Acoustically Grounded Cost Learning for Open-Vocabulary Audio-Visual Semantic Segmentation

Tianrui Hui
Hefei University of Technology
Hefei, Anhui, China

Shaofei Huang*
Hefei University of Technology
Hefei, Anhui, China

Qisong Han
Cardiff University
Cardiff, Wales, United Kingdom

Yaxiong Wang
Hefei University of Technology
Hefei, Anhui, China

Lechao Cheng
Hefei University of Technology
Hefei, Anhui, China

Zhedong Zheng
University of Macau
Macau, China

Zhun Zhong
Hefei University of Technology
Hefei, Anhui, China

Richang Hong
Hefei University of Technology
Hefei, Anhui, China

Meng Wang
Hefei University of Technology
Hefei, Anhui, China

Abstract

Open-Vocabulary Audio-Visual Semantic Segmentation (OV-AVSS) aims to perform pixel-level segmentation of sound-emitting objects from an open set of categories. The previous method relies on a class-agnostic foreground definition, which groups semantically diverse objects into a heterogeneous positive set, causing the model to learn unstable sounding patterns and produce unreliable proposals. To address this, we reformulate the objective to be category-specific and propose a novel Acoustically Grounded Cost Learning (AGCL) framework to transform the static, audio-agnostic visual-text priors into dynamic, audio-grounded cost representations. For intra-category soundingness discovery, we devise Audio-Modulated Cost Generation (AMCG) and Audio-Guided Temporal Aggregation (AGTA) modules to enable both frame-level sounding region highlighting and video-level temporal refinement with a low-intrusive audio injection mechanism. For inter-category distractor discrimination, we introduce a Synergistic Distractor Mining (SDM) strategy, which selectively penalizes acoustically and semantically confusing negative categories to learn more discriminative decision boundaries. Extensive experiments on the AVSBench-OV dataset demonstrate that our method significantly outperforms previous state-of-the-art approaches, particularly on unseen categories. Code is available at https://github.com/spyflying/AGCL.

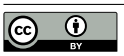
CCS Concepts

• **Computing methodologies** → **Video segmentation**; **Scene understanding**.

Keywords

Audio-Visual Semantic Segmentation, Open-Vocabulary, Cost Volume Learning

*Corresponding author.



This work is licensed under a Creative Commons Attribution 4.0 International License.
MM '26, Rio de Janeiro, Brazil
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2213-4/2026/11
https://doi.org/10.1145/3767308.3835028

ACM Reference Format:
Tianrui Hui, Shaofei Huang, Qisong Han, Yaxiong Wang, Lechao Cheng, Zhedong Zheng, Zhun Zhong, Richang Hong, and Meng Wang. 2026. Acoustically Grounded Cost Learning for Open-Vocabulary Audio-Visual Semantic Segmentation. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26), November 10–14, 2026, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3767308.3835028

1 Introduction

The goal of the Open-Vocabulary Audio-Visual Semantic Segmentation (OV-AVSS) [10] task is to achieve pixel-level segmentation and semantic recognition of sound-emitting objects from a given video and its associated audio across an open set of semantic categories. Unlike the conventional AVSS task [55, 56] that operates under a closed-set assumption, OV-AVSS requires the model to generalize beyond training categories and correctly identify both seen and unseen ones during inference, making this task substantially more challenging. Owing to this demand for generalization to unseen and unheard categories, OV-AVSS holds great potential for real-world applications such as robotic perception [11, 53], intelligent content creation [5, 12], embodied spatial reasoning [20, 42], etc.

As shown in Figure 1(a), the previous method [10] first generates class-agnostic sounding mask proposals via an audio-augmented segmentation model, then classifies them by offline feature matching with category texts using a frozen CLIP [33] model. However, foreground is defined event-wise as "***any object*** *that emits the given sound*" in the proposal generation stage, regardless of its semantic category. Such a category-agnostic definition groups semantically diverse objects under the same positive label, yielding a highly heterogeneous positive set with poor inner consistency. Consequently, the model tends to fit unstable soundingness rather than category-consistent objectness, resulting in unreliable proposals. These low-quality proposals further propagate errors to the proposal classification stage, degrading the performance. To address this issue, we reformulate the foreground definition in OV-AVSS as "***object of a specific category*** *that emits the given sound*", turning the original category-agnostic objective into a category-specific one. To accommodate the reformulated objective, we adopt the

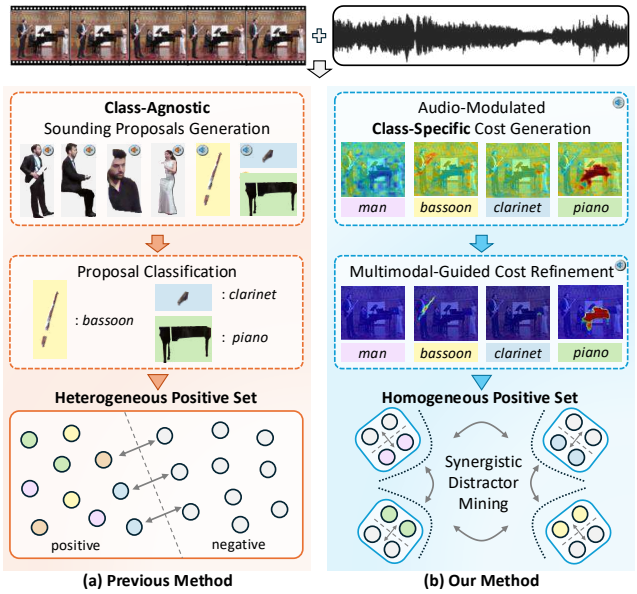


Figure 1: Motivation of our method. (a) The previous method adopts a class-agnostic foreground definition, grouping semantically diverse objects into a highly heterogeneous positive set. (b) Our method reformulates the foreground definition of OV-AVSS to be class-specific. By explicitly grounding acoustic events to cost representations, we achieve accurate intra-category soundingness discovery and robust inter-category distractor discrimination.

cost learning paradigm [4], where CLIP-derived visual-text matching costs naturally provide localization priors for individual categories. However, such priors only capture spatial existence rather than sounding dynamics for each category, yielding two primary challenges for the reformulation: **intra-category soundingness discovery**, *i.e.*, isolating objects "*emitting sound*" from silent ones within each specific category, and **inter-category distractor discrimination**, *i.e.*, distinguishing the category of "*given sound*" from acoustically confusing ones whose sounding characteristics are inadequately captured by visual-text priors. Therefore, we propose an Acoustically Grounded Cost Learning (AGCL) framework to transform the static, audio-agnostic visual-text priors into dynamic, audio-grounded cost representations.

For the first challenge, the AGCL framework features a low-intrusive audio injection mechanism that meticulously incorporates acoustic cues during both the generation and refinement of the cost maps, thereby highlighting sounding objects within each category while preserving the original open-vocabulary recognition capability. In the cost generation phase, we devise an Audio-Modulated Cost Generation (AMCG) module, which modulates frame-wise visual features in the CLIP vision encoder with audio cues before visual-text correlation. This modulation explicitly activates sounding regions on the cost map of each category without destructively shifting the original visual feature distributions. In the cost refinement phase, we further devise an Audio-Guided Temporal Aggregation (AGTA) module, which refines category-wise cost representations across the video by aggregating cost embeddings according to audio-derived temporal similarity, thereby improving the temporal consistency of soundingness discovery. Together, these two modules enable both frame-level sounding region highlighting and video-level temporal refinement, leading to more reliable sounding object discovery within each category.

