Delayed Bidirectional Alignment via Disentangled Audio Semantics 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], and Yansong Tang[1†]

[1] Tsinghua Shenzhen International Graduate School, Tsinghua University, China
[2] Peking University, China
{tjq25@mails, tang.yansong@sz}.tsinghua.edu.cn

Abstract. Audio–Visual Segmentation (AVS) aims to localize sound-producing objects at the pixel level by integrating auditory and visual cues. However, existing methods often struggle with multi-source entanglement and audio–visual misalignment, leading to a dominance bias toward acoustically or visually salient objects (i.e., louder or larger ones) at the expense of subtler or co-occurring sources. To address these challenges, we propose DDAVS: Delayed Bidirectional Alignment via Disentangled Audio Semantics for Audio-Visual Segmentation. 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 process 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 achieves state-of-the-art performance across single-source, multi-source, and multi-class multi-instance scenarios. These results validate the effectiveness and generalization ability of our framework under challenging real-world audio–visual segmentation conditions. Project page.

Keywords: Audio-Visual Segmentation · Multimodal Learning · Visual-Audio Representation

1 Introduction

Traditional Visual Segmentation (VS) focuses solely on appearance, partitioning all visible objects in an image regardless of their physical state or behavior [22, 24]. In contrast, Audio–Visual Segmentation (AVS) [33, 34] 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

† Corresponding author

Fig. 1: Qualitative comparison of DDAVS 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.

evidence, AVS moves toward a more holistic understanding of acoustic and visual multi-modal scenes, alongside recent advances in fine-grained grounding, temporal consistency and unified reasoning [6, 19–21, 29, 40, 48, 52, 61, 68].

Despite its potential, AVS introduces unique challenges illustrated in Fig. 1. First, multi-source entanglement in (a) multi-class and (b) multi-instance scenarios prevents precise isolation of individual sound-producers, leading to degraded segmentation performance. Second, audio-visual misalignment hinders cross-modal correspondence; specifically, (c) small or distant sources provide insufficient visual anchors, while (d) off-screen sources lack visual counterparts, often causing spurious activations or incorrect suppression.

Early approaches resort to an audio disentanglement module using learnable queries to disentangle the audio input into multiple semantics [7, 28, 54], followed by unidirectional audio-conditioned visual alignment [54, 65, 66] (see Fig. 2(a)). However, this pipeline faces two key limitations: its disentangled semantics reside in a self-organized latent space suboptimal for audio representation, and the unidirectional design prevents visual cues from enhancing scene-aligned audio components or suppressing irrelevant ones (e.g., off-screen sounds), thereby weakening cross-modal refinement. More recent studies, such as the audio bank-based framework [33], attempt to improve semantic clarity by approximating audio semantics via the K nearest centers from a multi-class feature bank (see Fig. 2(b)). However, when distinct audio sources coexist, weaker-source semantics are often lost due to the constrained semantic space of the K nearest classes, reducing output distinguishability; furthermore, bidirectional alignment relies on a gating mechanism that merely scales audio intensity without aligning to visual semantics or capturing spatial cues.

In this paper, we present DDAVS, a framework for audio-visual segmentation that leverages disentangled audio semantics to enable delayed bidirectional alignment (Fig. 2(c)). Our approach proceeds in two stages. First, in the audio disentanglement stage, we use learnable queries to extract multiple audio seman-

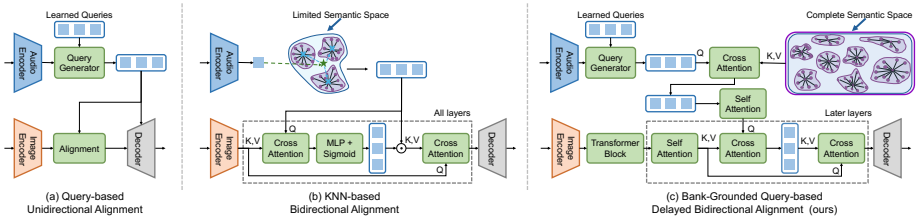

Fig. 2: Comparison of audio-visual segmentation approaches. (a) Query-based audio disentanglement followed by unidirectional audio-to-visual alignment. (b) KNN-based audio semantics approximation from a feature bank with gated bidirectional alignment. (c) Prototype memory bank-grounded disentanglement with delayed bidirectional cross-attention operating exclusively in later layers.

tics and perform cross-attention conditioned on a pre-built multi-class prototype memory bank of single-source audio embeddings. This anchors the extracted semantics to a structured and stable space, infusing prior knowledge and facilitating subsequent alignment. Additionally, we integrate contrastive learning during training to enhance the discriminability and robustness of disentangled audio semantic anchors. Second, in the alignment stage, unlike methods that apply cross-attention across all network layers (e.g., Fig. 2(b)), our delayed bidirectional cross-attention operates exclusively in later layers to align audio and visual modalities. 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 delayed bidirectional alignment via disentangled Audio Semantics for precise segmentation in challenging scenarios such as multi-source, subtle, distant, or off-screen sounds.
- We propose an audio disentanglement module 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 module 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 AVS-Objects and VPO benchmarks demonstrate that DDAVS consistently outperforms prior 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 [1, 9, 10, 25, 33–36, 54, 59, 65, 66]. As a pioneer, Zhou *et al.* [66] propose the audio-visual segmentation problem and

introduce the AVSBench benchmark. Typical AVS methods usually leverage learnable queries [7, 27, 28, 32, 34, 38, 39, 49, 54] 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 [37], counterfactual learning [60], audio enhancement and disentanglement [33], and robust audio-visual alignment [17, 30, 41]. In addition to architectural advances, several frameworks use contrastive learning [3, 11, 14] to enhance cross-modal alignment and training stability. CAVP [4], DiffusionAVS [41] and CQFormer [38] adopt an InfoNCE-based loss [44] to align audio and visual modalities. WS-AVS [43] 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 [7,28,33,39,54]. This is often achieved through learnable queries [7, 28, 39, 54] or K-nearest-neighbor-based decomposition [33], 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 [33, 46, 47, 54, 56, 65–67]. Typical AVS methods conduct unidirectional audio-conditioned visual alignment [54, 65, 66], ignoring the utilization of visual features to improve audio features. Recent methods [12, 33, 46, 51, 53, 56, 58, 62–64] introduce bidirectional alignment to improve inter-modal interaction. However, the gating mechanism [33] and the early alignment [46,51] might hinder effective alignment. Although AVESFormer [54] adopts delayed alignment, the alignment is unidirectional. In contrast, we propose delayed bidirectional alignment for 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 complementary 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

