This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

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

Chen Liu[1,4], Liying Yang[5], Peike Li[3], Dadong Wang[4], Lincheng Li[2], 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
yenanliu36@gmail.com, xin.yu@uq.edu.au

Abstract

Sound-guided object segmentation has drawn considerable attention for its potential to enhance multimodal perception. Previous methods primarily focus on developing advanced architectures to facilitate effective audio-visual interactions, without fully addressing the inherent challenges posed by audio natures, i.e., (1) feature confusion due to the overlapping nature of audio signals, and (2) audio-visual matching difficulty from the varied sounds produced by the same object. To address these challenges, we propose Dynamic Derivation and Elimination (DDESeg): a novel audio-visual segmentation framework. Specifically, to mitigate feature confusion, DDESeg reconstructs the semantic content of the mixed audio signal by enriching the distinct semantic information of each individual source, deriving representations that preserve the unique characteristics of each sound. To reduce the matching difficulty, we introduce a discriminative feature learning module, which enhances the semantic distinctiveness of generated audio representations. Considering that not all derived audio representations directly correspond to visual features (e.g., off-screen sounds), we propose a dynamic elimination module to filter out non-matching elements. This module facilitates targeted interaction between sounding regions and relevant audio semantics. By scoring the interacted features, we identify and filter out irrelevant audio information, ensuring accurate audio-visual alignment. Comprehensive experiments demonstrate that our framework achieves superior performance in AVS datasets. Our code is here.

1. Introduction

Audio-visual segmentation (AVS) aims to localize sound-emitting objects by generating either a binary map of audible regions or a semantic map categorizing the objects [53, 54]. Compared to other modality-guided segmentation tasks (*e.g.* text or visual) [10, 11, 42, 44], audio-driven segmentation presents unique challenges due to the inherent overlap and intra-class variability of sounds [1, 38, 51, 52]. While previous methods have focused on enhancing audio-visual interactions through learnable audio queries [17, 30, 42] or incorporating text [20, 21, 25] into audio semantics, they usually overlook the complexities introduced by audio's unique properties. In this work, we propose solutions specifically tailored to address the challenges that audio's inherent characteristics pose to AVS.

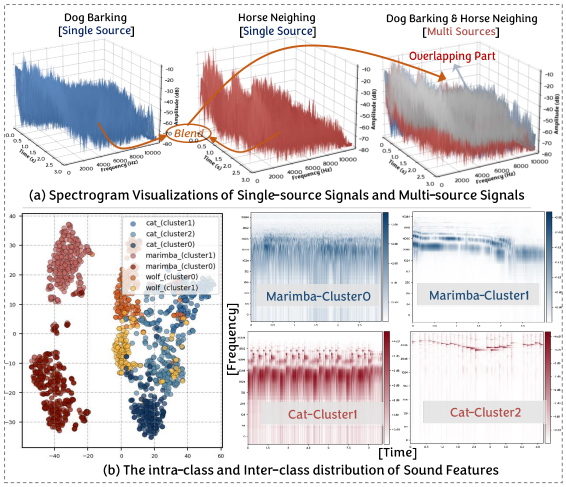


Figure 1. The illustration of ***feature confusion*** and ***audio-visual matching difficulty***. (a) ***Feature confusion*** denotes the challenge of distinguishing or separating individual sound sources in a mixed audio signal, especially with overlapping frequency, timbre, or spatial cues, which impedes accurate semantic extraction. (b) Significant amplitude and frequency variations within sounds produced by the same object lead to large intra-class variation, introducing ***audio-visual matching difficulty***. This complicates the model's ability to align audio and visual modalities.

We identify two key challenges posed by audio signals. The first is ***feature confusion due to overlapping audio signals***, where simultaneous sounds combine across time and

*Corresponding author.

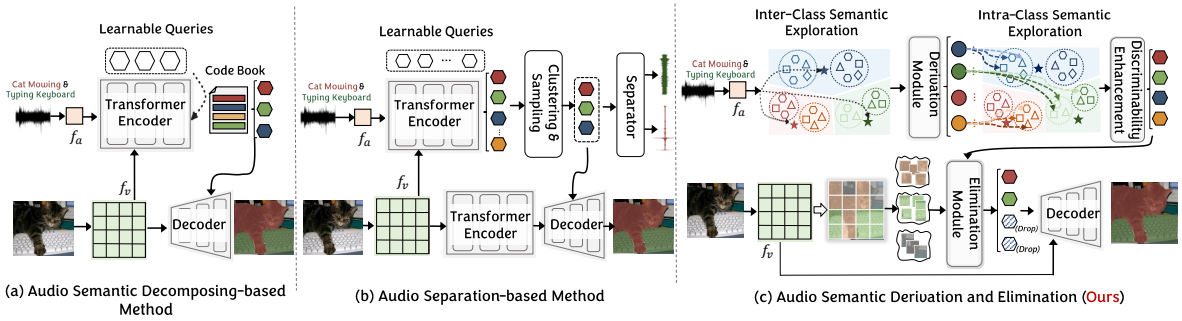


Figure 2. Three methods for achieving precise audio-visual alignment: (a) **Audio Semantic Decomposition [24]:** Models multi-source semantic space as a Cartesian product of single-source subspaces, employing product quantization and a shared codebook to decompose audio features into compact semantic tokens. (b) **Audio Separation [6]:** Devising a branch to decode audio-visual fused features into separated audio signals; (c) **Audio Semantic Derivation and Elimination (Ours):** Derives distinct semantic representations for each source from a mixed audio signal by exploring inter-class relationships. Furthermore, derived semantic representations are refined through intra-class relationships, while irrelevant audio representations are excluded under visual guidance.

frequency domains, forming a complex composite signal. As illustrated in Fig. 1 (a), the signals of ***horse neighing*** and ***dog barking*** exhibit significant overlap (*grey region*) in frequency and amplitude, complicating the isolation of unique features for each source. Consequently, extracted audio semantics may fail to capture the distinct semantic characteristics of each sound, leading to unreliable audio-visual alignment. Recent methods [6, 24] assume that sound events occur independently and attempt to decompose or separate audio semantics from audio-visual fused features, as illustrated in Fig. 2. However, different sources usually share frequencies within the spectrum, which poses challenges for these models to assign each signal component to the correct source accurately. This overlap leads to information loss, with key semantic details either diluted or missing, resulting in incomplete and unreliable representations.

The second challenge is ***audio-visual matching difficulty due to intra-class variations of sounds***. As depcited in Fig. 1 (b), sounds from the cat (● ● ●), marimba (● ●), and wolf (● ●) categories in the AVS dataset exhibit significant intra-class variation.* These variations primarily arise from considerable differences in the frequency and temporal characteristics of sounds produced by the same object [1, 38]. For example, cat sounds can be further classified into subcategories such as "`cat caterwauling`", "`cat growling`", "`cat hissing`", and "`cat meowing`". The varied sound produced by the same visual category object poses challenges for the model to consistently associate these diverse sounds with the same object, thereby hindering accurate audio-visual alignment.

