Do Audio-Visual Segmentation Models Truly Segment Sounding Objects?

Jia Li[1], **Wenjie Zhao**[1], **Ziru Huang**[2], **Yunhui Guo**[1], **Yapeng Tian**[1]

[1]Department of Computer Science, The University of Texas at Dallas, Richardson, TX, USA
[2]Tsinghua University, Beijing, China
{jia.li, wenjie.zhao, yunhui.guo, yapeng.tian}@utdallas.edu, huangzr21@mails.tsinghua.edu.cn

Abstract

Unlike traditional visual segmentation, audio-visual segmentation (AVS) requires the model not only to identify and segment objects but also to determine whether they are sound sources. Recent AVS approaches have achieved impressive performance on standard benchmarks. Yet, an important question remains: Do these models genuinely integrate audio-visual cues to segment sounding objects? Our study reveals a fundamental bias in current methods: they tend to generate segmentation masks based predominantly on visual salience, irrespective of the audio context, resulting in unreliable predictions when sounds are absent or irrelevant. To address this challenge, we introduce AVSBench-Robust, a comprehensive benchmark incorporating diverse negative audio scenarios including silence, noise, and off-screen sounds. We also propose a simple yet effective approach combining balanced training with negative samples and classifier-guided similarity learning. Our extensive experiments show that while state-of-the-art AVS methods consistently fail under negative audio conditions, our approach achieves remarkable improvements in both standard metrics and robustness measures, maintaining near-perfect false positive rates while preserving high-quality segmentation performance.

Code — https://github.com/jiali-home/AVSBench-Robust

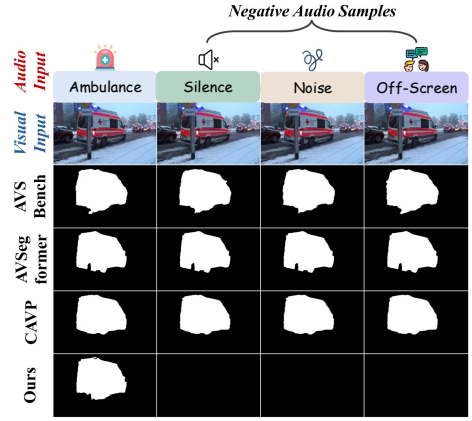


Figure 1: Performance in Different Audio Scenarios. We compare segmentation outputs from SOTA models (Zhou et al. 2022; Gao et al. 2024; Chen et al. 2024) and ours under four audio conditions. Existing models exhibit a strong "visual prior" bias, segmenting the salient ambulance even with negative audio (Silence, Noise, Off-screen). Our model correctly segments the object only in the presence of its corresponding sound, demonstrating true audio-visual alignment.

1 Introduction

Audio-Visual Segmentation (AVS) aims to identify and segment sounding objects within visual scenes (Zhou et al. 2022; Gao et al. 2024). This essential multimodal task mirrors a fundamental aspect of human perception: the integration of auditory and visual stimuli to focus attention on relevant sources (Small and Prescott 2005; Chen et al. 2020). For instance, when hearing a baby cry, people naturally locate the sound's visual source. Simulating this ability in machines could open up valuable cross-modal applications, such as multimedia analysis, human-computer interaction, and autonomous systems.

Recent years have witnessed remarkable progress in AVS. State-of-the-art (SOTA) methods leverage multimodal information, utilizing encoder-decoder structures with audio-visual interaction (Zhou et al. 2022), multimodal transformer architectures (Gao et al. 2024; Li et al. 2024b; Liu et al. 2023b), audio query-guided designs (Liu et al. 2023b; Sun, Zhang, and Hu 2024), and strong foundation models (Mo and Tian 2023; Liu et al. 2024; Wang et al. 2024a; Sun, Zhang, and Hu 2024) like SAM (Kirillov et al. 2023) and Mask2Former (Cheng et al. 2022). These works have achieved impressive performance on current standard benchmark datasets (Zhou et al. 2022).

However, a critical question arises: *are these models truly performing audio-visual segmentation, or simply conducting visual segmentation with minimal audio integration?* To systematically evaluate model robustness, we introduce AVS-Robust, a comprehensive benchmark featuring videos with four distinct audio conditions: normal sound, silence, distracting background noise, and off-screen sounds. Extensive experiments using our new benchmark yield several crucial findings. Recent SOTA methods, including SAMA-AVS (Liu et al. 2024), Stepping-Stones (Ma et al. 2024), and CAVP (Chen et al. 2024), consistently fail under neg-

ative audio conditions, exhibiting high False Positive Rates (FPR). When evaluated with metrics designed to holistically assess performance across both positive and negative audio conditions, these models show significant performance degradation compared to their reported results on standard benchmarks. Our study reveals the training paradigm for existing SOTA methods, which relies on datasets lacking negative or diverse audio conditions, inadvertently fosters a strong visual bias. Consequently, these models learn to equate segmentation with visual salience alone, rather than genuine audio-visual correspondence. For instance, these models segment an ambulance even in silence or with unrelated sounds, indicating an over-reliance on visual cues rather than genuine audio-visual integration (Fig. 1).

Building upon these insights, we explore solutions to address this visual bias. While incorporating negative audio-visual pairs into training seems like an intuitive solution, this approach alone is insufficient. The fundamental issue is that existing models lack a dedicated mechanism to determine whether a segmentation should be generated at all, leaving them unable to use negative audio cues to suppress visually salient but irrelevant objects.

