International Journal of Computer Vision (2026) 134:179
https://doi.org/10.1007/s11263-025-02687-x

# Hearing and Seeing Through CLIP: A Framework for Self-Supervised Sound Source Localization

**Sooyoung Park**[1,2] · **Arda Senocak**[3] · **Joon Son Chung**[1]

Received: 3 May 2025 / Accepted: 17 September 2025


**Abstract**

Large-scale vision-language models demonstrate strong multimodal alignment and generalization across diverse tasks. Among them, CLIP stands out as one of the most successful approaches. In this work, we extend the application of CLIP to sound source localization, proposing a self-supervised method operates without explicit text input. We introduce a framework that maps audios into tokens compatible with CLIP's text encoder, producing audio-driven embeddings. These embeddings are used to generate sounding region masks, from which visual features are extracted and aligned with the audio embeddings through a contrastive audio-visual correspondence objective. Our findings show that alignment knowledge of pre-trained multimodal foundation model enables our method to generate more complete and compact localization for sounding objects. We further propose an LLM-guided extension that distills object-aware audio-visual scene understanding into the model during training to enhance alignment. Extensive experiments across five diverse tasks demonstrate that our method, in all variants, outperforms state-of-the-art approaches and achieves strong generalization in zero-shot settings.

**Keywords** Audio-visual Learning · Sound Source Localization · Discriminative Learning

## 1 Introduction.

Localizing sound sources is a fundamental aspect of human and animal perception, allowing us to understand and navigate complex environments by integrating auditory and visual cues. Inspired by this capability, audio-visual sound source localization has gained significant attention in recent years (Senocak et al., 2018; Arandjelović & Zisserman, 2018; Senocak et al., 2021; Qian et al., 2020; Hu et al., 2020; Chen et al., 2021; Lin et al., 2023; Li et al., 2021; Senocak et al., 2022b; Song et al., 2022; Senocak et al., 2022a; Liu et al., 2022; Mo & Morgado, 2022b; Park et al., 2023; Mo & Morgado, 2022a; Sun et al., 2023; Senocak et al., 2023). A main challenge in this field is enabling machines to identify which visual entities are producing sounds without explicit supervision. To tackle this, many approaches leverage the natural semantic alignment between audio and visual modalities. Among them, self-supervised contrastive learning has emerged as a dominant paradigm, aligning representations of paired audio and visual inputs to learn robust, generalizable cross-modal embeddings without the need for labeled data.

While many sound source localization methods are built on the principle of learning audio-visual semantic correspondence, some incorporate additional priors to guide localization. These include visual objectness cues (Mo & Morgado, 2022b, a), object proposal networks (Xuan et al., 2022), and auxiliary modalities such as optical flow (Fedorishin et al., 2023). Although such heuristics may help localization performance, they often fail to improve true semantic alignment (Senocak et al., 2023, 2025) and can introduce biases, leading models to exploit shortcuts that undermine meaningful cross-modal reasoning (Oya et al., 2020; Mo &

Communicated by Zhixiang Wang.

Sooyoung Park and Arda Senocak: These authors contributed equally to this work.

✉ Joon Son Chung
joonsc@kaist.ac.kr

Sooyoung Park
sooyoung@etri.re.kr

Arda Senocak
arda.senocak@gmail.com

[1] School of Electrical Engineering, KAIST, Daejeon, South Korea

[2] Media Coding Research Section, ETRI, Daejeon, South Korea

[3] Graduate School of Artificial Intelligence, UNIST, Ulsan, South Korea

Published online: 09 March 2026

Morgado, 2022a; Arandjelović & Zisserman, 2018; Senocak et al., 2023, 2025). In contrast, our approach enhances audio-visual alignment by leveraging strong cross-modal priors learned from large-scale multimodal data. We employ the CLIP model (Radford et al., 2021), known for its robust alignment of visual content with natural language and its rich, flexible embedding space that encodes broad semantic information across modalities with strong descriptive power. We build our self-supervised sound source localization method on top of this pretrained alignment knowledge.

Most frameworks that leverage the CLIP model incorporate text prompts. However, in this work, we explore an approach that does not rely on explicit contextual text input. This decision is motivated by several key observations: (1) sound source localization datasets, such as SoundNet-Flickr, do not provide paired textual annotations; (2) the task itself is inherently unlabeled and audio-driven; and (3) a principled approach to sound source localization should rely on learning semantic alignment between audio and visual modalities through self-supervision, without requiring class labels. Methods that depend on class-label-derived text inputs are not universally applicable, as such labels are often unavailable in real-world, in-the-wild scenarios. Therefore, we adopt a core framework that employs the pre-trained CLIP model in a configuration that does not use class-label textual input, but instead relies solely on audio-driven inputs (as illustrated in Fig. 1), using audio-visual correspondence as the supervisory signal.

We propose the following pipeline as our core framework: First, we translate audio signals into tokens compatible with CLIP's text encoder, producing contextual audio-driven embeddings. These embeddings are then aligned with visual features using contrastive learning. Specifically, we use the audio-driven embeddings to highlight sounding regions in the visual scene and extract visual features from those regions at both the image and feature levels. These visual features are subsequently aligned with the audio-driven embeddings through an audio-visual correspondence objective. The entire model is trained end-to-end within this self-supervised contrastive learning framework.

We extensively evaluate our method across diverse set of sound source localization tasks – including single-source localization, audio-visual robustness, segmentation, interactive localization, and multi-source localization – on various datasets. Our experiments consistently show that the proposed approach outperforms existing methods by a significant margin across all benchmarks. These results highlight the potential of our method as a generalizable model in zero-shot settings.

As an extension of our core framework, we propose an alternative learning objective that enhances audio-visual correspondence by distilling object-aware scene understanding from a Large Language Model (LLM) via CLIP's text encoder, used only during training. Built on CLIP, our framework remains versatile and can incorporate text embeddings when auxiliary text inputs are available. This variant is designed to operate automatically on in-the-wild samples without requiring ground-truth class labels or manually annotated captions, making it broadly applicable to real-world scenarios where labeled metadata is unavailable. This alternative approach further improves our state-of-the-art results across various tasks.

We summarize our contributions as follows:

- We propose a self-supervised sound source localization framework that leverages CLIP without requiring direct textual prompts or class labels.
- We introduce an AudioTokenizer module that maps audio into tokens compatible with CLIP's text encoder, enabling effective audio-visual alignment.
- Our method achieves strong zero-shot generalization across five diverse tasks and consistently outperforms existing approaches.
- We extend our framework with an LLM-based training objective that distills object-aware audio-visual scene understanding to further boost performance.

## 2 Related work

Sound source localization. Audio-visual sound source localization aims to detect sound-emitting regions, objects, or events within a visual scene by understanding the semantic correspondences between audio and visual modalities. Initial studies (Senocak et al., 2018; Arandjelović & Zisserman, 2018; Senocak et al., 2021) investigated this problem by employing techniques such as cross-modal attention and contrastive learning to achieve effective alignment between the two modalities. However, in real-world scenarios, audio-visual pairs may be imperfectly aligned due to factors such as background noise, off-screen sound sources, or silent visual objects, all of which can introduce misleading associations and hinder accurate localization. To mitigate these challenges, previous approaches have proposed techniques such as false negative-aware learning (Sun et al., 2023), negative-free predictive learning (Song et al., 2022), and various regularization strategies in the contrastive learning formulation (Park et al., 2023; Mo & Morgado, 2022a). Another line of research in self-supervised contrastive learning focuses on enhancing audio-visual alignment by leveraging training data through effective strategies, such as sample mining (Senocak et al., 2022a; Chen et al., 2021), geometric equivalence learning (Liu et al., 2022), and multi-positive contrastive learning (Senocak et al., 2023, 2025). Furthermore, several sound localization methods explore using additional prior knowledge or post-processing techniques. For exam-

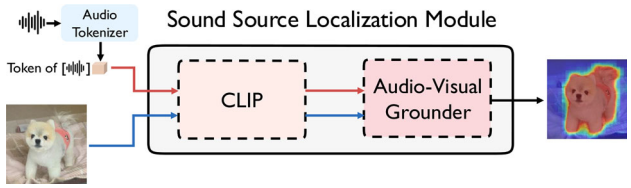


**Fig. 1** The proposed CLIP based sound source localization method

ple, Senocak et al. (2022b); Qian et al. (2020) incorporate label information to supervise the training of backbone audio and visual networks or to enhance audio-visual alignment. Xuan et al. (2022) uses object priors in the form of object proposals, while Mo and Morgado (2022b) applies a post-processing that refines localization outputs using pre-trained visual feature activation maps. In contrast, our approach – while adopting a fully self-supervised framework like existing methods – leverages CLIP's multimodal alignment knowledge as a prior to achieve robust audio-visual alignment.

Another important research direction in visual sound source localization is multi-source sound localization (Hu et al., 2020; Qian et al., 2020; Hu et al., 2022a; Mo & Tian, 2023; Mahmud et al., 2024; Kim et al., 2024), where multiple sound-emitting objects are present within a scene. These works primarily focus on localizing generic sound sources. More recent studies (Hamilton et al., 2024; Ryu et al., 2025) extend this direction by aiming to localize both generic sounds and spoken utterances from audio mixtures. In contrast, our method is designed for single sound source localization and is trained specifically for that objective. Nonetheless, due to its strong audio-visual alignment capabilities, it can be directly applied to multi-source generic sound localization tasks too.

CLIP in Audio-Visual Learning. Recent contrastive language-image pretraining (CLIP) models, which are pretrained on large-scale paired data (Radford et al., 2021; Jia et al., 2021), 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 (Radford et al., 2021) for audio-visual learning. WAV2CLIP (Wu et al., 2022) and AudioCLIP (Guzhov et al., 2022) 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 (Mahmud & Marculescu, 2023) and video parsing (Fan et al., 2023), as well as audio-visual source separation (Tan et al., 2023; Dong et al., 2022). While (Tan et al., 2023) employs text input for separation, CLIPSep (Dong et al., 2022) is trained based on the audio-visual relationship without text. Similarly, our proposed method is also trained solely with an audio-visual alignment objective. Another line of work (Yariv et al., 2023; Bhati et al., 2023) 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 with audio-driven input for the sound localization task.

More recently, T-VSL (Mahmud et al., 2024) proposed a text-guided multi-source sound localization framework based on the tri-modal joint embedding model, AudioCLIP. While their work also addresses sound source localization, our approach differs fundamentally in its design. Instead of relying on AudioCLIP, which operates within a tri-modal embedding space, we build upon the original CLIP model which does not include an audio modality. To use audio, we introduce a novel AudioTokenizer module that effectively mimics CLIP's textual embedding space. This design enables our model to directly align audio embeddings with visual features, without requiring explicit text embeddings during either training or inference.

This work extends Park et al. (2024) by introducing an enhanced training framework with an LLM-based objective that distills object-aware audio-visual scene understanding, conducting extensive experiments across five sound localization tasks to demonstrate the method's generalizability in zero-shot settings, and providing analysis showing AudioTokenizer enables learning semantically descriptive audio embeddings.

## 3 Method

### 3.1 Audio-Driven Embedder

Our goal is to use the CLIP text encoder to embed audio without requiring any text input. To this end, we introduce the **AudioTokenizer** module, which transforms audio into text-like tokens that can be directly processed by CLIP's pre-trained text encoder. Conceptually, the module translates audios into discrete embeddings that mimic natural language tokens. These 'audio tokens' serve as a bridge between the audio modality and CLIP's language space, enabling us to leverage CLIP's powerful multimodal alignment capabilities without relying on explicit textual annotations.

The **AudioTokenizer** consists of two main components: an **audio encoder** and a **projection network**. The audio encoder, $E_A$, is a transformer-based model pre-trained in a self-supervised manner, following Chen et al. (2023). It takes a spectrogram and produces audio embeddings that capture the semantic content of the sound. These embeddings are

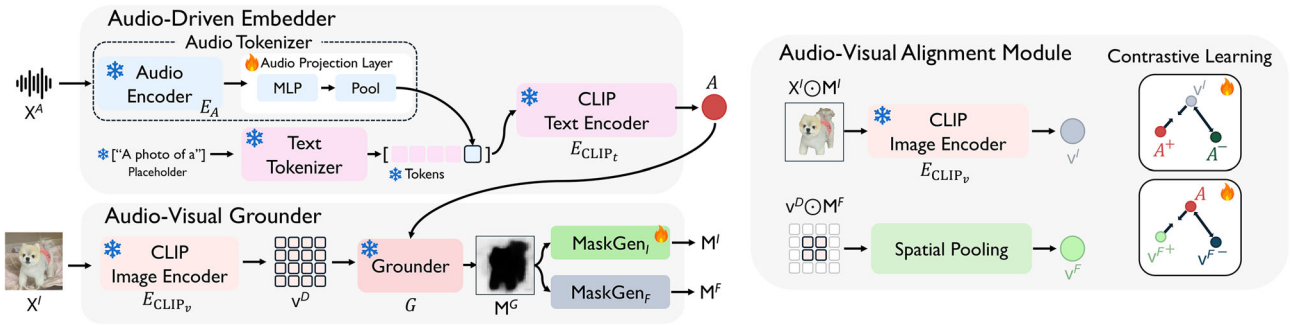


**Fig. 2 Our sound source localization framework with Audio-Visual Alignment.** The proposed method takes audio-visual pairs, translating audio signals into CLIP-compatible tokens via the Audio Tokenizer module to generate audio-driven embedding, **A**. This embedding highlights sounding regions within 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

then passed through a lightweight projection network, composed of two MLP layers and an attentive pooling layer (as in Yariv et al. (2023)), which transforms them into a form compatible with CLIP's token embeddings. While the audio encoder is pre-trained and kept frozen during training, the remaining layers are trained end-to-end in our sound source localization framework, with the objective of audio-visual alignment (see Eq. 5 or Eq 7).

The resulting 'audio token' serves as a semantic representation of the audio and is treated as a substitute for a word token. It is appended to a fixed textual prompt prefix, such as "A photo of a", forming a pseudo-text sequence as in Fig. 2. This hybrid token sequence is then passed through CLIP's frozen text encoder, $E_{\text{CLIP}_t}$, to produce a final audio-driven embedding, **A**. Importantly, because CLIP's text and vision encoders are jointly aligned during pretraining, the audio-driven embedding **A** inherits visual alignment properties by virtue of being processed through $E_{\text{CLIP}_t}$. This design enables seamless pairing or conditioning of **A** with any CLIP-based image encoder.

