T-VSL: Text-Guided Visual Sound Source Localization in Mixtures

Tanvir Mahmud
University of Texas at Austin

Yapeng Tian
University of Texas at Dallas

Diana Marculescu
University of Texas at Austin

Abstract

Visual sound source localization poses a significant challenge in identifying the semantic region of each sounding source within a video. Existing self-supervised and weakly supervised source localization methods struggle to accurately distinguish the semantic regions of each sounding object, particularly in multi-source mixtures. These methods often rely on audio-visual correspondence as guidance, which can lead to substantial performance drops in complex multi-source localization scenarios. The lack of access to individual source sounds in multi-source mixtures during training exacerbates the difficulty of learning effective audio-visual correspondence for localization. To address this limitation, in this paper, we propose incorporating the text modality as an intermediate feature guide using tri-modal joint embedding models (e.g., AudioCLIP) to disentangle the semantic audio-visual source correspondence in multi-source mixtures. Our framework, dubbed T-VSL, begins by predicting the class of sounding entities in mixtures. Subsequently, the textual representation of each sounding source is employed as guidance to disentangle fine-grained audio-visual source correspondence from multi-source mixtures, leveraging the tri-modal AudioCLIP embedding. This approach enables our framework to handle a flexible number of sources and exhibits promising zero-shot transferability to unseen classes during test time. Extensive experiments conducted on the MUSIC, VGGSound, and VGGSound-Instruments datasets demonstrate significant performance improvements over state-of-the-art methods. Code is released at `https://github.com/enyac-group/T-VSL/tree/main`.

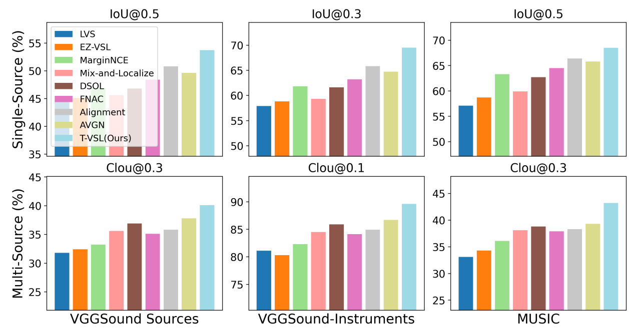

Figure 1. Comparison of our T-VSL with state-of-the-art methods on single-source (Top Row) and multi-source (Bottom Row) sound localization on VGGSound Sources [7], VGGSound-Instruments [21], and MUSIC [49] benchmark datasets. We use the same setup for the fair comparison.

1. Introduction

While observing a conversation between two individuals, we can easily associate the audio signal with the corresponding speaking person in the visual scene. This remarkable ability to perceive audio-visual correspondence stems from our extensive exposure to both single-source sounds and multi-source mixtures in everyday life. Inspired by this human capability, significant research efforts [21, 30, 42, 43] have been dedicated to developing visual sound source localization approaches in recent years.

Visual sound source localization aims to locate the visual regions representing sound sources present in a video. Earlier methods on single-source localization [1, 7, 19, 28, 29, 40, 41] mostly used audio-visual correspondence as guidance for localizing sounding objects in each frame. These approaches, developed for single sound source localization, suffer from significant performance drop in localizing multi-source mixtures. Multi-source localization from mixtures is very challenging, as the model must learn to distinguish each sounding source within a mixture and establish their cross-modal relationships, *without access to individual source sounds.* Earlier methods [20, 21, 30, 36] on multi-source localization attempt to infer fine-grained cross-modal associations directly from noisy multi-source mixtures, often leading to sub-optimal performance in practice for learning improper source correspondence. In this paper, we tackle the problem by introducing a unified solution for localizing visual sound sources in both single and multi-source mixtures.

The primary challenge of this task stems from the difficulty in disentangling the cross-modal correspondence of each sounding object from natural mixtures, given the absence of one-to-one corresponding single-source pairs. Additionally, the presence of silent visual objects and noise from invisible background sources further complicates this alignment process. To solve this problem, our key idea is to leverage the text modality as a coarse supervision for disentangling categorical audio-visual correspondence in natural mixtures, which differentiates our method from existing works. Unlike audio and visual modalities that may contain significant noises from different sources present in mixtures, textual representation can explicitly discriminate across multiple sources. However, a critical obstacle to utilizing text guidance in visual sound source localization is grounding fine-grained audio and visual features with their textual representation. Later, CLIP [39] introduced learning visual language grounding from web-scraped 400M image-text pairs. Recently, AudioCLIP [17] introduced the audio modality in the existing CLIP architecture, thereby generating tri-modal feature grounding through large-scale training. Our approach capitalizes on the tri-modal joint embedding space of AudioCLIP to disentangle one-to-one audio-visual correspondence in natural mixtures.

To this end, in this paper, we propose a novel text-guided multi-source localization framework that can discover fine-grained audio-visual semantic correspondence in multi-source mixtures. Initially, we detect the class instances of visual sounding objects in the frame using the noisy mixture features from AudioCLIP image and audio encoders. Then, the text representation of each detected sounding source class instance is extracted with AudioCLIP text encoder, which serves as a coarse guidance for audio and visual feature separation. Specifically, we disentangle the categorical audio and visual features of each sounding source using the coarse text-guidance through additional audio and image conditioning blocks, respectively. Finally, the extracted categorical audio-visual semantic features are further aligned through an audio-visual correspondence block for localizing each sounding source. In comparison with prior multi-source baselines, our approach can selectively localize the semantic visual region of each class of sounding objects in mixtures of varying number of sources. Extensive experiments on MUSIC [49], VGGSound [6], and VGGSound-Instruments [21] demonstrate significant performance improvements compared to existing single and multi-source baselines. Moreover, our method shows promising zero-shot transferability on unseen classes during test time. In addition, thorough ablation studies and qualitative analysis vividly showcase the effectiveness of the proposed framework.

Our main contributions can be summarized as follows:

- We propose a novel text-guided multi-source localization framework to disentangle one-to-one audio-visual correspondence from natural mixtures.
- Our work is the first to introduce AudioCLIP for utilizing the tri-modal feature grounding from large-scale pretraining in solving multi-source localization problems.
- Our proposed framework can operate with a flexible number of sources and shows promising zero-shot performance on unseen audio-visual classes.
- Extensive experiments on benchmark datasets clearly demonstrate the superiority of our proposed framework over other state-of-the-art methods.