To overcome this, we propose *Classifier-Guided Similarity Learning* to serve as the dedicated decision-making mechanism. This approach utilizes a classifier to guide feature learning, forcing the model to produce highly similar representations for valid audio-visual pairs and highly dissimilar ones for invalid pairs. This learned distance between features enables the model to determine whether a segmentation should be generated at all. The effectiveness of this mechanism is demonstrated by our approach's superior performance across all robustness metrics, where it also maintains competitive segmentation quality on positive audio inputs in both single- and multi-source scenarios.

Our main contributions are summarized as follows:

- We conduct a systematic study on audio robustness in AVS and introduce AVSBench-Robust along with our new robustness evaluation protocols. This benchmark evaluates AVS models under both standard conditions and challenging negative scenarios, assessing their ability to effectively integrate audio-visual information.
- We propose a novel training strategy that enhances AVS robustness by introducing a lightweight classifier-guided similarity learning module, which is trained alongside diverse negative audio sampling.
- Extensive experiments demonstrate that our approach substantially outperforms current SOTA methods in terms of our robustness metrics while achieving competitive performance on standard AVS benchmarks.

2 Related Work

Sound Source Localization. This task is closely related to AVS, focusing on localizing sound sources within visual scenes using bounding boxes or heatmaps (Arandjelovic and Zisserman 2018; Senocak et al. 2018; Mo and Morgado 2022; Chen et al. 2021; Mahmud, Tian, and Marculescu 2024). Recent approaches have improved sound discrimination through innovations like contrastive learning and class-aware feature alignment to handle complex and multi-source scenarios (Chen et al. 2021; Hu et al. 2020; Mo and Morgado 2022; Qian et al. 2020). However, the predicted sounding object heatmaps lack the fine-grained precision offered by AVS's pixel-level segmentation capabilities.

Audio-Visual Segmentation. AVS aims to produce pixel-level masks for sound-producing objects. The field has rapidly advanced from initial encoder-decoder designs with simple fusion (Zhou et al. 2022) to sophisticated architectures using multimodal transformers, audio-query guidance, and large foundation models to improve cross-modal understanding and segmentation accuracy (Gao et al. 2024; Liu et al. 2023b; Sun, Zhang, and Hu 2024; Wang et al. 2024c; Li et al. 2023; Ma et al. 2024; Yang et al. 2024; Liu et al. 2024; Wang et al. 2024a; He et al. 2024; Wang et al. 2024b).A recent survey provides an overview of AVS problem formulations, datasets, evaluation metrics, and methodological trends (Li and Tian 2025). Despite these developments, studies show that AVS models often suffer from a "visual prior" bias, predicting masks based on visual salience regardless of the audio (Sun, Zhang, and Hu 2024; Chen et al. 2024; Li et al. 2024b; Liu et al. 2023a). While recent efforts have tried to mitigate this with contrastive learning or semantic enhancement (Chen et al. 2024; Sun, Zhang, and Hu 2024; Li et al. 2024b), our findings (Fig. 1) show they still largely disregard audio cues under challenging conditions.

In this work, we systematically analyze this issue through AVSBench-Robust, a new benchmark protocol for evaluating models under both standard and challenging negative audio scenarios. A concurrent study (Juanola, Haro, and Fuentes 2024) also identified this visual bias in SSL models and introduced similar negative audio scenarios for evaluation. However, our contribution is twofold: we not only introduce a comprehensive benchmark for robustness but also propose a targeted training strategy to effectively mitigate this bias and enhance the model's reliance on genuine audio-visual correspondence.

3 Problem and Benchmark

In this section, we formulate the AVS task and its challenges, then introduce our AVSBench-Robust benchmark and its evaluation protocols.

Task and Challenges

Given T non-overlapping video and audio clips $\{V^t, A^t\}_{t=1}^{T}$, the goal of the AVS task is to predict a segmentation mask ${\mathcal{M}_{\text{pred}}}^t \in \mathbb{R}^{H \times W}$ that labels sounding pixels in each video frame of the clips, where H and W denote the frame dimensions, and the mask is binary. Following previous studies (Zhou et al. 2022; Gao et al. 2024; Li et al. 2023), we extract a single video frame at the end of each second and set $T = 5$ in practice, so each clip contains only one extracted frame.

Unlike purely visual segmentation, AVS must satisfy two key requirements: accurately segmenting sounding objects, and producing empty masks when no audio-visual correspondence exists (e.g., for silent but visually salient objects).

While current AVS models excel on standard benchmarks, their performance is misleading. These benchmarks, composed almost entirely of 'positive' audio-visual pairs where the sound source is visually salient, have taught models a flawed shortcut: to equate visual salience with the segmentation target. The fundamental issue is that this paradigm prevents models from developing a crucial mechanism for cross-modal validation—the ability to use audio as evidence to reject a visual hypothesis. As a result, they are architecturally incapable of handling realistic negative pairs (e.g., silence or noise) and erroneously segment visually prominent but silent objects. Our work addresses this by teaching models to also determine if a sound correspondence exists when deciding what to segment.

AVSBench-Robust

To address this challenge, we introduce **AVSBench-Robust**, a benchmark that repurposes existing datasets (Zhou et al. 2022, 2024) to systematically evaluate model robustness across single-source, multi-source, and semantic scenarios: (1) the single-source subset (S4), containing 4,932 videos (3,452 for training, 740 for validation, and 740 for testing), (2) the multi-source subset (MS3), comprising 424 videos (296 for training, 64 for validation, and 64 for testing), and (3) the semantic segmentation subset(AVSS), containing 12,356 videos (8,498 for training, 1,304 for validation, and 1,554 for testing) To evaluate model robustness, each video is paired with four types of audio conditions.

Positive Pair: Original audio of the video from AVS-Bench (Zhou et al. 2022) and AVS-Semantic (Zhou et al. 2024). , where the audio accurately reflects the visible objects.