## 3.2 Audio-Visual Grounder

After extracting the audio-driven embeddings, we also extract visual features of each audio-visual pair. These features are then passed to our audio-visual grounder, which performs grounding by identifying regions in the image associated with the sound and generating corresponding masks. These masks are subsequently used to extract visual embeddings at both the image-level and the feature-level, which are then utilized 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 $E_{\text{CLIP}_v}$. It is responsible for encoding the provided input images into spatial features. For our grounder, $G$, we employ off-the-shelf CLIP-based segmentation network known as CLIPSeg (Lüddecke & Ecker, 2022). 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, **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.

As a common approach, sound source localization methods use audio-attended visual embeddings alongside audio embeddings within the audio-visual alignment objective. Therefore, it is essential for our method to generate differentiable binary masks that accurately represent the sounding regions. We introduce two masking methods: Image-level Mask Generator ($MaskGen_I$) and Feature-level Mask Generator ($MasGen_F$), both of which serve to extract audio-grounded visual embeddings at the image and feature levels, respectively. Similar to Cha et al. (2023), $MaskGen_I$ 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 (Jang et al., 2017) to generate a differentiable binary mask, referred to as $\mathbf{M}^I$. This mask is used to identify sounding areas in the image. $MaskGen_F$ 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 mask generators is explained in the following section.

## 3.3 Alignment Module

After obtaining sounding region masks for the given audio-visual pairs from the audio-visual grounder, our method extracts contextual visual embeddings from the masked areas

at both image and feature levels, $\boldsymbol{v}^I$ and $\boldsymbol{v}^F$ respectively. These visual embeddings are then aligned with the audio-driven embedding, $\mathbf{A}$, as part of the audio-visual alignment objective. For this purpose, we define two contrastive learning losses: image-level and feature-level audio-grounded contrastive losses, $\mathcal{L}_{ACL_I}$ and $\mathcal{L}_{ACL_F}$. In a nutshell, our model learns to maximize the alignment between the visual features of sounding regions and the corresponding 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 Fig. 2. To begin, the mask $\mathbf{M}_i^I$ obtained from $MaskGen_I$ for the $i$th audio-visual pair is used to mask out the irrelevant areas in the image. This masked image is then transformed into a visual embedding by using CLIP image encoder $\boldsymbol{v}_i^I = E_{\text{CLIP}_v}\left(\mathbf{M}_i^I \odot \mathbf{X}_i^I\right)$. The audio-visual similarity between the audio-driven embedding $\mathbf{A}_i$ and the audio-grounded visual embedding $\boldsymbol{v}_i^I$ is computed using cosine similarity and defined as $S_{i,i}^I = (\boldsymbol{v}_i^I \cdot \mathbf{A}_i)$. 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:

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

where $\tau$ is the temperature 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, $\mathcal{L}_{ACL_F}$, allowing the use of masks in lower-resolution (on features), effectively bypassing the memory constraints. This strategic approach involves emphasizing regions within the spatial visual features, as shown in Fig. 2. To elaborate, the mask $\mathbf{M}_{i,j}^F \in \mathbb{R}^{h\times w}$ obtained from the $MaskGen_F$ 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:

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

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 $\mathcal{L}_{ACL_F}$ loss is defined as follows:

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

where $\mathbf{S}^F$ is feature-level audio-visual similarity matrix within batch. One may question the necessity of introducing a separate mask generator, $\mathbf{M}^F$, when $\mathbf{M}^I$ already exists. While it is technically possible to reuse the image-level mask $\mathbf{M}^I$ in Eq. 2, doing so can lead to unintended training dynamics. Specifically, $\mathbf{M}^I$ may produce masks that are nearly zero-valued when processing negative audio-visual pairs. This behavior can cause the numerator in Eq. 2 effectively being zero, resulting in $\boldsymbol{v}_{i,j}^F$ becoming arbitrary or uninformative. Consequently, this can degrade the contrastive training process by producing random or unstable similarity scores in the InfoNCE loss for negative pairs. To mitigate this issue and maintain robust training, we decouple the mask generators and define separate modules for image-level and feature-level masking.

### 3.4 Area Regularization

We observe that even when training with the $ACL_I$ and $ACL_F$ losses, the model may adopt a shortcut by generating masks that include both relevant and irrelevant regions – sometimes covering the entire image. This behavior arises because the CLIP image encoder in the Alignment module can still produce semantically meaningful visual features, even when the masked region is overly broad. For example, the entire image of a dog, a loosely cropped foreground, or an accurately segmented dog image can all yield similar high-level CLIP features corresponding to the 'dog' concept. As a result, the model receives a weak training signal and may

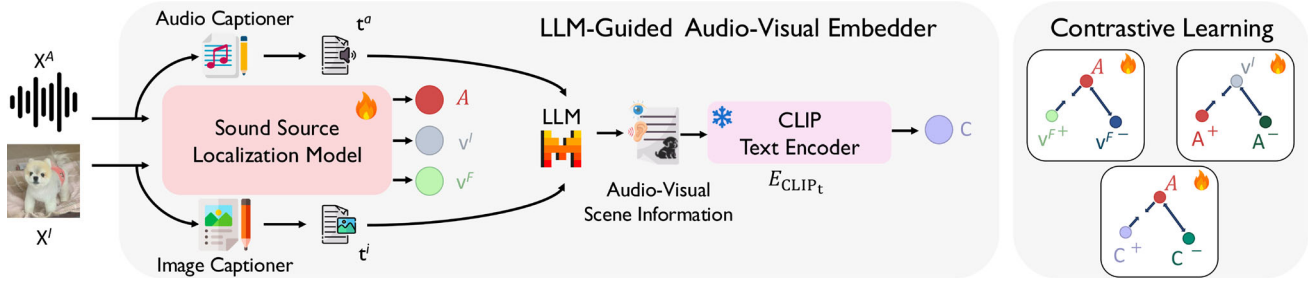


**Fig. 3** Our extended framework with LLM-Guided Object-Aware Alignment. The proposed alternative method extends our audio-visual alignment framework by incorporating LLM guidance. Given audio-visual pairs, captions from each modality, $\mathbf{t}^a$ and $\mathbf{t}^i$, are generated and processed by an LLM to extract object-aware scene information. This is then encoded using CLIP's text encoder, $\mathbf{c}$, and aligned with the audio-driven embedding $\mathbf{A}$, via contrastive learning, serving as an auxiliary objective. Since CLIP's text and visual features are aligned, this guidance implicitly reinforces audio-visual correspondence

not be incentivized to localize precisely. To address this, and following the regularization strategies proposed in Xie et al. (2022); Cha et al. (2023), we introduce an area regularization loss, defined as follows:

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

where $\mathbf{M}^{I}_{i,i}$ and $\mathbf{M}^{I}_{i,j}$ are the image masks from the positive and negative pairs respectively. The area of these masks are denoted as $\overline{\mathbf{M}}$. $p^{+}$ and $p^{-}$ represent the area prior hyperparameters, 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:

$$\mathcal{L} = \lambda_{ACL_I} \mathcal{L}_{ACL_I} + \lambda_{ACL_F} \mathcal{L}_{ACL_F} + \lambda_{REG} \mathcal{L}_{REG}, \tag{5}$$

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

### 3.6 LLM-Guided Object-Aware Alignment

In this section, we introduce an alternative learning approach to train our model by leveraging the audio-visual scene understanding capabilities of Large Language Models (LLMs). As demonstrated in Sung-Bin et al. (2025), recent LLMs have the ability to disentangle audio-visual information from metadata that describe a scene. Inspired by this, we aim to distill the object-aware audio-visual scene understanding of an LLM to enhance the audio-visual correspondence objective of our proposed method. To provide fine-grained descriptions of the scene to the LLM, we utilize off-the-shelf captioning models for each modality. For each given audio-image pair, the samples are fed into their respective captioning models, producing one caption per modality. These two captions are then provided to the LLM to obtain an object-aware understanding of the audio-visual scene. We design our LLM prompt as follows for a given image and audio captions:

> Identify the primary object in the 'image caption' most likely producing the sound like given 'audio caption', excluding background sounds which is hard to infer from given 'image caption'. Keep the answer concise and focused on general concepts, such as type. Limit the response to no more than 3 words. Image caption: {...}, Audio caption: {...}

We can formulate entire object-aware audio-visual scene knowledge stage of our pipeline as follows: $\mathcal{T}_{\text{image}} = \{t_i = G_{\text{I2T}}(\mathbf{x}_i)) \mid \forall \mathbf{x}_i \in \mathcal{D}\}$ and $\mathcal{T}_{\text{audio}} = \{t_a = G_{\text{A2T}}(\mathbf{x}_i)) \mid \forall \mathbf{x}_i \in \mathcal{D}\}$, where $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ is the dataset with $N$ samples (audio-visual pairs), and $G_{\text{I2T}}$ and $G_{\text{A2T}}$ represent the captioners for the image and audio modalities, respectively. Then, object-aware audio-visual scene information is obtained as $\mathcal{C}_{\text{object}} = \{c_i = LLM(p, t_i, t_a) \mid \forall t_i \in \mathcal{T}_{\text{image}}, \forall t_a \in \mathcal{T}_{\text{audio}}\}$.

As our proposed framework is built upon CLIP, it remains versatile, allowing the integration of CLIP's text encoder and its embeddings whenever text input is available. After the LLM generates its object-aware audio-visual scene understanding, $c_i$, we encode this information using CLIP's text encoder, $E_{\text{CLIP}_t}(c_i)$, to obtain an additional supervisory signal (Fig. 3). This representation is used alongside our original audio-visual alignment objective during training. Since CLIP's text and visual embeddings are already well aligned within the same embedding space, aligning our audio-driven embeddings to the CLIP text features inherently brings them closer to the visual features as well. By introducing this additional alignment path, we aim to strengthen the robustness of

the audio-visual correspondence learning, ultimately helping the audio features better align with the visual modality through the shared CLIP space.

To explicitly incorporate this additional alignment into our training objective, we define a caption-audio contrastive loss, $\mathcal{L}_{ACL_C}$. For each sample, the audio-caption similarity $S^C_{i,j} = (\boldsymbol{C}_i \cdot \mathbf{A}j)$ is calculated via cosine similarity between the caption embeddings, $\boldsymbol{C}_i = E_{\text{CLIP}_t}(c_i)$, and audio embeddings. Following the same InfoNCE formulation used in previous alignment objectives, the caption-level contrastive loss is defined as:

$$\mathcal{L}_{ACL_C} = \mathit{InfoNCE}(\mathbf{S}^C), \tag{6}$$

where $\mathbf{S}^C$ denotes the caption-audio similarity matrix within a batch. The total loss is formulated by adding the audio-caption similarity loss, $\mathcal{L}_{ACL_C}$, to the overall training loss term defined in Eq. 5, as follows:

$$\mathcal{L} = \lambda_{ACL_I}\mathcal{L}_{ACL_I} + \lambda_{ACL_F}\mathcal{L}_{ACL_F} + \lambda_{REG}\mathcal{L}_{REG} + \lambda_{ACL_C}\mathcal{L}_{ACL_C}. \tag{7}$$

It is important to emphasize that sound source localization is an unlabeled, audio-input-driven task, where ground-truth class labels are often unavailable across datasets. Consequently, methods that rely on text inputs derived from class labels are not universally applicable and can only be used in datasets where such annotations exist. In contrast, our proposed variant is designed to operate on any in-the-wild sample, as it does not require access to ground-truth class categories or manually annotated captions. Instead, object-aware audio-visual scene understanding is distilled through auxiliary encoders, rather than relying directly on ground-truth text labels. This design choice ensures that our method remains flexible and broadly applicable across diverse real-world scenarios where labeled metadata is not available.

### 3.7 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 $MaskGen_I$. 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 $MaskGen_I$ and $\sigma$ is sigmoid function. The final output mask is then used to obtain the localization result. Note that, regardless of the variant, our model uses only the given image and audio pairs during inference, without requiring any additional input.

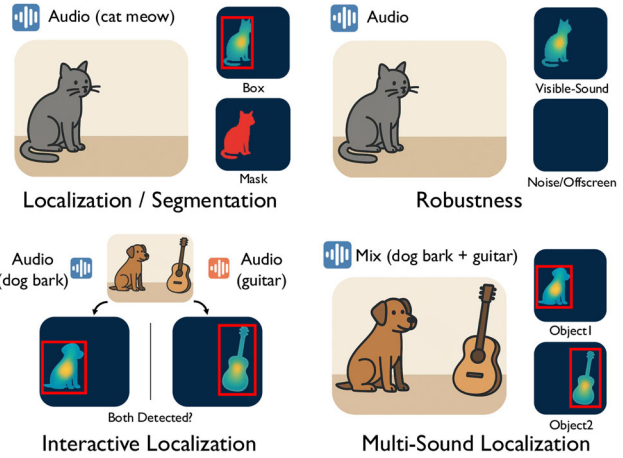


**Fig. 4** Illustration of each task

## 4 Experiments

### 4.1 Experimental Settings

Tasks. In this paper, we build on the evaluation protocol proposed by Senocak et al. (2025), which organizes sound source localization into three tasks: Single Sound Source Localization, Audio-Visual Segmentation, and Interactive Localization. We extend this by additionally including Audio-Visual Robustness and Multi-Source Localization. These tasks are illustrated in Fig. 4 and details are below:

- **Single Source Localization Task.** This task focuses on localizing sound sources from a detection perspective, specifically targeting single sounding objects. It involves localizing one object at a time rather than detecting multiple objects simultaneously, with evaluations based on ground-truth bounding boxes.
- **Audio-Visual Segmentation Task.** This task aims to localize sound sources from a segmentation perspective, measuring pixel-wise accuracy. Evaluations are based on ground-truth segmentation masks.
- **Audio-Visual Robustness Task.** This task includes the scenarios where the sound is not aligned with the visible content, such as silent audio, noisy audio (*e.g.*, white noise), or off-screen audio. This task is specifically designed to test whether the models can accurately determine if a sound is emitted by the visible object.
- **Interactive Sound Source Localization Task.** This task evaluates the model's ability to shift the localized region in the image when the same image is paired with a different sound present in the scene.
- **Multi-Source Localization Task.** Similar to the first task of localizing sound sources from a detection perspective, this task localizes multiple objects simultaneously based on given audio mixtures from a detection perspective.