In this paper, we present a novel AVS framework, Dynamic Derivation and Elimination (**DDESeg**), to address these challenges. Rather than separating or decoupling the audio semantics from the mixed one, DDESeg's Dynamic Derivation Module first identifies the possible sound semantic representations within the input audio signal. By supplementing incomplete information related to independent sound sources within the mixed audio signal, we derive multiple distinct representations, each conveying independent semantic information, as depicted in Fig. 2 (c). Once we obtain these derived audio semantic representations, we further enhance their discriminability by exploring intra-class semantic relationships. Inspired by the Generalized Laplace Operator [2, 50], our method captures feature distinctions by calculating differences and adaptively aggregating them. We then scale the learned discriminative features to refine the derived semantics. In this fashion, we enhance the discriminability of derived semantics while preserving the integrity of the original semantic space, effectively mitigating misalignment issues caused by the intra-class variation.

Furthermore, considering the sounds may not correspond to visual regions in the image (*e.g.,* off-screen sounds), we introduce the Dynamic Elimination Module (DEM) to eliminate ineffective audio semantics. Specifically, as shown in Fig. 2 (c), we first extract key semantics from the visual frame to isolate sound features associated with specific objects. The highly concentrated visual and audio semantic representations interact through an attention-based layer to generate the fused features. Afterward, we assess the relevance score of these fused features and utilize these scores to identify and weaken irrelevant audio semantics. By selectively emphasizing relevant audio features, we minimize interference from unrelated audio signals, ensuring robust audio-visual alignment.

In summary, our contributions are summarized as follows:

- We present an effective audio-visual segmentation framework to address the unique challenges in AVS arising from audio characteristics.
- We propose the DDM, which first derives multiple audio representations from the input audio signal to address the

*The features are extracted by the state-of-the-art audio classification model HTSAT [4], fine-tuned on the AVS dataset [54].

feature fusion problem, and then enhances these derived semantics by devising discriminative features to enhance the audio-visual alignment.
- We develop DEM to eliminate derived audio representations that lack a matching visual region in the frame, enabling precise audio-visual alignment.
- Extensive experiments demonstrate that our method achieves competitive segmentation performance, significantly outperforming previous state-of-the-art methods across all AVS benchmarks, *e.g.*, a 6.4% improvement in $\mathcal{J}\&\mathcal{F}_\beta$ over the second-best method on AVS-Semantic.

2. Related Work

Referring Image Segmentation. Referring Image Segmentation focuses on identifying and segmenting target objects within an image based on textual descriptions [9, 10, 16, 18, 29, 44, 45, 47]. The self-attention mechanism [40] in transformers has inspired many studies that utilize cross-attention blocks to enhance cross-modal alignment [19, 33, 43]. For instance, VLT [10] utilizes a cross-attention module to generate query vectors that capture multimodal features, which are then employed to query the image through the transformer decoder. Similarly, DMMI [16] introduces dual decoder branches to facilitate bidirectional information flow, enhancing image-text interactions.

Query-based transformer architectures [11, 37, 55, 56] have gained popularity in multimodal-guided segmentation tasks due to their ability to handle complex interactions between different data types, such as images and text. Models like X-Decoder [55], SEEM [56], and QaP [56] employ text tokens as queries, a design that has demonstrated strong performance. Inspired by these successes, some AVS architectures have adopted query-based transformers, utilizing audio tokens as queries [27, 28, 34]. However, the unique characteristics of audio, such as its overlapping and temporal nature, necessitate more specialized considerations when designing audio-visual alignment modules.

Audio Visual Segmentation. Given an audio signal, audio-visual segmentation aims to segment and identify sound-emitting objects in images or videos [5, 6, 12, 14, 17, 22, 30, 31, 35, 36, 39, 42, 48, 49, 53, 54]. Unlike images or text, which provide discrete and localized information, audio inherently introduces ambiguity due to its overlapping nature, complicating the alignment between audio cues and visual regions. Recent studies [6, 24, 27, 28] address this challenge by employing sound separation techniques to decompose mixed audio into individual sound components, or by leveraging audio classification results to identify relevant audio semantics. For instance, Liu et al. [28] and Sun et al. [39] extract semantics from mixed audio by first obtaining classification results and then applying a threshold to filter out irrelevant items from the multi-class outputs. QDFormer [24] assumes sound events are independent and utilizes a quantization-based decomposition method to separate them, while CPM [6] decodes audio-visual queries into spectrograms and applies an separation loss for supervision.

Despite these advancements, the challenges posed by the unique characteristics of audio remain unresolved. To address this, our method derives distinct, single-sounding source semantic representations from the input audio signal and explores intra-class relationships to enhance their discriminability.

3. Proposed Method

We propose **DDESeg**, a dynamic audio-visual segmentation framework that achieves precise audio-visual alignment through multi-stage cross-modal interactions (*cf.* Fig. 3 (a)). The framework consists of three key components: a dynamic derivation module (§3.1) that derives single-source audio semantic representations from the input audio signal by exploring inter- and intra-class audio semantic relationships, a dynamic elimination module (§3.2) that filters out audio representations irrelevant to the visual content, and a feature fusion module (§3.3) that progressively aligns the refined audio features with visual features.

The workflow of our framework is mathematically expressed as follows:

$$\begin{aligned} \hat{M} &= \mathcal{D}\left(\left\{\mathcal{F}_f\left(\hat{A}^l, V^l\right)\right\}_{l=1}^{L}\right), \\ \hat{A}^{(l)} &= \begin{cases} \Phi_D(F_a), & \text{if } l = 1, \\ \Phi_E(\hat{A}^{(l-1)}, V^l), & \text{if } l > 1. \end{cases} \end{aligned} \tag{1}$$

Here, $\hat{M} \in [0, C]^{H \times W}$ denotes the predicted segmentation mask for the input audio-visual pair, and L represents the number of stages in our framework. The modules Φ_D, Φ_E, $\mathcal{F}_f$, and $\mathcal{D}$ correspond to the audio semantic derivation, audio semantic elimination, feature fusion, and mask decoding operations, respectively. $\hat{A}^l$ represents the derived audio semantic representations at stage l. The visual features $V^l \in \mathbb{R}^{H_l \times W_l \times D}$ are either extracted by the visual backbone $\mathcal{E}_V$ (when $l = 1$) or obtained from the l-th feature fusion layer (when $l > 1$), where $H_l \times W_l$ denotes the spatial resolution and D is the channel dimension. F_a represents the audio feature extracted from the input audio signal. We employ the standard segmentation loss for training supervision, as detailed in §3.4.

