Supplementary Material for: How Can Contrastive Pre-training Benefit Audio-Visual Segmentation? A Study from Supervised and Zero-shot Perspectives

Jiarui Yu *[1]
yjr@mail.ustc.edu.cn
Haoran Li *[1]
lihaoran747@126.com
Yanbin Hao †[1]
haoyanbin@hotmail.com
Jinmeng Wu[2]
Jinmeng2004910@outlook.com
Tong Xu[1]
tongxu@ustc.edu.cn
Shuo Wang[1]
shuowang.edu@gmail.com
Xiangnan He[1]
xiangnanhe@gmail.com

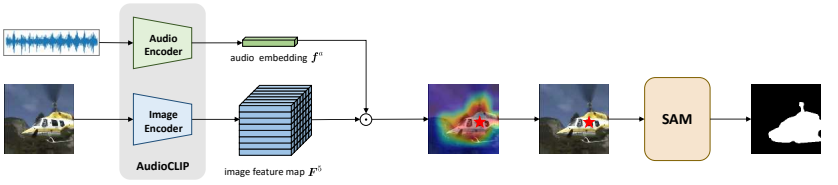
[1] University of Science and Technology of China
Hefei, China
[2] Wuhan Institute of Technology
Wuhan, China

The supplementary material provides additional content and more details, including (1) AC-FPN implementation details, (2) Heatmap-based Box-Prompt method, (3) Ablation study on Supervised AVS, (4) Ablation study on Zero-shot AVS, and (5) Failure case analysis of zero-shot AVS.

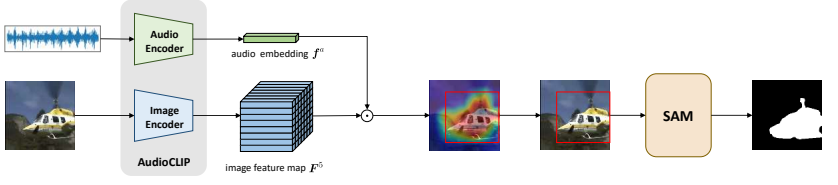
1 AC-FPN Implementation Details

In our proposed AC-FPN, we utilize the AudioCLIP backbones as the AC-FPN encoder and the FPN decoder as the AC-FPN decoder. For image(frame) encoding, the channel sizes of the four bottom-up visual feature maps $\{\boldsymbol{F}^i\}_{i=1}^4$ are $[256, 512, 1024, 2048]$ and the channel size of the last feature map $\boldsymbol{F}^5$ is 1024. Regarding audio encoding, we divide the audio signals into five one-second segments, assigning each segment to a frame. We then extend each one-second audio signal to 5s by repeating it five times. Each extended audio signal is fed into the AudioCLIP audio encoder, resulting in a 1024-dimension vector. For the decoding stage, we employ a structure the same as Semantic FPN [2], consisting of a neck with a channel size of 256 and a head with a channel size of 128. During training, we utilize

*Equal Contribution.
†Yanbin Hao is the corresponding author.



(a) Point-Prompt



(b) Heatmap-based Box-Prompt

Figure 1: The framework comparison of Point-Prompt and Heatmap-based Box-Prompt methods. (a) Point-Prompt mines the positive/negative points from the heatmap to prompt SAM. (b) Heatmap-based Box-Prompt generates bounding boxes of the positive region to prompt SAM. They share the same heatmap construction procedure.

the Adam [1] optimizer with a learning rate of 5×10^{-5}. Exceptionally, we adopt a learning rate of $0.1\times$ the global learning rate for the AudioCLIP image encoder, while keeping the AudioCLIP audio encoder fixed.

2 Heatmap-based Box-prompt

In the main paper, we presented Box-Prompt, a method that leverages CLAP and Grounding DINO to generate bounding boxes for prompting SAM. Despite Box-Prompt yielding satisfactory results for zero-shot AVS, it remains dependent on predefined category lists, such as the AudioSet category names.

In this section, we additionally introduce a method called Heatmap-based Box-Prompt, which directly generates bounding boxes from visual and audio features, eliminating the need for a category list to connect the two modalities. As shown in Figure 1 (b), we construct the heatmap following the same procedure introduced in Point-Prompt. Subsequently, we binarize the heatmap by applying a threshold, i.e., pixels with scores higher than the threshold are labelled as "positive", while the remaining pixels are labelled as "negative". Next, we extract the bounding box corresponding to the largest positive component in the binary map. Similar to Box-Prompt, we utilize these generated boxes to prompt SAM for mask predictions.