Datasets. Our model is trained on the VGGSound dataset (Chen et al., 2020), which contains $\sim$ 200K videos. After training, we evaluate its performance on the following datasets for single sound source localization: VGG-SS (Chen et al., 2021), SoundNet-Flickr Test (Senocak et al., 2018, 2021), IS4 (Senocak et al., 2025), VPO-SS (Chen et al., 2024), VPO-MS (Chen et al., 2024), AVSBench-S4 (Zhou et al., 2022), and DenseAV ADE20K (Hamilton et al., 2024). Similarly, IS4, VPO-SS, VPO-MS, AVSBench-S4, AVSBench-MS3 (Zhou et al., 2022), and DenseAV ADE20K are used for audio-visual segmentation. In addition, IS4 and VPO-MS are employed for evaluating interactive sound source localization. These evaluation perspectives and datasets are adopted from Senocak et al. (2025). Extended VGG-SS and SoundNet-Flickr datasets proposed by Mo and Morgado (2022a) and AVSBench-Robust proposed by Li et al. (2025) are used for audio-visual robustness. Finally, VGGSound-Instruments (Hu et al., 2022a) and VGGSound-Duet (Mo & Tian, 2023) datasets are used for multisource sound localization.

Evaluation Metrics. Following the evaluation protocols in Senocak et al. (2025); Mo and Morgado (2022a); Zhou et al. (2022); Hu et al. (2022a), we utilize a set of metrics for each task: for single sound source localization, we use cIoU, cIoU Adaptive, AUC, and AUC Adaptive; for audio-visual robustness, the metrics are AP, max-F1, LocAcc, FPR, G-mIoU, G-F and G-FPR; for audio-visual segmentation, we use mIoU, mIoU Adaptive, F-Score, and F-Score Adaptive; for interactive localization, the metrics are IIoU, IIoU Adaptive, IAUC, and IAUC Adaptive; and for multisource localization, we use CAP, PIAP, AUC and cIoU. The details are below:

- **cIoU and AUC** (Senocak et al., 2018). These metrics are commonly used for evaluating sound source localization. cIoU quantifies the overlap between the predicted sound source region and the ground truth bounding box, while AUC measures the area under the curve based on different cIoU values. Traditionally, cIoU considers the top 50% most highly activated pixels in the audio-visual attention map as the predicted area. The Intersection over Union (IoU) is then computed between this localized region and the ground truth bounding box. Finally, samples with an IoU greater than 0.5 are considered correctly localized.
- **cIoU Adaptive and AUC Adaptive** (Senocak et al., 2025). Recently, Senocak et al. (2025) introduced these evaluation metrics. The key issue arises from the heuristic 50% threshold used to select pixels for cIoU localization evaluation, which poses challenges for small sound-emitting objects. Since this method enforces 50% of the image as the predicted sounding region, it inevitably results in over-localization when the actual sound source is smaller than half the image size. Consequently, even highly accurate models may yield low IoU values despite precise localization. To address this, the revised evaluation metric determines the number of pixels, $K$, from the ground truth annotation and retains only the top $K$ pixels in the attention map, ensuring evaluation is done exclusively with those pixels.
- **mIoU and F-Score** (Zhou et al., 2022). These evaluation metrics are traditionally used for segmentation tasks and are also directly applied in Audio-Visual Segmentation (Zhou et al., 2022).
- **mIoU Adaptive and F-Score Adaptive** (Senocak et al., 2025). We follow Senocak et al. (2025) for these metrics, which are identical to the Adaptive approach discussed above for segmentation evaluation.
- **IIoU and IAUC** (Senocak et al., 2025). These two metrics were recently introduced by Senocak et al. (2025) to evaluate a model's interactive localization ability, where the same image is paired with two audio signals corresponding to two distinct sound sources in the image. Essentially, these metrics extend cIoU and Adaptive cIoU to assess the accuracy of each localized sound source (each pair). A sample is considered successful in IIoU if the model correctly localizes all sound sources (all pairs). Conversely, if any sound source is mislocalized, the entire sample is marked as a failure.
- **IIoU Adaptive and IAUC Adaptive** (Senocak et al., 2025). It is the adaptive version of IIoU and IAUC, as discussed above.
- **AP, max-F1, and LocAcc** (Mo & Morgado, 2022a). To handle non-sounding cases such as offscreen or silence in the extended audio–visual localization test set from Mo and Morgado (2022a), a confidence score is introduced, derived from a global audio–visual similarity score. When a visible sound source exists, a prediction is counted as correct if its IoU with the ground truth exceeds 0.5 and the confidence exceeds the threshold. In non-sounding cases, the prediction is considered correct if the confidence score is below the threshold. AP follows the standard object-detection definition: the area under the precision–recall curve obtained by varying the confidence threshold. max-F1 is computed by sweeping the same confidence threshold over all samples and taking the highest F1 under this rule. LocAcc is identical to the cIoU metric used in audio–visual localization.
- **FPR, G-mIoU, G-F and G-FPR** (Li et al., 2025). These metrics assess robustness on negative cases – *i.e.*, non-visible-audio conditions such as silence, noise, or offscreen. FPR is the share of pixels wrongly activated in negatives. Global mIoU (G-mIoU) and Global F-score (G-F) extend standard audio-visual segmentation metrics to jointly account for positives and negatives: each metric is calculated as the harmonic mean of the positive score (mIoU or F) and (1 - negative score), effectively treating

a perfectly clean negative mask as 1. Global FPR (G-F) is the average FPR over all negative samples and serves as a dedicated false-activation measure.

– **CAP, PIAP, CIoU and AUC** (Hu et al., 2022b). These metrics are used for multisource localization. CAP computes pixel-wise AP for each sounding object and then averages these values, choosing the pairing between predicted maps and ground-truth objects that maximizes the result. PIAP removes this pairing step by merging all predicted sounding maps into one and computing pixel-wise AP against the union of all ground-truth sounding regions. CIoU and AUC extend the single-source cIoU and AUC metrics by computing them per sounding object and averaging across all objects in the scene.

Implementation details. We employ frozen pre-trained "ViT-B/16" CLIP (Radford et al., 2021) model as image encoder, BEATs (Chen et al., 2023) for audio encoder and CLIPSeg (Lüddecke & Ecker, 2022) for grounder. In the LLM-guided variant, we use BLIP-2 (Li et al., 2023) and GAMA (Ghosh et al., 2024) as the vision and audio captioners, respectively, and Mistral-7B-Instruct-v0.2 (Jiang et al., 2023) as the language model based on their public availability and ease of integration. 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 $\lambda_{ACL_I}$, $\lambda_{ACL_F}$, $\lambda_{ACL_C}$, and $\lambda_{Reg}$ all to 1. Additionally, we used $\tau$ 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 $10^{-4}$ and a weight decay of $10^{-4}$. The model was trained on $2\times$A6000 GPUs, taking approximately two days. Our code is released at https://github.com/swimmiing/ACL-SSL, which includes all implementation details.

Baselines in Quantitative Comparisons. 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. Baseline details are 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 (w/ CLAP Text).** CLAP (Wu et al., 2023) is designed to learn audio representations by aligning them with natural language descriptions through contrastive learning, enabling models to perform tasks such as zero-shot audio classification and audio-text retrieval. Similarly, as described above, given an audio sample, the most relevant text (class label) can be retrieved and used with CLIPSeg to obtain the grounded visual region. Note that the three baselines mentioned above can only be applied when ground-truth class labels are available, and therefore cannot be used universally.
– **WAV2CLIP, AudioCLIP and ImageBind.** These multimodal models (Wu et al., 2022; Guzhov et al., 2022; Girdhar et al., 2023) embed text, vision, and audio into a shared feature space. To enable zero-shot sound source localization with these models, we utilize a pre-trained CLIP-like object detector (Li et al., 2022) 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.

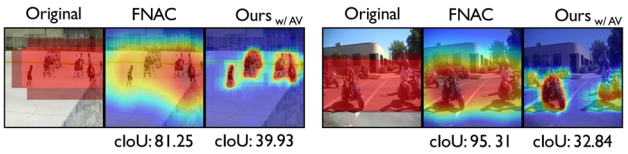


**Fig. 5** cIoU scores of SoundNet-Flickr samples

## 4.2 Results

### 4.2.1 Single Sound Source Localization

Comparison on standard benchmarks. This section presents a performance comparison between our method and existing approaches, including baselines. Experiments follow the standard evaluation protocol, based on previous works (Chen et al., 2021; Mo & Morgado, 2022b; Sun et al., 2023; Senocak et al., 2023). All models are trained on the VGGSound-144K dataset and evaluated on two widely used test sets: VGG-SS and SoundNet-Flickr. Notably, our approach does not utilize object-guided refinement (OGL). The quantitative results are reported in Table 1.

There is a substantial performance gap between our method and existing self-supervised approaches on the VGG-SS, regardless of the variant used. Notably, our approach achieves the highest scores even compared to previous methods that utilize OGL, a post-processing technique designed to refine localization outputs. Although our method is trained purely in a self-supervised manner using an audio-visual correspondence objective – similar to prior works – the results clearly demonstrate that leveraging CLIP's strong multimodal alignment significantly boosts performance. Furthermore, when LLM-based guidance is incorporated, our method gets additional gains.

**Table 1 Quantitative results on the VGG-SS and SoundNet-Flickr test sets**. All models are trained with 144K samples from VGG-Sound

| Method | VGG-SS | | | | Flickr-SoundNet | | | |
|---|---|---|---|---|---|---|---|---|
| | cIoU ↑ | w/ Adap. ↑ | AUC ↑ | w/ Adap. ↑ | cIoU ↑ | w/ Adap. ↑ | AUC ↑ | w/ Adap. ↑ |
| *Prior Works:* | | | | | | | | |
| $\text{Attention}_{\text{CVPR18}}$ | 18.50 | – | 30.20 | – | 66.00 | – | 55.80 | – |
| $\text{CoarseToFine}_{\text{ECCV20}}$ | 29.10 | – | 34.80 | – | – | – | – | – |
| $\text{LCBM}_{\text{WACV22}}$ | 32.20 | – | 36.60 | – | – | – | – | – |
| $\text{LVS}_{\text{CVPR21}}$ | 34.40 | – | 38.20 | – | 71.90 | - | 58.20 | – |
| $\text{HardPos}_{\text{ICASSP22}}$ | 34.60 | – | 38.00 | – | 76.80 | – | 59.20 | – |
| $\text{SSPL}_{\text{CVPR22}}$ | 33.90 | – | 38.00 | – | 76.70 | – | 60.50 | – |
| $\text{EZ-VSL}_{\text{ECCV22}}$ | 35.96 | 43.52 | 38.20 | 42.41 | 78.31 | 80.40 | 61.74 | 64.48 |
| $\text{EZ-VSL (w/ OGL)}_{\text{ECCV22}}$ | 38.85 | 57.77 | 39.54 | 49.00 | 83.94 | 88.80 | 63.60 | 68.90 |
| $\text{SSL-TIE}_{\text{ACM MM22}}$ | 38.63 | 51.92 | 39.65 | 48.06 | 79.50 | 84.80 | 61.20 | 65.64 |
| $\text{SLAVC}_{\text{NeurIPS22}}$ | 37.79 | 49.41 | 39.40 | 45.79 | 83.60 | 85.20 | – | 66.70 |
| $\text{SLAVC (w/ OGL)}_{\text{NeurIPS22}}$ | 39.80 | 59.08 | - | 49.73 | **86.00** | 90.00 | – | 69.28 |
| $\text{MarginNCE}_{\text{ICASSP23}}$ | 38.25 | 50.76 | 39.06 | 46.39 | 83.94 | 88.80 | 63.20 | 68.92 |
| $\text{MarginNCE (w/ OGL)}_{\text{ICASSP23}}$ | 39.78 | 59.29 | 40.01 | 50.23 | 85.14 | 91.60 | 64.55 | 70.78 |
| $\text{HearTheFlow}_{\text{WACV23}}$ | 39.40 | 54.56 | 40.00 | 48.01 | 84.80 | - | 64.00 | – |
| $\text{HearTheFlow (w/ OGL)}_{\text{WACV23}}$ | 40.24 | 58.07 | 40.23 | 49.28 | 84.80 | – | 64.00 | – |
| $\text{FNAC}_{\text{CVPR23}}$ | 39.50 | 47.00 | 39.66 | 43.30 | 84.73 | 89.20 | 63.76 | 69.78 |
| $\text{FNAC (w/ OGL)}_{\text{CVPR23}}$ | 41.85 | 58.78 | 40.80 | 49.66 | 85.14 | **92.40** | 64.30 | 70.54 |
| $\text{DenseAV}_{\text{CVPR24}}$ | 40.60 | 31.81 | 40.60 | 40.37 | 65.20 | 65.60 | 53.90 | 54.38 |
| $\text{Alignment}_{\text{TPAMI25}}$ | 41.42 | 57.25 | 40.76 | 49.32 | 83.20 | 88.00 | 64.00 | 69.36 |
| $\text{Alignment (w/ OGL)}_{\text{TPAMI25}}$ | 42.96 | 61.63 | 41.57 | 51.66 | 84.40 | 91.60 | **65.14** | **71.70** |
| *Baselines:* | | | | | | | | |
| 🔮 (***Oracle***) CLIPSeg (w/ GT Text) | 49.50 | 65.35 | 48.62 | 55.69 | – | – | – | – |
| CLIPSeg (w/ WAV2CLIP Text) | 24.84 | 20.16 | 26.01 | 24.76 | 37.20 | 36.80 | 32.14 | 42.64 |
| CLIPSeg (w/ CLAP Text) | 21.89 | 25.13 | 30.92 | 27.74 | 47.60 | 36.80 | 48.34 | 40.38 |
| $\text{WAV2CLIP}_{\text{ICASSP22}}$ | 37.71 | – | 39.93 | – | 26.00 | – | 29.60 | – |
| $\text{AudioCLIP}_{\text{ICASSP22}}$ | 44.15 | – | 46.23 | – | 47.20 | – | 45.22 | – |
| $\text{ImageBind}_{\text{CVPR23}}$ | 44.57 | – | 46.84 | - | 46.80 | – | 46.88 | - |
| ***Ours***: | | | | | | | | |
| ↪ w/ AV Alignment | 49.46 | 64.02 | 46.32 | 54.14 | 80.80 | 86.80 | 64.62 | 69.60 |
| ↪ w/ AV + LLM Alignment | **52.11** | **64.52** | **46.91** | **54.52** | 76.00 | 82.00 | 63.20 | 67.04 |

SLAVC (Mo & Morgado, 2022a) does not provide AUC scores. SoundNet-Flickr has no GT text. Baselines and our method *do not use OGL.* 🔮 method is Oracle

Interestingly, we observe that the zero-shot performance of our model on the SoundNet-Flickr test set lags behind existing models. We hypothesize this stems from the fact that our model generates more fine-grained outputs, resembling segmentation. However, the ground-truth bounding boxes for this test set are relatively coarse, causing our method to yield lower cIoU scores even when it successfully highlights the sounding region. As shown in Fig. 5, our model clearly identifies the sounding regions, yet these results still produce lower cIoU scores. This outcome is consistent with the quantitative results in Table 1, which demonstrate that our method on SoundNet-Flickr lags behind other methods due to the fact that the GT boxes and the localization results of competing methods are coarse.

