Whence the Voice? Self-supervised Dual-source Audio-Visual Localisation via Selective Convergence

Han Hu*, Dongheng Lin*, Yuqi Hou, Haotian Li, Hyung Jin Chang, and Jianbo Jiao

The MIx Group, School of Computer Science, University of Birmingham, UK
Project page: https://happy-new-bears.github.io/scav-project-page/

Abstract. Localising multiple sound sources in visual scenes remains a fundamental challenge in multimodal perception due to an inherent circular dependency: separating mixed audio requires knowing source locations, while identifying sound-producing regions requires separated audio signals. In this paper, we focus on the dual-source setting and discover a selective convergence in self-supervised audio-visual learning: when presented with multiple sound sources, contrastive models naturally converge to the most salient audio-visual correspondence rather than attempting to represent all sources equally. This emergent phenomenon, analogous to human selective auditory attention, enables us to break the above circular dependency through a progressive two-stage framework: first, leveraging selective convergence to identify dominant sources, and then exploiting these learned priors to uncover remaining sources. Our self-supervised approach achieves the best performance among self-supervised methods on dual-source benchmarks without requiring any manual annotations, and even surpasses some weakly-supervised approaches on certain metrics. Furthermore, we identify a fundamental evaluation inconsistency in existing benchmarks: comparing continuous localisation heatmaps against bounding-box annotations creates systematic biases, particularly for non-axis-aligned objects where the bounding box includes substantial background regions. To address this, we introduce pixel-level segmentation masks to the existing benchmark, enabling spatially-aligned evaluation. Together, these results suggest that embracing rather than suppressing selectivity offers a scalable, annotation-free route to multi-source localisation.

Keywords: Audio-Visual Localisation · Selective · Self-supervised

"They rose expectant: eye and ear waited..."

Charlotte Brontë, Jane Eyre

* Equal contribution.

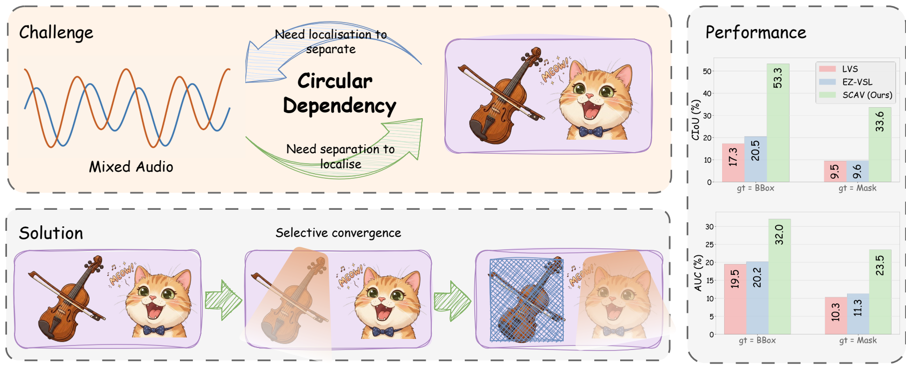

Fig. 1: Top-left: The fundamental paradox where visual localisation requires separated audio, while audio separation needs spatial visual information, creating an intractable circular dependency. **Bottom-left**: Our SCAV framework breaks this dependency by leveraging selective convergence. The model naturally focuses on one dominant source at first, then progressively reveals the other source. **Right**: Performance of the proposed SCAV surpasses other self-supervised methods across various metrics.

1 Introduction

Sound rarely occurs in isolation. In natural environments, multiple sources produce sounds simultaneously, creating rich acoustic scenes that humans navigate effortlessly by matching sounds to their visual origins. Such audio-visual correspondence forms the basis of how we understand and interact with the world. For machines, the challenge is concrete: *given a mixed audio signal and a visual scene with corresponding sounding objects, how can we determine which sounds come from which objects, when neither modality provides clear separation cues?*

Localising multiple simultaneous sound sources requires mixed-signal disentanglement in two different domains. In the audio domain, sounds mix through direct linear superposition; In contrast, the visual domain combines objects through spatial arrangement: different sound sources occupy distinct regions in an image. This domain asymmetry creates a challenging chicken-and-egg paradox: to decompose the additively mixed audio, we need spatial guidance from the visual domain to indicate where each source is located; yet to identify which spatial regions produce sound, we need separated audio signals to establish clear correspondences. This circular dependency makes the joint localisation task a severely ill-posed inverse problem: multiple valid decompositions exist for both the mixed audio and the visual scene, and without additional constraints, standard optimisation cannot determine which solution corresponds to the true source configuration.

While this circular dependency appears fundamentally intractable, human auditory perception offers an intriguing perspective. Humans effortlessly navigate cocktail parties through selective attention [4,5,36]: they focus on the most salient speaker first and then shift to others. This cognitive strategy suggests that perfect simultaneous separation may be neither necessary nor optimal. When we investigated whether neural networks exhibit similar behaviours in multi-source scenarios, extensive experiments with contrastive audio-visual learning revealed a remarkable phenomenon: Rather than struggle to represent all sources equally, networks consistently exhibit *selective convergence* and gravitate toward the most dominant audio-visual correspondence. We formalise this insight into a two-stage progressive framework that leverages selective convergence as a core mechanism for dual-source localisation. The first stage exploits this natural bias to identify dominant sources through contrastive learning, while the second stage uses these learned priors with targeted masking to progressively unmask subdominant sources. This approach effectively converts the ill-posed joint problem into a sequence of well-conditioned optimisations requiring no manual annotations.

To sum up, our contributions are threefold. **First**, we identify the *selective convergence* phenomenon: when learning from mixed audio, contrastive models naturally focus on the more dominant source rather than all sources equally. **Second**, we propose a two-stage progressive framework that leverages selective convergence to break the circular dependency problem. The first stage uses selective convergence to find dominant sources, while the second stage progressively unmasks remaining sources through targeted masking. **Third**, our approach not only achieves promising performance improvements on existing benchmarks, but the stable performance improvement brought by our method is also further validated under a newly introduced, more rigorous evaluation protocol.

2 Related Works

Audio-visual sounding source localisation aims to determine the spatial location of sounding objects within a video frame using the accompanying audio signal. The inherent co-occurrence between the auditory and visual modalities provides a powerful self-supervised signal for this task [1–3, 16, 29, 30, 37].

