Technical Contributions (R3, R4). Augmentation plays a key role in many machine learning systems. This paper has addressed discriminator overfitting — a fundamental problem in the GAN literature — via Differentiable Augmentation, which boosts GANs not only with **limited data** but also on the **large-scale** datasets. We believe that this is an important algorithmic contribution to the ML community. Though not theoretical contribution, many well-known empirical papers were also published at ML conferences (*e.g.*, DCGAN in ICLR, and LaplacianGAN in NeurIPS).

Image Diversity (R1). Our method can improve diversity. As suggested by R1, we use the recall metric [1] that estimates the coverage of the generated distribution and hence reflects the diversity. Ours with **20%** CIFAR-10 data (recall: **0.39**) is higher than the StyleGAN2 baseline with **20%** data (recall: **0.24**), even higher than its with **100%** data (recall: **0.33**). We will include it in our revision.

Clarification of T across G and D (R3). T is required to be the same random function but not necessarily the same random seed across G and D, since G and D are updated in different forward-backward iterations. We will clarify this in the revision.

Results of MineGAN (R3). We rerun MineGAN using their newly released code and will update their FID (obama: 50.63; grumpy cat: 34.54; panda: 14.84; cat: 54.45; dog: 93.03) and figures in the revision. Ours (StyleGAN2 + DiffAugment) outperforms MineGAN on 4 out of 5 datasets.

Why 100% Data in Fig. 6? (R3) We show that DiffAugment **even works** with 100% of CIFAR-10 data, where the discriminator still severely overfits the training set. This phenomenon is more severe with limited data. As the reviewer requests, with 10% CIFAR-10 data, at 10k iterations when the BigGAN baseline collapses, its D's training/validation accuracy is 99%/18% (**81%** difference, **severe** overfitting), while ours is 90%/41% (**49%** difference, **less** overfitting). Ours continues stable training for over 60k iterations, considerably alleviating the overfitting problem.

Naming of "Few-Shot" (R3). Thanks and we are happy to change it to "100-shot" in the revision.

Application to NLP (R4). Like image inpainting, the masking process in MaskGAN as R4 mentioned is used to construct the conditional input. This is not a form of discriminator augmentation, as D is still seeing the unmasked training set. We will cite MaskGAN and discuss its connection to our method. DiffAugment for NLP tasks is an interesting direction. We leave this for future work.

Choices of Augmentations (R1, R3). We mainly investigate the algorithmic perspective — where and how to apply augmentations to GANs; exhausting the set of augmentations is beyond the scope of this paper. In fact, we have tried many other augmentations like random scaling, rotations, shearing, smoothing, sharpening, and Gaussian noise but did not find them helpful. Moreover, when they are applied as "Augment reals only" or "Augment D only", all the results are consistently worse than the baseline. Thus we find that *Color*, *Translation*, and *Cutout* are especially effective for GANs. The simplicity also makes it easier to be deployed. We will discuss other augmentations in the revision.

Which Augmentations are Differentiable? (R4) Most existing augmentations could have a differentiable implementation, but they are currently absent in the widely used TensorFlow or PyTorch. Our code release provides differentiable implementations, which would benefit the community.

Grid Search (R2). Grid search can further optimize the performance, but our fixed *Color* + *Translation* + *Cutout* DiffAugment already works fairly well in most limited data settings, including CIFAR, few-shot, and our new results in Table 1. Although we used *Translation* + *Cutout* for the BigGAN models in the CIFAR tables, we later find that they can be further improved if *Color* is used as well (*e.g.*, from FID: 22 to 20 with 10% CIFAR-10 data). This combination is especially effective for GANs. Besides, we did not tune the level of each individual augmentation, which we found little beneficial, so the search space is significantly reduced.

Dataset:	FFHQ 256×256				LSUN-Cat 256×256			
# Training samples:	30k	10k	5k	1k	30k	10k	5k	1k
StyleGAN2	6.16	14.75	26.60	62.16	10.12	17.93	34.69	182.85
+ DiffAugment	**5.05**	**7.86**	**10.45**	**25.66**	**9.68**	**12.07**	**16.11**	**42.26**

Figure 1: FFHQ and LSUN-Cat results with the fixed *Color* + *Translation* + *Cutout* DiffAugment.

Results of Baseline for 100-Shot Generation (R3). Although each of our datasets contains only 1 object, their facial expressions, backgrounds, and poses are fairly diverse. D can easily memorize all those 100 training images and that's why the baseline StyleGAN2 is poor. What Fig. 3 presents is already the **best** training snapshot of the baseline model. It can be even worse if the training is longer.

Metrics for Overfitting (R3). As suggested by R3, the GAN-train/GAN-test metric is a good metric for assessing the generated images of the **generator**. E.g., the GAN-train/GAN-test of the BigGAN baseline with 10% data is 53.1%/72.4%, while ours achieves significantly better 62.7%/80.9%. However, in this paper, we only use the discriminator's accuracy on the real training/validation set to see if the **discriminator** overfits the real images. We will clarify this in the revision.

Typos (R1). Thanks for the suggestion. We will revise the paper thoroughly.

[1] Kynkäänniemi et al. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems*, 2019. 1