Next, we compare our method against the strong baselines introduced earlier and observe that it consistently outperforms them. Notably, our method does not explicitly utilize text input to highlight object regions via CLIPSeg, as these baselines do. This suggests that our audio-visual correspondence objective is effective in learning robust audio-visual relationships, enabling the AudioTokenizer and Audio-Driven Embedder to accurately project the true audio context into meaningful embeddings. Interestingly, our AV Alignment variant achieves performance on par with the

**Table 2** Sound source localization results on recent benchmarks. All models are trained on the VGGSound-144K dataset. Results without object guided refinement (OGL) are reported

| | Method | cIoU | w/ Adap. | AUC | w/ Adap. |
|---|---|---|---|---|---|
| IS4 | $\text{LVS}_{\text{CVPR21}}$ | 33.4 | 39.4 | 39.0 | 41.1 |
| | $\text{EZ-VSL}_{\text{ECCV22}}$ | 34.2 | 42.1 | 36.6 | 42.7 |
| | $\text{SSL-TIE}_{\text{ACM MM22}}$ | 38.5 | 49.3 | 41.7 | 46.7 |
| | $\text{SLAVC}_{\text{NeurIPS22}}$ | 36.9 | 45.0 | 40.2 | 42.7 |
| | $\text{MarginNCE}_{\text{ICASSP23}}$ | 40.6 | 52.6 | 42.5 | 47.7 |
| | $\text{FNAC}_{\text{CVPR23}}$ | 39.2 | 49.5 | 42.0 | 46.1 |
| | $\text{DenseAV}_{\text{CVPR24}}$ | 21.3 | 20.5 | 33.4 | 34.3 |
| | $\text{Alignment}_{\text{TPAMI25}}$ | 45.7 | 63.1 | 44.1 | 52.4 |
| | *Baselines:* | | | | |
| | (Oracle) CLIPSeg (w/ GT Text) | 74.3 | 88.5 | 59.3 | 69.1 |
| | CLIPSeg (w/ WAV2CLIP Text) | 26.7 | 55.4 | 25.8 | 49.8 |
| | CLIPSeg (w/ CLAP Text) | 34.1 | 61.5 | 29.9 | 54.9 |
| | $\text{WAV2CLIP}_{\text{ICASSP22}}$ | 40.5 | - | 43.9 | - |
| | $\text{AudioCLIP}_{\text{ICASSP22}}$ | 39.7 | - | 44.2 | - |
| | $\text{ImageBind}_{\text{CVPR23}}$ | 57.0 | - | **58.8** | - |
| | ***Ours***: | | | | |
| | ↪ w/ AV Alignment | 65.9 | 79.0 | 53.5 | 62.9 |
| | ↪ w/ AV + LLM Alignment | **66.7** | **79.1** | 53.7 | **62.9** |
| VPO-SS | $\text{LVS}_{\text{CVPR21}}$ | 28.5 | 31.8 | 29.1 | 32.7 |
| | $\text{EZ-VSL}_{\text{ECCV22}}$ | 26.6 | 30.3 | 29.0 | 32.9 |
| | $\text{SSL-TIE}_{\text{ACM MM22}}$ | 31.7 | 39.1 | 30.6 | 36.7 |
| | $\text{SLAVC}_{\text{NeurIPS22}}$ | 29.1 | 34.3 | 30.0 | 34.2 |
| | $\text{MarginNCE}_{\text{ICASSP23}}$ | 32.1 | 35.6 | 30.3 | 35.1 |
| | $\text{FNAC}_{\text{CVPR23}}$ | 31.5 | 36.3 | 30.8 | 34.8 |
| | $\text{DenseAV}_{\text{CVPR24}}$ | 22.8 | 19.6 | 27.6 | 25.6 |
| | $\text{Alignment}_{\text{TPAMI25}}$ | 30.4 | 38.7 | 30.4 | 36.2 |
| | *Baselines:* | | | | |
| | (Oracle) CLIPSeg (w/ GT Text) | 60.9 | 74.2 | 49.0 | 59.8 |
| | CLIPSeg (w/ WAV2CLIP Text) | 12.7 | 40.0 | 13.5 | 38.3 |
| | CLIPSeg (w/ CLAP Text) | 35.4 | 55.3 | 30.1 | 49.1 |
| | $\text{WAV2CLIP}_{\text{ICASSP22}}$ | 26.9 | - | 28.5 | - |
| | $\text{AudioCLIP}_{\text{ICASSP22}}$ | 43.5 | - | 44.2 | - |
| | $\text{ImageBind}_{\text{CVPR23}}$ | 44.5 | - | **44.3** | - |
| | ***Ours***: | | | | |
| | ↪ w/ AV Alignment | 38.1 | 52.3 | 33.9 | 44.9 |
| | ↪ w/ AV + LLM Alignment | **48.3** | **63.6** | 40.4 | **51.5** |
| VPO-MS | $\text{LVS}_{\text{CVPR21}}$ | 25.0 | 28.9 | 27.8 | 30.9 |
| | $\text{EZ-VSL}_{\text{ECCV22}}$ | 25.4 | 31.0 | 28.7 | 32.6 |
| | $\text{SSL-TIE}_{\text{ACM MM22}}$ | 27.5 | 35.4 | 29.4 | 35.1 |
| | $\text{SLAVC}_{\text{NeurIPS22}}$ | 27.1 | 33.9 | 29.2 | 34.0 |
| | $\text{MarginNCE}_{\text{ICASSP23}}$ | 28.7 | 33.5 | 29.5 | 34.1 |
| | $\text{FNAC}_{\text{CVPR23}}$ | 28.4 | 34.9 | 29.8 | 34.2 |
| | $\text{DenseAV}_{\text{CVPR24}}$ | 20.9 | 18.4 | 26.4 | 25.9 |
| | $\text{Alignment}_{\text{TPAMI25}}$ | 29.0 | 36.4 | 29.7 | 35.2 |
| | *Baselines:* | | | | |
| | (Oracle) CLIPSeg (w/ GT Text) | 54.7 | 70.6 | 45.6 | 58.0 |

**Table 2** continued

| | Method | cIoU | w/ Adap. | AUC | w/ Adap. |
|---|---|---|---|---|---|
| | CLIPSeg (w/ WAV2CLIP Text) | 13.4 | 40.1 | 14.0 | 38.1 |
| | CLIPSeg (w/ CLAP Text) | 31.3 | 51.5 | 27.9 | 45.7 |
| | $\text{WAV2CLIP}_{\text{ICASSP22}}$ | 27.4 | - | 29.0 | - |
| | $\text{AudioCLIP}_{\text{ICASSP22}}$ | 37.7 | - | 39.0 | - |
| | $\text{ImageBind}_{\text{CVPR23}}$ | 43.9 | - | **43.8** | - |
| | ***Ours***: | | | | |
| | ↪ w/ AV Alignment | 38.2 | 51.3 | 34.8 | 45.2 |
| | ↪ w/ AV + LLM Alignment | **45.0** | **59.1** | 39.6 | **49.6** |
| AVSBench S4 | $\text{LVS}_{\text{CVPR21}}$ | 42.0 | 51.2 | 41.0 | 47.2 |
| | $\text{EZ-VSL}_{\text{ECCV22}}$ | 44.9 | 52.4 | 41.9 | 47.5 |
| | $\text{SSL-TIE}_{\text{ACM MM22}}$ | 47.4 | 60.8 | 43.2 | 53.3 |
| | $\text{SLAVC}_{\text{NeurIPS22}}$ | 46.8 | 58.2 | 43.2 | 50.7 |
| | $\text{MarginNCE}_{\text{ICASSP23}}$ | 47.7 | 59.0 | 43.7 | 51.8 |
| | $\text{FNAC}_{\text{CVPR23}}$ | 48.4 | 58.5 | 43.8 | 50.7 |
| | $\text{Alignment}_{\text{ICCV23}}$ | 52.1 | 67.4 | 45.0 | 55.9 |
| | $\text{DenseAV}_{\text{CVPR24}}$ | 33.5 | 43.0 | 37.8 | 46.6 |
| | *Baselines:* | | | | |
| | (Oracle) CLIPSeg (w/ GT Text) | 66.5 | 80.2 | 54.1 | 64.9 |
| | CLIPSeg (w/ WAV2CLIP Text) | 36.3 | 62.6 | 31.9 | 64.9 |
| | CLIPSeg (w/ CLAP Text) | 56.4 | 74.0 | 47.2 | 61.5 |
| | $\text{WAV2CLIP}_{\text{ICASSP22}}$ | 43.2 | - | 45.1 | - |
| | $\text{AudioCLIP}_{\text{ICASSP22}}$ | 60.6 | - | 61.1 | - |
| | $\text{ImageBind}_{\text{CVPR23}}$ | 67.8 | - | **66.0** | - |
| | ***Ours***: | | | | |
| | ↪ w/ AV Alignment | 63.5 | **81.2** | 54.9 | 64.9 |
| | ↪ w/ AV + LLM Alignment | **68.2** | 80.9 | 55.7 | **65.3** |
| ADE20K | $\text{LVS}_{\text{CVPR21}}$ | 33.0 | 36.8 | 35.0 | 39.2 |
| | $\text{EZ-VSL}_{\text{ECCV22}}$ | 35.8 | 45.3 | 36.4 | 42.5 |
| | $\text{SSL-TIE}_{\text{ACM MM22}}$ | 38.7 | 48.1 | 37.9 | 44.9 |
| | $\text{SLAVC}_{\text{NeurIPS22}}$ | 38.7 | 45.3 | 38.5 | 45.7 |
| | $\text{MarginNCE}_{\text{ICASSP23}}$ | 34.9 | 44.3 | 37.4 | 44.8 |
| | $\text{FNAC}_{\text{CVPR23}}$ | 38.7 | 42.5 | 37.8 | 43.5 |
| | $\text{DenseAV}_{\text{CVPR24}}$ | 31.1 | 27.4 | 36.1 | 36.5 |
| | $\text{Alignment}_{\text{TPAMI25}}$ | 40.6 | 51.9 | 38.5 | 47.4 |
| | *Baselines:* | | | | |
| | (Oracle) CLIPSeg (w/ GT Text) | - | - | - | - |
| | CLIPSeg (w/ WAV2CLIP Text) | 19.8 | 37.7 | 18.3 | 35.3 |
| | CLIPSeg (w/ CLAP Text) | 22.6 | 40.6 | 21.4 | 39.1 |
| | $\text{WAV2CLIP}_{\text{ICASSP22}}$ | 18.9 | - | 24.4 | - |
| | $\text{AudioCLIP}_{\text{ICASSP22}}$ | 26.4 | - | 32.7 | - |
| | $\text{ImageBind}_{\text{CVPR23}}$ | 27.4 | - | 31.8 | - |
| | ***Ours***: | | | | |
| | ↪ w/ AV Alignment | 45.3 | 58.5 | 42.4 | 54.6 |
| | ↪ w/ AV + LLM Alignment | **53.8** | **67.0** | **48.2** | **58.6** |

CLIPSeg (w/ GT Text) baseline on VGG-SS, which serves as an oracle. This baseline represents a text-conditioned, open-world segmentation approach that leverages the ground-truth class labels of the test samples. This comparison highlights the importance of incorporating audio context in a descriptive and semantically rich manner. Furthermore, the performance gap between CLIPSeg (w/ GT Text) and its variants using WAV2CLIP or CLAP shows that the zero-shot performance of CLIPSeg is highly dependent on the quality of the text input. This is likely because the text retrieved from WAV2CLIP or CLAP tends to be noisier than the ground-truth text. Nevertheless, it is important to emphasize that sound source localization is an unlabeled, audio-input-driven task, where ground-truth class labels are often unavailable across datasets. As a result, methods that rely on class label text-inputs cannot be applied universally (only in the datasets with GT class labels). These class label–conditioned methods serve primarily as oracle baselines to illustrate the robustness and effectiveness of our model.

Finally, we compare our model with AudioCLIP and WAV2CLIP, both of which are trained contrastively on image-audio pairs using the pre-trained CLIP model. As shown in Table 1, our method outperforms both approaches. This suggests that our Audio-Driven Embedder, trained with the audio-visual alignment objective, is more effective at learning robust audio-visual correspondence than these prior methods, despite all leveraging pre-trained CLIP knowledge.

Comparison on recent benchmarks. We also compare our method on recent benchmarks, following the evaluation protocol proposed by Senocak et al. (2025). These benchmarks cover a diverse range of characteristics – for example, some datasets contain real-world samples, while others consist of synthetic data (*e.g.*, IS4), and they span distinct semantic categories. While these benchmarks provide segmentation maps for the sounding objects, we follow the conventional detection-based evaluation protocol for the single sound source localization task. Specifically, we convert the segmentation maps into bounding boxes and compute cIoU, AUC, and their adaptive variants, following the evaluation procedure described in Senocak et al. (2025). The results are presented in Table 2.

Similar to the previous comparison, our method demonstrates superior performance across all five benchmarks when compared to prior works and baselines (excluding the Oracle method). On some datasets, our *AV + LLM Alignment* variant even surpasses the Oracle. The consistent high performance of our method across all datasets indicates the generalizability of our approach. We also observe that incorporating additional LLM guidance during training brings significant improvements over the AV-only alignment model on most datasets (*e.g.*, +10.2 cIoU and +11.3 in cIoU Adaptive in VPO-SS). These findings further emphasize the state-of-the-art performance of our proposed method to date.

Qualitative Results. Fig. 6 shows the qualitative comparison between our method and recent prior works on various datasets. 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 row, columns 1–2).

#### 4.2.2 Audio-Visual Robustness

Existing benchmarks typically consist of sounding object-s/regions in the scene. However, in reality, silent objects or off-screen audio are also common occurrences. Mo and Morgado (2022a) proposes 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 alignment knowledge 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.

