# Robust Audio-Visual Segmentation via Audio-Guided Visual Convergent Alignment

Chen Liu[1,4], Peike Li[3], Liying Yang[5], 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

*Accurately localizing audible objects based on audio-visual cues is the core objective of audio-visual segmentation. Most previous methods emphasize spatial or temporal multi-modal modeling, yet overlook challenges from ambiguous audio-visual correspondences—such as nearby visually similar but acoustically different objects and frequent shifts in objects' sounding status. Consequently, they may struggle to reliably correlate audio and visual cues, leading to over- or under-segmentation. To address these limitations, we propose a novel framework with two primary components: an audio-guided modality alignment (AMA) module and an uncertainty estimation (UE) module. Instead of indiscriminately correlating audio-visual cues through a global attention mechanism, AMA performs audio-visual interactions within multiple groups and consolidates group features into compact representations based on their responsiveness to audio cues, effectively directing the model's attention to audio-relevant areas. Leveraging contrastive learning, AMA further distinguishes sounding regions from silent areas by treating features with strong audio responses as positive samples and weaker responses as negatives. Additionally, UE integrates spatial and temporal information to identify high-uncertainty regions caused by frequent changes in sound state, reducing prediction errors by lowering confidence in these areas. Experimental results demonstrate that our approach achieves superior accuracy compared to existing state-of-the-art methods, particularly in challenging scenarios where traditional approaches struggle to maintain reliable segmentation.*

## 1. Introduction

Given an audio signal, Audio-Visual Segmentation (AVS) aims to identify and segment audible objects within a visual scene [49, 50]. Existing studies [9, 16, 18] focus on modeling spatio-temporal audio-visual information to capture associations between audio and visual cues, such as dynamic sound properties and object motion. Although these methods demonstrate strong performance, they often overlook challenges arising from ambiguous audio-visual spatio-temporal correspondences in real-world scenarios, potentially resulting in suboptimal segmentation, *i.e.*, over- or under-segmentation.

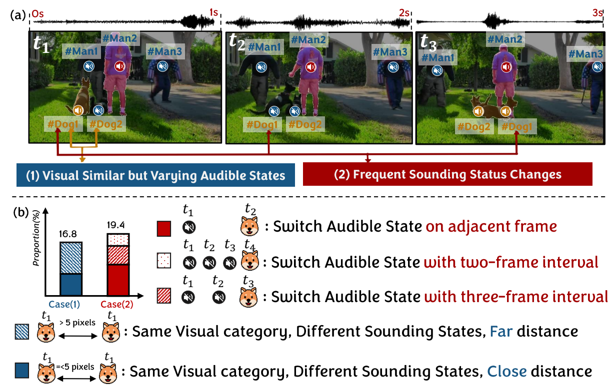


Figure 1. **(a) Illustration of ambiguous spatio-temporal correspondences.** ***Case (1):*** At time $t_1$, **#Dog1** and **#Dog2** are positioned closely but have differing sounding states, challenging the model to identify the genuine sounding one. ***Case (2):*** Over frames $t_1$, $t_2$, and $t_3$, **#Dog1** switches between sounding and silent states, posing challenges for models to capture the object's sounding status variations over time reliably. **(b) Distribution of Special Cases in AVSS Dataset.** We conduct a sample analysis utilizing a random 33.3% subset of AVSS-V2 [50], which reveals substantial occurrences of cases (1) and (2), indicating the frequent presence of challenging frames.

As illustrated in Fig. 1, we analyze and categorize the challenging cases in the AVS dataset, identifying two primary difficulties that arise from the audio-visual ambiguous correspondence. ❶ ***Incorrect Audio-Visual Associations Due to Visually Similar Objects with Different Sound States*** (**case 1**). In such cases, typical attention-based audio-

*Corresponding author.

visual alignment methods [16, 49, 50] rely on global context to weight information. However, for visually similar yet acoustically distinct objects, these methods struggle to adjust attention distribution adaptively based on their actual sounding states, causing both silent and audible objects to be mistakenly aligned with audio cues. To address this, some methods [5, 24] employ ground-truth masks to filter out silent regions or employ ground-truth masks as references to construct positive and negative samples for contrastive learning. Although these approaches can temporarily exclude interference from silent areas, the model cannot rely solely on the GT mask to learn each object's unique sound response, resulting in attention still being mistakenly allocated to silent areas. ❷ ***Learning Instability Due to Frequent Sounding Status Changes*** (**case 2**). As shown in Fig. 1, the sounding status of objects may frequently shift over short periods, complicating the learning of stable temporal audio-visual relationships. When sound states change quickly, existing methods [5, 9, 16, 24, 50] tend to smooth over variations based on the preceding continuous sounding state, leading over-segmentation results. In this paper, we propose a novel framework consisting of an audio-guided modality alignment (AMA) module aimed at addressing ❶ and an uncertainty estimation (UE) module targeting ❷.

Considering that employing global attention-based methods may lead to dispersed attention, AMA first groups visual features based on their semantic density, thereby restricting audio-visual interaction to a group level. Then, the sound-guided semantic merging module performs audio-guided weighted merging of group features. Each feature is assigned a weight based on its audio responsiveness, with features corresponding to sounding objects being amplified and those corresponding to silent objects being suppressed. This process introduces an audio-driven feature competition that effectively differentiates sounding and silent objects, generating compact representations. These representations are then reprojected onto the visual feature map. After performing the above process across multiple layers, the sounding regions are highlighted while the silent regions gradually weaken. To further strengthen the model's ability to differentiate between silent and sounding features, we treat compact representations with high audio responsiveness as positive samples and those with low responsiveness as negative samples. Contrastive learning is then employed to maximize the feature space distance between these positive and negative samples, ensuring better separation and discriminability [4, 10, 15, 33].