Silence Scenario: Test cases without audio, where objects are visually present but silent in the video.

Noise Condition: Background noise, we generated 5s white noise clips at 44.1 kHz by uniformly sampling values in [-1,1] for testing the model's ability to differentiate between meaningful and irrelevant audio signals.

Off-screen Audio: Semantically unrelated sounds from different major categories, testing the model's ability to maintain accurate audio-visual correspondence.

Evaluation Protocols

To comprehensively evaluate model performance on both positive and negative samples, we propose the following metrics. Let $\mathcal{P}$ and $\mathcal{N}$ denote sets of positive and negative samples, respectively. For positive samples, following established protocols (Zhou et al. 2022; Gao et al. 2024; Ma et al. 2024), we employ mean Intersection over Union (mIoU) and F-score to evaluate segmentation accuracy. For negative ones, we introduce complementary metrics to capture different aspects of model robustness.

False Positive Rate (FPR):

$$\text{FPR} = \frac{\sum_{x \in \mathcal{M}_{\text{pred}}} m(x)}{H \cdot W}, \tag{1}$$

where $m(x)$ denotes the binary indicator (0 or 1) for pixel x in the predicted mask. FPR measures the proportion of incorrectly activated pixels in negative scenarios, directly assessing the model's tendency to generate false predictions.

To evaluate overall performance across both positive and negative cases, we propose three global metrics.

Global mIoU (G-mIoU):

$$\text{G-mIoU} = \frac{2 \cdot \text{mIoU}_{\mathcal{P}} \cdot (1 - \text{mIoU}_{\mathcal{N}})}{\text{mIoU}_{\mathcal{P}} + (1 - \text{mIoU}_{\mathcal{N}})}. \tag{2}$$

where $\text{mIoU}_{\mathcal{P}}$ is the mIoU for positive samples, and $\text{mIoU}_{\mathcal{N}}$ is for negative samples. G-mIoU balances region-level accuracy, emphasizing the model's ability to maintain precise segmentation boundaries while suppressing false activations. A high score indicates accurate object delineation in positive cases and clean masks in negative cases.

Global F-score (G-F):

$$\text{G-F} = \frac{2 \cdot \text{F}_{\mathcal{P}} \cdot (1 - \text{F}_{\mathcal{N}})}{\text{F}_{\mathcal{P}} + (1 - \text{F}_{\mathcal{N}})}. \tag{3}$$

G-F provides a pixel-level assessment that equally weighs precision and recall, which is essential for evaluating both false positives and false negatives. This metric is particularly sensitive to small errors that may be overlooked by IoU-based measures.

Global False Positive Rate (G-FPR):

$$\text{G-FPR} = \frac{1}{|\mathcal{N}|} \sum_{i \in \mathcal{N}} \text{FPR}_i. \tag{4}$$

G-FPR provides a dedicated measure of a model's robustness by averaging the FPR across all negative conditions.

These metrics enable a thorough assessment of both segmentation quality and robustness against audio distractors.

4 Method

To address the architectural gap identified in prior models, we propose a targeted solution: Classifier-Guided Similarity Learning. While the components of this strategy (negative sampling and classifier-based metric learning) are well-established, our core contribution lies in identifying the root cause of the visual bias and demonstrating how this specific approach can be integrated as a simple, powerful decision-making gate. Instead of designing a complex architecture, we show that this explicit guidance is the crucial missing element needed to force existing AVS models to learn true audio-visual correspondence.

Preliminary: AVS Architecture

Encoder: We employ an encoder structure that separately processes audio clip A and visual frames V. Specifically, input audio is converted into spectrograms and processed through a VGGish-based network (Hershey et al. 2017), pre-trained on AudioSet (Gemmeke et al. 2017), to generate audio feature $\mathcal{F}_A \in \mathbb{R}^d$ where $d = 128$. For visual inputs V, we utilize a transformer-based backbone (Wang et al. 2022) to extract hierarchical visual features.$\mathcal{F}_{V_i} \in \mathbb{R}^{h_i \times w_i \times C_i}$, where $(h_i, w_i) = (H, W)/2^{i+1}, i = 1, \ldots, n$. The number of levels is set to $n = 4$ in all experiments.

Cross-Modal Fusion: Following (Zhou et al. 2022), the fusion module involves an Atrous Spatial Pyramid Pooling