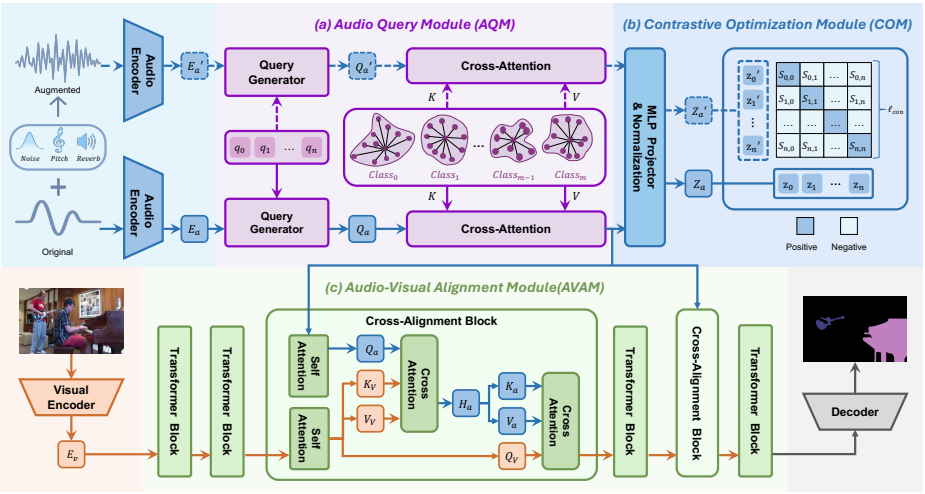

Fig. 3: Overview of the DDAVS framework. (a) Audio Query Module (AQM) encodes original and augmented waveforms into disentangled semantic queries. (b) Contrastive Optimization Module (COM) enhances query robustness through contrastive learning. (c) Audio-Visual Alignment Module (AVAM) fuses audio queries with visual features using alignment blocks.

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 COM 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 integrate 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 [26], 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 Construction. We construct a global prototype memory bank $\mathcal{M}$ to provide stable semantic anchors for query refinement. For each class $i \in \{1, \ldots, C\}$, we collect single-source audio clips where class i is the only audible sound, extract features using the HTSAT backbone, and obtain embeddings $E_i = \{x_k^i \in \mathbb{R}^d\}_{k=1}^{n_i}$. Applying K-means++ clustering with K_i clusters to E_i, we select the cluster centroids as class-specific prototypes $C_i = \{c_{i,j} \in \mathbb{R}^d\}_{j=1}^{K_i}$. Concatenating all class prototypes yields the global bank $\mathcal{M} = \text{concat}(C_1, \ldots, C_C) \in \mathbb{R}^{K_i \times d}$. Crucially, $\mathcal{M}$ remains fixed during both training and inference, ensuring consistent semantic grounding across all samples. Further implementation details are provided in the supplementary material.

Bank-Guided Refinement. The initial audio queries Q_a are refined through cross-attention with the prototype memory bank $\mathcal{M}$ described above. Specifically, Q_a interacts with $\mathcal{M}$ via:

$$A = \text{Softmax}\left(\frac{(Q_a W_Q)(\mathcal{M} W_K)^\top}{\sqrt{d}}\right) \tag{4}$$

$$\widetilde{Q} = A(\mathcal{M} W_V) \tag{5}$$

$$Q = \text{LN}(Q_a + \gamma \widetilde{Q}) \tag{6}$$

where $W_{Q/K/V} \in \mathbb{R}^{d \times d}$ are projection layers, γ is a scaling factor, and LN denotes layer normalization. This process anchors each query to its most relevant semantic prototype, injecting class-aware prior knowledge while preserving query diversity. Q denotes the final bank-grounded audio queries used for alignment.

While AQM effectively generates semantically aligned queries, their embeddings remain insufficiently discriminative and tend to be dominated by salient audio sources. In multi-sound scenarios (e.g., speech co-occurring with engine noise), dominant components often suppress weaker signals, resulting in poor inter-class separation. This limitation motivates the Contrastive Optimization Module (COM) introduced next.

3.2 Contrastive Optimization Module

To mitigate the issue of insufficiently discriminative audio embeddings, we design a Contrastive Optimization Module (COM), which employs contrastive learning to enhance semantic separation between different sound classes and improve robustness to acoustic variations.

Audio Signal Augmentation. To improve robustness under acoustic perturbations, we apply waveform-level augmentations to the raw audio signal.

The pipeline first resamples audio to 16 kHz, center-crops or pads to a fixed duration, and normalizes to $[-1, 1]$. We then generate an augmented counterpart $A'_a = g(A_a)$ using WavAugment [18, 23], where g applies a chain of time-domain effects: reverberation ($r \in [20, 40]$), pitch shift ($\Delta p \in [-150, 150]$ cents), dynamic-range compression, and volume jitter (SNR $\in [10, 20]$ dB). Parameter ranges are summarized in supplementary. Crucially, only COM processes the augmented branch: the clean waveform A_a drives the main segmentation path via $E_a = \mathcal{E}_a(A_a)$, while the augmented waveform A'_a is exclusively used by COM to produce $E'_a = \mathcal{E}_a(A'_a)$ and corresponding query set Q'. This design ensures semantic content remains intact while introducing moderate acoustic variations for robust representation learning.

