DDAVS: Disentangled Audio Semantics and Delayed Bidirectional Alignment for Audio-Visual Segmentation

Jingqi Tian[1] , Yiheng Du[2] , Haoji Zhang[1] , Yuji Wang[1] , Isaac Ning Lee[1] , Xulong Bai[1] , Tianrui Zhu[1] , Jingxuan Niu[1] , Yansong Tang[1†]

[1]Tsinghua Shenzhen International Graduate School, Tsinghua University [2]Peking University

{tjq25@mails.,tang.yansong@sz.}tsinghua.edu.cn

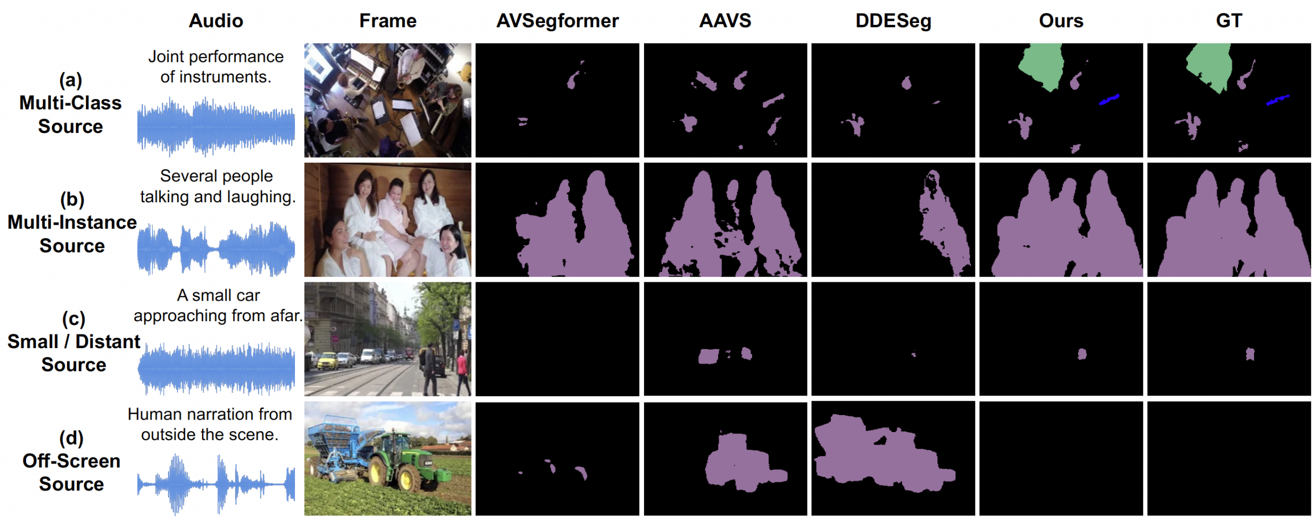


Figure 1. **Qualitative comparison of our DDAVS model and previous methods.** DDAVS consistently outperforms previous approaches in challenging scenarios involving multiple classes, multiple sources, small or distant sound sources, and off-screen audio cues.

Abstract

Audio–Visual Segmentation (AVS) aims to localize sound-producing objects at the pixel level by jointly leveraging auditory and visual information. However, existing methods often suffer from multi-source entanglement and audio–visual misalignment, which lead to biases toward louder or larger objects while overlooking weaker, smaller, or co-occurring sources. To address these challenges, we propose DDAVS, a Disentangled Audio Semantics and Delayed Bidirectional Alignment framework. To mitigate multi-source entanglement, DDAVS employs learnable queries to extract audio semantics and anchor them within a structured semantic space derived from an audio prototype memory bank. This is further optimized through contrastive learning to enhance discriminability and robustness. To alleviate audio–visual misalignment, DDAVS introduces dual cross-attention with delayed modality interaction, improving the robustness of multimodal alignment. Extensive experiments on the AVS-Objects and VPO benchmarks demonstrate that DDAVS consistently outperforms existing approaches, exhibiting strong performance across single-source, multi-source, and multi-instance scenarios. These results validate the effectiveness and generalization ability of our framework under challenging real-world audio–visual segmentation conditions.

1. Introduction

Traditional Visual Segmentation (VS) focuses solely on appearance, partitioning all visible objects in an image regardless of their physical state or behavior. [15, 17] In contrast, Audio–Visual Segmentation (AVS) [25, 26] introduces an additional auditory modality, aiming to identify and segment *sound-emitting* objects that are temporally and semantically linked to the accompanying audio signal. By enforcing pixel-level alignment between auditory cues and visual

†Correspondence to Yansong Tang.

