Audio−Visual Segmentation Supplementary Material

⋆Jinxing Zhou[1,2], ⋆Jianyuan Wang[2,3], Jiayi Zhang[2,4], Weixuan Sun[2,3], Jing Zhang[3], Stan Birchfield[5], Dan Guo[1], Lingpeng Kong[6,7], Meng Wang[1], and Yiran Zhong[2,7]

[1]Hefei University of Technology, [2]SenseTime Research, [3]Australian National University, [4]Beihang University, [5]NVIDIA, [6]The University of Hong Kong, [7]Shanghai Artificial Intelligence Laboratory
{eric.mengwang, zhongyiran}@gmail.com

In the supplementary material, we provide the following content: (A) the statistical details of the Single-source subset of the AVSBench; (B) additional ablation studies on our components and settings; (C) more discussion about the model training and inference; (D) additional qualitative results of the proposed AVS framework compared to the methods from SSL, VOS, and SOD tasks.

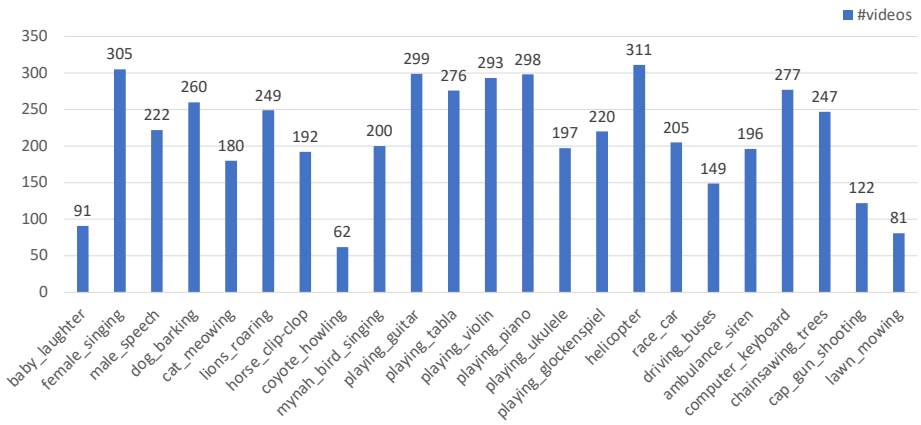

Fig. A1. Statistics of the Single-source subset of AVSBench. The texts represent the category names. For example, the 'helicopter' category contains 311 video samples.

A Dataset Statistics

The Single-source subset of AVSBench contains 4,932 videos over 23 categories. We provide the category names and their corresponding video numbers in Fig. A1. Some video samples are also listed in Fig. A2. Additionally, we provide some video samples and their annotations in the "dataset_sample.zip" file, containing

both the visual and audio contents, covering the Single-source and Multi-sources sets.