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:

$$z_i = \frac{\phi(q_i)}{\|\phi(q_i)\|_2}, \qquad z'_i = \frac{\phi(q'_i)}{\|\phi(q'_i)\|_2} \tag{7}$$

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

$$\mathcal{L}_{\text{con}} = -\frac{1}{n} \sum_{i=1}^{n} \log \frac{\exp(s_{i,i}/\tau)}{\sum_{j=1}^{n} \exp(s_{i,j}/\tau)} \tag{8}$$

where τ is the temperature coefficient. $\mathcal{L}_{\text{con}}$ pulls together positive pairs (z_i, z'_i) and pushes apart negative pairs $\{(z_i, z'_j)\}_{j \neq i}$, enlarging inter-query margins under acoustic variations. After contrastive optimization, we obtain enhanced audio embeddings $Q = \{q_i\}_{i=1}^n$that are more robust to noise and better disentangled across sound classes, directly strengthening downstream cross-modal alignment and segmentation stability.

3.3 Audio-Visual Alignment Module

The Audio-Visual Alignment Module (AVAM) aligns visual and auditory modalities to precisely localize sound-producing regions. As illustrated in Fig. 3, AVAM employs an alternating architecture of Cross Alignment Blocks and Transformer Blocks, progressively refining spatial coherence and cross-modal interactions.

Within the i-th Cross Alignment Block, given hidden state H_v^{i-1} and enhanced audio queries Q, alignment begins with audio queries attending to visual tokens. This design leverages the complementary nature of the modalities: audio delivers concise semantic cues about *what* is sounding, while vision supplies the spatial context for *where* it originates. Consequently, this directional attention naturally guides the model toward sound-relevant regions and suppresses

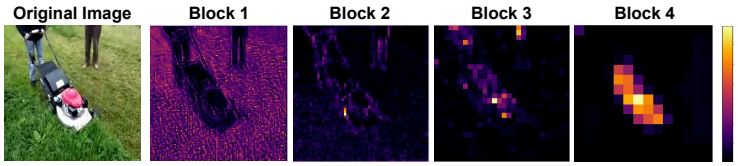

Fig. 4: Attention maps of audio-injected Transformer blocks across layers. It is observed that injecting audio features into block 3 and 4 bringing clearer instance-level attention, compared to the blurry pattern at earlier blocks.

background distractions without explicit supervision.

$$\tilde{H}_a^i = \text{SelfAttn}(Q) \tag{9}$$

$$\tilde{H}_v^i = \text{SelfAttn}(H_v^{i-1}) \tag{10}$$

$$H_a^i = \text{CrossAttn}(\tilde{H}_a^i,\, \tilde{H}_v^i,\, \tilde{H}_v^i) \tag{11}$$

$$H_v^i = \text{CrossAttn}(\tilde{H}_v^i,\, H_a^i,\, H_a^i) \tag{12}$$

Audio-Guided Filtering. Audio queries attend to visual tokens to extract sound-relevant visual evidence as shown in Eqs. (9) and (11). Here $\text{SelfAttn}(x)$ and $\text{CrossAttn}(x_Q, x_K, x_V)$ denote standard transformer self-attention and cross-attention. This step generates audio-conditioned visual features focused on regions visually correlated with emitted sounds. Using visual features from pre-trained encoder as keys and values simultaneously constrains the relatively noisier audio representations, providing an implicit denoising effect.

Visual-Guided Enhancement. The updated audio representations then act as keys and values to inject discriminative acoustic cues back into the visual stream following Eqs. (10) and (12). This reverse attention leverages the purified acoustic cues to further sharpen visual activations at sound-producing locations, completing a robust bidirectional alignment cycle. Ablation on this bidirectional ordering is provided in the supplementary material.

Delayed Cross-Modal Alignment. Cross-modal Alignment is applied exclusively between the third and fourth layers of the four-block architecture. As shown in Fig. 4, early fusion captures fragmented pixel patterns, while deeper fusion highlights coherent region- and instance-level structures essential for audio-visual grounding. Quantitative results in Tab. 5 confirm the block-3-and-4 configuration achieves peak $\mathcal{J}\&\mathcal{F}$ performance. The resulting audio-conditioned visual features H_v^L are then decoded into the segmentation mask $\hat{Y} = \mathcal{D}(H_v^L)$.

3.4 Optimization

We train the DDAVS model with a unified objective:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{ce}}\mathcal{L}_{\text{ce}} + \lambda_{\text{dice}}\mathcal{L}_{\text{dice}} + \lambda_{\text{iou}}\mathcal{L}_{\text{iou}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}}. \tag{13}$$

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

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 [66] [ECCV22]	83.3	78.7	87.9	59.3	54.0	64.5	32.5	29.8	35.2
CATR [27] [ACM-MM23]	87.9	84.4	91.3	68.6	62.7	74.5	35.7	32.8	38.5
AuTR [34] [Arxiv23]	82.1	77.6	86.5	72.0	66.2	77.7	–	–	–
AVSC [32] [ACM-MM23]	85.0	81.3	88.6	62.6	59.5	65.8	–	–	–
ECMVAE [42] [ICCV23]	85.9	81.7	90.1	64.3	57.8	70.8	–	–	–
AQFormer [16] [IJCAI23]	85.5	81.6	89.4	67.5	62.2	72.7	–	–	–
BAVS [31] [TMM24]	86.2	82.7	89.8	62.8	59.6	65.9	35.6	33.6	37.5
AVSegFormer [7] [AAAI24]	86.8	83.1	90.5	67.2	61.3	73.0	40.1	37.3	42.8
GAVS [51] [AAAI24]	85.1	80.1	90.0	70.6	63.7	77.4	–	–	–
AVSBG [13] [AAAI24]	86.1	81.7	90.4	61.0	55.1	66.8	–	–	–
AVESFormer [54] [Arxiv24]	84.5	79.9	89.1	63.3	57.9	68.7	34.0	31.2	36.8
QDFormer [28] [CVPR24]	83.9	79.5	88.2	64.0	61.9	66.1	–	–	–
CAVP [4] [CVPR24]	83.8	78.8	88.9	61.5	55.8	67.1	32.8	30.4	35.3
COMBO [57] [CVPR24]	88.3	84.7	91.9	65.2	59.2	71.2	44.1	42.1	46.1
AAVS [39] [ECCV24]	87.3	83.2	91.3	72.5	67.3	77.6	<u>50.9</u>	<u>48.5</u>	<u>53.2</u>
CPM [5] [ECCV24]	85.9	81.4	90.5	65.4	59.8	71.0	37.1	34.5	39.6
BiasAVS [49] [ACM-MM24]	88.2	83.3	93.0	74.0	67.2	<u>80.8</u>	47.2	44.4	49.9
DiffusionAVS [41] [TIP25]	85.9	81.5	90.3	65.4	59.6	71.2	40.6	38.1	43.0
VCT [17] [CVPR25]	88.5	84.7	92.3	73.4	67.5	79.3	50.4	47.9	52.9
DDESeg [33] [CVPR25]	<u>91.1</u>	<u>89.1</u>	93.1	72.2	68.1	76.2	49.6	47.1	52.1
TAViS [37] [ICCV25]	88.0	84.8	91.2	72.1	68.2	75.9	–	44.2	–
ICF [60] [ICCV25]	90.1	86.6	<u>93.5</u>	69.9	64.4	75.4	48.2	45.0	51.3
CCFormer [10] [TMM25]	88.8	84.9	92.7	<u>75.6</u>	<u>70.7</u>	80.5	45.4	42.3	48.5
DDAVS (Ours)	**92.4**	**90.6**	**94.2**	**76.0**	**70.9**	**81.1**	**52.9**	**50.2**	**55.6**

