Unraveling Instance Associations: A Closer Look for Audio-Visual Segmentation (Supplementary Material)

Yuanhong Chen [1] * Yuyuan Liu [1] * Hu Wang [1] Fengbei Liu [1]
Chong Wang [1] Helen Frazer[2] Gustavo Carneiro[3]
[1] Australian Institute for Machine Learning, University of Adelaide
[2] St Vincent's Hospital Melbourne
[3] Centre for Vision, Speech and Signal Processing, University of Surrey

1. Further Details for VPO Dataset

1.1. Dataset Statistics

We show the distribution of visual classes in VPO-SS, VPO-MS and VPO-MSMI in Figure 1. Similar to the AVSBench-Semantics [29], we also observe a data imbalance issue within our VPO dataset. We follow [25] to report an imbalance ratio ($\frac{N_{\text{max}}}{N_{\text{min}}}$) of 12.48% (female & zebra), 12.43% (female & zebra) and 12.62% (female & cow) on the three VPO subsets, and 59.57% (man & axe & missile-rocket) on AVSBench-Semantics [29]. These class imbalance issues can affect the model performance during testing, which will be discussed in Sec. 3.3. For the demonstration of training examples, please refer to the "**video_demo.mp4**" file within the supplementary materials.

1.2. Creation Procedure

We show a graphical illustration of our Visual Post-production (VPO) benchmark in Fig. 2. We divide the entire dataset generation process into three major steps:

- **Data collection**: We gather datasets from off-the-shelf segmentation datasets (e.g., COCO [15]) and audio datasets (e.g., VGGSound [2]), focusing on the overlapping classes listed in Tab. 1. We randomly match audio and video files to form new samples based on their semantic labels.
- **Data processing**: We prioritise the collection of images with multiple objects and incorporate spatial location information based on each selected instance mask.
- **Subset creation**: We organize subsets according to their keywords (e.g., single-source, multi-sources, multi-instances) and further partition each subset into training and testing sets.

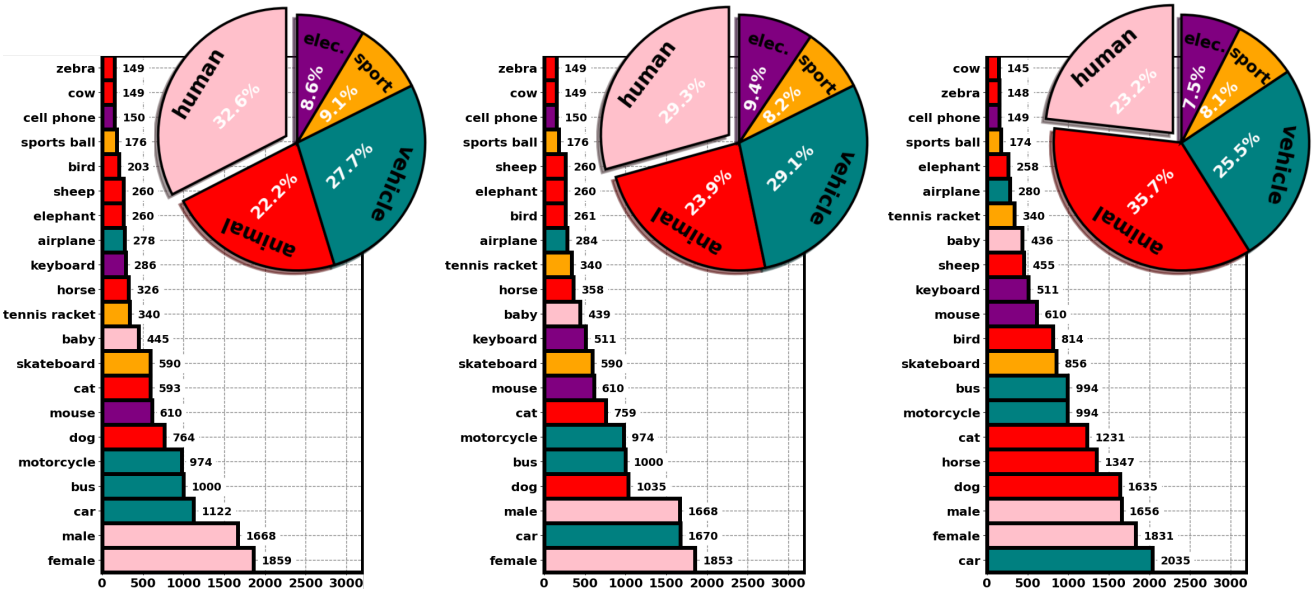
*First two authors contributed equally to this work.

Table 1. Detailed lookup table correlating audio tags with visual labels in the VPO creation process.

Visual-label	Audio-label
bird	mynah bird singing
keyboard	typing on computer keyboard
bus	driving buses
cat	cat purring / cat purring cat meowing / cat caterwaulin
dog	dog growling / dog bow-wow dog whimpering / dog howling dog barking / dog baying
horse	horse neighing / horse clip-clop
car	car passing by / car engine idling car engine starting race car, auto racing car engine knocking
sports ball	shot football
airplane	airplane / airplane flyby
sheep	sheep bleating
cow	cow lowing
motorcycle	driving motorcycle
mouse	mouse clicking
cell phone	cell phone buzzing
elephant	elephant trumpeting
zebra	zebra braying
tennis racket	playing tennis
skateboard	skateboarding
male	male speech, man speaking male singing
female	female speech, woman speaking
baby	baby babbling / baby crying baby laughter

1.3. Strengths and Weaknesses of VPO

As we discussed in the main paper, our VPO dataset enjoys the following **strengths**:



(a) VPO-SS (b) VPO-MS (c) VPO-MSMI

Figure 1. Visual class distribution in our proposed VPO-SS, VPO-MS and VPO-MSMI benchmark datasets.