2. Related Work

Visual Sound Source Localization. Visual sound source localization aims to locate the visual regions of sounding objects in video frames. Prior work explored diverse methods for audio-visual feature alignment, including traditional statistical machine learning approaches [12, 18, 23] to latest deep neural nets [1, 7, 19, 28, 29, 36, 40, 41]. However, most of these self-supervised and weakly-supervised methods are mostly designed for localizing single-source sounds. A two-stream architecture is proposed in Attention10k [40] which incorporates cross-attention to locate sounding sources. Later, motion cues present in videos are incorporated for finer localization [1]. Recently, various contrastive learning based methods are explored in LVS [7], EZ-VSL [28], and SLAVC [29] for single-source localization. Despite their promising performance, these single-source localization methods can hardly discriminate audio-visual correspondence in multi-source mixtures.

Sounds mostly appear in mixtures in the presence of significant background noises. The main challenges of multi-source localization is to disentangle the audio-visual correspondence in mixtures without having access to clean single-source audio-visual pairs. Several recent works have explored multi-source localization with explicit approaches for mixtures [20, 21, 30, 36]. To selectively isolate the silent objects in visual frames, DSOL [20] proposed a two-stage weakly-supervised training method. Mix-and-localize [21] proposed a self-supervised training with contrastive random walk to inherently align each sounding source with its visual representation. Recently, AVGN [30] introduced weakly-supervised grouping of category-aware features with learnable prompts. However, these approaches tackle the audio-visual correspondence in mixtures without any explicit guidance for single-source thereby resulting in sub-optimal performance. In contrast, we leverage text representation of fine-grained sounding sources to disentangle one-to-one audio-visual correspondence from multi-source mixtures. To the best of our knowledge, this work is the first to utilize text representation of sound sources for solving multi-source localization.

Audio, Visual, and Text Grounding. Large-scale pretraining on web-scraped data has been used to generate grounded features across modalities. CLIP [39] first introduced vision-language grounding using 400M image-text pairs, which has been extensively studied for zero-shot classification [2, 11, 14, 15, 22, 22, 44], text-video retrieval [3, 8, 25], open-vocabulary segmentation [9, 13, 26, 31, 37, 47, 48] and object detection [5, 10, 27, 45]. Re-

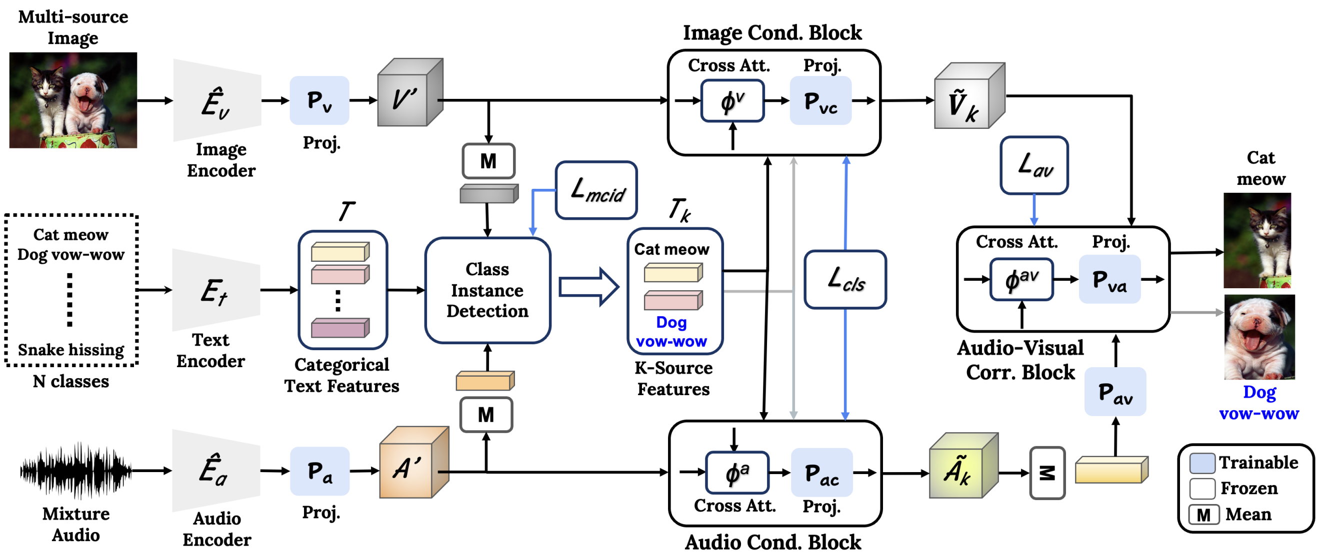

Figure 2. The proposed text-guided visual sound source localization (T-VSL) framework (for $K = 2$). We use the text modality to disentangle the fine-grained audio-visual correspondence from mixtures. Initially, we detect the audio-visual class instances from multi-source mixtures using AudioCLIP joint embedding space. Later, categorical text features of each detected K classes are used as coarse guidance in conditioning blocks to extract categorical visual and audio features. Afterwards, cross-modal feature alignment on extracted categorical features is performed in audio-visual correspondence block. Finally, cosine similarity of mean categorical audio features, and aligned visual features are used recursively to extract localization map of each class.

cently, AudioCLIP [17] and Wav2CLIP [46] introduced additional audio grounding in original CLIP model by learning on image, text, and audio triplets. Very recently, a CLIP-based single source localization method [34] is introduced that attempts replacing the text-conditioning branch of a supervised image segmentation baseline, trained with dense supervision, with audio conditioning. In relation with this concurrent work, our focus lies primarily on weakly supervised sound source localization in multi-source mixtures, leveraging the self-supervised AudioCLIP model.

3. Method

Given a mixture of audio and a video frame, our objective is to spatially localize each sounding object in the frame. We propose a novel text-guided multi-source localization framework to disentangle the fine-grained audio-visual correspondence, which is illustrated in Fig. 2. To suppress the background noises, we initially detect the common class instances in both audio and visual modality (Sec. 3.2). In particular, we leverage tri-modal AudioCLIP encoders to detect K visual-sounding class instances from N classes ($K \leq N$). The primary challenge in multi-source localization lies in establishing one-to-one audio-visual correspondences for individual sources without access to single-source samples. To overcome this, we leverage the text modality to generate coarse representation of the K detected sources in the mixture. Later, by exploiting the multi-modal grounding of AudioCLIP, we extract categorical audio and visual features from multi-source representation using text guidance (Sec. 3.3). Rather than aligning audio-visual mixture features as prior work, we introduce the categorical audio-visual feature alignment to further enhance one-to-one correspondence (Sec. 3.4). Finally, cosine similarity of categorical audio features and aligned visual features is used to extract semantic localization maps.