alignment. Beyond these segmentation losses, the contrastive term $\mathcal{L}_{\text{con}}$ (See Eq. (8)) 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

Implementation Details. The visual backbone is initialized from MiT-B5 [55], and the audio encoder adopts HTSAT [2] pretrained on AudioSet [8]. Following DDESeg [33], we construct the prototype memory bank from single-sounding source signals. Specifically, we build the bank from training splits by filtering 12356/12202 single-source clips across 71/21 categories for AVS-Semantic/VPO-SS, respectively. For each category, HTSAT embeddings are clustered via K-means++, and the $K_c = 5$ features nearest to the cluster center are fixed as prototypes, yielding 355/105 anchors in total. 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 [65, 66] and VPO [4], which cover single-source, multi-

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 underlined.

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 [66] [ECCV22]	44.63	41.64	47.62	45.68	42.30	49.06	43.19	40.03	46.34
AVSegFormer [7] [AAAI24]	45.94	43.81	48.06	43.72	47.30	40.14	49.93	47.19	52.67
CAVP [4] [CVPR24]	67.02	58.81	75.23	61.32	53.24	69.39	56.48	48.18	64.78
BiasAVS [49] [ECCV24]	67.46	59.14	75.78	63.42	55.61	71.23	57.94	49.60	66.27
CPM [5] [ECCV24]	73.49	67.09	79.88	72.91	65.91	79.90	68.07	60.55	75.58
AAVS [39] [ACM-MM24]	68.54	59.72	77.35	64.26	56.23	72.29	58.76	50.11	67.40
RAVS [30] [CVPR25]	74.97	68.03	81.90	73.49	66.97	80.01	69.30	61.89	76.70
DDESeg [33] [CVPR25]	74.38	67.55	81.20	74.30	67.64	80.96	68.39	62.11	74.67
DDAVS	**75.03**	**68.06**	**81.99**	**76.11**	**69.61**	**82.60**	**72.84**	**65.96**	**79.72**

Table 3: Quantitative component ablation. Performance and efficiency metrics of DDAVS variants on AVS-MS3 and AVSS.

Method	Params	FLOPs	*AVS-MS3*			*AVSS*		
			$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Baseline	129.02M	83.56G	69.71	65.88	73.54	48.63	45.83	51.42
+AQM	152.06M	85.39G	71.89	68.04	75.73	49.80	46.73	52.86
+AQM+COM	152.06M	85.39G	74.07	69.32	78.81	51.70	48.56	54.83
+AVAM	127.24M	83.91G	73.16	68.96	77.36	51.45	48.13	54.76
+AQM+AVAM	150.29M	85.72G	73.75	69.06	78.44	51.52	48.25	54.79
DDAVS	150.29M	85.72G	**76.01**	**70.92**	**81.10**	**52.94**	**50.20**	**55.67**

source, and semantic conditions. Following common practice [4, 66] 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 [66], while for VPO we follow the metric implementation of CAVP [4].

4.1 Quantitative Evaluation

AVSBench. Tab. 1 presents the experimental results on AVSBench. DDAVS achieves state-of-the-art performance across all subsets. On the semantic subset AVSS involving spatial and categorical ambiguity, DDAVS improves over previous best baseline by 2.0% $\mathcal{J}\&\mathcal{F}$, indicating that disentangled audio queries and dual-stage fusion effectively reduce interference between overlapping sources.
VPO. Tab. 2 presents the experimental results on VPO. 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 that the advantage of DDAVS in complex multi-source scenes that require robust cross-modal representation, while it also remains competitive under simple conditions such as VPO-SS.

