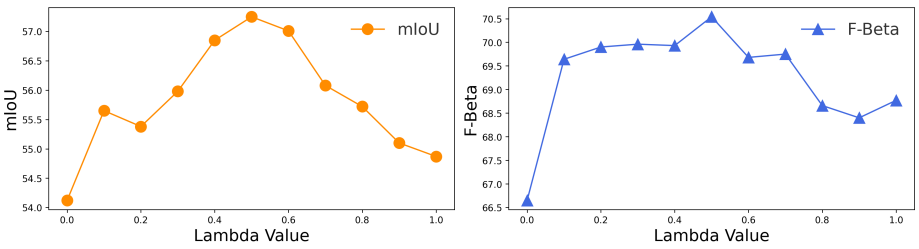
CPM: Class-conditional Prompting Machine for Audio-visual Segmentation

Yuanhong Chen[1], Chong Wang[1], Yuyuan Liu[1], Hu Wang[2], and Gustavo Carneiro[3]

[1] Australian Institute for Machine Learning, University of Adelaide, Australia
[2] Mohamed bin Zayed University of Artificial Intelligence, United Arab Emirates
[3] Centre for Vision, Speech and Signal Processing, University of Surrey
yuanhong.chen@adelaide.edu.au

1 Training and Inference Details

During training, we apply data augmentation for image inputs with colour jitter, horizontal flipping and random scaling between 0.5 and 2.0. We randomly crop images to 512×512 pixels. We downsample the audio data to 16 kHz for durations of 1 second [24] or 3 seconds [20] of the waveform on AVSBench and VPO. Subsequently, the resampled audio sequence is processed through the Short-Time Fourier Transform (STFT) using a 512 FFT length, a Hann window size of 400, and a hop length of 160. This results in 96×256 and 300×256 magnitude spectrograms on AVSBench and VPO, respectively. We use the AdamW [17] optimizer with a weight decay of 0.0001 and a polynomial learning-rate decay $(1 - \frac{\text{iter}}{\text{total\_iter}})^{\text{power}}$ with power = 0.9. We set the initial learning rate to 0.0001 with a mini-batch size of 16 and 100 epochs training length. During inference, we use the resized/or original resolution with a mini-batch size of 1. We set temperature τ as 0.1 and λ as 0.5. We adopted ResNet-50 [9] image backbones and a similar setting as [5] for the transformer decoder blocks in the segmentation head. For the audio backbones, we use VGGish [10] (following [24]) and ResNet-18 [9] (following [2, 20]) for AVSBench and VPO, respectively.



(a) mIoU ↑ **(b)** F_β ↑

Fig. 1: Ablation study on the weight coefficient λ using the AVSBench-Semantics dataset [23], with the ResNet50 [9] backbone.

Table 1: Quantitative (mIoU, F_β) audio-visual segmentation results (in %) on AVS-Bench test sets [23, 24] (resized to 224×224) with PVT-V2-B5 [21] backbone. Best results in **bold**, 2^{nd} best underlined. Improvements against the 2^{nd} best are in the last row.

Eval. Methods	Method	AVSBench-Object (SS)		AVSBench-Object (MS)		AVSBench-Semantics	
		mIoU ↑	F_β ↑	mIoU ↑	F_β ↑	mIoU ↑	F_β ↑
Per-image [23, 24]	TPAVI [24]	78.74	87.90	54.00	64.50	29.77	35.20
	AVSBG [8]	81.71	90.40	55.10	66.80	-	-
	ECMVAE [19]	81.74	90.10	57.84	70.80	-	-
	DiffusionAVS [18]	81.38	90.20	58.18	70.90	-	-
	CATR [12]	81.40	89.60	59.00	70.00	32.80	38.50
	AuTR [16]	80.40	89.10	56.20	67.20	-	-
	AQFormer [11]	81.60	89.40	61.10	72.10	-	-
	AVSegFormer [7]	82.06	89.90	58.36	69.30	36.66	42.00
	AVSC [14]	80.57	88.19	58.22	65.10	-	-
	BAVS [15]	81.96	88.60	58.63	65.49	32.64	36.42
	AVSAC [3]	84.51	91.56	64.15	76.60	36.98	42.39
	QSD [13]	-	-	-	-	-	-
	COMBO [22]	84.70	91.90	59.20	71.20	42.10	46.10
Per-dataset [6]	CAVP	<u>87.33</u>	<u>93.61</u>	<u>67.31</u>	<u>78.09</u>	<u>48.59</u>	<u>61.97</u>
	CPM	**91.26**	**95.43**	**68.41**	**79.09**	**55.08**	**69.01**