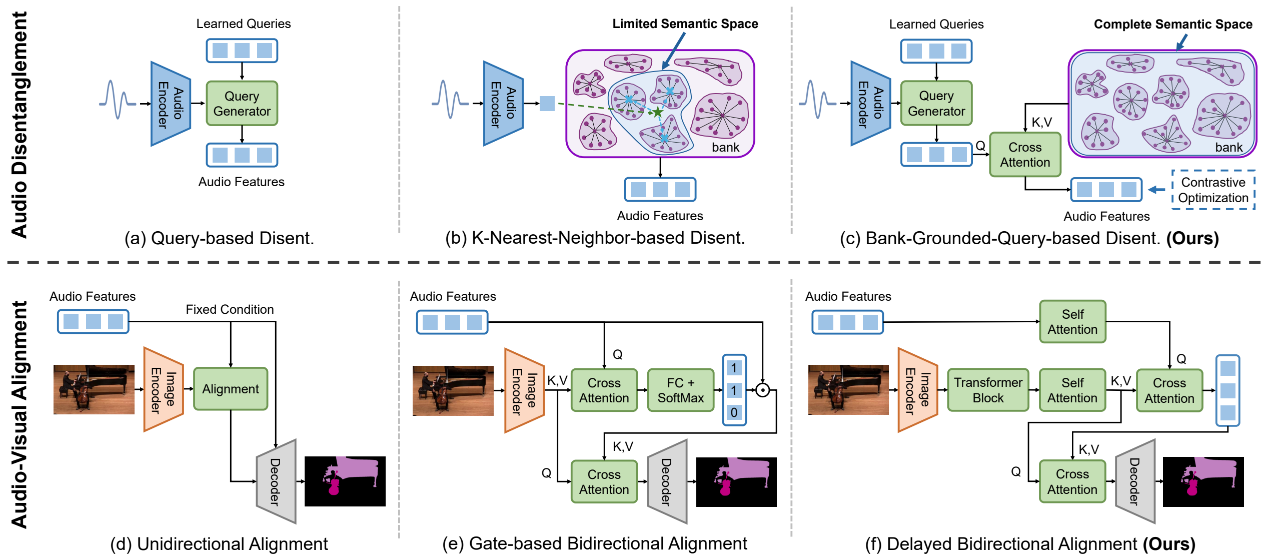


Figure 2. For **Audio Disentanglement**, prior methods (a) use learned queries for semantics [5, 21, 40] or (b) derive disentangled features from K-nearest classes [25]. In contrast, our method (c) uses an audio prototype memory bank to ground audio queries, coupled with contrastive optimization to enhance their discriminability and robustness. For **Audio-Visual Alignment**, existing methods either (d) treat audio features as a fixed condition [40, 46, 47], or (e) apply gating mechanisms to scale audio features for dual cross-attention [25]. Our method, however, (f) performs dual cross-attention with delayed modality interaction to improve multimodal alignment robustness.

evidence, AVS moves toward a more holistic understanding of visual and acoustic multi-modal scenes [4, 36, 39, 45].

The introduction of audio modality in AVS presents unique challenges compared with VS. Fig. 1 illustrates such cases, where the audio input originates from (a) multi-class sources, (b) multi-instance sources, (c) small or distant sources, or (d) off-screen sources. The challenges posed by multi-class and multi-instance sources stem from the **multi-source entanglement problem**, where the audio input entangles multiple sources (e.g., a guitar and a piano), hindering the precise segmentation of individual sound-producing objects. While the challenges posed by small or distant sources and off-screen sources stem from the **audio-visual misalignment problem**, where the audio and vision modality cannot be accurately associated with each other.

To address the multi-source entanglement problem, existing methods typically resort to an audio disentanglement module to disentangle the audio input into multiple semantics by using learnable queries [5, 21, 40] to generate audio semantics (see Fig. 2(a)). However, the resulting semantics often lie in a self-organized latent space that is suboptimal for representing audio. The recent audio bank-based method [25] approximates audio semantics by selecting the K nearest centers from a multi-class audio feature bank (see Fig. 2(b)). However, when the audio input includes a distinct source, the semantics of the weaker source are often lost, as the entangled representations are derived from the limited semantic space of the K nearest classes. This also reduces the distinguishability of the output results. To address the audio-visual misalignment problem, existing methods [40, 46, 47] typically perform unidirectional audio-conditioned visual alignment (see Fig. 2(d)). However, this unidirectional design prevents vision from enhancing visually associated audio components and suppressing non-associated ones, such as off-screen sounds. Although the current SOTA method [25] introduces bidirectional alignment to suppress off-screen sounds, it relies on a gating mechanism that merely scales audio intensity, without aligning to visual semantics or capturing visual spatial information (see Fig. 2(e)).

In this paper, we present *DDAVS*, a two-stage framework that first disentangles audio semantics and then performs delayed bidirectional alignment with visual features. Specifically, in the audio disentanglement stage, we propose to use learnable queries to extract multiple audio semantics and perform cross-attention conditioned on a pre-built multi-class prototype memory bank comprising single-source audio embeddings (Fig. 2(c)). This anchors the extracted semantics to the bank's structured and stable semantic space, infusing the bank's knowledge and facilitating subsequent alignment. During the training, we additionally introduce contrastive learning, where the anchor is a single audio semantic derived from the clean audio input, the positive is the corresponding audio semantic derived from the waveform augmented audio input, and the negative is other audio semantics derived from the clean and augmented audio inputs. The motivation is to enhance the distinguishability of disentangled audio semantics and improve robustness to audio inputs. During the alignment stage, we perform delayed bidirectional cross-attention to align audio and visual modalities (Fig. 2(f)). The delayed interaction filters low-level noise, while the bidirectional design enables symmetric cross-attention between audio and video, capturing mutual dependencies for precise segmentation.

In summary, our technical contributions are as follows:

- We propose an AVS framework with disentangled audio semantics and a delayed bidirectional alignment for precise segmentation in challenging scenarios such as multi-source, subtle, distant, or off-screen sounds.
- We propose an audio disentanglement method that anchors query-extracted audio semantics to a prototype memory bank for global consistency, and uses contrastive learning to enhance discriminability and robustness.
- We propose an audio-visual alignment method using cascaded bidirectional cross-attention to enhance inter-modal interaction and delayed alignment for precise high-level correspondence while reducing low-level noise.
- Experiments on the AVS-Objects and VPO benchmarks demonstrate that DDAVS consistently outperforms previous methods especially in challenging scenarios.

2. Related Work

Audio-Visual Segmentation. Given an audio signal and an accompanying image or video, audio-visual segmentation aims to produce the segmentation mask of the sounding objects in the image [7, 8, 18, 25, 26, 40, 43, 46, 47]. As a pioneer, Zhou *et al*. [46] propose the audio-visual segmentation problem and introduce the AVSBench benchmark. Typical AVS methods usually leverage learnable queries [5, 20–22, 26, 28, 29, 37, 40] to extract audio or visual semantics and perform an audio-visual alignment to achieve visual segmentation based on audio cues. Recent AVS methods focus on text-bridged strategy [27], counterfactual learning [44], audio enhancement and disentanglement [25], and robust audio-visual alignment [13, 24, 31]. In addition to architectural advances, several frameworks use contrastive learning [2, 9, 11] to enhance cross-modal alignment and training stability. CAVP [3], DiffusionAVS [31] and CQ-Former [28] adopt an InfoNCE-based loss [33] to align audio and visual modalities. WS-AVS [32] applies contrastive learning under weak supervision. Our method differs from existing approaches by using a bank to anchor and enrich query-generated audio semantics and a delayed bidirectional alignment to guide segmentation. Moreover, we leverage contrastive learning to enhance the discriminability and robustness of audio features rather than to align audio and visual modalities.

Multi-Source Audio Disentanglement. In multi-source scenarios, AVS methods usually employ representation-level audio disentanglement mechanisms to separate overlapping sound sources [5, 21, 25, 29, 40]. This is often achieved through learnable queries [5, 21, 29, 40] or K-nearest-neighbor-based decomposition [25], where the audio feature is decomposed into multiple audio semantics representing distinct sound emitters. However, existing query-based methods produce semantic tokens in a self-organized space without explicit structure. While the K-nearest-neighbor-based method might be limited in discriminability. We embed audio semantics into an audio-preferred semantic space using a prototype memory bank and enhance their discriminability via contrastive learning.