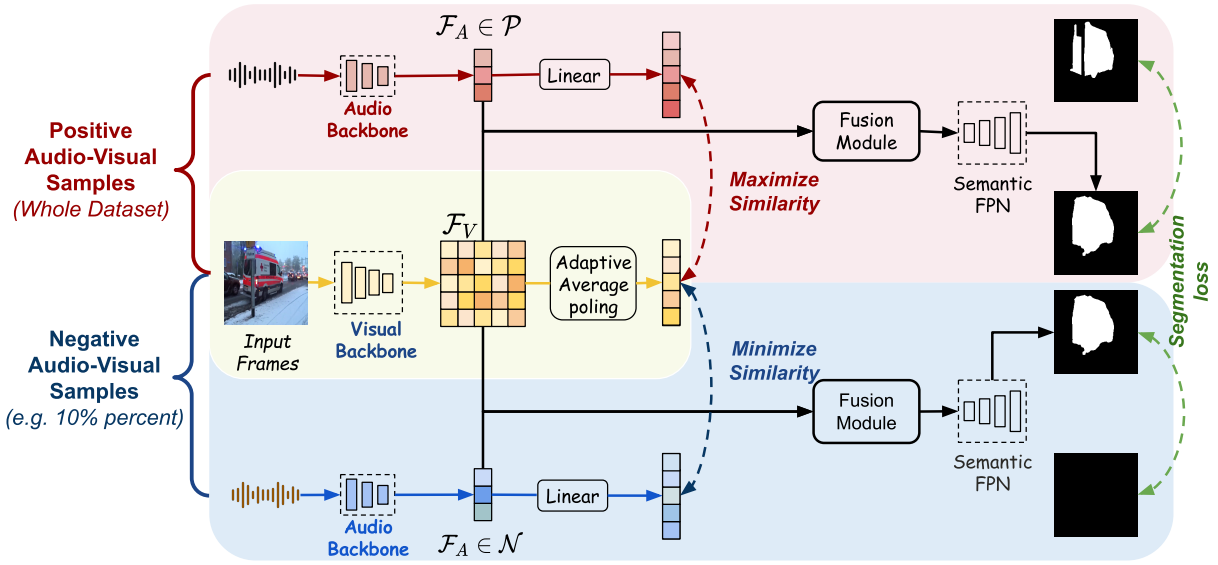


Figure 2: Framework Overview. Given video frames and an audio clip as inputs, our approach can robustly identify and segment sounding objects in video frames. Positive audio-visual pairs represent aligned sound sources, while negative pairs, such as silence or offscreen sounds, correspond to empty masks. The model uses separate visual and audio encoders to extract modality-specific features, applies similarity-based alignment optimized with classifier guidance in a contrastive manner, and integrates features through a fusion module. Positive pairs maximize similarity, while negative pairs minimize it, using a small portion (10%) of the dataset for improved boundary delineation. This dual-stream design facilitates segmentation by distinguishing sound-relevant regions in complex scenes.

(ASPP) module (Chen et al. 2017) that computes the audio-visual similarity α_i through a normalized dot-product between V_i and $\hat{A}$. This similarity guides the interaction, updating the visual features to $Z_i \in \mathbb{R}^{T\times h_i\times w_i\times C}$ as $Z_i = \text{Softmax}(\theta(V_i)\cdot\phi(\hat{A}))\cdot g(V_i) + \mu(V_i)$. Here, θ, ϕ, g, and μ denote $1\times1\times1$ convolutions.

Decoder: The decoder leverages Panoptic-FPN (Kirillov et al. 2019) architecture, which processes outputs from the fusion stage and refines them through upsampling, aiming to recover detailed segmentations at original scale.

Segmentation Loss: the segmentation objective is the binary cross-entropy loss for basic segmentation accuracy.

$$\mathcal{L}_{\text{Seg}} = \mathcal{L}_{BCE}(\mathcal{M}_{pred}, \mathcal{M}_{gt}), \tag{5}$$

where $\mathcal{M}_{pred}$ is the predicted segmentation mask, $\mathcal{M}_{gt}$ is the ground-truth (GT) mask. Additionally, an audio-visual regularization term is added to ensure that if the audio features of some frames are close in the feature space, the corresponding sounding objects are expected to be close in the feature space.

$$\mathcal{L}_{AVM} = \sum_{i=1}^{n} \text{KL}(avg(M_{pred}\odot Z_i), A_i). \tag{6}$$

Framework Overview

Our framework, as illustrated in Fig. 2, processes both positive and negative audio-visual pairs to learn robust correspondence for segmentation. Built upon the presented AVS architecture, our model achieves balanced training by incorporating negative audio-visual pairs, enhancing robustness in AVS. Within this framework, audio and visual features are extracted and used to compute cosine similarity scores for both positive pairs $\mathcal{P}$ and negative pairs $\mathcal{N}$, allowing the model to differentiate aligned from unaligned audio-visual pairs. For mask prediction, we employ a segmentation module that combines a fusion module and an FPN decoder, enabling precise segmentation of sound-producing objects. The dual-stream design allows the model to accurately identify sound-relevant regions in complex scenes while suppressing predictions when no valid audio-visual correspondence exists. The following sections detail each component and their integration within the framework.

Learning with Balanced Audio-Visual Pairs

Given a video clip with its corresponding audio signal, we construct two types of pairs: *Positive Pairs ($\mathcal{P}$)* and *Negative Pairs ($\mathcal{N}$)*. We maintain 10% of negative pairs during training, which we empirically found to optimally balance robustness, segmentation accuracy, and training efficiency. Expanding the diversity of training samples is anticipated to further enhance the model's robustness.

Classifier-Guided Feature Alignment

However, we observed that simply introducing negative pairs is insufficient to mitigate the visual bias, as show in Table 3. Due to the inherent bias in existing models, which often fail to effectively utilize audio information, the model