Furthermore, to enhance the model's reliability in handling frequent sound state changes, we introduce an uncertainty estimation (UE) module. When an object's sound state shifts rapidly, the model may experience high uncertainty in mask predictions on transition frames. To address this, we first apply temporal modeling through an attention-based layer to capture temporal dependencies across frames. The resulting features are then fed into two separate heads in UE: a mask prediction head and an uncertainty estimation head. By incorporating uncertainty estimation into the mask prediction, the model adaptively reduces prediction confidence in regions with high uncertainty, effectively mitigating the impact of unreliable predictions during transitions.

In summary, our contributions are three-fold:

- We introduce an audio-guided modality alignment module that adaptively identifies and highlights sounding regions through audio-driven feature competition and semantic grouping, improving the model's robustness against incorrect audio-visual associations.
- We develop an uncertainty estimation module to mitigate prediction errors caused by frequent changes in the sounding states of objects.
- Extensive experiments on benchmark datasets show that our framework outperforms previous state-of-the-art methods across all AVS datasets, with a 4.2% gain in $\mathcal{J}\&\mathcal{F}_\beta$ on the AVS-Semantic dataset and an 11.5% gain on the challenging VPO-MSMI dataset.

## 2. Related Work

### 2.1. Audio-Visual Alignment.

Effective audio-visual alignment is crucial for achieving precise sound source segmentation. Many studies have shown that attention-based modules [9, 16, 18, 24, 50] and contrastive learning are powerful methods for modal alignment. For instance, TPAVI [49, 50] utilizes an attention mechanism to align audio and visual modalities in both spatial and temporal dimensions. To minimize interference from background noise and irrelevant regions, stepstone [24] introduces ground-truth masks to filter out the silent visual regions, enhancing audio-visual alignment. However, pre-filtering these silent regions prevents the model from learning to differentiate between silent and active objects, limiting its ability to fully explore multimodal integration.

Contrastive learning is another popular approach for audio-visual alignment, particularly in audio-visual localization [25, 31, 32, 38]. Inspired by this, recent works [5, 6] adapt contrastive learning to AVS to enhance audio-visual correlation. Traditional AVL methods [2, 12, 14, 28, 30–32, 34, 35, 39, 42] treat audio-visual pairs from the same video as positive samples and pairs from different videos as negative. However, this approach neglects the potential semantic alignment between features from different videos, mistakenly treating them as negative pairs and disregarding cross-video consistency. To address this, Chen et al. [5] employ the ground truths to select positive samples, linking them with the corresponding audio signals, and treating background regions or mismatched audio-visual pairs as negative samples. While this method improves sample precision, it still overlooks semantic mismatches caused by visually similar but acoustically different objects in a frame.

In this work, we employ audio as the guidance to construct positive and negative samples, facilitating the model to be distinguishable for sounding and silent regions.

### 2.2. Audio Visual Segmentation.

Audio-visual segmentation aims to correctly partition the regions and label them with specific sound types [49, 50]. Existing methods can be categorized by input granularity into two main types: frame-level input [5, 6, 20, 21] (*i.e.*, a single frame paired with one-second of audio), and segment-level input [9, 11, 13, 16, 18, 23, 27, 37, 41, 44, 45, 49, 50] (*i.e.*, multiple frames paired with an audio segment of matching length). For example, BAVS [21] utilizes frame-level input to mitigate the model's tendency to over-segment silent objects. In contrast, methods emphasizing temporal information, such as CATR [16], claim that incorporating temporal dynamics better captures sound source variations over time, employing temporal audio-visual interactions through cross-attention mechanisms. Moreover, with the emergence of query-based video segmentation models [7, 46], several approaches have developed AVS frameworks built upon this architecture, incorporating audio information into the queries to facilitate audio-visual alignment [18, 24, 37], thereby effectively modeling the temporal dynamics of the fused features.

While these approaches have improved AVS performance, there are still limitations. Frame-level methods fail to fully utilize multimodal information across time, limiting segmentation accuracy. Meanwhile, temporal modeling may favor generating smooth prediction results and overlook abrupt sound changes, resulting in suboptimal segmentation performance at sound transition points. In this work, we address these issues by estimating the prediction uncertainty caused by sound transitions and adjusting the mask prediction confidence based on the uncertainty map.

## 3. Proposed Method

Fig. 2 depicts a diagram of our proposed framework. Before elaborating on our method, we first provide a preliminary in §3.1. Then we introduce the key components of our method: ***(i) Audio-Guided Modality Alignment*** (*cf.* §3.2), which details the progress of facilitating the model in learning to distinguish between sounding and silent objects. ***(ii) Uncertainty Estimation*** (*cf.* §3.3), which measures the model's confidence changes over time regarding sounding state transitions. The training objective is detailed in §3.4.

### 3.1. Notation and Preliminaries

**Task Setup.** Formally, given a video consisting of $T$ one-second clips, where each second's data is represented as

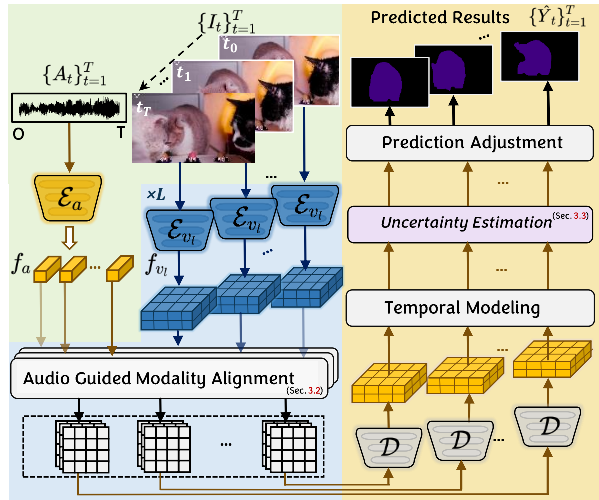


Figure 2. **Method Overview.** Our framework takes video frames $\{I_t\}_{t=1}^T$ and audio signals $\{A_t\}_{t=1}^T$ as input to segment masks $\{\hat{Y}_t\}_{t=1}^T$ for audible objects. Visual and audio features extracted by the visual block $\mathcal{E}_{v_l}$ and audio encoder $\mathcal{E}_a$ are aligned through audio-guided modality alignment. The multi-scale features from each frame are then fed into the mask decoder to generate a fused feature map. Through temporal modeling, the feature maps are processed by the uncertainty estimation module to obtain the uncertainty map and mask confidence predictions. The final predicted results are generated by integrating the uncertainty map with mask confidence predictions.