3.1. Dynamic Derivation Module

The Dynamic Derivation Module (DDM) generates audio semantic representations with distinct single-source semantics based on the input audio signal and semantic memory (*cf.* Fig. 3 (b)). This process involves three key steps: **Step1**: Constructing a robust audio semantic memory from single-sounding source signals to capture clean and distinct semantic features. **Step2**: Retrieving the K nearest

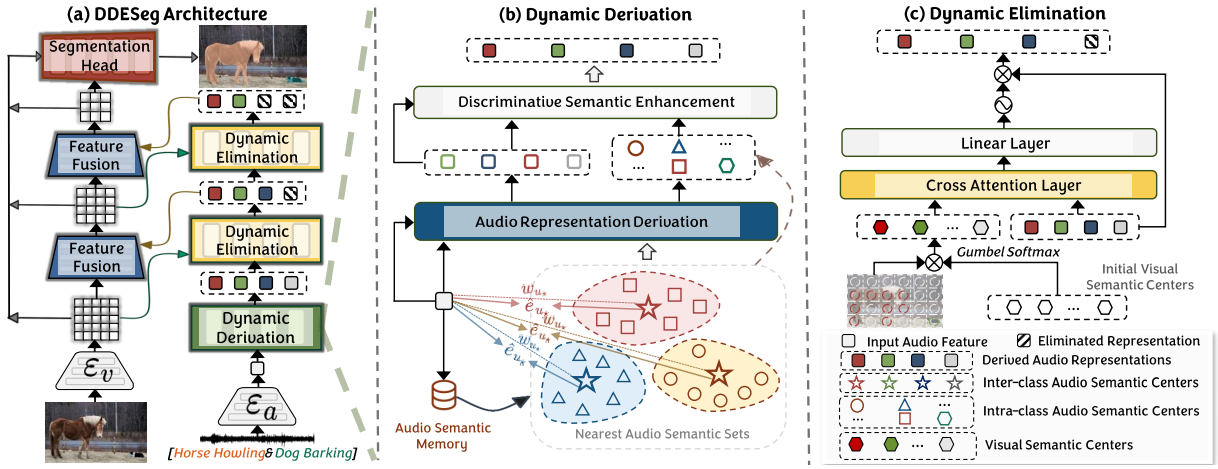


Figure 3. Overview of DDESeg architecture and its key components: (a) **Framework Architecture.** Overview of our dual-branch framework that hierarchically processes and aligns audio-visual features. Through progressive multi-stage alignment, the framework fuses cross-modal information and generates precise pixel-wise segmentation maps via the segmentation head. (b) **Dynamic Derivation Module.** This module generates multiple audio representations from the input audio feature and explores intra-class relationships to equip discriminative features for each derived representation. (c) **Dynamic Elimination Module.** DEM eliminates audio representations that do not correspond to visual regions by evaluating the relevance between audio representations and learned image semantic representations.

semantic centers of the input audio signal and using them to transform the input audio features into new representations with enhanced semantic distinctiveness; and **Step3**: Learning intra-class feature differences and adaptively refining the derived audio representations to further enhance their discriminability.

Step1: Semantic Memory Construction. To construct a robust semantic memory, we employ a pre-trained foundation model to extract audio features from single-source audio signals, providing a solid basis for capturing distinct and meaningful semantics. For each class c, we analyze its feature distribution through hierarchical clustering, starting with the global centroid μ^c, which is computed as:

$$\mu^c = \frac{1}{n_c} \sum_{i=1}^{n_c} x_i^c, \tag{2}$$

where x_i^c are the extracted audio features, and n_c is the number of features in class c. This global centroid serves as a reference for deriving audio representations in Step 2.

To capture intra-class variations, we partition each class c into k clusters $\{C_j^c\}_{j=1}^m$ via the k-means algorithm, with sub-cluster centroids μ_j^c computed as:

$$\mu_j^c = \frac{1}{|C_j^c|} \sum_{x_i^c \in C_j^c} x_i^c. \tag{3}$$

For each cluster C_j^c, we select m feature vectors closet to the cluster centroid μ_j^c as the discriminative features:

$$x_{rep_j}^c = arg \min_{x_i^c \in C_j^c, |X_j|=m} \sum_{x_i^c \in X_j} ||x_i^c - \mu_j^c||, \tag{4}$$

where $X_j \subseteq C_j^c$. The resulting discriminative features $x_{rep_j}^c \in \mathbb{R}^{m \times d}$ form a rich semantic memory that will be leveraged in Step 3 to enhance the derived audio prototypes.

Step2: Audio Prototype Derivation. The goal of this step is to derive multiple audio representations with unique audio semantics from the input audio feature F_a. Given F_a and the audio semantic memory, we first compute the Euclidean distance d_c to each class center μ^c, and identify the K nearest audio semantic centers $\{\mu_i\}_{i=1}^K$.

Inspired by the Generalized Laplace Operator, which emphasizes distinctions between the features of a central point and its neighboring points to accurately measure and adjust for semantic differences, we first compute the edge feature $e_{u_i} \in \mathbb{R}^d$ for each selected nearest class center μ_i. This edge feature captures the relationship between F_a and u_i, expressed by:

$$e_{u_i} = \phi_{\text{GELU}}(W_e(u_i - F_a) + b_e), \tag{5}$$

where $W_e \in \mathbb{R}^{d \times d}$ and $b_e \in \mathbb{R}^d$ are learnable parameters. ϕ_{GELU} is the Gaussian error linear unit activation function. Then, we compute a fusion weight w_{u_i} for each edge feature to quantify its relevance and further attain the weighted edge features $\hat{e}_{u_i} \in \mathbb{R}^d$. The process is defined as:

$$w_{u_i} = \texttt{softmax}(W_{f1} e_{u_i}), \hat{e}_{u_i} = w_{u_i} e_{u_i}, \tag{6}$$

where $W_{f1} \in \mathbb{R}^{1 \times d}$ is learnable weight matrix.

Based on the weighted edge features $\hat{e}_{u_i}$, we compute the compensation $\Delta a_{u_i} \in \mathbb{R}^d$ to bridge the semantic discrepancy between the input audio feature F_a and class centers. This is mathematically formulated as:

$$\Delta a_{u_i} = \phi_{\text{tanh}}(W_{o1}(F_a + \hat{e}_{u_i}) + b_{o1}), \tag{7}$$

where $W_{o1} \in \mathbb{R}^{d\times d}$ and $b_{o1} \in \mathbb{R}^d$ are learnable parameters. The derived audio representation a_i is then given by:

$$a_i = F_a + \Delta a_{u_i}. \tag{8}$$

Leveraging this design, we generate K distinct audio representations $A = \{a_i\}_{i=1}^K \in \mathbb{R}^{K\times d}$ from the input feature F_a through precise, semantically coherent adjustments.

