We would like to thank all reviewers for their time and consideration in reviewing our paper. We were impressed by the quality of the feedback and insights, and appreciate the reviewers' recognition of the benefits afforded by our work. We include here some comments from the reviewers. R1: "This work is perhaps the most effective in achieving [training time] reduction that I have yet seen". "This paper will spark discussion... and the discussion it sparks will have value". R2: "This work will no doubt be of substantial interest to the image generation community". R3: "Quite interesting and refreshingly new". "It is impressive that a very simple preprocessing strategy can result in substantial improvements in the performance". R4: "Exceptionally well-written". "Very handy and simple, which is a virtue". The reviewers also raised some questions to which we respond below. In the rebuttal, IS stands for Instance Selection (**not Inception Score**), while P, R, C and D stand for Precision, Recall, Density and Coverage metrics.

R1 Progress should be made with model adaptation, not dataset adaptation. Indeed, this is a philosophical question. Does it make sense to separate the model and the data if we can benefit downstream tasks by considering them jointly? We look forward to the discussions that our work will generate in this regard. Inception embedding yields advantages under certain evaluation metrics. We agree. We will add a note to clarify this. Small-GAN. We will cite this as related work in the camera-ready.

R2 Effectiveness of diversity-sensitive evaluation metrics. Of the metrics considered, two specifically target diversity captured by the model: R and C. R can be thought of as a measure of *unconstrained diversity*, where we care less about image quality and more about the spread of the samples. C, by definition, reflects *realism-constrained diversity*, where we are interested in modeling diversity among realistic looking model samples. In all of our experiments we observe that the application of IS does reduce R, which is expected given that the diversity of the training set is reduced. With C we observe an initial increase in performance, but when too much of the original dataset is removed it too begins to drop. We see this in Figure 3 in the C subplot, where retention ratio 20 hovers at a lower C value than retention ratio 30 before its collapse. This example demonstrates that we need to lose a lot of the diversity in the dataset before it becomes more of a problem than the gains from better image quality. This would also explain why FID improves when IS is applied, despite reduced diversity in the training set. We believe this follows the conclusion that "we are still in a regime where we can improve these metrics purely by improving sample quality, and have yet to hit the ceiling where we would have to increase diversity to push the numbers further". We will add discussion of this to the paper. Not enough emphasis on diversity trade-off. In the camera-ready we will emphasize that IS is a tool for trading diversity captured in the dataset for image fidelity, and that this trade-off is controlled by the IS retention rate.

R3 Post IS the model cannot generate tails of the distribution. While this is true, we have shown experimentally that for current GANs these outliers might result in a worse fit of the true data distribution. This gives practitioners a tool that they can choose to use depending on whether they prefer image fidelity, or the chance to generate distribution tails. Inception Score is not a good measure. We agree. We include it for completeness since it is a standard metric to report, despite its many shortcomings. BigGAN does not match reported results. We train our BigGAN models using the official PyTorch code on a GPU server. We change a single hyperparameter, the channel multiplier, from 96 to 64 to lower the memory requirements and speed up training, which is why our baseline does not match the FID of 9.77 achieved by the official implementation. As a note of interest, after the submission deadline we found that training IS models with a smaller batch size of 256 further improved performance (likely a regularization effect), achieving an FID of 9.61 in less than 4 days. Thus, we can outperform the official full capacity BigGAN model while using half the trainable parameters, and still train in ⅓ the time. Baseline evaluation metrics on real data. This is a great idea, we will include it in the camera-ready! CelebA HD. Datasets with less diversity already have high manifold density. As such, we expect the impact of IS to be less pronounced in these cases. CelebA HD has already been heavily pre-processed with face alignment (which might be considered a form of IS), so we would not expect much additional benefit from IS on this dataset. We train several SNGAN models (baseline and 95% IS) on CelebA HD and observe average improvements in P (0.83 to 0.85) and D (1.09 to 1.27) and impairments to FID (11.92 to 13.25) and R (0.41 to 0.35). C is unchanged (0.84 for both). Thus IS still achieves the desired effect of trading diversity for visual fidelity, but the trade-off is less appealing for well curated datasets. We note that these results were achieved using an ImageNet pretrained embedding, so better performance may still be achieved by using a domain-specific face embedding.

R4 Classification Accuracy Score (CAS). Generally, classifiers prefer diverse data over realistic data, which is why data augmentation techniques such as Mixup and CutMix perform well even though they produce unrealistic images. Since the goal of IS is to improve image quality by reducing diversity, we do not expect it to be helpful for data augmentation. We measure CAS using a ResNet18 architecture for quicker training. Our baseline 128×128 BigGAN model achieves a Top-1 accuracy of 8.69 and Top-5 of 18.51, while our IS model reaches a slightly lower Top-1 accuracy of 6.31 and Top-5 of 16.04. This falls in line with our expectations. Unsupervised embeddings. We trained a model using a ResNet50 SwAV embedding (Caron et al., 2020), which was recently released and shown to outperform SimCLR. This model achieves an FID of 16.65 after 200k iterations (compared to the baseline of 21.66), demonstrating that unsupervised embeddings can be used for IS. Learning useful unsupervised embeddings for domains outside of images is still an open problem, but we see this as an exciting direction for future work. PRDC for BigGAN experiments. We computed P&R and D&C for our BigGAN experiments and observe a similar trend to our SAGAN experiments: P, D, and C increase over the baseline, while R decreases. Baseline/IS: P=0.76/0.88, R=0.66/0.50, D=0.90/1.34, C=0.84/0.90.