To tackle the second challenge, we introduce the Synergistic Distractor Mining (SDM) strategy for robust inter-category discrimination. Since visual-text priors inherently lack sounding characteristics, standard cost learning struggles to establish clear decision boundaries among acoustically confusing categories. To overcome this limitation, rather than uniformly suppressing all negative categories during training, our SDM leverages external audio-language knowledge to selectively penalize the most confusing distractors based on comprehensive acoustic and semantic cues. Specifically, acoustic distractors are identified by the similarity between the input audio and category texts in the embedding space of a pretrained audio-language model, while semantic distractors are identified by the textual feature similarity between ground-truth and other categories. By explicitly penalizing these highly confusing distractors, SDM effectively sharpens the decision boundaries among easily confused categories, thereby encouraging better discrimination of unseen categories.

Our contributions are summarized as follows: (1) We reformulate the foreground definition in OV-AVSS to be class-specific and propose a novel Acoustically Grounded Cost Learning (AGCL) framework to transform the static, audio-agnostic visual-text priors into dynamic, audio-grounded cost representations. (2) For intra-category soundingness discovery, AGCL includes Audio-Modulated Cost Generation and Audio-Guided Temporal Aggregation modules to highlight sounding regions in each category while preserving the open-vocabulary capabilities. For inter-category distractor discrimination, we introduce the Synergistic Distractor Mining strategy to penalize semantically and acoustically confusing distractors for better generalization. (3) Extensive experiments on the AVSBench-OV dataset verify that our method outperforms previous state-of-the-art methods by large margins, especially for unseen categories (mIoU 29.14% *vs.* 45.59%).

2 Related Work

2.1 Audio-Visual Segmentation

Zhou *et al.* [55, 56] first introduce the Audio-Visual Segmentation (AVS) task and collect two datasets, namely AVSBench-Object and AVSBench-Semantic. The former focuses on binary foreground-background segmentation of sounding objects, while the latter further requires identifying the semantic categories of the sounding objects in addition to segmentation. Many subsequent studies [2, 8, 15–17, 22, 23, 27, 36, 40, 41, 50] adopt the audio-query design and achieve consistent performance gains. COMBO [50] introduces a bilateral entanglement mechanism at the pixel, modality, and temporal levels for more effective audio-visual fusion. VCT [16] constructs queries in a vision-centric manner, alleviating the perception ambiguity of audio queries and the loss of visual details. Beyond

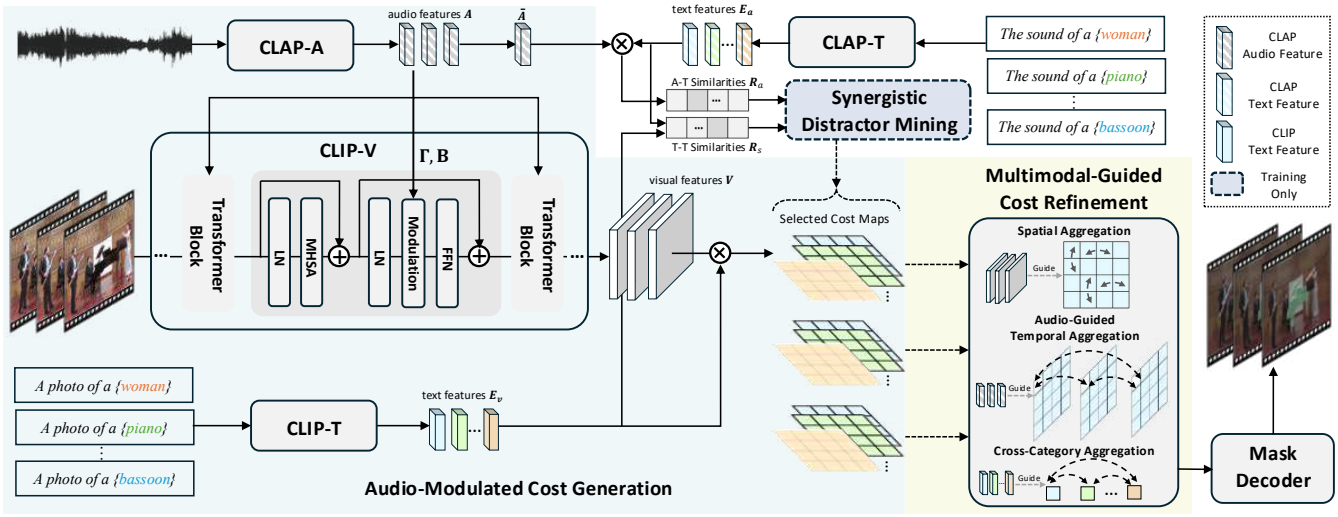


Figure 2: Overall architecture of our method. In the Audio-Modulated Cost Generation module, category-wise cost maps are obtained by correlating visual features from CLIP-V with category text features from CLIP-T. Audio features from CLAP-A are injected into CLIP-V to modulate visual features and softly highlight sounding regions. The resulting cost maps are further refined in the Multimodal-Guided Cost Refinement module to produce sharper spatial boundaries and better temporal consistency. The refined cost maps are then fed into a mask decoder to generate category-wise sounding object segmentation masks. During training, we apply the Synergistic Distractor Mining (SDM) strategy to the generated cost maps to select acoustically and semantically confusing distractor categories for more discriminative learning.

architectural evolution, UFE [28] exploits motion and semantic cues from unlabeled neighboring and distant frames to improve performance, while ICF [52] introduces an implicit counterfactual framework that balances audio-visual modalities through implicit text bridging and counterfactual learning. This paper focuses on the OV-AVSS [10] task, which extends the closed-set semantic categories in AVS to an open-vocabulary scenario. Different from the previous method [10] that relies on a class-agnostic foreground definition, we reformulate the task to model category-specific sounding situations. By explicitly grounding acoustic events to cost representations, our AGCL framework tackles the challenges of intra-category soundingness discovery and inter-category distractor discrimination for more effective open-vocabulary segmentation.

2.2 Open-Vocabulary Semantic Segmentation

Open-Vocabulary Semantic Segmentation (OVSS) aims to enable models to perform pixel-level segmentation for arbitrary semantic categories, including those unseen during training. Current OVSS methods can generally be categorized into two paradigms: two-stage pipelines [6, 7, 18, 24, 25, 29, 48] and end-to-end pipelines [4, 39, 44–47, 54]. Two-stage pipelines typically first extract object proposals and then classify them using CLIP [33]. For example, SCAN [29] enhances the alignment between proposals and category texts by incorporating CLIP-based semantic priors and contextual calibration. Mask-Adapter [24] extracts semantic activation maps from masks and enforces consistency to better align them with CLIP features. Among end-to-end pipelines, SAN [47] designs side-adapters on top of a frozen CLIP to predict masks and attention biases, achieving efficient OVSS. CAT-Seg [4] constructs and aggregates a cost volume between visual and textual features to achieve pixel-level vision-language alignment. In this work, we propose an Acoustically Grounded Cost Learning (AGCL) framework, which transforms the static, audio-agnostic visual-text priors in plain cost learning paradigm [4] into dynamic, audio-grounded cost representations.

