Can CLIP Help Sound Source Localization?

Sooyoung Park[*1,2]   Arda Senocak[*1]   Joon Son Chung[1]

[1] Korea Advanced Institute of Science and Technology, South Korea

[2] Electronics and Telecommunications Research Institute, South Korea

Abstract

Large-scale pre-trained image-text models demonstrate remarkable versatility across diverse tasks, benefiting from their robust representational capabilities and effective multimodal alignment. We extend the application of these models, specifically CLIP, to the domain of sound source localization. Unlike conventional approaches, we employ the pre-trained CLIP model without explicit text input, relying solely on the audio-visual correspondence. To this end, we introduce a framework that translates audio signals into tokens compatible with CLIP's text encoder, yielding audio-driven embeddings. By directly using these embeddings, our method generates audio-grounded masks for the provided audio, extracts audio-grounded image features from the highlighted regions, and aligns them with the audio-driven embeddings using the audio-visual correspondence objective. Our findings suggest that utilizing pre-trained image-text models enable our model to generate more complete and compact localization maps for the sounding objects. Extensive experiments show that our method outperforms state-of-the-art approaches by a significant margin.

1. Introduction

The ability of humans and other animals to pinpoint the locations of sound sources is crucial for perceiving the world around us. We receive continuous multisensory information, such as auditory and visual inputs, understand their relationships, infer which object/event is producing sound, and focus on sounding objects/events. To provide machine perception with similar abilities, audio-visual sound source localization has been extensively explored in recent years [1, 4, 12, 16–18, 21, 22, 24, 25, 27–33]. One fundamental approach in this direction involves leveraging the natural
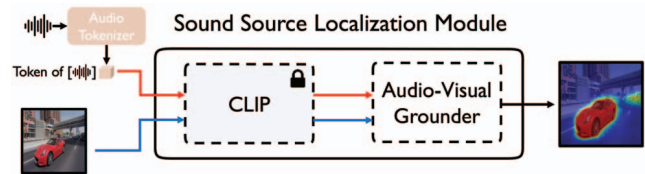
Figure 1. **The proposed text input-free CLIP based sound source localization method.**

correspondence between audio and visual signals without explicit supervision or the need for annotated data. The most predominant method for achieving this is by aligning audio-visual representations as a self-supervision signal within a contrastive learning framework.

While sound localization methods are trained with the aforementioned fundamental assumption, some additional prior knowledge is also incorporated. These pieces of prior knowledge are introduced in the form of using visual objectness [21, 22] and object proposal networks [38], or other modalities such as optical flow [10]. As true sound source localization methods necessitate a strong audio-visual semantic alignment, the previously mentioned priors might not contribute to improved alignment, as they can introduce visual objectness or motion bias that may lead to shortcuts [1, 21, 23]. In this work, our focus is to leverage strong multimodal alignment knowledge as a prior to improve audio-visual alignment for genuine sound source localization. From this perspective, we employ the Contrastive Language-Image Pretraining (CLIP) [26] model for the sound source localization task. This choice is due to its robust representation and multimodal alignment capability, stemming from learning directly from raw text about images on large scale data. Thus, it provides a broader source of supervision rather than limited category labels.

The frameworks that leverage the CLIP model generally include text queries/prompts. However, we aim to explore this approach without using explicit contextual text information. The reasons we do not intuitively utilize direct text inputs are as follows: (1) There is no available paired text data in sound source localization benchmark datasets, (2) the sound source localization task is unlabeled, (3) a genuine sound source localization approach necessitates learn-

ing pure audio-visual alignment through self-supervision. Therefore, in this paper, we employ the pre-trained CLIP model in a textless manner (as illustrated in Figure 1), relying solely on audio-visual correspondence.

To utilize CLIP in a text input-free manner and train our sound source localization method through self-supervised audio-visual alignment, we propose the following steps (depicted in Figure 2): First, we introduce a framework that translates audio signals into tokens compatible with CLIP's text encoder. This process yields contextual embeddings for the provided audio input, a concept we refer to as audio-driven embedding. Second, our key idea involves aligning audio and visual features in a self-supervised manner using contrastive learning. Consequently, we seamlessly integrate this audio-driven embedding to emphasize the sounding regions within the visual scenes. Subsequently, audio-grounded visual features on both the image and feature levels are extracted from these regions. These features are then aligned with the audio-driven embedding through audio-visual correspondence within a contrastive learning framework. The entire model is trained at once with the audio-visual alignment objective. Through our experiments, we validate that the proposed method outperforms existing approaches and baselines. In some instances, it even achieves competitive results when compared to fully supervised or text-queried sound source localization baselines.

We summarize the contributions of our work as follows:

- We present a novel self-supervised sound source localization framework that exploits the large-scale pre-trained CLIP model.

- We propose an end-to-end textless approach, *i.e.* no explicit text input. Our framework translates audio signals into tokens that are compatible with CLIP to obtain audio-driven embeddings.

- We utilize the audio-driven embeddings to emphasize the sounding regions and align them with the audio content for the objective of audio-visual correspondence.

- We conduct extensive experiments on the VGG-SS, SoundNet-Flickr, VGG-SS OpenSet, AVSBench, and Extended VGG-SS/SoundNet-Flickr datasets, collectively demonstrate the effectiveness of our proposed method.

2. Related work

Sound source localization. The predominant technique employed for audio-visual sound source localization involves cross-modal attention [27, 28, 35], often coupled with contrastive loss. Following the contrastive learning paradigm, subsequent enhancements have been made by explicitly incorporating hard negatives from background regions [4], utilizing iterative contrastive learning with pseudo-labels obtained from the same model in previous epochs [17], applying transformation invariance and equiv-

ariance through data augmentations and geometric consistency [18], considering semantically similar hard positives [29], implementing negative-free contrastive learning [32] similar to SiamSiam [7], using momentum encoders to mitigate overfitting [21], adding negative margin into contrastive learning alleviate the effect of noisy correspondences [24], and applying false negative-aware contrastive learning via intra-modal similarities [33]. Following a similar trend, our method also integrates self-supervised contrastive learning.