Audio-Visual Alignment. This module establishes spatial and semantic correspondences between the audio and visual modalities before decoding [25, 34, 35, 40, 42, 46–48]. Typical AVS methods conduct unidirectional audio-conditioned visual alignment [40, 46, 47], ignoring the utilization of visual features to improve audio features. Recent methods [25, 34, 38, 42] introduce bidirectional alignment to improve inter-modal interaction. However, the gating mechanism [25] and the early alignment [34, 38] might hinder effective alignment. Although AVESFormer [40] performs delayed alignment, it performs an unidirectional alignment. In contrast, we propose a delayed bidirectional alignment to achieve dynamic cross-modal feature matching, effectively improving segmentation accuracy.

3. Method

We propose an end-to-end Disentangled Audio Semantics and Delayed Bidirectional Alignment framework (DDAVS). As shown in Fig. 3, the framework comprises three key components: (1) Audio Query Module (AQM) converts audio features into a compact set of disentangled semantic queries anchored to a prototype bank; (2) Contrastive Optimization Module (COM) refines these queries via contrastive learning; and (3) Audio-Visual Alignment Module (AVAM) employs multi-stage dual cross-attention to align both modalities progressively and bidirectionally.

Formally, given a raw audio waveform A_a and its corresponding video clip of frames I_v, the audio feature $E_a = \mathcal{E}_a(A_a)$ and visual feature $E_v = \mathcal{E}_v(I_v)$ are extracted by their encoders $\mathcal{E}_a$ and $\mathcal{E}_v$. H_v^i denotes the visual feature after the i-th decoder stage. The inference pipeline is:

$$Q = \text{AQM}(E_a) \tag{1}$$

$$H_v^0 = E_v \tag{2}$$

$$H_v^i = \text{AVAM}^i(Q, H_v^{i-1}), \quad i = 1, \ldots, L \tag{3}$$

The AQM transform E_a into a disentangled representation Q, while E_v serves as the initial visual input H_v^0. The CQM is only used during training to provide an additional contrastive loss. DDAVS then performs L iterative stages of alignment and fusion through the Audio-Visual Alignment Module (AVAM), each applying dual cross-attention followed by Transformer refinement to synchronize and inte-

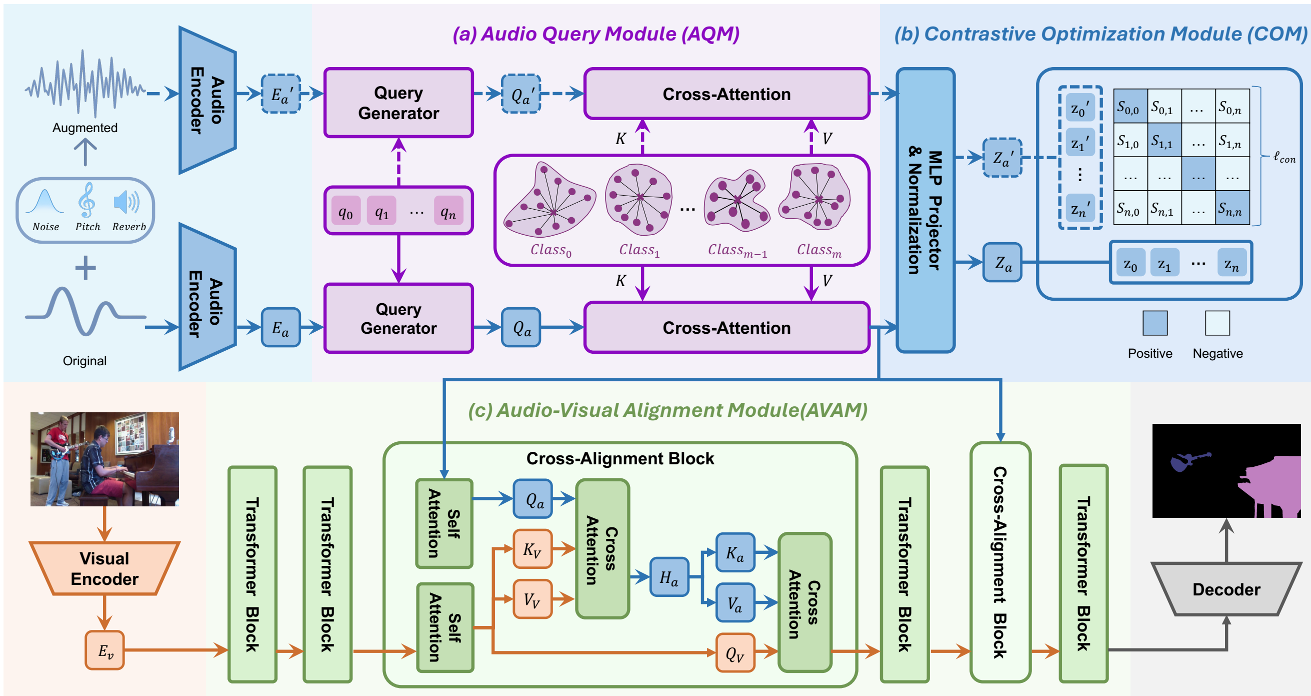


Figure 3. **Overview of the DDAVS framework.** (a) The Audio Query Module (AQM) encodes original and augmented waveforms into disentangled semantic queries anchored to a prototype memory bank. (b) The Contrastive Optimization Module (COM) enhances query robustness through contrastive learning, used only during training. (c) The Audio-Visual Alignment Module (AVAM) fuses audio queries with visual features via stacked alignment blocks, and a lightweight decoder outputs the sound-conditioned segmentation masks.

grate the two modalities. This progressive alignment yields increasingly discriminative and spatially coherent representations. Finally, a lightweight decoder $\mathcal{D}$ generates the pixel-level segmentation mask $\hat{Y} = \mathcal{D}(H_v^L)$ that highlights audible regions within the scene.

3.1. Audio Query Module

The Audio Query Module (AQM) transforms the encoded audio features $E_a \in \mathbb{R}^{L\times d}$ into compact and disentangled representations by learned queries. It aims to decouple overlapping sound sources and map them into a stable semantic space anchored by a global prototype memory bank.