$\{I_t, A_t\}_{t=1}^T$, with $I_t$ denoting the last video frame of the $t$-th second and $A_t$ representing the corresponding one-second audio segment. A robust audio-visual segmentation model $f$ is designed to process the input pairs and generate pixel-wise localization maps $\{\hat{Y}_t\}_{t=1}^T$ for audible regions:

$$\{\hat{Y}_t\}_{t=1}^T = f(\{I_t, A_t\}_{t=1}^T). \tag{1}$$

The optimal solution $f^*$ is typically determined by minimizing a composite segmentation loss $\mathcal{L}_{SEG}$ across a dataset of $N$ videos $\{\mathcal{V}_n\}_{n=1}^N$:

$$f^* = \operatorname*{argmin}_f \frac{1}{NT} \sum_n \sum_t \mathcal{L}_{\text{seg}}(\hat{Y}_t, Y_t), \tag{2}$$

where $\mathcal{L}_{\text{seg}}$ typically includes cross-entropy loss, Dice loss [17], and IoU loss [47].

**Architecture Overview.** Similar to other established audio-visual segmentation frameworks [5, 18, 50], our architecture consists of a visual encoder with $L$ blocks, an audio encoder, and a mask decoder, defined as follows:

- *Visual Encoder* processes a single image $I_t \in \mathbb{R}^{H\times W\times 3}$ with resolution $H \times W$. Each block $\mathcal{E}_{v_l}$ extracts a feature map $f_{V_l} \in \mathbb{R}^{D_l\times H_l\times W_l}$, yielding a total of $L$ multi-scale feature maps, where $H_l$ and $W_l$ denote the spatial dimensions at level $l$, and $D_l$ is the number of channels.
- *Audio encoder* $\mathcal{E}_a$ extracts an audio representation $f_a$ from the Mel spectrogram of the audio segment $A_t \in$

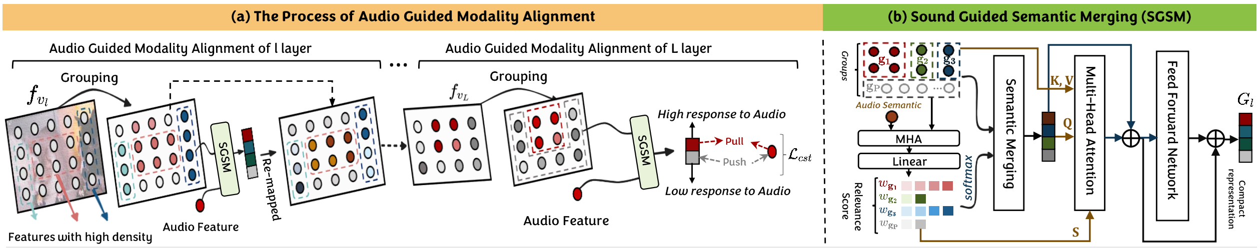


Figure 3. **(a)** Image features are first grouped based on their semantic similarity. Audio and visual features interact at the group level, where features within each group are merged into compact representations guided by the audio signal. Through multiple layers of interaction, the sounding regions are progressively highlighted. The compact representations from the final layer are then used to perform contrastive learning with audio cues. **(b)** Guided by audio features, the features within each group merge into compact semantic representations. These grouped semantics are then remapped onto the feature map to perform the next level of alignment.

$\mathbb{R}^{T\times F}$, where $T$ is the temporal duration and $F$ is the number of Mel filter banks. The output feature is $f_a = \mathcal{E}_a(A_t) \in \mathbb{R}^D$, with $D$ as the feature dimension.

- *Mask decoder* $\mathcal{D}$ takes the multi-scale feature maps $\{f_l\}_{l=1}^{L}$ as input and produces a fused feature map $f_f \in \mathbb{R}^{D_f\times H_f\times W_f}$. Here, $H_f$ and $W_f$ denote the spatial dimension, and $D_f$ is the number of channels.

### 3.2. Audio Guided Modality Alignment

To focus the model's attention on sound-relevant regions and prevent it from spreading to irrelevant areas, we first group visual features by semantic similarity, and then merge these groups based on their responsiveness to sound, thereby achieving compact representations Contrastive learning is subsequently applied to reinforce the distinction between sounding and silent objects by treating sound-responsive compact representations as positive samples and less responsive ones as negatives. Fig. 3 (a) illustrates the process intuitively.

**Visual Feature Grouping.** We employ a k-nearest neighbor variant of Density Peaks Clustering (DPC-KNN) as [8, 48] for visual feature grouping due to its ability to effectively capture complex data distributions through local density. Visual features are divided into $P$ groups, $\mathbf{G} = \{\mathbf{g_1}; \mathbf{g_2}; \ldots; \mathbf{g_P}\}$, where $\mathbf{g_p} \in \mathbb{R}^{N_p\times D}$. Here, $N_p$ denotes the token amount in group $p$, which varies across groups. Principally, DPC-KNN works as follows:

① For each visual token $f_{V_l}^i$ (where $i = 1, 2, \ldots, H_l W_l$), DPC-KNN begins by calculating the local density $\rho_i$ based on its $k$ nearest neighbors $\mathcal{N}_k(i)$:

$$\rho_i = \sum_{j\in\mathcal{N}_k(i)} exp(-||f_{v_l}^i - f_{v_l}^j||). \tag{3}$$

② Each token $f_{V_l}^i$ is then assigned to a cluster by identifying its nearest neighbor $h_i$ with a higher density:

$$h_i = \underset{j\in\mathcal{H}(i)}{\arg\min} \|f_{v_l}^i - f_{v_l}^j\|, c_i = \begin{cases} i, & \text{if } \rho_i = \max(\rho) \\ c_{h_i}, & \text{otherwise} \end{cases}, \tag{4}$$

where $\mathcal{H}(i) = \{j|\rho_i > \rho_j\}$ is the set of tokens with higher density than token $i$, and $c_i$ is the label assigned to token $i$.