Besides this trend, some other sound localization methods attempt to utilize additional prior knowledge or post-processing approaches. [25, 30] incorporate label information to learn backbone audio and visual networks or to refine the audio-visual alignment. Xuan *et al*. [38] use object priors in the form of object proposals, while Mo *et al*. [22] employ a post-processing approach to refine audio-visual localization results using pre-trained visual feature activation maps. In our work, we leverage CLIP's multimodal alignment knowledge as a prior in a textless and fully self-supervised manner without any post-processing.

CLIP in Audio-Visual Learning. Recent contrastive language-image pretraining (CLIP) models, which are pretrained on large-scale paired data [14, 26], demonstrate robust generalization ability and have been successfully used in numerous downstream tasks across various research topics. In this section, we review related works that incorporate CLIP [26] for audio-visual learning. WAV2CLIP [36] and AudioCLIP [11] expand the pre-trained CLIP model by aligning audio features with text and visual features in a shared embedding space, *i.e.* representation learning. They achieve this either using paired data or by utilizing the visual modality as a bridge. Beyond representation learning, CLIP models are also employed in audio-visual event localization [20] and video parsing [9], as well as audio-visual source separation [8, 34]. While [34] employs text input for separation, CLIPSep [8] is trained based on the audio-visual relationship without text query. Similarly, our proposed method is also trained solely with an audio-visual alignment objective. Another line of work [2, 39] adapt pre-trained CLIP models and text encoders for audio. They achieve this by mimicking contextual text tokens using audio signals, enabling the CLIP text encoder to embed audio signals. Our work also employs a similar approach to leverage the CLIP model without text input for the sound localization task.

3. Method

3.1. Audio-Driven Embedder

Our goal is to use the CLIP text encoder to embed audios without any text input. We employ the Audio Tokenizer module for this purpose, which transforms audio
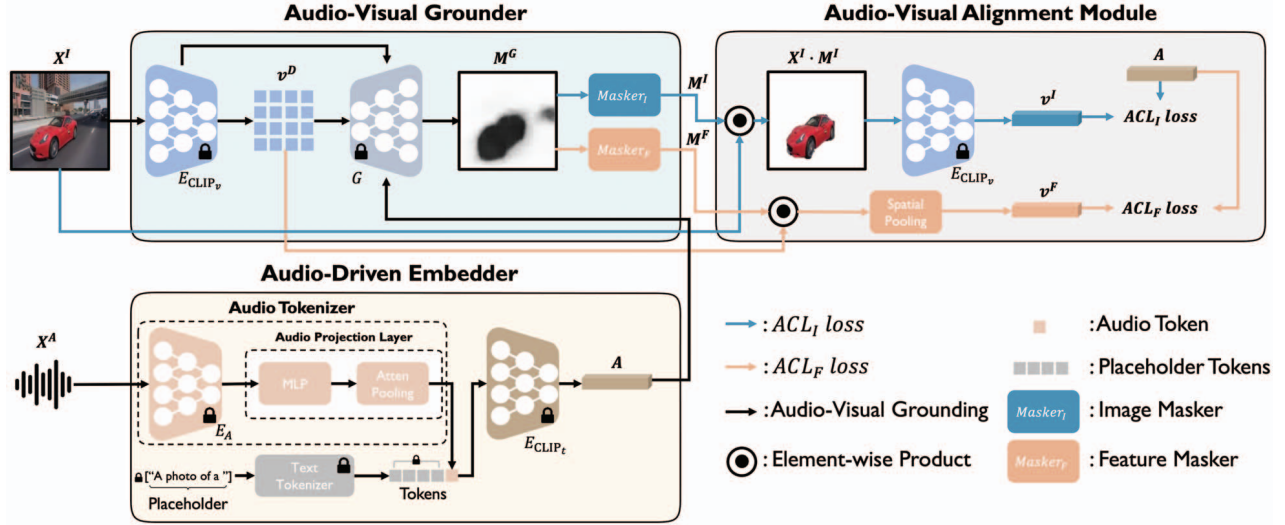
Figure 2. Our sound source localization framework. The proposed method takes audio-visual pairs, translating audio signals into CLIP-compatible tokens via the Audio Tokenizer module to generate audio-driven embedding, $\mathbf{A}$. This embedding highlights sounding regions within the Audio-Visual Grounder module. With the sounding area masks, the Audio-Visual Alignment module extracts audio-grounded visual features at both image-level ($\boldsymbol{v}^I$) and feature-level ($\boldsymbol{v}^F$). These visual features and audio feature are aligned via contrastive learning.

context into text-like tokens. In essence, an audio segment is translated into a word token, which can subsequently undergo processing by the pre-trained CLIP text encoder. The module has two key components: an **audio encoder** and a **projection network**. The projection network contains two MLP layers and one attentive pooling layer, similar to [39]. While the audio encoder is pre-trained and fixed during training, the remaining layers are trained end-to-end in our sound source localization approach with the objective of audio-visual alignment.

Audio Encoder, E_A, is a transformer-based network pre-trained in a self-supervised manner, following [6]. It takes an audio spectrogram and extracts audio embeddings. Once the audio embedding is extracted, it undergoes processing in a small **projection network**, effectively mimicking the textual tokens through the audio. The outcome is an "audio token" aligned with textual tokens. This token is then appended to the fixed placeholder text tokens of "A photo of a" to complete the input token representation as in Figure 2. This way, the audio signal with its proper context can be fitted as an additional token for CLIP text encoder (E_{CLIP_t}). This combination of fixed placeholder text and audio tokens is processed in the pre-trained E_{CLIP_t}, and the audio-driven embedding $\mathbf{A}$ is obtained. This audio-driven embedding can then be paired or conditioned with any CLIP image encoder-based approaches as it contains the visual alignment knowledge due to E_{CLIP_t}.

3.2. Audio-Visual Grounder

For a given input batch of audio-visual pairs, which consist of images and their corresponding audios, our audio-visual grounder performs grounding to detect the regions with sound and then generates masks. These masks are subsequently utilized to extract visual embeddings at both the image-level and feature-level, which are used in the audio-visual alignment objective. Our Audio-Visual grounding module is designed with three components: 1) an image encoder, 2) a grounder, and 3) mask generators.