Query Generation. As shown in Fig. 3, the *Query Generator* is implemented with a Q-Former [19], which maps the sequential audio tokens E_a into n audio query vectors: $Q_a = f_{\text{QG}}(E_a; \{q_i\}_{i=1}^n) \in \mathbb{R}^{n\times d}$. Each learnable query q_i is a latent slot that focuses on a distinct sound component, allowing AQM to separate co-occurring acoustic patterns.

Bank-based Refinement. The initial audio query Q_a is refined by cross-attention with a pre-constructed prototype memory bank $\mathcal{M}$, where each class $C_i = \{c_{i,j} \in \mathbb{R}^d\}_{j=1}^{K_i}$ stores K_i cluster feature centroids extracted from single-source audios of the i-th class. All prototypes are concatenated to form $\mathcal{M} = \text{concat}(\{c_{i,j}\}) \in \mathbb{R}^{P\times d}$ ($P = \sum_i K_i$). Q_a interacts with the memory bank via cross-attention:

$$A = \text{Softmax}\left(\frac{(Q_a W_Q)(\mathcal{M} W_K)^\top}{\sqrt{d}}\right), \qquad (4)$$
$$\widetilde{Q} = A(\mathcal{M} W_V), \qquad Q = \text{LN}(Q_a + \gamma \widetilde{Q})$$

Here $W_{Q/K/V} \in \mathbb{R}^{d\times d}$ are projection layers, γ is a hyper-parameter, and LN denotes layer normalization. Q denotes the bank-grounded audio queries. This process aligns each query with its most relevant prototype anchor that encapsulates class-wise prior knowledge.

Although AQM generates semantically aligned queries, their embeddings are insufficiently discriminative and tend to be dominated by the most salient audio source. When multiple sounds types (e.g., speech and engine) co-occur, the components of dominant sounds often suppress weaker signals, resulting in poor inter-class separation.

3.2. Contrastive Optimization Module

To mitigate this problem, we design a Contrastive Optimization Module (COM), which employs contrastive learning to increase the semantic separation between different sound classes and improve robustness to acoustic variations.

Audio Signal Augmentation. To improve robustness under acoustic perturbations, we randomly apply waveform-level

Table 1. **Quantitative comparisons on the AVSBench dataset**, including single-source (AVS-Objects-S4), multi-source (AVS-Objects-MS3), and semantic-source (AVS-Semantic) settings. Best results in **Bold**, while second best underlined.

Method	AVS-Objects-S4			AVS-Objects-MS3			AVS-Semantic		
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
TPAVI [46] [ECCV22]	83.3	78.7	87.9	59.3	54.0	64.5	32.5	29.8	35.2
CATR [20] [ACM-MM23]	87.9	84.4	91.3	68.6	62.7	74.5	35.7	32.8	38.5
AuTR [26] [Arxiv23]	82.1	77.6	86.5	72.0	66.2	77.7	–	–	–
AVSC [22] [ACM-MM23]	85.0	81.3	88.6	62.6	59.5	65.8	–	–	–
ECMVAE [30] [ICCV23]	85.9	81.7	90.1	64.3	57.8	70.8	–	–	–
AQFormer [12] [IJCAI23]	85.5	81.6	89.4	67.5	62.2	72.7	–	–	–
AVSegFormer [5] [AAAI24]	86.8	83.1	90.5	67.2	61.3	73.0	40.1	37.3	42.8
BAVS [23] [TMM24]	86.2	82.7	89.8	62.8	59.6	65.9	35.6	33.6	37.5
GAVS [38] [AAAI24]	85.1	80.1	90.0	70.6	63.7	77.4	–	–	–
AVSBG [10] [AAAI24]	86.1	81.7	90.4	61.0	55.1	66.8	–	–	–
AVESFormer [40] [Arxiv24]	84.5	79.9	89.1	63.3	57.9	68.7	34.0	31.2	36.8
QDFormer [21] [CVPR24]	83.9	79.5	88.2	64.0	61.9	66.1	–	–	–
CAVP [3] [CVPR24]	83.8	78.8	88.9	61.5	55.8	67.1	32.8	30.4	35.3
AAVS [29] [ECCV24]	87.3	83.2	91.3	72.5	67.3	77.6	50.9	48.5	53.2
BiasAVS [37] [ACM-MM24]	88.2	83.3	93.0	74.0	67.2	**80.8**	47.2	44.4	49.9
DiffusionAVS [31] [TIP25]	85.7	81.4	90.0	64.6	58.2	70.9	–	–	–
VCT [13] [CVPR25]	88.5	84.7	92.3	73.4	67.5	79.3	50.4	47.9	52.9
DDESeg [25] [CVPR25]	91.1	89.1	93.1	72.2	68.1	76.2	49.6	47.1	52.1
TAViS [27] [ICCV25]	88.0	84.8	91.2	72.1	68.2	75.9	–	44.2	–
ICF [44] [ICCV25]	90.1	86.6	93.5	69.9	64.4	75.4	48.2	45.0	51.3
DDAVS (Ours)	**92.4**	**90.6**	**94.2**	**75.1**	**70.6**	79.5	**52.6** (↑1.7)	**49.7** (↑1.2)	**55.5** (↑2.3)

perturbations ϵ (additive noise, reverb, pitch shift) to the raw waveform A_a to obtain $A'_a = g(A_a)$, here g is the augmentation operator following WavAugment [14, 16]. The augmented audio is encoded as $E'_a = \mathcal{E}_a(A'_a, \epsilon)$ and passed through AQM with shared parameters, producing the corresponding query set Q'. Both Q and Q' provide disentangled and semantically grounded audio representations.

Contrastive Learning. Given the refined query set $Q = \{q_i\}_{i=1}^n$ and its augmented counterpart $Q' = \{q'_i\}_{i=1}^n$, we apply a projector $\phi(\cdot)$ and ℓ_2-normalization to each query:

$$z_i = \frac{\phi(q_i)}{\|\phi(q_i)\|_2}, \qquad z'_i = \frac{\phi(q'_i)}{\|\phi(q'_i)\|_2}, \tag{5}$$

Let $s_{i,j} = z_i^\top z'_j$. The contrastive loss is:

$$\mathcal{L}_{\text{con}} = -\frac{1}{n}\sum_{i=1}^{n} \log \frac{\exp(s_{i,i}/\tau)}{\sum_{j=1}^{n}\exp(s_{i,j}/\tau)}, \tag{6}$$

Here τ is the temperature coefficient. $\mathcal{L}_{\text{con}}$ pulls together positives (z_i, z'_i) and pushes apart negatives $\{(z_i, z'_j)\}_{j \neq i}$, enlarging inter-query margins under acoustic variations. After contrastive optimization, we obtain the enhanced audio embedding $Q = \{q_i\}_{i=1}^n$, which is discriminative and noise-resilient, making downstream modality alignment and segmentation more robust and stable.