2.3 Cost Volume Learning

Cost volumes [26, 34, 35] were originally introduced to represent dense matching costs between the spatial features of image pairs, measured using metrics such as cosine similarity. This volume is subsequently aggregated to establish fine-grained correspondences for tasks like optical flow and stereo matching [3, 13, 14, 49]. CAT-Seg [4] constructs a multi-modal cost volume between image and text features extracted by CLIP and aggregates it, achieving superior open-vocabulary semantic segmentation performance compared to directly operating on image-text features. However, existing cost learning paradigms predominantly focus on visual-textual alignments, leaving the potential of acoustic cues unexplored. In this paper, we further incorporate audio information into the multi-modal cost volume learning process to collaborate with category texts, thereby improving the segmentation of sounding objects from both seen and unseen categories.

3 Acoustically Grounded Cost Learning

Given a video clip $\mathcal{V} = \{\mathcal{V}_t\}_{t=1}^{T}$ of T frames and its corresponding audio clip $\mathcal{A}$, the Open-Vocabulary Audio-Visual Semantic Segmentation (OV-AVSS) task aims to predict frame-level segmentation masks $Y = \{Y_t\}_{t=1}^{T}$ of sounding objects over the *total* category

set $C_{\text{total}} = C_{\text{seen}} \cup C_{\text{unseen}}$, while only the *seen* categories C_{seen} are annotated and used for training. The overall architecture of our method is illustrated in Figure 2. The proposed AGCL framework addresses the two primary challenges of the reformulated OV-AVSS task, namely intra-category soundingness discovery and inter-category distractor discrimination. To tackle the first challenge, we introduce the Audio-Modulated Cost Generation (AMCG) and Audio-Guided Temporal Aggregation (AGTA) modules. First, the AMCG module softly injects audio features extracted by the audio encoder into the vision encoder of the CLIP model to produce sound-aware visual features. These features are then correlated with text features to obtain category-wise cost maps that serve as localization priors of sounding objects. Subsequently, the AGTA module refines these cost embeddings across the temporal dimension to produce temporally consistent representations. The refined cost embeddings are then fed into a mask decoder to generate pixel-wise segmentation masks for sounding objects in each category. To tackle the second challenge, we introduce the Synergistic Distractor Mining (SDM) strategy during training. Rather than learning an optimization bias of universal suppression on most categories, SDM leverages external audio-language knowledge to dynamically mine the most confusing negative categories across both acoustic and semantic dimensions. By selectively penalizing these distractors, SDM effectively guides the model to learn highly discriminative decision boundaries for better generalization.

3.1 Multimodal Feature Extraction

We adopt two pretrained models for feature extraction, namely the CLIP model [33] to extract aligned vision features and category text features, and the CLAP model [43] to extract aligned audio features and category text features. Specifically, the audio clip is first fed into the audio encoder of the CLAP model to extract dense audio features, which are then temporally aligned to the video length T, resulting in audio features $\boldsymbol{A} \in \mathbb{R}^{T \times D_a}$, where D_a is the channel number. For the video clip, each frame is passed through the vision encoder of the CLIP model to obtain dense visual features $\boldsymbol{V} \in \mathbb{R}^{T \times H \times W \times D_v}$, where H, W and D_v denote the height, width, and channel number of visual features. During this process, the audio features $\boldsymbol{A}$ are used as modulation factors to inject audio-aware cues into the visual representations, as detailed in the next section. For the category names, we obtain two types of text features, *i.e.*, a visual-perspective feature $\boldsymbol{E}_v \in \mathbb{R}^{C_{\text{k}} \times D_v}$ from the text encoder of the CLIP model using the template *"A photo of a {category}"*, and an acoustic-perspective feature $\boldsymbol{E}_a \in \mathbb{R}^{C_{\text{k}} \times D_a}$ from the text encoder of the CLAP model using the template *"The sound of a {category}"*. Unless otherwise specified, in the following sections of this paper, we use C_{k} to denote $C_{\text{seen}} = |C_{\text{seen}}|$ during training and $C_{\text{total}} = |C_{\text{total}}|$ during inference.

3.2 Audio-Modulated Cost Generation

To obtain audio-aware cost maps with strong generalization, we softly inject audio features into CLIP's vision encoder through audio-conditioned modulation layers, thus discovering sounding regions in each category while preserving the vision-language alignment properties. Concretely, CLIP's vision encoder consists of L Transformer blocks, where the l-th block follows the sequence $\text{LN}_1^l(\cdot) \rightarrow \text{MHSA}^l(\cdot) \rightarrow \text{LN}_2^l(\cdot) \rightarrow \text{FFN}^l(\cdot)$, for $l = 1, \ldots, L$. We modify this structure by inserting the audio-conditioned modulation layer after $\text{LN}_2^l(\cdot)$ in selected blocks to modulate the normalized visual activations. Let $\tilde{\boldsymbol{V}}^l \in \mathbb{R}^{T \times H \times W \times D_v}$ denote the visual features output by $\text{LN}_2^l(\cdot)$ in the l-th block. We first apply two linear layers to the audio feature $\boldsymbol{A}$ to predict modulation factors:

$$\boldsymbol{\Gamma}^l = 1 + \phi_s^l(\boldsymbol{A}), \quad \mathbf{B}^l = \phi_b^l(\boldsymbol{A}), \tag{1}$$

where $\boldsymbol{\Gamma}^l, \mathbf{B}^l \in \mathbb{R}^{T \times D_v}$ are the scale and bias factors, and $\phi_s^l(\cdot)$ and $\phi_b^l(\cdot)$ denote linear layers. Note that the weights and biases of both $\phi_s^l(\cdot)$ and $\phi_b^l(\cdot)$ are initialized to zero, so that the modulation starts with no effect on visual features and gradually takes effect as training progresses, ensuring a low-intrusive audio injection. The modulation is then performed as:

$$\bar{\boldsymbol{V}}^l = \boldsymbol{\Gamma}^l \odot \tilde{\boldsymbol{V}}^l + \mathbf{B}^l, \tag{2}$$

where broadcasting is applied along the spatial dimensions, and $\odot$ denotes element-wise multiplication. The modulated visual features $\bar{\boldsymbol{V}}^l$ are subsequently fed into the $\text{FFN}^l(\cdot)$ to produce the output features of the l-th block.

After passing through L blocks, the final output $\boldsymbol{V}$ of vision encoder becomes audio-aware. We then calculate the cosine similarity between $\boldsymbol{V}$ and the text features $\boldsymbol{E}_v$ from CLIP text encoder, to generate audio-aware category-wise cost maps $\mathcal{M} \in \mathbb{R}^{T \times H \times W \times C_{\text{k}}}$:

$$\mathcal{M} = \text{Sim}(\phi^v(\boldsymbol{V}), \phi^e(\boldsymbol{E}_v)), \tag{3}$$

where $\text{Sim}(\cdot)$ denotes the cosine similarity, and $\phi^v(\cdot)$ and $\phi^e(\cdot)$ are linear projection layers. We then apply a convolutional layer to the cost map for each category to obtain the cost embeddings $\boldsymbol{M} \in \mathbb{R}^{T \times H \times W \times C_{\text{k}} \times D_m}$, where D_m is the cost embedding dimension.

3.3 Multimodal-Guided Cost Refinement

After obtaining the coarse cost embedding $\boldsymbol{M}$, we extend the prior cost-learning paradigm [4] from a single-frame formulation to a temporal one and introduce three aggregation components that progressively refine the embeddings with multimodal cues, thereby improving spatial boundary precision and temporal consistency.