3.1. Preliminaries

In this section, we formulate the multi-source localization problem, and revisit the AudioCLIP [17] for audio, visual, and text grounding.

Problem Formulation. Let's assume the dataset contains a total of N classes of sounding events across all videos. Given a video frame and an audio mixture, our objective is to localize $K (K \leq N)$ sound sources within the frame. For each audio-visual mixture pair, we can extract $Y = \{y_i\}_{i=1}^{N}$ binary ground truth labels, along with categorical text representations for the on-screen sound sources in the mixture. During training, we are limited to video-level sounding class entities, excluding bounding boxes or mask-level annotations, thus constituting weakly-supervised learning.

Revisit AudioCLIP for multi-modal grounding. AudioCLIP [17] learns a joint embedding space of audio, visual, and text triplets through contrastive learning. Let's consider the dataset $\mathcal{D} = \{a, v, t\}_{i=1}^{Z}$ containing audio spectrogram $a \in \mathbb{R}^{M \times F}$, text $t \in \mathbb{R}^{l}$, and image $v \in \mathbb{R}^{C \times H \times W}$ triplets.

Also, separate audio encoder ($E_a : \mathbb{R}^{M\times F} \rightarrow \mathbb{R}^D$), visual encoder ($E_v : \mathbb{R}^{C\times H\times W} \rightarrow \mathbb{R}^D$), and text encoder ($E_t : \mathbb{R}^l \rightarrow \mathbb{R}^D$) are used for each modality. Hence, the extracted audio, visual, and text features are represented by $A = E_a(a)$, $V = E_v(v)$, and $T = E_t(t)$, respectively, where $A, V, T \in \mathbb{R}^D$. Later, InfoNCE contrastive loss ($\mathcal{L}_{\text{cnt}}$) [33] across each modality pair is used for feature grounding given by:

$$\mathcal{L}_{\text{aclip}} = \mathcal{L}_{\text{cnt}}(A, T) + \mathcal{L}_{\text{cnt}}(V, T) + \mathcal{L}_{\text{cnt}}(A, V) \quad (1)$$

This training objective, combined with large-scale pretraining on web-scraped data, establishes a grounded joint embedding space across all three modalities.

3.2. Audio-Visual Class Instance Detection

The target sounding objects for localization must be present in both the audio mixture and the corresponding video frame. However, audio mixtures may contain noise from off-screen background sources, while video frames may include non-sounding objects. To address these challenges, we first detect the class instances of visible sounding objects in the frame. Then, we utilize this detection to suppress redundant features in the mixture, ensuring that the localization focuses on the relevant sounding objects.

Token extraction from Uni-modal encoders. Multiple sounding objects can be spatially located in different regions of the reference frame, while the audio mixture may contain the temporally distributed signal of different sources. This makes it challenging to detect multiple class instances with uni-dimensional single-token features generated from AudioCLIP encoders. Instead of a single pooled token representation, we extract numerous patch tokens from audio and visual encoders with simple modifications, such that $\widehat{E}_a : \mathbb{R}^{M\times F} \rightarrow \mathbb{R}^{n_a\times D}$ and $\widehat{E}_v : \mathbb{R}^{C\times H\times W} \rightarrow \mathbb{R}^{n_v\times D}$. Here, $n_a = m \times f$ and $n_v = h \times w$ represent the number of audio and visual patch tokens, respectively. Afterwards, we apply linear projectors on these extracted patch tokens, such that $P_a : \mathbb{R}^{n_a\times D} \rightarrow \mathbb{R}^{n_a\times D}$ and $P_v : \mathbb{R}^{n_v\times D} \rightarrow \mathbb{R}^{n_v\times D}$. Thus, extracted audio and visual patch tokens are represented as $A' = P_a(\widehat{E}_a(a))$, $V' = P_v(\widehat{E}_v(v))$, respectively, where $A' \in \mathbb{R}^{n_a\times D}$ and $V' \in \mathbb{R}^{n_v\times D}$. Simultaneously, categorical text features $\mathcal{T} = \{e_i^t\}_{i=1}^N \in \mathbb{R}^{N\times D}$ representing all N sound source classes in the dataset are extracted.

Multi-label class instance detection. With the categorical text features $\mathcal{T}$, mixture audio A', and visual V' patch tokens, we detect class entities present in the mixture. Initially, we extract the mean-pooled audio-visual features $X' \in \mathbb{R}^{2D}$ by simple concatenation, and then, apply fusion projector $P_f : \mathbb{R}^{2D} \rightarrow \mathbb{R}^D$ to extract mixture audio-visual features $F_{av} \in \mathbb{R}^D$. Finally, cosine similarity between N class text features $\mathcal{T}$ and F_{av} is used to detect mixture class entities $\widehat{Y} \in \mathbb{R}^N$, given by:

$$\begin{aligned} &F_{av} = P_f(X'),\ X' = [\texttt{Mean}(A') \oplus \texttt{Mean}(V')], \\ &\widehat{Y} = \texttt{sim}(\mathcal{T}, F_{av}) \end{aligned} \quad (2)$$

where $[\ \oplus\ ]$ denotes the concatenation operator, and $\texttt{sim}(X, Y)$ denotes cosine similarity between X and Y. We formulate a multi-label classification objective $\mathcal{L}_{\text{mcid}}$ by applying binary cross-entropy loss $\mathcal{L}_{\text{bce}}(\cdot)$ on each prediction using video-level ground-truth labels $Y \in \mathbb{R}^N$ as:

$$\mathcal{L}_{\text{mcid}} = \sum_{i=1}^{N} \mathcal{L}_{\text{bce}}(y_i, \widehat{y}_i) \quad (3)$$

In particular, we train audio, visual, and fusion projectors (P_a, P_v, P_f) on noisy multi-source data keeping the AudioCLIP backbones frozen. Hence, these projectors assist in extracting relevant foreground features from noisy audio and visual mixture features extracted from AudioCLIP.

3.3. Audio and Visual Conditioning Blocks

Establishing one-to-one correspondences for each source in multi-source mixtures is particularly challenging, as both audio and visual features contain mixed representations of various sources. Prior works [21, 30] on multi-source localization have attempted to learn audio-visual alignment intrinsically without explicit single-source guidance. However, in the absence of single-source data, these methods struggle to adequately disentangle multi-source features. In contrast, we leverage the categorical text features of K classes present in the mixture as coarse guidance for fine-grained feature disentanglement from multi-source mixtures. We use ground truth weak labels Y of representative audios in the mixture to extract text embedding of K visual sounding sources $\mathcal{T}_K = \{e_k^t\}_{k=1}^K \subseteq \mathcal{T}$ where $e_k^t \in \mathbb{R}^{1\times D}$.