We use a pre-trained CLIP image encoder as our image encoder, denoted as E_{CLIP_v}. It is responsible for encoding the provided input images into both global features and spatial features. For our grounder, G, we employ off-the-shelf CLIP-based segmentation network known as CLIPSeg [19]. It is important to note that CLIPSeg requires CLIP-based visual features and text conditioning to perform segmentation. We leverage the outputs from our image encoder as visual features for grounder. However, since our approach does not use any text input directly, we utilize our audio-driven embedding, $\mathbf{A}$, for conditioning. The result of the grounder G, $\mathbf{M}^G$, is potential sounding regions. Both the image encoder and the grounder remain fixed during training.

To obtain audio-grounded visual embeddings for the provided paired images $\mathbf{X}^I$ and audios $\mathbf{X}^A$ within the Audio-Visual Alignment module during training, it is essential to have differentiable binary masks for sounding regions. We introduce two masking methods: Image Masker ($Masker_I$) and Feature Masker ($Masker_F$), both of which serve to extract audio-grounded visual embeddings at the image-level and feature-level, respectively. Similar to [3], $Masker_I$ utilizes a learnable scalar projection ($w \cdot \mathbf{M}^G + b$) on the output of the grounder, $\mathbf{M}^G$, and then applies the Gumbel-Max technique [13] to generate a differentiable binary mask, referred to as $\mathbf{M}^I$. This mask is used to identify sounding areas in the image. $Masker_F$ is designed with min-

max normalization and soft-thresholding functions applied to $\mathbf{M}^G$ to obtain $\mathbf{M}^F$, which allows the extraction of audio-visually correlated areas at the feature level. The utilization of these maskers is explained in the following section.

3.3. Audio-Visual Alignment

After obtaining sounding area masks for the given audio from the audio-visual grounder, our method extracts visual embeddings from the masked areas at both the image-level and feature-level, aligning them with the audio-driven embedding, $\mathbf{A}$, for the audio-visual alignment objective. For this purpose, we define two contrastive learning losses: image-level and feature-level audio-grounded contrastive losses, ACL_I and ACL_F respectively. In a nutshell, our model learns to maximize the alignment between the visual features of sounding regions and audio features.

Image-Level Audio-Grounded Contrastive Loss. Different from typical global image and audio correspondence, our focus is on alignment between sounding region and audio. One approach to achieve this is by highlighting the sounding regions (foreground pixels) in the image and masking out the background areas, as depicted in Figure 2. To begin, the mask $\mathbf{M}_i^I$ obtained from $Masker_I$ for audio-visual pair of ith clip to mask out the irrelevant areas in the image. Our image-level audio-grounded contrastive loss, ACL_I, consists of CLIP image encoder E_{CLIP_v}. This masked image is then transformed into a visual embedding, $\boldsymbol{v}_i^I = E_{\text{CLIP}_v}\left(\mathbf{M}_i^I \cdot \mathbf{X}_i^I\right)$. The audio-visual similarity between the audio-driven embedding $\mathbf{A}_j$ from jth clip and the audio-grounded visual embedding $\boldsymbol{v}_i^I$ is computed using cosine similarity and defined as $S_{i,j}^I = (\boldsymbol{v}_i^I \cdot \mathbf{A}_j)$. We employ symmetric InfoNCE for the contrastive loss. We note that image-level masks are computed only for positive pairs. Thus, the objective of this loss is to maximize the similarity between the positive sounding region and the corresponding audio pair, while also ensuring dissimilarity between negative audios and the actual sounding region. The ACL_I loss is defined as follows:

$$
\begin{aligned}
\mathcal{L}_{ACL_I} &= \textit{InfoNCE}(\mathbf{S}^I) \\
&= -\frac{1}{2B}\sum_i^B \log \frac{\exp(S_{i,i}^I/\tau)}{\sum_j^B \exp(S_{i,j}^I/\tau)} \\
&\quad - \frac{1}{2B}\sum_i^B \log \frac{\exp(S_{i,i}^I/\tau)}{\sum_j^B \exp(S_{j,i}^I/\tau)}
\end{aligned} \quad (1)
$$

where τ is the temperature parameter and $\mathbf{S}^I$ is image-level audio-visual similarity matrix within batch. With the help of this loss, the sounding region and the generated mask $\mathbf{M}^I$ gradually cover the target sounding area. However, we observe that ACL_I alone can not enable the model to completely suppress the background regions.

Feature-Level Audio-Grounded Contrastive Loss. Suppressing masks derived from negative pairs is essential for enhancing robustness against background regions. However, due to memory constraints, generating high-resolution image-level masks for all negative pair combinations within a batch is infeasible. As an alternative, we introduce the feature-level audio-grounded contrastive loss, ACL_F, allowing the use of masks in lower-resolution (on features), effectively bypassing the memory constraints. A strategic approach involves emphasizing regions within the spatial visual features, as shown in Figure 2. To elaborate, the mask $\mathbf{M}_{i,j}^F \in \mathbb{R}^{h \times w}$ obtained from the $Masker_F$ for given image $\mathbf{X}_i^I$ and audio $\mathbf{X}_j^A$ is applied during spatial pooling of the spatial visual features $\boldsymbol{v}_i^D \in \mathbb{R}^{c \times h \times w}$ to focus on regions within the features that exhibit high correlation with the paired audio. Feature-level audio-grounded visual embedding $\boldsymbol{v}_{i,j}^F \in \mathbb{R}^c$ is as follows:

$$
\boldsymbol{v}_{i,j}^F = \frac{\sum_{h,w}\mathbf{M}_{i,j,h,w}^F \cdot \boldsymbol{v}_{i,h,w}^D}{\sum_{h,w}\mathbf{M}_{i,j,h,w}^F}. \quad (2)
$$

In contrast to ACL_I, which focuses on the sounding region, ACL_F focuses on the highly correlated area, regardless of positive or negative audio-visual pairs. The audio-visual similarity between the audio-driven embedding $\mathbf{A}$ and the feature-level audio-grounded visual embedding $\boldsymbol{v}^F$ for both positive and negative pairs is computed using cosine similarity defined as $S_{i,j}^F = (\boldsymbol{v}_{i,j}^F \cdot \mathbf{A}_j)$. The ACL_F loss is defined as follows:

$$
\mathcal{L}_{ACL_F} = \textit{InfoNCE}(\mathbf{S}^F), \quad (3)
$$

where $\mathbf{S}^F$ is feature-level audio-visual similarity matrix within batch. While it is possible to replace the mask $\mathbf{M}^F$ with $\mathbf{M}^I$ in Equation 2, this may lead to unintended training. The reason is that $\mathbf{M}^I$ may generate a mask that is close to a zero matrix when dealing with negative pairs. This can result in the numerator of Equation 2 effectively being zero, making $\boldsymbol{v}_{i,j}^F$ arbitrary. To simplify, this replacement may cause the InfoNCE loss to generate random similarity scores for negative pairs.