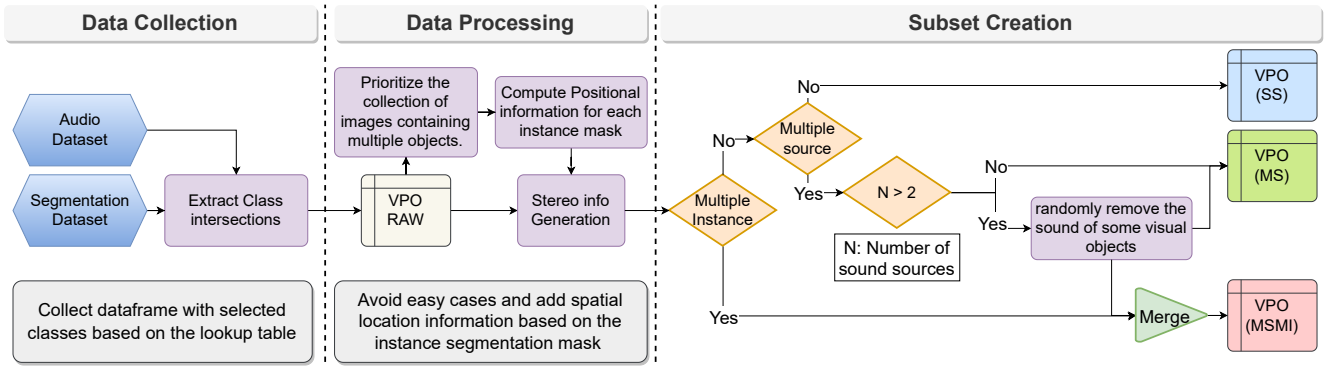


Figure 2. Illustration of VPO collection and production process.

- *Cost-effectiveness*. The manually labelled AVS dataset needs annotators to watch and listen to an entire video so they can provide labels. The VPO production process can significantly reduce such annotation costs.
- *Data scalability*. We can easily increase the amount of data by leveraging existing visual (ADE20K [27], Pascal VOC [6], etc.) and audio (AudioSet [8], ESC-50 [23], etc.) datasets.
- *More diverse scenarios.* In our proposed VPO datasets, every object within the scene will have a chance to be the sounding object, which is crucial to mitigate the "commonsense" bias that is observed in AVSBench-Semantics [29] and addressing the assessment of spurious correlation [1, 5, 19].
- *Isolation of motion information.* The disentanglement of motion and sound in our VPO benchmark prevents the model from solely relying on motion information to make predictions, encouraging the learning and evaluation of cross-modal alignment [24].
- *Introduction of stereo audio.* The use of stereo audio encourages the study of spatial prompts.

We also identified the following **weaknesses** of VPO:

- *Data imbalance issue*. Such imbalance can affect the segmentation accuracy, particularly for the tail classes.
- *Lack of temporal image data.* Since we match still images with audio, we cannot use image motion information.
- *Comprehensive simulations of spatial audio.* Our VPO does not include the modelling of arrival time differences and microphone distance.

2. Experiments

2.1. Implementation Details

Training & Inference: During training, we apply data augmentation for image inputs with colour jitter, horizon-

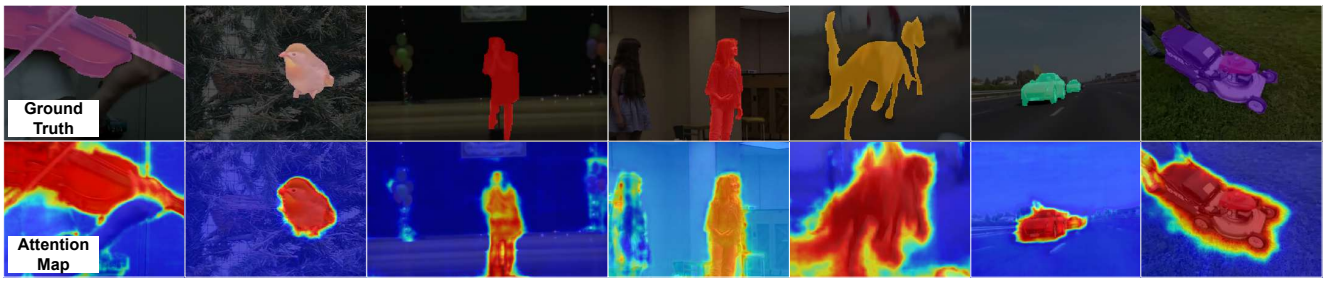
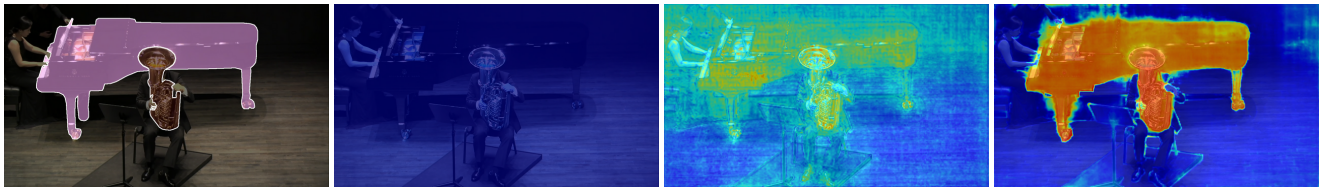
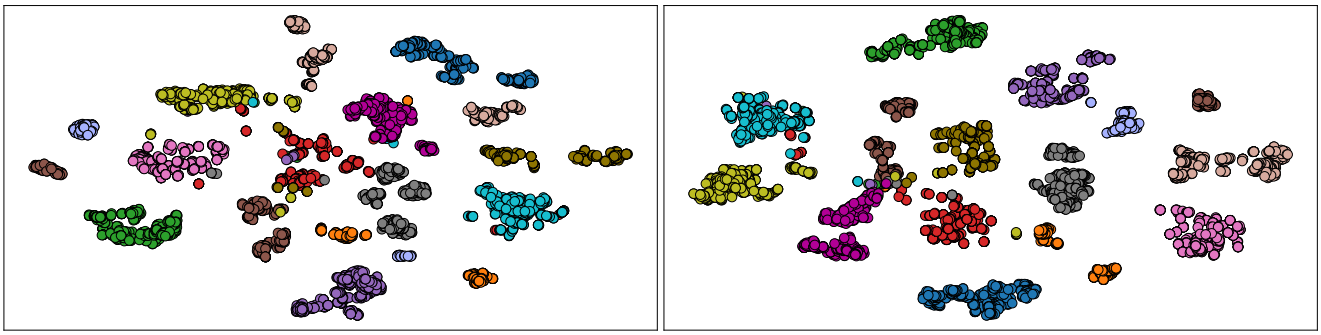


