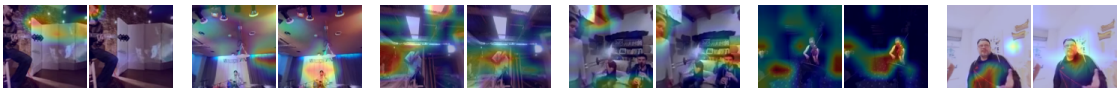

Figure 1: Sound localization maps from AVC (**1st** of each pair) and AVSA (**2nd** of each pair) models.

We thank reviewers for the constructive feedback. We're glad reviewers found AVSA novel, well-motivated (R1,R4), and superior to prior work in several tasks (R1,R2,R3,R4); and appreciated our dataset (R2,R4). We'll address minor points in the paper. Some requested ablations were already in suppl. Table 3. We'll summarize them in the paper.

Significance (R1) / Why 360 videos? (R3) We respectfully disagree with R1. The standard of "I see no practical significance" could be used to kill most SSL research, as it rarely outperforms full supervision. Applying this standard to 360°SSL seems equally short sighted. It is true that 360°SSL currently underperforms other forms of SSL for tasks like action recognition. This is because 360°datasets are much smaller. However, with the push for AR/VR and self-driving cars, this is temporary. The 360°camera market is projected to grow 25.6% annually until 2027 [link], and content availability should grow even faster. Furthermore, 360°SSL will enable new applications. AVSA can be used, as is, to detect and correct misalignments between 360°cameras and external microphones commonly used when shooting high quality 360°content. AVSA should also enable better audio localization, since it provides direct spatial supervision by matching co-located audio and video crops. However, no labeled video datasets exist for evaluation (only a small set of static images [50]). While we aim to annotate a sound localization dataset, this is beyond the scope of this work.

In conclusion, there is a flourishing literature on audio-visual SSL methods, which are SOTA for SSL in tasks like action recognition, and more will appear at NeurIPS. We show that, on the same dataset, AVSA clearly beats these methods, even on the AVC task. Hence, there is no reason to believe that 360°SSL will not be SOTA in due time.

Alternative implementations: Context-aware puzzles (R1) / Regressive AVSA (R2) / Other SSL losses (R3). These are great suggestions, although they shouldn't challenge the main contribution (showing that spatial information boosts representation learning). Given the short rebuttal cycle, we didn't try all, but will add a full comparison to the camera ready. One advantage of our method is the seamless combination of audio-visual correspondence (AVC) and spatial alignment (SA) by simply extending the negative set. Suppl. Table 3c shows that this combination is crucial. A puzzle-like approach (R1), where the arrangement of audio/video crops is predicted *simultaneously* by an MLP, only handles SA and thus performs worse (61.7% video@1 on UCF vs. 73.8% for AVSA). Combining with AVC improves performance (70.3%), but is still less effective. We tried Regressive AVSA (R2), but it performed much worse as well (66.7%). Per R3 suggestion, we experimented with masking random K audio/video crops and using the masked features to setup a multiple choice problem. The best design (K=4) was worse than AVSA (63.3% w/o AVC, 70.4% w/ AVC), likely due to unmasked video crops making SA of other video crops easier. We didn't try single modality SSL losses, eg next frame prediction, as prior work [1,3,30,41] show the superiority of cross-modal over within-modal supervision.

"Spatial alignment" in title (R2) There is a misunderstanding. Spatial alignment is our main source of information for representation learning (L154-161), as we seek to match audio and video crops *from the same location*. And we do *evaluate* on audio-visual spatial alignment directly (Table 2, AVSA-Bin task). Not surprisingly, AVSA performs best.

Evaluation by sound source localization (R2) / visualizing alignment (R3) Fig. 1 shows the GradCAMs of audio-visual alignment scores for AVC and AVSA models. AVSA localization is more accurate. We'll add these to the paper. Quantitative comparison to prior work is infeasible, since localization ability has historically been shown qualitatively.

Semantic segmentation (R2) We disagree that boost is marginal. We emphasize that only AVSA can learn context features, which yield a 3% mIOU boost over AVC and 1% over AVTS. Comparisons to ImageNet pre-training are not informative, since segmentation targets are from an FPN model *pre-trained on ImageNet* and tuned on COCO. Kinetics pre-training is slightly better than AVSA (75.5% acc and 36.9% mIoU using video alone) but relies on human labels. Random initialization performs poorly (29.7% acc and 8.5% mIoU) since we evaluate frozen representations (common SSL practice). Full training is shown in last row of Table 3, but not comparable to frozen representations (other rows).

Crop selection (R2) Crops are sampled randomly, with a minimum distance enforced (L186). More sophisticated crop selection is an interesting direction, eg guiding selection toward objects using objectness scores, moving objects using optical flow, or sound sources using audio energy from different viewpoints. We'll add future work to the paper.

Impact of batch size (R3) Yes, large batches should improve AVSA. This would require (1) a lager GPU cluster for distributed training [9] or (2) maintaining a memory bank (NPID [61]) or buffer (MoCo [23]) of features. However, unlike [61,23], our features vary with location, and storing several features per video (one per location) is infeasible.

Why using mono? Audio input ablation (R4) There is a misunderstanding here. We use full ambisonics aligned with each viewpoint by a 3D rotation (L195-196), and compare to mono/stereo inputs in suppl. Table 3a.

Is AVSA complementary to AVC/AVTS? (R4) Suppl. Table 3c shows that AVSA is complementary to AVC. New experiments shows the same for AVTS only on segmentation task. Spatial negatives alone (AVSA) yield 35.8% / 68.5% on segmentation / UCF. Spatio-temporal negatives (ie AVSA+AVTS) yield 36.1% / 65.8%. We'll add to the paper.

Understanding AVSA gains (R4) We compared class accuracies of AVSA and AVC on segmentation. AVSA indeed improves the most on background classes: rocks (27.7→34.7), window (41.2→46.0), pavement (33.3→36.8), sand (38.8→42.1), sea (46.8→50.1) and road (45.1→47.1). This suggests that AVSA disentangles common backgrounds more effectively. We'll add a complete error analysis to the paper. Thanks for the suggestion!