Spatial aggregation. The spatial aggregation (SA) component refines the category-wise cost embeddings $\boldsymbol{M}$ along the spatial dimension to refine object boundary delineation and suppress background noise. Following [4], it applies two Swin Transformer blocks [30] to each frame of category-wise cost embedding $\boldsymbol{M}(t, :, :, c) \in \mathbb{R}^{H \times W \times D_m}$ for spatial context aggregation, where $t \in [1, T]$, and $c \in [1, C_{\text{k}}]$. Visual features from the t-th frame $\boldsymbol{V}(t) \in \mathbb{R}^{H \times W \times D_v}$ are linearly projected and concatenated with each $\boldsymbol{M}(t, :, :, c)$, serving as query-key inputs for the self-attention layers of the Swin Transformer block to provide spatial structural guidance in an attention manner. Formally, the process can be expressed as:

$$\boldsymbol{M}_{\text{sa}} = \Psi_{\text{sa}}(\boldsymbol{M}, \phi_{\text{sa}}^v(\boldsymbol{V})), \tag{4}$$

where $\phi_{\text{sa}}^v(\cdot)$ denotes the linear projection layer.

Audio-guided temporal aggregation. Since the input of OV-AVSS consists of multiple video frames, temporal context plays an important role in localizing sound-emitting objects. Therefore, we further propose an audio-guided temporal aggregation (AGTA) component, which aggregates category-wise cost embeddings across

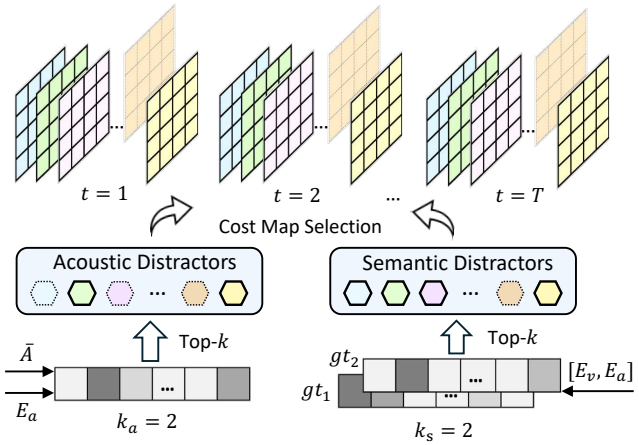


Figure 3: A simplified example of Synergistic Distractor Mining. k_a acoustic distractors are selected based on the feature similarities between the input audio and category texts. k_s semantic distractors are selected based on text feature similarities between ground-truth and other categories. Here, k_a and k_s are set to 2 for better illustration.

frames based on their audio similarity, thereby enhancing the temporal consistency of the cost representation. In detail, we apply a Transformer [37] layer on each $\boldsymbol{M}_{\text{sa}}(:, h, w, c) \in \mathbb{R}^{T \times D_m}$ to aggregate multi-frame temporal context, where $h \in [1, H], w \in [1, W], c \in [1, C_{\text{k}}]$. Formally, the process can be expressed as:

$$\boldsymbol{M}_{\text{ta}} = \Psi_{\text{ta}}(\boldsymbol{M}_{\text{sa}}, \phi^a_{\text{ta}}(\boldsymbol{A})), \tag{5}$$

where the audio feature projected by $\phi^a_{\text{ta}}(\cdot)$ is utilized as temporal guidance. Similarly, instead of directly incorporating audio features, AGTA aggregates information through audio-derived cross-frame attention, which helps alleviate overfitting to seen acoustic patterns.

Cross-category aggregation. We then apply a cross-category aggregation (CCA) component to model the relationships among different categories to enhance category discrimination. Similar to the previous two modules, the CCA component aggregates cost embeddings among categories for each spatial location using a Transformer layer, guided by category text features $\boldsymbol{E}_v$. The CCA component $\Psi_{\text{ca}}(\cdot)$ is formulated as follows:

$$\boldsymbol{M}_{\text{ca}} = \Psi_{\text{ca}}(\boldsymbol{M}_{\text{ta}}, \phi^e_{\text{ca}}(\boldsymbol{E}_v)). \tag{6}$$

During training, we also design a category description diversification scheme that leverages an LLM [1] to generate multiple semantically equivalent but syntactically diverse descriptions for each category name. In each iteration, one description is randomly selected to replace the plain category name for text feature extraction as category guidance in the CCA component, thereby enriching the diversity of inter-class interaction patterns.

3.4 Synergistic Distractor Mining

To tackle the challenge of inter-category distractor discrimination, we propose a Synergistic Distractor Mining (SDM) strategy, which selectively mines acoustic and semantic distractor categories, to sharpen the decision boundaries during training. We present a simplified example in Figure 3 for illustration. For acoustic distractor mining, we first calculate the similarity between CLAP audio features $\boldsymbol{A}$ and text features $\boldsymbol{E}_a$ as:

$$\boldsymbol{R}_a = \text{Sim}(\phi^a_{\text{sim}}(\bar{\boldsymbol{A}}), \phi^e_{\text{sim}}(\boldsymbol{E}_a)), \tag{7}$$

where $\bar{\boldsymbol{A}} \in \mathbb{R}^{1 \times D_a}$ is the temporal average of $\boldsymbol{A}$, $\phi^a_{\text{sim}}(\cdot)$ and $\phi^e_{\text{sim}}(\cdot)$ are projection layers. The top-k_a categories with the highest similarities in $\boldsymbol{R}_a$ are regarded as acoustic distractors, whose indices are denoted as $\mathcal{I}_a$. To make the category selection process differentiable, we apply Gumbel-TopK [19, 21, 31] to $\boldsymbol{R}_a$. A soft audio-category distribution is first generated as:

$$\tilde{\boldsymbol{R}}_a = \text{Softmax}(\boldsymbol{R}_a + \boldsymbol{G}), \boldsymbol{G} \sim \text{Gumbel}(0, 1). \tag{8}$$

We then define a multi-row one-hot selection matrix $\text{S}^{\text{hard}}_a \in \mathbb{R}^{k_a \times C_{\text{seen}}}$ as follows:

$$\text{S}^{\text{hard}}_a(i, j) = \begin{cases} 1, & \text{if } j = \mathcal{I}_a(i), \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

To keep gradients to $\boldsymbol{R}_a$, we replicate $\tilde{\boldsymbol{R}}_a$ into $\boldsymbol{R}'_a \in \mathbb{R}^{k_a \times C_{\text{seen}}}$ and apply the straight-through estimator:

$$\boldsymbol{S}_a = \text{S}^{\text{hard}}_a + \boldsymbol{R}'_a - \text{sg}(\boldsymbol{R}'_a) \tag{10}$$

where $\text{sg}(\cdot)$ stops gradients. The cost maps for acoustic distractors can be obtained by matrix multiplication between $\mathcal{M}$ and $\boldsymbol{S}_a$.

For semantic distractor mining, we compute the pairwise similarity matrix $\boldsymbol{R}_s \in \mathbb{R}^{C_{\text{seen}} \times C_{\text{seen}}}$ among the seen categories based on the concatenation of their CLIP text features $\boldsymbol{E}_v$ and CLAP text features $\boldsymbol{E}_a$. For each ground-truth category, the top-k_s categories having the highest similarities to it in $\boldsymbol{R}_s$ are selected as semantic distractors, as they are semantically similar to the true sounding category but belong to different categories. The cost maps for semantic distractors are obtained similarly to those for acoustic distractors. Note that the SDM strategy is applied only during training, while all categories are included for cost computing during inference.