Figure 3. Qualitative results for cross-attention heatmap on AVSBench-Semantics [29].



(a) Ground Truth (b) Softmax (c) Min-max (d) Sigmoid

Figure 4. Visualisation of attention map process by Softmax and Sigmoid activation.



(a) Without CAVP (b) With CAVP

Figure 5. T-SNE visualisation of features after cross-attention fusion layer, trained with or without our CAVP.

tal flipping and random scaling between 0.5 to 2.0. We randomly crop images to 512×512 pixels. For the audio data, we extract the log Mel-Spectrogram with 64 mel filter banks over 1s [28] or 3s [22] of the waveform at 16 kHz on AVSBench and VPO. We set temperature τ as 0.1. We use the SGD optimizer with a momentum of 0.9, weight decay of 0.0005, and a polynomial learning-rate decay $(1 - \frac{\text{iter}}{\text{total\_iter}})^{\text{power}}$ with power $= 0.9$. The initial learning rate is set to 0.001, the mini-batch size is 16 and training lasts for 80 epochs. During inference, we use the original resolution resizing and cropping with a mini-batch size of 1. We adopt two image backbones (ResNet [10], PVT-V2-B5 [26]) and DeepLabV3+ [4] for the segmentation network. For the audio backbones, we use VGGish [12] (following [28]) and ResNet-18 [10] (following [3, 22]) for AVSBench and VPO, respectively. The overwhelming amount of negative samples is mitigated by maintaining a memory bank [11] to store raw waveform data for each class. During training, we proportionally transfer negative pairs from the negative set to the positive set by pairing positive audio with respect to the image label.

3. Attention Map Visualisation

To demonstrate the effectiveness of our cross-attention module, we visualize the audio-visual attention heatmap. We employ models pre-trained on AVSBench-Semantics [29], utilizing a full-resolution set-up and equipped with a ResNet50 [10] backbone. As shown in Fig. 3, the heatmaps illustrate that our module can effectively retrieve the foreground object by leveraging the interaction between audio and visual embeddings. Additionally, we show a visual comparison amount of the application of $\mathrm{softmax}(.)$, $\mathrm{Minmax}(.)$ and $\mathrm{sigmoid}(.)$ activation functions. As depicted in Fig. 4, using Softmax under spatial dimension may lead to a diminished attention map (Fig. 4b), while incorporating Minmax activation over the dot product of the audio-visual feature could yield a noisy and limited discernment of relevant audio-visual correspondences. (Fig. 4c). Our method in Fig. 4d, utilizing $\mathrm{sigmoid}(.)$, demonstrates a better efficacy compared to these two methods in terms of cross-modal feature activation.

Table 2. Quantitative (mIoU, F_β) audio-visual segmentation results (in %) on AVSBench dataset [28, 29] (resized to 224×224) with PVT-V2-B5 [26] backbone. The best results are in **bold** and the second best are underlined.

PVT-V2-B5 [26]	Method	AVSBench-Object (SS)		AVSBench-Object (MS)		AVSBench-Semantics	
		mIoU ↑	F_β ↑	mIoU ↑	F_β ↑	mIoU ↑	F_β ↑
Transformer	CATR [14]	81.40	89.60	59.00	70.00	32.80	38.50
	AuTR [18]	80.40	89.10	56.20	67.20	-	-
	AQFormer [13]	81.60	89.40	61.10	72.10	-	-
	AVSegFormer [7]	82.06	89.90	58.36	69.30	36.66	42.00
	AVSC [16]	80.57	88.19	58.22	65.10	-	-
	BAVS [17]	81.96	88.60	58.63	65.49	32.64	36.42
Per-pixel Classification	TPAVI [28]	78.74	87.90	54.00	64.50	29.77	35.20
	AVSBG [9]	81.71	90.40	55.10	66.80	-	-
	ECMVAE [21]	81.74	90.10	57.84	70.80	-	-
	DiffusionAVS [20]	81.38	90.20	58.18	70.90	-	-
	Ours	**87.33** (+5.27)	**93.61** (+3.41)	**67.31** (+6.21)	**78.09** (+5.99)	**48.59** (+11.93)	**61.97** (+19.97)

Table 3. Quantitative (mIoU, F_β, FDR) audio-visual segmentation results (in %) on VPO dataset with ResNet50 [10] backbone and mono audio. Best results are in **bold**, and second best areunderlined. Improvements against the second best are in the last row.