Early efforts in audio-visual sound source localisation primarily focused on only one sound source per scene [33, 34]. Arandjelovic and Zisserman [2] and Owens and Efros [29] employed mechanisms akin to Class Activation Map to measure the similarity between image regions and audio features. More recent approaches adopted contrastive learning frameworks [28] to improve discriminability at the region level. Chen *et al*. [6] introduced a contrastive learning framework with a differentiable thresholding mechanism to automatically mine hard negative image fragments. Mo and Morgado [25] proposed EZ-VSL, a multiple-instance contrastive learning strategy that aligns audio and visual spaces by maximising the similarity between the audio and the most responsive region in the corresponding image. Recent efforts have addressed practical challenges such as false negatives [35], off-screen sounds and background noise [9, 23], learning from silence [18], and feature decomposition through slot attention [20]. Park *et al*. [31] explored leveraging CLIP's pre-trained visual representations for sound source localisation, demonstrating that vision-language models can provide useful semantic priors for audio-visual correspondence. However, these methods fail when multiple sources sound simultaneously, which is the common case in real-world scenarios.

When multiple sound sources are present in a visual scene, separating their mixed audio requires knowing where each source is located, but finding these locations requires having the separated audio to establish correspondences, which creates a fundamental circular dependency. To break this dependency, researchers have explored different strategies. Mix-and-Localize [17, 27] proposed a self-supervised random walk on an image-sound graph to force source disentanglement through cycle consistency. Other methods introduce explicit semantic guidance. The self-supervised approach DSOL [15] tackles the paradox by generating a pseudo-class dictionary of visual representations via unsupervised clustering in a single-source setting. More recently, Kim *et al.* [19] introduced an iterative curriculum learning framework that progressively discovers multiple sound sources through object-aware contrastive learning and negative exclusion, though it requires multiple hand-tuned thresholds. The weakly-supervised method AVGN [26] directly uses true video-level class labels to guide learnable class tokens for explicit source grouping and disentanglement, while Dual Mean-Teacher [12] employs semi-supervised learning with dual teacher-student structures to generate high-quality pseudo-labels. More recent work introduces external modalities: T-VSL [24] leverages text representations from AudioCLIP's tri-modal embedding space to guide category-specific feature extraction, while OA-SSL [37] employs Multimodal Large Language Models to generate contextual descriptions that distinguish actively sounding objects from silent ones. In contrast to methods that engineer complex mechanisms, rely on weakly-supervised labels, or introduce external text modalities, our work leverages the emergent property of contrastive learning to progressively break the circular dependency in a fully self-supervised manner with only audio-visual data.

3 Method

3.1 Preliminaries

Given a video containing two simultaneously sounding sources, our goal is to localise both sound sources in the visual scene without manual annotations. Let $\mathcal{X} = (I, \mathbf{a}_{\text{mix}})$ denote an audio-visual pair, where the visual frame $I \in \mathbb{R}^{H\times W\times 3}$ captures a scene with two sound-producing objects, and the mixed audio spectrogram $\mathbf{A}_{\text{mix}} \in \mathbb{R}^{T\times F}$ contains T time steps and F frequency bins. Our objective is to produce two continuous-valued spatial localisation maps $\mathbf{M}_1, \mathbf{M}_2 \in [0,1]^{H\times W}$, where each map indicates the location of the corresponding sound source.

Localising sound sources requires computing audio-visual correspondences between $\mathbf{a}_i$ and I to obtain $\mathbf{M}_i$, yet we only observe the mixed audio $\mathbf{a}_{\text{mix}} = \sum_i \mathbf{a}_i$ where individual source contributions $\mathbf{a}_i$ are unknown. One might attempt to first separate $\mathbf{a}_i$ from $\mathbf{a}_{\text{mix}}$, but separation requires knowing $\mathbf{M}_i$ to assign frequency components to spatial locations. This creates a *chicken-egg paradox*: localisation needs separated audio, while separation needs spatial maps. Without additional constraints, this circular dependency makes the problem ill-posed.

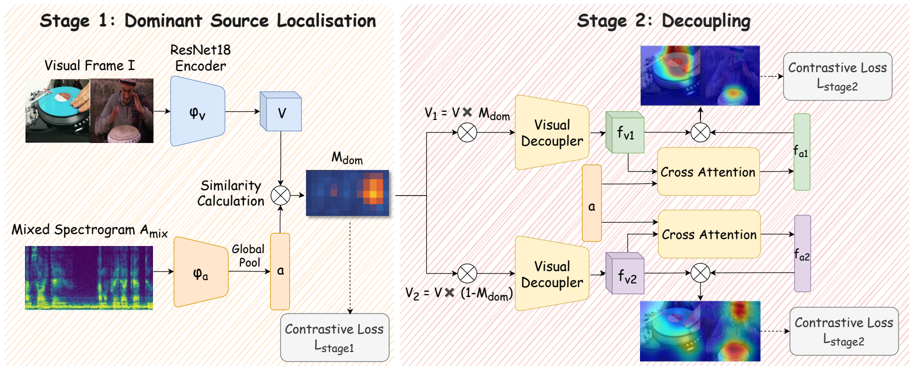

Fig. 2: Architecture of our Selective Convergence Audio-Visual localisation (SCAV) framework. Stage 1 (left): We extract visual features V using a frozen ResNet-18 encoder. The visual features are aligned with audio features $\mathbf{a}$ via contrastive learning to generate a coarse localisation heatmap $\mathbf{M}_{\text{dom}}$ that identifies the more dominant sounding source. **Stage 2 (right)**: Using the Stage 1 heatmap as a spatial prior, visual and audio decouplers progressively separate the mixed features into source-specific representations that yield refined localisation maps for both sources.

While this circular dependency appears intractable, we observe that contrastive learning naturally exhibits *selective convergence*: rather than representing all sources equally, models gravitate toward a single, most salient audio-visual correspondence. We propose the Selective Convergence Model for Audio-Visual localisation (SCAV), a two-stage framework that leverages this phenomenon. Through contrastive learning, the first stage identifies the source that the model is biased toward, which we term the *dominant source* $\mathbf{M}_{\text{dom}}$; the other is the *subdominant source* $\mathbf{M}_{\text{sub}}$. This dominance is determined by the model's bias, instead of any predefined property of the audio or visual input. Importantly, our framework relies only on the fact that selective convergence isolates one source as the spatial prior; it is agnostic to which of the two sources becomes dominant. The second stage uses $\mathbf{M}_{\text{dom}}$ as the spatial prior to progressively reveal the subdominant source $\mathbf{M}_{\text{sub}}$. This approach effectively converts the ill-posed joint problem into a sequence of well-conditioned optimisations in a fully self-supervised manner.

