Reviewer#1 - 1) Adding variability measures for the results As recommended, we will add variability measures, for example, harmonic mean (average of seen and unseen), $39.9 \pm 11.3\%$([18]), $49.6 \pm 10.1\%$([29]), $51.5 \pm 13.2\%$([15]) and $52.4 \pm 13.9\%$(ours). **2) Visualizing the learned classes with low-dimensional toy simulation studies** We notice that the reviewer encouraged us to provide some low-dimensional toy simulations which could help other readers easily understanding what is going on. Since our network encodes, for example in AwA dataset, 4096D datapoints into 64D latent variables, we could display the structure of several datasets directly using T-SNE as shown in Fig. 1, especially for unseen classes. We hope it will be better supplements than recommended, then we will include this figure.

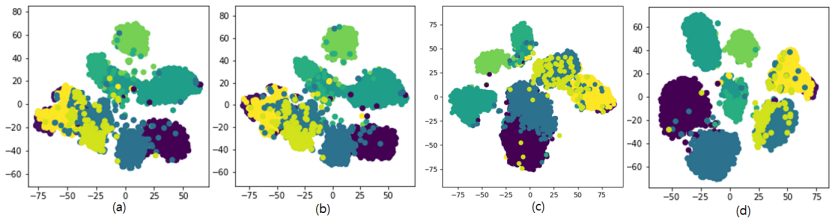

Figure 1: Structure visualization of learned dataset AwA1,2. Each color denotes unseen classes. Results of (a) mmVAE on AwA1, (b) SGAL on AwA1, (c) mmVAE on AwA2 and (d) SGAL on AwA2. While harmonic mean score is increased from 52.2% to 62.2% on AwA1, there are less drastic changes between (a) and (b). On the other hand, increased from 26.9% to 65.6% on AwA2, clusters are more separated from each other in (d) compared to (c).

Reviewer#2 - 1) Notation As recommended, we will reexamine the notations and try to modify them to simple forms. **2) Why VAE?** To generate datapoints and perform feedback training to catch intractable distributions, we need encoder-decoder structured non-parametric generative model like VAE. Although GAN is also a powerful model, absence of encoder limits our intend to implement feedback training and regularization of encoded latent variables and multi-modal prior distribution. **3) Reliance of the unseen classes** Previous works such as [8,19,21] aim to have semantic embedding model to cope with unknown attributes. On the other hand, [3,15,18,29] exploit generalization of generative models for zero-shot problems, assuming known attributes in order to generate samples for unseen classes on which classifiers are trained. Therefore, in our opinion, unknown attributes would be the better assumption for zero-shot problems but it is still worth studying with known attributes similar to other works mentioned above. **(a-e)** Please note that we aim to train a generative model for both seen and unseen classes, and overcome lack of training data for unseen by approximating missing samples. In other words, in one iteration for training, the model generates missing samples and is trained on both missing samples for unseen and existing ones for seen. This one iteration is formulated by EM, thus **(d)** one EM step is equivalent to train our network with just one iteration for unseen which will be insufficient to converge. **(a)** In order to examine EM, we show the results of mmVAE which is trained only on seen classes without EM. And we do not have any deep-NN classifier, since we exploit encoded features for classification, which is our other contribution. Specifically, **(b,d)** when our model is completely trained, we encode datapoints to latent variables and determine their classes by calculating Euclidean distances to each multi-modal and choosing the modal with minimum distance as Eqn. (3). **(d)** Even if the cited work [15] uses a deep-NN rather than the SVM, or if we use additional classifier, still our purpose is different from others since we aim to have the model with both seen and unseen; [15] also performs feedback with generated unseen datapoints, but only for updating decoder in order to restrict their encoded latent variables partially. **(c)** In [15], for example, they use 2 and 1 hidden layers for encoder and decoder with 512 hidden units, while ours use 1 and 1 with 512 for AwA dataset. **(e)** To perform generation and MLE, discriminative deep-NN could hardly be adopted, but generative model could be.

Reviewer#3 - 1) It is not fully clear how the SGAL strategy generates the missing datapoints As the reviewer commented, the multi-modal prior does play a crucial role. To generate missing datapoints by implementing Eqn. (4), 1) multi-modal prior generates the latent variables of each unseen class, and 2) decoder predicts the missing datapoints by decoding these latent variables. Subsequently the whole model is trained on both the generated datapoints of unseen and existing ones of seen. We will add this additional explanation in our paper. **2) The number of iterations for the benchmarks** For mmVAE and SGAL(EM): 170,000 and 1,300 for AwA1, 64,000 and 900 for AwA2, 17,000 and 2,000 for CUB1 and 1,450,000 and 1,500 for SUN1. We will add this with a table in our paper. **3) Fake samples?** As the reviewer suggested, approximations would be the better expression compared to fake samples, therefore we will modify it as recommended. **4) Computational complexity and memory requirements** Network structure and parameters can be a standard to examine the complexity and memory requirements, and we compare ours with other generative-based methods: For ours on AwA2, 1 hidden layer with 512 units is used for both encoder and decoder. In [15], 2 and 1 with both 512 units are used for encoder and decoder, respectively. In [18], 2 with 512 and 1 with 1024 are used for encoder and decoder respectively. [29] uses 1 with 4096 for generator, and 1 with 1024 for discriminator. We will add this evaluation to our paper.