If we examine Table 3 carefully, we observe that our model lags behind prior works only in Ext. Flickr with the LocAcc metric, despite achieving a large performance gap in other metrics. To investigate the reason for this discrepancy, we conduct a further analysis. As in Table 4, our model's max-F1 score reflects a balanced contribution from both precision and recall, with a gap of 22.37. In contrast, previous methods show much larger gaps – exceeding 40 – which indicates that their performance is dominated by recall. This suggests that their max-F1 scores are largely driven by inflated recall, reflecting a tendency to over-detect sounding regions - even in inaudible or mismatched frames. This over-detection behavior is also evident in the precision-recall (PR) curves shown in Fig. 7. While prior methods such as SLAVC (Mo & Morgado, 2022a) and FNAC (Sun et al., 2023) achieve high recall in the max-F1 metric, they suffer from low precision, resulting in a smaller area under the curve. In contrast, both our *AV Alignment* and *AV + LLM Alignment* variants consistently maintain higher precision at comparable recall levels. This

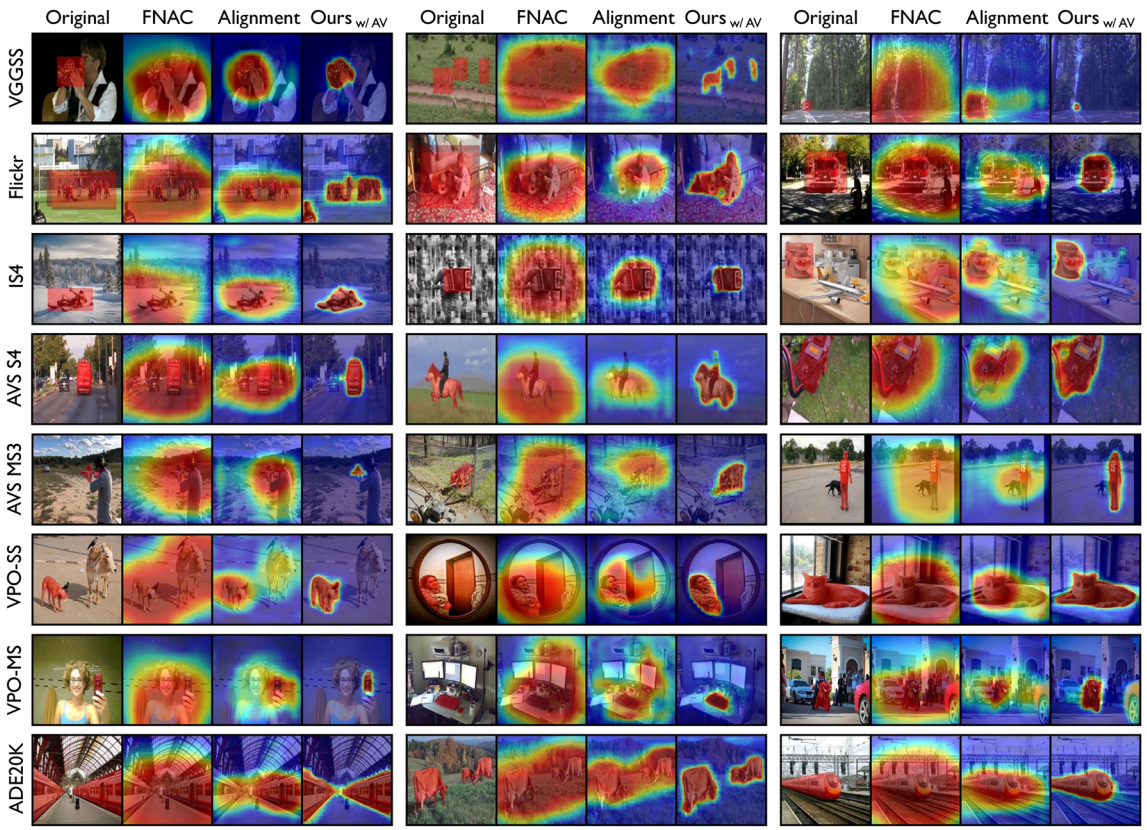


**Fig. 6** Qualitative sound source localization results on various datasets

**Table 3 Quantitative results on Extended VGG-SS and Extended Flickr-SoundNet benchmark** All models are trained with 144K samples from VGG-Sound

| Method | Extended VGG-SS | | | Extended Flickr | | |
|---|---|---|---|---|---|---|
| | AP | max-F1 | LocAcc | AP | max-F1 | LocAcc |
| $\text{SLAVC}_{\text{NeurIPS22}}$ | 32.95 | 40.00 | 37.79 | 51.63 | 59.10 | 83.60 |
| $\text{MarginNCE}_{\text{ICASSP23}}$ | 30.58 | 36.80 | 38.25 | 57.99 | 61.80 | <u>83.94</u> |
| $\text{FNAC}_{\text{CVPR23}}$ | 23.48 | 33.70 | 39.50 | 50.40 | 62.30 | **84.73** |
| $\text{DenseAV}_{\text{CVPR24}}$ | 22.07 | 30.60 | 26.21 | 70.69 | 69.50 | 65.20 |
| $\text{Alignment}_{\text{TPAMI25}}$ | 34.73 | 40.70 | 39.94 | 64.43 | 66.90 | 79.60 |
| *Baselines:* | | | | | | |
| $\text{WAV2CLIP}_{\text{ICASSP22}}$ | 26.67 | 33.00 | 37.71 | 20.99 | 24.80 | 29.60 |
| $\text{AudioCLIP}_{\text{ICASSP22}}$ | 23.79 | 32.80 | 44.15 | 34.00 | 38.80 | 45.22 |
| $\text{ImageBind}_{\text{CVPR23}}$ | 35.32 | 44.60 | 44.57 | 51.28 | 54.40 | 46.80 |
| ***Ours***: | | | | | | |
| ↪ w/ AV Align. | <u>40.79</u> | <u>49.10</u> | <u>49.46</u> | **76.07** | **73.20** | 80.80 |
| ↪ w/ AV + LLM Align. | **46.72** | **51.90** | **52.11** | <u>75.21</u> | <u>72.40</u> | 76.00 |

suggests that our model more accurately localizes sounding regions by reducing false activations in silent or irrelevant areas. As shown in Fig. 5, these prior methods often highlight broad visual regions that overlap with the coarsely annotated bounding box and yield high LocAcc scores. However, this does not mean that the model has correctly identified the presence of sound. This discrepancy highlights that high scores on LocAcc neither guarantee accurate detection of sound presence nor reflect true audio-visual understanding.

In addition, we conduct experiments on the newly introduced and more challenging AVSBench-Robust dataset (Li et al., 2025). It includes samples where the sound is not aligned with the visible content, such as silent audio, noisy audio (*e.g.*, white noise), or off-screen audio. This dataset is

**Table 4** Analysis on Extended SoundNet-Flickr dataset

| Method | Extended Flickr | | | | | |
|---|---|---|---|---|---|---|
| | AP ↑ | max-F1 ↑ | Precision ↑ | Recall ↑ | LocAcc ↑ | Recall - Precision ↓ |
| $\text{SLAVC}_{\text{NeurIPS22}}$ | 51.63 | 59.10 | 43.78 | **90.78** | 83.60 | 47.00 |
| $\text{FNAC}_{\text{CVPR23}}$ | 50.40 | 62.3 | 47.00 | 89.52 | **84.73** | 42.52 |
| **Ours:** | | | | | | |
| ↪ w/ AV Alignment | **76.07** | **73.20** | **63.67** | 86.04 | 80.80 | **22.37** |
| ↪ w/ AV + LLM Alignment | 75.21 | 72.40 | 61.33 | 88.46 | 76.00 | 27.13 |

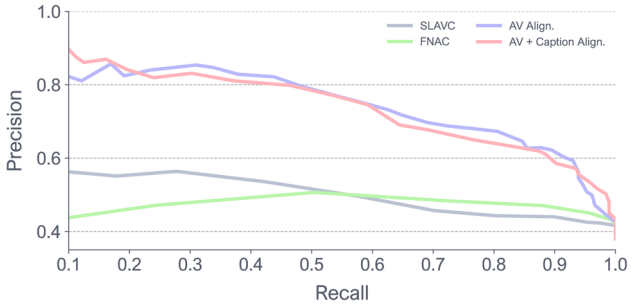


**Fig. 7** Precision-Recall Curve on Extended Flickr-SoundNet benchmark

specifically used to test whether our model can accurately determine if a sound is emitted by the visible object. As shown in Table 5, both variants of our method achieve state-of-the-art performance compared to existing works. We also provide qualitative results on the AVSBench-Robust dataset in Fig. 8. All of these results demonstrate our method's ability to correctly capture the presence or absence of sound.

#### 4.2.3 Audio-Visual Segmentation

Our proposed method produces sound source localization outputs that are more compact and fine-grained (see Fig.