		Global metric			Positive audio input		Unseen Audio		
Test set	**Model**	**G-mIoU↑**	**G-F↑**	**G-FPR↓**	mIoU ↑	F-score ↑	mIoU ↓	F-score ↓	FPR ↓
AVSBench-S4	TPAVI (Zhou et al. 2022)	35.032	21.479	0.186	78.7	87.9	78.22	87.52	18.69
	AVSegFormer (Gao et al. 2024)	28.199	17.355	0.188	82.1	89.9	82.98	90.40	18.81
	Stepping-Stones (Ma et al. 2024)	28.980	15.806	0.190	83.2	91.3	82.51	91.27	19.02
	SAMA-AVS (Liu et al. 2024)	52.688	40.417	0.155	83.1	90.0	69.7	79.9	17.00
	CAVP (Chen et al. 2024)	33.526	19.891	0.185	78.7	88.8	78.7	88.8	18.53
	CAVP (Chen et al. 2024)	33.526	19.891	0.185	78.7	88.8	78.7	88.8	18.53
	COMBO (Yang et al. 2024)	26.062	14.888	0.190	84.7	91.9	84.6	91.9	18.9
	Selm (Li et al. 2024a)	41.543	30.054	0.175	76.6	86.2	71.5	81.8	17.67
	VCT (Huang et al. 2025)	23.791	12.329	0.193	**86.2**	**93.4**	86.2	93.4	19.71
	TPAVI + Ours	**87.672**	**82.461**	**0.000**	78.1	88.2	0.61	**17.31**	**0.00**
	AVSegFormer + Ours	85.069	80.849	0.001	74.2	84.8	**0.16**	22.59	**0.00**
AVSBench-MS3	TPAVI (Zhou et al. 2022)	59.468	51.036	0.072	54.0	64.5	42.24	62.16	9.53
	AVSegFormer (Gao et al. 2024)	54.889	46.571	0.103	61.3	73.8	51.14	66.24	11.73
	Stepping-Stones (Ma et al. 2024)	61.439	43.937	0.114	67.3	77.6	45.81	68.04	12.48
	SAMA-AVS (Liu et al. 2024)	65.308	65.038	0.125	**68.6**	**78.3**	41.9	47.5	14.36
	CAVP (Chen et al. 2024)	49.647	47.262	0.110	45.8	61.7	45.81	61.72	11.02
	Selm (Li et al. 2024a)	51.281	44.021	0.127	60.5	71.5	55.0	68.2	13.56
	VCT (Huang et al. 2025)	47.908	33.387	0.131	67.6	81.4	62.9	79.0	14.03
	TPAVI + Ours	65.427	70.911	0.001	51.3	64.5	9.06	33.96	0.82
	AVSegFormer + Ours	**73.354**	**78.244**	**0.000**	61.5	74.0	**9.06**	**16.96**	**0.00**

Table 1: Performance comparison of different models on AVSBench-S4 and AVSBench-MS3 under global metrics and positive / unseen audio inputs. Higher is better for G-mIoU, G-F, mIoU, and F-score, while lower is better for G-FPR and FPR.

Model	**G-mIoU↑**	**G-F↑**	**G-FPR↓**
TPAVI (Zhou et al. 2024)	41.90	45.51	0.103
AVSegFormer (Gao et al. 2024)	47.41	49.69	0.116
Stepping-Stones (Ma et al. 2024)	49.10	48.03	0.184
AVSegFormer + Ours	**54.62**	**65.09**	**0.003**

Table 2: Summary of AVSBench-Semantic (AVSS) performance. Higher is better for G-mIoU and G-F, while lower is better for G-FPR.

	Negative samples	$\mathcal{L}_{\text{BCE}}$	G-mIoU↑	G-F↑	G-FPR↓
S4	×	×	35.032	21.479	0.186
	✓	×	34.847	21.993	0.189
	✓	✓	**87.672**	**82.461**	**0.000**
MS3	×	×	59.468	51.036	0.072
	✓	×	55.489	30.057	0.095
	✓	✓	**66.605**	**70.590**	**0.004**

Table 3: Effects of negative samples and classifier guidance.

tends to behave more like a purely visual segmentation model. Without explicit guidance, adding negative pairs can lead to confusion during training, as the model alternates between predicting object masks and empty masks. We therefore introduce a classifier to directly supervise audio-visual similarity learning, creating clear decision boundaries for correspondence detection.

Given multi-scale visual features $\mathcal{F}_i \in \mathbb{R}^{h_i \times w_i \times C_i}$ from the backbone, we use the final-stage features $\mathcal{F}_4 \in \mathbb{R}^{h_4 \times w_4 \times C_4}$ and audio features $\mathcal{F}_A \in \mathbb{R}^{D_a}$ for similarity computation. We project $\mathcal{F}_A$ to C_4 dimensions via a linear layer and apply spatial pooling to $\mathcal{F}_4$ to obtain aligned features $\hat{\mathcal{F}}_A, \hat{\mathcal{F}}_V \in \mathbb{R}^{C_4}$. Their correspondence is then computed through cosine similarity:

$$s(F_A, F_V) = \cos(\hat{\mathcal{F}}_A, \hat{\mathcal{F}}_V). \tag{7}$$

We then apply BCE loss to explicitly guide similarity learning in a contrastive manner:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{|\mathcal{P}| + |\mathcal{N}|} \sum_{j=1}^{|\mathcal{P}|+|\mathcal{N}|} [y_j \log \sigma(s_j) + (1 - y_j) \log(1 - \sigma(s_j))], \tag{8}$$

where $\sigma(\cdot)$ is the sigmoid function, y_j is the binary label (1 for positive pairs, 0 for negative pairs), and $|\mathcal{P}| + |\mathcal{N}|$ is the total number of positive pairs and the total number of negative pairs respectively. This loss pushes the similarity score towards 1 for positive pairs and 0 for negative pairs, teaching the model to rely on audio cues only when a meaningful correspondence exists.

Joint Training with Segmentation

Our total objective function $\mathcal{L}$ can be computed as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{Seg}} + \lambda_2 \mathcal{L}_{AVM}, \tag{9}$$

where λ_1 and λ_2 are balancing weights. The three terms jointly enforce robust and effective learning in AVS models: 1) The first term determines whether segmentation should occur based on audio-visual correspondence; 2) The second term ensures correct segmentation masks when correspondence exists; 3) The third term ensures the masked visual features have similar distributions with the corresponding audio features; 4) For negative pairs, the empty GT masks would guide the segmentation loss to suppress predictions.