The experimental results, shown in Table 1, demonstrate that Heatmap-based Box-Prompt achieves comparable performance with Point-Prompt but is inferior to Box-Prompt.

Method	S4		MS3	
	mIoU	F-score	mIoU	F-score
Point-Prompt(dense)	40.3	.515	28.8	.333
Heatmap-based Box-Prompt	41.1	.547	24.0	.322
Box-Prompt	**51.2**	**.615**	**41.8**	**.478**

Table 1: Performance comparison of Point-Prompt, Heatmap-based Box-Prompt and Box-Prompt on AVSBench test split in the zero-shot setting.

Id	Visual-Enc.	Pre-train	Audio-Enc.	Pre-train	S4_mIoU	MS3_mIoU
1	R50	Contrastive	ESResNeXt	Contrastive	77.12	49.95
2	R50	Contrastive	ESResNeXt	AudioSet	76.89	49.20
3	R50	Contrastive	ESResNeXt	None	76.92	48.93
4	R50	Contrastive	No audio	No audio	76.31	47.66

Table 2: AC-FPN performance on AVSBench test split with different audio encoder pre-training tasks.

Audio-Encoder	S4_mIoU	S4_F-score	MS3_mIoU	MS3_F-Score
Frozen	77.12	.874	49.95	.635
Trainable	77.31	.875	53.33	.646

Table 3: AC-FPN performance with frozen/trainable Audio-Encoder.

Data Ratio	AC-FPN (Ours)		TPAVI-R50	
	S4_mIoU	MS3_mIoU	S4_mIoU	MS3_mIoU
100%	77.12	49.95	72.79	47.88
50%	74.78 (-3.0%)	43.27 (-13.4%)	70.50 (-3.1%)	37.94 (-20.8%)
20%	72.01 (-6.6%)	37.77 (-24.4%)	66.77 (-8.3%)	34.18 (-28.6%)

Table 4: Performance varies with different amounts of data.

3 Ablation Study on Supervised AVS

Varying audio encoder pre-training tasks. As illustrated in Table 2, it is evident that the performance declines when using a different audio encoder pre-training task (e.g., AudioSet) or when training from scratch (i.e., None). If we remove the audio encoder, the drop in performance is more obvious (i.e., No audio). These findings underscore the valuable role of contrastive pre-training in AVS.

Fix audio encoder or not. We fix our audio encoder to ensure a fair comparison with the baseline TPAVI, which freezes the audio encoder. From Table 3, we find that the performance can be further improved when training both audio and visual encoders.

Amount of training data. We test AC-FPN and TPAVI with 20% and 50% of training data. Tab. 4 shows that our AC-FPN consistently surpasses TPAVI with smaller performance drops, which indicates contrastive pre-training is helpful for low-data scenarios.

Metric	S4					MS3				
	0.5	0.6	0.7	0.8	0.9	0.5	0.6	0.7	0.8	0.9
mIoU	19.5	20.7	22.0	23.3	23.8	16.8	18.0	19.3	19.7	19.0
F-score	.281	.298	.321	.343	.358	.210	.220	.233	.242	.247

Table 5: Performance changes of No-Prompt with different thresholds on AVSBench test split in the zero-shot setting.

Metric	S4					MS3				
	0.55	0.60	0.65	0.70	0.75	0.55	0.60	0.65	0.70	0.75
mIoU	30.7	30.7	30.3	29.7	29.3	19.2	19.6	20.0	19.8	19.3
F-score	.414	.416	.413	.408	.407	.257	.265	.270	.270	.267

Table 6: Performance changes of Point-Prompt(local) with different thresholds on AVS-Bench test split in the zero-shot setting.

Metric	S4					MS3				
	0.70	0.75	0.80	0.85	0.90	0.65	0.70	0.75	0.80	0.85
mIoU	39.6	40.0	40.2	40.3	39.5	28.6	28.8	28.3	27.6	26.2
F-score	.484	.494	.504	.515	.521	.326	.333	.335	.335	.326

Table 7: Performance changes of Point-Prompt(dense) with different thresholds on AVS-Bench test split in the zero-shot setting.