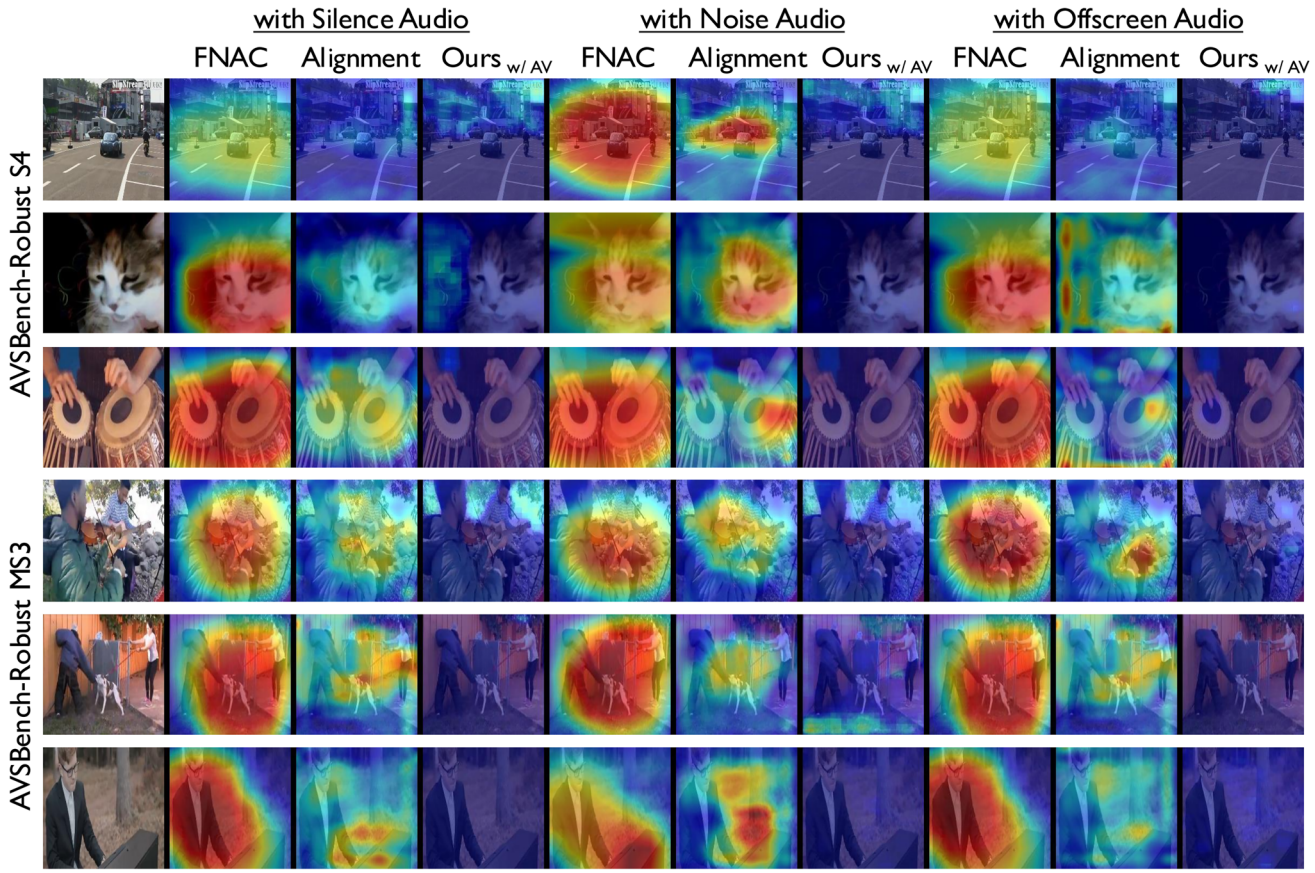


**Fig. 8** Qualitative sound source localization results on AVSBench-Robust

**Table 5** Quantitative comparison on AVSBench-Robust S4 and AVSBench-Robust MS3

| | Model | Positive audio input | | Silence | | | Noise (negative) | | | Offscreen sound | | | Global metric | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | mIoU↑ | F-score↑ | mIoU↓ | F-score↓ | FPR↓ | mIoU↓ | F-score↓ | FPR↓ | mIoU↓ | F-score↓ | FPR↓ | G-mIoU↑ | G-F↑ | G-FPR↓ |
| AVSBench-Robust S4 | $\text{SLAVC}_{\text{NeurIPS22}}$ | 28.0 | 34.4 | 27.7 | 35.7 | 0.50 | 28.2 | 36.4 | 0.50 | 27.7 | 34.9 | 0.50 | 40.343 | 44.860 | 0.500 |
| | $\text{MarginNCE}_{\text{ICASSP23}}$ | 28.9 | 35.3 | 28.1 | 28.0 | 0.50 | 26.1 | 34.7 | 0.50 | 27.8 | 39.8 | 0.50 | 41.349 | 44.822 | 0.500 |
| | $\text{FNAC}_{\text{CVPR23}}$ | 28.8 | 35.3 | 27.8 | 32.3 | 0.50 | 27.9 | 31.0 | 0.50 | 28.3 | 33.3 | 0.50 | 41.148 | 46.474 | 0.500 |
| | $\text{Alignment}_{\text{TPAMI25}}$ | 30.1 | 36.3 | 27.7 | 38.4 | 0.50 | 26.0 | 35.3 | 0.50 | 25.2 | 33.0 | 0.50 | 42.745 | 46.484 | 0.500 |
| | $\text{DenseAV}_{\text{CVPR24}}$ | 34.8 | 52.4 | 21.2 | 31.4 | 0.37 | 17.2 | 24.5 | 0.51 | 15.0 | **22.6** | 0.69 | 48.878 | 61.655 | 0.525 |
| | *Baselines:* | | | | | | | | | | | | | | |
| | $\text{WAV2CLIP}_{\text{ICASSP22}}$ | 28.7 | 35.4 | 21.1 | 28.0 | 0.17 | 21.5 | 28.3 | 0.19 | 20.8 | 27.7 | 0.18 | 42.084 | 47.464 | 0.183 |
| | $\text{Audio2CLIP}_{\text{ICASSP22}}$ | 36.6 | 42.2 | 32.6 | 38.1 | 0.36 | 34.6 | 40.1 | 0.38 | 35.6 | 41.2 | 0.38 | 47.026 | 49.618 | 0.375 |
| | $\text{ImageBind}_{\text{CVPR23}}$ | 41.1 | 47.3 | 27.9 | 34.0 | 0.25 | 22.5 | 28.7 | 0.22 | 21.7 | 27.8 | 0.20 | 53.347 | 56.405 | 0.222 |
| | ***Ours:*** | | | | | | | | | | | | | | |
| | ↪ w/ AV Alignment | 59.8 | 69.0 | **0.2** | **22.6** | 0.08 | 0.4 | **22.6** | 0.08 | **0.9** | **22.6** | 0.01 | 74.677 | 72.972 | 0.043 |
| | ↪ w/ AV + LLM Alignment | **61.8** | **69.6** | 0.4 | **22.6** | **0.05** | **0.1** | **22.6** | **0.01** | **0.9** | **22.6** | **0.00** | **76.249** | **73.298** | **0.024** |
| AVSBench-Robust MS3 | $\text{SLAVC}_{\text{NeurIPS22}}$ | 22.1 | 24.3 | 21.9 | 25.5 | 0.50 | 22.3 | 25.6 | 0.50 | 21.5 | 23.9 | 0.50 | 34.451 | 36.714 | 0.500 |
| | $\text{MarginNCE}_{\text{ICASSP23}}$ | 23.1 | 25.5 | 23.1 | 30.0 | 0.50 | 21.5 | 24.8 | 0.50 | 22.4 | 28.8 | 0.50 | 35.610 | 37.738 | 0.500 |
| | $\text{FNAC}_{\text{CVPR23}}$ | 23.2 | 25.5 | 22.9 | 24.1 | 0.50 | 22.2 | 23.7 | 0.50 | 23.1 | 25.7 | 0.50 | 35.684 | 38.166 | 0.500 |
| | $\text{Alignment}_{\text{TPAMI25}}$ | 23.7 | 26.2 | 21.6 | 25.8 | 0.50 | 21.2 | 26.1 | 0.50 | 20.0 | 22.7 | 0.50 | 36.470 | 38.867 | 0.500 |
| | $\text{DenseAV}_{\text{CVPR24}}$ | 26.5 | 33.1 | 16.9 | 17.3 | 0.31 | 18.9 | 22.4 | 0.34 | 11.8 | **16.9** | 0.50 | 40.363 | 47.174 | 0.412 |
| | *Baselines:* | | | | | | | | | | | | | | |
| | $\text{WAV2CLIP}_{\text{ICASSP22}}$ | 25.1 | 23.8 | 18.4 | **16.9** | 0.10 | 19.1 | 17.0 | 0.15 | 16.8 | **16.9** | 0.12 | 38.425 | 36.998 | 0.124 |
| | $\text{Audio2CLIP}_{\text{ICASSP22}}$ | 27.1 | 26.5 | 23.6 | 22.4 | 0.26 | 25.4 | 24.7 | 0.29 | 24.7 | 24.0 | 0.29 | 39.876 | 39.334 | 0.277 |
| | $\text{ImageBind}_{\text{CVPR23}}$ | 33.3 | 35.9 | 18.8 | **16.9** | 0.17 | 19.5 | 17.0 | 0.13 | 21.8 | 19.2 | 0.16 | 47.024 | 50.000 | 0.150 |
| | ***Ours:*** | | | | | | | | | | | | | | |
| | ↪ w/ AV Alignment | 41.1 | 46.3 | **7.8** | **16.9** | 0.11 | **7.6** | **16.9** | 0.07 | **8.6** | **16.9** | 0.02 | 56.803 | 59.495 | 0.069 |
| | ↪ w/ AV + LLM Alignment | **43.6** | **48.4** | 8.1 | **16.9** | **0.01** | 9.4 | **16.9** | **0.01** | 9.4 | **16.9** | **0.00** | **58.945** | **61.181** | **0.047** |

**Table 6** Quantitative results on the segmentation test sets

| | Method | mIoU | w/ Adap. | F-Score | w/ Adap. |
|---|---|---|---|---|---|
| IS4 | $\text{LVS}_{CVPR21}$ | 23.8 | 23.8 | 29.7 | 34.9 |
| | $\text{EZ-VSL}_{\text{ECCV22}}$ | 24.5 | 26.3 | 30.3 | 37.7 |
| | $\text{SSL-TIE}_{\text{ACM MM22}}$ | 26.0 | 32.6 | 32.1 | 45.4 |
| | $\text{SLAVC}_{\text{NeurIPS22}}$ | 24.3 | 26.7 | 30.1 | 37.9 |
| | $\text{MarginNCE}_{\text{ICASSP23}}$ | 26.1 | 30.4 | 31.9 | 42.9 |
| | $\text{FNAC}_{\text{CVPR23}}$ | 25.3 | 27.4 | 31.1 | 39.1 |
| | $\text{DenseAV}_{\text{CVPR24}}$ | 28.5 | 32.5 | 41.8 | 44.7 |
| | $\text{Alignment}_{\text{TPAMI25}}$ | 27.3 | 35.7 | 33.1 | 49.0 |
| | *Baselines:* | | | | |
| | (Oracle) CLIPSeg (w/ GT Text) | 58.4 | 74.2 | 65.1 | 82.9 |
| | CLIPSeg (w/ WAV2CLIP Text) | 18.5 | 41.9 | 22.2 | 51.9 |
| | CLIPSeg (w/ CLAP Text) | 28.8 | 54.7 | 33.9 | 64.7 |
| | $\text{WAV2CLIP}_{\text{ICASSP22}}$ | 25.6 | - | 31.7 | - |
| | $\text{AudioCLIP}_{\text{ICASSP22}}$ | 23.7 | - | 28.6 | - |
| | $\text{ImageBind}_{\text{CVPR23}}$ | 33.9 | - | 40.6 | - |
| | ***Ours***: | | | | |
| | ↪ w/ AV Alignment | <u>55.6</u> | <u>65.9</u> | <u>62.9</u> | **74.1** |
| | ↪ w/ AV + LLM Alignment | **56.3** | **66.0** | **63.3** | <u>73.8</u> |
| VPO-SS | $\text{LVS}_{\text{CVPR21}}$ | 20.3 | 21.4 | 25.5 | 29.7 |
| | $\text{EZ-VSL}_{\text{ECCV22}}$ | 20.0 | 26.3 | 25.3 | 37.7 |
| | $\text{SSL-TIE}_{\text{ACM MM22}}$ | 21.0 | 26.5 | 26.4 | 36.2 |
| | $\text{SLAVC}_{\text{NeurIPS22}}$ | 20.6 | 21.7 | 25.8 | 30.2 |
| | $\text{MarginNCE}_{\text{ICASSP23}}$ | 20.8 | 24.6 | 26.1 | 33.9 |
| | $\text{FNAC}_{\text{CVPR23}}$ | 21.1 | 22.5 | 26.3 | 31.2 |
| | $\text{DenseAV}_{\text{CVPR24}}$ | 17.7 | 18.0 | 24.2 | 25.8 |
| | $\text{Alignment}_{\text{TPAMI25}}$ | 21.0 | 23.6 | 26.3 | 32.8 |
| | *Baselines:* | | | | |
| | (Oracle) CLIPSeg (w/ GT Text) | 51.6 | 62.4 | 57.9 | 69.9 |
| | CLIPSeg (w/ WAV2CLIP Text) | 9.6 | 30.8 | 18.4 | 38.9 |
| | CLIPSeg (w/ CLAP Text) | 28.8 | <u>48.5</u> | 32.2 | <u>55.9</u> |
| | $\text{WAV2CLIP}_{\text{ICASSP22}}$ | 17.7 | - | 22.3 | - |
| | $\text{AudioCLIP}_{\text{ICASSP22}}$ | 26.1 | - | 30.5 | - |
| | $\text{ImageBind}_{\text{CVPR23}}$ | 27.4 | - | 32.6 | - |
| | ***Ours***: | | | | |
| | ↪ w/ AV Alignment | <u>32.8</u> | 42.9 | <u>37.7</u> | 49.9 |
| | ↪ w/ AV + LLM Alignment | **39.5** | **51.7** | **44.7** | **58.4** |
| VPO-MS | $\text{LVS}_{\text{CVPR21}}$ | 17.8 | 18.2 | 22.7 | 25.6 |
| | $\text{EZ-VSL}_{\text{ECCV22}}$ | 18.5 | 19.4 | 23.4 | 27.4 |
| | $\text{SSL-TIE}_{\text{ACM MM22}}$ | 19.1 | 24.2 | 24.0 | 32.7 |
| | $\text{SLAVC}_{\text{NeurIPS22}}$ | 18.7 | 20.1 | 23.6 | 28.0 |
| | $\text{MarginNCE}_{\text{ICASSP23}}$ | 19.2 | 22.4 | 24.1 | 30.9 |
| | $\text{FNAC}_{\text{CVPR23}}$ | 19.1 | 20.7 | 24.0 | 28.7 |
| | $\text{DenseAV}_{\text{CVPR24}}$ | 17.5 | 20.0 | 25.0 | 28.1 |

**Table 6** continued

| | Method | mIoU | w/ Adap. | F-Score | w/ Adap. |
|---|---|---|---|---|---|
| | $\text{Alignment}_{\text{TPAMI25}}$ | 19.7 | 23.1 | 24.6 | 31.7 |
| | *Baselines:* | | | | |
| | (Oracle) CLIPSeg (w/ GT Text) | 46.3 | 58.6 | 53.4 | 66.9 |
| | CLIPSeg (w/ WAV2CLIP Text) | 11.0 | 30.2 | 16.9 | 37.8 |
| | CLIPSeg (w/ CLAP Text) | 26.5 | 45.1 | 30.6 | 52.8 |
| | $\text{WAV2CLIP}_{\text{ICASSP22}}$ | 18.9 | - | 24.1 | - |
| | $\text{AudioCLIP}_{\text{ICASSP22}}$ | 24.6 | - | 29.0 | - |
| | $\text{ImageBind}_{\text{CVPR23}}$ | 28.9 | - | 35.0 | - |
| | ***Ours***: | | | | |
| | ↪ w/ AV Alignment | 34.3 | 43.2 | 40.0 | 50.1 |
| | ↪ w/ AV + LLM Alignment | **39.5** | **49.3** | **45.5** | **56.2** |
| AVSBench S4 | $\text{LVS}_{\text{CVPR21}}$ | 27.0 | 30.5 | 33.4 | 42.4 |
| | $\text{EZ-VSL}_{\text{ECCV22}}$ | 27.7 | 30.7 | 34.1 | 42.8 |
| | $\text{SSL-TIE}_{\text{ACM MM22}}$ | 28.9 | 38.9 | 35.2 | 52.5 |
| | $\text{SLAVC}_{\text{NeurIPS22}}$ | 28.0 | 32.8 | 34.4 | 45.5 |
| | $\text{MarginNCE}_{\text{ICASSP23}}$ | 28.9 | 35.4 | 35.3 | 48.6 |
| | $\text{FNAC}_{\text{CVPR23}}$ | 28.8 | 33.0 | 35.3 | 45.6 |
| | $\text{DenseAV}_{\text{CVPR24}}$ | 34.8 | 43.1 | 52.4 | 57.7 |
| | $\text{Alignment}_{\text{TPAMI25}}$ | 30.1 | 40.6 | 36.3 | 54.3 |
| | *Baselines:* | | | | |
| | (Oracle) CLIPSeg (w/ GT Text) | 51.3 | 63.5 | 58.0 | 72.8 |
| | CLIPSeg (w/ WAV2CLIP Text) | 26.5 | 46.7 | 30.6 | 57.4 |
| | CLIPSeg (w/ CLAP Text) | 41.3 | 58.5 | 46.8 | 67.9 |
| | $\text{WAV2CLIP}_{\text{ICASSP22}}$ | 28.7 | - | 35.4 | - |
| | $\text{AudioCLIP}_{\text{ICASSP22}}$ | 36.6 | - | 42.2 | - |
| | $\text{ImageBind}_{\text{CVPR23}}$ | 41.1 | - | 47.3 | - |
| | ***Ours***: | | | | |
| | ↪ w/ AV Alignment | 59.8 | 66.8 | 69.0 | **75.9** |
| | ↪ w/ AV + LLM Alignment | **61.8** | **67.1** | **69.6** | 75.2 |
| AVSBench MS3 | $\text{LVS}_{\text{CVPR21}}$ | 22.8 | 26.8 | 25.1 | 28.9 |
| | $\text{EZ-VSL}_{\text{ECCV22}}$ | 22.6 | 27.8 | 25.0 | 30.9 |
| | $\text{SSL-TIE}_{\text{ACM MM22}}$ | 23.5 | 32.7 | 25.9 | 37.8 |
| | $\text{SLAVC}_{\text{NeurIPS22}}$ | 22.1 | 26.1 | 24.3 | 28.5 |
| | $\text{MarginNCE}_{\text{ICASSP23}}$ | 23.1 | 30.1 | 25.5 | 35.4 |
| | $\text{FNAC}_{\text{CVPR23}}$ | 23.2 | 30.4 | 25.5 | 34.2 |
| | $\text{DenseAV}_{\text{CVPR24}}$ | 26.5 | 26.9 | 33.1 | 36.2 |
| | $\text{Alignment}_{\text{TPAMI25}}$ | 23.7 | 31.4 | 26.2 | 35.9 |
| | *Baselines:* | | | | |
| | (Oracle) CLIPSeg (w/ GT Text) | 50.9 | 55.3 | 55.9 | 66.2 |
| | CLIPSeg (w/ WAV2CLIP Text) | 30.8 | 39.8 | 30.0 | 49.5 |
| | CLIPSeg (w/ CLAP Text) | 43.0 | 46.5 | 45.4 | 56.3 |
| | $\text{WAV2CLIP}_{\text{ICASSP22}}$ | 25.1 | - | 23.8 | - |
| | $\text{AudioCLIP}_{\text{ICASSP22}}$ | 27.1 | - | 26.5 | - |
| | $\text{ImageBind}_{\text{CVPR23}}$ | 33.3 | - | 35.9 | - |