3.2 Dominant Source Localisation via Selective Convergence

We first extract features from different modalities:

$$V = \Phi_v(I) \in \mathbb{R}^{D\times h\times w}, \quad \mathbf{a} = \text{GAP}(\Phi_a(\mathbf{A}_{\text{mix}})) \in \mathbb{R}^D \tag{1}$$

where $h = H/32$, $w = W/32$, and GAP denotes global average pooling. During training, we initialise the visual encoder Φ_v with ImageNet pretrained weights and keep it frozen to leverage robust visual representations, while the audio encoder Φ_a is trained from scratch to learn audio-visual correspondence

For audio features $\mathbf{a}_i$ from sample i and visual features V_j from sample j, we calculate the dot product between $\mathbf{a}_i$ and the feature vector at each spatial patch location (h, w) of V_j:

$$\mathcal{S}_{i \to j} = \langle V_j, \mathbf{a}_i \rangle \in \mathbb{R}^{h \times w} \tag{2}$$

where each element $\mathcal{S}_{i \to j}^{(h,w)}$ represents the similarity score between the audio and the visual features at position (h, w).

To train the encoders in a self-supervised manner, we adopt the contrastive learning framework from [6] with differentiable thresholding to automatically identify positive and negative spatial regions. The soft masks $\hat{m}_p$ and $\hat{m}_n$ are computed via differentiable thresholding:

$$\hat{m}_p = \sigma((\mathcal{S}_{i \to i} - \epsilon_p)/\tau), \quad \hat{m}_n = 1 - \sigma((\mathcal{S}_{i \to i} - \epsilon_n)/\tau) \tag{3}$$

with $\epsilon_p = 0.65$, $\epsilon_n = 0.4$, temperature $\tau = 0.03$ and σ means Sigmoid function.

These masks partition spatial locations into positive pairs (regions with strong audio-visual correspondence, $\hat{m}_p$) and negative pairs (regions with weak correspondence, $\hat{m}_n$). The positive and negative scores are then computed as:

$$P_i = \frac{1}{|\hat{m}_p|} \langle \hat{m}_p, \mathcal{S}_{i \to i} \rangle, \tag{4}$$

$$N_i = \frac{1}{|\hat{m}_n|} \langle \hat{m}_n, \mathcal{S}_{i \to i} \rangle + \frac{1}{hw} \sum_{j \neq i} \langle \mathbf{1}, \mathcal{S}_{i \to j} \rangle \tag{5}$$

where P_i aggregates similarity over positive regions (high audio-visual correspondence) and N_i aggregates over negative regions (low correspondence within sample i, plus cross-sample negatives from $j \neq i$).

The contrastive loss is:

$$\mathcal{L}_{\text{stage1}} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(P_i)}{\exp(P_i) + \exp(N_i)}. \tag{6}$$

Training with such contrastive loss on mixed audio leads the model to exhibit *selective convergence*: the learned similarity map $\mathcal{S}_i = \langle V_i, \mathbf{a}_i \rangle \in \mathbb{R}^{h \times w}$ naturally concentrates on the dominant source with the most obvious audio-visual correspondence. This arises from the simplicity bias of gradient-based optimisation [38].

To extract a clean spatial prior $\mathbf{M}_{\text{dom}}$ from this similarity map, we further apply a simple post-processing step (detailed in supplementary Algorithm 1) .

3.3 Progressive Multi-source Localisation with Spatial Priors

Having identified the dominant source $\mathbf{M}_{\text{dom}}$ in Stage 1, we now leverage it as a spatial prior to break the circular dependency. The key insight is simple: with $\mathbf{M}_{\text{dom}}$ telling us where the dominant source is located, we can (1) partition visual features into source-specific regions, (2) use these partitioned visual features to guide audio separation via cross-attention, and (3) localise each source independently with the decoupled audio-visual pairs.

Feature extraction and projection. We reuse the encoders from the first stage to extract visual features $V \in \mathbb{R}^{512\times h\times w}$ and audio features $\mathbf{a} \in \mathbb{R}^{512}$. To enable effective decoupling, we project these features into a unified embedding space and add 2D positional encoding to preserve spatial structure:

$$\tilde{V} = \mathrm{Proj}_v(V) + \mathrm{PE}_{2D} \in \mathbb{R}^{D\times h\times w}, \tag{7}$$

$$\tilde{\mathbf{a}} = \mathrm{Proj}_a(\mathbf{a}) \in \mathbb{R}^{D} \tag{8}$$

where Proj_v and Proj_a are learnable linear projection layers, and PE_{2D} denotes 2D sinusoidal positional encoding.

Visual feature decoupling. With the spatial prior $\mathbf{M}_{\mathrm{dom}}$, we partition the visual features into two complementary regions:

$$V_1 = \tilde{V} \odot \mathbf{M}_{\mathrm{dom}}, \quad V_2 = \tilde{V} \odot (1 - \mathbf{M}_{\mathrm{dom}}) \tag{9}$$

where V_1 captures the dominant source region and V_2 captures the complementary region (other possible source region and the background).

A visual decoupler Φ_v^{dec} then refines these coarsely separated features to produce more discriminative source-specific visual features:

$$f_{v1}, f_{v2} = \Phi_v^{\mathrm{dec}}(V_1, V_2), \quad f_{vi} \in \mathbb{R}^{D\times h\times w}. \tag{10}$$

Audio feature decoupling. With decoupled visual features (f_{v1}, f_{v2}) ready, we can now separate the mixed audio feature $\tilde{\mathbf{a}}$ into source-specific representations via the audio decoupler Φ_a^{dec}, shown as the cross-attention branch in Stage 2 (right) of Fig. 2. The key idea is to use each visual feature as a spatial guide: the audio decoupler Φ_a^{dec} employs cross-attention where the mixed audio serves as the query, and each visual feature provides keys and values:

$$f_{a1}, f_{a2} = \Phi_a^{\mathrm{dec}}(\tilde{\mathbf{a}}, f_{v1}, f_{v2}) \in \mathbb{R}^{D}. \tag{11}$$

This allows each audio feature f_{ai} to attend to its corresponding visual region, effectively extracting the audio component associated with that spatial location.