3.4. Area Regularization

We observe that even using ACL_I and ACL_F losses during training, the model can take a shortcut and output masks that contain both irrelevant and sounding regions, such as the entire image. In this case, the CLIP image encoder in the Audio-Visual Alignment module can still generate relevant visual features. Therefore, similar to [3, 37], we formulate an area regularizer loss, as defined below:

$$
\mathcal{L}_{Reg} = \sum_i \|p^+ - \overline{\mathbf{M}_{i,i}^I}\|_1 + \sum_{i \neq j} \|p^- - \overline{\mathbf{M}_{i,j}^I}\|_1, \quad (4)
$$

where $\mathbf{M}_{i,i}^I$ and $\mathbf{M}_{i,j}^I$ are the image masks from the positive and the negative pairs respectively. The area of these masks are denoted as $\overline{\mathbf{M}}$. p^+ and p^- represent the area prior hyperparameters, which are set to 0.4 and 0.0. The area regularizer constrains the size of the mask during learning to ensure that the intended sounding regions are contained while irrelevant areas are discarded.

3.5. Training

The overall training loss term is defined as follows:

$$\mathcal{L} = \lambda_{ACL_I}\mathcal{L}_{ACL_I} + \lambda_{ACL_F}\mathcal{L}_{ACL_F} + \lambda_{REG}\mathcal{L}_{REG}, \quad (5)$$

where λ_{ACL_I}, λ_{ACL_F}, and λ_{REG} are the hyper-parameters weighting the loss terms.

3.6. Inference

For the provided image and audio pairs, an audio-driven embedding is acquired and fed into the grounder G along with the visual features obtained from the image encoder. The resulting output of the grounder, $\mathbf{M}^G$, is subsequently used in $Masker_I$. Unlike training, during inference, it is adjusted using $\sigma\left(\mathbf{M}^G + b/w\right)$, where w, b are scalar projection parameters learned during training in the image masker $Masker_I$ and σ is sigmoid function. The final output mask is then thresholded using the hyperparameter t to obtain the localization result.

4. Experiments

Datasets. Our approach is trained using the VGGSound dataset [5], comprising around ~200K videos. After training, we evaluate sound localization performance on VGG-SS [4] and SoundNet-Flickr-Test [27, 28] datasets. These evaluation sets provide bounding box annotations for sound sources, totaling about 5K and 250 samples, respectively. Further evaluations are conducted using AVS-Bench [40] and Extended VGG-SS/SoundNet-Flickr [21] datasets. AVSBench includes binary segmentation maps indicating audio-visually related pixels and is divided into Single-source (S4) and Multi-sources (MS3) subsets, categorized by the number of sounding objects. These subsets contain around 5K samples in (S4) and about 400 samples in (MS3). Lastly, the Extended VGG-SS/SoundNet-Flickr datasets proposed by [21] are used to explore non-visible sound sources.

Implementation details. We employ frozen pre-trained "ViT-B/16" CLIP [26] model as image encoder, BEATs [6] for audio encoder and CLIPSeg [19] for grounder. During training, we used 10-second audio segments sampled at 16kHz, and the center frame of the video resized to 352x352. For the overall loss, we set the parameters λ_{ACL_I}, λ_{ACL_F}, and λ_{Reg} all to 1. Additionally, we used τ as 0.07

in Equation 1. The model is optimized for 20 epochs with a batch size of 16, using the Adam optimizer with a learning rate of 10^{-3} and a weight decay of 10^{-3}.

4.1. Quantitative Results

Baselines. Besides the existing works, we also compare our proposed method with closely-related baselines that can be obtained using different components of our overall architecture. The details of these baselines are introduced below:

- **CLIPSeg w/ GT Text.** We utilize the ground truth class labels of test samples as text conditions to obtain the segmentation results from CLIPSeg, essentially serves as an oracle method.

- **CLIPSeg w/ WAV2CLIP Text.** WAV2CLIP aligns text, vision, and audio embeddings together in the CLIP space. For a given audio, the most relevant text (class label) can be retrieved. This retrieved text is used with CLIPSeg to highlight the sounding region in the image.

- **CLIPSeg - Sup. AudioTokenizer** We train AudioTokenizer module in a supervised manner rather than in a self-supervised way like our proposed model. The predicted audio-driven embedding is supervised using the corresponding GT text $\mathbf{X}^T$ of each sample directly with $\mathcal{L}_{Sup} = \|E_{\mathrm{CLIP}_T}(\mathbf{X}^T) - \mathbf{A}\|_1$. The audio-driven embeddings obtained from this model are used with zero-shot CLIPSeg to obtain sound localization results.

- **WAV2CLIP and AudioCLIP.** These models leverage the pre-trained CLIP model to align text, vision, and audio embeddings. To enable zero-shot sound source localization with these models, we utilize a pre-trained CLIP-like object detector [15] to extract region proposals from the images and calculate the cosine similarity between the visual features of those regions and the audio features. The region with the highest similarity is employed as the localization result.

Comparison on standard benchmarks. In this section, we perform a comparative analysis of our method for localizing sound sources in comparison to existing approaches and the strong baselines. Our evaluations are conducted within the established setting, similar to prior methodologies [4, 22, 29, 33]. We train our model on the VGGSound-144K dataset and subsequently assess its performance on the VGG-SS and SoundNet-Flickr test sets. It is worth noting that all the models we compare are trained using equivalent amounts of data. However, note that our model does not use object guided refinement (OGL). We present our findings in Table 1.