3.3. Audio-Visual Alignment Module

The Audio-Visual Alignment Module is designed to align visual and auditory modalities for precise localization of sound-producing regions. As shown in Fig. 3, AVAM alternates between Cross Alignment Block and Transformer Block, which progressively enhance spatial coherence and cross-modal interactions. Given the visual tokens E_v and enhanced audio embedding Q, each Cross Alignment Block performs a double cross-attention $\mathrm{Attn}(Q, K, V)$ pipeline.

Audio-Guided Filtering. Audio queries attend to visual tokens to extract sound-relevant visual evidence:

$$H_a^i = \mathrm{Attn}(Q, H_v^{i-1}W_K^1, H_v^{i-1}W_V^1), \tag{7}$$

This allows the audio tokens to focus on visually correlated regions that are consistent with the emitted sound.

Visual-Guided Enhancement. The updated audio representations then act as keys and values to inject discriminative acoustic cues back into the visual stream:

$$H_v^i = \mathrm{Attn}(H_v W_Q^2, H_a^i W_K^2, H_a^i W_V^2), \tag{8}$$

This step enhances the discriminative capacity of the visual features, emphasizing sound-producing regions.

Our pipeline achieves superior modality alignment by interleaving Cross Alignment Blocks and Transformer Blocks, with delayed cross-modal fusion applied exclusively between the third and fourth layers of the four-block architecture. The motivation is that audio sources are inherently associated with instance-level visual features, while the initial Transformer layers focus on pixel-level information. Delaying the cross-modal alignment to the later layers allows the model to leverage richer, more global representa-

Table 2. **Quantitative comparisons on the VPO dataset**, including single-source (VPO-SS), multi-source (VPO-MS), and multi-source multi-instance (VPO-MSMI) settings. Best results in **Bold**, while second best <u>underlined</u>.

Method	VPO-SS			VPO-MS			VPO-MSMI		
	$\mathcal{J}\&\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\&\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\&\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$
TPAVI [46] [ECCV22]	44.63	41.64	47.62	45.68	42.30	49.06	43.19	40.03	46.34
AVSegFormer [5] [AAAI24]	45.94	43.81	48.06	43.72	47.30	40.14	49.93	47.19	52.67
CAVP [3] [CVPR24]	67.02	58.81	75.23	61.32	53.24	69.39	56.48	48.18	64.78
BiasAVS [37] [ECCV24]	67.46	59.14	75.78	63.42	55.61	71.23	57.94	49.60	66.27
AAVS [29] [ACM-MM24]	68.54	59.72	77.35	64.26	56.23	72.29	58.76	50.11	67.40
RAVS [24] [CVPR25]	**74.97**	**68.03**	**81.90**	73.49	66.97	80.01	<u>69.30</u>	61.89	<u>76.70</u>
DDESeg [25] [CVPR25]	74.38	67.55	81.20	<u>74.30</u>	<u>67.64</u>	<u>80.96</u>	68.39	<u>62.11</u>	74.67
DDAVS	<u>74.80</u>	<u>67.82</u>	<u>81.77</u>	**76.11**	**69.61**	**82.60**	**72.84** (↑3.54)	**65.96** (↑3.85)	**79.72** (↑3.02)

tions, thereby facilitating more accurate and robust modality alignment. The final audio-conditioned visual representation H_v^L is passed to the segmentation decoder $\mathcal{D}$ to generate the predicted segmentation mask $\hat{Y} = \mathcal{D}(H_v^L)$.

3.4. Optimization

We train the DDAVS model with a unified objective:

$$\begin{aligned}\mathcal{L}_{\text{total}} = \lambda_{\text{ce}}\mathcal{L}_{\text{ce}} + \lambda_{\text{dice}}\mathcal{L}_{\text{dice}} \\ + \lambda_{\text{iou}}\mathcal{L}_{\text{iou}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}}.\end{aligned} \tag{9}$$

The cross-entropy loss $\mathcal{L}_{\text{ce}}$ provides pixel-wise supervision, while the Dice and IoU losses $\mathcal{L}_{\text{dice}}$ and $\mathcal{L}_{\text{iou}}$ encourage region completeness and accurate boundary alignment. Beyond these segmentation losses, the contrastive term $\mathcal{L}_{\text{con}}$ (Eq. (6)) enforces discriminative audio queries by enlarging inter-query margins under acoustic perturbations. The segmentation losses (Cross-entropy, Focal, Dice, IoU) and λ coefficients are detailed in the supplementary material.

4. Experiments

4.1. Experimental Setup

Implementation Details. The visual backbone is initialized from MiT-B5 [41], and the audio encoder adopts HTSAT [1] pretrained on AudioSet [6]. Our prototype memory bank is constructed following DDESeg [25] from single-sounding source signals. All experiments are conducted on a workstation equipped with eight NVIDIA RTX 4090 GPUs (24 GB each). Training uses the AdamW optimizer with an initial learning rate of 1×10^{-4} and a batch size of 64. We also fix random seeds to ensure reproducibility.

Datasets and Metrics. We evaluate DDAVS on two audio-visual segmentation benchmarks: AVSBench [46, 47] and VPO [3], which cover single-source, multi-source, and semantic conditions. Following common practice [3, 46] in AVS, we adopt the Jaccard index ($\mathcal{J}$), the F-score ($\mathcal{F}$) and their average $\mathcal{J}\&\mathcal{F}$ as evaluation metrics. The F-score is $\mathcal{F} = \frac{(1+\beta^2)\cdot\text{Precision}\cdot\text{Recall}}{\beta^2\cdot\text{Precision}+\text{Recall}}$, where $\beta^2 = 0.3$, which places more emphasis on recall. For AVSBench (including AVS-Object and AVS-Semantic), the scores are computed using the official TPAVI evaluation protocol [46], while for VPO we follow the metric implementation of CAVP [3].