**Table 6** continued

| | Method | mIoU | w/ Adap. | F-Score | w/ Adap. |
|---|---|---|---|---|---|
| | ***Ours***: | | | | |
| | ↪ w/ AV Alignment | 41.1 | 44.4 | 46.7 | 54.2 |
| | ↪ w/ AV + LLM Alignment | **43.6** | **47.4** | **48.4** | **56.9** |
| ADE20K | $\text{LVS}_{\text{CVPR21}}$ | 22.1 | 20.0 | 27.2 | 29.5 |
| | $\text{EZ-VSL}_{\text{ECCV22}}$ | 22.3 | 20.4 | 27.3 | 29.8 |
| | $\text{SSL-TIE}_{\text{ACM MM22}}$ | 23.6 | 26.2 | 28.7 | 37.3 |
| | $\text{SLAVC}_{\text{NeurIPS22}}$ | 24.2 | 22.0 | 29.0 | 32.4 |
| | $\text{MarginNCE}_{\text{ICASSP23}}$ | 24.1 | 22.3 | 28.9 | 33.8 |
| | $\text{FNAC}_{\text{CVPR23}}$ | 23.9 | 21.8 | 28.7 | 32.1 |
| | $\text{DenseAV}_{\text{CVPR24}}$ | 24.8 | 25.8 | 36.1 | 37.2 |
| | $\text{Alignment}_{\text{TPAMI25}}$ | 25.0 | 26.9 | 30.0 | 38.3 |
| | *Baselines:* | | | | |
| | (Oracle) CLIPSeg (w/ GT Text) | - | - | - | - |
| | CLIPSeg (w/ WAV2CLIP Text) | 16.6 | 34.9 | 20.8 | 43.3 |
| | CLIPSeg (w/ CLAP Text) | 20.0 | 38.4 | 24.0 | 48.0 |
| | $\text{WAV2CLIP}_{\text{ICASSP22}}$ | 16.9 | - | 23.3 | - |
| | $\text{AudioCLIP}_{\text{ICASSP22}}$ | 20.3 | - | 26.3 | - |
| | $\text{ImageBind}_{\text{CVPR23}}$ | 20.5 | - | 27.4 | - |
| | ***Ours***: | | | | |
| | ↪ w/ AV Alignment | 37.2 | 45.6 | 44.1 | 54.6 |
| | ↪ w/ AV + LLM Alignment | **41.1** | **50.6** | **48.1** | **59.5** |

6). To further demonstrate the precision of our localization capabilities, we evaluate our method, along with its variants, on the audio-visual segmentation task. This task is particularly suitable for validating fine-grained localization ability, as it requires pixel-level accuracy. Following Senocak et al. (2025), we use the IS4, VPO-SS, VPO-MS, AVSBench-S4, and AVSBench-MS3 datasets, and additionally evaluate on the DenseAV ADE20K dataset. All of these benchmarks provide segmentation masks. These experiments are conducted in a zero-shot setting, where both our model and the compared baselines are trained on VGGSound-144K and directly tested on the target datasets without further fine-tuning. The audio-visual segmentation results are summarized in Table 6. Consistent with earlier quantitative evaluations, our method achieves superior performance compared to existing approaches; however, the performance gap is even more pronounced in the segmentation setting, with, for instance, a 27.0% cIoU improvement on the S4 dataset. This is expected, as our model tends to generate pixel-level accurate localization maps, in contrast to the coarser, blob-shaped outputs typical of previous methods. Moreover, we observe that the *AV + LLM Alignment* variant consistently provides additional performance gains across all evaluated datasets. Some qualitative results are also presented in Fig. 6.

#### 4.2.4 Interactive Sound Source Localization

The desired outcome for sound source localization models is to accurately pinpoint objects that correlate with the sound, rather than simply focusing on prominent or salient objects. Senocak et al. (2023, 2025) reveal that the majority of existing sound source localization methods fail to localize objects interactively, despite their strong performance in single sound source localization tasks. This task assesses a model's ability to adjust the localized region within an image when the same image is paired with different sounds present in the scene. For this evaluation, we use the IS4 and VPO-MS benchmarks in a zero-shot setting. The interactive sound source localization capabilities of our model, compared with existing methods, are presented in Table 7. Our method outperforms all prior works and baselines by a large margin, consistent with previous experiments. This further demonstrates that our model learns strong audio-visual correspondence by effectively building on CLIP's cross-modal alignment.

Qualitative Results.Fig. 9 visualizes the results of the Interactive Localization task, comparing our method with recent state-of-the-art approaches, FNAC (Sun et al., 2023) and Alignment (Senocak et al., 2025), on the IS4 and VPO-MS datasets. Both our method and the Alignment approach demonstrate interactive localization capabilities, but ours

**Table 7** Interactive sound source localization results. All models are trained on VGGSound-144K dataset

| | Method | IIoU | w/ Adap. | IAUC | w/ Adap. |
|---|---|---|---|---|---|
| IS4 | $\text{LVS}_{\text{CVPR21}}$ | 6.5 | 11.2 | 26.0 | 25.3 |
| | $\text{EZ-VSL}_{\text{ECCV22}}$ | 7.4 | 13.0 | 26.4 | 26.6 |
| | $\text{SSL-TIE}_{\text{ACM MM22}}$ | 9.4 | 19.0 | 28.4 | 31.5 |
| | $\text{SLAVC}_{\text{NeurIPS22}}$ | 7.5 | 14.5 | 26.3 | 25.5 |
| | $\text{MarginNCE}_{\text{ICASSP23}}$ | 11.5 | 23.7 | 29.4 | 32.5 |
| | $\text{FNAC}_{\text{CVPR23}}$ | 11.5 | 22.4 | 28.9 | 31.0 |
| | $\text{DenseAV}_{\text{CVPR24}}$ | 2.3 | 2.5 | 20.7 | 22.4 |
| | $\text{Alignment}_{\text{TPAMI25}}$ | 15.8 | 37.6 | 31.4 | 39.5 |
| | *Baselines:* | | | | |
| | (Oracle) CLIPSeg (w/ GT Text) | 54.7 | 78.4 | 47.3 | 60.4 |
| | CLIPSeg (w/ WAV2CLIP Text) | 7.4 | 30.7 | 11.6 | 34.9 |
| | CLIPSeg (w/ CLAP Text) | 16.5 | 49.1 | 18.8 | 44.7 |
| | $\text{WAV2CLIP}_{\text{ICASSP22}}$ | 16.6 | - | 22.6 | - |
| | $\text{AudioCLIP}_{\text{ICASSP22}}$ | 4.8 | - | 15.6 | - |
| | $\text{ImageBind}_{\text{CVPR23}}$ | 30.7 | - | 36.1 | - |
| | ***Ours:*** | | | | |
| | ↪ w/ AV Alignment | **43.4** | **62.3** | **38.9** | **50.5** |
| | ↪ w/ AV + LLM Alignment | 43.2 | 62.1 | 38.4 | 49.9 |
| VPO-MS | $\text{LVS}_{\text{CVPR21}}$ | 21.2 | 24.2 | 24.8 | 27.0 |
| | $\text{EZ-VSL}_{\text{ECCV22}}$ | 20.9 | 25.4 | 25.4 | 28.3 |
| | $\text{SSL-TIE}_{\text{ACM MM22}}$ | 23.8 | 30.6 | 26.4 | 31.0 |
| | $\text{SLAVC}_{\text{NeurIPS22}}$ | 22.4 | 28.4 | 25.8 | 29.3 |
| | $\text{MarginNCE}_{\text{ICASSP23}}$ | 24.9 | 28.1 | 26.4 | 29.9 |
| | $\text{FNAC}_{\text{CVPR23}}$ | 24.9 | 29.7 | 26.8 | 30.1 |
| | $\text{DenseAV}_{\text{CVPR24}}$ | 17.0 | 13.4 | 23.4 | 21.6 |
| | $\text{Alignment}_{\text{TPAMI25}}$ | 24.9 | 30.5 | 26.6 | 30.9 |
| | *Baselines:* | | | | |
| | (Oracle) CLIPSeg (w/ GT Text) | 47.5 | 65.0 | 39.8 | 53.6 |
| | CLIPSeg (w/ WAV2CLIP Text) | 8.4 | 32.8 | 10.9 | 32.9 |
| | CLIPSeg (w/ CLAP Text) | 26.9 | 50.5 | 25.6 | 43.2 |
| | $\text{WAV2CLIP}_{\text{ICASSP22}}$ | 16.8 | - | 19.7 | - |
| | $\text{AudioCLIP}_{\text{ICASSP22}}$ | 21.6 | - | 24.5 | - |
| | $\text{ImageBind}_{\text{CVPR23}}$ | 35.2 | - | **36.4** | - |
| | ***Ours:*** | | | | |
| | ↪ w/ AV Alignment | 28.8 | 42.2 | 28.4 | 38.4 |
| | ↪ w/ AV + LLM Alignment | **37.4** | **51.9** | 33.7 | **43.9** |

yields more accurate results, whereas the FNAC method fails with largely unchanged outputs.

#### 4.2.5 Sound Source Localization in Mixtures

Until now, all tasks presented in this paper have focused on the single sound source localization scenario. However, another important research direction in visual sound source localization is multi-source sound localization, where multiple sound-emitting objects are present in a scene. In this section, we demonstrate the capability of our proposed method to handle this setting too.

Datasets and evaluation metrics. Following AVGN (Mo & Tian, 2023) and T-VSL (Mahmud et al., 2024), we use the VGGSound-Instruments (Hu et al., 2022a) and VGGSound-Duet (Mo & Tian, 2023) datasets for evaluation, along with the metrics defined in Hu et al. (2022a): CAP (Class-Aware Average Precision), PIAP (Permutation-Invariant Average Precision), CIoU (Class-Aware Intersection over Union), and AUC (Area Under the Curve). Refer to the original papers for detailed metric definitions. We note that prior works report

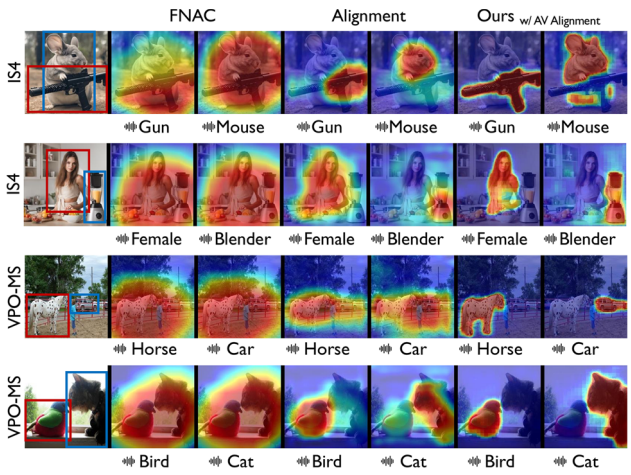


**Fig. 9** Qualitative results for interactive sound source localization. Our model interactively perform accurate localization for given different sounds, whereas FNAC results are remain unchanged

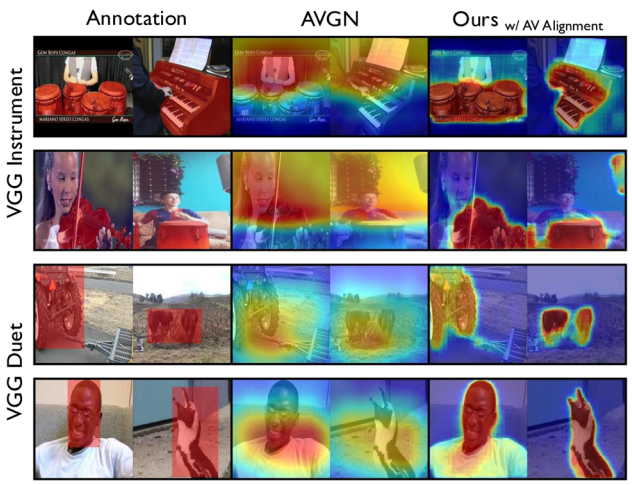


**Fig. 10** Qualitative results for multi-source localization

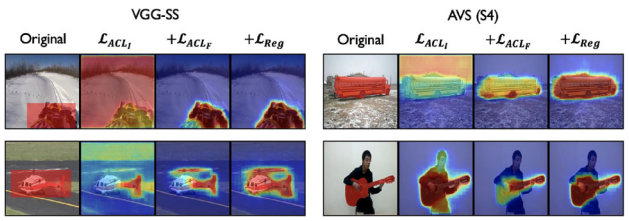


**Fig. 11** Sound localization results by using different combinations of loss functions

results using CIoU@10 for VGGSound-Instruments and CIoU@30 for VGGSound-Duet. However, these thresholds – particularly 10% – represent relatively small overlapping ratios, which may be insufficient for robust evaluation. Therefore, in addition to the original settings, we also evaluate our models using CIoU@10, CIoU@30 and CIoU@50 on both datasets, alongside other methods with publicly available models.

Results. Unlike existing methods, which are individually trained on each target dataset for multi-source localization with audio mixtures, we directly apply our approach in a zero-shot manner – just as we have done in previous tasks and experiments. Prior works such as AVGN and T-VSL incorporate additional supervision, including class labels (*e.g.*, as class tokens in visual transformers or as category-level text inputs), during both training and inference. Following the protocol of T-VSL, we use a predefined category list and feed these categories into our grounding module, $G$, which is capable of conditioning on text input.

Given $N$ classes, our grounder outputs a sounding region mask for each class category, denoted as $\mathbf{M}_n^G$, where $n \in 1, \ldots, N$. Following the inference procedure described in Section 3.7, these masks are subsequently passed to $MaskGen_I$ to obtain the corresponding image-level visual features, denoted as $\boldsymbol{v}_n^I$. Once the visual features are obtained, we estimate the cosine similarity between the audio-driven embedding, $\mathbf{A}$, and each of the $n$ candidate localization regions. Finally, we select the top-$k$ heatmaps corresponding to the highest similarities, and report these as our multi-source localization results. We present the results in Table 8. Notably, although our model is not specifically trained for the multi-source sound localization task, our proposed zero-shot setting outperforms methods that are explicitly trained for it in both datasets. This ability can be attributed to two key factors: (1) the versatility of our approach in seamlessly incorporating text input when available, and (2) the semantically rich audio-driven embeddings, which are highly aligned with CLIP's visual features. We also provide qualitative results in Fig. 10.

Here, we observe that the *AV+LLM alignment* variant performs relatively lower than *AV Alignment* in the multi-source setting. This is mainly due to the design of the LLM-based captions, which are intended to focus on the primary sounding object, as defined by the prompt in Section 3.6. As a result, the model tends to localize only the dominant source, rather than capturing all sounding objects in the scene.

## 4.3 Ablation Results

### 4.3.1 Ablation on different combinations of loss functions

Our method is optimized by a combination of three loss functions, *i.e.* $\mathcal{L}_{ACL_I}$, $\mathcal{L}_{ACL_F}$, and area regularization. Here, we perform ablation experiments to understand the impact of each loss function. We primarily conduct experiments on VGG-SS, AVSBench S4 and Extended VGG-SS datasets. Results are in Table 9.