Step3: Discriminative Semantic Enhancement. In this step, our goal is to capture the subtle differences between derived audio semantics and discriminative intra-class audio features to mitigate semantic bias arising from large intra-class variations. Specifically, for each derived audio representation a_i, we compute the edge features $e_{c_i,j} \in \mathbb{R}^{m\times d}$ between a_i and the discriminative intra-class feature $x^c_{rep_j}$, defined as follows:

$$e_{c_i,j} = \phi_{\text{GELU}}(W_d(x^c_{rep_j} - a_i) + b_d), \tag{9}$$

where $W_d \in \mathbb{R}^{d\times d}$ and $b_d \in \mathbb{R}^d$ are learnable parameters.

Next, we calculate the fusion weights for the intra-class edge features, resulting in edge weights $w_{c_i,j} \in \mathbb{R}^{1\times m}$ and the corresponding weighted edge features $\hat{e}_{c_i,j} \in \mathbb{R}^{1\times d}$, expressed as:

$$w_{c_i,j} = \texttt{softmax}(W_{f2}e_{c_i,j}), \quad \hat{e}_{c_i,j} = w_{c_i,j}\cdot e_{c_i,j}, \tag{10}$$

where $W_{f2} \in \mathbb{R}^{1\times d}$ is a learnable weight matrix.

Afterwards, we compute the offset vectors $\Delta a_{c_i,j} \in \mathbb{R}^{1\times d}$ for intra-class refinement, defined as:

$$\Delta a_{c_i,j} = \phi_{\text{tanh}}(W_{o2}(\hat{e}_{c_i,j}) + b_{o2}), \tag{11}$$

where $W_{o2} \in \mathbb{R}^{d\times d}$ and $b_{o2} \in \mathbb{R}^d$ are learnable parameters. Finally, we enhance each derived audio representation by incorporating the learned offset. Unlike the additive adjustment applied in Step 2, this step refines $\hat{a}_i$ by scaling its features to emphasize discriminative aspects without deviating from the original semantic space, as defined by:

$$\hat{a}_i = a_i \odot (1 + \Delta a_{c_i,j}), \tag{12}$$

where $\odot$ denotes element-wise multiplication, and $\hat{a}_i \in \mathbb{R}^{1\times d}$ represents the refined audio semantics. After this process, we obtain the set of enhanced audio representations $\hat{A} = \{\hat{a}_i\}_{i=1}^K \in \mathbb{R}^{K\times d}$. Notably, we employ addition to adjust inter-class semantic gaps, as it facilitates shifts between categories. In contrast, we apply scaling for intra-class enhancement, as it preserves the adjusted features within the same class.

3.2. Dynamic Elimination Module

The Dynamic Elimination Module addresses a key challenge in audio-visual segmentation: ensuring that derived audio representations correspond to actual sound-emitting objects in the visual scene. Given derived audio representation $\hat{A} \in \mathbb{R}^{K\times d}$ and visual features $V \in \mathbb{R}^{H\times W\times d}$, this module leverages visual guidance to filter out irrelevant audio tokens, enhancing the discriminability of audio-visual associations (*cf.* Fig. 3 (c)). This filtering step is crucial for achieving consistent audio-visual alignment.

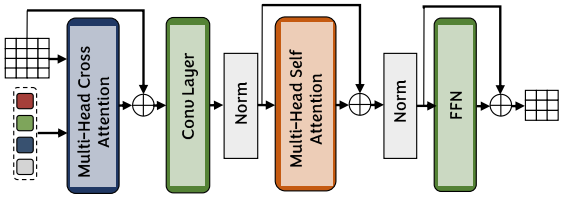


Figure 4. Structure of the Feature Fusion Block in DDESeg.

Soft Clustering of Visual Features. To obtain compact yet representative visual features, we employ a soft clustering approach with K visual semantic centers $C_v \in \mathbb{R}^{K\times d}$. The assignment of visual features to these centers is computed through a Gumbel-Softmax operation:

$$O = \frac{\exp((VC_v^T) + g)/\tau}{\sum_{k=1}^N \exp((VC_v^T) + g)/\tau}, \tag{13}$$

where $O \in \mathbb{R}^{HW\times K}$ represents the assignment matrix, g consists of i.i.d. random samples from a `Gumbel(0, 1)` distribution, and τ serving as a temperature parameter.

The visual semantic centers are then refined through weighted aggregation:

$$C_v = OC_v. \tag{14}$$

Semantic Scoring and Elimination. To evaluate the relevance of each audio token to the visual centers, we introduce a discriminative scoring mechanism. Specifically, we first perform cross-modal interaction between the audio features $\hat{A}$ and the visual centers C_v, producing an audio-visual interacted feature $F_{av} \in \mathbb{R}^{k\times d}$. This process is formulated as:

$$F_{av} = \texttt{MCA}(\hat{A}, C_v). \tag{15}$$

where MCA denotes Multi-head Cross-Attention.

After that, we pass the interacted feature F_{av} through a linear layer followed by a sigmoid activation function to obtain the discriminative score $S \in [0, 1]^k$ for each derived audio semantic, expressed by:

$$S = \sigma(\texttt{MLP}(F_{av})). \tag{16}$$

These scores serve as weights to weaken audio representations where no visual match can be found. The updated process is formulated as follows:

$$\hat{A} = S \cdot \hat{A}. \tag{17}$$

This approach ensures that relevant audio semantics contribute more significantly to the audio-visual fused features.

Table 1. Quantitative comparisons on the AVS-Objects and AVS-Semantic datasets (see analysis details in §4.2). 'Trans.': Transformer-based architecture. Best results in **Bold**, while second best underlined.