Algorithm 1 Stability Score (STS)

N: number of queries
C: number of classes
$f_{\text{Hungarian}}$: Hungarian algorithm
$H(\cdot)$: the function used to compute the negative entropy score
require: Training set $\mathcal{D}$ with D samples, model f_θ, class-agnostic query $\mathbf{q}$, an empty list R.
# Get assigned labels for each sample and each query.
for $(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{D}$ **do**
    $\mathbf{p} = f_\theta(\mathbf{x}_i)$
    $\tilde{\mathbf{y}}_i = f_{\text{Hungarian}}(\mathbf{p})$ # $\tilde{\mathbf{y}}_i : 1 \times N$
    R.append($\tilde{\mathbf{y}}_i$)
end for
# Concatenate all the assigned labels
$\mathbf{R} = \mathsf{Concat}(R)$ # $\mathbf{R} : D \times N$
$\mathsf{STS} = 0$
# Iterate through all the classes and compute the average STS score.
for c in C **do**
    $\mathbf{s}_c = (\mathbf{R} == c).\text{sum}(0)$ # $\mathbf{s}_c : 1 \times N$
    $\mathbf{s}_c = \text{clamp}(\mathbf{s}_c / \text{sum}(\mathbf{s}_c), \text{min} = 1e-12, \text{max} = 1)$
    $\mathsf{STS} = \mathsf{STS} + H(\mathbf{s}_c)$
end for
$\mathsf{STS} = \mathsf{STS}/C$

1.1 Additional Results

Results on Resized AVSBench We follow previous methods [4, 7, 8, 11, 12, 14–16, 18, 19, 24] to evaluate our model with PVT-V2-B5 [21] backbone on AVSBench-Objects (SS & MS) [24] and AVSBench-Semantics [23] with resized

image resolution (224×224). The results in Tab. 1 show that we improve mIoU by 3.93%, 1.10% and 6.49% on the respective benchmarks. Please note that Tab. 1 includes two evaluation protocols. We employ standard semantic segmentation protocols to compute both mIoU and F_β same as PascalVOC, initially mentioned in [24].

