R1, **R2**, **R3**, **R4**: We thank the reviewers for the numerous positive comments. **R4**: **‘‘The proposed tool can have a good impact on the community and help standardize several experiments with synthetic data. I was impressed by the versatility of the framework”. R3**: **“The task of constructing harder and non-fixed datasets for training and evaluation is of great practical important.”**. **R1**: **“There is a paradigm shift happening from datasets to dataset generators (e.g. simulators). This tool is aligned with that shift and might be broadly useful.”. R2**: **“Very well written and structured.’**

R1: **“one is left wondering whether this insight generalizes beyond the specifics of this experiments/dataset?”**
R3: **“Are the results obtained on Synbols dataset generalizable to large-scale datasets?”**
In the general case, one should always be careful on how scientific findings can generalize to other setups. We made sure to communicate this properly in the paper and encourage a good experiment design and, when possible, to test on real world datasets. Specifically, when a failure case is detected with Synbols, it is expected to hold in at least some real world scenarios. For example, we expect that P-Entropy would fail in real world scenarios when some objects are occluded in the dataset. Next, an algorithm that solves the problem in various simulation setup is expected to at least offer insights on real world solutions.

R1: **“It is difficult to characterize what new scientific understanding or knowledge was presented in this paper.”**
We agree, many of the presented results are part of the wisdom of the more experimented researchers. While the aim of the paper is to provide a new tool for generating datasets (part of NeurIPS’s CFP), we seized the opportunity to solidify and quantify this knowledge, which would otherwise remain intuitions.

R1: **“The value of such tools is often clear only in hindsight... If they are useful, they see organic adoption.”**
For this purpose, we are investing effort on a high quality github repository with good documentation and ease of use. We will also properly advertise the tool to make sure it reaches the potential users.

R1: **“why you can’t render (and use) lower than 64x64 images from say CLEVR?”**
In CLEVR multiple objects of different sizes are present in a single scene. When resizing the image, some of these objects become so small that they are reduced to a single pixel.

R2: **“Authors use a very limited number of learning nnet models used for evaluation”**
There is a total of 14 different backbones implemented across all the experiments. We are happy to add results from other backbones such as Squeeze-and-Excite Networks or other recommendations. The main limitation is the readability of the tables and the space available for describing the backbones, but we are happy to add many more in appendix, such as our few-shot learning experiments.

R2: **“differences between those more proficient models are similar for some of the tasks addressed. Is it possible to “scale” the generation tool to address (future) more complex learning tasks”**
Yes, the generator is prepared to be extended with harder benchmarks such as video generation, and VQA. Other attributes such as symbol border, shadow, and texture are also planned (see conclusion). Note that the font classification task already provides a challenging benchmark for current learning algorithms. Results in the last column of Table 1 report 80.35% vs 67.20% for the two best performing models.

R2: **“further analysis to determine what instances (images) are more difficult/discriminant (e.g. performing IRT analysis ...”**
This would make a really interesting experiment. We will add this for the camera ready. In addition, we will investigate how the overall difficulty of the dataset is affected by those attributes with complexity measures (Ho and Basu 2002).

R3: **“The paper proposed a new data augmentation method to generate low resolution digit/text images with rich composition of latent features.” R3**: **“The proposed method seems only works for digit or text images, such as MNIST and SVHN. Can it be used on natural images, such as CIFAR10”**
Our work is not a tool for data augmentation. It is a tool for discovering model biases and it is not designed to generate natural images.

R3: **“The relation to prior data augmentation works [1-4] (and many others in their references) is not well discussed”** These data augmentation techniques are really interesting but orthogonal to our work. We now cite them in the related work section with a description of the differences between data augmentation and the proposed generator.

R4: **“It would be interesting to discuss why and which attributes were selected and the impact on the bias of the generated datasets.”** In the methodology sub-sections of each experiments, we describe the distribution of each attributes and the intent of this choice. We also show samples of each generated dataset in Appendix to qualitatively verify the biases. We also updated the supplementary material to extend the discussion on this bias as requested.

R4: **“In table 2, all scores in the last column should be red since they have all a drop of 5% compared to col. 1.”**
No, for font classification, the reference point is last column of Table 1. We made it more clear in the paper.