Stepping Stones: A Progressive Training Strategy for Audio-Visual Semantic Segmentation (Supplementary Material)

Juncheng Ma[1], Peiwen Sun[2], Yaoting Wang[3], and Di Hu[✉3]

[1] University of Chinese Academy of Sciences
majuncheng21@mails.ucas.ac.cn
[2] Beijing University of Posts and Telecommunications
sunpeiwen@bupt.edu.cn
[3] Gaoling School of Artificial Intelligence, Renmin University of China, Beijing
yaoting.wang@outlook.com
dihu@ruc.edu.cn

1 More Qualitative Comparison

In this section, we conduct a more comprehensive qualitative comparison across all three subtasks, contrasting our approach with AVSbench [6] and AVSegformer [2]. For the S4 task, illustrated in Fig. A1, all three methods demonstrate satisfactory performance, yet ours exhibits slight superiority in terms of segmentation accuracy and audio-visual alignment. For the MS3 task, depicted in Fig. A2, our method notably surpasses other approaches in sound source determination and segmentation accuracy. Finally, for the AVSS task, as depicted in Fig. A3, our approach markedly outperforms prior methods in both establishing audio-visual correspondence and semantic comprehension.

2 Generalization of Stepping Stones Training Strategy

To further validate the generalization of our approach, we present some results to illustrate the enhancement on AVSBench and AVSegformer through the *Stepping Stones* strategy. As depicted in Figs. B4 and B5, the application of the *Stepping Stones* strategy yields significant improvements on both audio-visual alignment and semantic comprehension.

In the main paper, we simulated three levels of accuracy for the first-stage results during experiments to validate the generalization of *Stepping Stones*. These levels, termed *low*, *high* and *oracle*, correspond to first-stage results with IoU values of 77.10%, 83.65%, and 100% (averaged across S4, MS3, and V2 subset), respectively. Figs. B4 and B5 present results obtained at *high* level.

✉Corresponding author.

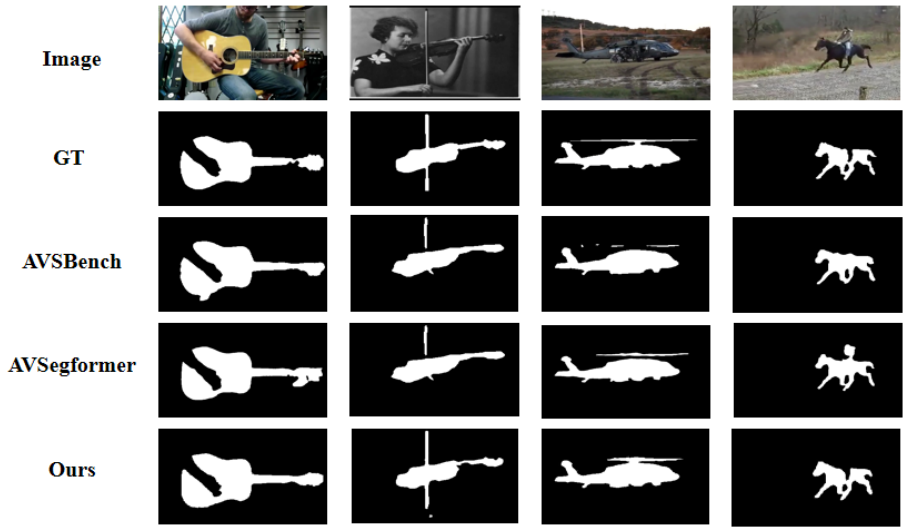

Fig. A1: Qualitative comparison of the S4 task ignoring the presentation of input audio.

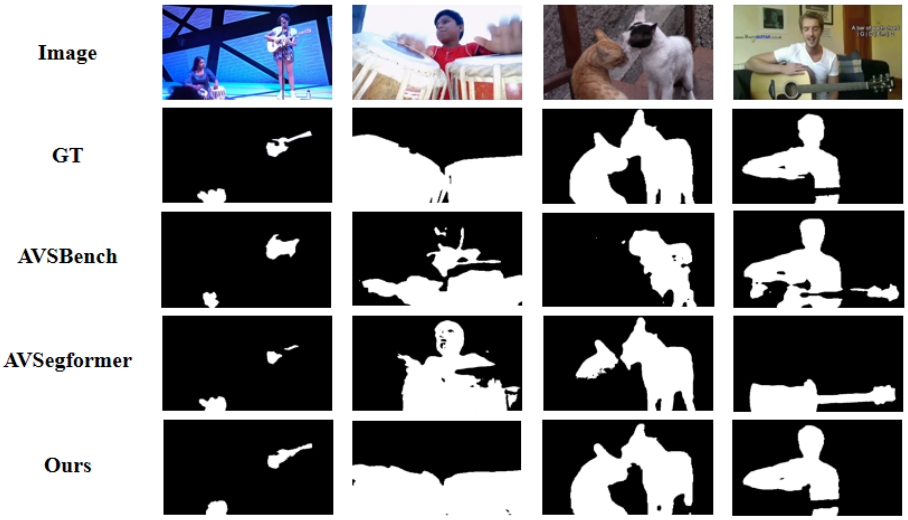

Fig. A2: Qualitative comparison of the MS3 task ignoring the presentation of input audio.