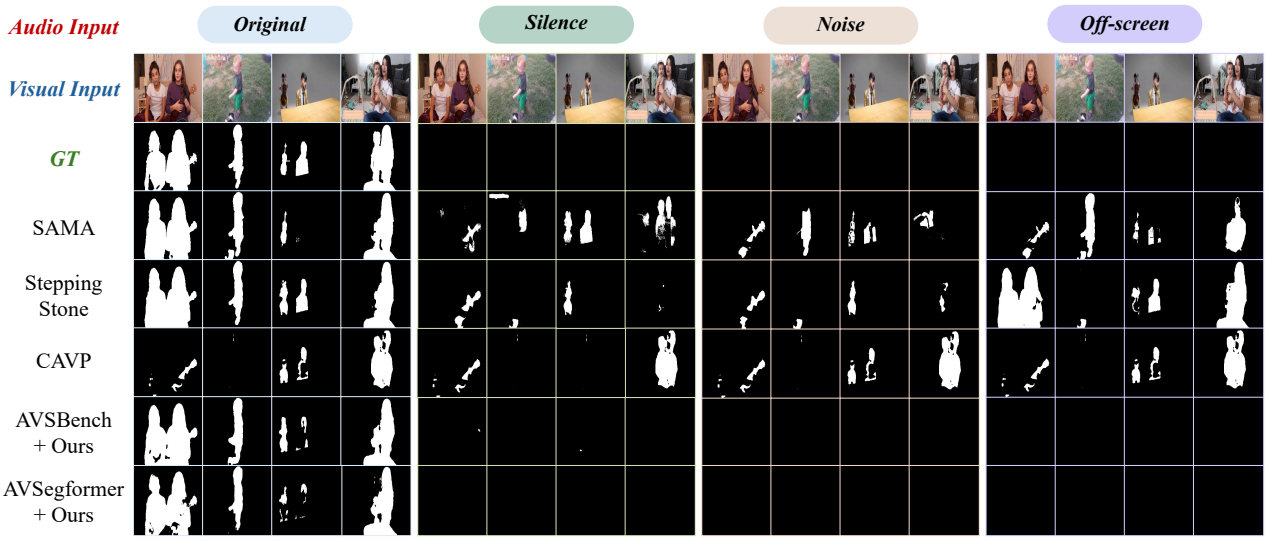


Figure 3: Performance comparison of different AVS models under various audio conditions on the Robust-MS3 dataset. Existing state-of-the-art methods (Liu et al. 2024; Ma et al. 2024; Chen et al. 2024) mainly segment objects based on visual salience, exhibiting a strong visual bias. In contrast, our approach achieves accurate segmentation with original audio while successfully rejecting predictions in negative scenarios (e.g., silence, noise, and off-screen sounds).

This simple, well-motivated approach can achieve strong performance without relying on complex model modifications, making our method easier to implement, tune, and integrate with existing AVS architectures.

5 Experiment

Dataset. We conduct experiments on our *AVSBench-Robust* benchmark. As detailed previously, this benchmark evaluates models under positive (original audio) and three negative (silence, noise, off-screen) audio conditions across single-source, multi-source, and semantic scenarios.

Baselines. We benchmark our model against notable methods including TPAVI (Zhou et al. 2022) and AVSegFormer (Gao et al. 2024), representing fusion-based and prompt-based approaches, respectively. We also compared our method with the CAVP (Chen et al. 2024), SteppingStones (Ma et al. 2024), SAMA-AVS (Liu et al. 2024) and COMBO (Yang et al. 2024). These baselines allow us to demonstrate the broad applicability of our method by comparing it against state-of-the-art models designed to address different aspects of audio-visual segmentation.

Evaluation Metrics. Evaluation metrics, including mIoU, F1 score, FPR, and G-mIou, G-F, G-FPR, are used to assess the segmentation accuracy and robustness of AVS models.

Experimental Comparison

Our experiments, summarized in Table 1 and Table 2, reveal critical flaws in current AVS models.

SOTA methods fail under negative audio conditions. Existing methods exhibit a strong visual bias, producing nearly identical segmentations regardless of the audio context (positive, silent, noise, or off-screen). This is evident from their high False Positive Rates (FPR) and poor global metrics (e.g., G-mIoU between 28.19 and 35.03 on AVSBench-S4), indicating they largely ignore the audio signal and instead perform visual object segmentation.

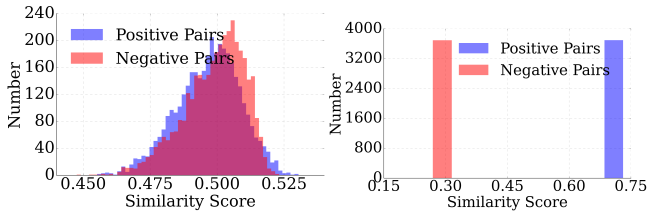
Our method can resolve bias. Our approach dramatically improves robustness across single-source (S4), multi-source (MS3), and semantic (AVSS) settings. When integrated with both TPAVI and AVSegFormer, our method reduces the FPR to near-zero under all negative conditions while maintaining competitive performance on positive samples. This leads to state-of-the-art global scores; for instance, our TPAVI variant achieves a G-mIoU of 87.672 on S4, substantially outperforming all baselines. These results confirm the effectiveness and generality of our approach. Visualizations are provided in Figure 3 and Figure 5.

Trade-off for Enhanced Reliability. While our method significantly boosts robustness, it can cause a slight dip in positive-only metrics. This occurs because it forces the model to abandon its over-reliance on visual cues in favor of genuine audio-visual correspondence. We consider this an advantageous trade-off, as the resulting model is far more reliable. This enhanced reliability is quantitatively confirmed by the dramatic improvements in our global scores, which assess performance beyond ideal, positive-only conditions.

Ablation Study and Analysis

We conducted ablation study using TPAVI (Zhou et al. 2022) as the baseline to analyze the individual contributions of our proposed components, with results summarized in Table 3.