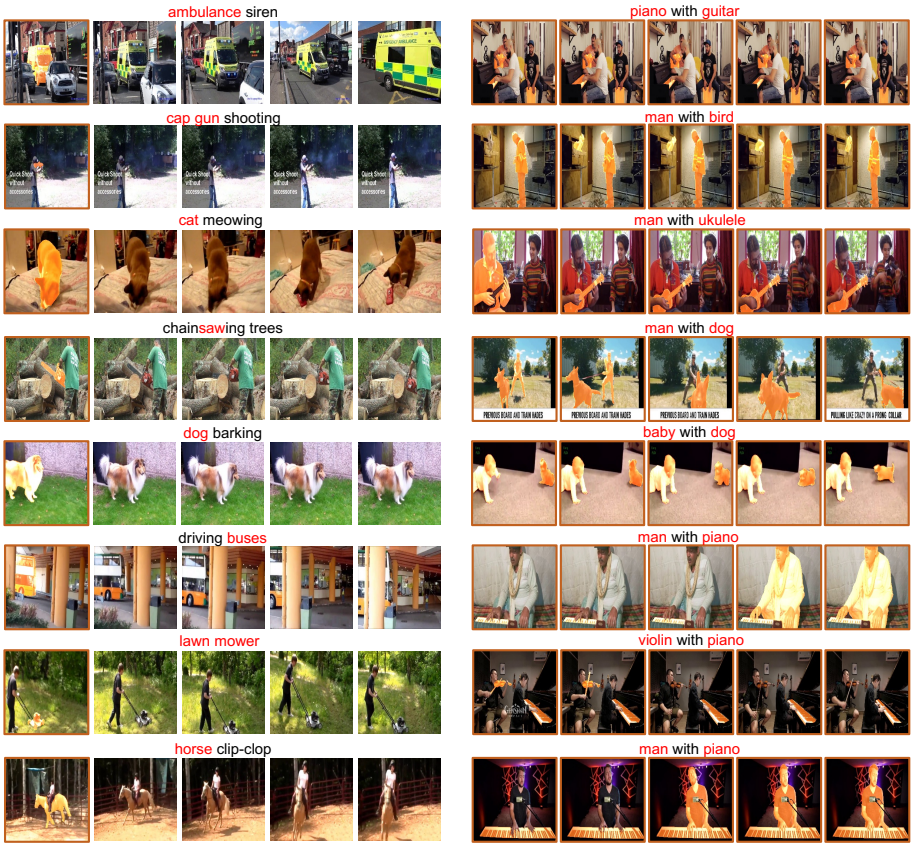

Fig. A2. AVSBench samples. The AVSBench dataset contains the Single-source subset (Left) and Multi-sources subset (Right). Each video is divided into 5 clips. The annotated clips are indicated by brown rectangles and the name of sounding objects is highlighted by red texts. For the Single-source training set, only the first frame of each video is annotated, whereas for other sets all the five frames are annotated.

B Ablation Study

Cross-modal fusion at various stages. The detailed architecture of the proposed TPAVI module is provided in Fig. A3. It is a plug-in architecture that can be applied in any stage for cross-modal fusion. As shown in Table A1, when the TPAVI module is used in different single stage, the segmentation performance fluctuates. For the variant based on the ResNet50 backbone, the model performs

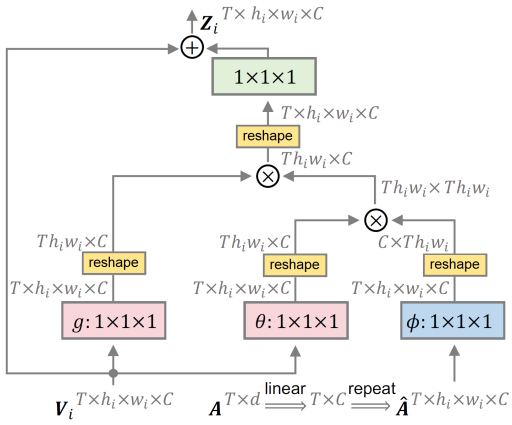

Fig. A3. The TPAVI module takes the i-th stage visual feature $\boldsymbol{V}_i$ and the audio feature $\boldsymbol{A}$ as inputs. The colored boxes represent $1 \times 1 \times 1$ convolutions, while the yellow boxes indicate reshaping operations. The symbols "⊗" and "⊕" denote matrix multiplication and element-wise addition, respectively.

Table A1. Cross-modal fusion at various stages, measured by mIoU. For both the S4 and MS3 settings, the model achieves the best performance when the TPAVI module is used in all four stages.

Setting	Backbone	i-th stage of Encoder, $i \in \{1,2,3,4\}$						
		1	2	3	4	3,4	2,3,4	1,2,3,4
S4	ResNet50	68.55	69.56	**71.30**	69.99	71.29	71.98	**72.79**
	PVT-v2	78.30	**78.58**	78.02	77.70	78.19	78.47	**78.74**
MS3	ResNet50	41.62	42.37	**43.02**	42.29	44.84	45.98	**47.88**
	PVT-v2	46.16	48.79	47.35	**49.01**	49.79	50.53	**54.00**

best when employing the TPAVI module at the third stage under both S4 and MS3 settings. As for the PVT-v2 based model, it is better to use the TPAVI module at the second stage under the S4 setting, and at the fourth stage under the MS3 setting. For all the settings, using TPAVI at the first stage cannot achieve the best performance, and we attribute it to that the visual features at the first stage enjoy limited semantics. Since our decoder architecture adopts a skip-connection, it would be beneficial to apply the TPAVI modules in multiple stages, as verified in the right part of Table A1. For example, under the MS3 setting, applying the TPAVI modules at all the four stages would increase the metric mIoU from 49.01% to 54.00%, with a gain of 4.99%. It indicates the model has the ability to fuse and balance the features from multiple stages.

Table A2. Performance with different initialization strategies under the MS3 setting. Compared to training from scratch on the Multi-sources subset, we observe a significant performance improvement if pre-training the model on the Single-source subset first. Note the proposed $\mathcal{L}_{\text{AVM}}$ loss is used in all the experiments of the Table. The metric is mIoU.