With the categorical text features $\mathcal{T}_K$ of K sources, we independently disentangle the multi-object visual patch tokens V' and mixture audio patch tokens A' by using cross-attention $\phi^v(\cdot)$ and $\phi^a(\cdot)$, respectively, as:

$$\begin{aligned} &\widetilde{V}_k = \phi^v(V', e_k^t, V'),\ f_k^v = \texttt{Mean}(\widetilde{V}_k), \\ &\widetilde{f}_k^v = P_{vc}(f_k^v),\ \ \forall k = \{1, \ldots, K\} \end{aligned} \quad (4)$$

$$\begin{aligned} &\widetilde{A}_k = \phi^a(A', e_k^t, A'),\ f_k^a = \texttt{Mean}(\widetilde{A}_k), \\ &\widetilde{f}_k^a = P_{ac}(f_k^a),\ \ \forall k = \{1, \ldots, K\} \end{aligned} \quad (5)$$

$$\phi(A, B, C) = \left(\frac{AB^\top}{\|A\|\|B\|}\right) \odot C \quad (6)$$

where $\widetilde{f}_k^a, \widetilde{f}_k^v, f^a, f^v \in \mathbb{R}^{1\times D}$, $\widetilde{V}_k \in \mathbb{R}^{n_v\times D}$, $\widetilde{A}_k \in \mathbb{R}^{n_a\times D}$, and $P_{ac}, P_{vc} : \mathbb{R}^D \rightarrow \mathbb{R}^D$ are linear audio and visual projectors on conditioned mean features, respectively, and $\odot$ is the Hadamard product.

Method	VGGSound-Single			VGGSound-Instruments			MUSIC-Solo		
	AP(%)	IoU@0.5(%)	AUC(%)	AP(%)	IoU@0.3(%)	AUC(%)	AP(%)	IoU@0.5(%)	AUC(%)
Current SOTA Based Methods									
OTS [4] ECCV18	38.9	43.2	43.7	50.9	56.9	38.4	76.2	57.8	48.9
CoarsetoFine [36] ECCV20	39.4	43.1	44.8	51.1	57.5	39.1	76.3	58.1	49.7
LVS [7] CVPR21	39.7	44.6	45.3	51.5	57.9	39.4	76.1	57.1	49.3
EZ-VSL [28] ECCV22	40.6	45.1	47.2	52.3	58.8	39.9	78.4	58.7	51.1
Mix-and-Localize [21] CVPR22	41.8	45.6	47.4	53.8	59.3	41.5	78.9	59.9	52.7
DSOL [20] NeurIPS20	–	46.8	47.9	–	61.6	43.7	–	62.7	54.6
MarginNCE [35] ICASSP23	42.5	46.9	48.3	56.9	61.8	44.2	83.1	63.3	55.2
FNAC [43] CVPR23	43.3	48.4	49.1	57.2	63.2	45.3	84.6	64.5	56.4
AVGN [30] CVPR23	44.1	49.6	49.5	59.3	64.7	46.1	85.4	65.8	56.9
Alignment [42] ICCV23	45.3	50.8	50.2	60.4	65.8	48.7	86.1	66.4	57.2
CLIP-Based Baseline Methods									
AudioCLIP [17] ICASSP22	42.8	47.4	48.5	58.3	62.7	45.2	83.8	63.1	55.7
Wav2CLIP [46] ICASSP22	39.3	43.6	44.7	53.8	57.2	42.0	78.9	59.9	51.2
T-VSL (Ours)	**48.1**	**53.7**	**52.9**	**64.6**	**69.5**	**51.4**	**88.2**	**68.5**	**60.1**

Table 1. Performance comparison of single-source localization on VGGSound-single, VGGSound-Instruments, and MUSIC-Solo datasets. For the fair comparison, we use pre-trained AudioCLIP audio and image encoders for baseline methods.

Additionally, to guide extracting categorical audio ($\widetilde{A}_k$) and visual ($\widetilde{V}_k$) patch tokens, we introduce coarse supervision on the output of projected mean patch tokens $\widetilde{f}_k^a, \widetilde{f}_k^v$ using N-class text embedding $\mathcal{T} \in \mathbb{R}^{N\times D}$. Hence, the class conditioning loss ($\mathcal{L}_{\text{cls}}$) is given by:

$$\begin{aligned} &\mathbf{e}_k^v = \texttt{sim}(\mathcal{T}, \widetilde{f}_k^v);\ \mathbf{e}_k^a = \texttt{sim}(\mathcal{T}, \widetilde{f}_k^a) \\ &\mathcal{L}_{\text{cls}} = \sum_{k=1}^{K} \mathcal{L}_{\text{ce}}(\mathbf{e}_k^v, \mathbf{h}_k) + \mathcal{L}_{\text{ce}}(\mathbf{e}_k^a, \mathbf{h}_k) \end{aligned} \tag{7}$$

where $\mathbf{h}_k \in \mathbb{R}^N$ denotes an one-hot encoding of the class label, and $\mathbf{e}_k^v \in \mathbb{R}^N$, $\mathbf{e}_k^a \in \mathbb{R}^N$ are visual and audio class predictions for each k^{th} source, respectively.

3.4. Audio-Visual Correspondence Block

While text representation provides coarse guidance for disentangling fine-grained spatio-temporal audio and visual features, it has inherent limitations in practice. For example, multiple instances of a sounding object class can appear in the visual frame, including silent instances, which are difficult to distinguish using simple text class representation. To address this challenge, we introduce an audio-visual correspondence block to further refine the alignment of audio and visual features based on extracted categorical features.