At the outset, we compare our method with other existing sound source localization models. There is a substantial gap between the existing self-supervised methods and ours in VGG-SS evaluation task. Although our model is also

Method	VGG-SS		SoundNet-Flickr	
	cIoU ↑	AUC ↑	cIoU ↑	AUC ↑
Attention [27]_{CVPR18}	18.50	30.20	66.00	55.80
CoarseToFine [25]_{ECCV20}	29.10	34.80	-	-
LCBM [30]_{WACV22}	32.20	36.60	-	-
LVS [4]_{CVPR21}	34.40	38.20	71.90	58.20
HardPos [29]_{ICASSP22}	34.60	38.00	76.80	59.20
SSPL [32]_{CVPR22}	33.90	38.00	76.70	60.50
EZ-VSL (w/o OGL) [22]_{ECCV22}	35.96	38.20	78.31	61.74
EZ-VSL (w/ OGL) [22]_{ECCV22}	38.85	39.54	83.94	63.60
SSL-TIE [18]_{ACM MM22}	38.63	39.65	79.50	61.20
SLAVC (w/o OGL) [21]_{NeurIPS22}	37.79	39.40	83.60	-
SLAVC (w/ OGL) [21]_{NeurIPS22}	39.80	-	**86.00**	-
MarginNCE (w/o OGL) [24]_{ICASSP23}	38.25	39.06	83.94	63.20
MarginNCE (w/ OGL) [24]_{ICASSP23}	39.78	40.01	85.14	64.55
HearTheFlow [10]_{WACV23}	39.40	40.00	84.80	64.00
FNAC (w/o OGL) [33]_{CVPR23}	39.50	39.66	84.73	63.76
FNAC (w/ OGL) [33]_{CVPR23}	41.85	40.80	85.14	64.30
Alignment (w/o OGL) [31]_{ICCV23}	39.94	40.02	79.60	63.44
Alignment (w/ OGL) [31]_{ICCV23}	42.64	41.48	82.40	64.60
Baselines:				
WAV2CLIP [36]_{ICASSP22}	37.71	39.93	26.00	29.60
AudioCLIP [11]_{ICASSP22}	44.15	46.23	47.20	45.22
CLIPSeg (w/ GT Text)	49.50	48.62	-	-
CLIPSeg (w/ WAV2CLIP Text)	24.84	26.01	37.20	32.14
CLIPSeg (Sup. AudioTokenizer)	49.09	45.75	68.00	54.96
Ours (w/o OGL)	**49.46**	**46.32**	80.80	**64.62**

Table 1. **Quantitative results on the VGG-SS and SoundNet-Flickr test sets**. All models are trained with 144K samples from VGG-Sound. SLAVC [21] does not provide AUC scores. SoundNet-Flickr has no GT text.

Test Class	Method	cIoU ↑	AUC ↑
Heard 110	LVS [4]_{CVPR21}	28.90	36.20
	EZ-VSL (w/o OGL) [22]_{ECCV22}	31.86	36.19
	EZ-VSL (w/ OGL) [22]_{ECCV22}	37.25	38.97
	SLAVC (w/o OGL) [21]_{NeurIPS22}	35.84	-
	SLAVC (w/ OGL) [21]_{NeurIPS22}	38.22	-
	FNAC (w/ OGL) [33]_{CVPR23}	39.54	39.83
	Alignment (w/o OGL) [31]_{ICCV23}	38.31	39.05
	Alignment (w OGL) [31]_{ICCV23}	41.85	40.93
	CLIPSeg (w/ GT Text)	49.65	45.74
	CLIPSeg (w/ WAV2CLIP Text)	23.24	24.78
	CLIPSeg (Sup. AudioTokenizer)	49.73	45.35
	Ours (w/o OGL)	48.44	45.06
Unheard 110	LVS [4]_{CVPR21}	26.30	34.70
	EZ-VSL (w/o OGL) [22]_{ECCV22}	32.66	36.72
	EZ-VSL (w/ OGL) [22]_{ECCV22}	39.57	39.60
	SLAVC (w/o OGL) [21]_{NeurIPS22}	36.50	-
	SLAVC (w/ OGL) [21]_{NeurIPS22}	38.87	-
	FNAC (w/ OGL) [33]_{CVPR23}	42.91	41.17
	Alignment (w/o OGL) [31]_{ICCV23}	39.11	39.80
	Alignment (w OGL) [31]_{ICCV23}	42.94	41.54
	CLIPSeg (w/ GT Text)	49.13	44.77
	CLIPSeg (w/ WAV2CLIP Text)	26.25	27.03
	CLIPSeg (Sup. AudioTokenizer)	43.65	41.05
	Ours (w/o OGL)	41.98	41.55

Table 2. **Comparison results on open-set audio-visual localization experiments trained and tested on the splits of [4, 22, 24].**

purely trained in a self-supervised manner with the audio-visual correspondence objective, it is evident that leveraging CLIP's strong multimodal alignment knowledge significantly impacts the performance. However, note that even though we leverage CLIP, we do not employ any explicit text input. These results thus demonstrate that our AudioTokenizer module effectively encodes the audio context, enabling proper learning of the audio-visual correspondence objective. Interestingly, we observe that the zero-shot performance of our model on the SoundNet-Flickr test lags behind that of the existing models. We hypothesize that this result stems from the fact that our model generates more fine-grained outputs, resembling segmentation. Nonetheless, the ground-truth bounding boxes are relatively coarse, causing our method to yield lower cIoU scores despite successfully highlighting the sounding region. We provide some illustrative qualitative results for this in Section 4.3.

Next, we conduct comparisons against the strong baselines introduced earlier. Our method outperforms or achieves on-par performance with these baselines. It is worth noting that our method does not explicitly utilize text information to highlight object regions via CLIPSeg or learn audio-driven embeddings in a supervised fashion, as done by these baselines. This indicates that our audio-visual correspondence objective effectively learns robust audio-visual correspondence and drives the AudioTokenizer and Audio-Driven Embedder to accurately project the true audio context into audio-driven embeddings. Interestingly,

our model gives on-par performance with the CLIPSeg w/ GT Text baseline on VGG-SS, which serves as an Oracle. This model is text-conditioned open-world segmentation approach and utilizes the ground-truth class labels of the test samples. This signifies that it is important to incorporate the audio context properly to enhance performance. Additionally, the performance difference between CLIPSeg w/ GT Text and CLIPSeg w/ WAV2CLIP highlights that the text-queried zero-shot performance of CLIPSeg in sound source localization is highly dependent on the quality of the text input. This is due to the fact that the text retrieved from WAV2CLIP for given audio tends to be noisier compared to GT text. Nevertheless, it is important to note that sound source localization is unlabeled task, and these methods serve as Oracle baselines. Furthermore, the results demonstrate that training the AudioTokenizer in a supervised way with GT texts and employing audio-driven embeddings with CLIPSeg also gives on-par performance to the Oracle. This implies that audio-driven embeddings indeed provide accurate information for highlighting the sounding regions. Finally, we acknowledge that our method, which employs an audio-visual correspondence objective for self-supervised learning, outperforms CLIPSeg - Sup. AudioTokenizer. This suggests that our Audio-Visual alignment contrastive losses offer effective supervision for the model as using explicit text input, compelling the AudioTokenizer module to generate richer audio-driven embeddings.