AVS method	From scratch		Pretrained on Single-source	
	ResNet50	PVT-v2	ResNet50	PVT-v2
wo. TPAVI	43.56	48.21	**45.50**	**50.59**
w. TPAVI	47.88	54.00	**54.33**	**57.34**

C Discussion of the model training and inference

Pre-training on the Single-source subset. As introduced in Sec. 3 of the paper, the videos in the Multi-sources subset share similar categories to those in the Single-source subset. A natural idea is whether we can pre-train the model on the Single-source subset to help deal with the MS3 problem. As shown in Table A2, we test two initialization strategies, *i.e.*, from scratch or pretrained on the Single-source subset. It is verified that the pre-training strategy is beneficial in all the settings, whether we use the audio information ("AVS w. TPAVI") or not ("AVS wo. TPAVI"). Taking the PVT-v2 based AVS model for example, the mIoU is improved from 48.21% to 50.59% (by 2.38%) and from 54.00% to 57.34% (by 3.34%), respectively without or with TPAVI. The phenomenon is more obvious if using ResNet50 as the backbone and adopting the TPAVI module, where the mIoU increases from 47.88% to 54.33% (by 6.45%). With pre-training on the Single-source subset, the model can learn prior knowledge about the audio-visual correspondence, *i.e.*, the matching relationship between the visual objects and sounds. This kind of knowledge is naturally beneficial.

Segmenting unseen objects. The proposed AVS framework is trained without accessing the category labels of the sounding objects, and hence it can be used to predict the videos which do not strictly lie in the category vocabulary of AVSBench dataset, though may have a performance drop for unseen objects. We display some qualitative examples on real-world videos whereas the sounding objects are barely not appeared in the training set of AVS model. As shown in Fig. A4, the pretrained AVS model has a certain ability to segment the correct sounding objects in the case of single sound source (a), multiple visible objects (b, c), and multiple sound sources (d). We speculate that the pretrained AVS model learned some prior knowledge about audio-visual correspondence from AVSBench dataset that helps it to generalize to even unseen videos and give possibly accurate pixel-level segmentation.

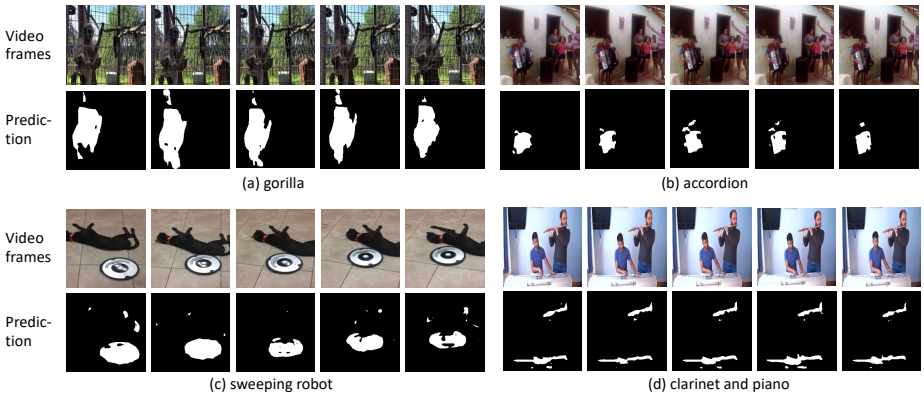

Fig. A4. Qualitative examples of applying the pretrained AVS model to unseen videos. The caption in each sub-figure indicates the sounding object(s) accordingly. There are almost no videos having the same category as these sounding objects during AVS model training. The pretrained AVS model gains the ability to segment the correct sounding object(s) in both single and multi sources.

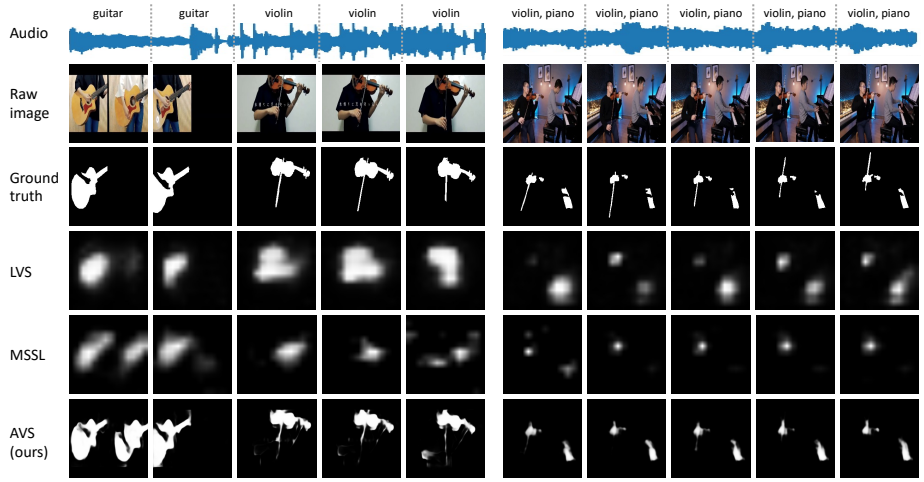

Fig. A5. Qualitative examples of the SSL methods and our AVS framework, under the fully-supervised MS3 setting. The SSL methods (LVS [1] and MSSL [3]) can only generate rough location maps, while the AVS framework can accurately segment the pixels of sounding objects and nicely outline their shapes.