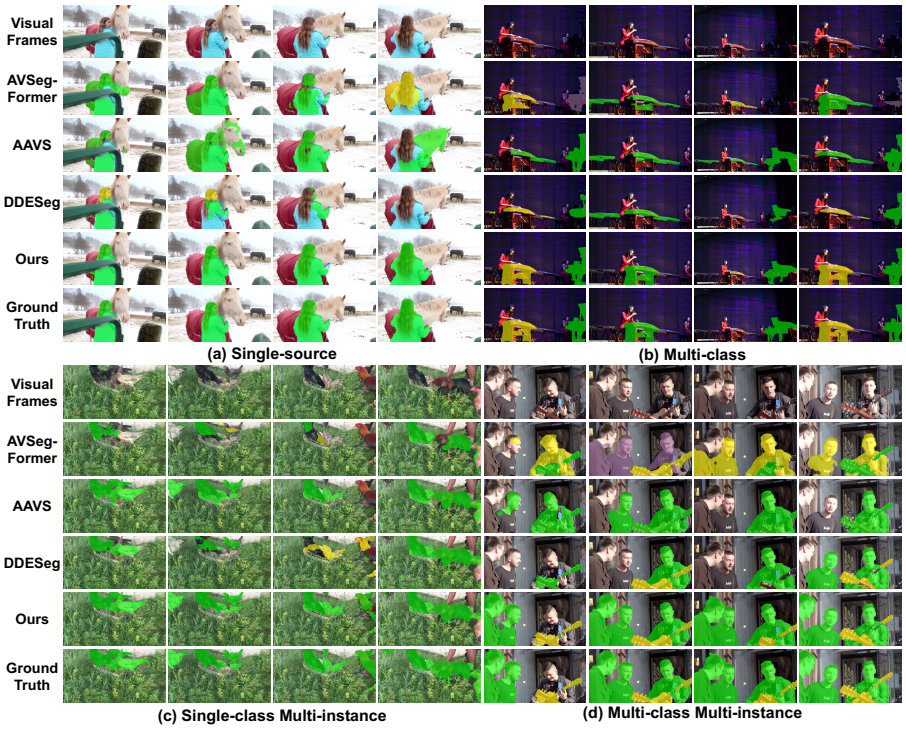

Fig. 5: Qualitative results on AVSBench. DDAVS produces cleaner and more source-consistent masks than previous baselines AVSegFormer, AAVS, and DDESeg, especially in complex multi-source scenes (multi-class, multi-instance).

4.2 Qualitative Evaluation

Qualitative Results. As shown in Fig. 5, DDAVS produces cleaner and more precise segmentation masks than baselines across diverse scenarios: (a) isolates the speaking human and suppresses the silent horse while prior methods leak activation to the horse. (b) segments all sounding instruments(e.g., saxophone, guitar) without activating silent objects whereas others miss sources or misassign segments. (c) distinctly identifies each person and their guitar while competing models merge individuals or omit instruments. (d) robustly handles complex multi-person multi-instrument scenes under occlusion while baselines exhibit misassignment and fragmentation. consistently demonstrates superior multi-source disentanglement and mask fidelity.

Robustness in Multi-Source Scenarios. We evaluate weak target localization under acoustic imbalance. As shown in Fig. 6, the baseline focuses diffusely on the dominant sound (dog) and misses the weak target (person), whereas DDAVS generates sharp, complete activations for both sources, confirming its robust disentanglement of overlapping audio-visual signals.

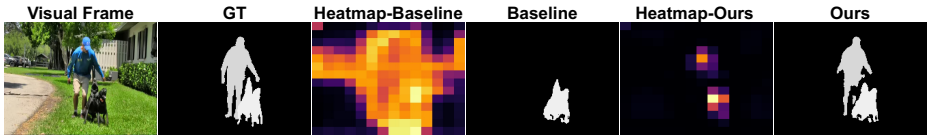

Fig. 6: Weak target localization. DDAVS precisely activates both dominant (dog) and weak (person) targets, while the baseline is biased toward the louder sound.

Fig. 7: Qualitative component ablation. Visual comparison of segmentation masks: AQM improves localization, COM improves audio-query robustness, AVAM strengthens spatial alignment. Full model achieves precise multi-source segmentation.

4.3 Ablation Study

Component Ablation. We evaluate DDAVS components quantitatively (Tab. 3) and qualitatively (Fig. 7). The baseline only contains encoders, transformer blocks and the segmentation deocder. Adding AQM yields moderate gains with minimal overhead, attributable to effective bank-based audio extraction. When COM is enabled, performance improvements become significantly more pronounced, confirming that contrastive learning is essential for robust audio representation without adding inference cost. AVAM alone delivers strong improvements by efficiently injecting audio cues for cross-modal fusion. Crucially, adding AQM+COM on top of AVAM further boosts performance from 73.16/51.45 to 76.01/52.94, confirming that optimizing query separability is essential even when spatial alignment is already established. The full DDAVS framework achieves optimal performance with $\mathcal{J}\&\mathcal{F}$ gains of +6.30 points on AVS-MS3 and +4.31 points on AVSS, while maintaining practical feasibility with only +2.16G FLOPs overhead over the baseline. Further efficiency details (FPS, training time, etc) are provided in the supplementary material. Qualitative results (Fig. 7) corroborate these findings: AQM sharpens source localization, COM improves audio-query robustness and reduces frame-wise instability, and AVAM refines spatial alignment. Their synergy yields precise, clean masks even in complex multi-source scenarios.

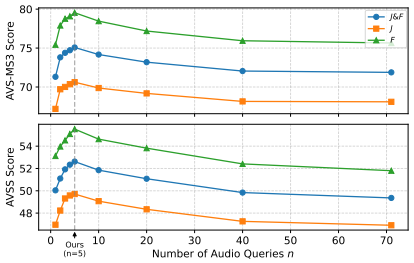

Fig. 8: Effect of audio query number. $n = 5$ yields the best results, while higher values cause degradation.

Table 4: Out-of-bank robustness. Performance when removing x% of class prototypes in the memory bank.

Setting	*MS3* ($\mathcal{J}\&\mathcal{F}$)	*AVSS* ($\mathcal{J}\&\mathcal{F}$)
Baseline	69.71	48.63
OOB-100%	73.16	51.45
OOB-80%	73.61	51.58
OOB-50%	74.18	51.72
OOB-20%	75.29	52.30
Full Bank	**76.01**	**52.94**

Table 5: Injection block placement ablation. Blocks 1–4 denote the first to fourth Transformer blocks (from input to output) in the visual backbone [55].

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	74.27	69.42	79.12	51.78	49.03	54.52
4	73.82	68.85	78.79	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)	**76.01**	**70.92**	**81.10**	**52.94**	**50.20**	**55.67**
1,2,3	72.07	67.82	76.32	50.29	47.24	53.33
2,3,4	74.67	69.78	79.55	51.65	48.52	54.77
1,2,3,4	72.57	68.48	76.65	51.07	48.21	53.93

Table 6: Backbone ablation. DDAVS consistently boosts performance across all visual/audio encoder combinations.

Visual Backbone	Audio Backbone	AVS-S4			AVS-MS3			AVSS		
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
PVTv2-B5 [50]	VGGish [15]	91.35	89.36	93.34	73.60	68.77	78.43	47.55	44.11	50.98
PVTv2-B5 [50]	HTSAT	91.18	89.12	93.23	74.69	69.49	79.88	48.62	45.53	51.71
MiT-B5	VGGish [15]	92.34	90.56	94.11	74.45	69.34	79.56	52.15	49.23	55.07
MiT-B5 (Ours)	**HTSAT (Ours)**	**92.43**	**90.61**	**94.24**	**76.01**	**70.92**	**81.10**	**52.94**	**50.20**	**55.67**

Audio Query Quantity. Fig. 8 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.