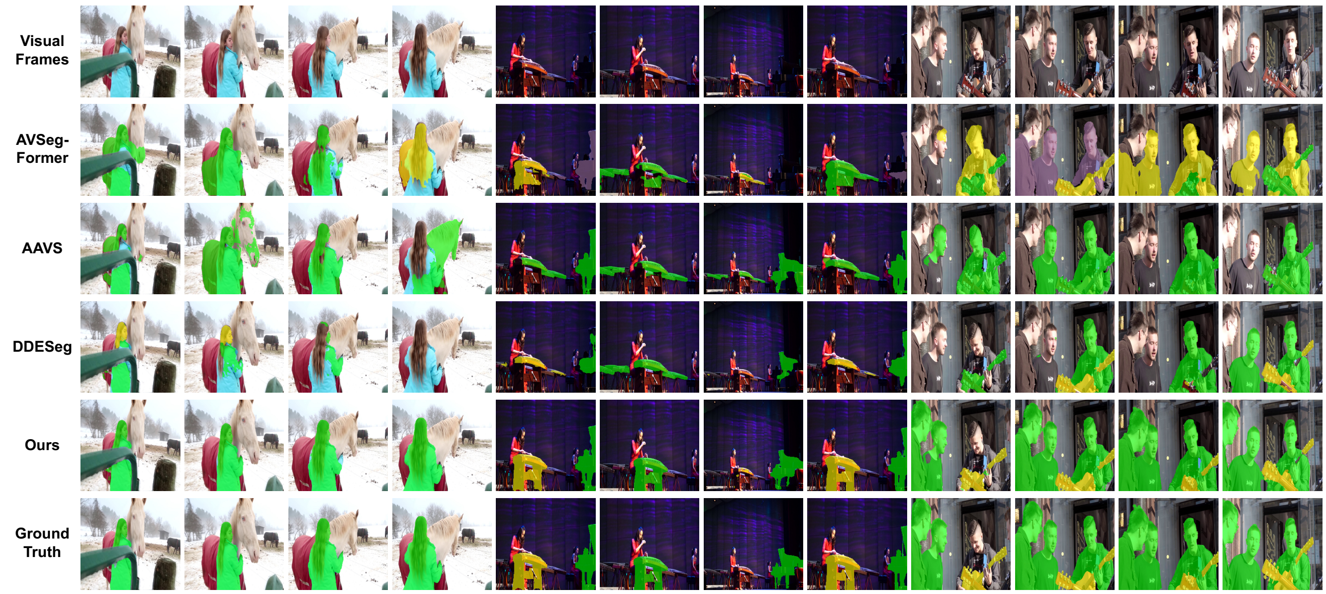
4.2. Comparison with State-of-the-art Methods

AVSBench. Tab. 1 presents the experimental results on AVSBench. DDAVS outperforms the previous best results by 1.30% on AVS-S4, 1.10% on AVS-MS3, and 1.70% on AVSS in terms of $\mathcal{J}\&\mathcal{F}$. In the multi-source setting (MS3), DDAVS exceeds the flagship method AVSegFormer [5] by 7.90% and the recent DDESeg [25] by 2.90%. On the semantic subset AVSS involving spatial and categorical ambiguity, DDAVS improves over previous best baseline by 1.20% and 2.30% for $\mathcal{J}$ and $\mathcal{F}$ respectively, indicating that disentangled audio queries and dual-stage fusion effectively reduce interference between overlapping sources.

VPO. As shown in Tab. 2, DDAVS outperforms the state-of-the-art by 1.81% and 3.54% $\mathcal{J}\&\mathcal{F}$ on VPO-MS (multi-source) and VPO-MSMI (multi-source and multi-instance). These results demonstrate the advantage of DDAVS in complex multi-source scenes that require robust cross-modal representation, while DDAVS also remains competitive under simple acoustic conditions such as VPO-SS.

4.3. Qualitative Evaluation

Fig. 4 presents qualitative comparisons on AVS-Semantic. (a) DDAVS correctly isolates the speaking human and suppresses the silent horse, while prior models often leak activation onto the horse. (b) DDAVS segments all sounding instruments without activating silent ones, whereas other methods miss sources or segment incorrect objects. (c) DDAVS identifies all sounding persons and their guitars, while competing models merge people, lose details or miss guitars. Overall, DDAVS delivers cleaner masks and handles multi-sources more reliably than previous methods.



(a) Non-sounding distractor (horse–human) (b) Multiple instruments (piano, guzheng) (c) Multi-person guitar performances

Figure 4. **Qualitative results on AVSBench-Semantic.** DDAVS produces cleaner and more source-consistent masks than previous baselines AVSegFormer, AAVS, and DDESeg, especially in complex multi-source scenes with non-sounding distractors (horse–human), multiple active instruments (piano–guzheng), or multi-person guitar performances.

Table 3. **Ablation study on the framework components on AVS-MS3 and AVSS Benchmark.**

Method	*AVS-MS3*			*AVSS*		
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Baseline	69.71	65.88	73.54	48.63	45.83	51.42
+ AQM	70.89	67.04	74.73	49.80	46.73	52.86
+ AQM + COM	73.47	69.32	77.61	51.70	48.56	54.83
+ AVAM	73.16	68.96	77.36	51.45	48.13	54.76
DDAVS (Ours)	**75.08**	**70.62**	**79.54**	**52.62**	**49.71**	**55.53**

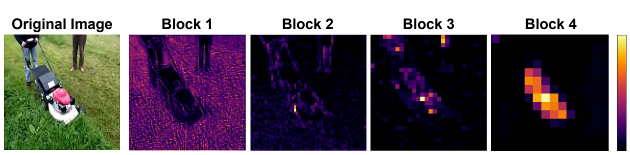


Figure 5. **Visualization of attention maps of audio-injected transformers blocks at different layers.** It is observed that injecting audio features into blocks 3 and 4 bringing clearer instance-level attention, compared to the blurry pattern at earlier blocks.

4.4. Ablation Study

Framework Components. Tab. 3 analyzes the contribution of AQM, COM, and AVAM on AVS-MS3 and AVSS. The baseline only contains encoders, transformer blocks and the segmentation deocder. Adding AQM brings moderate gains, due to the effective bank-based audio feature extraction. When COM is enabled on top of AQM, the gains become much more pronounced, showing that contrastive learning is crucial for robust and disentangled audio representation. Only adding AVAM alone also yields strong improvements, which demonstrates the effectiveness of injecting audio information before the last two transformer blocks for cross-modal fusion. Finally, combining all three modules yields the best overall performance, with $\mathcal{J}\&\mathcal{F}$ improvements of 5.37% on AVS-MS3 and 3.99% on AVSS.