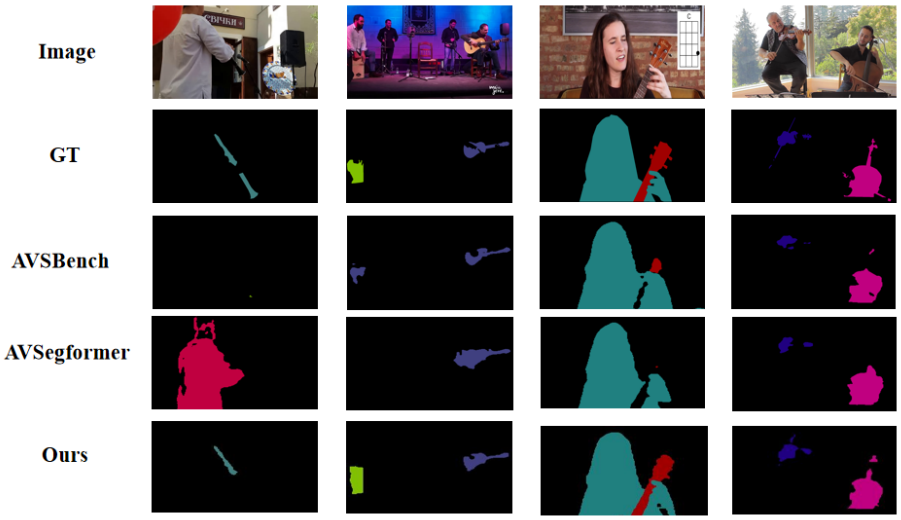

Fig. A3: Qualitative comparison of the AVSS task ignoring the presentation of input audio. *Ours* donates the AAVS model with *Stepping Stones* strategy.

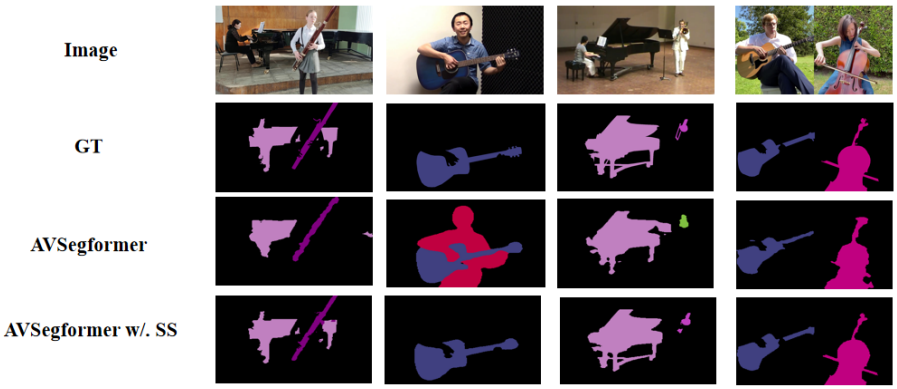

Fig. B4: Effectiveness of the *Stepping Stones* strategy on AVSegformer. "SS" denotes the abbreviation for *Stepping Stones* here.

Fig. B5: Effectiveness of the *Stepping Stones* strategy on AVSBench. "SS" denotes the abbreviation for *Stepping Stones* here.

3 Effectiveness of Stepping Stones Training Strategy

We present a comparison of results obtained before and after applying *Stepping Stones* strategy to the AAVS model to further validate its effectiveness, as depicted in Fig. C6. Notably, in the second column, despite the inaccurate sound source localization information provided by the first-stage results, the robustness of the second-stage model enables accurate final prediction of semantic mask. Consistent with the setting of ablation experiments, we utilize the first-stage results inferred from the trained AAVS model, with IoU of 83.18% for S4 labels, 67.50% for MS3 labels, and 72.77% for V2 labels.

4 Comparison with CAVP [1]

In a contemporaneous study, Chen, *et al.* [1] introduced a novel dataset for audio-visual segmentation and proposed an informative sample mining method based on contrastive learning. However, we observed their calculations of evaluation metrics (mIoU, F-score) during testing are fundamentally different from ours. For mIoU, CAVP accumulates the intersection and union for all categories across the test set, then divides and averages along the categories. In contrast, following the protocols of AVSBench [6], CATR [3], AVSegformer [2], GAVS [4], and MUTR [5], we accumulate the IoU for each category across the test set, divide by the total number of valid categories, and then average along the categories. These two calculation methods result in significant differences in mIoU and F-score.