Source-specific contrastive learning. After having decoupled both modalities into (f_{v1}, f_{v2}) and (f_{a1}, f_{a2}), we train each source pair independently using contrastive learning. For each source $i \in \{1, 2\}$, we compute the similarity map $\mathcal{S}_i = \langle f_{vi}, f_{ai}\rangle$ and optimie:

$$\mathcal{L}_i = -\log \frac{\exp(P_i)}{\exp(P_i) + \exp(N_i')} \tag{12}$$

where P_i and N_i' are positive and negative scores computed via differentiable thresholding (Eq. (3)). Crucially, unlike Stage 1, we exclude cross-sample negative terms in N_i' because the decoupling task focuses on within-sample source assignment rather than cross-sample discrimination. The total loss is:

$$\mathcal{L}_{\mathrm{Stage2}} = \mathcal{L}_1 + \mathcal{L}_2. \tag{13}$$

Inference stage. At inference, the trained model produces decoupled audio-visual pairs (f_{v1}, f_{a1}) and (f_{v2}, f_{a2}). The final localisation maps are obtained by computing similarity and upsampling:

$$\mathbf{M}_k = \mathrm{Upsample}(\langle f_{vk}, f_{ak} \rangle), \quad k \in \{1, 2\} \tag{14}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product.

4 Experiments

4.1 Experimental Setup

Implementation details. For visual inputs I, each image is resized to 224×224 resolution. For audio inputs, signals are sampled at 22.05 kHz and converted to log-spectrograms $\mathbf{A}_{\mathrm{mix}}$ using Short-Time Fourier Transform with 50ms windows and 25ms hop size for 3-second clips. The visual encoder Φ_v employs separate ResNet18 [13] networks initialised with ImageNet pre-trained weights. The audio encoder Φ_a uses another ResNet18 architecture with the first convolutional layer adapted to a single channel. The visual decoupler Φ_v^{dec} employs a 4-layer Transformer encoder with 8 attention heads, while the audio decoupler Φ_a^{dec} uses cross-attention mechanisms to produce separated audio representations.

Training protocol. We employ a two-stage progressive training strategy. In the first stage, we train encoders (Φ_v, Φ_a) using $\mathcal{L}_{\mathrm{stage1}}$ to learn dominant source localisation through selective convergence. In the second stage, we train the full architecture using $\mathcal{L}_{\mathrm{stage2}}$, with spatial priors $\mathbf{M}_{\mathrm{dom}}$ provided by the trained first-stage model. We train our model using the Adam optimizer [21] with an initial learning rate of 10^{-4} and a batch size of 256.

Note that the two-stage design refers to two *training* stages only; at inference time, the trained model executes a single forward pass without interruption, achieving real-time performance at 43.2 FPS on a single NVIDIA A100 GPU.

4.2 Evaluation Metrics

Following prior works [15,17], we evaluate dual-source localisation using CIoU@X, AUC, and CAP. Note that for synthetic dual-source benchmarks (*e.g.* VGGSound-Duet), we evaluate each predicted heatmap over the *full* concatenated image domain against zero-padded ground truth masks without any cropping on the predicted heatmap. This ensures that cross-source activations are properly penalised (see supplementary material for more details).

(1) CAP (Class-Aware Precision) measures the percentage of correctly localised sources at the class level:

$$CAP = \frac{\sum_{k=1}^{K} \delta_k \mathrm{AP}_k}{\sum_{k=1}^{K} \delta_k}. \tag{15}$$

(2) CIoU (Class-Aware IoU) is the intersection-over-union between predicted heatmaps and ground truth:

$$CIoU = \frac{\sum_{k=1}^{K} \delta_k \, IoU_k}{\sum_{k=1}^{K} \delta_k} \tag{16}$$

where IoU_k is calculated based on the predicted sounding object area and the annotated bounding box for the k-th class.

(3) AUC is the area under the precision-recall curve:

$$\mathrm{AUC} = \int_0^1 R(t)\,\mathrm{d}t, \quad \text{where} \quad R(t) = \frac{\sum_{k=1}^{K} \delta_k \cdot \mathbb{1}[\mathrm{IoU}_k \geq t]}{\sum_{k=1}^{K} \delta_k}. \tag{17}$$

4.3 Evaluation on Existing Benchmarks

We first evaluate our approach on three established multi-source localisation benchmarks, *i.e.* MUSIC-Duet, VGGSound-Instruments, and VGGSound-Duet:

MUSIC-Duet. The dataset originally contains 149 untrimmed duet videos from the MUSIC dataset [39], the accessible set comprises 141 videos that cover 11 classes of musical instruments due to the YouTube availability issue. The dataset provides bounding-box annotations for dual sources generated using the Faster R-CNN detector by Hu *et al.* [15]. Following Hu *et al.* [15], the first two videos per instrument category form the test set, and the rest are for training.

As shown in Tab. 1, our self-supervised SCAV achieves the best localisation quality (CIoU/AUC) among self-supervised methods. However, SCAV exhibits weaker detection performance (CAP) compared to AVGN [26].

VGGSound-Instruments. Hu *et al.* [17] filtered 37 musical instrument categories from VGGSound [7] with 32k training videos. For multi-source evaluation, Hu *et al.* [17] annotated segmentation masks for 446 high-quality frames. The benchmark synthesises multi-source scenarios by randomly concatenating two frames into 224×448 images while summing their audio waveforms. As shown in Tab. 1, the proposed SCAV method achieves the best performance among self-supervised methods on most metrics (CAP/AUC) and a competitive CIoU@0.1.

VGGSound-Duet. To assess scalability, we evaluate on VGGSound-Duet [26], the largest multi-source benchmark with 5,158 test videos spanning 220 diverse categories beyond musical instruments. Like VGGSound-Instruments, it synthesises dual-source scenarios but uses bounding box annotations from the VGGSound-Source test set [6].

SCAV achieves the best performance among self-supervised methods across all metrics in Tab. 1 and Tab. 2, even surpassing some weakly-supervised methods on certain metrics. These consistent gains on the larger-scale dataset suggest that our approach particularly benefits from increased data diversity and scale.

Table 1: Quantitative results on MUSIC-Duet (BBox) and VGGSound-Duet (BBox).