3.5 Loss Functions

During training, the refined cost embedding $\boldsymbol{M}_{ca}$ is fed into a mask decoder [4] to predict segmentation logits $\boldsymbol{Y}$. We compute BCE loss $\mathcal{L}_{\text{mask}}$ between $\boldsymbol{Y}$ and the ground-truth mask $\boldsymbol{Y}^*$ to supervise the semantic segmentation learning on seen categories. To ensure semantic consistency across modalities, we further design an alignment loss between the CLIP- and CLAP-encoded text features, encouraging implicit alignment between visual and audio representations. The similarity matrix $\boldsymbol{P} \in \mathbb{R}^{C_{\text{seen}} \times C_{\text{seen}}}$ is computed as:

$$\boldsymbol{P} = \text{Sim}(\phi^e_{\text{clip}}(\boldsymbol{E}_v), \phi^e_{\text{clap}}(\boldsymbol{E}_a)), \tag{11}$$

where $\phi^e_{\text{clip}}(\cdot)$ and $\phi^e_{\text{clap}}(\cdot)$ are projection layers. The ground-truth for feature similarity is an identity matrix $\boldsymbol{P}^*$. The alignment loss $\mathcal{L}_{\text{align}}$ is also implemented as a BCE loss between $\boldsymbol{P}$ and $\boldsymbol{P}^*$. $\mathcal{L}_{\text{align}}$ encourages features from the same category to be close while pushing apart those from different categories, thereby preserving semantic consistency across modalities. The overall loss function of our model is summarized as:

$$\mathcal{L} = \lambda_{mask}\mathcal{L}_{\text{mask}} + \lambda_{align}\mathcal{L}_{\text{align}}, \tag{12}$$

where λ_{mask} and λ_{align} are the loss coefficients.

Table 1: Comparison with state-of-the-art methods on the AVSBench-OV dataset. Our method obtains significant performance gains for the challenging unseen categories with the same CLIP model as previous methods. * denotes the results are reproduced by us using the authors' released code. Best results are presented in bold.

Method	Reference	Vision Encoder	Text Encoder	mIoU			
				Unseen	Seen	Harmonic	Overall
Closed-Set Methods							
TPAVI [55, 56]	[ECCV'22]	ResNet-50	-	0.00	28.45	0.00	18.08
CATR [23]	[ACM MM'23]	ResNet-50	-	0.00	30.64	0.00	19.73
VCT [16]*	[CVPR'25]	Swin-Base	-	0.00	52.09	0.00	32.00
Zero-Shot Method							
Sam4AVS [51]	[BMVC'23]	Grounded-SAM	-	8.53	13.55	10.47	12.47
Open-Vocabulary Methods							
OV2Seg [38]	[ICCV'23]	ResNet-50	CLIP	12.49	36.25	18.58	26.93
OpenVIS [9]	[AAAI'25]	ResNet-50	CLIP-ViT-B/32	17.54	45.23	25.31	34.40
CLIP-VIS [57]	[TCSVT'24]	ResNet-50	CLIP	15.00	40.35	21.87	30.31
OV-AVSS [10]	[ACM MM'24]	R-50+CLIP-B	CLIP-ViT-B/16	22.00	49.77	30.71	38.91
		Swin-B+CLIP-L	CLIP-ViT-L/14	29.14	55.43	38.20	44.81
AGCL (Ours)	-	CLIP-ViT-B/16	CLIP-ViT-B/16	33.81	61.03	43.51	50.69
		CLIP-ViT-L/14	CLIP-ViT-L/14	**45.59**	**61.47**	**52.35**	**55.20**

4 Experiment

4.1 Dataset and Evaluation Metrics

We train and evaluate our models on the AVSBench-OV dataset [10], which is derived from AVSBench-Semantic [55]. AVSBench-OV contains 70 semantic categories in total, of which 40 frequently occurring categories are used as seen categories, and the other 30 uncommon ones are reserved as unseen categories. Videos containing the unseen categories are excluded from the training set to avoid data leakage. The dataset includes 5,184 real-world videos for training, 1,240 for validation, and 1,490 for testing. Following the prior work [10], we adopt mean Intersection-over-Union (mIoU) as the evaluation metric. Four variants are reported, namely the overall mIoU over all 70 categories, Seen-mIoU for seen categories, Unseen-mIoU for unseen categories, and Harmonic-mIoU defined as the harmonic mean between Seen- and Unseen-mIoUs. We also conduct out-of-distribution (OOD) evaluation on the VPO dataset [2] to validate the cross-dataset generalization capability of our method.

4.2 Implementation Details

Our model is implemented using PyTorch [32]. We employ CLIP-ViT-B/16 and CLIP-ViT-L/14 models [33] as the pre-trained vision-language models for fair comparison with the previous method [10], and adopt the CLAP model [43] pre-trained on the LAION-Audio-630K dataset as the audio-language model. During training, only the key and value projection layers are finetuned for both CLIP and CLAP models. For each video clip, the total number of video frames T is set to 5 during training, while we sample 1 frame per second during testing. The input video frame resolution is 336×336, and the cost embedding dimension D_m is 128. The numbers of selected categories in SDM are $k_a = 10$ and $k_s = 3$. The loss coefficients λ_{mask} and λ_{align} are set to 1.0 and 0.01. The AdamW optimizer is employed with an initial learning rate of $1e^{-4}$ for CLIP-ViT-B/16 and $2e^{-4}$ for CLIP-ViT-L/14 with a weight decay of $1e^{-4}$ and a cosine learning-rate schedule. We train our model for a total of 40,000 iterations with a batch size of 1. All experiments are conducted on a single RTX 4090 GPU with 24 GB of memory.

4.3 Comparison with State-of-the-art Methods

Since OV-AVSS is a relatively new task, there has been only one dedicated work [10] prior to ours. Following its evaluation protocol, we report results on the AVSBench-OV dataset using the same set of adapted baselines, which include closed-set AVS methods [16, 23, 56], a zero-shot AVS method [51], and open-vocabulary semantic segmentation methods [9, 38, 57]. As shown in Table 1, our AGCL framework achieves significant improvements over all prior methods across both seen and unseen categories. In particular, AGCL (CLIP-ViT-B/16) achieves a harmonic mIoU of 43.51%, exceeding the previous best OV-AVSS (CLIP-ViT-L/14) by nearly 5% on the unseen set. When using the same CLIP model (*i.e.*, CLIP-ViT-B/16), our method outperforms OV-AVSS by over 10% on both seen and unseen categories. These improvements indicate that our AGCL framework learns more stable and transferable representations of sounding objects. Furthermore, when equipped with a stronger CLIP model, AGCL delivers an additional performance gain, reaching 45.59% unseen mIoU and 61.47% seen mIoU.

4.4 Ablation Studies and Further Discussion

We conduct ablation studies on the AVSBench-OV dataset with CLIP-ViT-B/16 as the basic model to verify the effectiveness of different designs in our method.

Component Analysis. We report the ablation study results of each proposed components of our AGCL framework in Table 2. The first row represents the baseline cost-learning paradigm without audio injection or distractor mining. In the second row, generating sound-aware cost maps with AMCG brings notable improvements

Table 2: Ablation study on different components of our AGCL framework. *SDM-A* and *SDM-S* denote mining acoustically and semantically confusing distractors, respectively.

