We thank all four reviewers for their constructive comments. We respond to each reviewer's comments separately below. We also report some results from new experiments suggested by the reviewers.

R1: (1) More context from computer vision: We thank the reviewer for suggesting these references, which we will incorporate in an improved discussion of the related work on representation learning from videos.

R2: (1) More details about temporal classification: The basic idea in temporal classification is to divide the entire longitudinal video dataset into shorter "episodes" and use these episode labels as implicit training signals: i.e. the model tries to learn which episode a given frame belongs to. This idea is illustrated in Figure 2 in the paper. In the revised paper, we will describe the temporal classification model in more detail. **(2) Training temporal unsupervised models:** First, we emphasize that our temporal classification algorithm is an example of unsupervised (or self-supervised) learning. Second, we followed the reviewer's suggestion and developed a temporal version of MoCo as a new baseline: for each frame, we used the two neighboring frames as "positive" examples with respect to that frame and the remaining frames as "negative" examples. The temporal MoCo model thus tries to learn similar embeddings for neighboring frames. The temporal MoCo model performed slightly better than the purely image-based MoCo model reported in the paper, but still substantially worse than our temporal classification model (e.g. TC-S: 60.4%, MoCo-Temp: 49.3%, MoCo-Img: 46.6% in the labeled S dataset). These new results will be included in the revised paper. **(3) Exemplar split in Toybox:** In Toybox, each of the 12 categories contains 30 exemplars. In the exemplar split, we use 27 of these exemplars to train linear classifiers on top of the pre-trained and frozen features. We use the remaining 3 exemplars for testing. This split is challenging because it requires few-shot generalization, i.e. learning each category from 27 exemplars only. **(4) ImageNet-pretrained models on labeled S vs. Toybox:** ImageNet-pretrained models perform slightly better on Toybox than on labeled S, because it is likely that the Toybox dataset is "more similar" to ImageNet (both are taken by photographers/camerapeople, with a central object, etc.) than the naturalistic headcam videos. **(5) Single feature selectivity analysis:** The main goal of this analysis is to investigate the extent to which the high-level visual categorical information is distributed vs. localized in the self-supervised models. We will do a better job of motivating this important question in the revised paper. The particular analysis used here was inspired by Leavitt & Morcos (2020), who show that distributed representations lead to better object recognition performance. In our case, we have observed a mostly distributed representation of visual categorical information (Figure 7a), giving us assurance that our self-supervised models behave as expected from a high-performing object recognition model.

R3: (1) Linear evaluation on ImageNet: ImageNet results reported in the Supplement were obtained without any data augmentation. We have found that it is possible to get slightly better results (22.3% vs. 20.9% top-1 for child S; and 25.2% for a temporal classification model trained on data from all three children) with standard data augmentation methods typically used for this benchmark, i.e. random resized crops and random horizontal flips. Second, there was a typo in the Supplement: top-1 accuracy for a random net should be 1.2%, not 10.2%; so, the self-supervised models perform much better than the random model, suggesting the learned features are indeed quite useful. Third, our goal is not to build a sota model for ImageNet or for any other benchmark. Our main goal in this work is rather to investigate, for the first time, whether it is possible to learn useful, high-level visual representations from developmentally naturalistic videos. SAYCam is much closer to the early visual experience children receive, compared to any other dataset available. So, for our purposes, it doesn't make sense to use some other dataset, as the reviewer suggests, just because it happens to contain more data or higher quality data. If machine learning is to help us understand how the human mind develops, it must account for learning from datasets like SAYCam. **(2) More intuitive discoveries:** In the current work, our focus was explicitly on learning high-level visual categorical representations. In future work, we are also very much interested in exploring the capabilities of the trained self-supervised models further. We thank the reviewer for suggesting object grouping and optical flow as possible properties to investigate further. **(3) Missing references:** We thank the reviewer for pointing out the missing references. After the submission, we have also noticed that the visualization technique we reported in the paper is identical to CAM. We have revised the paper to acknowledge and cite this earlier work. **(4) Transfer learning on further datasets:** We would be happy to include these in the revised paper.

R4: (1) Application to other video datasets: The temporal classification algorithm most naturally applies only to longitudinal video datasets and unfortunately we are not aware of very many such longitudinal datasets. Most video datasets in computer vision consist of relatively short video clips instead (typically on the order of tens of seconds); in this setting, the temporal classification model becomes similar to a video instance embedding model, which has been explored before (cf. Zhuang et al., 2020). Also, please note that our main goal in the paper is not to demonstrate the effectiveness of a particular algorithm across a range of datasets, but rather to investigate whether it is possible to learn useful, high-level visual representations from developmentally naturalistic videos (cf. our response **#1** to **R3** above). **(2) Longer segments:** We have indeed tried 576s segments and observed that the model's performance seems to saturate around this point (e.g. 58.6% vs. 58.4% top-1 accuracy in the labeled S dataset for 576s vs. 288s segments, respectively). We expect that the performance will start to deteriorate at some point as the segment length is increased further, but we have not had a chance to try longer segments yet. We would be happy to report new results with longer segments in the revised paper. We thank the reviewer for this suggestion.