Injection Blocks in AVAM. To determine the optimal placement of cross-modal alignment within AVAM, we conduct a controlled ablation by freezing the visual backbone and training only the cross-attention modules, thereby iso-

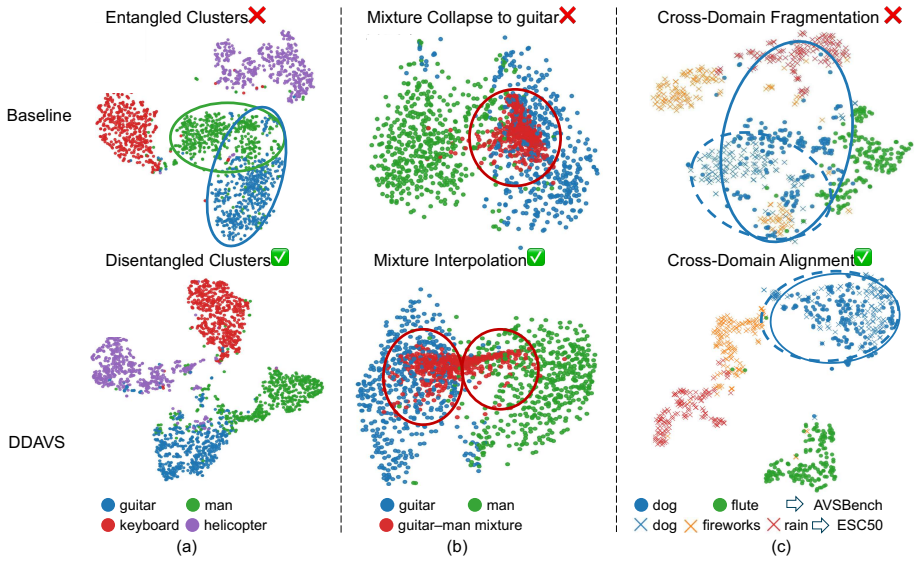

Fig. 9: t-SNE visualization of audio representations. Left and middle columns illustrate multi-source disentanglement on AVSBench. Right column evaluates cross-domain consistency for the shared class *dog*.

lating the effect of injection position from backbone updates. Quantitative results in Tab. 5 show that later-layer injection ({3,4}) consistently outperforms early fusion and other alternatives, achieving peak $\mathcal{J}\&\mathcal{F}$ performance. This configuration is adopted as our default setting. Attention maps in Fig. 4 provide qualitative evidence for the ablation results, revealing that deeper injection enables precise instance-level focusing rather than diffuse background responses.

Out-of-Bank Robustness. We evaluate robustness by progressively removing $x\%$ of class prototypes from the memory bank. As Tab. 4 shows, DDAVS degrades gracefully: even with *all* prototypes absent (OOB-100%), it consistently outperforms the baseline while retaining a clear margin below the full-bank setting. This confirms prototype anchoring provides stable grounding for unseen sounds.

Backbone Variants. Tab. 6 evaluates DDAVS across diverse visual/audio backbone combinations. HTSAT consistently outperforms VGGish [15], confirming its superior acoustic representation capability. MiT-B5 surpasses PVTv2-B5 [50] across all benchmarks, validating its effectiveness for dense prediction tasks. Critically, DDAVS improves every configuration, proving the framework's architecture-agnostic design and strong generalization across backbone choices.

4.4 Representation Analysis

Fig. 9 presents t-SNE visualizations that validate the quality of audio representations in terms of multi-source disentanglement (left and middle) and cross-domain consistency (right).

For **multi-source disentanglement**, Fig. 9(a) shows single-source categories (*guitar*, *man*, *keyboard*, *helicopter*). The baseline exhibits entangled clusters with substantial overlap, while DDAVS achieves well-separated, distinct clusters. In Fig. 9(b), the baseline suffers mixture collapse, where the *guitar-man* mixture collapses into the *guitar* cluster. DDAVS places mixtures along smooth interpolations between their source components, explicitly representing all constituent sources without collapsing to any single one.

For **cross-domain consistency**, we evaluate *dog* embeddings across AVSBench and ESC-50 [45] (a dataset of 2,000 environmental audio clips). The baseline exhibits **cross-domain fragmentation**, with *dog* embeddings dispersed across the embedding space according to dataset origin, undermining intra-class cohesion. In contrast, DDAVS achieves **cross-domain alignment**, consolidating all *dog* embeddings into a single compact cluster that remains distinctly separated from other semantic classes (e.g., *flute*, *fireworks*).

These results show that DDAVS's disentangled mixture representations and domain-invariant embeddings underlie its consistent performance gains.

5 Discussion

Conclusion. In this work, we presented DDAVS, a novel 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.

Limitations and Future Work. Current validation is limited to benchmark scenarios with well-defined misalignment ranges and curated sound categories. In future work, we plan to extend this paradigm to open-domain video analysis and streaming audio applications, paving the way for real-time, scalable, and broadly generalizable audio-visual perception systems.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grant 62572270, and in part by the Guangdong Natural Science Funds for Distinguished Young Scholar (No. 2025B1515020012).