With these assigned labels, we group the visual features, defining each group as $\mathbf{g_p} = \{f_{v_l}^i | c_i = p\}$, $\forall p \in \{1, 2, \ldots, P\}$.

**Sound Guided Semantic Merging.** To prevent attention dispersion caused by global attention, we restrict the audio-visual attention calculation to local regions. We further merge visual features based on their responsiveness to audio signals, creating a compact representation that emphasizes features with high responsiveness to audio signals and more clearly reflects their association with audio cues, as depicted in Fig. 3 (b).

To be specific, we first obtain the audio-visual interaction feature map $\hat{f_{v_l}}$ through a Multi-Head Cross-Attention (MCA) layer:

$$\hat{f_{v_l}} = \text{MCA}(f_{v_l}, f_{a_l}), \tag{5}$$

where $f_{a_l} \in \mathbb{R}^{D_l}$ is the audio feature projected through a linear layer to ensure dimensional consistency with $f_{v_l}$. Next, we estimate the relevance score $S \in \mathbb{R}^{H_l W_l}$ of $\hat{f_{v_l}}$ via a multi-perception layer and calculate the weight vector $w_{\mathbf{g_p}} \in \mathbb{R}^{N_p}$ for each group by applying an exponential function. This process is formulated as:

$$w_x = \frac{\exp(s_x)}{\sum_{j\in\mathbf{g_p}} \exp(s_j)}, g_p = \sum_{x\in\mathbf{g_p}} w_x \cdot x, \tag{6}$$

where $\mathbf{g_p}$ is the set of visual tokens assigned to the group $p$, and $w_x$ is a scalar weight for $x \in \mathbb{R}^{D_l}$ in group $p$. The compact representation of the input audio-visual pair is then defined as $G_l \in \mathbb{R}^{P\times D_l}$, where $P$ is the group number. This operation enables the model to focus on the most relevant tokens within each group, enhancing the distinction of regions with high audio response.

To highlight the sounding regions in the visual feature map, we enhance the compact representations using a query-based scheme and then project the updated representations back onto the visual feature map based on the token positions within each group. Specifically, the compact representations are updated by a stack of transformer decoders to converge high-level semantics:

$$G_l \leftarrow G_l + \texttt{softmax}(G_l f_{v_l}^T / \sqrt{D_l} + S) f_{v_l}. \tag{7}$$

Here, the relevance score $S$ is added to the attention weights to ensure that visual tokens with high responsiveness to sound contribute more significantly to the output. The feature map obtained by remapping the compact representations is then passed into the next block. Specifically, we first retrieve the corresponding group feature from $G_l$ for each visual token $f_{v_l}^i$ based on its group index, and then add the retrieved feature to the visual token to enhance the representation of sounding regions.

**Contrastive-based Audio-visual Alignment.** While the audio-guided feature merging yields compact visual representations that are more responsive to audio signals, it does not fully separate silent objects from sounding ones in the feature space. To address this, we employ InfoNCE contrastive loss [33] to align audio and visual features for sounding objects while repelling those of silent objects.

Let $\hat{G} \in \mathbb{R}^{P \times D}$ represents the normalized compact features obtained in the final block, and $\hat{f}_a \in \mathbb{R}^D$ the normalized audio features. We compute similarity scores between audio and visual features via a dot product, followed by a sigmoid to derive alignment scores. These scores are then utilized to differentiate positive and negative samples based on their response to audio signals, as defined below:

$$A = f_a \hat{G}^\top, I = \text{sigmoid}(A \cdot \sigma_p), \tag{8}$$

where $A$ represents the similarity scores, and $\sigma_p$ is a scaling parameter that controls the sharpness of alignment. The contrastive loss is then formulated as follows:

$$\mathcal{L}_{\text{cst}} = -\log \frac{\sum_{i \in \mathcal{P}} \exp(A_i/\tau)}{\sum_{i \in \mathcal{P}} \exp(A_i/\tau) + \sum_{j \in \mathcal{N}} \exp(A_j/\tau)}. \tag{9}$$

Here, $\mathcal{P}$ is the set of indices for positive samples, defined by $I_i > \epsilon_a$, $\mathcal{N}$ denotes the set of indices for negative samples, defined by $I_j < \epsilon_a$, and $\tau$ is the temperature parameter.

### 3.3. Uncertainty Estimation.

When the sounding status of objects frequently changes in a video, the model may struggle to generate accurate and stable predictions during these transitions. To address this, we introduce uncertainty estimation to assess changes in the model's prediction confidence across all video frames. The final predictions are then refined by weighting uncertainty estimates alongside mask prediction probabilities.

**Generation of Uncertainty Map and Mask Probability Map.** After passing the multi-scale feature maps from all audio-visual pairs in a video through the mask decoder $\mathcal{D}$, we obtain fused feature maps. These fused maps are subsequently processed by a Multi-Head Attention layer to capture semantic dependencies across all time steps. The resulting feature maps, denoted as $F_f = \{f_f\}_{t=1}^T \in \mathbb{R}^{T \times C \times H \times W}$, are input to the segmentation head, which generates the class logits $m \in \mathbb{R}^{T \times C \times H \times W}$ for each pixel.

To quantify the model's prediction uncertainty, we adopt the Dirichlet distribution, as it provides granular, robust confidence information, enabling accurate uncertainty estimation and class competition [29, 36]. Specifically, $F_f$ is processed by the uncertainty head to obtain the uncertainty logits $\alpha \in \mathbb{R}^{T \times C \times H \times W}$. We then apply a softplus activation to the logits to ensure non-negativity. These values serve as the parameters for the Dirichlet distribution, which allows us to compute pixel-wise uncertainty as follows:

$$\delta = \frac{\alpha \left( \sum_{c=1}^{C} \alpha^c - \alpha \right)}{\left( \sum_{c=1}^{C} \alpha^c \right)^2 \left( \sum_{c=1}^{C} \alpha^c + 1 \right)}, \tag{10}$$

where $\delta \in \mathbb{R}^{T \times C \times H \times W}$ represents the prediction uncertainty across spatial locations. Higher values of $\delta$ indicate regions with greater uncertainty in predictions.