With the disentangled categorical audio ($\widetilde{A}_k$) and visual ($\widetilde{V}_k$) patch tokens, we apply cross-attention $\phi^{av}(\cdot)$ to enhance audio-visual patch alignment as:

$$\begin{aligned} &g_k^a = \texttt{Mean}(\widetilde{A}_k),\ \widehat{g}_k^a = P_{av}(g_k^a) \\ &\widehat{V}_k = \phi^{av}(\widetilde{V}_k, \widehat{g}_k^a, \widetilde{V}_k),\ g_k^v = \texttt{Mean}(\widehat{V}_k) \\ &\widehat{g}_k^v = P_{va}(g_k^v),\ \forall k = \{1, \dots, K\} \end{aligned} \tag{8}$$

where $P_{av}, P_{va} : \mathbb{R}^D \to \mathbb{R}^D$ are linear projectors. Afterwards, to correspond between projected mean categorical audio and visual features $\widehat{g}_k^a, \widehat{g}_k^v \in \mathbb{R}^D$, we apply audio-visual contrastive loss ($\mathcal{L}_{\text{av}}$) given by:

$$\begin{aligned} &\mathcal{L}_{\text{av}} = \mathcal{L}_{\text{cnt}}(\widehat{g}_k^v, \widehat{g}_k^a) \\ &\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{av}} + \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{mcid}} \end{aligned} \tag{9}$$

where $\mathcal{L}_{\text{total}}$ is the total accumulated loss.

Inference: During inference, we estimate cosine similarity across projected mean audio feature $\widehat{g}_k^a \in \mathbb{R}^D$ and aligned visual patch tokens ($\widehat{V}_k = \{\widehat{v}_k^j\}_{j=1}^{n_v}, \widehat{v}_k^j \in \mathbb{R}^D$). Finally, after reshaping and up-scaling through bilinear interpolation, we generate K-source heat-maps $\mathbf{H} \in \mathbb{R}^{K\times H\times W}$.

4. Results and Discussions

4.1. Experimental setup

Datasets. We used MUSIC, VGGSound-Instruments, and VGGSound datasets for the performance evaluation[1]. MUSIC [49] dataset contains 445 solo music videos of 11 instruments and 142 duet music videos of 8 instruments from YouTube. Following prior works [30], we use MUSIC-Solo for single-source and MUSIC-Duet for multi-source localization. From MUSIC-Solo, we use 350 videos for training and remaining 95 for evaluation. For MUSIC-Duet, we use 120 videos for training and remaining 22 videos for evaluation. VGGSound-Instruments [21] is a subset of VGGSound dataset [6] containing around 32k video clips from 37 music instruments of 10s duration. Apart from the musical instrument datasets, we gather around 150k videos of 10s duration from original VGGSound dataset [6] representing single-source sounds of 221 categories, such as animals, people, nature, instruments, etc. For the single-source evaluation on VGGSound, we use the full test

[1]Since many videos are no longer available for public use, these datasets become smaller. We use the same data split for all baselines and reproduce them under the same setting for a fair comparison.

Method	VGGSound-Duet CAP(%)	VGGSound-Duet CIoU@0.3(%)	VGGSound-Duet AUC(%)	VGGSound-Instruments CAP(%)	VGGSound-Instruments CIoU@0.1(%)	VGGSound-Instruments AUC(%)	MUSIC-Duet CAP(%)	MUSIC-Duet CIoU@0.3(%)	MUSIC-Duet AUC(%)
Current SOTA Single-Source Methods									
OTS [4] ECCV18	23.9	26.4	27.5	28.7	77.8	18.2	51.2	29.4	24.3
CoarsetoFine [36] ECCV20	–	28.4	27.9	–	78.8	19.7	–	31.8	24.1
LVS [7] CVPR21	–	31.8	30.1	–	81.1	23.2	–	33.1	26.9
EZ-VSL [28] ECCV22	–	32.4	30.6	–	80.3	22.9	–	34.3	26.7
MarginNCE [35] ICASSP23	27.1	33.2	31.2	33.5	82.3	24.7	55.1	36.1	28.5
FNAC [43] CVPR23	29.8	35.1	33.4	35.4	84.1	26.8	57.3	37.9	30.6
Alignment [42] ICCV23	30.4	35.8	33.9	35.6	84.9	27.1	57.8	38.3	30.9
Current SOTA Multi-Source Methods									
Mix-and-Localize [21] CVPR22	29.5	35.6	34.1	36.3	84.5	26.7	57.4	38.1	30.7
DSOL [20] NeurIPS20	–	36.9	34.6	–	85.9	28.4	–	38.8	31.3
AVGN [30] CVPR23	31.9	37.8	35.4	38.1	86.7	28.8	59.2	39.3	32.4
CLIP-Based Baseline Methods									
AudioCLIP [17] ICASSP22	28.4	34.9	32.8	34.7	83.8	25.9	56.1	36.9	29.2
Wav2CLIP [46] ICASSP22	26.3	31.4	28.9	31.2	80.3	22.2	52.6	32.2	27.2
T-VSL (ours)	**35.7**	**40.1**	**37.9**	**41.8**	**89.6**	**31.5**	**62.9**	**43.2**	**35.9**

Table 2. Performance comparison of multi-source localization on VGGSound-Duet, VGGSound-Instruments, and MUSIC-Duet datasets. For the fair comparison, we use pre-trained AudioCLIP audio and image encoders for baseline methods.

set of 5158 videos, while on VGGSound-Instruments, we use 446 videos following prior work [21, 30]. For the multi-source training and evaluation in both VGGSound and VGGSound-Instruments, we randomly concatenate two frames from different sounding sources to produce the multi-source input image with a size of 448×224, and mix the single-source audios, following prior work [21, 30].

Evaluation Metrics. For evaluating single-source localization performance, we use Intersection over Union (IoU), average precision (AP), and Area Under Curve (AUC) following prior work [21, 30]. For the multi-source localization evaluation, we use Class-aware Average precision (CAP), Class-aware IoU (CIoU), and Area Under Curve (AUC) following prior work [21, 30]. Similarly, for fair comparison with the prior work [30], we use the same thresholds in these metrics: we use IoU@0.5 and CIoU@0.3 for the MUSIC-Solo and MUSIC-Duet datasets, IoU@0.3 and IoU@0.5 for single-source VGGSound-Instruments and VGGSound datasets, and CIoU@0.1 and CIoU@0.3 for multi-source VGGSound-Instruments and VGGSound.

Implementation details. We use the AudioCLIP [17] pre-trained encoders containing ResNet-50 [19] image encoder, ESResNeXt [16] audio encoder, and transformer [38] based text encoder. For the fair comparison, we reproduce the result of existing methods with AudioCLIP image and audio encoders, instead of using separately pre-trained encoders. We use 224×224 resolution for the single-source input image, $D = 1024$, $n_v = 49$ for $7 \times 7 (h \times w)$ spatial maps generated from the visual encoder, and $n_a = 60$ for $10 \times 6 (m \times f)$ time-frequency map generated from the audio encoder. We use Adam optimizer [24] with a batch size of 256 and with a learning rate of $1e-4$.