Methods	Visual Backbone	Audio Backbone	*AVS-Objects-S4*			*AVS-Objects-MS3*			*AVS-Semantic*		
			$\mathcal{J}\&\mathcal{F}_m \uparrow$	$\mathcal{J} \uparrow$	$\mathcal{F}_m \uparrow$	$\mathcal{J}\&\mathcal{F}_m \uparrow$	$\mathcal{J} \uparrow$	$\mathcal{F}_m \uparrow$	$\mathcal{J}\&\mathcal{F}_m \uparrow$	$\mathcal{J} \uparrow$	$\mathcal{F}_m \uparrow$
TPAVI [53] [ECCV22]	PVT-V2-B5	VGGish	83.3	78.7	87.9	59.3	54.0	64.5	32.5	29.8	35.2
AVSC [27] [ACM-MM23]	Swin-Base	VGGish	85.0	81.3	88.6	62.6	59.5	65.8	-	-	-
CATR [22] [ACM-MM23]	PVT-V2-B5	VGGish	87.9	84.4	91.3	68.6	62.7	74.5	35.7	32.8	38.5
DiffusionAVS [35] [Arxiv23]	PVT-V2-B5	VGGish	85.7	81.4	90.0	64.6	58.2	70.9	-	-	-
ECMVAE [36] [ICCV23]	PVT-V2-B5	VGGish	85.9	81.7	90.1	64.3	57.8	70.8	-	-	-
AuTR [30] [Arxiv23]	PVT-V2-B5	VGGish	82.1	77.6	86.5	72.0	66.2	77.7	-	-	-
AQFormer [17] [IJCAI23]	PVT-V2-B5	VGGish	85.5	81.6	89.4	67.5	62.2	72.7	-	-	-
AVSegFormer [12] [AAAI24]	PVT-V2-B5	VGGish	86.8	83.1	90.5	67.2	61.3	73.0	40.1	37.3	42.8
AVSBG [14] [AAAI24]	PVT-V2-B5	VGGish	86.1	81.7	90.4	61.0	55.1	66.8	-	-	-
GAVS [42] [AAAI24]	ViT-Base	VGGish	85.1	80.1	90.0	70.6	63.7	77.4	-	-	-
BAVS [28] [TMM24]	Swin-Base	Beats	86.2	82.7	89.8	62.8	59.6	65.9	35.6	33.6	37.5
COMBO [49] [CVPR24]	PVT-V2-B5	VGGish	88.3	84.7	91.9	65.2	59.2	71.2	44.1	42.1	46.1
QDFormer [24] [CVPR24]	Swin-Tiny	VGGish	83.9	79.5	88.2	64.0	61.9	66.1	-	53.4	-
CAVP [5] [CVPR24]	PVT-V2-B5	VGGish	90.5	87.3	93.6	72.7	67.3	78.1	55.3	48.6	62.0
AVSStone† [34] [ECCV24]	Swin-Base	VGGish	87.3	83.2	91.3	72.5	67.3	77.6	61.5	56.8	66.2
BiasAVS [39] [ACM-MM24]	Swin-Base	VGGish	88.2	83.3	93.0	74.0	67.2	80.8	47.2	44.4	49.9
DDESEG (OURS)	PVT-V2-B5	VGGish	92.0	90.3	92.2	74.7	69.7	79.6	65.7	60.8	70.6
DDESEG (OURS)	Trans.	VGGish	92.7	91.8	93.5	75.6	70.1	81.0	67.0	62.3	71.6
DDESEG (OURS)	Trans.	HTSAT	**94.2**	**92.4**	**95.9**	**77.9**	**72.3**	**83.4**	**67.9**	**63.4**	**72.3**

3.3. Feature Fusion

The Feature Fusion module performs hierarchical cross-modal interaction through multiple fusion blocks (*cf.* Fig. 4). Each block employs a cross-attention architecture, with derived audio representations serving as queries and visual features as keys and values, enabling precise audio-visual alignment. After the cross-attention operation, a convolutional layer reduces spatial dimensions, distilling essential semantic information as in [7, 8]. The module concludes with a self-attention mechanism and feed-forward network to capture global contextual relationships, ensuring comprehensive feature integration across modalities.

3.4. Loss Function

Given the predicted segmentation map $\hat{m}$ and its corresponding ground truth m, we optimize our network using a weighted combination of three complementary loss terms:

$$\mathcal{L}(m, \hat{m}) = \lambda_{dice}\mathcal{L}_{dice} + \lambda_{bce}\mathcal{L}_{bce} + \lambda_{iou}L_{iou}, \quad (18)$$

where $\mathcal{L}_{dice}$ [23] and $\mathcal{L}_{iou}$ [50] both optimize region-based overlap metrics from different perspectives, while $\mathcal{L}_{bce}$ provides pixel-wise supervision. The balancing weights λ_{dice}, λ_{bce}, and λ_{iou} are set to 5, 5, and 2, respectively.

4. Experiments

Architecture Details. Our framework includes four feature fusion modules, each containing 3, 6, 40, and 3 fusion blocks, respectively. The feature fusion module is modified based on MiT [46]. For audio feature extraction, we adopt HTSAT [4], a state-of-the-art audio classification model pre-trained on the large-scale AudioSet dataset [13], which provides robust audio representations across diverse acoustic environments.

Metrics. We employ the Jaccard Index ($\mathcal{J}$) and F-score ($\mathcal{F}_\beta$) to evaluate model performance, following the setup in [5], with β set to 0.3.

Datasets. We evaluate our approach on three benchmark datasets for audio-visual segmentation:

- **AVS-Object** [53] includes two subsets: Single Sound Source (S4) with 4,932 five-second video clips and Multi Sound Sources (MS3) with 424 five-second clips. For each video, the last frame of each second is paired with the corresponding audio signal as input. Ground truth is a binary mask, where 1 indicates the sounding region.
- **AVS-Semantic** [54] comprises 12,356 video clips across 71 categories, each lasting 5s or 10s. A semantic mask is provided as ground truth for each audio-visual pair, identifying both the sounding region and the audio event type.
- **VPO** [5] consists of audio-visual pairs created from single images in COCO [26] combined with three-second audio signals from VGGSound [3], with semantic masks provided as ground truth. It covers 21 categories and includes three subsets: VPO-SS (12,202 pairs with a single sound source), VPO-MS (9,817 pairs with multiple sound sources), and VPO-MSMI (12,855 pairs with multiple silent and audible objects of the same category).

4.1. Quantitative Results

Performance on AVS-Object and AVS-Semantic. Table 1 compares our method with previous state-of-the-art approaches on AVS-Object and AVS-Semantic using transformer-based visual backbones. Our method shows substantial improvements across all evaluation metrics, achieving increases in $\mathcal{J}\&\mathcal{F}_\beta$ compared to the second-best method: from 90.5% to 94.2% on AVS-Object-S4, 74.0% to 77.9% on AVS-Object-MS3, and 61.5% to 67.9% on AVS-Semantic. These gains highlight our method's capacity for

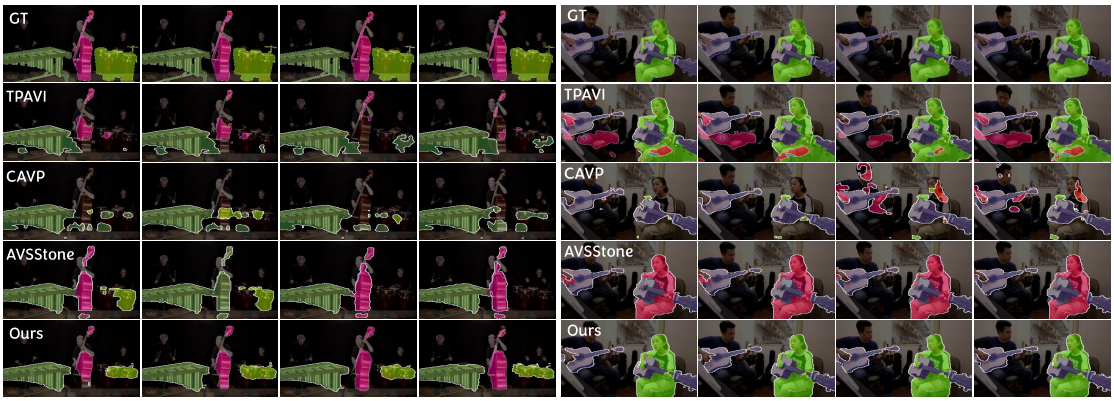


Figure 5. Qualitative results on AVSBench-Semantic (§4.2). Our method achieves precise localization in multi-source cases.

