We thank all the reviewers for the valuable and thorough feedback. We appreciate the recognition of *"addressing one of the biggest problems with the existing methods for unsupervised disentangled representation learning"* by [R2] with the most expertise. Our results beat benchmarks [R1, R2, R3] by roughly doubling the disentanglement scores while more than halving the variance across all results [R2, R3] and therefore sets a new state-of-the-art for the field [R1] and is *"likely to pave way for a large body of applied disentanglement work."* as [R2] puts it. We respond to selected comments below but will address all concerns in the final version.

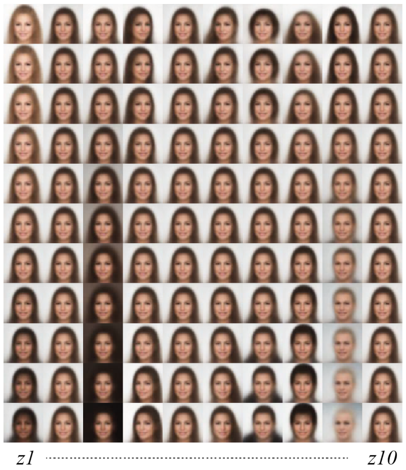

Figure 1: qualitative evaluation of *celebA* disentanglement.

[R2, R3] CelebA: We recognise that the demonstration of our method on *celebA*, beyond the benchmarks, would add some value. We therefore include preliminary results, particularly traversals of an exemplary model on *celebA* (Figure 1). z_{active} seem to disentangle the following: **z1**-Skin Color, **z3**-Background Color, **z4**-Head Rotation, **z6**-Hair Orientation, **z7**-Haircut, **z8**-Bangs, **z9**-Baldness. In the final submission we will include results from all models trained on *celebA*. **[R1] Remark Figure 2:** Indeed, this figure illustrates the consistency across models, which all learn to disentangle in the first `metaEpoch` the x,y position. Only in subsequent `metaEpochs` the other factors are learned. Model 1 of Figure 2 is used for the learning across `metaEpochs` in Figure 5, where one can see that in `metaEpoch` 2 scale is learned and subsequently shape/rotation in `metaEpoch` 3. **[R1] Eval/step definitions:** We will add formal definitions in the final manuscript, additionally to the ones already present in appendix E.2 (step) and Section 3.1 (eval). **[R1] Recursive linear algebra:** We appreciate the link to recursive linear algebra methods like Gram-Schmidt and will discuss them in the background section. **[R1] Figure 4 skewness:** One reason could be the number of models used for the violin plot. As described in the text, the reference figures consist of 50 models, as opposed to our 5 models. In the final paper, we will overlay a swarmplot over the violinplot to directly visualize that difference. **[R2] Latent Traversals:** We included full latent traversals for *shapes3d* in the manuscript (Figure 12), we will add in the same section full traversals for *dsprites* and *celebA*. **[R3] (Semi-)supervised comparison:** Your understanding of the (semi-)supervised setting is correct. The comparison with β-TCVAE is not fair, its intention is to serve as an example of vanilla unsupervised performance. In the (semi-)supervised setting, we mainly validate our approach and set an **upper limit** for the realistically achievable performance in the unsupervised setting, as the model trained with the surrogate labels cannot be better than models trained with the real labels. We nevertheless agree that the model selection β-TCVAE is a better baseline and will include it in Figure 1b) of the final submission. Preliminary results indicate that 1000-PBT-S-VAE still consistently outperforms them. **[R3] Underlying factors:** For computing the MIG and DCI scores, having images and labels is sufficient. Contrary to that, the FactorVAE metric as well as β-VAE metric require to sample from the generative model and perform interventions (Sec. 3.1 in [15]). For instance fixing one factor and varying all others randomly is used to estimate empirical variation in each dimension, in case of FactorVAE, similarly for the β-VAE (Sec. 4 in [23]). **[R3, R4] Description of `leaf-runs`:** Each `leaf-run` consists of multiple `metaEpochs`, where after each `metaEpoch`, the learned factors are removed from the dataset and training. Think of the intersection of z_{active} intervals as the set intersection of the sets of all images where the respective latent gets mapped onto a given interval. During each `leaf-run`, a subset of the original dataset gets surrogate labels, increasing the numbers of `leaf-runs` therefore increases the amount of total surrogate labels. In a complete dataset, we could recursively generate a tree from all z_{active} intervals, where the height of the tree is the number of all learned factors. For efficiency reasons, we restrict the labeling to `leaf-runs`, going from the root directly to the leaf, without further branching out. We will clarify in the revision. **[R3] Figure 3 LEV**: x-axis will be correctly labelled with 'image indices' and not latent encoding value, hence the large numbers. **[R4] Only PBT:** The introduction of PBT to variational training is only the first step of our work, introduced to get robust results across models in each pass. As pointed out by **[R1,R2,R3]** the subsequent recursive approach is really the essential part of our work, which allows it to beat all current benchmarks in unsupervised disentanglement learning by tremendous margins. **[R4] Hyperparameters:** When we refer to hyperparameter sensitivity, we are particularly referring to the very fundamental problem with the variable outcome of VAEs across datasets as pointed out by Locatello et al. [7], whereas our algorithms unoptimized hyperparameters worked across all datasets equally and exceptionally well. We will expand the discussion of hyperparameter schedules in supplementary section "B PBT Training Details". **[R4] Motivation for recursive model:** Models did consistently only learn x,y position in the first `metaEpoch` (Figure 3). Only in subsequent `metaEpochs` more factors are learned. More `leaf-runs` increase the amount of surrogate labels, therefore increasing performance and robustness significantly. The exact amount of `leaf-runs` needed will ultimately depend on the dataset. **[R4] Active latents:** We'll add the following formal definition of active latent factors, following [8]: if $KL(q(z_a|x)||p(z_a)) > 0.01$, z_a is active, otherwise not. **[R4] Latent representations:** Supplementary C describes the learned factors for *shapes3d*, we will add a similar figure for *dsprites*, a proxy can be seen in Figure 5. We will add a sentence in the discussion of the final manuscript, describing these learned factors as well as *celebA* results.