Finally, we compare our model with AudioCLIP and WAV2CLIP, both of which are contrastively trained on image-audio pairs, leveraging the pre-trained CLIP. The results in Table 1 demonstrate that our method outperforms these approaches. This indicates that our Audio-Driven Embedder module, with the audio-visual alignment objective, is more effective in learning a stronger audio-visual alignment than these previous approaches, as they also leverage

Method	Extended VGG-SS			Extended Flickr		
	AP ↑	max-F1 ↑	LocAcc ↑	AP ↑	max-F1 ↑	LocAcc ↑
SLAVC (w/o OGL) [21]NeurIPS22	32.95	40.00	37.79	51.63	59.10	83.60
MarginNCE (w/o OGL) [24]ICASSP23	30.58	36.80	38.25	57.99	61.80	83.94
FNAC (w/o OGL) [33]CVPR23	23.48	33.70	39.50	50.40	62.30	**84.73**
Alignment (w/o OGL) [31]ICCV23	34.73	40.70	39.94	64.43	66.90	79.60
WAV2CLIP [36]ICASSP22	26.67	33.00	37.71	20.99	24.80	29.60
AudioCLIP [11]ICASSP22	23.79	32.80	44.15	34.00	38.80	45.22
CLIPSeg (Sup. AudioTokenizer)	34.96	41.00	49.09	55.14	57.00	68.00
Ours (w/o OGL)	**40.79**	**49.10**	**49.46**	**76.07**	**73.20**	80.80

Table 3. **Quantitative results on Extended VGG-SS and Extended Flickr-SoundNet benchmark.** All models are trained with 144K samples from VGG-Sound. The results of the prior approaches are obtained from [21].

pre-trained CLIP knowledge. Additionally, note that these baseline approaches incorporate powerful object detectors to obtain object proposals/areas that correspond with the given audio, in order to achieve sound localization results.

Open Set Audio-Visual Localization. Chen et al. [4] propose an open-set evaluation scenario to assess the generalization ability of sound source localization methods. This evaluation setting involves testing the models on categories present in the training data (heard), as well as categories that are absent (unheard). For this evaluation, 110 randomly selected categories from the VGGSound dataset are used for training, while an entirely separate set of 110 categories is held for testing. This ensures that the model encounters new and previously unseen categories during the evaluation process. To make a fair comparison, we conduct the experiments using the the same train/test split as [4, 22, 24]. It is important to note that unlike previous methods, we do not utilize object-guided refinement (OGL). The results are presented in Table 2, showing that our method outperforms existing approaches in the Heard categories. However, it lags behind FNAC [33] in the Unheard category, due to the usage of OGL in their method, which we do not employ.

Extended Flickr-SoundNet/VGG-SS. Existing benchmarks typically consist of sounding objects/regions in the scene. However, in reality, silent objects or off-screen audio are also common occurrences. Mo et al. [21] propose a new evaluation that extends the existing benchmarks to include non-audible frames, non-visible sound sources, and mismatched audio-visual pairs. In this evaluation scenario, it is expected that sound localization methods should not highlight an object/region if the audio and visual signals are mismatched. The experiments conducted using the extended Flickr-SoundNet/VGG-SS datasets in Table 3 demonstrate that our method outperforms all the existing methods and baselines. The superiority of our method indicates that it learns a strong alignment of audio and visual embeddings with the help of our AudioTokenizer and leveraging CLIP without text input , as this task requires a robust semantic relationship between the cross-modalities. One interesting observation is that, even though baseline approaches leverage CLIP, their performance is lower than ours due to the absence of audio-visual alignment supervision.

Method	S4		MS3	
	mIoU ↑	F-Score ↑	mIoU ↑	F-Score ↑
SLAVC (w/o OGL) [21]NeurIPS22	28.10	34.60	24.37	25.56
MarginNCE (w/o OGL) [24]ICASSP23	33.27	45.33	27.31	31.56
FNAC (w/o OGL) [33]CVPR23	27.15	31.40	21.98	22.50
Alignment (w/o OGL) [31]ICCV23	29.60	35.90	-	-
Baselines:				
WAV2CLIP [36]ICASSP22	28.70	35.35	25.09	23.84
AudioCLIP [11]ICASSP22	36.57	42.15	27.06	26.48
CLIPSeg (w/ GT Text)	51.32	58.02	**50.93**	**55.41**
CLIPSeg (w/ WAV2CLIP Text)	26.52	30.60	30.82	29.97
CLIPSeg (Sup. AudioTokenizer)	49.82	56.43	42.57	46.72
Ours (w/o OGL)	**59.76**	**69.03**	41.08	46.67

Table 4. **Quantitative results on the AVSBench test sets.**

	ACL_I	ACL_F	Reg	VGG-SS		AVS (S4)		Extended VGG-SS	
				cIoU ↑	AUC ↑	mIoU ↑	F-score ↑	AP ↑	max-F1 ↑
(A)	✓	✗	✗	40.42	40.84	38.55	45.94	28.59	35.90
(B)	✗	✓	✗	2.30	7.46	4.08	22.59	0.86	1.80
(C)	✓	✓	✗	46.61	44.71	53.06	63.01	40.72	47.90
(D)	✓	✗	✓	41.08	41.01	41.93	48.99	33.37	41.30
(E)	✗	✓	✓	35.15	38.36	32.06	41.05	39.91	47.20
(F)	✓	✓	✓	**49.46**	**46.32**	**59.76**	**69.03**	**40.79**	**49.10**

Table 5. **Ablative experiments on our method by using different combinations of loss functions.**

AVSBench [40]. We conduct additional experiments using the AVSBench S4 and MS3 datasets to demonstrate the precise sound localization ability of our model. These datasets are designed to identify audio-visual correspondences at the pixel level, *i.e.* audio-visual segmentation. In these experiments, all models are trained on VGGSound-144K and then tested on the AVSBench datasets in a zero-shot setting. Our results, presented in Table 4, draw a substantial performance gap compared to existing methods. This gap is more pronounced on audio-visual segmentation datasets than on standard benchmarks, as our model tends to generate more fine-grained localization maps due to the grounder and learnable maskers it employs. Our proposed method also demonstrates competitive or stronger performance compared to most of the baselines. While our method outperforms others on the S4 dataset, CLIPSeg w/ GT Text and CLIPSeg w/ WAV2CLIP Text (Oracles) achieve better segmentation performance on the MS3 dataset. However, we emphasize that our model does not employ any direct supervision or usage of the text. Instead, it relies solely on audio-visual alignment. Also, note that sound source localization task is theoretically unlabeled.