Metric	S4				MS3			
	0.50	0.55	0.60	0.65	0.50	0.55	0.60	0.65
mIoU	41.7	41.1	40.1	39.0	23.2	24.0	23.9	22.9
F-score	.538	.547	.548	.548	.312	.322	.325	.312

Table 8: Performance changes of Heatmap-based Box-Prompt with different thresholds on AVSBench test split in the zero-shot setting.

4 Ablation Study on Zero-shot AVS

Threshold influence. In the zero-shot scenario, the threshold is the crucial hyper-parameter used to select positive prompts. Table 5, 6, 7 and 8 display the performance changes of four threshold-related methods by varying threshold. The results indicate the sensitivity of the performance to the chosen threshold. For the S4 subset, we set the thresholds as 0.9, 0.6, 0.85 and 0.55 for No-Prompt, Point-Prompt(local), Point-Prompt(dense) and Heatmap-based Box-Prompt, respectively. For the MS3 subset, the thresholds are set as 0.8, 0.65, 0.70 and 0.55 for No-Prompt, Point-Prompt(local), Point-Prompt(dense) and Heatmap-based Box-Prompt, respectively.

Grounded SAM without predicted category in Box-Prompt. In Box-Prompt, Grounded SAM receives the textual prompt, i.e., audio category predicted by CLAP, and segments the given image. However, as shown in Tab. 9 (Original Image), simply prompting Grounded SAM with trivial textual prompts, like "the object in the middle", achieves better performance in the S4 sub-task. This is primarily due to the *location bias* of the dataset, i.e., most

Prompt	Original Image		Concatenated Image	
	S4_mIoU	MS3_mIoU	S4_mIoU	MS3_mIoU
Predicted Category	51.2	41.8	25.5	30.0
"the sounding object in the image"	8.5	7.2	8.6	11.2
"the object in the middle"	59.3	31.6	14.5	16.4

Table 9: Performance changes of Box-Prompt with the predicted category and the trivial textual prompts for Grounded SAM in the zero-shot setting. Original Image refers to the original dataset. Concatenated Image refers to the constructed dataset by concatenating each original image with three images from other categories.

Prompt	S4_mIoU	S4_F-score
Predicted Category	51.2	.615
Look! Object of [Predicted Category]	51.9	.628
Caption from AudioCaps*	54.5	.638

Table 10: Performance changes of Box-Prompt with different text prompts for Grounded SAM in the zero-shot setting.

Prompt	S4_mIoU	S4_F-Score
Predicted Category	51.2	.615
Groundtruth Category	63.9	.749

Table 11: Upper limit testing for Box-Prompt in the zero-shot setting.

objects are centrally located. To simulate a more intricate and unbiased scenario, we concatenate the original image with three random images from different categories, resulting in a bigger image with a 2x2 original image size, and conduct the testing again. As observed in Tab. 9 (Concatenated Image), we find that the performance of Grounded SAM with trivial textual prompts substantially lags behind it with the predicted category prompts.

Natural sentence as text prompt in Box-Prompt. In addition to the category name, Grounded SAM can be prompted by a natural sentence. For further exploration, we use two methods to convert the audio signal to a natural sentence for prompting Grounded SAM: (1) add a prefix for the predicted label, such as Look! Object of [label]" and (2) a caption retrieved from the AudioCaps* dataset using CLAP. In Tab. 10, natural sentences can improve the performance, and the caption retrieval outperforms the prefix.

Upper limit of Box-Prompt. To test the performance upper limit of Box-Prompt, we use the groundtruth audio category as the prompt for Grounded SAM. Tab. 11 shows the influence of CLAP error predictions on Box-Prompt's performance.

5 Zero-shot AVS Failure Case Analysis

In this section, we focus on the failure case analysis for zero-shot AVS. In Figure 2, we visualize the failure cases of Point-Prompt(dense) (row 5) and Heatmap-based Box-Prompt (row 7), along with the prompts (i.e., points and boxes) used for SAM in these methods (rows

*Kim et al. AudioCaps: Generating Captions for Audios in The Wild. NAACL 2019

4 and 6). Additionally, rows 1, 2, and 3 showcase the raw images, ground truth labels, and heatmaps generated by AudioCLIP, respectively.