CLIP-Based Baseline methods. We adopt several CLIP-based baseline methods for the sound source localization task, which are described as follows:

- **AudioCLIP** [17]: We fine-tune pre-trained image and audio encoders of AudioCLIP on the target dataset. We estimate the localization heatmap for the sounding sources using the similar cosine similarity across patch tokens.
- **Wav2CLIP** [46]: Similar to AudioCLIP baseline, we experiment with Wav2CLIP audio and image encoders to extract localization heatmaps of sounding sources.

4.2. Comparison to SOTA Baselines

Single-source localization: We present the quantitative results on single-source localization in Tab. 1. We apply the AudioCLIP image and audio encoders in all baseline methods for the fair comparison with our methods. We observe some performance improvements with several SOTA methods [30, 42, 43] over the AudioCLIP baseline. However, we also notice some reduction in performance from the AudioCLIP baseline in several methods [4, 21, 28]. We hypothesize that these can be due to the alterations of AudioCLIP grounding with the modified contrastive loss. In contrast, by utilizing all three modalities for sound source localization, the proposed T-VSL achieves the best performance over all CLIP-based baselines as well as SOTA methods. T-VSL outperforms the current SOTA single source localization method Alignment [42] by +2.9, +3.7, +2.1 higher IoUs on VGGSound-Single, VGGSound-Instruments, and MUSIC-solo datasets, respectively. Moreover, T-VSL achieves +6.3, +6.8, and +5.4 higher IoUs over the AudioCLIP baseline [32]. We hypothesize that the enhanced refinement of noisy audio and visual features with text guidance in T-VSL effectively contributes to performance improvement.

Multi-source localization: We present performance comparison in multi-source sound localization in Tab. 2. As be-

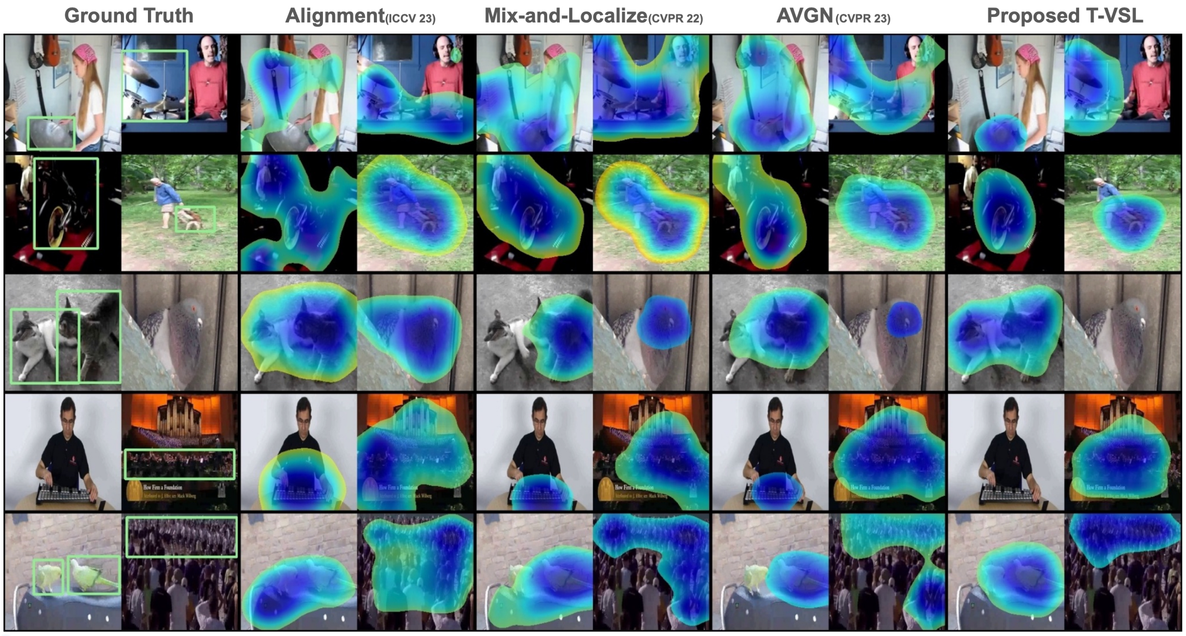

Figure 3. We present qualitative comparisons on challenging multi-source localization with SOTA single and multi-source baseline methods. Here, blue color represents high-attention values to the sounding object, and red color represents low-attention values. The proposed T-VSL can selectively isolate the sounding regions from the background and generates more precise localization maps for sounding sources.

fore, we apply the AudioCLIP image and audio encoders in all baseline methods for the fair comparison with our methods. We observe some performance improvements, mostly in SOTA multi-source methods [20, 21, 30], over the AudioCLIP baseline. In most single source methods, we observe performance loss in multi-source cases over the baseline. The proposed T-VSL achieves the best performance that outperforms the SOTA multi-source baseline AVGN [30] by $+2.3$, $+2.9$, and $+3.9$ CIoUs on multi-source VGGSound-Duet, VGGSound-Instruments, and MUSIC-Duet datasets, respectively. Also, T-VSL achieves $+5.2$, $+5.8$, and $+6.3$ higher CIoUs over the AudioCLIP baseline. These significant performance gains on multi-source benchmarks demonstrate the effectiveness of the proposed T-VSL in disentangling multi-source mixtures.

Qualitative comparisons: In addition, we present qualitative comparisons on multi-source localization in Fig. 3 among single-source baseline Alignment [42], self-supervised multi-source baseline Mix-and-Localize [21], weakly-supervised multi-source baseline AVGN [30], and proposed T-VSL. Single-source SOTA baseline Alignment [42] struggles in multi-source localization due to the lack of audio-visual correspondence, mostly in the presence of noisy audios and silent visual objects. However, without any guidance on fine-grained source localization, multi-source baselines [21, 30] underperform on localizing challenging multi-source mixtures. Notably, our method generates superior localization maps, which shows the effectiveness of T-VSL in multi-source feature disentanglement.

4.3. Ablation on T-VSL building blocks

We present the ablation study of different building blocks of T-VSL in Tab. 3 under both single-source and multi-source settings. The vanilla baseline contains AudioCLIP image and audio encoders that directly operate on the input data. By integrating the audio-visual correspondence (AVC) block on extracted mixture patch tokens, we notice considerable performance increase in single-source (by $+1.1$ AP and $+0.7$ IoU@0.5) and multi-source localization (by $+1.3$ CAP and $+1.4$ CIoU@0.3). By integrating the audio and image conditioning blocks following audio-visual instance detection for separating fine-grained class features, we observe additional improvements of $+1.8$ AP and $+2.3$ AP in single-source, and $+3.1$ CAP and $+3.6$ AP in multi-source localization, respectively. Finally, the combination of all building blocks results in the best performance that improves the baseline by $+5.3$ AP, and 6.3 IoU@0.5 in single-source and by $+7.3$ CAP and $+5.2$ CIoU@0.3 in multi-source localization. This further demonstrates the effective disentanglement of audio-visual features facilitated by intermediate text guidance.