Method	Self-Sup.	MUSIC-Duet (BBox)			VGGSound-Duet (BBox)		
		CIoU@0.3(%)	AUC(%)	CAP(%)	CIoU@0.3(%)	AUC(%)	CAP(%)
CoarsetoFine (ECCV'2020) [32]	✗	17.6	20.6	-	14.7	18.5	-
AVGN (CVPR'2023) [26]	✗	32.5	24.6	50.6	26.2	23.8	21.9
OA-SSL (CVPR'2025) [37]	✗	45.9	36.1	64.1	55.2	44.8	45.9
Attention10k (CVPR'2018) [33]	✓	21.6	19.6	-	11.5	15.2	-
OTS (ECCV'2018) [2]	✓	13.3	18.5	11.6	12.2	15.8	10.7
DMC (CVPR'2019) [14]	✓	17.5	21.1	-	13.8	17.1	-
DSOL (NeurIPS'2020) [15]	✓	<u>30.1</u>	<u>22.3</u>	-	<u>22.3</u>	<u>21.1</u>	-
LVS (CVPR'2021) [6]	✓	22.5	20.9	-	17.3	19.5	-
EZ-VSL (ECCV'2022) [25]	✓	24.3	21.3	-	20.5	20.2	-
Mix-and-Localize (CVPR'2022) [17]*	✓	26.5	21.5	<u>47.5</u>	21.1	20.5	16.3
NoPrior (CVPR'2024) [19]	✓	2.5†	13.2	21.7	18.1	16.8	<u>26.8</u>
Slot-Attn (CVPR'2025) [20]	✓	9.9	12.7	19.6	15.5	16.2	25.0
SCAV (Ours)	✓	**47.1**	**28.6**	**47.6**	**53.3**	**32.0**	**53.0**

*Mix-and-Localise is re-run under our frame-wise evaluation for VGGSound-Duet; OA-SSL and AVGN are taken as originally reported.
†Please see the supplementary material for detailed implementation and analysis.

Table 2: Quantitative comparison on VGGSound-Instruments (Mask) and VGGSound-Duet$^{\text{Mask}}$ with segmentation-mask-based evaluation.

Dataset	Method	Self-Sup.	CAP(%)	CIoU@0.1(%)	CIoU@0.3(%)	AUC(%)
VGGSound-Instruments	CoarsetoFine (ECCV'2020) [32]	✗	-	54.2	-	12.9
	AVGN (CVPR'2023) [26]	✗	27.3	77.5	-	18.2
	Attention10k (CVPR'2018) [33]	✓	-	52.3	-	11.7
	OTS (ECCV'2018) [2]	✓	<u>23.3</u>	51.2	-	11.2
	DMC (CVPR'2019) [14]	✓	-	53.7	-	12.5
	DSOL (NeurIPS'2020) [15]	✓	-	<u>74.3</u>	-	<u>15.9</u>
	LVS (CVPR'2021) [6]	✓	-	57.3	-	13.3
	EZ-VSL (ECCV'2022) [25]	✓	-	60.2	-	14.2
	Mix-and-Localize (CVPR'2022) [17]	✓	21.5	73.2	-	15.6
	SCAV (Ours)	✓	**32.0**	**76.0**	-	**21.1**
VGGSound-Duet$^{\text{Mask}}$	AVGN (CVPR'2023) [26]	✗	33.61	55.43	18.44	17.39
	LVS (CVPR'2021) [6]	✓	18.00	31.51	9.53	10.28
	EZ-VSL (ECCV'2022) [25]	✓	16.82	37.97	9.58	11.35
	NoPrior (CVPR'2024) [19]	✓	20.26	37.02	5.77	10.25
	Slot-Attn (CVPR'2025) [20]	✓	15.21	30.87	3.63	8.92
	SCAV (Ours)	✓	**39.30**	**69.42**	**33.59**	**23.52**

Important observation. During analysis of the VGGSound-Duet results, we discovered a fundamental flaw in the evaluation of the bounding box that may have been overlooked by the community. Bounding boxes treat all enclosed pixels equally: both object and background receive the same weight in IoU computation. This evaluation protocol essentially rewards over-activation rather than precise localisation. Fig. 3 vividly demonstrates this issue with several representative examples. In the *playing flute* case, while our model precisely localises the sound-producing region around the instrument, the ground truth bounding box includes substantial empty corners due to the diagonal instrument orientation. Similarly, for *turkey gobbling*, our heatmap identifies the turkey as the sound source, yet the bounding box encompasses large background areas. These examples reveal a possible misalignment: current box-based metrics paradoxically penalise precise localisation while rewarding models that indiscriminately activate background regions within bounding boxes.

Recent work in object detection [8] identified a similar misalignment of the evaluation and demonstrated that segmentation masks provide a more faithful

Fig. 3: Bounding box evaluation can be misleading. Shown by four representative examples. Top: Ground-truth bounding boxes encompass substantial background regions beyond the actual sound sources. Bottom: Our predicted localisation heatmaps precisely identify sounding regions.

assessment by precisely delineating object boundaries. Although VGGSound-Instruments [17] introduced segmentation masks, its limited scale (446 test samples) and restriction to musical instruments prevent comprehensive evaluation across diverse acoustic scenarios. However, large-scale pixel-precise benchmarks remain scarce, as segmentation annotation was prohibitively expensive before recent advances like SAM [22]. This observation motivates us to establish a rigorous evaluation benchmark for the audio-visual localisation task using segmentation masks at scale, which could enable accurate measurement of what models truly localise across diverse sound categories.

4.4 A New Evaluation Protocol

VGGSound-Duet^{Mask}. To address the evaluation limitations identified above, we introduce VGGSound-Duet$^{\mathrm{Mask}}$, a pixel-precise version of the VGGSound-Duet benchmark [26]. We start from the VGGSound-Source (VGG-SS) test set [6], which originally contains 5,158 single-source videos with bounding box annotations. Due to YouTube's availability, 4,390 videos remain accessible. For each video, we generate high-quality segmentation masks using SAM [22], with the original bounding boxes as initial prompts. To construct the multi-source evaluation, we follow the exact pairing protocol of VGGSound-Duet [26], yielding 3,951 dual-source test pairs by concatenating frames and mixing audio from two videos. Unlike VGGSound-Instruments, which offers only 446 mask-annotated samples restricted to musical categories, our VGGSound-Duet$^{\mathrm{Mask}}$ benchmark provides nearly 10× more samples across 220 diverse sound categories. Tab. 3 summarises existing multi-source localisation benchmarks; our VGGSound-Duet$^{\mathrm{Mask}}$ provides the largest mask-annotated test set. We detail the construction of this benchmark in the supplementary material.