4.2. Ablation Results

Our proposed method is optimized by a combination of three loss functions, *i.e.* ACL_I, ACL_F, and area regularization. Here, we perform ablation experiments to understand the impact of each loss function. We primarily conduct the experiments by training our model on VGGSound-144K and evaluating it on VGG-SS, AVSBench and Extended VGG-SS datasets. Results are in Table 5.

As revealed by results (A) and (B), using ACL_I is crucial to enable our model to learn the corresponding audio-visual alignment. On the other hand, relying solely on ACL_F is not effective for learning audio-visual align-
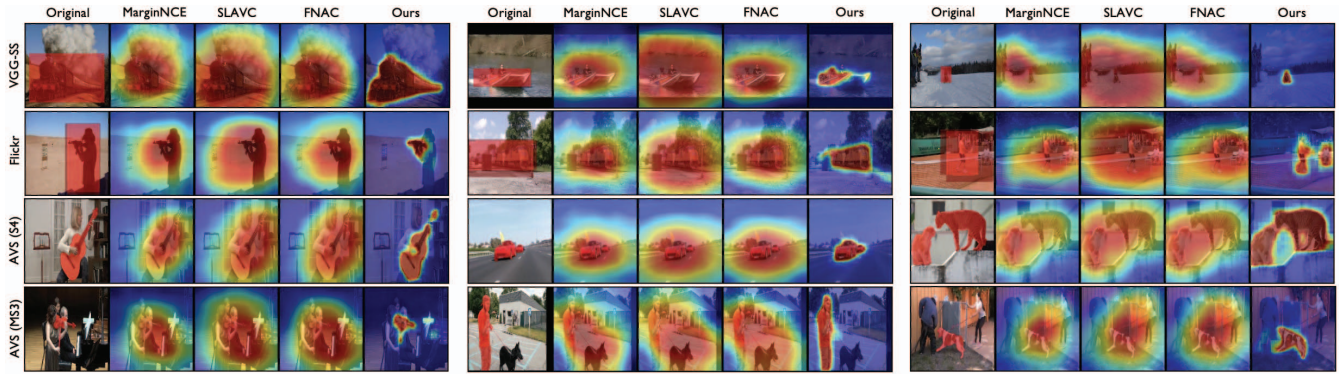
Figure 3. **Sound localization results on VGG-SS, SoundNet-Flickr, and AVSBench datasets, along with a comparison with previous methods.**
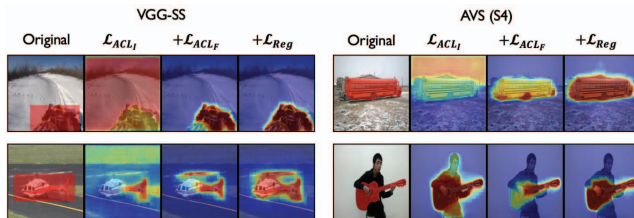

Figure 4. **Sound localization results by using different combinations of loss functions.**
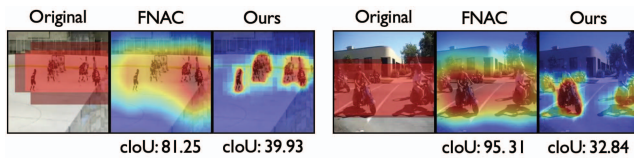

Figure 5. **cIoU scores of SoundNet-Flickr samples.**

ment, as it primarily focuses on suppressing unrelated areas. However, as demonstrated by the results of (A *vs.* C) and (B *vs.* C), the combination of these two loss functions are complementary. As mentioned earlier, ACL_F contributes to performance enhancement by suppressing background areas. Furthermore, an examination of the results from the experiments (C *vs.* F) highlights that area regularization provides additional improvements by constraining the size of the activated regions. Visualization of these ablative studies can be found in Figure 4.

4.3. Qualitative Results

Comparison to the existing approaches. Figure 3 displays the comparison results between our method and recent prior works. The visualized samples illustrate that the localized regions from our proposed method are more compact and fine-grained compared to the other methods. For example, regardless of the test set, our model can accurately localize small-sized sounding objects compared to recent methods. Moreover, our model accurately highlights multiple sound sources and separates them, while other methods tend to cover the entire area as one large region (last column of the second and third rows).

Visualization of the ablation experiments. The visual results are presented in Figure 4. As demonstrated, when using only ACL_I, we observe that background areas remain activated (also discussed in Section 3.3). As evident in the third column, the addition of ACL_F helps eliminate the background pixels (non-sounding areas). However, it is noticeable that the outputs of ACL_I+ACL_F can be relatively less completed. With the area regularizer, the final output of our model becomes more complete and fine-grained.

Visualization of fine-grained localization with lower cIoU. We present our localization results along with the cIoU scores on SoundNet-Flickr. As depicted in Figure 5, despite our model successfully highlighting the sounding regions, these results yield lower cIoU scores. This outcome is consistent with the quantitative results in Table 1, which demonstrate that our method on SoundNet-Flickr lags behind the other methods due to the fact that the GT boxes and the localization results of competing methods are coarse.