Classifier guidance is essential for leveraging negative samples. Our results show that simply adding negative samples without explicit guidance is detrimental, degrading performance (e.g., on MS3, G-mIoU drops from 59.47 to 55.49). The model becomes confused, unable to decide whether to segment a salient object or not. However, when



(a) Before (b) After

Figure 4: Cosine similarity distributions between paired features before and after training. (a) Similar distributions in positive and negative pairs indicate the model's limited ability to distinguish audio-visual correspondence. (b) Well-separated distributions after classifier-guided similarity learning demonstrate the model's enhanced capability to identify valid audio-visual pairs.

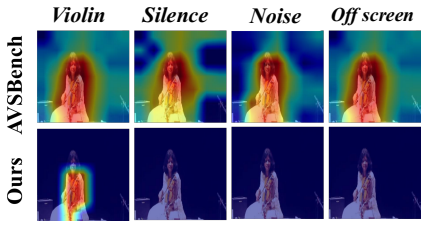


Figure 5: Qualitative comparison of audio-visual attention maps. Brighter colors indicate stronger model attention. *Baseline (Top Row):* The original model exhibits strong visual bias, incorrectly focusing on the musician even with silent, noisy, or off-screen audio. *Our Method (Bottom Row):* Our classifier-guided model successfully learns correspondence, generating a focused attention map only for the correct audio and suppressing for all negative scenarios.

combined with our classifier guidance, performance improves dramatically across all metrics (e.g., on S4, G-mIoU reaches 87.672 and G-FPR drops to 0). As shown in Fig. 4, the classifier successfully separates the feature similarity distributions for positive and negative pairs, enabling the model to learn true audio-visual correspondence and robustly suppress false predictions.

Why use 10% negative pairs? As shown in Table 5, we found that *adding even a small fraction of negative samples provides a massive performance boost*. On S4, introducing just 10% negative pairs increases G-mIoU from 35.032 to 87.672 and reduces FPR from 0.186 to 0.000. While higher proportions (20-30%) yield similar results, the gains are marginal. We therefore adopt a 10% ratio as it offers the best trade-off between performance and training efficiency.

Verifying True Correspondence vs. Pattern Memorization. To verify that model learns genuine audio-visual correspondence rather than simply memorizing negative audio patterns, we conducted a control experiment. Here, negative audio (noise) was digitally mixed with the original positive audio to act as a distractor. As shown in Table 4, the model's performance remains high (e.g., 77.18 mIoU on S4), with only a negligible drop compared to using the original audio alone. This result confirms that the model correctly prioritizes the relevant audio cues, demonstrating true correspondence learning instead of a simple suppression mechanism.

Audio Input Type	S4 Dataset mIoU	S4 Dataset F	MS3 Dataset mIoU	MS3 Dataset F
Original audio	78.15	88.22	51.27	64.51
Fused audio (original + noise)	77.18	86.44	50.76	63.60

Table 4: Performance with fused audio. We mix negative audio (noise) with the original signal to test whether the model simply rejects known negative patterns. The small performance drop indicates that the model correctly prioritizes the relevant audio source.

	Model	Pos pairs (%)	Neg pairs (%)	G-mIoU↑	G-F↑	G-FPR↓
S4	Baseline	100	0	35.032	21.479	0.186
	Ours	90	10	87.672	82.461	0.000
		80	20	**87.780**	**82.114**	0.000
		70	30	87.204	82.233	0.000
MS3	Baseline	100	0	59.468	51.036	0.072
	Ours	90	10	65.427	70.911	0.001
		80	20	**67.909**	**72.572**	0.003
		70	30	66.251	72.908	0.000

Table 5: Impact of the positive-negative pair ratio on model.

For a more detailed exploration of our work, we provide an extensive appendix. This includes (A) comprehensive implementation details for reproducibility, (B) further experimental analysis with qualitative visualizations and evaluations, (C) a full suite of ablation studies and sensitivity analyses for our proposed components, and (D) a broader discussion on the limitations, failure cases, computational efficiency and potential future applications of our method.

Limitations. The performance of our method is limited by the capacity of the backbone model, and although our benchmark is comprehensive, it does not fully capture the complexity of real-world scenarios.

6 Conclusion

Our comprehensive study using AVSBench-Robust reveals that current SOTA methods exhibit strong visual bias, generating segmentation masks based predominantly on visual salience regardless of audio context. To address this issue, we introduce a simple yet effective approach combining balanced training with negative audio-visual pairs and classifier-guided feature alignment, which improves model robustness while maintaining competitive performance on standard AVS tasks. We hope our work could serve as a foundation for future research on robust audio–visual segmentation, inspire further research in this worthwhile field.

Acknowledgments

This work was supported in part by an Amazon Research Award Fall 2023. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not reflect the views of Amazon.

References

Arandjelovic, R.; and Zisserman, A. 2018. Objects that sound. In *Proceedings of the European conference on computer vision (ECCV)*, 435–451.

Chen, H.; Xie, W.; Afouras, T.; Nagrani, A.; Vedaldi, A.; and Zisserman, A. 2021. Localizing visual sounds the hard way. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16867–16876.

Chen, H.; Xie, W.; Vedaldi, A.; and Zisserman, A. 2020. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 721–725. IEEE.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4): 834–848.

Chen, Y.; Liu, Y.; Wang, H.; Liu, F.; Wang, C.; Frazer, H.; and Carneiro, G. 2024. Unraveling Instance Associations: A Closer Look for Audio-Visual Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26497–26507.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1290–1299.