D-ResNet50 [10]	Method	VPO (SS)			VPO (MS)			VPO (MSMI)		
		mIoU ↑	F_β ↑	FDR ↓	mIoU ↑	F_β ↑	FDR ↓	mIoU ↑	F_β ↑	FDR ↓
Transformer	AVSegFormer [7]	52.96	67.89	25.50	56.46	71.89	24.66	50.96	64.96	32.72
Per-pixel Classification	TPAVI [28]	51.84	68.77	23.64	44.08	58.14	30.82	50.37	66.80	29.82
	Ours	**61.48**	**75.53**	**18.79**	**61.85**	**74.60**	**20.24**	**57.22**	**72.26**	**24.04**
Improvements	**Ours**	**+8.52**	**+6.76**	**-4.85**	**+5.39**	**+2.71**	**-4.42**	**+6.26**	**+5.46**	**-5.78**

3.1. T-SNE Visualisation

To present the qualitative results of our CAVP method in the latent space, we extract the features computed before the classification layers, and generate T-SNE plots in Fig. 5. We employ models pre-trained on AVSBench-Semantics [29], utilizing a full-resolution set-up and equipped with a ResNet50 [10] backbone. We consider both scenarios: with the proposed CAVP method (Fig. 5b) and without it (Fig. 5a). The results demonstrate that our method can enhance intra-class compactness while preserving intra-class separability.

3.2. Additional Results on Resized AVSBench

We follow previous methods [7, 9, 13, 14, 16–18, 20, 21, 28] to evaluate our model with PVT-V2-B5 [26] backbone on AVSBench-Objects (SS & MS) [28] and AVSBench-Semantics [29] with resized image resolution (224×224). The results in Tab. 2 show that we improve mIoU by 5.27%, 6.21% and 11.93% on the respective three benchmarks.

3.3. Additional Results

We present supplementary results for the paper, showcasing the performance of AVSegformer [7], TPAVI [28], and our model on **VPO with mono audio** with ResNet50 [10] backbone, as depicted in Tab. 3. Our method outperforms the baseline methods on all experimental settings by a minimum of 5.39%, 2.71% and 4.42% for mIoU, F_β score and false detection rate, respectively. We also provide **class-wise results** on AVSBench-Semantics [29] in Tables 4,5,6. We observe that the tail classes, such as clippers, axe, missile-rocket and utv, show significantly worse results than the remaining classes, illustrating the importance of addressing the imbalance issue in the AVS task.

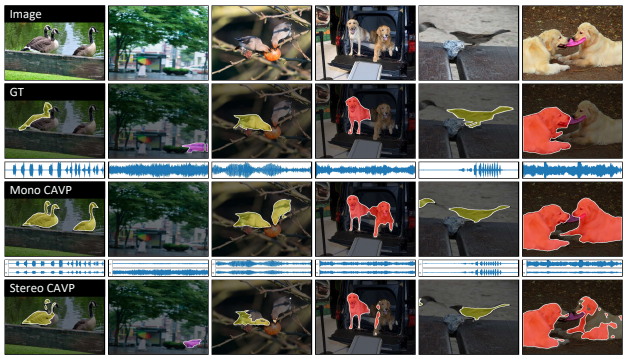


Figure 6. Visual comparison between CAVP models trained with mono audio and stereo audio.

To showcase the model's performance with both mono and stereo audio inputs, we opted for two CAVP models utilizing the ResNet50 [10] backbone. Trained on VPO-MSMI, these models were chosen to assess prediction outcomes in multi-instance scenarios, illustrated in Fig. 6. Our observations reveal that, when aided by stereo audio, the model effectively diminishes its focus on the incorrect spatial direction. However, challenges persist in handling images with multiple instances, exemplified in the last column (depicting two dogs) of Fig. 6. This underscores a significant challenge in the audio-visual segmentation (AVS) task.

To further demonstrate the effectiveness of our VPO dataset, we used the TPAVI [28] trained on VPO and applied

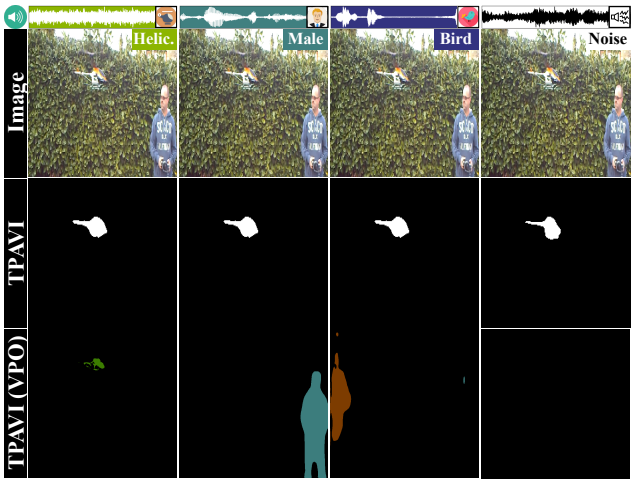


Figure 7. Visual comparison between TPAVI models trained on AVSBench [28] (2nd row) and VPO (3rd row), using an identical set of synthetic test samples. The columns display the original audio (1st column) and alternative sound types, including a person speaking (2nd column), bird chirping (3rd column), or background noise (4th column).

it to test the synthetic examples, shown in Fig. 7. Despite hallucinations persisting in the 3rd column (class "Bird"), there is a noticeable improvement, with the correct segment of "Helicopter", "Male", and "Noise", compared to the results in the 2nd row.

Finally, we show a qualitative comparison visualization among TPAVI [28], AVSegFormer [7], and our CAVP on VPO in Fig. 8 and on AVSBench-Semantics in Fig. 9. We demonstrate that our method consistently provides a more effective approximation of the true segmentation of objects in the scene compared to alternative methods. For the demonstration of full video examples on AVSBench-Semantics, please refer to the "**video_demo.mp4**" file within the attached supplementary materials.