D Qualitative Samples

Qualitative comparison between AVS and SSL. We provide some qualitative examples to compare our AVS framework with the SSL methods, LVS [1]

and MSSL [3]. As shown in the left sample of Fig. A5, LVS over-locates the sounding object *violin*. At the same time, MSSL fails to locate the *piano* of the right sample. Both the results of these two methods are blurry and they cannot accurately locate the sounding objects. Instead, our AVS framework can not only accurately segment all the sounding objects, but also nicely outline the object shapes.

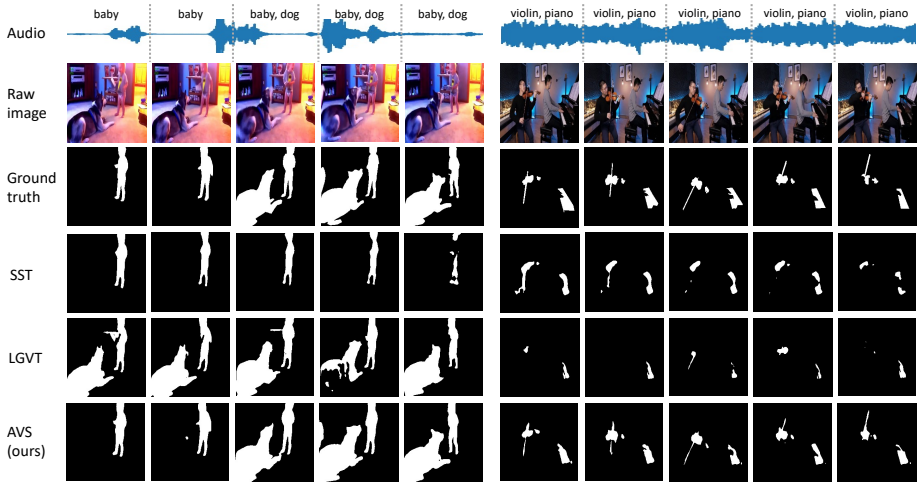

Fig. A6. Qualitative examples of the VOS, SOD, and our AVS methods, under the fully-supervised MS3 setting. We pick the state-of-the-art VOS method SST [2] and SOD method LGVT [4]. As can be verified in the left sample, SST or LGVT cannot capture the change of sounding objects (from 'baby' to 'baby and dog'), while the AVS accurately conducts prediction under the guidance of the audio signal.

Qualitative comparison between AVS and VOS/SOD. We compare the proposed AVS framework with the state-of-the-art methods from VOS and SOD, *i.e.*, SST [2] and LGVT [4], respectively. As shown in Fig. A6, SST and LGVT can predict their objects of interest in a pixel-wise manner. However, their predictions rely on the visual saliency and the dataset prior, which cannot satisfy our problem setting. For example, in the left sample of Fig. A6, the *dog* keeps quiet in the first two frames and should not be viewed as an object of interest. Our AVS method correctly follows the guidance of the audio signal, *i.e.*, accurately segmenting the *baby* at the first two frames and both the sounding objects at the last three frames, with their shapes complete. Instead, the VOS method SST misses the barking dog at the last three frames. The SOD method LGVT masks out both the *baby* and *dog* over all the frames mainly because these two objects usually tend to be 'salient', which is not desired in this sample. When it comes to the right sample of Fig. A6, we can observe that LGVT almost fails to capture the *violin*, since the violin is relatively small. The VOS method SST

can find the rough location of the violin, with the help of the information from temporal movement. In contrast, our AVS framework can accurately depict the shapes and locations of the violin and piano.

References

1. Chen, H., Xie, W., Afouras, T., Nagrani, A., Vedaldi, A., Zisserman, A.: Localizing visual sounds the hard way. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16867–16876 (2021)
2. Duke, B., Ahmed, A., Wolf, C., Aarabi, P., Taylor, G.W.: Sstvos: Sparse spatiotemporal transformers for video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5912–5921 (2021)
3. Qian, R., Hu, D., Dinkel, H., Wu, M., Xu, N., Lin, W.: Multiple sound sources localization from coarse to fine. In: Proceedings of the European conference on computer vision (ECCV). pp. 292–308 (2020)
4. Zhang, J., Xie, J., Barnes, N., Li, P.: Learning generative vision transformer with energy-based latent space for saliency prediction. Advances in Neural Information Processing Systems **34** (2021)