Gao, S.; Chen, Z.; Chen, G.; Wang, W.; and Lu, T. 2024. Avsegformer: Audio-visual segmentation with transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 12155–12163.

Gemmeke, J. F.; Ellis, D. P.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R. C.; Plakal, M.; and Ritter, M. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 776–780. IEEE.

He, J.; Wang, Y.; Wang, L.; Lu, H.; He, J.-Y.; Lan, J.-P.; Luo, B.; and Xie, X. 2024. Multi-modal Instruction Tuned LLMs with Fine-grained Visual Perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13980–13990.

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. 2017. CNN architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, 131–135. IEEE.

Hu, D.; Qian, R.; Jiang, M.; Tan, X.; Wen, S.; Ding, E.; Lin, W.; and Dou, D. 2020. Discriminative sounding objects localization via self-supervised audiovisual matching. *Advances in Neural Information Processing Systems*, 33: 10077–10087.

Huang, S.; Ling, R.; Hui, T.; Li, H.; Zhou, X.; Zhang, S.; Liu, S.; Hong, R.; and Wang, M. 2025. Revisiting Audio-Visual Segmentation with Vision-Centric Transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 8352–8361.

Juanola, X.; Haro, G.; and Fuentes, M. 2024. A Critical Assessment of Visual Sound Source Localization Models Including Negative Audio. *arXiv preprint arXiv:2410.01020*.

Kirillov, A.; Girshick, R.; He, K.; and Dollár, P. 2019. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6399–6408.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Li, J.; and Tian, Y. 2025. From waveforms to pixels: A survey on audio-visual segmentation. *arXiv preprint arXiv:2508.03724*.

Li, J.; Yu, S.; Wang, Y.; Wang, L.; and Lu, H. 2024a. Selm: Selective mechanism based audio-visual segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 3926–3935.

Li, K.; Yang, Z.; Chen, L.; Yang, Y.; and Xiao, J. 2023. Catr: Combinatorial-dependence audio-queried transformer for audio-visual video segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*, 1485–1494.

Li, X.; Wang, J.; Xu, X.; Peng, X.; Singh, R.; Lu, Y.; and Raj, B. 2024b. 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*, 3402–3413.

Liu, C.; Li, P. P.; Qi, X.; Zhang, H.; Li, L.; Wang, D.; and Yu, X. 2023a. Audio-visual segmentation by exploring cross-modal mutual semantics. In *Proceedings of the 31st ACM International Conference on Multimedia*, 7590–7598.

Liu, J.; Ju, C.; Ma, C.; Wang, Y.; Wang, Y.; and Zhang, Y. 2023b. Audio-aware query-enhanced transformer for audio-visual segmentation. *arXiv preprint arXiv:2307.13236*.

Liu, J.; Wang, Y.; Ju, C.; Ma, C.; Zhang, Y.; and Xie, W. 2024. Annotation-free audio-visual segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5604–5614.

Ma, J.; Sun, P.; Wang, Y.; and Hu, D. 2024. Stepping stones: A progressive training strategy for audio-visual semantic segmentation. *arXiv preprint arXiv:2407.11820*.

Mahmud, T.; Tian, Y.; and Marculescu, D. 2024. T-vsl: Text-guided visual sound source localization in mixtures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26742–26751.

Mo, S.; and Morgado, P. 2022. A closer look at weakly-supervised audio-visual source localization. *Advances in Neural Information Processing Systems*, 35: 37524–37536.

Mo, S.; and Tian, Y. 2023. Av-sam: Segment anything model meets audio-visual localization and segmentation. *arXiv preprint arXiv:2305.01836*.

Qian, R.; Hu, D.; Dinkel, H.; Wu, M.; Xu, N.; and Lin, W. 2020. Multiple sound sources localization from coarse to fine. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, 292–308. Springer.

Senocak, A.; Oh, T.-H.; Kim, J.; Yang, M.-H.; and Kweon, I. S. 2018. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4358–4366.

Small, D. M.; and Prescott, J. 2005. Odor/taste integration and the perception of flavor. *Experimental brain research*, 166: 345–357.

Sun, P.; Zhang, H.; and Hu, D. 2024. Unveiling and Mitigating Bias in Audio Visual Segmentation. *arXiv preprint arXiv:2407.16638*.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2022. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3): 415–424.

Wang, Y.; Liu, W.; Li, G.; Ding, J.; Hu, D.; and Li, X. 2024a. Prompting segmentation with sound is generalizable audio-visual source localizer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5669–5677.

Wang, Y.; Sun, P.; Li, Y.; Zhang, H.; and Hu, D. 2024b. Can Textual Semantics Mitigate Sounding Object Segmentation Preference? *arXiv preprint arXiv:2407.10947*.

Wang, Y.; Sun, P.; Zhou, D.; Li, G.; Zhang, H.; and Hu, D. 2024c. Ref-avs: Refer and segment objects in audio-visual scenes. *arXiv preprint arXiv:2407.10957*.

Yang, Q.; Nie, X.; Li, T.; Gao, P.; Guo, Y.; Zhen, C.; Yan, P.; and Xiang, S. 2024. 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*, 27134–27143.

Zhou, J.; Shen, X.; Wang, J.; Zhang, J.; Sun, W.; Zhang, J.; Birchfield, S.; Guo, D.; Kong, L.; Wang, M.; et al. 2024. Audio-visual segmentation with semantics. *International Journal of Computer Vision*, 1–21.

Zhou, J.; Wang, J.; Zhang, J.; Sun, W.; Zhang, J.; Birchfield, S.; Guo, D.; Kong, L.; Wang, M.; and Zhong, Y. 2022. Audio–visual segmentation. In *European Conference on Computer Vision*, 386–403. Springer.