A related task that also produces pixel-level masks is audio-visual segmentation (AVS) [10, 11, 40, 41]. AVS targets a different question from ours: they segment the sounding objects in a scene as a whole, either as a single foreground mask in AVSBench [41], by semantic category in AVSS [40], or as separate instances in AVISeg [11], and the audio serves only as a cue for whether an object is sounding. Our VGGSound-Duet$^{\mathrm{Mask}}$ instead evaluates the decomposition of a

Table 3: Summary of Multi-Source Sound Source Localisation Benchmarks.

Dataset	Year	Test Samples	Annotation Type	#Classes
MUSIC-Duet [15]	2018	17[a]	Bounding Box	11
MUSIC-Synthetic [15]	2020	455	Bounding Box	15
VGGSound-Instruments [17]	2022	446	Segmentation Mask	37
VGGSound-Duet [26]	2023	~~5,158~~ 3,951[b]	Bounding Box	220
VGGSound-Duet$^{\mathbf{Mask}}$ (Ours)	**2025**	**3,951**	**Segmentation Mask**	**220**

[a] MUSIC-Duet reports video-level counts, where each video contains multiple annotated frames, while all other benchmarks count individual frames.
[b] Among the videos released, 3,951 samples are currently (as of Oct. 2025) available; the remaining videos are unavailable on YouTube due to copyright issues or removal by uploaders.

two-source audio mixture into a separate localisation for each source, which no existing AVS benchmark provides.

Quantitative results. Tab. 2 presents the results under our pixel-precise evaluation protocol. Transitioning from bounding boxes to segmentation masks changes the evaluation dynamics: models can no longer benefit from activating irrelevant background pixels within rectangular regions. SCAV's consistent improvement under mask-based evaluation confirms genuine, pixel-level localisation of sound-producing regions rather than reliance on coarse spatial heuristics.

Qualitative results. Fig. 4 presents the localisation maps generated by different methods on representative multi-source scenarios. The baseline models LVS [6] and EZ-VSL [25] struggle to localise both sources simultaneously. As shown in the final two columns, our method generates two independent localisation maps that accurately identify each individual sound source.

Interestingly, in the third row, where multiple birds are present, our method produces connected regions that appropriately cover all birds when localising bird sounds. This demonstrates its ability to handle spatially distributed sources of the same category. We also observe that each localisation map exhibits a primary activation pattern: while one source receives strong activation, the other source's region may show weaker residual activations. This soft separation reflects our progressive decoupling design: our method focuses on each source but maintains contextual awareness rather than enforcing hard boundaries. This helps when sources have overlapping characteristics or ambiguous boundaries, as the model handles uncertainty naturally. This qualitative evidence, corroborating our quantitative results, demonstrates that our SCAV method successfully breaks the circular dependency problem and allows each decoupled branch to focus on distinct acoustic sources without interference. More qualitative comparisons and failure cases are provided in the supplementary material.

4.5 Empirical Analysis of Selective Convergence

To validate selective convergence in the first stage, we conduct an empirical analysis that complements the theoretical justification provided in the supplementary material. Specifically, when trained with contrastive learning on mixed

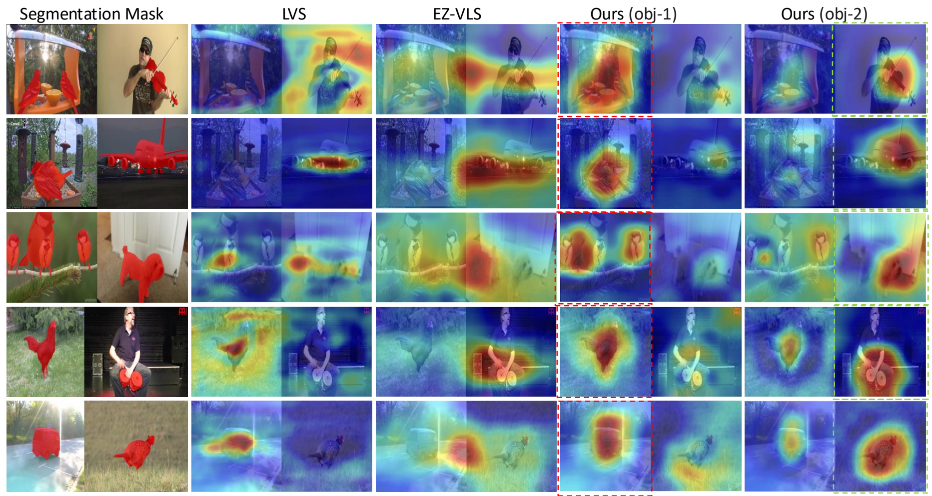

Fig. 4: Qualitative comparison of multi-source sound localisation. For each test sample containing mixed audio from two sources, from left to right, we show: **First**, input frame with segmentation masks indicating both sound sources. **Second and Third**, localisation heatmaps from baseline models LVS [6] and EZ-VSL [25]. **Fourth and Last**, localisation heatmaps for each detected source from our SCAV method. The heatmaps visualise predicted sound source locations, where warmer colours indicate higher activation values. Best viewed in colour.

audio, the first-stage model naturally focuses on one dominant source rather than attempting to cover all sources equally. To this end, we design a controlled experiment that directly measures whether the model's predictions concentrate on a single source or spread across multiple sources.

We evaluate the first-stage model on VGGSound-Duet, where each test sample is constructed by concatenating two frames into a 224×448 image while mixing their corresponding audio waveforms. For each sample, we have two ground truth masks GT_1 and GT_2, where each mask covers one source region (either the left or right half of the concatenated image) and is zero-padded on the other half. The first-stage model takes the concatenated image and mixed audio as input, and produces a single heatmap $\mathbf{M}_{\text{dom}}$ over the full 224×448 image domain. We compute the IoU between the predicted heatmap $\mathbf{M}_{\text{dom}}$ and each ground truth:

$$\mathrm{IoU}_1 = \mathrm{IoU}(\mathbf{M}_{\text{dom}}, \mathrm{GT}_1), \quad \mathrm{IoU}_2 = \mathrm{IoU}(\mathbf{M}_{\text{dom}}, \mathrm{GT}_2). \tag{18}$$

For each sample, we identify the dominant source as the one with a higher IoU: $i^* = \arg\max(\mathrm{IoU}_1, \mathrm{IoU}_2)$ and denote the other source as j^*. We then compute metrics separately for the dominant source i^* and the second source j^* across all test samples.