**Uncertainty-Weighted Prediction Adjustment.** To ensure compatibility between uncertainty values and predicted probability distributions, we first normalize the spatial uncertainty by linearly scaling it to the [0, 1] range. The final predictions are then obtained via the following equation:

$$\{\hat{Y}\}_{t=1}^T = \sigma(m)/(\delta + \epsilon), \tag{11}$$

where $\sigma$ represents the softmax operation for semantic segmentation (or the sigmoid operation for binary segmentation), $m$ denotes the class logits from the segmentation head, and $\epsilon$ is set to $1 \times 10^{-6}$ to ensure numerical stability.

### 3.4. Loss function

The final learning objective combines the typical segmentation loss $\mathcal{L}_{\text{seg}}$ (*cf.* Eq. 2) and the contrastive loss (*cf.* Eq. 9):

$$\mathcal{L} = \lambda_{\text{seg}} \mathcal{L}_{\text{seg}} + \lambda_{\text{cst}} \mathcal{L}_{\text{cst}}, \tag{12}$$

where $\lambda_{cst}$=0.1 and $\lambda_{seg}$=1 are the weight parameters.

## 4. Experiments

### 4.1. Experimental Setup

**Datasets and Metrics.** We conduct experiments on: *AVS-Object* [49], *AVS-Semantic* [50], and *VPO* [5].

- **AVS-Objects** comprises 5,356 audio-visual pairs, each with five frames extracted from a video and a corresponding 5-second audio segment. Among these pairs, 4,932 contain a single audio event, while 424 involve multiple audio events. Binary masks are provided as ground truth to identify regions as either sounding or silent.
- **AVS-Semantic** extends the AVS-Object dataset, consisting of 12,356 audio-visual pairs. This extended dataset includes ten frames paired with a 10-second audio segment. It provides semantic annotations that identify both the sounding regions and the specific audio type.

Table 1. Quantitative results on AVS-Objects and AVS-Semantic (resized to 224 × 224). The best results are highlighted in **bold**, while the second-best results are <u>underlined</u>. Results are obtained using our complete design but with different backbones. See analysis in §4.2.

| 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 [49] [ECCV22] | PVT-V2-B5 | VGGish | 83.3 | 78.7 | 87.9 | 59.3 | 54.0 | 64.5 | 32.5 | 29.8 | 35.2 |
| AVSC [20] [ACM-MM23] | Swin-Base | VGGish | 85.0 | 81.3 | 88.6 | 62.6 | 59.5 | 65.8 | – | – | – |
| CATR [16] [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 [26] [Arxiv23] | PVT-V2-B5 | VGGish | 85.7 | 81.4 | 90.0 | 64.6 | 58.2 | 70.9 | – | – | – |
| ECMVAE [27] [ICCV23] | PVT-V2-B5 | VGGish | 85.9 | 81.7 | 90.1 | 64.3 | 57.8 | 70.8 | – | – | – |
| AuTR [22] [Arxiv23] | PVT-V2-B5 | VGGish | 82.1 | 77.6 | 86.5 | 72.0 | 66.2 | 77.7 | – | – | – |
| AQFormer [13] [IJCAI23] | PVT-V2-B5 | VGGish | 85.5 | 81.6 | 89.4 | 67.5 | 62.2 | 72.7 | – | – | – |
| AVSegFormer [9] [AAAI24] | PVT-V2-B5 | VGGish | 86.8 | 83.1 | 90.5 | 67.2 | 61.3 | 73.0 | 40.1 | 37.3 | 42.8 |
| AVSBG [11] [AAAI24] | PVT-V2-B5 | VGGish | 86.1 | 81.7 | 90.4 | 61.0 | 55.1 | 66.8 | – | – | – |
| GAVS [41] [AAAI24] | ViT-Base | VGGish | 85.1 | 80.1 | 90.0 | 70.6 | 63.7 | 77.4 | – | – | – |
| BAVS [21] [TMM24] | Swin-Base | Beats | 86.2 | 82.7 | 89.8 | 62.8 | 59.6 | 65.9 | 35.6 | 33.6 | 37.5 |
| COMBO [45] [CVPR24] | PVT-V2-B5 | VGGish | 88.3 | 84.7 | 91.9 | 65.2 | 59.2 | 71.2 | 44.1 | 42.1 | 46.1 |
| QDFormer [18] [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 | <u>90.5</u> | <u>87.3</u> | <u>93.6</u> | 72.7 | <u>67.3</u> | 78.1 | 55.3 | 48.6 | 62.0 |
| AVSStone [24] [ECCV24] | Swin-Base | VGGish | 87.3 | 83.2 | 91.3 | 72.5 | <u>67.3</u> | 77.6 | <u>61.5</u> | <u>56.8</u> | <u>66.2</u> |
| BiasAVS [37] [ACM-MM24] | Swin-Base | VGGish | 88.2 | 83.3 | 93.0 | <u>74.0</u> | 67.2 | <u>80.8</u> | 47.2 | 44.4 | 49.9 |
| RAVS (OURS) | PVT-V2-B5 | VGGish | 91.3 | 90.3 | 92.2 | 74.7 | 69.7 | 79.6 | 63.9 | 58.7 | 69.0 |
| RAVS (OURS) | MIT-B5 | VGGish | 92.2 | 91.6 | 92.8 | 74.2 | 68.3 | 80.1 | **65.4** | **59.8** | **71.0** |
| **RAVS (OURS)** | MIT-B5 | HTSAT | **93.5** | **93.1** | **93.8** | **76.4** | **70.6** | **82.1** | **65.7** | **60.8** | **70.6** |

- **VPO** is a synthetic dataset that combines single-frame images from COCO [19] with 3-second audio clips from VGGSound [1]. It includes three subsets: VPO-SS (12,202 single-source samples), VPO-MS (9,817 multi-source samples), and VPO-MSMI (12,855 samples), where weighted single-source data represent spatial information for multi-source sounds.