AMCG	AGTA	SDM-A	SDM-S	mIoU			
				Unseen	Seen	Harmonic	Overall
				26.47	56.66	36.08	44.40
✓				29.11	58.66	38.91	46.91
	✓			28.44	58.29	38.23	46.67
✓	✓			30.45	60.65	40.54	48.87
✓	✓	✓		32.32	59.27	41.83	48.58
✓	✓		✓	31.81	60.30	41.65	48.72
✓	✓	✓	✓	**33.81**	**61.03**	**43.51**	**50.69**

on both seen and unseen categories. In the third row, incorporating the AGTA component in the cost-refinement phase further enhances temporal coherence by propagating sounding-region features across frames. Although its gain is slightly smaller than AMCG due to audio being injected at a later phase, the improvement in harmonic mIoU confirms that temporal consistency benefits overall segmentation stability. When combining AMCG and AGTA in the fourth row, the model achieves an overall gain of nearly +4.5% mIoU compared to the baseline, showing the superiority of our low-intrusive audio-injection mechanism, which discovers sounding objects without disturbing the vision-language alignment.

We further analyze the design of the synergistic distractor mining strategy in the last three rows. When only acoustically confusing distractors or only semantically confusing distractors are mined, the model achieves higher Unseen-mIoU, suggesting improved generalization to unseen categories. However, their performance on seen categories drops, possibly because limiting the negative scope reduces the diversity of seen category supervision and weakens intra-category compactness. By jointly considering both acoustically and semantically confusing distractors, the full AGCL framework increases the performance on seen categories and further improves unseen category segmentation, yielding the best harmonic mIoU of 43.51%. This indicates that combining complementary types of distractors exposes the model to more discriminative boundaries, thus improving open-vocabulary generalization.

Table 3: Ablation study on different approaches of incorporating audio cues for cost generation in our AMCG module.

Method	mIoU			
	Unseen	Seen	Harmonic	Overall
-	26.47	56.66	36.08	44.40
SA	26.68	57.78	36.51	45.40
CA	27.43	**58.92**	37.44	46.10
AMCG	**29.11**	58.66	**38.91**	**46.91**

Audio Cues Incorporation. We further compare different approaches for incorporating audio cues in our AMCG module in Table 3. To make visual features audio-aware while retaining vision-language alignment, we explore three low-intrusive designs: (1) SA, which concatenates audio and visual tokens in the self-attention layers, (2) CA, which injects audio as keys and values in cross-attention adapters, and (3) AMCG, which generates modulation factors with audio features. All inserted modules are zero-initialized to avoid disturbing the pretrained representation at the early training iterations. Among the three, AMCG achieves the best performance, particularly on unseen categories (+2.64% *vs.* +0.96% or +0.21%), indicating stronger open-vocabulary generalization. Compared with SA and CA, which directly alter the feature composition of CLIP representations, AMCG performs soft feature rescaling that highlights sounding regions while maintaining the original distribution.

Table 4: Ablation study on different types of guidance features used in the AGTA module.

Guidance	mIoU			
	Unseen	Seen	Harmonic	Overall
w/o AGTA	29.11	58.66	38.91	46.91
No Guide	29.27	60.00	39.34	48.00
Visual	29.70	60.01	39.73	48.19
Audio	**30.45**	**60.65**	**40.54**	**48.87**

Temporal Guidance Feature. As shown in Table 4, we ablate the types of guidance features used in the AGTA module. Compared with the baseline without temporal aggregation, using either no guidance (the second row) or visual features to guide attention (the third row) improves performance mainly on seen categories, as both variants enhance the temporal consistency of cost embeddings. However, their gain on unseen categories remains marginal. In contrast, using audio features as guidance yields the best results, especially for unseen categories, as the audio-based cross-frame similarity serves as a robust indicator of sounding continuity and enables more precise temporal aggregation.

Table 5: Evaluation results of our AGCL *w/o* training on the VPO dataset. "OOD" and "ID" denotes out-of-distribution and in-distribution, respectively. Best and Second Best.

Method	Type	VPO (MS)			VPO (MSMI)		
		MIoU↑	F_β ↑	FDR ↓	MIoU↑	F_β ↑	FDR ↓
TPAVI [56]	ID	44.08	58.14	30.82	50.37	66.80	29.82
AVSegFormer [8]	ID	56.46	71.89	24.66	50.96	64.96	32.72
CAVP [2]	ID	**61.85**	**74.60**	**20.24**	**57.22**	**72.26**	**24.04**
OV-AVSS [10]	OOD	45.04	56.42	43.51	42.30	53.18	47.59
AGCL (Ours)	OOD	<u>57.72</u>	<u>72.91</u>	<u>24.05</u>	<u>54.24</u>	<u>68.65</u>	<u>29.52</u>

Out-of-Distribution Evaluation. As shown in Table 5, we evaluate the generalization capability of our method on the VPO dataset [2] using mono audios under an Out-of-Distribution (OOD) setting. Specifically, OV-AVSS [10] and our AGCL are trained solely on the AVSBench-OV dataset using CLIP-ViT-L/14 as backbone and directly evaluated on the VPO dataset. The results indicate that our method significantly outperforms the OOD baseline OV-AVSS across both MS and MSMI subsets. Furthermore, our approach even surpasses TPAVI [56] and AVSegFormer [8], which are evaluated under an In-Distribution (ID) setting with training on the VPO dataset, achieving the second-best performance overall. Outperforming an in-distribution-trained model without target-domain fine-tuning demonstrates our method's robust generalization.

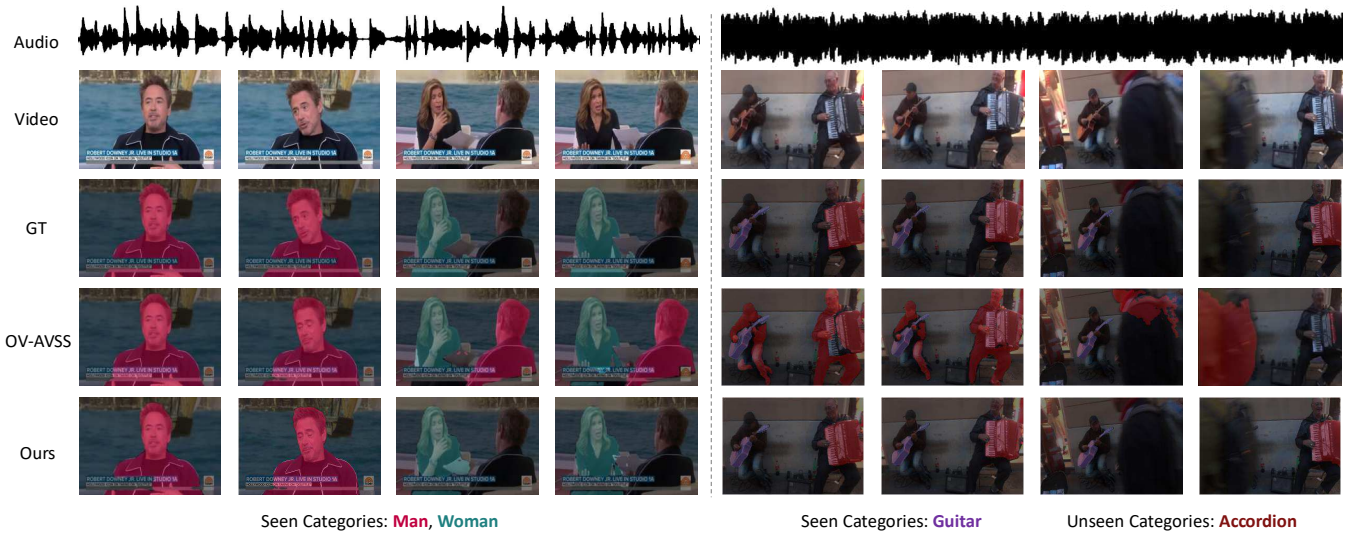


Figure 4: Qualitative comparison between OV-AVSS [10] and our AGCL on both seen and unseen categories. Sounding categories correspond to masks of the same colors.