5. Conclusion

In this work, we explore using large-scale pre-trained image-text models, specifically CLIP, for sound source localization. Our aim is to integrate CLIP's multimodal alignment knowledge in a text input-free form through self-supervised audio-visual correspondence. To this end, we translate audio signals into CLIP-compatible tokens and use the resulting audio-driven embeddings for audio-visual grounding. This process is integrated with contrastive learning, enabling self-supervised audio-visual alignment learning. We show that our proposed model significantly outperforms existing methods in audio-visual coarse sound source localization and fine-grained segmentation tasks. Moreover, it compares favorably with fully supervised or text-queried baselines. Our study suggests that the true essence of sound source localization, characterized by strong audio-visual alignment, can take advantage from the already structured multimodal alignment offered by large-scale pre-trained image-text models.

References

[1] Relja Arandjelović and Andrew Zisserman. Objects that sound. In *European Conference on Computer Vision*, 2018. 1

[2] Saurabhchand Bhati, Jesús Villalba, Laureano Moro-Velazquez, Thomas Thebaud, and Najim Dehak. Segmental speechclip: Utilizing pretrained image-text models for audio-visual learning. In *INTERSPEECH*, 2023. 2

[3] Junbum Cha, Jonghwan Mun, and Byungseok Roh. Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 3, 4

[4] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 1, 2, 5, 6, 7

[5] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2020. 5

[6] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. BEATs: Audio pretraining with acoustic tokenizers. In *International Conference on Machine Learning*, 2023. 3, 5

[7] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2

[8] Hao-Wen Dong, Naoya Takahashi, Yuki Mitsufuji, Julian McAuley, and Taylor Berg-Kirkpatrick. Clipsep: Learning text-queried sound separation with noisy unlabeled videos. In *International Conference on Learning Representations*, 2022. 2

[9] Yingying Fan, Yu Wu, Yutian Lin, and Bo Du. Revisit weakly-supervised audio-visual video parsing from the language perspective. *arXiv preprint arXiv:2306.00595*, 2023. 2

[10] Dennis Fedorishin, Deen Dayal Mohan, Bhavin Jawade, Srirangaraj Setlur, and Venu Govindaraju. Hear the flow: Optical flow-based self-supervised visual sound source localization. In *IEEE Winter Conf. on Applications of Computer Vision*, 2023. 1, 6

[11] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. AudioCLIP: Extending CLIP to image, text and audio. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2022. 2, 6, 7

[12] 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. *Advances in Neural Information Processing Systems, NeurIPS*, 2020. 1

[13] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*, 2017. 3

[14] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, 2021. 2

[15] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 5

[16] Sizhe Li, Yapeng Tian, and Chenliang Xu. Space-time memory network for sounding object localization in videos. In *British Machine Vision Conference*, 2021. 1

[17] Yan-Bo Lin, Hung-Yu Tseng, Hsin-Ying Lee, Yen-Yu Lin, and Ming-Hsuan Yang. Unsupervised sound localization via iterative contrastive learning. *Computer Vision and Image Understanding*, 2023. 1, 2

[18] Jinxiang Liu, Chen Ju, Weidi Xie, and Ya Zhang. Exploiting transformation invariance and equivariance for self-supervised sound localisation. In *ACM International Conference on Multimedia*, 2022. 1, 2, 6

[19] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 3, 5

[20] Tanvir Mahmud and Diana Marculescu. Ave-clip: Audioclip-based multi-window temporal transformer for audio visual event localization. In *IEEE Winter Conf. on Applications of Computer Vision*, 2023. 2

[21] Shentong Mo and Pedro Morgado. A closer look at weakly-supervised audio-visual source localization. *Advances in Neural Information Processing Systems, NeurIPS*, 2022. 1, 2, 5, 6, 7

[22] Shentong Mo and Pedro Morgado. Localizing visual sounds the easy way. In *European Conference on Computer Vision*, 2022. 1, 2, 5, 6, 7

[23] Takashi Oya, Shohei Iwase, Ryota Natsume, Takahiro Itazuri, Shugo Yamaguchi, and Shigeo Morishima. Do we need sound for sound source localization? In *Asia Conference on Computer Vision*, 2020. 1

[24] Sooyoung Park, Arda Senocak, and Joon Son Chung. MarginNCE: Robust sound localization with a negative margin. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2023. 1, 2, 6, 7

[25] Rui Qian, Di Hu, Heinrich Dinkel, Mengyue Wu, Ning Xu, and Weiyao Lin. Multiple sound sources localization from coarse to fine. In *European Conference on Computer Vision*, 2020. 1, 2, 6

[26] 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*, 2021. 1, 2, 5

[27] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2, 5, 6

[28] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound

sources in visual scenes: Analysis and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5):1605–1619, 2021. 1, 2, 5

[29] Arda Senocak, Hyeonggon Ryu, Junsik Kim, and In So Kweon. Learning sound localization better from semantically similar samples. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2022. 1, 2, 5, 6

[30] Arda Senocak, Hyeonggon Ryu, Junsik Kim, and In So Kweon. Less can be more: Sound source localization with a classification model. In *IEEE Winter Conf. on Applications of Computer Vision*, 2022. 1, 2, 6

[31] 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 *IEEE International Conference on Computer Vision*, 2023. 1, 6, 7

[32] Zengjie Song, Yuxi Wang, Junsong Fan, Tieniu Tan, and Zhaoxiang Zhang. Self-supervised predictive learning: A negative-free method for sound source localization in visual scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2, 6

[33] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 1, 2, 5, 6, 7

[34] Reuben Tan, Arijit Ray, Andrea Burns, Bryan A Plummer, Justin Salamon, Oriol Nieto, Bryan Russell, and Kate Saenko. Language-guided audio-visual source separation via trimodal consistency. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 2

[35] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *European Conference on Computer Vision*, 2018. 2

[36] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2CLIP: Learning robust audio representations from CLIP. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2022. 2, 6, 7

[37] Jinheng Xie, Xianxu Hou, Kai Ye, and Linlin Shen. Clims: Cross language image matching for weakly supervised semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 4

[38] Hanyu Xuan, Zhiliang Wu, Jian Yang, Yan Yan, and Xavier Alameda-Pineda. A proposal-based paradigm for self-supervised sound source localization in videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2

[39] Guy Yariv, Itai Gat, Lior Wolf, Yossi Adi, and Idan Schwartz. AudioToken: Adaptation of text-conditioned diffusion models for audio-to-image generation. In *INTERSPEECH*, 2023. 2, 3

[40] Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio-visual segmentation. In *European Conference on Computer Vision*, 2022. 5, 7