In all cases, the heatmaps generally identify the location of the sounding object. This raises the question: *Why do they fail despite having reasonably good heatmaps?* In cases (a, b), where the sounding object (helicopter and car) is quite small, the low-resolution heatmap activates a much larger area than the actual object. As a result, a larger box or false positive points in the background are generated. Due to these inaccurate prompts, SAM incorrectly segments the background (sky and road). In cases (c, d), where the sounding object (car and horse) is large, the heatmaps fail to activate all the components of the sounding objects. Thus, the generated box and points are insufficient to accurately describe the shape of the sounding objects, particularly when using boxes as prompts. In case (e), the Heatmap-based Box-Prompt precisely locates the object. However, SAM misinterprets the box and mistakenly segments the bottom part of the keyboard. In case (f), we observed the heatmaps sometimes are distracted by the text in the frame. We suspect that this is because of the co-occurrence of text and auditory signals in some pre-training paired data. When the object size is neither too small nor too large, both Point-Prompt and Heatmap-based Box-Prompt perform well, as observed in cases (g, h, i).

We also show the failure cases of Box-Prompt in Figure 3, where Grounding DINO and CLAP are combined to generate the boxes as prompts. Besides the inaccurate localization observed in cases (a, b), the SAM model fails to segment objects based on human preference despite having reasonably good bounding boxes, as illustrated in cases (c, d). Fine-tuning the SAM to adapt it to the AVS domain is an avenue for future exploration.

References

[1] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[2] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *CVPR*, pages 6399–6408, 2019.

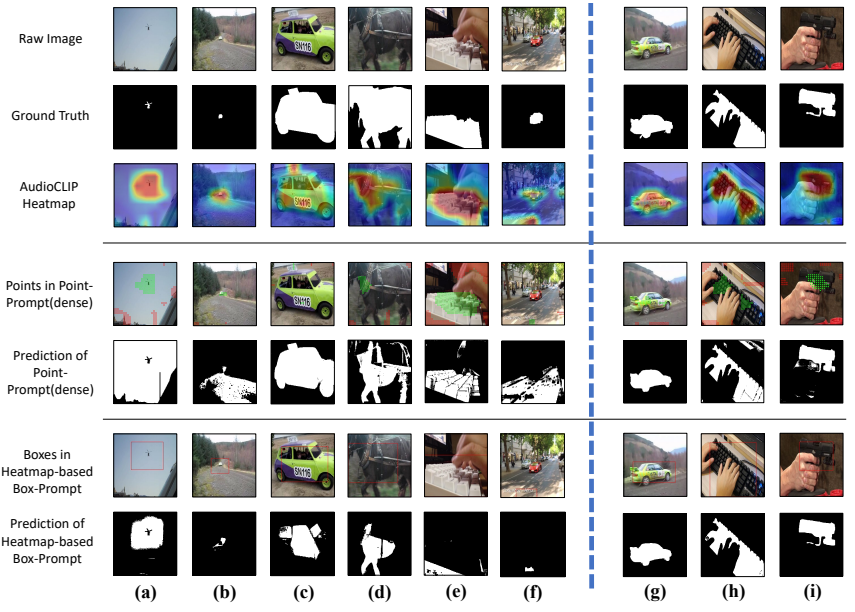


Figure 2: Failure cases of Point-Prompt and Heatmap-based Box-Prompt. Rows 1, 2 and 3 display the raw images, ground truth labels, and heatmaps generated by AudioCLIP. Rows 4 and 5 depict the point prompts (green points are positive and red ones are negative) and the predictions of Point-Prompt. Row 6 and 7 showcase the box prompts (red box) and the predictions of Heatmap-based Box-Prompt. The cases on the right side of the blue line are good cases.

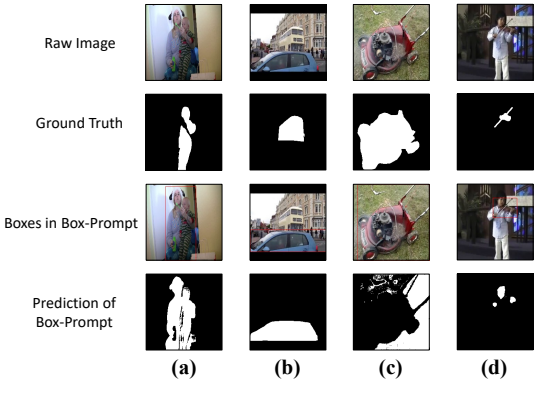


Figure 3: Failure cases of Box-Prompt. Rows 1 and 2 showcase the raw images and ground truth. Rows 3 and 4 display the box prompts (red box) generated by Grounding DINO and the predictions of Box-Prompt.