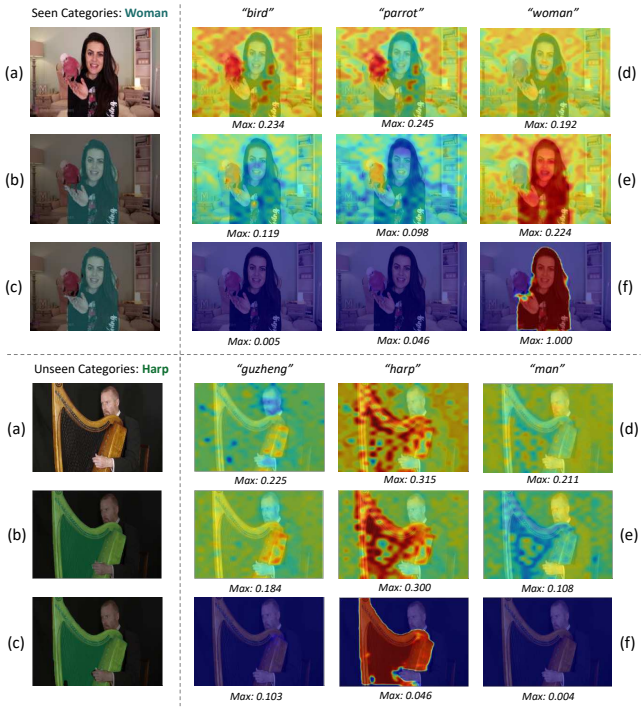


Figure 5: Visualization of cost maps. (a) Original video frame. (b) Ground-truth mask. (c) Our prediction. (d) Cost maps generated by the original CLIP model. (e) Cost maps generated in our AMCG module. (f) Cost maps after refinement.

4.5 Qualitative Analysis

Figure 4 compares our method with OV-AVSS [10]. For seen categories such as *man* and *woman*, OV-AVSS fails to identify the active speaker when the speaker changes, whereas our method accurately follows these dynamic transitions. This benefits from AGTA, which aggregates temporal context using audio similarities, and AMCG, which enhances sounding regions while suppressing silent but visually salient areas. For the unseen *accordion* category, OV-AVSS produces over-expanded masks that include surrounding people and incompletely cover the instrument, while our method segments the full instrument with clear boundaries. Figure 5 visualizes cost maps for seen and unseen categories. After AMCG, visually present but non-sounding categories are substantially suppressed, demonstrating that audio-conditioned modulation highlights sounding regions and attenuates visual distractors. In the top row, the maximum responses of "*bird*" and "*parrot*" decrease to about half of those from the original CLIP model. In the bottom row, the model distinguishes the sounding "*harp*" from the acoustically similar "*guzheng*" and the silent "*man*". Subsequent multimodal-guided refinement further suppresses non-sounding entities and concentrates responses on "*woman*" and "*harp*", leading to accurate predictions.

5 Conclusion

Open-Vocabulary Audio-Visual Semantic Segmentation (OV-AVSS) segments sound-emitting objects from both seen and unseen categories. We reformulate its foreground definition from generic soundingness to category-specific audio-visual patterns and propose Acoustically Grounded Cost Learning (AGCL) to transform static, audio-agnostic visual-text priors into dynamic, audio-grounded cost representations. Our designed Audio-Modulated Cost Generation (AMCG) and Audio-Guided Temporal Aggregation (AGTA) modules enable low-intrusive audio injection for intra-category soundingness discovery, while our introduced Synergistic Distractor Mining (SDM) strategy mines acoustically and semantically confusing negatives for inter-category discrimination. Extensive experiments on the AVSBench-OV dataset show substantial improvements of our AGCL over previous methods.

Acknowledgments

This research was supported in part by the National Natural Science Foundation of China (Grant No. 62502142, 62572166, 62302140, 62472139), the Natural Science Foundation of Anhui Province (Grant No. 2508085QF226), and the Fundamental Research Funds for the Central Universities (Grant No. JZ2025HGTA0161). The computation is completed on the HPC Platform of Hefei University of Technology.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Yuanhong Chen, Yuyuan Liu, Hu Wang, Fengbei Liu, Chong Wang, Helen Frazer, and Gustavo Carneiro. 2024. Unraveling instance associations: A closer look for audio-visual segmentation. In *CVPR*.

[3] Seokju Cho, Sunghwan Hong, and Seungryong Kim. 2022. Cats++: Boosting cost aggregation with convolutions and transformers. *TPAMI* (2022).

[4] Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. 2024. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *CVPR*.

[5] Shangzhe Di, Zeren Jiang, Si Liu, Zhaokai Wang, Leyan Zhu, Zexin He, Hongming Liu, and Shuicheng Yan. 2021. Video background music generation with controllable music transformer. In *ACM MM*.

[6] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. 2022. Decoupling zero-shot semantic segmentation. In *CVPR*.

[7] Zheng Ding, Jieke Wang, and Zhuowen Tu. 2022. Open-vocabulary universal image segmentation with maskclip. *arXiv preprint arXiv:2208.08984* (2022).

[8] Shengyi Gao, Zhe Chen, Guo Chen, Wenhai Wang, and Tong Lu. 2024. Avsegformer: Audio-visual segmentation with transformer. In *AAAI*.

[9] Pinxue Guo, Hao Huang, Peiyang He, Xuefeng Liu, Tianjun Xiao, and Wenqiang Zhang. 2025. OpenVIS: Open-vocabulary video instance segmentation. In *AAAI*.

[10] Ruohao Guo, Liao Qu, Dantong Niu, Yanyu Qi, Wenzhen Yue, Ji Shi, Bowei Xing, and Xianghua Ying. 2024. Open-vocabulary audio-visual semantic segmentation. In *ACM MM*.

[11] Jingxuan He, Lechao Cheng, Chaowei Fang, Zunlei Feng, Tingting Mu, and Mingli Song. 2024. Progressive feature self-reinforcement for weakly supervised semantic segmentation. In *AAAI*.

[12] Runze He, Shaofei Huang, Xuecheng Nie, Tianrui Hui, Luoqi Liu, Jiao Dai, Jizhong Han, Guanbin Li, and Si Liu. 2024. Customize your nerf: Adaptive source driven 3d scene editing via local-global iterative training. In *CVPR*.

[13] Sunghwan Hong, Seokju Cho, Jisu Nam, Stephen Lin, and Seungryong Kim. 2022. Cost aggregation with 4d convolutional swin transformer for few-shot segmentation. In *ECCV*.

[14] Sunghwan Hong, Jisu Nam, Seokju Cho, Susung Hong, Sangryul Jeon, Dongbo Min, and Seungryong Kim. 2022. Neural matching fields: Implicit representation of matching fields for visual correspondence. *NeurIPS* (2022).

[15] Shaofei Huang, Han Li, Yuqing Wang, Hongji Zhu, Jiao Dai, Jizhong Han, Wenge Rong, and Si Liu. 2023. Discovering sounding objects by audio queries for audio visual segmentation. In *IJCAI*.

[16] Shaofei Huang, Rui Ling, Tianrui Hui, Hongyu Li, Xu Zhou, Shifeng Zhang, Si Liu, Richang Hong, and Meng Wang. 2025. Revisiting Audio-Visual Segmentation with Vision-Centric Transformer. In *CVPR*.