References

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 2

[2] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020. 1

[3] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16867–16876, 2021. 3

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 3

[5] Volkan Cirik, Louis-Philippe Morency, and Taylor Berg-Kirkpatrick. Visual referring expression recognition: What do systems actually learn? *arXiv preprint arXiv:1805.11818*, 2018. 2

[6] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010. 2

[7] Shengyi Gao, Zhe Chen, Guo Chen, Wenhai Wang, and Tong Lu. Avsegformer: Audio-visual segmentation with transformer. *arXiv preprint arXiv:2307.01146*, 2023. 4, 5, 6, 7

[8] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017. 2

[9] Dawei Hao, Yuxin Mao, Bowen He, Xiaodong Han, Yuchao Dai, and Yiran Zhong. Improving audio-visual segmentation with bidirectional generation. *arXiv preprint arXiv:2308.08288*, 2023. 4

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3, 4, 8

[11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 3

[12] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017. 3

[13] Shaofei Huang, Han Li, Yuqing Wang, Hongji Zhu, Jiao Dai, Jizhong Han, Wenge Rong, and Si Liu. Discovering sounding objects by audio queries for audio visual segmentation. *arXiv preprint arXiv:2309.09501*, 2023. 4

[14] Kexin Li, Zongxin Yang, Lei Chen, Yi Yang, and Jun Xun. Catr: Combinatorial-dependence audio-queried transformer for audio-visual video segmentation. *arXiv preprint arXiv:2309.09709*, 2023. 4

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1

[16] Chen Liu, Peike Li, Xingqun Qi, Hu Zhang, Lincheng Li, Dadong Wang, and Xin Yu. Audio-visual segmentation by exploring cross-modal mutual semantics, 2023. 4

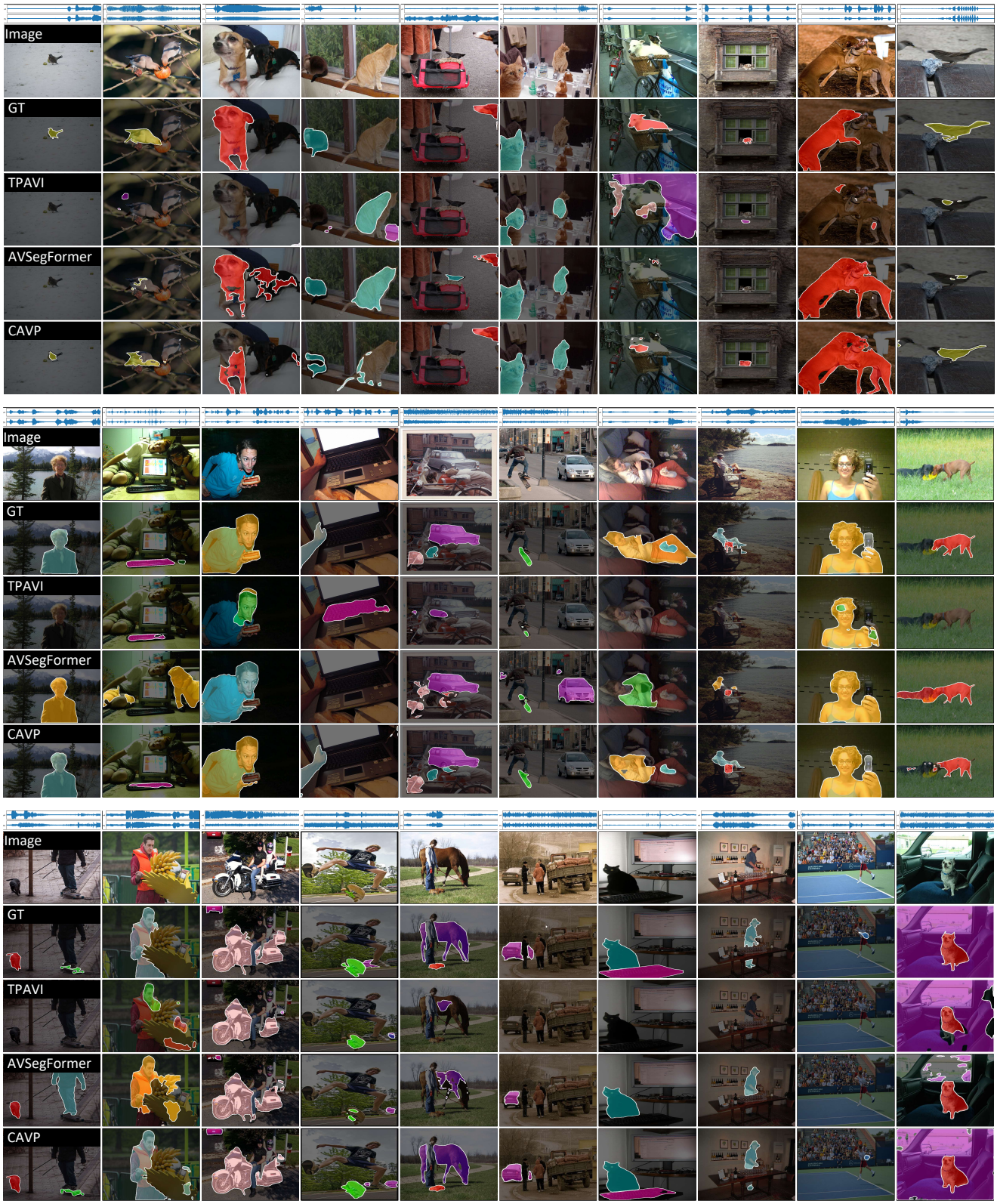


Figure 8. Qualitative audio-visual segmentation results on VPO by TPAVI [28], AVSegFormer [7], and our CAVP. The prediction results can be compared with the ground truth (GT) of the first row of each sample.