4.4. Zero-shot transfer across datasets

We can perform zero-shot transfer across datasets with proposed T-VSL by simply replacing the N-class text prompts in Fig. 2. Since other weakly-supervised methods cannot perform zero-shot transfer, we present comparisons with self-supervised Mix-and-Localize [21], Alignment [42], and FNAC [43] methods in Tab. 4. To ensure a fair comparison, we conducted uniform training with 50%-50% class

Audio Cond.	Image Cond.	AVC	VGGSound-Single		VGGSound-Duet	
			AP(%)	IoU@0.5(%)	CAP(%)	CIoU@0.3(%)
✗	✗	✗	42.8	47.4	28.4	34.9
✗	✗	✓	43.9	48.1	29.7	36.3
✓	✗	✓	45.7	50.9	32.8	38.1
✗	✓	✓	46.2	51.4	33.3	38.6
✓	✓	✓	**48.1**	**53.7**	**35.7**	**40.1**

Table 3. Ablation study on audio conditioning block, image-conditioning block, and audio-visual correspondence block (AVC) of T-VSL in single-source and multi-source localization tasks.

Target Dataset	Proposed T-VSL	Mix-and-Localize [21]	Alignment [42]	FNAC [43]
MUSIC-Duet	**55.8**	50.9	51.7	51.2
MUSIC-Solo	**80.5**	75.1	77.6	77.4
VGGSound-Instr.	**81.7**	76.8	76.5	76.1

Table 4. Ablation on zero-shot transfer across single and multi-source datasets with 50%-50% class split. We report IoU@0.5(%) for MUSIC-Solo, CIoU@0.3(%) for MUSIC-Duet, and CIoU@0.1(%) for VGGSound-Instruments. We use same AudioCLIP encoders for all baselines.

split for all methods with same AudioCLIP encoders. We note that our method achieves significantly better performance on zero-shot transfer to target datasets than these compared approaches. In particular, we achieve +4.9, +5.4, +4.9 improvements on target datasets than the multi-source Mix-and-Localize [21]. Also, we achieve +4.1, +2.9, and +5.2 higher scores than the SOTA single source baseline, Alignment [42]. This result suggests that our T-VSL has superior generalization capabilities for zero-shot localization tasks in unseen classes.

4.5. Robustness to a higher number of sources

The proposed T-VSL is not limited by the number of sources present in training mixtures unlike prior work [21]. We present comparative analysis with SOTA multi-source AVGN [30], and self-supervised single source methods FNAC [43] and Alignment [42] in Tab. 5 on robustness to higher number of test sources in VGGSound dataset. We train all methods on the VGGSound-Single dataset with same encoders for a fair comparison. We note that our method consistently maintains significantly higher CIoU score irrespective of test scenarios, and also achieves notably smaller relative performance drop for a large number of sources in mixtures. These results further demonstrate the robustness of the proposed T-VSL in disentangling multi-source mixtures.

4.6. Use of learnable text prompts

To disentangle multi-source mixtures, we primarily use the class label of each sounding source as text prompts (*e.g.* dog barking, snake hissing). Inspired by the conditional text prompt learning in [50], we study the use of learnable prompts along with class label representations. Instead of only using class label representation, a set of learnable embedding is integrated with class label embedding that provides additional flexibility to adapt the text prompt to target objective. We analyze the sound source localization performance with different length of learnable prompts. The performances are reported in Table 6. We notice considerable performance improvements by optimizing learnable prompts in both single and multi-source cases. However, such prompts limit zero-shot use cases to unseen classes.

Test Source No.	Proposed T-VSL	AVGN [30]	Alignment [42]	FNAC [43]
2	**39.6**	36.8	35.4	34.1
3	**35.7**	30.8	28.9	27.1
4	**29.5**	22.4	19.3	18.8

Table 5. Robustness to higher number of sources in VGGSound dataset when trained with single-source data. CIoU@0.3(%) score is reported. Same AudioCLIP encoders are used for all baselines.

Prompt Length	VGGSound-Single		VGGSound-Duet	
	AP(%)	IoU@0.5(%)	CAP(%)	CIoU@0.3(%)
0	48.1	53.7	35.7	40.1
2	48.5	54.1	36.0	40.4
4	49.2	54.6	36.8	41.3
8	**50.3**	**55.2**	37.1	42.5
16	50.2	55.0	**37.4**	**42.9**
32	49.7	54.5	37.1	42.7

Table 6. Ablation study on the effect of learnable text prompts on VGGSound-Single and VGGSound-Duet datasets. Integration of learnable prompts results in noticable performance improvements in both single and multi-source localization.

5. Conclusion

In this paper, we propose T-VSL, a novel text-guided multi-source visual sound source localization framework that can disentangle one-to-one audio-visual correspondence from multi-source mixtures. We leverage the text modality to guide fine-grained feature separation and localization, from noisy audio and visual features of mixtures, by exploiting the joint embedding space of AudioCLIP. In comparison with SOTA multi-source baselines, our method shows superior zero-shot transfer across datasets. Moreover, our method demonstrates notable robustness on challenging test mixtures with higher number of sources than training scenarios. In both single-source and multi-source localization, our method largely outperforms existing weakly-supervised and self-supervised baselines on three benchmark datasets.

Acknowledgements

This research was supported in part by ONR Minerva program, iMAGiNE - the Intelligent Machine Engineering Consortium at UT Austin, and a UT Cockrell School of Engineering Doctoral Fellowship.