References

1. Bai, S., Liu, Y., Han, Y., Zhang, H., Tang, Y.: Self-calibrated clip for training-free open-vocabulary segmentation. arXiv preprint arXiv:2411.15869 (2024)
2. Chen, K., Du, X., Zhu, B., Ma, Z., Berg-Kirkpatrick, T., Dubnov, S.: 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. pp. 646–650 (2022)
3. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PmLR (2020)
4. Chen, Y., Liu, Y., Wang, H., Liu, F., Wang, C., Frazer, H., Carneiro, G.: Unraveling instance associations: A closer look for audio-visual segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26497–26507 (2024)
5. Chen, Y., Wang, C., Liu, Y., Wang, H., Carneiro, G.: Cpm: Class-conditional prompting machine for audio-visual segmentation. arXiv preprint arXiv:2407.05358 (2024)
6. Du, H., Li, G., Zhou, C., Zhang, C., Zhao, A., Hu, D.: Crab: A unified audio-visual scene understanding model with explicit cooperation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18804–18814 (2025)
7. Gao, S., Chen, Z., Chen, G., Wang, W., Lu, T.: Avsegformer: Audio-visual segmentation with transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 12155–12163 (2024)
8. Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audioset: An ontology and human-labeled dataset for audio events. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 776–780. IEEE (2017)
9. Gong, S., Zhuge, Y., Zhang, L., Wang, Y., Zhang, P., Wang, L., Lu, H.: Avs-mamba: Exploring temporal and multi-modal mamba for audio-visual segmentation. IEEE Transactions on Multimedia (2025)
10. Gong, S., Zhuge, Y., Zhang, L., Zhang, P., Lu, H.: Complementary and contrastive learning for audio-visual segmentation. IEEE Transactions on Multimedia (2025)
11. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems **33**, 21271–21284 (2020)
12. Gu, X., Zhang, H., Fan, Q., Niu, J., Zhang, Z., Zhang, L., Chen, G., Chen, F., Wen, L., Zhu, S.: Thinking with bounding boxes: Enhancing spatio-temporal video grounding via reinforcement fine-tuning. arXiv preprint arXiv:2511.21375 (2025)
13. Hao, D., Mao, Y., He, B., Han, X., Dai, Y., Zhong, Y.: Improving audio-visual segmentation with bidirectional generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 2067–2075 (2024)
14. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
15. Hershey, S., Chaudhuri, S., Ellis, D.P., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., et al.: Cnn architectures for large-scale audio classification. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 131–135. IEEE (2017)

16. Huang, S., Li, H., Wang, Y., Zhu, H., Dai, J., Han, J., Rong, W., Liu, S.: Discovering sounding objects by audio queries for audio visual segmentation. In: Proceedings of the International Joint Conference on Artificial Intelligence (2023)
17. Huang, S., Ling, R., Hui, T., Li, H., Zhou, X., Zhang, S., Liu, S., Hong, R., Wang, M.: Revisiting audio-visual segmentation with vision-centric transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8352–8361 (2025)
18. Jiang, D., Li, W., Cao, M., Zou, W., Li, X.: Speech simclr: Combining contrastive and reconstruction objective for self-supervised speech representation learning. arXiv preprint arXiv:2010.13991 (2020)
19. Jin, H., Lin, K., Zhang, W., Jin, Y., Li, G.: Videocurl: Video curriculum reinforcement learning with orthogonal difficulty decomposition. arXiv preprint arXiv:2601.00887 (2025)
20. Jin, H., Wang, Q., Zhang, W., Liu, Y., Cheng, S.: Videomem: Enhancing ultra-long video understanding via adaptive memory management. arXiv preprint arXiv:2512.04540 (2025)
21. Jin, H., Zhu, R., Du, Z., Jiang, X., Tian, J., Zhang, Q., Ding, J.: Dgpo: Distribution guided policy optimization for fine grained credit assignment. arXiv preprint arXiv:2605.03327 (2026)
22. Ke, L., Ye, M., Danelljan, M., Tai, Y.W., Tang, C.K., Yu, F., et al.: Segment anything in high quality. Advances in Neural Information Processing Systems **36**, 29914–29934 (2023)
23. Kharitonov, E., Rivière, M., Synnaeve, G., Wolf, L., Mazaré, P.E., Douze, M., Dupoux, E.: Data augmenting contrastive learning of speech representations in the time domain. In: 2021 IEEE Spoken Language Technology Workshop (SLT). pp. 215–222. IEEE (2021)
24. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
25. Li, J., Tian, Y.: From waveforms to pixels: A survey on audio-visual segmentation. arXiv preprint arXiv:2508.03724 (2025)
26. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models (2023)
27. Li, K., Yang, Z., Chen, L., Yang, Y., Xiao, J.: Catr: Combinatorial-dependence audio-queried transformer for audio-visual video segmentation. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 1485–1494 (2023)
28. Li, X., Wang, J., Xu, X., Peng, X., Singh, R., Lu, Y., Raj, B.: 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. pp. 3402–3413 (2024)
29. Lin, H., Lv, K., Jiang, X., Tian, J., Du, Z., Ding, J., Zhang, Q., Jin, H.: Visd: Enhancing video reasoning via structured self-distillation. arXiv preprint arXiv:2605.06094 (2026)
30. Liu, C., Li, P., Yang, L., Wang, D., Li, L., Yu, X.: Robust audio-visual segmentation via audio-guided visual convergent alignment. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28922–28931 (2025)
31. Liu, C., Li, P., Zhang, H., Li, L., Huang, Z., Wang, D., Yu, X.: Bavs: Bootstrapping audio-visual segmentation by integrating foundation knowledge. IEEE Transactions on Multimedia (2024)