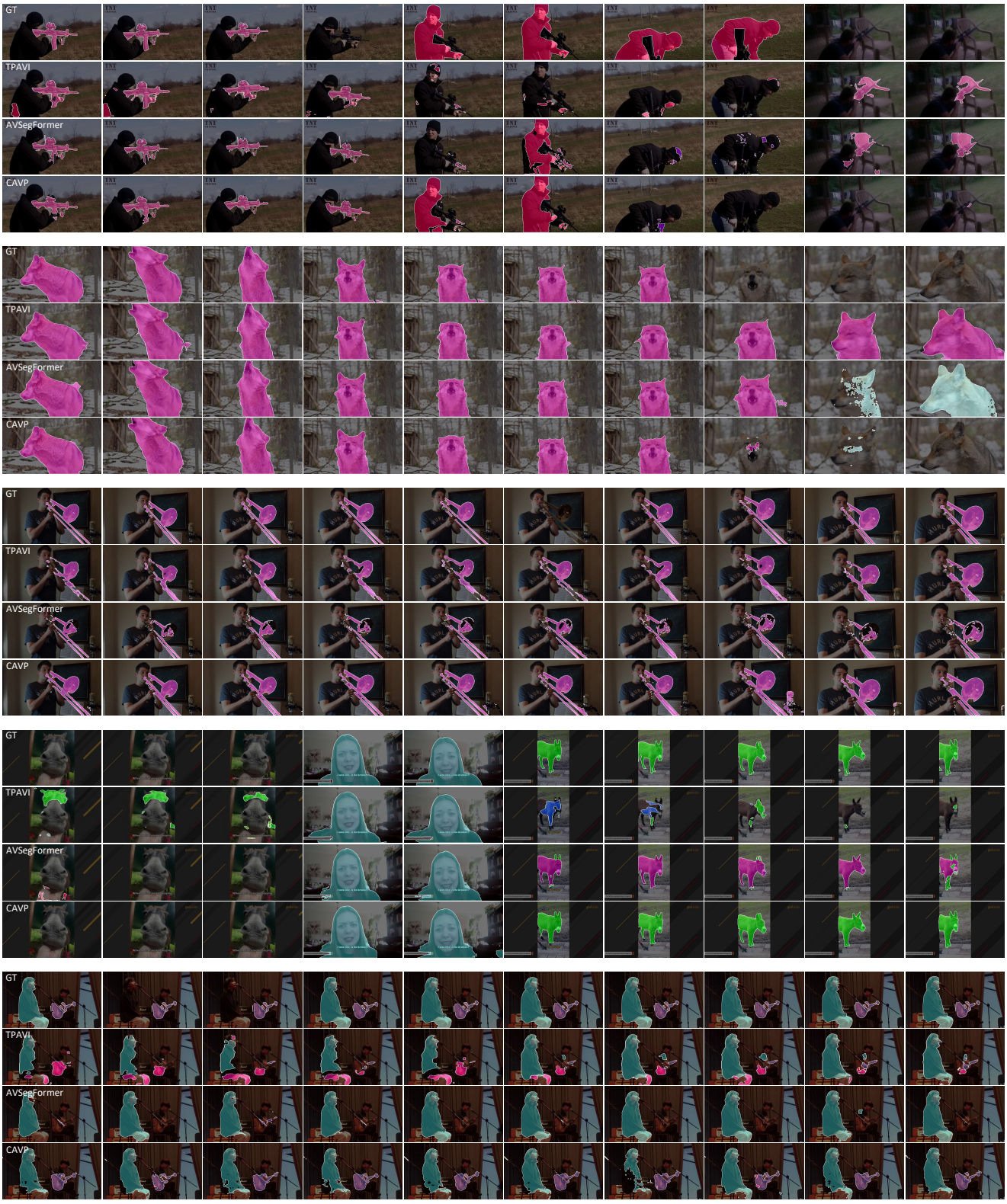


Figure 9. Qualitative audio-visual segmentation results on AVSBench-Semantics [29] by TPAVI [28], AVSegFormer [7], and our CAVP. The prediction results can be compared with the ground truth (GT) of the first row of each video.

Table 4. Class-level (mIoU) audio-visual segmentation results (in %) on AVSBench-Semantic dataset [29] (original resolution) with ResNet50 [10] backbone.

Class	background	accordion	airplane	axe	baby	bassoon	bell	bird	boat	boy	bus	car	cat	cello	clarinet
AVSegformer	0.9093	0.8610	0.9003	0.0000	0.4178	0.1683	0.0618	0.2251	0.7751	0.0130	0.4240	0.2800	0.3952	0.4553	0.0033
TPAVI	0.9111	0.7484	0.9261	0.0000	0.4147	0.1578	0.0795	0.3814	0.7149	0.0251	0.7019	0.4214	0.5209	0.4537	0.1624
CAVP	0.9168	0.9229	0.9291	0.0000	0.4212	0.2580	0.3376	0.2481	0.8096	0.1194	0.3486	0.3786	0.4968	0.5490	0.0180
Class	clipper	clock	dog	donkey	drum	duck	elephant	emergency-car	erhu	flute	frying-food	girl	goose	guitar	
AVSegformer	0.0000	0.5482	0.2847	0.3876	0.0288	0.5581	0.8649	0.5377	0.2396	0.2287	0.5187	0.1163	0.2171	0.7520	
TPAVI	0.0000	0.4179	0.2329	0.1913	0.3180	0.4620	0.7492	0.5342	0.3697	0.3134	0.4161	0.1737	0.0054	0.7203	
CAVP	0.0000	0.7088	0.3259	0.4141	0.2987	0.5154	0.7597	0.6329	0.5898	0.3358	0.4761	0.1691	0.6841	0.8172	
Class	gun	guzheng	hair-dryer	handpan	harmonica	harp	helicopter	hen	horse	keyboard	leopard	lion	man	marimba	
AVSegformer	0.2456	0.4480	0.4926	0.8611	0.0000	0.6583	0.6952	0.1423	0.1709	0.7823	0.6721	0.7946	0.4050	0.8194	
TPAVI	0.2355	0.4999	0.6422	0.7801	0.0000	0.5573	0.7185	0.2600	0.3104	0.7592	0.6018	0.7331	0.4229	0.8223	
CAVP	0.3059	0.5587	0.6075	0.8885	0.0000	0.6473	0.7762	0.7078	0.2506	0.7800	0.6797	0.8328	0.4300	0.8392	
Class	missile-rocket	motorcycle	mower	parrot	piano	pig	pipa	saw	saxophone	sheep	sitar	sorna	squirrel	tabla	
AVSegformer	0.0000	0.0501	0.6662	0.1493	0.6031	0.2891	0.5213	0.2953	0.3404	0.0837	0.4364	0.2090	0.4190	0.6088	
TPAVI	0.1086	0.0106	0.6237	0.1239	0.5406	0.2505	0.6602	0.4154	0.4560	0.0507	0.5225	0.6431	0.4567	0.6963	
CAVP	0.0000	0.1063	0.6487	0.1158	0.6178	0.6247	0.6518	0.5625	0.5141	0.1069	0.6552	0.4723	0.6357	0.6383	
Class	tank	tiger	tractor	train	trombone	truck	trumpet	tuba	ukulele	utv	vacuum-cleaner	violin	wolf	woman	
AVSegformer	0.3381	0.5467	0.4995	0.8996	0.4133	0.2434	0.3514	0.7967	0.5930	0.0000	0.1720	0.4008	0.7839	0.3837	
TPAVI	0.4718	0.5849	0.4586	0.8529	0.4425	0.1831	0.2631	0.7744	0.5388	0.0000	0.1558	0.5302	0.6924	0.4285	
CAVP	0.6514	0.6226	0.4922	0.9427	0.6415	0.1948	0.3839	0.8249	0.7038	0.0000	0.5680	0.6005	0.8510	0.4157	