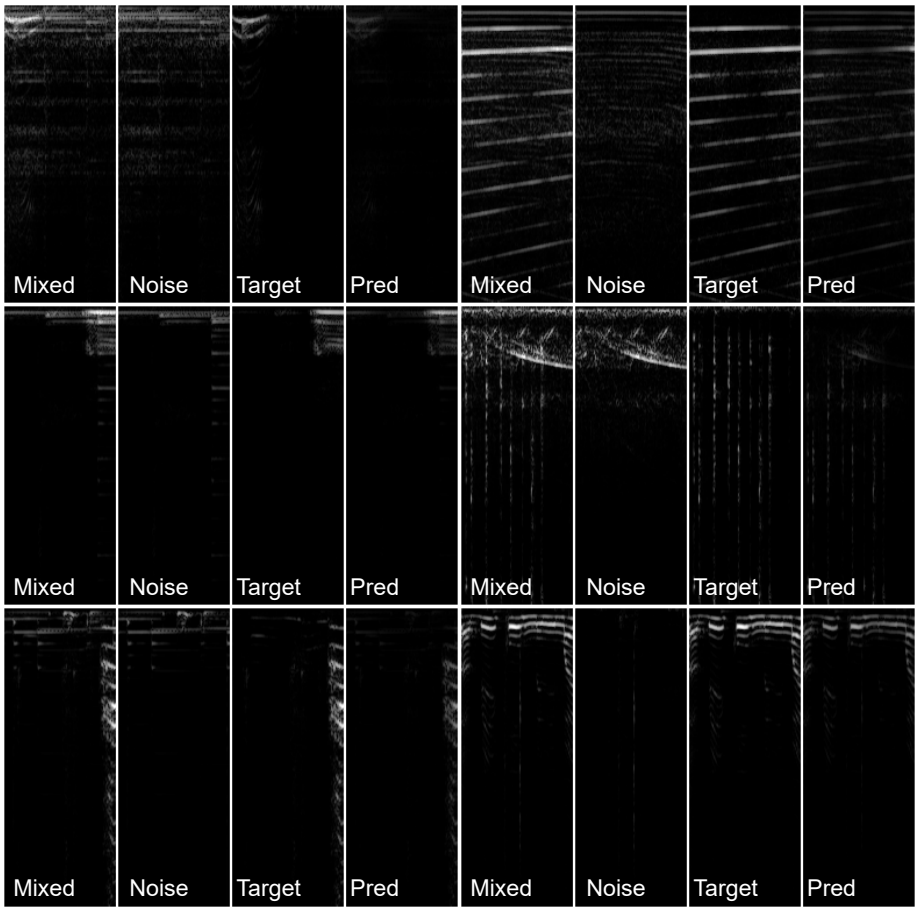


Fig. 2: Visualization of Audio Conditional Prompting (ACP) process on six samples from the AVSBench-Semantics dataset [16] (two examples per row). Each example includes a mixture of magnitude spectrogram (Mixed), noise beyond the visible range (Noise), the ground-truth target (Target), and the prediction generated by the CPM model (Pred).

Average Pooling Vs. Max Pooling We employed masked average pooling (MAP) in our prompting-based contrastive learning (PCL) module to extract correlated audio features. While an alternative option, masked max pooling

