R1: Thank you for your review and understanding the potential impact of this dataset. We will clarify the sentence which is stating that we study the effect that diversity has on recognition and add discussion on the long-tailed distribution of actions and way for it to be addressed, such as weighting the loss per-class or sampling classes in the tail more frequently.

R2: Thank you for your review, you raised some good points we would like to further address.

1. Yes, we agree that different actions in different cultures have different meanings. This is one of the major motivations for creating the dataset. When selecting the action classes, we did our best to avoid situations where the action meaning could be ambiguous. For example, we would label an action as ‘nodding head’ rather than including the meaning (i.e., ‘nodding yes’). We will add further discussion of this point and closely check all the classes to find any that may have ambiguity.

2. AViD is roughly the same size as Kinetics, HACS, etc., as we tried to illustrate in Table 1. Note that pre-training with AViD outperforms pre-training with Kinetics-600, despite that Kinetics-600 is a bit larger in terms of the video hours. We will add further experiments using subsets of AViD to more closely analyze how the size impacts performance. Similarly, we will add more experiments training on different diversity splits of AViD to determine how that compares, in the final version of the paper. The video quality (image resolution and frame rate) of AViD matches existing datasets like Kinetics.

3. We believe we provide statistical measurements showing that the AViD dataset sufficiently differs from the existing datasets in terms of the diversity. These include the heatmaps in Figure 1 as well as the data comparison in Table 3. In addition, we believe the experiments in Table 4 explicitly confirms the downside of the existing datasets compared to AViD. We will add more fine-grained experiments in the final version of the paper, such as training on one country/continent and testing on another.

4. We will add more discussion on the diversity in the paper. One discussion point is the population prior; the metric in Table 3 assumes a uniform distribution across countries, but this ignores aspects like population density. Places like the Sahara desert have a very low population but are weighted the same as places like India. We will add another metric comparing the distribution of videos to a distribution based on population density. We will further discuss these metrics and their drawbacks to more clearly state and explain the diversity of AViD and any potential limitations of it.

Overall, we believe the introduction of the AViD dataset will make a strong positive impact to the community, particularly compared to the standard practice today: using heavily biased video datasets (e.g., Kinetics with 90% of videos from North America) for both training and evaluation.

R3: Thank you for your review. To answer your questions:

1. All the videos are trimmed clips from longer ones. These intervals were annotated and checked by humans. We do not have the original untrimmed clips due to data storage limits (1TB for the full, untrimmed version) and anonymization difficulties. Using only the trimmed clips has been standard on many other datasets like Kinetics, moments in time, etc., and we tried to follow this while enhancing the country diversity, privacy, and stability in AViD.

2. Yes, during the annotation process, we did have attention checks of the annotators where they were asked to label videos we had manually done. If they failed those videos, their annotations were discarded. The annotators were asked to label 10-15 videos per task, and we further manually checked one video from each reviewer (in addition to the attention check) to ensure they were doing a good job. While there is some noise in the annotation process, overall we are confident the annotations are accurate. We will add this description to the paper.