Table 2. Quantitative comparisons on the VPO datasets (§4.2). 'CPM*' indicates that training and testing are performed on images with the original resolution, while other methods employ images resized to 224 × 224 pixels. Best results in **Bold**, while second best underlined.

Methods	Visual Backbone	Audio Backbone	Trainable Params	*VPO-SS*			*VPO-MS*			*VPO-MSMI*		
				$\mathcal{J}\&\mathcal{F}_\beta\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}_\beta\uparrow$	$\mathcal{J}\&\mathcal{F}_\beta\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}_\beta\uparrow$	$\mathcal{J}\&\mathcal{F}_\beta\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}_\beta\uparrow$
TPAVI [53] [ECCV22]	PVT-V2-B5	VGGish	101.32M	44.63	41.64	47.62	45.68	42.3	49.06	43.19	40.03	46.34
AVSegFormer [12] [AAAI24]	PVT-V2-B5	VGGish	186.05M	45.94	43.81	48.06	43.72	47.3	40.14	49.93	47.19	52.67
CAVP [5] [CVPR24]	ResNet50	VGGish	119.79M	67.02	58.81	75.23	61.32	53.24	69.39	56.48	48.18	64.78
BiasAVS [39] [ECCV24]	SwinBase	VGGish	107.12M	67.46	59.14	75.78	63.42	55.61	71.23	57.94	49.6	66.27
AVSStone [34] [ACM-MM24]	SwinBase	VGGish	114.63M	68.54	59.72	77.35	64.26	56.23	72.29	58.76	50.11	67.40
CPM* [6] [ECCV24]	ResNet50	VGGish	-	73.49	67.09	79.88	72.91	65.91	79.9	68.07	60.55	75.58
DDESeg (Ours)	Trans.	VGGish	103.84M	73.78	67.02	80.54	73.62	67.01	80.23	67.90	61.78	74.01
DDESeg (Ours)	Trans.	HTSAT	103.84M	**74.38**	**67.55**	**81.20**	**74.30**	**67.64**	**80.96**	**68.39**	**62.11**	**74.67**

precise sound localization, even in complex multi-source semantic segmentation scenarios. Note that our visual backbone has a comparable parameter (85M) to Swin-Base [32] (88M) and PVT-V2-B5 [41] (82M). To further validate the effectiveness of our approach, we also adopt PVT-V2-B5 as the visual encoder, which achieves consistent superior performance across all evaluation metrics.

Additionally, we replace VGGish [15] with the state-of-the-art audio classification model HTSAT [4] as the audio backbone, leading to performance improvements of 1.5% in $\mathcal{J}$ and 0.7% in $\mathcal{F}_\beta$. This indicates that enhanced audio representations can positively impact sound localization.

Performance on VPO. We evaluate recent state-of-the-art (SOTA) methods on the VPO dataset equipped with synthesized stereo audios. As illustrated in Table 2, DDESeg consistently outperforms all existing methods. Since the code of CPM [6] is unavailable, we report its results at the original image resolution, denoted as CPM* in Table 2. Despite being trained on lower-resolution images (224×224 pixels), our model still outperforms CPM. Moreover, DDESeg delivers significant performance gains over other methods with comparable parameter counts and identical input resolution. Specifically, compared to the previous best method AVSStone [34], DDESeg achieves 6%, 10.04%, and 4.9% higher $\mathcal{J}\&\mathcal{F}_\beta$ scores on the SS, MS, and MSMI sub-datasets, respectively. These substantial improvements on the VPO datasets underscore the effectiveness of our audio enhancement method.

4.2. Qualitative Results

Fig. 5 presents qualitative comparisons between our method and recent state-of-the-art approaches on AVS-Semantic. In challenging multi-source scenarios, our method accurately segments multiple simultaneously sounding objects, whereas other methods suffer from segmentation omissions. Notably, AVSStone fails to segment the `'drum'` in the first frame and misclassifies the `'cello'` as a `'marimba'` in the second frame. These results highlight the effectiveness of our approach in attaining high-quality audio semantic representations. We posit that deriving clearer sound semantics not only enhances localization accuracy but also ensures more reliable sound class identification. More visualizations are provided in the *Supplementary Material*.

4.3. Further Analysis

Ablation Study. For detailed analysis, we conduct a series of ablation studies on AVS-Objects-MS3 and AVS-Semantic, as shown in Table 3. We first establish a baseline performance without DDM and DEM. Building on this baseline, the integration of audio representation derivation (Derr.) has a significant improvement, improving the $\mathcal{J}\&\mathcal{F}_\beta$ score by 4.6% on AVS-Semantic, validating the effectiveness of our representation derivation design. Adding discriminative semantic enhancement further boosts performance, providing an additional 2.95% improvement in $\mathcal{J}\&\mathcal{F}_\beta$ on AVS-Semantic. On this basis, incorporating DEM yields a further 2.33% increase in the $\mathcal{J}$ score and

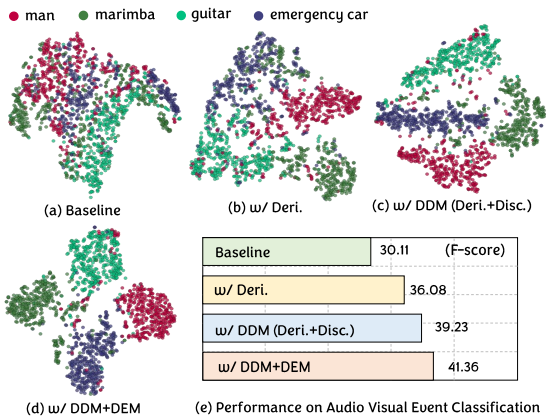


Figure 6. Qualitative and quantitative comparisons of audio semantic quality. (a)-(c) show t-SNE visualizations of audio features under different settings. (d) presents the performance of employing audio-visual interacted semantics from different settings in audio-visual event classification. See §4.3 for analysis details.

Table 3. Ablation study (see details in §4.3) on AVS datasets.

