

Dynamic Derivation and Elimination: Audio Visual Segmentation with Enhanced Audio Semantics

Chen Liu^{1,4}, Liying Yang⁵, Peike Li³, Dadong Wang⁴, Lincheng Li², Xin Yu^{1*}

¹ The University of Queensland, ² NetEase Fuxi AI Lab, ³ Matrix Verse AI,

⁴ CSIRO Data61, ⁵ Macau University of Science and Technology

yenianliu36@gmail.com, xin.yu@uq.edu.au

1. More Quantitative Results

1.1. Efficiency Comparison

We report training time and inference FPS for recent transformer-based methods in Table 1, on AVSS with batch size 1 (A100 40GB, 224² pixels).

Table 1. Comparisons of inference and training time.

Methods	TPAVI	AVSegFormer	AVSSStone	AVSBias	CAVP	Ours
Training	77h	46h	231h	1660h	56h	41h
Inference	15.4 fps	23.4 fps	3.9 fps	1.4 fps	25.9 fps	58.5 fps

1.2. More Backbone Comparisons

With PVT-V2-B5 and VGGish, our method still performs best on VPO, i.e., 73.58%, 73.35%, and 73.58% $\mathcal{J}\&\mathcal{F}_\beta$ on VPO-SS, -MS, and -MSMI, respectively.

1.3. Comparison with AVSAC

AVSAC [1] addresses modality imbalance via bidirectional AV interaction and frame-wise synchrony, while DDESeg tackles feature fusion and matching difficulty by enhancing audio representations and dynamically eliminating non-visual audio elements. Our method outperforms AVSAC: 88.0% / 94.2% (S4), 70.4% / 77.9% (MS3), 39.7% / 67.9% (AVSS) on $\mathcal{J}\&\mathcal{F}_m$.

1.4. Performance on Difficult Cases

❶ **Tiny objects:** DDESeg may produce inferior masks when audible objects have weak visual cues. ❷ **Disappeared objects:** After sounding objects disappear, DDESeg will not segment silent objects as there is no AV correlation. ❸ **Noise accumulation:** Noise does not accumulate, as DEM eliminates visually irrelevant audio semantics.



Figure 1. Visualizations of difficult cases.

*Corresponding author.

2. More Implementation Details

We train our model for 100 epochs using the AdamW optimizer [2] and an initial learning rate of 1e-4, distributed across eight NVIDIA A100 GPUs. The learning rate is scheduled with a cosine annealing schedule, gradually decreasing throughout training. To ensure a fair comparison, all images are resized to a resolution of 224 × 224. Audio samples are processed at a sampling rate of 22,050 Hz, with a window size of 1024, a hop size of 320, and 644 mel bins, to compute STFTs and mel spectrograms.

References

- [1] Tianxiang Chen, Zhentao Tan, Tao Gong, Qi Chu, Yue Wu, Bin Liu, Nenghai Yu, Le Lu, and Jieping Ye. Bootstrapping audio-visual video segmentation by strengthening audio cues. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 1
- [2] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1