To ensure a fair comparison, we compared the performance of CAVP and our method on three sub-tasks using the above two kinds of calculation methods, as shown in Tabs. D1 and D2. Among them, mIoU and F-score represent the evaluation metrics used in our paper, while mIoU* and F-score* represent the evaluation metrics used in CAVP. It is evident that regardless of calcula-

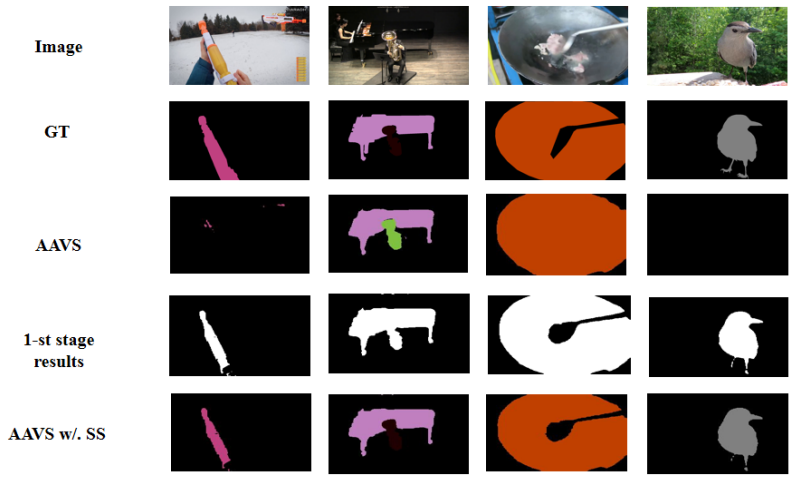

Fig. C6: Effectiveness of the *Stepping Stones* strategy on AAVS. "SS" denotes the abbreviation for *Stepping Stones* here.

tion methods, our method consistently outperforms CAVP. Further quantitative comparison can be seen in Fig. D7.

Table D1: Quantitative comparison on S4 and MS3 subtask.

Method	S4				MS3			
	mIoU	F-score	mIoU*	F-score*	mIoU	F-score	mIoU*	F-score*
CAVP [1]	60.45	70.90	87.20	93.40	43.48	51.07	67.57	77.80
Ours	**83.18**	**91.33**	**91.39**	**95.49**	**67.30**	**77.63**	**78.39**	**85.98**

Table D2: Quantitative comparision on AVSS subtask.

Method	mIoU	F-score	mIoU*	F-score*
CAVP [1]	35.21	39.34	50.75	64.33
Ours	**48.50**	**53.20**	**60.45**	**70.90**

5 More Ablation Study

The results of the experiments on the initialization method of mask attention in the transformer decoder are presented in Tab. E3. Audio initialization shows a slight improvement.

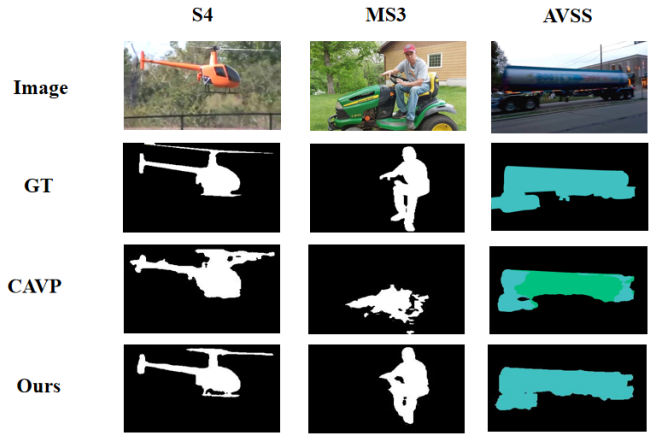

Fig. D7: Qualitative comparision with CAVP [1].

Table E3: Ablation Experiment of initialization of attention mask. "Actual" refers to pseudo labels inferred from the trained AAVS, while "Oracle" denotes the utilization of ground truth binary labels.

Method	Actual		Oracle	
	mIoU	F-score	mIoU	F-score
w/. Audio Initialization	43.43	48.33	51.19	56.19
Origin Initialization	41.79	47.18	-	-

References

1. Chen, Y., Liu, Y., Wang, H., Liu, F., Wang, C., Frazer, H., Carneiro, G.: Unraveling instance associations: A closer look for audio-visual segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26497–26507 (2024)
2. Gao, S., Chen, Z., Chen, G., Wang, W., Lu, T.: Avsegformer: Audio-visual segmentation with transformer (2024)
3. Li, K., Yang, Z., Chen, L., Yang, Y., Xiao, J.: Catr: Combinatorial-dependence audio-queried transformer for audio-visual video segmentation. In: ACM MM. pp. 1485–1494 (2023)
4. Wang, Y., Liu, W., Li, G., Ding, J., Hu, D., Li, X.: Prompting segmentation with sound is generalizable audio-visual source localizer (2024)
5. Yan, S., Zhang, R., Guo, Z., Chen, W., Zhang, W., Li, H., Qiao, Y., He, Z., Gao, P.: Referred by multi-modality: A unified temporal transformer for video object segmentation. AAAI (2023)
6. Zhou, J., Wang, J., Zhang, J., Sun, W., Zhang, J., Birchfield, S., Guo, D., Kong, L., Wang, M., Zhong, Y.: Audio–visual segmentation. In: ECCV. pp. 386–403. Springer (2022)