[17] Shaofei Huang, Rui Ling, Hongyu Li, Tianrui Hui, Zongheng Tang, Xiaoming Wei, Jizhong Han, and Si Liu. 2025. Unleashing the temporal-spatial reasoning capacity of gpt for training-free audio and language referenced video object segmentation. In *AAAI*.

[18] Dat Huynh, Jason Kuen, Zhe Lin, Jiuxiang Gu, and Ehsan Elhamifar. 2022. Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling. In *CVPR*.

[19] Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144* (2016).

[20] Hao Ju, Shaofei Huang, Hongyu Li, Zihan Ding, Si Liu, Meng Wang, and Zhedong Zheng. 2026. From Instruction to Event: Sound-Triggered Mobile Manipulation. *arXiv preprint arXiv:2601.21667* (2026).

[21] Wouter Kool, Herke Van Hoof, and Max Welling. 2019. Stochastic beams and where to find them: The gumbel-top-k trick for sampling sequences without replacement. In *ICML*.

[22] Jiaxu Li, Songsong Yu, Yifan Wang, Lijun Wang, and Huchuan Lu. 2024. Selm: Selective mechanism based audio-visual segmentation. In *ACM MM*.

[23] Kexin Li, Zongxin Yang, Lei Chen, Yi Yang, and Jun Xiao. 2023. Catr: Combinatorial-dependence audio-queried transformer for audio-visual video segmentation. In *ACM MM*.

[24] Yongkang Li, Tianheng Cheng, Bin Feng, Wenyu Liu, and Xinggang Wang. 2025. Mask-Adapter: The Devil is in the Masks for Open-Vocabulary Segmentation. In *CVPR*.

[25] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. 2023. Open-vocabulary semantic segmentation with mask-adapted clip. In *CVPR*.

[26] Biyang Liu, Huimin Yu, and Guodong Qi. 2022. Graftnet: Towards domain generalized stereo matching with a broad-spectrum and task-oriented feature. In *CVPR*.

[27] Chen Liu, Peike Patrick Li, Xingqun Qi, Hu Zhang, Lincheng Li, Dadong Wang, and Xin Yu. 2023. Audio-visual segmentation by exploring cross-modal mutual semantics. In *ACM MM*.

[28] Jinxiang Liu, Yikun Liu, Fei Zhang, Chen Ju, Ya Zhang, and Yanfeng Wang. 2024. Audio-visual segmentation via unlabeled frame exploitation. In *CVPR*.

[29] Yong Liu, Sule Bai, Guanbin Li, Yitong Wang, and Yansong Tang. 2024. Open-vocabulary segmentation with semantic-assisted calibration. In *CVPR*.

[30] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*.

[31] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. 2016. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712* (2016).

[32] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Alban Desmaison, Andreas Kopf, Edward Gibson, Chien Yu Wang, Zachary DeVito, Martin Auer, Nikita Nikitin, Sergiy Popov, Jonathan Dynan, Colby Childers, Arnaud Lefebvre, and Soumith Chintala. 2024. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Journal of Machine Learning Research* (2024).

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

[34] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. 2017. Convolutional neural network architecture for geometric matching. In *CVPR*.

[35] Xiao Song, Guorun Yang, Xinge Zhu, Hui Zhou, Zhe Wang, and Jianping Shi. 2021. Adastereo: A simple and efficient approach for adaptive stereo matching. In *CVPR*.

[36] Peiwen Sun, Honggang Zhang, and Di Hu. 2024. Unveiling and Mitigating Bias in Audio Visual Segmentation. In *ACM MM*.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *NeurIPS* (2017).

[38] Haochen Wang, Cilin Yan, Shuai Wang, Xiaolong Jiang, Xu Tang, Yao Hu, Weidi Xie, and Efstratios Gavves. 2023. Towards open-vocabulary video instance segmentation. In *ICCV*.

[39] Yuanbin Wang, Shaofei Huang, Yulu Gao, Zhen Wang, Rui Wang, Kehua Sheng, Bo Zhang, and Si Liu. 2023. Transferring CLIP's knowledge into zero-shot point cloud semantic segmentation. In *ACM MM*.

[40] Yaoting Wang, Weisong Liu, Guangyao Li, Jian Ding, Di Hu, and Xi Li. 2024. Prompting segmentation with sound is generalizable audio-visual source localizer. In *AAAI*.

[41] Yaoting Wang, Peiwen Sun, Dongzhan Zhou, Guangyao Li, Honggang Zhang, and Di Hu. 2024. Ref-avs: Refer and segment objects in audio-visual scenes. In *ECCV*.

[42] Fangwen Wu, Jingxuan He, Yufei Yin, Yanbin Hao, Gang Huang, and Lechao Cheng. 2024. Masked collaborative contrast for weakly supervised semantic segmentation. In *WACV*.

[43] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP*.

[44] Bin Xie, Jiale Cao, Jin Xie, Fahad Shahbaz Khan, and Yanwei Pang. 2024. Sed: A simple encoder-decoder for open-vocabulary semantic segmentation. In *CVPR*.

[45] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. 2022. Groupvit: Semantic segmentation emerges from text supervision. In *CVPR*.

[46] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. 2023. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *CVPR*.

[47] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. 2023. Side adapter network for open-vocabulary semantic segmentation. In *CVPR*.

[48] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. 2022. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *ECCV*.

[49] Gengshan Yang, Joshua Manela, Michael Happold, and Deva Ramanan. 2019. Hierarchical deep stereo matching on high-resolution images. In *CVPR*.

[50] Qi Yang, Xing Nie, Tong Li, Pengfei Gao, Ying Guo, Cheng Zhen, Pengfei Yan, and Shiming Xiang. 2024. Cooperation does matter: Exploring multi-order bilateral relations for audio-visual segmentation. In *CVPR*.

[51] Jiarui Yu, Haoran Li, Yanbin Hao, Jinmeng Wu, Tong Xu, Shuo Wang, and Xiangnan He. 2023. How Can Contrastive Pre-training Benefit Audio-Visual Segmentation? A Study from Supervised and Zero-shot Perspectives.. In *BMVC*.

[52] Mingfeng Zha, Tianyu Li, Guoqing Wang, Peng Wang, Yangyang Wu, Yang Yang, and Heng Tao Shen. 2025. Implicit Counterfactual Learning for Audio-Visual Segmentation. In *ICCV*.

[53] Pingrui Zhang, Xianqiang Gao, Yuhan Wu, Kehui Liu, Dong Wang, Zhigang Wang, Bin Zhao, Yan Ding, and Xuelong Li. 2025. Moma-kitchen: A 100k+ benchmark for affordance-grounded last-mile navigation in mobile manipulation. In *ICCV*.

[54] Ziyu Zhao, Xiaoguang Li, Lingjia Shi, Nasrin Imanpour, and Song Wang. 2025. DPSeg: Dual-Prompt Cost Volume Learning for Open-Vocabulary Semantic Segmentation. In *CVPR*.

[55] Jinxing Zhou, Xuyang Shen, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, et al. 2025. Audio-visual segmentation with semantics. *IJCV* (2025).

[56] Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. 2022. Audio–visual segmentation. In *ECCV*.

[57] Wenqi Zhu, Jiale Cao, Jin Xie, Shuangming Yang, and Yanwei Pang. 2024. CLIP-VIS: Adapting CLIP for open-vocabulary video instance segmentation. *TCSVT* (2024).