Table 4. **Ablation study on the position of audio injection blocks in the AVAM.** Blocks 1–4 denote the first to fourth Transformer blocks (from input to output) in the visual backbone [41].

Injected blocks	*AVS-MS3*			*AVSS*		
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
1	68.37	64.33	72.41	47.96	44.53	51.39
2	72.02	67.76	76.27	50.03	46.92	53.13
3	73.77	69.42	78.12	51.78	49.03	54.52
4	73.32	68.85	77.79	51.13	48.11	54.15
1,2	70.69	66.25	75.13	49.53	46.54	52.52
2,3	73.29	68.82	77.15	51.05	48.13	53.97
3,4 (Ours)	**75.08**	**70.62**	**79.54**	**52.62**	**49.71**	**55.53**
1,2,3	72.07	67.82	76.32	50.29	47.24	53.33
2,3,4	74.17	69.78	78.55	51.65	48.52	54.77
1,2,3,4	72.57	68.48	76.65	51.07	48.21	53.93

Cross Alignment Blocks. To answer the question of *where to perform cross-modal alignment* within AVAM, we conducted a in-depth analysis. We observe that injecting audio

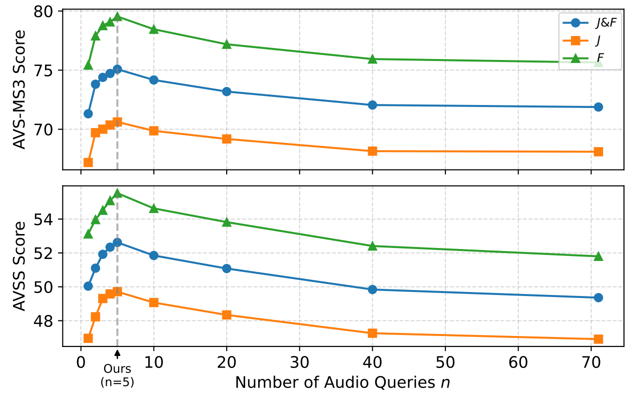


Figure 6. **Effect of the number of audio queries.** Performance on AVS-MS3 (top) and AVSS (bottom) as the number of audio queries n varies, where our choice $n = 5$ (ours) achieves the best overall results before larger n leads to performance degradation due to redundant queries.

information into transformer block at different layers result in essential different attention pattern. As shown in Fig. 5, earlier cross-modal fusion focuses on pixel-level local feature, while later fusion focuses on region- or instance-level global features, which is crucial for audio source association. We further conducted a comprehensive ablation study regarding to the position arrangement of Cross Alignment Blocks. As shown in Tab. 4, injection audio information at later layers produce high performance and the combination of injecting to block 3 and 4 results in the highest $\mathcal{J}\&\mathcal{F}$ score, which is adopted in the DDAVS model.

Number of Audio Queries. Fig. 6 illustrates the effect of varying the number of audio queries n on AVS-MS3 and AVSS. As n increases from 1 to 5, $\mathcal{J}\&\mathcal{F}$ rises rapidly on both datasets, indicating that a small set of diverse queries helps capture different sounding patterns. When n becomes larger than 5, the performance starts to decrease, suggesting that using too many queries is unnecessary in practice. Based on this empirical observation, we adopt a moderate value $n = 5$ as the default setting of DDAVS.

4.5. Representation Analysis

To further examine the learned audio semantics, we visualize t-SNE projections of the audio queries, as shown in Fig. 7. We consider four single-source categories (`guitar`, `man`, `keyboard`, `helicopter`) and a mixed "`guitar-man`" source. For the baseline model, single-source clusters are partially entangled and the mixed samples (red) collapse into the `guitar` cluster, indicating that the representation is dominated by a single source and fails to preserve mixture semantics. In contrast, DDAVS produces four compact and clearly separated clusters for single-source classes, while the mixed samples are located

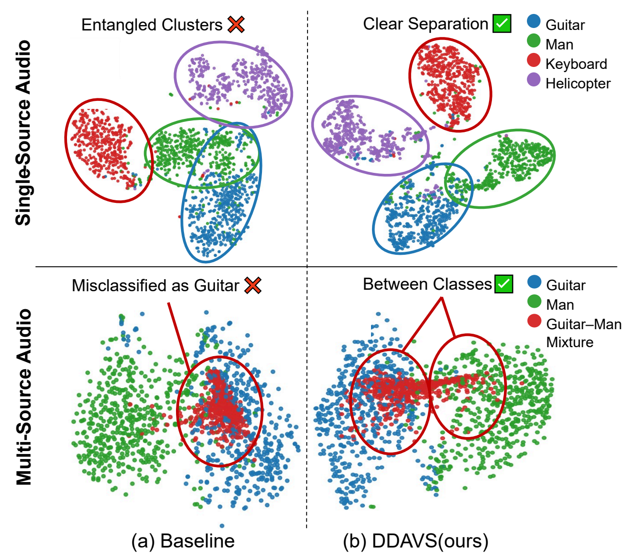


Figure 7. **t-SNE visualization of audio representations. Top**: DDAVS produces well-separated single-source clusters compared with the baseline. **Bottom**: for mixed "guitar–man" audio, baseline features collapse toward the guitar cluster, while DDAVS positions mixture samples between the two classes, better capturing multi-source semantics.

between the `guitar` and `man` clusters, suggesting that our disentangled queries explicitly encode contributions from multiple sources and alleviate multi-source entanglement.

5. Conclusion

In this work, we presented DDAVS, a disentangled audio–visual segmentation framework that explicitly addresses the challenges posed by multi-source mixtures and audio–visual misalignment. By introducing a prototype-guided Audio Query Module (AQM), a waveform-level Contrastive Optimization Module (COM), and a delayed bidirectional Audio–Visual Alignment Module (AVAM), our method improves semantic separation, preserves weak or mixed audio cues, and achieves more reliable cross-modal alignment. We further validate the effectiveness of this disentanglement–alignment paradigm through comprehensive experiments on AVSBench and VPO, where DDAVS establishes state-of-the-art performance across single-source, multi-source, and semantic-source settings. These results demonstrate the value of structured audio semantics and robust alignment strategies for advancing audio–visual segmentation.

In future work, we plan to extend the disentanglement–alignment paradigm to open-domain videos and streaming audio, enabling real-time, scalable, and more generalizable audio–visual perception.

