We thank the reviewers for their valuable feedback. All reviewers agree that the problem of evaluating unsupervised representations/groupings is of high relevance and interest to the representation learning community. However, R3 might have misunderstood the goal/setup of the approach, which we hope to clarify further in this rebuttal.

Reviewer 1. 1. Why SeLa and MoCo?: They are representative of two important classes of unsupervised representation learning algorithms: (1) deep clustering (SeLa) and (2) contrastive learning (MoCo). 2. Qualitative results (Fig. 3) & failure cases: The results in Fig. 3 are average-case results. Additional results shown and discussed in the Appendix also include failure cases (*e.g*. Fig. A1-h,j,l). In general, lower-purity clusters are less semantically coherent, thus the class-level description is moe likely to fail. We will discuss this in the main paper. 3. Purity (L241): Entropy and purity are measures adopted from cluster analysis. They are also briefly discussed in [3,8]. A cluster is highly pure if the vast majority of its samples share a common label. Since in this case this label is manually provided, high purity strongly correlates with high interpretability. 4. Error bars (Fig. 2): 95% CI, estimated per cluster.

Reviewer 2. 1. Soundness: choice of negative samples: Different functions for sampling negatives correspond to exploring different aspects of the learned classes. In the paper we considered the simplest choice (random negatives), which explores the overall learnability of a class. We have since also considered hard negatives, *i.e*. sampling negatives only from the class "most similar" to the target one, based on class centroids in feature space (SeLa results in Tab. 1). This tests whether there are sufficient *fine-grained* differences between classes to be learnable by humans. The outcome of this experiment suggests that this is often *not* the case, which is unsurprising given that the algorithms find a large number of clusters (3000) and thus likely over-fragment the data. It also indicates that, while clusters are often semantically coherent, they are not necessarily "complete", in the sense of encompassing *all* the images that should be grouped together — finding the right number of clusters remains an open challenge in literature, and this experiments emphasizes that. We will add these comments and results to the paper. 2. Significance/Usefulness/Scalability: The goal of most studies in *interpretability* is to analyse a model independently of downstream tasks. For that, the use of manual assessment is widespread despite its cost. Our contribution is of significance because it removes subjectivity in this popular category of assessment methods [5,24,72]. The method we propose acts complementarily to downstream tasks (*e.g*. training linear probes to compare against a pre-labelled dataset). Our findings are significant: we show that these fixed labels do not necessarily align with the ones discovered automatically, *i.e*. less pure clusters are also human-interpretable. As an example, in Fig. A1-g (Appendix), SeLa discovers a "newborn" class in ImageNet, which is not part of the existing label set. 3. Cluster size: We agree; SeLa returns clusters of approximately the same size (min: 418, max: 435 samples), MoCo's k-means clusters vary in size (min: 1, max: 1238, median: 407). For fair comparisons, we selected MoCo clusters with a min. size of 200 samples (mean: 465). 4. Desc. length and coherence: We observed no correlation between sentence length and coherence. However, human-written descriptions tend to be short for pure clusters since they can be easily described as a single concept (Pearson's $r = -0.38, p = 0.001$ between length/*purity*).

Table 1: Semantic coherence with hard negatives.

Π	Random	Hard
$(0.3, 0.4]$	71.8	55.3
$(0.4, 0.5]$	94.2	60.0
$(0.5, 0.6]$	97.2	71.8
$(0.6, 0.7]$	99.7	63.2
$(0.7, 0.8]$	98.0	65.3
$(0.8, 0.9]$	99.8	63.8
$(0.9, 1.0]$	98.8	72.2

Reviewer 3. 1. Generating Visual Explanations; Reed/Park/Kim: There seems to be a misunderstanding about the setting of our paper. The work and datasets on generating visual explanations address a very different problem, namely to generate a description justifying individual image predictions. Instead: (1) we answer the question of whether a *given* image grouping (unsupervised or not) is interpretable (can be learned by a human) and describable (can be captured by a description; L26-35); (2) we consider descriptions for image *groups*, not individual images (L57-58); (3) our primary goal is not to generate justifications, but to provide a human-based assessment method of interpretability in unsupervised algorithms, instead of matching their output to pre-scripted labels (see R2.2). 2. Class label included in the description: We use the term "class" to refer to an arbitrary image grouping, not a pre-specified category (L28-31). Similarly, by "class-level description" we refer to the group caption; it is *not* a description of a known label and does not reveal a "solution" — this is not applicable in our setting. 3. Few qualitative examples (descriptions): We have provided 4 pages of qualitative examples and an interactive demo with image- and class-level captions for *all* clusters in the sup. mat. 4. How are pairs sampled?: Positives are sampled uniformly at random from the cluster in question; the negative is sampled uniformly among all other clusters (L127-129). We include here experiments with hard negatives (see R2.1). 5. Where descriptions come from: They are either human-written or automatic (see Sec. 4, L204-207, 214-217). We obtain one description per cluster in each case. 6. Sentence encoder: We use Sentence-BERT (L269-271).

Reviewer 4. 1. Scalability: Please see R2.2. 2. Exploring hard negatives: We have included experiments with hard negatives (see R2.1). 3. Asking participants to assess captions: Note that our method is designed precisely to avoid asking annotators to express a judgment (L36-40) to remove subjectivity from the evaluation. Nevertheless, following your suggestion, we also conducted an experiment asking participants to rate how well the auto-generated caption matches an image group as a whole (scale: 1-worst to 5-best). The results (Fig. 1) follow the same trend as the results presented in the paper. 4. "Going up a level" in abstractness: This is indeed a very interesting direction that we have already started investigating as part of our future work.

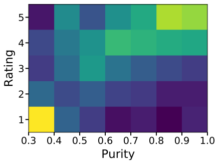

Figure 1: Histogram of user ratings.