Following [5, 6, 18], we evaluate the performance of all methods using the mean Intersection over Union $\mathcal{J}$ and F-score ($\mathcal{F}_\beta$), where $\beta$ is set as 0.3. Additionally, due to space constraints, we provide comparisons using the metrics from TPAVI [49, 50] in the *supplementary material*.

**Network Configuration.** Our framework is end-to-end trainable, with all components parameterized by neural networks. The spatial dimensions of the three feature maps used for sound-guided semantic merging are 1/4, 1/8, and 1/16 of the input image size. The number of groups for each feature map is set to 14, 7, and 5. The threshold $\epsilon_a$ for selecting positive and negative samples is set to 0.5.

## 4.2. Quantitative Performance

**Performance on AVS-Objects and AVS-Semantic.** Table 1 presents a comparative analysis of our approach against 16 recent audio-visual segmentation (AVS) methods evaluated on AVS-Objects and AVS-Semantic. In addition to the commonly adopted PVT-V2-B5 [40] and VGGish [1], we leverage advanced visual and audio models, MIT-B5 [43] and HTSAT [3], to further enhance segmentation performance. Our approach achieves impressive results, with $\mathcal{J}\&\mathcal{F}_\beta$ scores of **93.0%** on S4 and **76.4%** on MS3 for AVS-Objects, demonstrating precise localization of audible objects. In particular, on the AVS-Semantic dataset, our method achieves a substantial **4.2%** improvement in $\mathcal{J}\&\mathcal{F}_\beta$ over the previous state-of-the-art, AVSStone. Notably, even with the standard backbones PVT-V2-B5 and VGGish, our model sets new benchmarks across all metrics.

**Performance on VPO.** In Table 2, we report the performance comparison on VPO series datasets. Our approach greatly outperforms the second-best one AVSStone. Specifically, for the VPO-SS dataset, we achieve $\mathcal{J}\&\mathcal{F}_\beta$ scores of **74.97%** compared to 68.54%, for VPO-MS: **73.48%** *vs.* 64.26%, and for VPO-MSMI: **69.30%** *vs.* 58.76%. Notably, VPO-MSMI includes many cases where objects from the same visual category appear in a frame but have different auditory states [5]. This substantial improvement over AVSStone—$\mathcal{J}$ on VPO-MSMI: **61.89%** *vs.* 50.11% and $\mathcal{F}$: **76.70%** *vs.* 67.40%—demonstrates that our method effectively directs attention to the genuine sounding regions.

## 4.3. Qualitative Performance

Fig. 4 presents qualitative comparisons between our method and three competing approaches: TPAVI [50], CAVP [5], and AVSStone [24]. In scenario (a), two men are positioned closely but are in different sounding states. AVSStone, CAVP, and TPAVI all tend to over-segment the silent man's region while under-segmenting the audible man's, reflecting their difficulty in distinguishing between visually similar objects with differing sound states. In contrast, our method accurately localizes the speaking man, effectively avoiding these mis-segmentations. In the more complex scenario (b), the sound status of the bird switches between the second and third frames. TPAVI and AVSStone tend to assume the bird is sounding in all frames, leading to segmentation of the bird throughout, while CAVP assumes silence and barely segments the bird region. By contrast, our method robustly handles these rapid sound transitions, mitigating these challenges effectively.


Figure 4. Visual comparison of challenging cases (illustrated ❶ and ❷ in §1) in AVSBench-Semantic. Refer to §4.3 for detailed analysis.

Table 2. Quantitative comparisons on 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. The best results are highlighted in **bold**, and the second-best results are in underlined. Results are obtained using our complete design but with different audio-visual backbones.

| 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 [49] [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 [9] [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 [37] [ECCV24] | SwinBase | VGGish | 107.12M | 67.46 | 59.14 | 75.78 | 63.42 | 55.61 | 71.23 | 57.94 | 49.6 | 66.27 |
| AVSStone [24] [ACM-MM24] | SwinBase | VGGish | 114.63M | <u>68.54</u> | <u>59.72</u> | <u>77.35</u> | <u>64.26</u> | <u>56.23</u> | <u>72.29</u> | <u>58.76</u> | <u>50.11</u> | <u>67.40</u> |
| CPM* [6] [ECCV24] | ResNet50 | VGGish | - | 73.49 | 67.09 | 79.88 | 72.91 | 65.91 | 79.9 | 68.07 | 60.55 | 75.58 |
| RAVS (OURS) | MIT-B5 | VGGish | 103.84M | 74.27 | 67.51 | 81.02 | 72.92 | 66.33 | 79.51 | 68.69 | 61.32 | 76.05 |
| **RAVS (OURS)** | MIT-B5 | HTSAT | 103.84M | **74.97** | **68.03** | **81.90** | **73.49** | **66.97** | **80.01** | **69.30** | **61.89** | **76.70** |

### 4.4. Further Analysis

**Ablation Study.** We carry out ablation studies on AVS-Object-MS3 and AVS-Semantic to evaluate the effectiveness of individual components and design choices in our framework. Our baseline model includes audio and visual encoders, cross-attention layers for fusion, and a mask decoder. Based on this baseline, we set up two configurations to examine the effectiveness of SGSM. The first configuration, shown in Row 2 (Baseline+GTF), utilizes ground truth masks to filter silent regions during modality alignment. The second configuration, shown in Row 3 (Baseline+SGSM), incorporates our proposed sound-guided semantic merger (SGSM). As results indicate, SGSM improves the baseline by 7.4% in $\mathcal{J}\&\mathcal{F}_\beta$ on AVS-Semantic and achieves 3% better performance than Baseline+GTF. This demonstrates that SGSM effectively directs the model's attention to audible regions, mitigating interference from silent areas.