Tab. 4 reveals a substantial performance gap between the dominant source and the second source, which provides direct empirical evidence of selective convergence: the model's heatmap predominantly focuses on one source rather than

Table 4: Localisation performance comparison between one dominant source and across dual sources. It shows that the first-stage model tends to selectively converge on one of the sounding source objects.

Setting	CAP (%)	CIoU@0.3 (%)	AUC (%)
Dominant Source	63.39	61.70	37.82
Second Source	13.91 (-49.48)	3.62 (-58.08)	8.12 (-29.70)

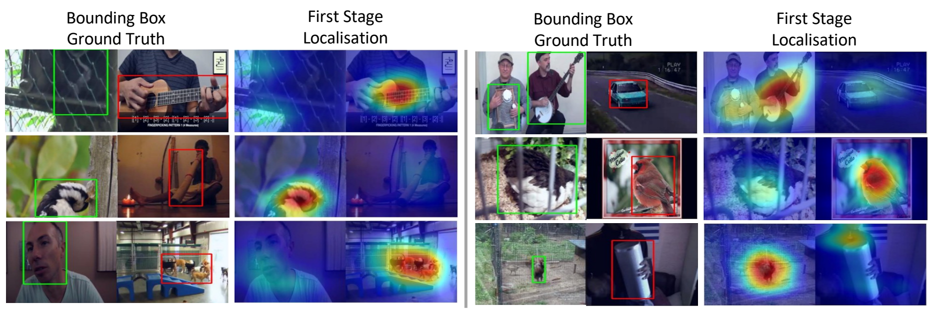

Fig. 5: Examples of selective convergence behaviour of the first stage. We show several input sample frames with both ground truth sources (the first and third columns) and the first stage output heatmap (the second and fourth columns). Green boxes indicate the sounding object in the left frame, while red boxes indicate the sounding object in the right frame. The heatmaps consistently show strong activation on only one source while producing weak or negligible responses on the other.

distributing attention across both. Fig. 5 further visualises this behaviour across diverse acoustic scenarios. Both quantitative and qualitative results indicate that the first stage produces focused, dominant-source localisations rather than diffused activations that attempt to cover multiple sources.

4.6 Ablation Study and Discussion

We conduct ablation studies on the VGGSound-Duet dataset to analyse the contribution of each component in our SCAV framework, as shown in Tab. 5. The ablation results reveal the interdependencies among components. Comparing Settings (a) and (b) shows that the decoupling architecture relies heavily on spatially informative guidance to function effectively. Setting (c) uses the same architecture as our full model but differs only in training strategy: joint end-to-end training versus progressive training. Despite having identical components, joint training achieves only 41.84% CIoU@0.3 compared to our 53.27%. This demonstrates that the training strategy plays a crucial role in our framework. In joint training, the first-stage encoder has not yet converged during early training. This leads to unstable spatial priors that mislead the second-stage decoupler. Progressive training addresses this by first training the first stage to convergence before training the second stage, which yields more stable

Table 5: Ablation study on VGGSound-Duet.

Setting #	Components				VGGSound-Duet (BBox-based)				VGGSound-Duet$^{\text{Mask}}$			
	Stage 2	Spatial Prior	Progr. Train	Cross-neg.	CAP (%)	CIoU@0.1(%)	CIoU@0.3(%)	AUC (%)	CAP	CIoU@0.1	CIoU@0.3	AUC
(a)	✓	✗*	✓	✗	37.23	73.38	28.94	21.94	27.39	52.58	16.15	15.47
(b)	✗	—	—	✗	38.65	75.15	32.66	22.97	28.93	53.41	18.84	16.27
(c)	✓	✓	✗	✗	45.17	79.41	41.84	26.88	32.71	63.33	25.62	19.68
(d)	✓	✓	✓	✓	35.68	67.38	26.69	20.93	26.00	50.92	17.40	15.75
Ours	✓	✓	✓	✗	**52.96**	**83.17**	**53.27**	**31.96**	**39.30**	**69.42**	**33.59**	**23.52**

*Uses uniform mask instead of learned spatial prior.

spatial priors. Setting (d) adds cross-sample negatives on top of our full model and degrades performance, since Stage 2 is essentially a within-sample source-assignment problem where the useful contrast lies between the dominant source and the other source inside the same mixture. Such cross-sample negatives instead dilute this within-sample alignment.

5 Conclusion

In this work, we introduced selective convergence for dual-source audio-visual localisation, an emergent property of contrastive audio-visual learning where models naturally focus on dominant sources rather than attempting a joint multi-source representation. By leveraging selective convergence, we transformed the circular dependency problem inherent in multi-source localisation into a tractable sequence of conditional optimisations. The proposed SCAV framework achieved the best performance among self-supervised methods in the literature, and competitive performance compared to weakly-supervised approaches. Notably, the performance gap over baselines widens on larger-scale datasets, which indicates that our method effectively embraces the richness of large-scale audio-visual data.

Our empirical analysis also revealed systematic biases in current bounding box evaluation protocols, following which we introduced pixel-precise benchmarks that better reflect actual localisation quality. Future work could explore whether selective convergence generalises to other multi-modal learning scenarios beyond audio-visual correspondence, potentially for handling ambiguous many-to-many associations in a self-supervised manner.

Acknowledgements

This project is partially supported by an Amazon Research Award. The computations in this research were partially performed using the Baskerville Tier 2 HPC service. Baskerville was funded by the EPSRC and UKRI through the World Class Labs scheme (EP\T022221\1) and the Digital Research Infrastructure programme (EP\W032244\1) and is operated by Advanced Research Computing at the University of Birmingham.