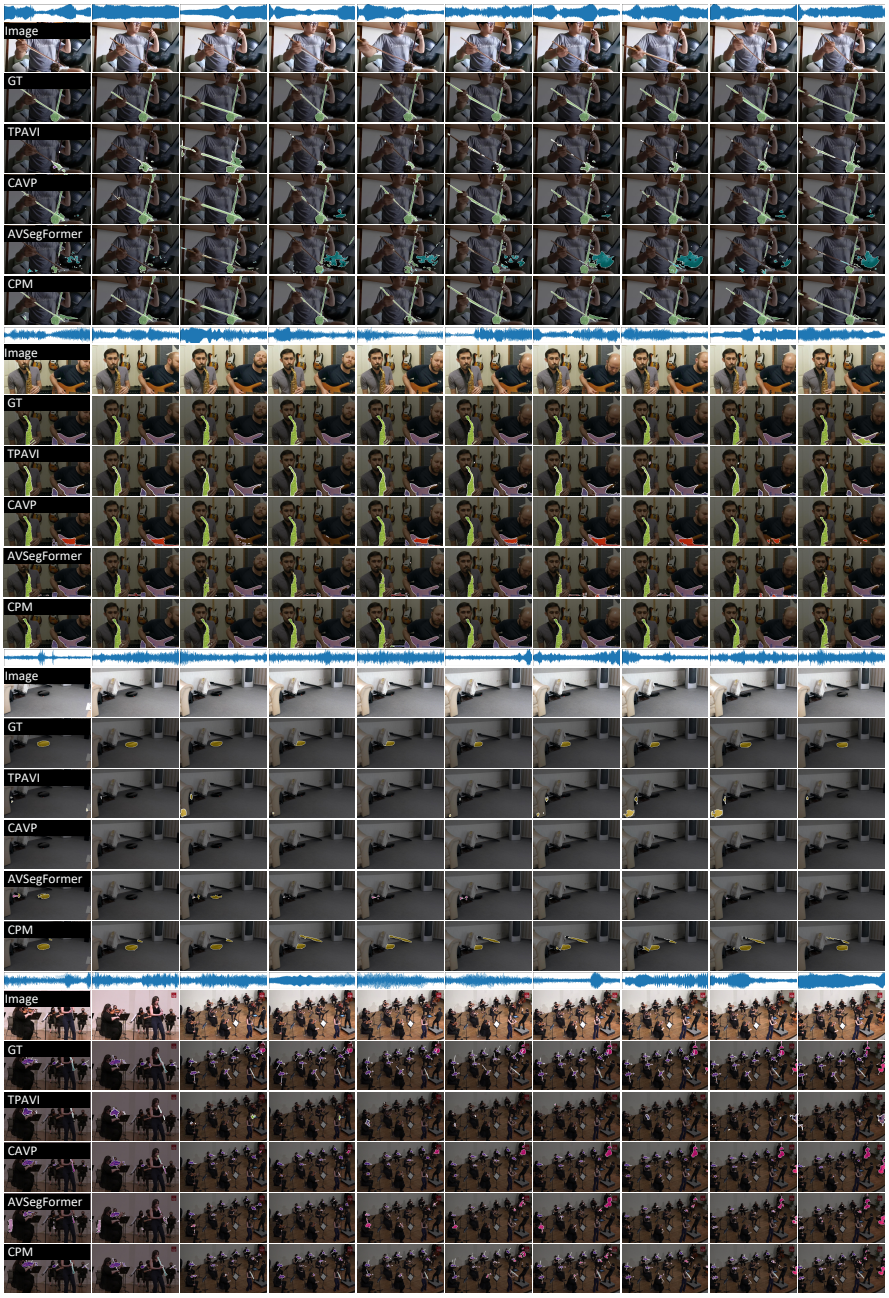


Fig. 3: Qualitative audio-visual segmentation results on AVSBench-Semantics [23] by TPAVI [24], AVSegFormer [7], CAVP [4] and our CPM. The prediction results can be compared with the original frame and the ground truth (GT) of the first two rows of each video.

	Per-pixel Classification		Transformer	
Methods	TPAVI [24]	CAVP [4]	AVSegFormer [7]	CPM
Training	48 h	22 h	23 h	30 h
Inference	9 fps	20 fps	7 fps	14 fps

Table 2: Efficiency comparison for training and inference on AVSBench-Semantic [16] using a ResNet50 backbone at original resolution.

(MMP), may seem viable, we argue it is less effective due to its sensitivity to false activation and incapability to capture spatially distributed information. To validate our choice, we conducted an ablation study replacing MAP with MMP in PCL during training and compared F_β scores on AVSBench-Semantics [16]. The results showed a 1.07% performance drop with MMP, indicating the effectiveness of MAP.

1.2 Evaluation of Training and Inference Efficiency

We measure the overall training time and frame per second (FPS) required for inference in Tab. 2. The experiments are conducted on the AVSBench-Semantic (AVSS) dataset [23] using a ResNet50 backbone at the original resolution. During inference, we evaluate models on 10 videos, sampling 100 frames per video (*i.e.*, a total of 10×100 frames) to compute the number of frames per second (FPS). The results show that our CPM requires additional training time (e.g., +8 hours) and has a slower inference speed (i.e., −6 fps) compared to CAVP. However, CPM achieves a +7 fps faster inference speed than AVSegFormer for a similar model architecture.

Hyper-parameter Analysis We further conduct an ablation study to investigate the sensitivity of the hyper-parameter (λ) on AVSBench-Semantics [16], as depicted in Fig. 1. Our findings demonstrate that a moderate value of λ is conducive to model training, whereas excessively small values ($\lambda = 0.1$) may result in learning stagnation, and overly large weights ($\lambda = 1.0$) could potentially be influenced by noisy gradients during the initial stages of learning.

2 Pseudo-code for Stability Score

We present a detailed explanation of the computation methodology for the stability score STS, as outlined in Alg.1. The STS is calculated after each epoch to evaluate the stability of the bipartite matching process through an average entropy score across all classes. The entire process comprises two main steps. Initially, we gather all assigned labels (generated by the Hungarian Algorithm [1]) for each training sample. Subsequently, we iterate through all classes and compute the respective STS based on the negative entropy of the probability distribution of the assignments.

3 Audio Conditional Prompting Visualisation

We show six training examples sourced from the AVSBench Semantics [16] dataset, as depicted in Figure 2. The results demonstrate our successful utilization of class conditional prompts to search for corresponding semantic information within the perturbed magnitude spectrogram.