Method	*AVS-Objects-MS3*			*AVS-Semantic*		
	$\mathcal{J}\&\mathcal{F}_\beta\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}_\beta\uparrow$	$\mathcal{J}\&\mathcal{F}_\beta\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}_\beta\uparrow$
Baseline	65.81	60.03	71.59	58.00	52.69	63.31
Deri.	71.51	65.20	77.81	62.60	57.14	68.06
DEM (Deri.+Disc.)	73.40	67.50	79.30	65.55	61.07	70.03
DEM+DDM (Ours)	**75.57**	**70.10**	**81.03**	**67.85**	**63.40**	**72.30**

a 2.27% increase in the $\mathcal{F}_\beta$ score. This confirms that our dynamic elimination design effectively reduces interference from non-visually matching audio representations.

Quality of Generated Audio Representations. Fig. 6 presents the qualitative and quantitative evaluations of audio representation quality across different settings. A comparison between (a) and (b) shows a distinct clustering trend in audio features with the introduction of audio representation derivation (*w/* Deri.). This clustering effect is further strengthened when both audio representation derivation and discriminative semantic enhancement are applied. Notably, the intra-class distribution in (b) reveals sub-clustering patterns, as observed in classes such as `'marimba'` and `'emergency car'`. These variations are effectively mitigated in (c), where the dynamic elimination module (*w/* DDM+DEM) results in more pronounced class-wise clustering, underscoring the substantial improvements our modules bring to audio quality.

To quantitatively validate our semantic representations, we conducted linear probing experiments by assessing their performance in audio-visual event classification. Using features from each model variant as input to a linear classifier, we observe that the *w/* DDM+DEM configuration achieved an 11.25% F-score improvement over the baseline. This notable performance increase in linear probing provides further evidence that our approach generates high-quality audio-visual representations.

Analysis of Elimination Designs. We systematically investigate various elimination designs to optimize audio-visual correlation learning, with results presented in Table 4. This analysis explores five progressive architectural variants:

Table 4. Further study of elimination designs (§4.3). FC: Fully Connected layer, CA: Cross-Attention Layer, SK: Soft-Kmeans, and GS: Gumbel Softmax.

	Scheme	*AVS-Objects-MS3*			*AVS-Semantic*		
		$\mathcal{J}\&\mathcal{F}_\beta\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}_\beta\uparrow$	$\mathcal{J}\&\mathcal{F}_\beta\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}_\beta\uparrow$
①	wo/Elimination	73.4	67.5	79.3	65.6	61.1	70.0
②	FC+Sigmoid	72.1	66.7	77.4	64.0	59.8	68.2
③	CA+FC+Sigmoid	74.4	68.4	80.3	65.9	61.8	70.0
④	SK+CA+FC+Sigmoid	74.8	69.0	80.6	66.5	62.1	70.9
⑤	GS+CA+FC+Sigmoid	**77.9**	**72.3**	**83.4**	**67.9**	**63.4**	**72.3**

- *Scheme* ① (Baseline): Directly interacts derived audio semantic representations with visual feature maps through a feature fusion module.
- *Scheme* ②: Incorporates the FC-Sigmoid filtering mechanism, which results in degraded performance.
- *Scheme* ③: Enhances audio-visual correlation by applying cross-attention prior to filtering, yielding a significant improvement of 0.3% in $\mathcal{J}\&\mathcal{F}_\beta$ compared to Scheme ①.
- *Scheme* ④: Adds Soft-Kmeans to attain more compact visual semantic representations, leading to a 0.6% improvement in $\mathcal{J}\&\mathcal{F}_\beta$ over Scheme ③.
- *Scheme* ⑤: Replaces Soft-Kmeans with Gumbel Softmax for enhanced visual semantic compactness, achieving a 1.4% improvement in $\mathcal{J}\&\mathcal{F}_\beta$ over scheme ④.

As the results suggested, Scheme ⑤ attains the best performance, demonstrating that visual semantic compaction is essential for robust audio-visual correlation learning.

5. Conclusion

In this paper, we propose an effective framework for audio-visual segmentation that addresses challenges posed by complex audio characteristics. Central to our approach are two key modules: the Dynamic Derivation Module (DDM), which generates multiple audio representations with distinct semantics, and the Dynamic Elimination Module (DEM), which filters out non-visually matching audio cues. Within the DDM, the derivation module addresses issues in feature fusion due to the overlapping nature of audio, while the discriminative semantic enhancement component mitigates the audio-visual matching difficulty caused by great intra-class variations. Leveraging DEM, our framework achieves precise audio-visual alignment. Extensive experiments on three benchmark datasets demonstrate that our approach significantly advances the state-of-the-art, showing particular strength in challenging multi-source scenarios.

Acknowledgements. This work is supported by ARC-Discovery (DP220100800 to XY) and ARC-DECRA (DE230100477 to XY). Chen Liu is funded by the China Scholarship Council and CSIRO top-up (50092128).

References

[1] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of the European conference on computer vision (ECCV)*, pages 435–451, 2018. 1, 2

[2] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003. 2

[3] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020. 6

[4] Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 646–650. IEEE, 2022. 2, 6, 7

[5] Yuanhong Chen, Yuyuan Liu, Hu Wang, Fengbei Liu, Chong Wang, Helen Frazer, and Gustavo Carneiro. Unraveling instance associations: A closer look for audio-visual segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26497–26507, 2024. 3, 6, 7

[6] Yuanhong Chen, Chong Wang, Yuyuan Liu, Hu Wang, and Gustavo Carneiro. Cpm: Class-conditional prompting machine for audio-visual segmentation. *arXiv preprint arXiv:2407.05358*, 2024. 2, 3, 7

[7] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. 2021. 6

[8] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. 2022. 6

[9] Yong Xien Chng, Henry Zheng, Yizeng Han, Xuchong Qiu, and Gao Huang. Mask grounding for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26573–26583, 2024. 3

[10] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16321–16330, 2021. 1, 3

[11] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. Mevis: A large-scale benchmark for video segmentation with motion expressions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2694–2703, 2023. 1, 3

[12] Shengyi Gao, Zhe Chen, Guo Chen, Wenhai Wang, and Tong Lu. Avsegformer: Audio-visual segmentation with transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12155–12163, 2024. 3, 6, 7

[13] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017. 6

[14] Dawei Hao, Yuxin Mao, Bowen He, Xiaodong Han, Yuchao Dai, and Yiran Zhong. Improving audio-visual segmentation with bidirectional generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2067–2075, 2024. 3, 6

[15] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson. Cnn architectures for large-scale audio classification, 2017. 7

[16] Yutao Hu, Qixiong Wang, Wenqi Shao, Enze Xie, Zhenguo Li, Jungong Han, and Ping Luo. Beyond one-to-one: Rethinking the referring image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4067–4077, 2023. 3