Building on SGSM, we introduce contrastive modality alignment (CST), shown in Row 4. Compared to SGSM alone, this alignment design improves $\mathcal{J}\&\mathcal{F}_\beta$ by 1.6% on AVS-Semantic, indicating that aligning compact features with audio representations enhances audio-visual alignment. Additionally, we integrate our uncertainty estimation module (UE) on top of SGSM+CMA, yielding further performance gains. On AVS-Semantic, $\mathcal{J}$ and $\mathcal{F}_\beta$ increase by 1.7% and 1.3%, respectively, confirming that UE effectively mitigates errors caused by frequent changes in an object's audible state, resulting in more accurate segmentation.

**Effectiveness of Audio Guided Alignment.** In Fig. 5, we map the compact representation identified as a positive or negative sample back onto the image based on the locations of their visual combinations and show their responsiveness to the audio cues. When two objects of the same visual category but different sound states appear in a video frame, our method effectively identifies the truly sounding object as a positive sample and the silent object as a negative sample. For example, in (a)-(c), in scenes with two women where only one is speaking, and in scenes with multiple adjacent cars where only one is making sounds, our method accurately associates the sound with the truly audible objects.

When multiple objects of the same category are sounding simultaneously, our method consistently classifies all sounding objects as positive samples, as shown in (d)-(f). Additionally, by analyzing the sound responsiveness of all samples, we observe that regions around positive samples typically show lower sound responsiveness. This indicates that the audio-guided feature merging not only focuses ef-

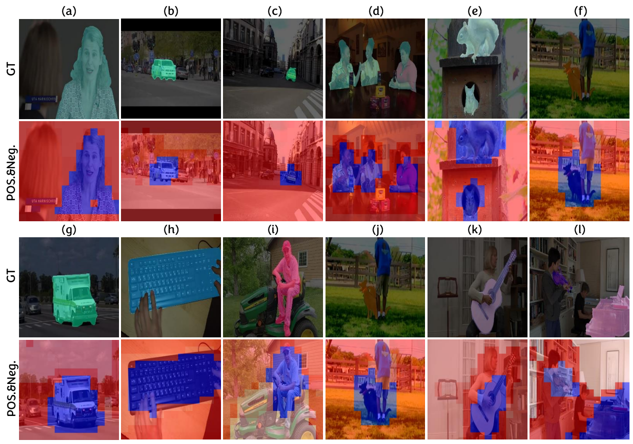


Figure 5. Visualization of positive (Pos.) and negative (Neg.) samples generated by the AMA module. Red indicates negative samples, while blue represents positive samples. Darker blue indicates regions with a higher responsiveness to audio cues, while darker red indicates lower responsiveness. See detailed analysis in §4.4.

Table 3. Evaluation of component effectiveness (§4.4). GTF: Ground-Truth mask filtering; SGSM: sound-guided semantic merging; CST: contrastive-based alignment; AMA: audio-guided modality alignment; UE: uncertainty estimation.

| Settings | *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 | 64.3 | 59.8 | 68.7 | 57.2 | 51.4 | 62.9 |
| +GTF | 67.9 | 62.2 | 73.6 | 59.6 | 55.1 | 64.1 |
| +SGSM | 70.9 | 65.3 | 76.4 | 62.6 | 57.7 | 67.5 |
| +SGSM+CST (AMA) | 72.2 | 66.1 | 78.3 | 64.2 | 59.1 | 69.3 |
| +AMA+UE (Ours) | **76.4** | **70.6** | **82.1** | **65.7** | **60.8** | **70.6** |

fectively on relevant sounding regions but also assists in identifying hard samples around the sounding regions.

**Effectiveness of Uncertainty Estimation.** Fig. 6 presents the visualizations of uncertainty maps generated by our uncertainty estimation (UE) module. In (a), where the sound state of all objects remains constant, the model shows consistently low uncertainty values across the entire image, indicating high confidence in its segmentation predictions. In contrast, in (b), where a `'leopard'` undergoes a transition in the sound state between adjacent frames, the model exhibits higher uncertainty for both sound-emitting and silent frames, with even higher uncertainty values in the silent frames. This higher uncertainty effectively lowers the prediction probability in the `'leopard'` region in the second frame, reducing over-segmentation in silent frames. These findings demonstrate that frequent sound state transitions significantly impact the model's confidence in its predictions, suggesting that uncertainty estimation is an effective way to mitigate mis-segmentation issues.

**Effects of Group Number.** We conduct a study on the impact of different group numbers when conducting audio-guided alignment. The grouping setting of [14, 7, 5] (approximately 1/4 of the feature map size) achieves the best performance, attaining the highest $\mathcal{J}\&\mathcal{F}_\beta$ scores on both datasets. In contrast, smaller grouping numbers like [5, 5, 5] led to decreased performance. This suggests that progressively reducing the group number supports more effective semantic merging of sound-emitting regions under audio guidance. Additionally, a larger group number, such as [28, 14, 5], divides the feature map into many small groups, dispersing attention across numerous small regions and thus leading to poorer results.

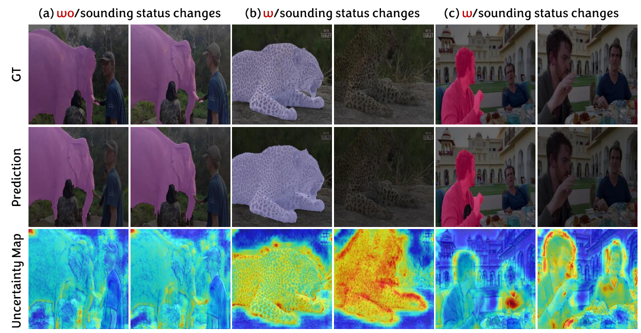


Figure 6. Visualizations of uncertainty maps, with colors ranging from red (high uncertainty) to blue (low uncertainty).