4 Video Visualisation

We present a qualitative comparison visualization among TPAVI [24], AVSegFormer [7], CAVP [4], and our CPM on AVSBench-Semantics in Fig. 3. Our findings demonstrate that our method offers a more accurate 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. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
2. 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. pp. 16867–16876 (2021)
3. Chen, T., Tan, Z., Gong, T., Chu, Q., Wu, Y., Liu, B., Lu, L., Ye, J., Yu, N.: Bootstrapping audio-visual segmentation by strengthening audio cues. arXiv preprint arXiv:2402.02327 (2024)
4. Chen, Y., Liu, Y., Wang, H., Liu, F., Wang, C., Carneiro, G.: A closer look at audio-visual semantic segmentation. arXiv e-prints pp. arXiv–2304 (2023)
5. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1290–1299 (2022)
6. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**, 303–338 (2010)
7. Gao, S., Chen, Z., Chen, G., Wang, W., Lu, T.: Avsegformer: Audio-visual segmentation with transformer. arXiv preprint arXiv:2307.01146 (2023)
8. Hao, D., Mao, Y., He, B., Han, X., Dai, Y., Zhong, Y.: Improving audio-visual segmentation with bidirectional generation. arXiv preprint arXiv:2308.08288 (2023)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
10. Hershey, S., Chaudhuri, S., Ellis, D.P., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., et al.: Cnn architectures for large-scale audio classification. In: 2017 ieee international conference on acoustics, speech and signal processing (icassp). pp. 131–135. IEEE (2017)

11. Huang, S., Li, H., Wang, Y., Zhu, H., Dai, J., Han, J., Rong, W., Liu, S.: Discovering sounding objects by audio queries for audio visual segmentation. arXiv preprint arXiv:2309.09501 (2023)
12. Li, K., Yang, Z., Chen, L., Yang, Y., Xun, J.: Catr: Combinatorial-dependence audio-queried transformer for audio-visual video segmentation. arXiv preprint arXiv:2309.09709 (2023)
13. Li, X., Wang, J., Xu, X., Peng, X., Singh, R., Lu, Y., Raj, B.: Towards robust audiovisual segmentation in complex environments with quantization-based semantic decomposition. arXiv preprint arXiv:2310.00132 (2023)
14. Liu, C., Li, P., Qi, X., Zhang, H., Li, L., Wang, D., Yu, X.: Audio-visual segmentation by exploring cross-modal mutual semantics (2023)
15. Liu, C., Li, P., Zhang, H., Li, L., Huang, Z., Wang, D., Yu, X.: Bavs: Bootstrapping audio-visual segmentation by integrating foundation knowledge. arXiv preprint arXiv:2308.10175 (2023)
16. Liu, J., Ju, C., Ma, C., Wang, Y., Wang, Y., Zhang, Y.: Audio-aware query-enhanced transformer for audio-visual segmentation. arXiv preprint arXiv:2307.13236 (2023)
17. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
18. Mao, Y., Zhang, J., Xiang, M., Lv, Y., Zhong, Y., Dai, Y.: Contrastive conditional latent diffusion for audio-visual segmentation. arXiv preprint arXiv:2307.16579 (2023)
19. Mao, Y., Zhang, J., Xiang, M., Zhong, Y., Dai, Y.: Multimodal variational auto-encoder based audio-visual segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 954–965 (2023)
20. Mo, S., Morgado, P.: A closer look at weakly-supervised audio-visual source localization. arXiv preprint arXiv:2209.09634 (2022)
21. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pvt v2: Improved baselines with pyramid vision transformer. Computational Visual Media **8**(3), 415–424 (2022)
22. Yang, Q., Nie, X., Li, T., Gao, P., Guo, Y., Zhen, C., Yan, P., Xiang, S.: Cooperation does matter: Exploring multi-order bilateral relations for audio-visual segmentation. arXiv preprint arXiv:2312.06462 (2023)
23. Zhou, J., Shen, X., Wang, J., Zhang, J., Sun, W., Zhang, J., Birchfield, S., Guo, D., Kong, L., Wang, M., et al.: Audio-visual segmentation with semantics. arXiv preprint arXiv:2301.13190 (2023)
24. Zhou, J., Wang, J., Zhang, J., Sun, W., Zhang, J., Birchfield, S., Guo, D., Kong, L., Wang, M., Zhong, Y.: Audio–visual segmentation. In: European Conference on Computer Vision. pp. 386–403. Springer (2022)