[17] Shaofei Huang, Han Li, Yuqing Wang, Hongji Zhu, Jiao Dai, Jizhong Han, Wenge Rong, and Si Liu. Discovering sounding objects by audio queries for audio visual segmentation. *arXiv preprint arXiv:2309.09501*, 2023. 1, 3, 6

[18] Tianrui Hui, Si Liu, Zihan Ding, Shaofei Huang, Guanbin Li, Wenguan Wang, Luoqi Liu, and Jizhong Han. Language-aware spatial-temporal collaboration for referring video segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8646–8659, 2023. 3

[19] Namyup Kim, Dongwon Kim, Cuiling Lan, Wenjun Zeng, and Suha Kwak. Restr: Convolution-free referring image segmentation using transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18145–18154, 2022. 3

[20] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19108–19118, 2022. 1

[21] Guangyao Li, Wenxuan Hou, and Di Hu. Progressive spatio-temporal perception for audio-visual question answering. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7808–7816, 2023. 1

[22] Kexin Li, Zongxin Yang, Lei Chen, Yi Yang, and Jun Xiao. Catr: Combinatorial-dependence audio-queried transformer for audio-visual video segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1485–1494, 2023. 3, 6

[23] Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. Dice loss for data-imbalanced nlp tasks. *arXiv preprint arXiv:1911.02855*, 2019. 6

[24] Xiang Li, Jinglu Wang, Xiaohao Xu, Xiulian Peng, Rita Singh, Yan Lu, and Bhiksha Raj. Qdformer: Towards robust audiovisual segmentation in complex environments with quantization-based semantic decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3402–3413, 2024. 2, 3, 6

[25] Zhangbin Li, Dan Guo, Jinxing Zhou, Jing Zhang, and Meng Wang. Object-aware adaptive-positivity learning for audio-visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3306–3314, 2024. 1

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 6

[27] Chen Liu, Peike Patrick Li, Xingqun Qi, Hu Zhang, Lincheng Li, Dadong Wang, and Xin Yu. Audio-visual segmentation by exploring cross-modal mutual semantics. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7590–7598, 2023. 3, 6

[28] Chen Liu, Peike Li, Hu Zhang, Lincheng Li, Zi Huang, Dadong Wang, and Xin Yu. Bavs: bootstrapping audio-visual segmentation by integrating foundation knowledge. *IEEE Transactions on Multimedia*, 2024. 3, 6

[29] Chang Liu, Xiangtai Li, and Henghui Ding. Referring image editing: Object-level image editing via referring expressions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13128–13138, 2024. 3

[30] Jinxiang Liu, Chen Ju, Chaofan Ma, Yanfeng Wang, Yu Wang, and Ya Zhang. Audio-aware query-enhanced transformer for audio-visual segmentation. *arXiv preprint arXiv:2307.13236*, 2023. 1, 3, 6

[31] Jinxiang Liu, Yu Wang, Chen Ju, Chaofan Ma, Ya Zhang, and Weidi Xie. Annotation-free audio-visual segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5604–5614, 2024. 3

[32] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021. 7

[33] Zhuoyan Luo, Yicheng Xiao, Yong Liu, Shuyan Li, Yitong Wang, Yansong Tang, Xiu Li, and Yujiu Yang. Soc: Semantic-assisted object cluster for referring video object segmentation. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[34] Juncheng Ma, Peiwen Sun, Yaoting Wang, and Di Hu. Stepping stones: A progressive training strategy for audio-visual semantic segmentation. *arXiv preprint arXiv:2407.11820*, 2024. 3, 6, 7

[35] Yuxin Mao, Jing Zhang, Mochu Xiang, Yunqiu Lv, Yiran Zhong, and Yuchao Dai. Contrastive conditional latent diffusion for audio-visual segmentation. *arXiv preprint arXiv:2307.16579*, 2023. 3, 6

[36] Yuxin Mao, Jing Zhang, Mochu Xiang, Yiran Zhong, and Yuchao Dai. Multimodal variational auto-encoder based audio-visual segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 954–965, 2023. 3, 6

[37] Bo Miao, Mohammed Bennamoun, Yongsheng Gao, and Ajmal Mian. Spectrum-guided multi-granularity referring video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 920–930, 2023. 3

[38] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 631–648, 2018. 1, 2

[39] Peiwen Sun, Honggang Zhang, and Di Hu. Unveiling and mitigating bias in audio visual segmentation. *arXiv preprint arXiv:2407.16638*, 2024. 3, 6, 7

[40] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 3

[41] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022. 7

[42] Yaoting Wang, Weisong Liu, Guangyao Li, Jian Ding, Di Hu, and Xi Li. Prompting segmentation with sound is generalizable audio-visual source localizer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5669–5677, 2024. 1, 3, 6

[43] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11686–11695, 2022. 3

[44] Dongming Wu, Wencheng Han, Tiancai Wang, Xingping Dong, Xiangyu Zhang, and Jianbing Shen. Referring multi-object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14633–14642, 2023. 1, 3

[45] Jianzong Wu, Xiangtai Li, Xia Li, Henghui Ding, Yunhai Tong, and Dacheng Tao. Towards robust referring image segmentation. *IEEE Transactions on Image Processing*, 2024. 3

[46] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34: 12077–12090, 2021. 6

[47] Li Xu, Mark He Huang, Xindi Shang, Zehuan Yuan, Ying Sun, and Jun Liu. Meta compositional referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19478–19487, 2023. 3

[48] Shilin Yan, Renrui Zhang, Ziyu Guo, Wenchao Chen, Wei Zhang, Hongyang Li, Yu Qiao, Hao Dong, Zhongjiang He, and Peng Gao. Referred by multi-modality: A unified temporal transformer for video object segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6449–6457, 2024. 3

[49] Qi Yang, Xing Nie, Tong Li, Pengfei Gao, Ying Guo, Cheng Zhen, Pengfei Yan, and Shiming Xiang. Cooperation does matter: Exploring multi-order bilateral relations for audio-visual segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27134–27143, 2024. 3, 6

[50] Jiahui Yu, Yuning Jiang, Zhangyang Wang, Zhimin Cao, and Thomas Huang. Unitbox: An advanced object detection network. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 516–520, 2016. 2, 6

[51] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of

pixels. In *Proceedings of the European conference on computer vision (ECCV)*, pages 570–586, 2018. 1

[52] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1735–1744, 2019. 1

[53] Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio–visual segmentation. In *European Conference on Computer Vision*, pages 386–403. Springer, 2022. 1, 3, 6, 7

[54] Jinxing Zhou, Xuyang Shen, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, et al. Audio-visual segmentation with semantics. *International Journal of Computer Vision*, pages 1–21, 2024. 1, 2, 3, 6

[55] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15116–15127, 2023. 3

[56] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *Advances in Neural Information Processing Systems*, 36, 2024. 3