Table 4. Comparison of different group numbers for three-scale feature maps. See analysis in §4.4.

| Groups | *AVS-Objects-MS3* | | | *AVS-Semantic* | | |
|---|---|---|---|---|---|---|
| | $\mathcal{J}\&\mathcal{F}_\beta \uparrow$ | $\mathcal{J} \uparrow$ | $\mathcal{F}_\beta \uparrow$ | $\mathcal{J}\&\mathcal{F}_\beta \uparrow$ | $\mathcal{J} \uparrow$ | $\mathcal{F}_\beta \uparrow$ |
| [28, 14, 5] | 73.7 | 68.6 | 78.8 | 63.4 | 57.8 | 68.9 |
| [14, 7, 5] | **76.4** | **70.6** | **82.1** | **65.7** | **60.8** | **70.6** |
| [10, 7, 5] | 75.1 | 69.3 | 80.8 | 64.4 | 58.8 | 70.0 |
| [5,5,5] | 71.5 | 66.7 | 76.3 | 62.6 | 56.5 | 68.7 |

## 5. Conclusion

In this paper, we propose a novel framework to address modality misalignment caused by ambiguous spatio-temporal correspondences. Unlike previous methods that rely on global attention or employ masks to exclude silent regions, our approach first utilizes audio to guide the merging of sounding regions, followed by contrastive learning to align compact visual representations with audio cues. This design enhances the model's ability to distinguish between audible and silent objects, even when visually similar but acoustically different objects are in close proximity. Additionally, we introduce uncertainty estimation to mitigate predictive instability arising from frequent changes in an object's sound state. The two items enable our framework to prevent the over-segmentation of silent regions in complex scenarios. Extensive experiments demonstrate the effectiveness of our approach in handling challenging cases.

**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] 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

[2] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16867–16876, 2021. 2

[3] 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. 6

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2

[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. 2, 3, 5, 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, 6, 7

[7] Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G. Schwing. Mask2former for video instance segmentation, 2021. 3

[8] Mingjing Du, Shifei Ding, and Hongjie Jia. Study on density peaks clustering based on k-nearest neighbors and principal component analysis. *Knowledge-Based Systems*, 99: 135–145, 2016. 4

[9] 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. 1, 2, 3, 6, 7

[10] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 2

[11] 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

[12] Di Hu, Rui Qian, Minyue Jiang, Xiao Tan, Shilei Wen, Errui Ding, Weiyao Lin, and Dejing Dou. Discriminative sounding objects localization via self-supervised audiovisual matching. *Advances in Neural Information Processing Systems*, 33:10077–10087, 2020. 2

[13] 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. 3, 6

[14] EA King, A Tatoglu, D Iglesias, and A Matriss. Audio-visual based non-line-of-sight sound source localization: A feasibility study. *Applied Acoustics*, 171:107674, 2021. 2

[15] Phuc H Le-Khac, Graham Healy, and Alan F Smeaton. Contrastive representation learning: A framework and review. *Ieee Access*, 8:193907–193934, 2020. 2

[16] 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. 1, 2, 3, 6

[17] 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. 3

[18] 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. 1, 2, 3, 6

[19] 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

[20] 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

[21] 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

[22] 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. 6

[23] 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

[24] 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. 2, 3, 6, 7

[25] Tanvir Mahmud, Yapeng Tian, and Diana Marculescu. T-vsl: Text-guided visual sound source localization in mixtures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26742–26751, 2024. 2

[26] 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. 6

[27] 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

[28] Yoshiki Masuyama, Yoshiaki Bando, Kohei Yatabe, Yoko Sasaki, Masaki Onishi, and Yasuhiro Oikawa. Self-supervised neural audio-visual sound source localization via probabilistic spatial modeling. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4848–4854. IEEE, 2020. 2

[29] Thomas Minka. Estimating a dirichlet distribution, 2000. 5

[30] Shentong Mo and Pedro Morgado. A closer look at weakly-supervised audio-visual source localization. *Advances in Neural Information Processing Systems*, 35:37524–37536, 2022. 2

[31] Shentong Mo and Pedro Morgado. Localizing visual sounds the easy way. In *European Conference on Computer Vision*, pages 218–234. Springer, 2022. 2

[32] Shentong Mo and Yapeng Tian. Audio-visual grouping network for sound localization from mixtures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10565–10574, 2023. 2

[33] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2, 5

[34] Rui Qian, Di Hu, Heinrich Dinkel, Mengyue Wu, Ning Xu, and Weiyao Lin. Multiple sound sources localization from coarse to fine. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 292–308. Springer, 2020. 2

[35] Kranthi Kumar Rachavarapu, Vignesh Sundaresha, AN Rajagopalan, et al. Localize to binauralize: Audio spatialization from visual sound source localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1930–1939, 2021. 2

[36] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018. 5

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

[38] Weixuan Sun, Jiayi Zhang, Jianyuan Wang, Zheyuan Liu, Yiran Zhong, Tianpeng Feng, Yandong Guo, Yanhao Zhang, and Nick Barnes. Learning audio-visual source localization via false negative aware contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6420–6429, 2023. 2

[39] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 247–263, 2018. 2

[40] 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. 6

[41] 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. 3, 6

[42] Yan Xia and Zhou Zhao. Cross-modal background suppression for audio-visual event localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19989–19998, 2022. 2

[43] 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

[44] 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

[45] 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

[46] Kaining Ying, Qing Zhong, Weian Mao, Zhenhua Wang, Hao Chen, Lin Yuanbo Wu, Yifan Liu, Chengxiang Fan, Yunzhi Zhuge, and Chunhua Shen. Ctvis: Consistent training for online video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 899–908, 2023. 3

[47] 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. 3

[48] Wang Zeng, Sheng Jin, Wentao Liu, Chen Qian, Ping Luo, Wanli Ouyang, and Xiaogang Wang. Not all tokens are equal: Human-centric visual analysis via token clustering transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11101–11111, 2022. 4

[49] 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, 2, 3, 5, 6, 7

[50] 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, 5, 6