We would like to thank all the reviewers for their constructive feedbacks, we will edit the paper to include the missing discussion and implementation details.

[All Reviewers] Related work. We agree with the reviewers that a more extended discussion is required for related work, we will include all the suggested citations and discuss them properly in our final version.

[R1,R3,R4] Additional baselines. (1) Training infoNCE for more epochs. As shown in Table 1, the effective training epochs for CoCLR-RGB and CoCLR-Flow were 500 epochs, ending up with 70.2% and 68.7% respectively for the linear probe. In contrast, training infoNCE for RGB and Flow for the same number of epochs (500), the linear probe results are only 52.3% (RGB) and 66.8% (Flow). **(2) Cross-domain co-training is effective.** Our CoCLR mines positives across RGB and optical flow domain. **R3** suggests comparing against the baseline that mines positives within the domain, similar to the deep clustering algorithm. We experiment with mining positives within the domain, specifically, using RGB representation from infoNCE to construct positive sets for each sample in the RGB domain. This model is trained on UCF101 with the same schedule as CoCLR for a fair comparison. The linear probe results on UCF101 is 48.7%, significantly lower than that of CoCLR-RGB (70.2%). This is due to the fact that such within-domain sample mining process will not provide *hard* positives, as the samples in the positive set are already close to its corresponding query sample in the same representation domain, so they are actually *easy* positives, violating our basic hypothesis on the important role of hard positives. However, with multiple domains, easy positives in one domain are very likely to be hard in the other domain. **(3) Compare against other methods with optical flow input.** We note a recent arXiv paper (MemDPC, to appear in ECCV2020 by Han *et al.*) has also used both RGB and optical flow, yet CoCLR still outperforms it on both classification (90.7% vs. 86.1%, for pretrain on K400 and finetune on UCF101) and retrieval tasks (55.9% vs. 40.2% on R@1 on UCF101). We will add these discussions.

[R1,R3] UberNCE upperbound claim. UberNCE is in fact a supervised learning objective that encourage samples from the same category to be well-clustered, and samples from different categories to be separated on a unit sphere (forming spherical caps). Intuitively, this will enable the classes to be *linearly separable* in the feature space, which is the common criterion for evaluating representation quality. We will rephrase this claim.

[R1] (1) Test time augmentation. We actually used the same augmentation as DPC in their released codebase. **(2) Training cycles.** For the K400 experiments, we only managed to finish 1-cycle of training at the time of submission, but we conjecture more cycles will be beneficial, we will add the n-cycle results in the final version.

[R2] Relation and comparison to CMC. (1) Difference. CMC is an excellent paper, but its training scheme is fundamentally different to that of CoCLR. CMC aims to maximize the mutual information between **multiple domains of the same video clip**, *i.e.* learning the correspondence between its RGB and Flow representations, thus, the proxy task defined by CMC is still limited to be *instance discrimination*. However, CoCLR goes beyond this by explicitly allowing **multiple instances from the same domain** to be positive pairs, specifically, the most similar clip to one query sample based on flow representation is treated as a positive pair in the RGB domain, this design allows to exploit the complementary nature between different modalities, in addition, we also adopts a noise-tolerant MIL-NCE loss. These are the core contributions of our paper. **(2) Experiments.** We note that CMC will appear at ECCV2020, so it is *not* officially published yet. Nevertheless, we have carried out a comparison on UCF101, by training a CMC model with the same backbone (S3D) as ours for the same number of epochs (fair comparison with CoCLR). Specifically, two networks for both RGB and Flow streams are trained simultaneously by maximizing the mutual information among three views $\texttt{RGB}_t, \texttt{RGB}_{t+k}, \texttt{Flow}_t$, as in CMC section 4.2. We evaluated the RGB model by **linear probing**: the CMC model gets 55.2% on UCF101 (note that, the model in original CMC paper actually finetunes the entire network, and only get 59.1%, this is significantly lower than all of our finetune models in Table 1, above 78%), whereas the CoCLR-RGB model gets 70.2% by linear probing. We hope this has resolved the major concern raised by R2.

[R4] (1) CoCLR performance relies on flow as input. As shown in Table2, our CoCLR-RGB achieves 87.3% top1 accuracy, even without using flow for inference. Considering the size of the training data, we argue that CoCLR is amongst the most efficient and effective approaches. In addition, the general approach of co-training idea is not limited to flow (L64-68), and it remains unclear to us why flow is a stronger representation than text – after all, one can treat supervised learning as one special case of data with a text modality. **(2) Reproducibility.** This concern can be **fully resolved** by releasing all the training codes and models, we have promised to do so in the draft (as mentioned in L328). **(3) Training epochs.** CoCLR is trained longer only on the small dataset (UCF101) for the ablation study. But for K400, CoCLR was trained for 150 epochs (100 for infoNCE and 50 for the 1st cycle), we will clarify this. Furthermore, measuring epochs is only meaningful when all approaches use the same dataset for training. As shown in Table 2, other competitive approaches have been trained on orders of magnitude more data than K400, *e.g.* IG65M (273×), HTM (196×), Youtube8M-2 (169×), 1 epoch on these datasets will be equivalent to over 100 epochs on K400. **(4) K400 linear probe result.** The evaluation from "Watching the World Go By" mentioned by R4 actually trains both **a LSTM and a linear layer** to get 39% accuracy, so it is *not* a linear probe. At the time of submission, we reported 33.8% linear probing accuracy on K400 with RGB input. We have re-evaluated the same CoCLR-RGB representation with a proper regularization and learning rate schedule, achieving **40.5** top1-accuracy. In addition, we will continue CoCLR training on K400 for more epochs (match other competitive approaches) and this accuracy is expected to be further boosted.