References

1. Arandjelovic, R., Zisserman, A.: Look, listen and learn. In: Proceedings of the IEEE international conference on computer vision. pp. 609–617 (2017)
2. Arandjelovic, R., Zisserman, A.: Objects that sound. In: Proceedings of the European conference on computer vision (ECCV). pp. 435–451 (2018)
3. Aytar, Y., Vondrick, C., Torralba, A.: Soundnet: Learning sound representations from unlabeled video. Advances in neural information processing systems **29** (2016)
4. Brattico, P., Brattico, E., Vuust, P.: Global sensory qualities and aesthetic experience in music. Frontiers in Neuroscience **11**, 159 (2017)
5. Bronkhorst, A.W.: The cocktail-party problem revisited: early processing and selection of multi-talker speech. Attention, Perception, & Psychophysics **77**(5), 1465–1487 (2015)
6. Chen, H., Xie, W., Afouras, T., Nagrani, A., Vedaldi, A., Zisserman, A.: Localizing visual sounds the hard way. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16867–16876 (2021)
7. Chen, H., Xie, W., Vedaldi, A., Zisserman, A.: Vggsound: A large-scale audio-visual dataset. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 721–725. IEEE (2020)
8. Choe, J., Oh, S.J., Lee, S., Chun, S., Akata, Z., Shim, H.: Evaluating weakly supervised object localization methods right. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
9. Choi, H., Lee, J., Kwak, N.: What's making that sound right now? video-centric audio-visual localization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2025)
10. 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 (AAAI). vol. 38, pp. 12155–12163 (2024)
11. Guo, R., Ying, X., Chen, Y., Niu, D., Li, G., Qu, L., Qi, Y., Zhou, J., Xing, B., Yue, W., Shi, J., Wang, Q., Zhang, P., Liang, B.: Audio-visual instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
12. Guo, Y., Ma, S., Su, H., Wang, Z., Zhao, Y., Zou, W., Sun, S., Zheng, Y.: Dual mean-teacher: An unbiased semi-supervised framework for audio-visual source localization. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 36, pp. 1–14 (2023)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
14. Hu, D., Nie, F., Li, X.: Deep multimodal clustering for unsupervised audiovisual learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9248–9257 (2019)
15. Hu, D., Qian, R., Jiang, M., Tan, X., Wen, S., Ding, E., Lin, W., Dou, D.: Discriminative sounding objects localization via self-supervised audiovisual matching. Advances in Neural Information Processing Systems **33**, 10077–10087 (2020)
16. Hu, H., Lin, D., Huang, Q., Hou, Y., Chang, H.J., Jiao, J.: Audio-visual separation with hierarchical fusion and representation alignment. In: British Machine Vision Conference (BMVC) (2025)
17. Hu, X., Chen, Z., Owens, A.: Mix and localize: Localizing sound sources in mixtures. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)

18. Juanola, X., Morais, G., Fuentes, M., Haro, G.: Learning from silence and noise for visual sound source localization. In: British Machine Vision Conference (BMVC) (2025)
19. Kim, D., Um, S.J., Lee, S., Kim, J.U.: Learning to visually localize sound sources from mixtures without prior source knowledge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26467–26476 (2024)
20. Kim, J., Lee, J., Seo, S.e., Yang, J., Kim, J.U.: Improving sound source localization with joint slot attention on image and audio. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 26477–26486 (2025)
21. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
22. 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)
23. Liu, X., Qian, R., Zhou, H., Hu, D., Lin, W., Liu, Z., Zhou, B., Zhou, X.: Visual sound localization in the wild by cross-modal interference erasing. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 1801–1809 (2022)
24. Mahmud, T., Li, Y., Tian, Y.: T-vsl: Text-guided visual sound source localization in mixtures. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 23456–23465 (2024)
25. Mo, S., Morgado, P.: Localizing visual sounds the easy way. In: European Conference on Computer Vision. pp. 218–234. Springer (2022)
26. Mo, S., Tian, Y.: Audio-visual grouping network for sound localization from mixtures. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10565–10574 (2023)
27. Mo, S., Wang, H.: Multi-scale multi-instance visual sound localization and segmentation. arXiv preprint arXiv:2409.00486 (2024)
28. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
29. Owens, A., Efros, A.A.: Audio-visual scene analysis with self-supervised multisensory features. In: Proceedings of the European conference on computer vision (ECCV). pp. 631–648 (2018)
30. Owens, A., Wu, J., McDermott, J.H., Freeman, W.T., Torralba, A.: Ambient sound provides supervision for visual learning. In: European conference on computer vision. pp. 801–816. Springer (2016)
31. Park, S., Senocak, A., Chung, J.S.: Can clip help sound source localization? In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 5699–5708 (2024)
32. Qian, R., Hu, D., Dinkel, H., Wu, M., Xu, N., Lin, W.: Multiple sound sources localization from coarse to fine. In: European Conference on Computer Vision. pp. 292–308. Springer (2020)
33. Senocak, A., Oh, T.H., Kim, J., Yang, M.H., Kweon, I.S.: Learning to localize sound source in visual scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4358–4366 (2018)
34. Senocak, A., Oh, T.H., Kim, J., Yang, M.H., Kweon, I.S.: Learning to localize sound sources in visual scenes: Analysis and applications. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(8), 1984–1997 (2020)
35. Sun, W., Zhang, J., Wang, J., Liu, Z., Zhong, Y., Feng, T., Guo, Y., Zhang, Y., Barnes, N.: Learning audio-visual source localization via false negative aware con-

trastive learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6420–6429 (2023)
36. Sussman, E.S.: Auditory scene analysis: An attention perspective. Journal of Speech, Language, and Hearing Research **60**(10), 2989–3000 (2017)
37. Um, S.J., Kim, D., Lee, S., Kim, J.U.: Object-aware sound source localization via audio-visual scene understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8342–8351 (2025)
38. Xue, Y., Joshi, S., Gan, E., Chen, P.Y., Mirzasoleiman, B.: Which features are learnt by contrastive learning? on the role of simplicity bias in class collapse and feature suppression. In: International Conference on Machine Learning (2023)
39. Zhao, H., Gan, C., Rouditchenko, A., Vondrick, C., McDermott, J., Torralba, A.: The sound of pixels. In: Proceedings of the European conference on computer vision (ECCV). pp. 570–586 (2018)
40. Zhou, J., Shen, X., Wang, J., Zhang, J., Sun, W., Zhang, J., Birchfield, S., Guo, D., Kong, L., Wang, M., Zhong, Y.: Audio-visual segmentation with semantics. arXiv preprint arXiv:2301.13190 (2023)
41. Zhou, J., Wang, J., Zhang, J., Sun, W., Zhang, J., Birchfield, S., Guo, D., Kong, L., Wang, M., Zhong, Y.: Audio-visual segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV) (2022)