We thank all reviewers for their thorough feedback that further strengthened our paper. The reviewers find the motivation for this novel clustering method and task clear (R1, R3) and the experiments convincing (R1, R2, R3, R4). In particular, they pointed out our contributions of a) clustering into non-uniform bins (R2, R3), b) the modality alignment (R2) and c) the non-parametric analysis of the results (R1). We first reply to main comments and then provide minor clarifications.

@R1 - **Intuition behind multiple heads when final NMI to ground-truth is similar.** We computed the average NMI between all pairs of heads as $(77.8 \pm 4\%)$. This means that the different heads do learn fairly different clusterings (NMI takes permutations into account) whilst being at a similar distance to the 'ground-truth' $(53.1 \pm 0.1\%)$.
@R1 - **NMI metric in Tab. 2 is hard to judge.** While NMI is standard in the clustering literature, we agree and have added the more interpretable Top-1 accuracy (Acc) from Sec. 4.2 to our now extended ablation table, see Tab. 1.
@R1 - **Scalability to larger datasets.** While it is possible to do the clustering step in smaller randomized chunks, even the current method can scale well to large datasets: Storage requires $64\text{bit} \cdot N \cdot K$, *i.e.* with $K = 300$ and $N = 2\text{M}$ for AudioSet, we only require 4.8GB on a GPU for the SK clustering. Hence, even HowTo100M could be clustered on a node with $8 \times 32\text{GB}$ GPUs without multi-node distributed processing.
@R1 - **Intuition behind better performance compared to XDC.** First, as shown in Tab. 1(b), simply using SK (24.7% NMI) instead of k-means (18.1% NMI) improves performance, likely for the same reasons that SeLa [2] outperforms DeepCluster [3]. Secondly, XDC (X variant) produces different labels for each modality, whereas we obtain consistent modality-invariant labels which allow the discovery of multi-modal semantics. XDC (concat variant) produces a single set of labels, but does not enforce modality invariance (concat did not work well for us either: Tab. 1c).
@R1 - **Fundamentally, [why] clustering vs direct representation learning?** We tackle clustering as an important problem in its own right. The motivation (`l.16-24`) is to learn to assign data to discrete buckets, basically *naming* things automatically, which is important for symbolic reasoning and language/human-centric applications. For this, we show in our evaluation that our end-to-end clustering approach works better than representation learning followed by clustering. On the other hand, the clustering task also leads to strong feature representations (see Tab. A.5).
@R2 - **An ablation experiment with both [alignment and non-uniform marginals] disabled would be helpful.** We have extended Tab. 2 from the paper: see Tab. 1 (c-f) below. The results are consistent.
@R2, R3 - **What is the baseline of SeLa using visual and audio features?** We have provided this as the SeLa "concat" variant in Tab. 1(b) below which uses both modalities, *i.e.* it concatenates the output of both modalities before having a joint fully-connected layer for classification. While the performance increases by almost five percent compared to the visual only variant, our method still outperforms this by more than 25% points in NMI.
@R4 - **Results are on curated, action-focused clips.** True, even if we do not use the labels, the datasets are pre-curated, so we have toned down `l.23-24`. Likewise, we assume that videos focus on single actions, whereas in many applications it would also be necessary to segment the videos in appropriate clips beforehand. However, platforms such as TikTok/Instagram etc. generate a wealth of short, action-oriented clips where our method could be applied directly.
@R4 - **It would be good to discuss potential biases in the few label setting [...].** Thank you for the suggestions. Indeed models trained using our method will inherit the biases present in the dataset, which could be known but also unknown, potentially leading to propagation of biases/inequalities. We will add this and more details to the BI section.

Clarifications. In the following, we address the minor clarifications raised by the reviewers.
@R1 - **Is it a pure alternate approach? [...] How many examples and reclustering steps?** Yes, we alternate between clustering and training (`l.128-130`). As we mention in the Appendix on `l.5-25` (referred to on `l.227` in the paper), we use all training examples and 100 reclustering steps using the same alternating schedule as in SeLa [2].
@R1 - **No need to reinit the last linear layer, is that correct?** Yes, no resetting is required.
@R1- **What does Alignment (A) stand for? Is it the technique of l.182.** That is correct. For clarity, we now denote this with modality alignment (MA) instead of just (A).
@R1 - **Why not directly use a Zipf rather than Gaussian?** Zipf distributions are well suited to videos in the wild. However, as we wish to compare clustering quality to human-generated labels, we use the introduced datasets, which are more curated and we therefore demonstrate the effect with a Gaussian prior in this paper.
@R1 - **The setup for retrieval is not clear. What does R@1 mean?** The setup follows the standard protocol from [22] (`l.281`) and is described in detail in App. `l.69-74`. R@k refers to retrieval performance using k nearest neighbors.
@R1 - **Rounding applied to Q?** Yes – we take the argmax of Q after SK.
@R3 - **Small time increase of decorrelated heads (DH) unclear.** DH only requires more forward passes when clustering, but not when training the CNN, which takes the majority of computing time ($\approx 90\%$). The overall compute time increases only by $10\% = 9.4h$.
@R4 - **Better clarify differences to ELo [58].** ELo is very different: It evolves and meta-learns a loss function by computing common self-supervised losses and distilling these. Clustering is used *as an evaluation metric* for meta-learning, as the distribution of labels in Youtube-8M is known to be Zipfian. We will add these details to Sec. 2 and Tab. A5.

Table 1: **[Updated Table 2]**

Method	🖼	(🔊 📹)	MA?	G.?	DH?	**NMI**	**Acc.**
(a) SeLa	✓	✗	–	–	–	20.1	5.9
(b) Concat	✗	✓	–	✗	✗	24.7	7.6
(c) Ours	✗	✓	✗	✗	✗	51.9	27.5
(d) Ours	✗	✓	✗	✓	✓	52.6	27.8
(e) Ours	✗	✓	✓	✗	✓	52.7	28.0
(f) Ours	✗	✓	✓	✓	✗	52.4	27.3
(g) Ours	✗	✓	✓	✓	✓	53.2	28.7