Table 5. Class-level (F_β) audio-visual segmentation results (in %) on AVSBench-Semantic dataset [29] (original resolution) with ResNet50 [10] backbone.

Class	background	accordion	airplane	axe	baby	bassoon	bell	bird	boat	boy	bus	car	cat	cello	clarinet
AVSegformer	0.9484	0.9395	0.9488	0.0000	0.5795	0.4054	0.1657	0.2971	0.8914	0.0351	0.5099	0.3539	0.5359	0.6597	0.0129
TPAVI	0.9445	0.8090	0.9581	0.0000	0.6223	0.4018	0.2440	0.5241	0.8458	0.0786	0.8474	0.5883	0.7307	0.6560	0.3185
CAVP	0.9528	0.9579	0.9505	0.0000	0.6393	0.5019	0.5094	0.3887	0.8846	0.2930	0.4237	0.4606	0.6165	0.7143	0.0535
Class	clipper	clock	dog	donkey	drum	duck	elephant	emergency-car	erhu	flute	frying-food	girl	goose	guitar	
AVSegformer	0.0000	0.7194	0.4178	0.6265	0.0956	0.7510	0.9100	0.6769	0.5244	0.3931	0.7594	0.2153	0.4183	0.8626	
TPAVI	0.0000	0.6666	0.4233	0.4442	0.6056	0.6664	0.8798	0.7391	0.6092	0.5772	0.7110	0.3057	0.0105	0.8266	
CAVP	0.0000	0.8454	0.5669	0.6258	0.5230	0.7024	0.8076	0.7269	0.7465	0.6553	0.7321	0.2821	0.8127	0.9053	
Class	gun	guzheng	hair-dryer	handpan	harmonica	harp	helicopter	hen	horse	keyboard	leopard	lion	man	marimba	
AVSegformer	0.4728	0.7464	0.5995	0.9431	0.0000	0.7688	0.8334	0.3946	0.2563	0.8703	0.7653	0.8616	0.5782	0.9098	
TPAVI	0.5106	0.7231	0.7612	0.9119	0.0000	0.7621	0.8379	0.5141	0.4684	0.8806	0.7856	0.8201	0.6245	0.9119	
CAVP	0.5433	0.7352	0.6854	0.9427	0.0000	0.7616	0.8725	0.8698	0.3430	0.8783	0.7884	0.8958	0.6217	0.9223	
Class	missile-rocket	motorcycle	mower	parrot	piano	pig	pipa	saw	saxophone	sheep	sitar	sorna	squirrel	tabla	
AVSegformer	0.0000	0.1409	0.8527	0.3047	0.7714	0.3990	0.7633	0.4944	0.5863	0.2197	0.7148	0.4113	0.6475	0.7012	
TPAVI	0.2028	0.0316	0.8035	0.2810	0.7535	0.4973	0.8320	0.6863	0.7137	0.1161	0.7707	0.8317	0.6656	0.7887	
CAVP	0.0000	0.2981	0.8605	0.2403	0.8134	0.7346	0.8311	0.8037	0.7085	0.2074	0.8023	0.5639	0.7770	0.7387	
Class	tank	tiger	tractor	train	trombone	truck	trumpet	tuba	ukulele	utv	vacuum-cleaner	violin	wolf	woman	
AVSegformer	0.6493	0.6707	0.6241	0.9607	0.6450	0.5623	0.6298	0.9178	0.7482	0.0000	0.2465	0.5906	0.8749	0.5682	
TPAVI	0.7202	0.7762	0.5412	0.9239	0.6158	0.4848	0.4430	0.8686	0.7479	0.0000	0.2122	0.7115	0.8528	0.6369	
CAVP	0.8315	0.7188	0.5650	0.9710	0.7491	0.4830	0.6486	0.8840	0.8353	0.0000	0.6977	0.7634	0.9193	0.6123	

Table 6. Class-level (FDR) audio-visual segmentation results (in %) on AVSBench-Semantic dataset [29] (original resolution) with ResNet50 [10] backbone.