32. Liu, C., Li, P.P., Qi, X., Zhang, H., Li, L., Wang, D., Yu, X.: Audio-visual segmentation by exploring cross-modal mutual semantics. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 7590–7598 (2023)
33. Liu, C., Yang, L., Li, P., Wang, D., Li, L., Yu, X.: 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). pp. 3131–3141 (2025)
34. Liu, J., Ju, C., Ma, C., Wang, Y., Wang, Y., Zhang, Y.: Audio-aware query-enhanced transformer for audio-visual segmentation. arXiv preprint arXiv:2307.13236 (2023)
35. Liu, Y., Bai, S., Li, G., Wang, Y., Tang, Y.: Open-vocabulary segmentation with semantic-assisted calibration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3491–3500 (2024)
36. Liu, Y., Wu, S.L., Bai, S., Wang, J., Wang, Y., Tang, Y.: Stepping out of similar semantic space for open-vocabulary segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22664–22674 (2025)
37. Luo, Z., Liu, N., Yang, X., Khan, S., Anwer, R.M., Cholakkal, H., Khan, F.S., Han, J.: Tavis: Text-bridged audio-visual segmentation with foundation models. arXiv preprint arXiv:2506.11436 (2025)
38. Lv, Y., Liu, Z., Chang, X.: Consistency-queried transformer for audio-visual segmentation. IEEE Transactions on Image Processing (2025)
39. Ma, J., Sun, P., Wang, Y., Hu, D.: Stepping stones: A progressive training strategy for audio-visual semantic segmentation. arXiv preprint arXiv:2407.11820 (2024)
40. Ma, S., Zhang, C., Wang, C., Wang, Y., Wu, Y., Wang, Z., Tian, J., Zhu, Z., Tang, Y.: Safe-pruner: Semantic attention-guided future-aware token pruning for efficient vision-language-action manipulation. arXiv preprint arXiv:2605.29662 (2026)
41. Mao, Y., Zhang, J., Xiang, M., Lv, Y., Li, D., Zhong, Y., Dai, Y.: Contrastive conditional latent diffusion for audio-visual segmentation. IEEE Transactions on Image Processing (2025)
42. Mao, Y., Zhang, J., Xiang, M., Zhong, Y., Dai, Y.: Multimodal variational auto-encoder based audio-visual segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 954–965 (2023)
43. Mo, S., Raj, B.: Weakly-supervised audio-visual segmentation. Advances in Neural Information Processing Systems **36**, 17208–17221 (2023)
44. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
45. Piczak, K.J.: ESC: Dataset for Environmental Sound Classification. In: Proceedings of the 23rd Annual ACM Conference on Multimedia. pp. 1015–1018. ACM Press
46. Seon, J., Im, W., Lee, S., Lee, J., Yoon, S.E.: Extending segment anything model into auditory and temporal dimensions for audio-visual segmentation. In: 2024 IEEE International Conference on Image Processing (ICIP). pp. 2480–2486. IEEE (2024)
47. Shi, Z., Wu, Q., Meng, F., Xu, L., Li, H.: Cross-modal cognitive consensus guided audio-visual segmentation. IEEE Transactions on Multimedia (2024)
48. Sun, G., Yu, W., Tang, C., Chen, X., Tan, T., Li, W., Lu, L., Ma, Z., Wang, Y., Zhang, C.: video-salmonn: Speech-enhanced audio-visual large language models. arXiv preprint arXiv:2406.15704 (2024)
49. Sun, P., Zhang, H., Hu, D.: Unveiling and mitigating bias in audio visual segmentation. arXiv preprint arXiv:2407.16638 (2024)

50. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pvt v2: Improved baselines with pyramid vision transformer. Computational visual media **8**(3), 415–424 (2022)
51. Wang, Y., Liu, W., Li, G., Ding, J., Hu, D., Li, X.: Prompting segmentation with sound is generalizable audio-visual source localizer. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 5669–5677 (2024)
52. Wang, Y., Sun, P., Zhou, D., Li, G., Zhang, H., Hu, D.: Ref-avs: Refer and segment objects in audio-visual scenes. European Conference on Computer Vision (2024)
53. Wang, Y., Zhang, H., Tian, J., Tang, Y.: Ponder & press: Advancing visual gui agent towards general computer control. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 1461–1473 (2025)
54. Wang, Z., Yang, Q., Shi, L., Yu, J., Liang, Q., Li, F., Xiang, S.: Avesformer: Efficient transformer design for real-time audio-visual segmentation. arXiv preprint arXiv:2408.01708 (2024)
55. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in Neural Information Processing Systems **34**, 12077–12090 (2021)
56. Yang, Q., Nie, X., Li, T., Gao, P., Guo, Y., Zhen, C., Yan, P., Xiang, S.: 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. pp. 27134–27143 (2024)
57. Yang, Q., Nie, X., Li, T., Gao, P., Guo, Y., Zhen, C., Yan, P., Xiang, S.: 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 (CVPR). pp. 27134–27143 (June 2024)
58. Yang, Z., Wang, J., Tang, Y., Chen, K., Zhao, H., Torr, P.H.: Lavt: Language-aware vision transformer for referring image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18155–18165 (2022)
59. Ying, K., Ding, H., Jie, G., Jiang, Y.G.: Towards omnimodal expressions and reasoning in referring audio-visual segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22575–22585 (2025)
60. Zha, M., Li, T., Wang, G., Wang, P., Wu, Y., Yang, Y., Shen, H.T.: Implicit counterfactual learning for audio-visual segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22349–22360 (2025)
61. Zhang, B., Li, K., Cheng, Z., Hu, Z., Yuan, Y., Chen, G., Leng, S., Jiang, Y., Zhang, H., Li, X., et al.: Videollama 3: Frontier multimodal foundation models for image and video understanding. arXiv preprint arXiv:2501.13106 (2025)
62. Zhang, H., Gu, X., Li, J., Ma, C., Bai, S., Zhang, C., Zhang, B., Zhou, Z., He, D., Tang, Y.: Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning. arXiv preprint arXiv:2508.04416 (2025)
63. Zhang, H., Wang, Y., Tang, Y., Liu, Y., Feng, J., Jin, X.: Flash-vstream: Efficient real-time understanding for long video streams. arXiv preprint arXiv:2506.23825 (2025)
64. Zhang, J., Du, Y., Wang, Q., Li, W., Gu, Y., Zhang, J.: Alignedgen: Aligning style across generated images. arXiv preprint arXiv:2509.17088 (2025)
65. Zhou, J., Shen, X., Wang, J., Zhang, J., Sun, W., Zhang, J., Birchfield, S., Guo, D., Kong, L., Wang, M., et al.: Audio-visual segmentation with semantics. International Journal of Computer Vision (2024)

66. Zhou, J., Wang, J., Zhang, J., Sun, W., Zhang, J., Birchfield, S., Guo, D., Kong, L., Wang, M., Zhong, Y.: Audio-visual segmentation. In: European Conference on Computer Vision. pp. 386–403. Springer (2022)
67. Zhou, Y.H., Huang, H., Guo, C., Tu, R.C., Xiao, Z., Wang, B., Mao, X.L.: 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)
68. Zhu, T., Zhang, S., Sun, Z., Tian, J., Tang, Y.: Memorize-and-generate: Towards long-term consistency in real-time video generation. arXiv preprint arXiv:2512.18741 (2025)