References

[1] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 208–224, 2020. 1, 2

[2] Muhammad Ali and Salman Khan. Clip-decoder: Zeroshot multilabel classification using multimodal clip aligned representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4675–4679, 2023. 2

[3] Tayfun Alpay, Sven Magg, Philipp Broze, and Daniel Speck. Multimodal video retrieval with clip: a user study. *Information Retrieval Journal*, 26(1-2):6, 2023. 2

[4] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 435–451, 2018. 5, 6

[5] Maria A Bravo, Sudhanshu Mittal, and Thomas Brox. Localized vision-language matching for open-vocabulary object detection. In *DAGM German Conference on Pattern Recognition*, pages 393–408. Springer, 2022. 2

[6] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020. 2, 5

[7] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16867–16876, 2021. 1, 2, 5, 6

[8] Chaorui Deng, Qi Chen, Pengda Qin, Da Chen, and Qi Wu. Prompt switch: Efficient clip adaptation for text-video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15648–15658, 2023. 2

[9] Zheng Ding, Jieke Wang, and Zhuowen Tu. Open-vocabulary panoptic segmentation with maskclip. *arXiv preprint arXiv:2208.08984*, 2022. 2

[10] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14084–14093, 2022. 2

[11] Sepideh Esmaeilpour, Bing Liu, Eric Robertson, and Lei Shu. Zero-shot out-of-distribution detection based on the pre-trained model clip. In *Proceedings of the AAAI conference on artificial intelligence*, pages 6568–6576, 2022. 2

[12] John W Fisher III, Trevor Darrell, William Freeman, and Paul Viola. Learning joint statistical models for audio-visual fusion and segregation. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2000. 2

[13] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European Conference on Computer Vision*, pages 540–557. Springer, 2022. 2

[14] Sachin Goyal, Ananya Kumar, Sankalp Garg, Zico Kolter, and Aditi Raghunathan. Finetune like you pretrain: Improved finetuning of zero-shot vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19338–19347, 2023. 2

[15] Ziyu Guo, Renrui Zhang, Longtian Qiu, Xianzheng Ma, Xupeng Miao, Xuming He, and Bin Cui. Calip: Zero-shot enhancement of clip with parameter-free attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 746–754, 2023. 2

[16] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Esresne (x) t-fbsp: Learning robust time-frequency transformation of audio. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021. 6

[17] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE, 2022. 2, 3, 5, 6, 1

[18] John Hershey and Javier Movellan. Audio vision: Using audio-visual synchrony to locate sounds. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 1999. 2

[19] Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9248–9257, 2019. 1, 2, 6

[20] Di Hu, Rui Qian, Minyue Jiang, Xiao Tan, Shilei Wen, Errui Ding, Weiyao Lin, and Dejing Dou. Discriminative sounding objects localization via self-supervised audiovisual matching. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pages 10077–10087, 2020. 1, 2, 5, 6, 7

[21] Xixi Hu, Ziyang Chen, and Andrew Owens. Mix and localize: Localizing sound sources in mixtures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10483–10492, 2022. 1, 2, 4, 5, 6, 7, 8

[22] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19606–19616, 2023. 2

[23] Einat Kidron, Yoav Y Schechner, and Michael Elad. Pixels that sound. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005. 2

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[25] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022. 2

[26] Huaishao Luo, Junwei Bao, Youzheng Wu, Xiaodong He, and Tianrui Li. Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In *International Conference on Machine Learning*, pages 23033–23044. PMLR, 2023. 2

[27] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh

Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *European Conference on Computer Vision*, pages 728–755. Springer, 2022. 2

[28] Shentong Mo and Pedro Morgado. Localizing visual sounds the easy way. In *Proceedings of European Conference on Computer Vision (ECCV)*, page 218–234, 2022. 1, 2, 5, 6

[29] Shentong Mo and Pedro Morgado. A closer look at weakly-supervised audio-visual source localization. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1, 2

[30] Shentong Mo and Yapeng Tian. Audio-visual grouping network for sound localization from mixtures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10565–10574, 2023. 1, 2, 4, 5, 6, 7, 8

[31] Jishnu Mukhoti, Tsung-Yu Lin, Omid Poursaeed, Rui Wang, Ashish Shah, Philip HS Torr, and Ser-Nam Lim. Open vocabulary semantic segmentation with patch aligned contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19413–19423, 2023. 2

[32] Jishnu Mukhoti, Tsung-Yu Lin, Omid Poursaeed, Rui Wang, Ashish Shah, Philip HS Torr, and Ser-Nam Lim. Open vocabulary semantic segmentation with patch aligned contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19413–19423, 2023. 6

[33] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4

[34] Sooyoung Park, Arda Senocak, and Joon Son Chung. Can clip help sound source localization? *arXiv preprint arXiv:2311.04066*, 2023. 3, 1

[35] Sooyoung Park, Arda Senocak, and Joon Son Chung. Marginnce: Robust sound localization with a negative margin. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 5, 6

[36] Rui Qian, Di Hu, Heinrich Dinkel, Mengyue Wu, Ning Xu, and Weiyao Lin. Multiple sound sources localization from coarse to fine. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 292–308, 2020. 1, 2, 5, 6

[37] Jie Qin, Jie Wu, Pengxiang Yan, Ming Li, Ren Yuxi, Xuefeng Xiao, Yitong Wang, Rui Wang, Shilei Wen, Xin Pan, et al. Freeseg: Unified, universal and open-vocabulary image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19446–19455, 2023. 2

[38] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 6

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[40] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4358–4366, 2018. 1, 2

[41] Arda Senocak, Hyeonggon Ryu, Junsik Kim, and In So Kweon. Learning sound localization better from semantically similar samples. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022. 1, 2

[42] Arda Senocak, Hyeonggon Ryu, Junsik Kim, Tae-Hyun Oh, Hanspeter Pfister, and Joon Son Chung. Sound source localization is all about cross-modal alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7777–7787, 2023. 1, 5, 6, 7, 8

[43] Weixuan Sun, Jiayi Zhang, Jianyuan Wang, Zheyuan Liu, Yiran Zhong, Tianpeng Feng, Yandong Guo, Yanhao Zhang, and Nick Barnes. Learning audio-visual source localization via false negative aware contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6420–6429, 2023. 1, 5, 6, 7, 8

[44] Hualiang Wang, Yi Li, Huifeng Yao, and Xiaomeng Li. Clipn for zero-shot ood detection: Teaching clip to say no. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1802–1812, 2023. 2

[45] Tao Wang. Learning to detect and segment for open vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7051–7060, 2023. 2

[46] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2clip: Learning robust audio representations from clip. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4563–4567. IEEE, 2022. 3, 5, 6

[47] Jilan Xu, Junlin Hou, Yuejie Zhang, Rui Feng, Yi Wang, Yu Qiao, and Weidi Xie. Learning open-vocabulary semantic segmentation models from natural language supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2935–2944, 2023. 2

[48] Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, and Lei Zhang. A simple framework for open-vocabulary segmentation and detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1020–1031, 2023. 2

[49] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 570–586, 2018. 1, 2, 5

[50] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 8