References

[1] 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 *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 646–650, 2022. 6

[2] 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. 3

[3] 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, 5, 6

[4] Henghui Du, Guangyao Li, Chang Zhou, Chunjie Zhang, Alan Zhao, and Di Hu. Crab: A unified audio-visual scene understanding model with explicit cooperation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18804–18814, 2025. 2

[5] 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. 2, 3, 5, 6

[6] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audioset: An ontology and human-labeled dataset for audio events. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 776–780. IEEE, 2017. 6

[7] Sitong Gong, Yunzhi Zhuge, Lu Zhang, Yifan Wang, Pingping Zhang, Lijun Wang, and Huchuan Lu. Avs-mamba: Exploring temporal and multi-modal mamba for audio-visual segmentation. *IEEE Transactions on Multimedia*, 2025. 3

[8] Sitong Gong, Yunzhi Zhuge, Lu Zhang, Pingping Zhang, and Huchuan Lu. Complementary and contrastive learning for audio-visual segmentation. *IEEE Transactions on Multimedia*, 2025. 3

[9] 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. 3

[10] 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. 5

[11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 3

[12] 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. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2023. 5

[13] Shaofei Huang, Rui Ling, Tianrui Hui, Hongyu Li, Xu Zhou, Shifeng Zhang, Si Liu, Richang Hong, and Meng Wang. Revisiting audio-visual segmentation with vision-centric transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8352–8361, 2025. 3, 5

[14] Dongwei Jiang, Wubo Li, Miao Cao, Wei Zou, and Xiangang Li. Speech simclr: Combining contrastive and reconstruction objective for self-supervised speech representation learning. *arXiv preprint arXiv:2010.13991*, 2020. 5

[15] Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu, et al. Segment anything in high quality. *Advances in Neural Information Processing Systems*, 36: 29914–29934, 2023. 1

[16] Eugene Kharitonov, Morgane Rivière, Gabriel Synnaeve, Lior Wolf, Pierre-Emmanuel Mazaré, Matthijs Douze, and Emmanuel Dupoux. Data augmenting contrastive learning of speech representations in the time domain. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 215–222. IEEE, 2021. 5

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 1

[18] Jia Li and Yapeng Tian. From waveforms to pixels: A survey on audio-visual segmentation. *arXiv preprint arXiv:2508.03724*, 2025. 3

[19] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. 4

[20] 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, 5

[21] 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, 5

[22] 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, 5

[23] 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. 5

[24] Chen Liu, Peike Li, Liying Yang, Dadong Wang, Lincheng Li, and Xin Yu. Robust audio-visual segmentation via audio-guided visual convergent alignment. In *Proceedings of the*

Computer Vision and Pattern Recognition Conference, pages 28922–28931, 2025. 3, 6

[25] Chen Liu, Liying Yang, Peike Li, Dadong Wang, Lincheng Li, and Xin Yu. Dynamic derivation and elimination: Audio visual segmentation with enhanced audio semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3131–3141, 2025. 1, 2, 3, 5, 6

[26] 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, 5

[27] Ziyang Luo, Nian Liu, Xuguang Yang, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Fahad Shahbaz Khan, and Junwei Han. Tavis: Text-bridged audio-visual segmentation with foundation models. *arXiv preprint arXiv:2506.11436*, 2025. 3, 5

[28] Ying Lv, Zhi Liu, and Xiaojun Chang. Consistency-queried transformer for audio-visual segmentation. *IEEE Transactions on Image Processing*, 2025. 3

[29] 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, 5, 6

[30] 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. 5

[31] Yuxin Mao, Jing Zhang, Mochu Xiang, Yunqiu Lv, Dong Li, Yiran Zhong, and Yuchao Dai. Contrastive conditional latent diffusion for audio-visual segmentation. *IEEE Transactions on Image Processing*, 2025. 3, 5

[32] Shentong Mo and Bhiksha Raj. Weakly-supervised audio-visual segmentation. *Advances in Neural Information Processing Systems*, 36:17208–17221, 2023. 3

[33] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[34] Juhyeong Seon, Woobin Im, Sebin Lee, Jumin Lee, and Sung-Eui Yoon. Extending segment anything model into auditory and temporal dimensions for audio-visual segmentation. In *2024 IEEE International Conference on Image Processing (ICIP)*, pages 2480–2486. IEEE, 2024. 3

[35] Zhaofeng Shi, Qingbo Wu, Fanman Meng, Linfeng Xu, and Hongliang Li. Cross-modal cognitive consensus guided audio-visual segmentation. *IEEE Transactions on Multimedia*, 2024. 3

[36] Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, Yuxuan Wang, and Chao Zhang. video-salmonn: Speech-enhanced audio-visual large language models. *arXiv preprint arXiv:2406.15704*, 2024. 2

[37] Peiwen Sun, Honggang Zhang, and Di Hu. Unveiling and mitigating bias in audio visual segmentation. *arXiv preprint arXiv:2407.16638*, 2024. 3, 5, 6

[38] 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, 5

[39] Yaoting Wang, Peiwen Sun, Dongzhan Zhou, Guangyao Li, Honggang Zhang, and Di Hu. Ref-avs: Refer and segment objects in audio-visual scenes. *European Conference on Computer Vision*, 2024. 2

[40] Zili Wang, Qi Yang, Linsu Shi, Jiazhong Yu, Qinghua Liang, Fei Li, and Shiming Xiang. Avesformer: Efficient transformer design for real-time audio-visual segmentation. *arXiv preprint arXiv:2408.01708*, 2024. 2, 3, 5

[41] 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, 7

[42] 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

[43] Kaining Ying, Henghui Ding, Guangquan Jie, and Yu-Gang Jiang. Towards omnimodal expressions and reasoning in referring audio-visual segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22575–22585, 2025. 3

[44] Mingfeng Zha, Tianyu Li, Guoqing Wang, Peng Wang, Yangyang Wu, Yang Yang, and Heng Tao Shen. Implicit counterfactual learning for audio-visual segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22349–22360, 2025. 3, 5

[45] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025. 2

[46] 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. 2, 3, 5, 6

[47] 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*, 2024. 2, 3, 6

[48] Yang-Hao Zhou, Heyan Huang, Cunhan Guo, Rong-Cheng Tu, Zeyu Xiao, Bo Wang, and Xian-Ling Mao. Aloha: Adapting local spatio-temporal context to enhance the audio-visual semantic segmentation. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21 (6):1–23, 2025. 3