Class	background	accordion	airplane	axe	baby	bassoon	bell	bird	boat	boy	bus	car	cat	cello	clarinet
AVSegformer	0.0550	0.0479	0.0502	0.0000	0.4288	0.3772	0.7397	0.7448	0.0926	0.9487	0.5460	0.6959	0.4878	0.3081	0.9300
TPAVI	0.0631	0.2274	0.0449	0.0000	0.3432	0.3230	0.4422	0.4977	0.1435	0.8376	0.1323	0.4156	0.2250	0.3140	0.6382
CAVP	0.0505	0.0437	0.0601	0.0000	0.3147	0.3791	0.4865	0.6186	0.1240	0.5693	0.6331	0.5954	0.4191	0.2809	0.9047
Class	clipper	clock	dog	donkey	drum	duck	elephant	emergency-car	erhu	flute	frying-food	girl	goose	guitar	
AVSegformer	1.0000	0.2708	0.6018	0.3006	0.7584	0.2165	0.1045	0.3412	0.2448	0.5871	0.1603	0.7783	0.5090	0.1339	
TPAVI	1.0000	0.2492	0.5280	0.3382	0.2250	0.3010	0.0993	0.2199	0.3155	0.2966	0.1332	0.6854	0.9897	0.1824	
CAVP	1.0000	0.1405	0.3474	0.3350	0.4075	0.2775	0.2348	0.3099	0.2496	0.1147	0.1721	0.7238	0.1870	0.0896	
Class	gun	guzheng	hair-dryer	handpan	harmonica	harp	helicopter	hen	horse	keyboard	leopard	lion	man	marimba	
AVSegformer	0.4300	0.0935	0.4443	0.0412	0.0000	0.2514	0.1549	0.2107	0.7680	0.1360	0.2649	0.1580	0.4202	0.0822	
TPAVI	0.2801	0.2203	0.2559	0.0553	1.0000	0.1932	0.1607	0.3488	0.5360	0.1038	0.1826	0.2010	0.3472	0.0798	
CAVP	0.3705	0.2483	0.3653	0.0559	0.0000	0.2581	0.1288	0.0918	0.6947	0.1201	0.2287	0.1150	0.3599	0.0692	
Class	missile-rocket	motorcycle	mower	parrot	piano	pig	pipa	saw	saxophone	sheep	sitar	sorna	squirrel	tabla	
AVSegformer	1.0000	0.7616	0.0958	0.6423	0.2116	0.6355	0.1542	0.4672	0.3244	0.6554	0.1578	0.5089	0.2943	0.3404	
TPAVI	0.7909	0.9448	0.1636	0.6326	0.1957	0.3731	0.1337	0.1972	0.1895	0.8594	0.1386	0.1214	0.2973	0.2370	
CAVP	1.0000	0.4343	0.0645	0.7223	0.1387	0.2925	0.1293	0.1074	0.2643	0.7786	0.1885	0.4891	0.2233	0.2929	
Class	tank	tiger	tractor	train	trombone	truck	trumpet	tuba	ukulele	utv	vacuum-cleaner	violin	wolf	woman	
AVSegformer	0.1411	0.3576	0.4080	0.0274	0.2927	0.1022	0.2312	0.0539	0.2487	1.0000	0.7833	0.3927	0.1284	0.4196	
TPAVI	0.1946	0.1879	0.5165	0.0732	0.3823	0.0581	0.5315	0.1349	0.2057	1.0000	0.8205	0.2718	0.1152	0.3276	
CAVP	0.1282	0.3183	0.4969	0.0285	0.2766	0.1775	0.2415	0.1326	0.1567	1.0000	0.3237	0.2250	0.0810	0.3646	

[17] Chen Liu, Peike Li, Hu Zhang, Lincheng Li, Zi Huang, Dadong Wang, and Xin Yu. Bavs: Bootstrapping audio-visual segmentation by integrating foundation knowledge. *arXiv preprint arXiv:2308.10175*, 2023. 4

[18] Jinxiang Liu, Chen Ju, Chaofan Ma, Yanfeng Wang, Yu Wang, and Ya Zhang. Audio-aware query-enhanced transformer for audio-visual segmentation. *arXiv preprint arXiv:2307.13236*, 2023. 4

[19] Runtao Liu, Chenxi Liu, Yutong Bai, and Alan L Yuille. Clevr-ref+: Diagnosing visual reasoning with referring ex-

pressions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4185–4194, 2019. 2

[20] Yuxin Mao, Jing Zhang, Mochu Xiang, Yunqiu Lv, Yiran Zhong, and Yuchao Dai. Contrastive conditional latent diffusion for audio-visual segmentation. *arXiv preprint arXiv:2307.16579*, 2023. 4

[21] Yuxin Mao, Jing Zhang, Mochu Xiang, Yiran Zhong, and Yuchao Dai. Multimodal variational auto-encoder based audio-visual segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 954–965, 2023. 4

[22] Shentong Mo and Pedro Morgado. A closer look at weakly-supervised audio-visual source localization. *arXiv preprint arXiv:2209.09634*, 2022. 3

[23] Karol J Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018, 2015. 2

[24] Arda Senocak, Hyeonggon Ryu, Junsik Kim, Tae-Hyun Oh, Hanspeter Pfister, and Joon Son Chung. Sound source localization is all about cross-modal alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7777–7787, 2023. 2

[25] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. *Advances in Neural Information Processing Systems*, 33:1513–1524, 2020. 1

[26] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022. 3, 4

[27] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 2

[28] Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio–visual segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, pages 386–403. Springer, 2022. 3, 4, 5, 6, 7

[29] Jinxing Zhou, Xuyang Shen, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, et al. Audio-visual segmentation with semantics. *arXiv preprint arXiv:2301.13190*, 2023. 1, 2, 3, 4, 7, 8