical results, and thorough ablations.

R1 questioned whether BigBiGAN truly represents an "important step forward toward answering some fundamental question in ML" rather than merely an empirical contribution or "tech demo". In our view, the question of whether generative models can learn interesting representations and be applied beyond generation is rather fundamental. We believe our work makes a strong case that unsupervised generative models of images are capable of learning interesting semantics which are practically useful for downstream tasks. With the exception of our work, the unsupervised representation learning field is currently dominated by methods based on self-supervision. Showing that state-of-the-art (at the moment of submission) results can be achieved using an entirely different family of methods is likely to be impactful, and will further motivate the community (especially those working on generative models) to explore representation learning applications further (R3 also mentioned this as a benefit). Beyond improving empirical results, both R1 & R2 noted our improved joint discriminator with unary loss terms, with R2 describing this as a "well-motivated methodological contribution".

On R1's specific concern that we only demonstrate "good linear probe classification performance", we now have additional results which we'll include in a future revision showing solid classification performance (43.3% top-1 accuracy) from non-parametric k-nearest neighbors (k-NN) classifier in the learned representation space.

R1 & R3 suggested adding demonstrations in smaller settings. We haven't experimented with smaller datasets (e.g. MNIST, CIFAR) mainly because the prior work we primarily build upon (ALI, BiGAN) has addressed these settings, and we expect little to be gained there. Note also that among recent work in unsupervised representation learning for images, ImageNet-focused evaluations are quite standard and small low-resolution datasets are seldom considered; there is simply not enough semantic "juice" in these tiny datasets to enable interesting representation learning. On model size: for the encoder, most of our experiments use the standard ResNet50 architecture – on the smaller end of widely-used models in visual recognition today – and for the generator, our ablations show that we can't afford to reduce its size much without significant cost to representation learning performance. That said, we do acknowledge that training these models involves significant computational costs and consider improved scalability an important goal for future research.

R2 & R3 suggested providing code for reproducibility. In a future revision, we are planning to release pre-trained models, and ensure that all training hyperparameters are fully specified.

R2 suggested visualizing iterated reconstruction results, where the reconstructed input image is passed through reconstruction again multiple times. This is an interesting experiment, and in fact we were curious about this as well and have created visualizations which we will add to a future revision. We did not observe an exact fixed point, but did notice that the process goes through stages, each of which produces images semantically similar to each other, while the images are quite different across the stages. We hypothesize this is due to the iterated reconstruction eventually reaching a point outside of the natural image distribution the model has been trained on, which leads to semantically different reconstructions at the next step. R2 also suggested an intuitive explanation of the losses – we will add this in a future revision as well.

R1 suggested adding comparisons to other methods, including results published after the submission deadline. We will include the results from concurrent work in a future revision of the paper. Regarding the results included in the submission, we attempted to include in Table 2 all recent competitive results on ImageNet from unsupervised approaches we were aware of, as well as a result from the original BiGAN due to its relevance to our method.

Finally, R1 asked for comparisons with VQ-VAE-2 on generation and representation learning. For ImageNet generation, BigBiGAN is trained unconditionally (without class information), while VQ-VAE-2 is class-conditional, so it's not fair to compare the two. Generally, BigBiGAN generators underperform state-of-the-art class-conditional generative models (BigGAN, VQ-VAE-2, etc.), which is expected, since our generator does not receive information about which class it should generate. VQ-VAE-2 did not report representation learning results (unsupervised or otherwise).