Why XDC outperforms CDC and MDC? [all reviewers]. We have shown in Study I (Table 1) that XDC quantitatively outperforms both CDC and MDC on three downstream tasks. We provide the following intuition on why XDC is the best of the three models. XDC groups samples together when they are similar in one of the two modalities (video to supervise the audio encoder, audio to supervise the visual encoder). Instead, CDC groups samples together only if they are similar according to both the audio *and* the video modality (to supervise both encoders). Thus, XDC visual and audio clusters allow for more diversity than those of CDC. We hypothesize that this diversity allows XDC to learn richer representations, which translates into better performance on the downstream tasks. Also, recent work [A1] has shown that models trained on different modalities learn and generalize at different speeds, and that training them jointly (as done in MDC which uses two-modality heads) is sub-optimal. We believe that this could contribute to MDC performing worse than XDC, which optimizes for each modality independently.

Table A: XDC using two encoders of the same modality. We use Kinetics for pretraining and report the top-1 accuracy on split-1 of each dataset.

Method	UCF101	HMDB51	ESC50
XDC-visual-encoders	61.3	30.5	N/A
XDC-audio-encoders	N/A	N/A	66.0
SDC	61.8	31.4	66.5
XDC	74.2	39.0	78.0

Table B: XDC with ResNet3D-18 pretrained on Kinetics. We compare against the baselines: Scratch and fully-supervised pretraining (Superv) on the same backbone.

Method	UCF101	HMDB51	ESC50
Scratch	60.1	25.7	54.3
Superv	87.5	54.5	82.3
XDC	68.0	36.3	75.5

Table C: Temporal action localization on THUMOS14. We compare G-TAD [A2] algorithm using XDC features vs. using fully-supervised pretrained (Superv) features.

	mAP @ tIoU				
Method	0.3	0.4	0.5	0.6	0.7
Superv (Kinetics)	50.9	44.4	36.6	28.4	19.8
XDC (IG-Random)	**51.5**	44.8	36.9	28.6	**20.0**
XDC (IG-Kinetics)	**51.5**	**44.9**	**37.2**	**28.7**	**20.0**

Cross-modality vs. single-modality [R1, R2, R4]. We thank R2 for suggesting the insightful baseline corresponding to training XDC with the two encoders defined on the same modality (either visual or audio). Table A compares this baseline to SDC. It can be seen that the same-modality-XDC baselines perform similarly to SDC and are 8-12% worse than multi-modal-XDC. This suggests that cross-modality provides a superior supervisory signal for self-supervised learning and that multi-modal-XDC is the best model not because of its optimization strategy but rather because of the use of the other modality for pseudo-labeling.

XDC using a different backbone [R2]. We pretrain XDC on Kinetics with ResNet3D-18 as the visual backbone and keep the same audio encoder. The results are compared with those of baselines in Table B. XDC with the ResNet3D-18 backbone outperforms the training from scratch baseline by good margins on three downstream tasks.

XDC for other tasks [R1]. Table C provides the results of transferring XDC to the task of temporal action localization on THUMOS14 dataset. We employ the recent G-TAD [A2] algorithm, where we replace the clip features (originally extracted from a TSN model pretrained on Kinetics) with XDC features from the R(2+1)D-18 model pretrained on IG-Kinetics or IG-Random. We compare against the features from the R(2+1)D-18 model fully-supervised pretrained on Kinetics. We do not finetune any of the feature extractors used in this experiment. Both XDC variants outperform the fully-supervised features across all temporal Intersection over Union (tIoU) thresholds. This confirms the same trend observed in the tasks discussed in the paper and suggests that XDC can also be used for other tasks.

Learning using audio rather than text from ASR [R2]. We note that while our approach was demonstrated by leveraging audio, the method is general and is easy to adapt to other modalities, including text. While video and text are semantically correlated, audio and video are temporally correlated. Thus, these two form of correlations are likely to provide different forms of self-supervision, potentially leading to further gains when used in combination. A disadvantage of text from ASR is that it is only available for videos with speech. Audio provides information about environmental sounds beyond speech (*e.g.* walking steps, playing guitar, and dog barking) and allows us to train on uncurated datasets of arbitrary Web videos, as we demonstrated with IG-Random.

AVTS pretrained on IG-Kinetics and IG-Random [R4]. Training on such large datasets is expensive and unfortunately cannot be done within the short rebuttal period. However, we extensively compared XDC against AVTS (Section 6) pretrained on Kinetics, AudioSet-240K, and AudioSet using the same backbone. These results suggest that XDC outperforms AVTS consistently under the same settings on UCF101 and HMDB51.

Other comments. We thank R2 for suggesting an alternative pseudo-labeling initialization method. We will investigate this approach. Training on IG-Kinetics or IG-Random takes about 360 hours on 160 V100 GPUs. We will add the suggested references (by R1, R2, R3) to the final version and adjust the claim on using more data (by R1). We truly appreciate the constructive feedback from all reviewers.

References

[A1] Wang *et al.* What makes training multi-modal classification networks hard? In *CVPR*, 2020.

[A2] Xu *et al.* G-TAD: Sub-graph localization for temporal action detection. In *CVPR*, 2020.