As revealed by results (A) and (B), using $\mathcal{L}_{ACL_I}$ is crucial to enable our model to learn the corresponding audio-visual alignment. On the other hand, relying solely on $\mathcal{L}_{ACL_F}$ is

**Table 8** Performance comparison of multi-source localization

| | Method | CAP | PIAP | AUC | CIoU@10 | CIoU@30 | CIoU@50 |
|---|---|---|---|---|---|---|---|
| VGG Inst. | $\text{OTS}_{\text{ECCV18}}$ | 23.3 | 37.8 | 11.7 | 51.2 | - | - |
| | $\text{Mix-and-Localize}_{\text{CVPR22}}$ | 21.5 | 37.5 | 15.6 | 73.2 | - | - |
| | $\text{AVGN}_{\text{CVPR23}}$ | 27.3 | 42.8 | 18.2 | 77.5 | 21.4 | 5.71 |
| | $\text{NoPrior}_{\text{CVPR24}}$ | - | - | - | - | - | - |
| | $\text{T-VSL}_{\text{CVPR24}}$ | 41.8 | - | 31.5 | 89.6 | - | - |
| | ***Ours:*** | | | | | | |
| | ↪ w/ AV Alignment | **64.1** | **78.9** | 51.6 | **97.2** | 81.6 | 55.6 |
| | ↪ w/ AV + LLM Alignment | 63.4 | 78.4 | **53.1** | 97.0 | **83.7** | **57.2** |
| VGG Duet | $\text{OTS}_{\text{ECCV18}}$ | 10.5 | 12.7 | 15.8 | - | 12.2 | - |
| | $\text{Mix-and-Localize}_{\text{CVPR22}}$ | 16.3 | 22.6 | 20.5 | - | 21.1 | - |
| | $\text{AVGN}_{\text{CVPR23}}$ | 21.9 | 28.1 | 23.8 | 84.5 | 26.2 | 22.2 |
| | $\text{NoPrior}_{\text{CVPR24}}$ | 32.5 | 44.4 | 29.2 | - | 46.9 | |
| | $\text{T-VSL}_{\text{CVPR24}}$ | 35.7 | - | 37.9 | - | 40.1 | |
| | ***Ours:*** | | | | | | |
| | ↪ w/ AV Alignment | **56.8** | **63.2** | **42.2** | **84.9** | **62.7** | **42.6** |
| | ↪ w/ AV + LLM Alignment | 53.7 | 61.5 | 38.7 | 77.7 | 55.9 | 37.2 |

**Table 9** Ablative experiments on our method by using different combinations of loss functions

| | $\mathcal{L}_{ACL_I}$ | $\mathcal{L}_{ACL_F}$ | *Reg* | VGG-SS | | AVS (S4) | | Ext. 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** |

not effective for learning audio-visual alignment, 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 is complementary. As mentioned earlier, $\mathcal{L}_{ACL_F}$ contributes to performance enhancement by suppressing background areas. Furthermore, the results from the experiments (C *vs.* F) show that area regularization provides additional improvements by constraining the size of the activated regions.

Visualization of the ablation experiments. The visual results are presented in Fig. 11. As demonstrated, when using only $\mathcal{L}_{ACL_I}$, we observe that background areas remain activated (also discussed in Section 3.3). As evident in the third column, the addition of $\mathcal{L}_{ACL_F}$ helps eliminate the background pixels (non-sounding areas). However, it is noticeable that the outputs of $\mathcal{L}_{ACL_I}$+$\mathcal{L}_{ACL_F}$ can be relatively less completed. With the area regularizer, the final output of our model becomes more complete and fine-grained.

#### 4.3.2 Ablation on different modality captions for LLM-based guidance alignment

Our *AV + LLM Alignment* variant uses the output of an LLM as auxiliary self-supervision, where the LLM's response is generated based on obtained captions from vision and audio samples. To see the impact of each caption modality, we trained our model by providing the LLM with captions from only a single modality. We present the results in Table 10.

We observe that captions derived from the vision modality yield better performance than those from the audio modality. However, incorporating single-modality captions along with their corresponding LLM responses does not lead to improved performance; in fact, they underperform compared to the AV Alignment variant. This can be attributed to the fact that single-modality captions prompt the LLM to infer object-aware audio-visual understanding using only one modality, which may be insufficient to accurately describe the scene or identify the sound-emitting object. As a result, this weakens the audio-visual correspondence the model is

**Table 10** Ablation on caption modalities for caption alignment variant.

| Caption | VGG-SS | | AVS (S4) | | Ext. VGG-SS | |
|---|---|---|---|---|---|---|
| Modality | cIoU | AUC | mIoU | F-Score | AP | max-F1 |
| AV Align. | 49.5 | 46.3 | 59.8 | 69.0 | 40.8 | 49.1 |
| Audio | 46.7 | 43.3 | 58.3 | 66.2 | 42.8 | 47.4 |
| Vision | 49.0 | 45.0 | 60.7 | 68.3 | 44.3 | 49.4 |
| Audio + Vision | **52.1** | **46.9** | **61.8** | **69.6** | **46.7** | **51.9** |

**Table 11** Ablation on audio token length

| Audio Token | VGG-SS | | AVS (S4) | | Ext. VGG-SS | |
|---|---|---|---|---|---|---|
| | cIoU | AUC | mIoU | F-Score | AP | max-F1 |
| Length = 1 | **49.5** | **46.3** | 59.8 | **69.0** | 40.8 | **49.1** |
| Length = 2 | 48.4 | 44.1 | 59.3 | 67.5 | 42.3 | 47.6 |
| Length = 4 | 48.6 | 44.2 | 57.2 | 65.6 | **44.8** | 48.3 |
| Length = 8 | 46.6 | 44.3 | **60.8** | 57.2 | 44.2 | 48.3 |

learning between audio-driven embeddings and visual features. In contrast, when captions from both modalities are provided to the LLM, the response is likely to be more accurate and semantically aligned with the scene. This enriched multi-modal supervision enables the model to surpass the performance of the AV Alignment variant alone.

#### 4.3.3 Ablation on audio token length

We analyze how different audio token lengths from the Audio Projection layer affect performance in Table 11. While some datasets (*e.g.*, Extended VGG-SS) benefit from longer tokens, the overall performance difference is minor. Considering the trade-off with increased model complexity, using a single token is sufficient and does not significantly compromise performance.

#### 4.3.4 Ablation on prompt template

We ablate the impact of the audio tokenizer prompt template. Each model is trained using a specific template, and the same template is used for inference. As shown in Table 12, there is no significant performance gap between the different templates, and both "A photo of a ⟨∗⟩" and "An image of a ⟨∗⟩" perform strongly. We use "A photo of a ⟨∗⟩" as it is a good default template in the literature (Radford et al., 2021).

#### 4.3.5 Ablation on the reuse of $\mathbf{M}^I$ for feature-level alignment

We perform an ablation study on the choice of masks used in the image-level ($ACL_I$) and feature-level ($ACL_F$) contrastive objectives, as shown in Table 13. When using the image-level mask $\mathbf{M}^I$ for feature-level alignment (first row in Table 13), performance degrades across all metrics. As discussed in the main text, this is likely because $\mathbf{M}^I$ may fail to activate any region for negative samples, which can lead to zeroed feature representations and introduce undesirable guidance during training. This results in suboptimal optimization and reduced performance.

#### 4.3.6 Ablation on the reuse of $\mathbf{M}^F$ for image-level alignment

Conversely, applying the feature-level mask $\mathbf{M}^F$ to the image-level objective (second row in Table 13) also leads to degraded performance. Since $\mathbf{M}^F$ is defined at a lower resolution, upsampling it for image-level use produces blurred boundaries, resulting in less precise object–background separation.

#### 4.3.7 Ablation on binary vs. soft masks

We perform an ablation between a binary image-level mask $\mathbf{M}^I$ obtained by Gumbel-Max from $\mathbf{M}^G$ and a soft mask $\sigma(\mathbf{M}^G)$ (last row vs. third row in Table 13). The soft mask performs worse because it blurs object–background boundaries and partly includes many background pixels, which weakens region-level contrastive learning.

#### 4.3.8 Ablation on area priors

Our training is based on the VGGSound dataset, where the average ground-truth bounding box area in VGG-SS is $0.3124 \pm 0.2342$. Accordingly, we set the positive area prior $p^+$ to 0.4, which is slightly above the average to adequately cover the typical object regions and maintain stable learning.

We first perform an ablation by varing $p^+$ while fixing $p^-$. On VGGSound, decreasing $p^+$ below 0.4 makes the model focus on narrow regions, while increasing $p^+$ above 0.4 admits more background. Both cases fail to provide proper informative feature for contrastive learning, leading to degraded performance (see first and third rows of Table 14). Overall, $p^+ = 0.4$ aligns with the VGGSound distribution and provides the best choice.

We also perform an ablation by varying $p^-$ while fixing $p^+ = 0.4$. For negative pairs, the ideal mask is all zeros. Set-

**Table 12** Ablation on prompt variants

| Prompt | VGG-SS | | AVS (S4) | | Ext. VGG-SS | |
|---|---|---|---|---|---|---|
| | cIoU | AUC | mIoU | F-Score | AP | max-F1 |
| `"A photo of a <*>"` | 49.5 | 46.3 | 59.8 | 69.0 | 40.8 | 49.1 |
| `"A photograph of a <*>"` | 46.7 | 43.5 | 60.5 | 68.7 | 42.0 | 47.6 |
| `"An image of a <*>"` | 48.8 | 45.0 | **61.0** | **69.2** | 43.7 | 49.0 |
| `"<*>"` | 47.6 | 44.3 | 59.2 | 67.2 | **43.8** | 48.7 |

**Table 13** Ablation on mask usage variant in objective function $\mathcal{L}_{ACL_I}$ and $\mathcal{L}_{ACL_F}$.

| Mask | | VGG-SS | | AVS (S4) | | Ext. VGG-SS | |
|---|---|---|---|---|---|---|---|
| $\mathcal{L}_{ACL_I}$ | $\mathcal{L}_{ACL_F}$ | cIoU | AUC | mIoU | F-Score | AP | max-F1 |
| $\mathbf{M}^I$ | $\mathbf{M}^I$ | 26.9 | 34.2 | 25.0 | 30.4 | 13.3 | 22.6 |
| $\mathbf{M}^F$ | $\mathbf{M}^F$ | 41.5 | 32.9 | 28.1 | 35.3 | 24.8 | 31.1 |
| $\sigma(\mathbf{M}^G)$ | $\mathbf{M}^F$ | 46.0 | 43.8 | 59.1 | 67.0 | **41.9** | 45.9 |
| $\mathbf{M}^I$ | $\mathbf{M}^F$ | **49.5** | **46.3** | **59.8** | **69.0** | 40.8 | **49.1** |

**Table 14** Ablation on area priors

| Area Prior | | VGG-SS | | Flickr | | Ext. Flickr | |
|---|---|---|---|---|---|---|---|
| $p^+$ | $p^-$ | cIoU | AUC | cIoU | AUC | AP | max-F1 |
| 0.2 | 0.0 | 44.9 | 43.1 | 70.0 | 59.0 | 71.2 | 67.9 |
| 0.4 | 0.0 | **49.5** | **46.3** | **80.8** | **64.6** | **76.1** | **73.2** |
| 0.6 | 0.0 | 45.5 | 42.6 | 70.0 | 58.7 | 75.6 | 69.2 |
| 0.4 | 0.2 | 47.5 | 44.4 | 71.6 | 61.3 | 73.1 | 69.2 |

ting $p^- = 0.2$ forces non-zero regions in negatives, injecting unnecessary noise into the negative features and providing poor guidance, which degrades performance (see the last row of Table 14).

#### 4.3.9 Ablation on different loss balance weights

We ablate the dynamics of the combined loss functions. The overall training loss is defined as $\mathcal{L} = (\mathcal{L}_{ACL_I} + \mathcal{L}_{ACL_F} + \mathcal{L}_{REG}) + \lambda_{ACL_C} \mathcal{L}_{ACL_C}$ where the terms in parentheses represent the AV alignment loss and $\mathcal{L}_{ACL_C}$ is the LLM alignment loss. We show the effect of the weight $\lambda_{ACL_C}$ of the LLM alignment loss within the overall training regime, as shown in Table 15. The results show that incorporating LLM alignment consistently improves performance across all datasets, with particularly notable gains on Extended VGG-SS, where it helps more accurately determine whether a sound is present in the scene. However, setting $\lambda_{ACL_C}$ too high (e.g., $> 10$) leads to performance degradation. We interpret this as the model overemphasizing alignment with the primary object described in the caption, thereby reducing the diversity of audio–visual associations and limiting the model's ability to capture alternative sounding regions.

**Table 15** Ablation on $\lambda_{ACL_C}$

| | VGG-SS | | AVS (S4) | | Ext. VGG-SS | |
|---|---|---|---|---|---|---|
| $\lambda_{ACL_C}$ | cIoU | AUC | mIoU | F-Score | AP | max-F1 |
| 0.0 | 49.5 | 46.3 | 59.8 | 69.0 | 40.8 | 49.1 |
| 0.01 | 49.9 | 45.8 | 60.7 | 68.3 | 44.8 | 49.7 |
| 0.1 | 51.1 | 46.5 | 61.0 | 69.4 | 45.4 | 51.0 |
| 1 | **52.1** | **46.9** | **61.8** | **69.6** | **46.7** | **51.9** |
| 10 | 50.2 | 46.0 | 60.8 | 69.0 | 43.8 | 50.0 |
| 100 | 46.5 | 44.8 | 57.0 | 65.2 | 39.4 | 45.3 |

### 4.4 Further Analysis on Audio Tokenizer

In this section, we analyze the capabilities of our AudioTokenizer and audio-driven embeddings, focusing on their ability to describe audio samples in a richer and more detailed manner. We also provide qualitative examples to emphasize that the 🔮 Oracle method (CLIPSeg w/ GT text) method is not universally applicable and has inherent limitations. Throughout the paper, we have discussed that the sound source localization task is inherently unlabeled, and that ground-truth class information for benchmark samples is not always available, making the Oracle method a hypothetical reference. To further demonstrate why solving sound source localization through text input-based segmentation approaches is problematic, we present qualitative examples showing that class labels are often insufficiently descriptive to accurately localize sound sources. In contrast, learning audio-visual alignment equips the audio encoder with greater descriptive power, enabling our model – enhanced by the proposed AudioTokenizer – to represent audio more richly and localize sound sources more effectively. Note that this anal-

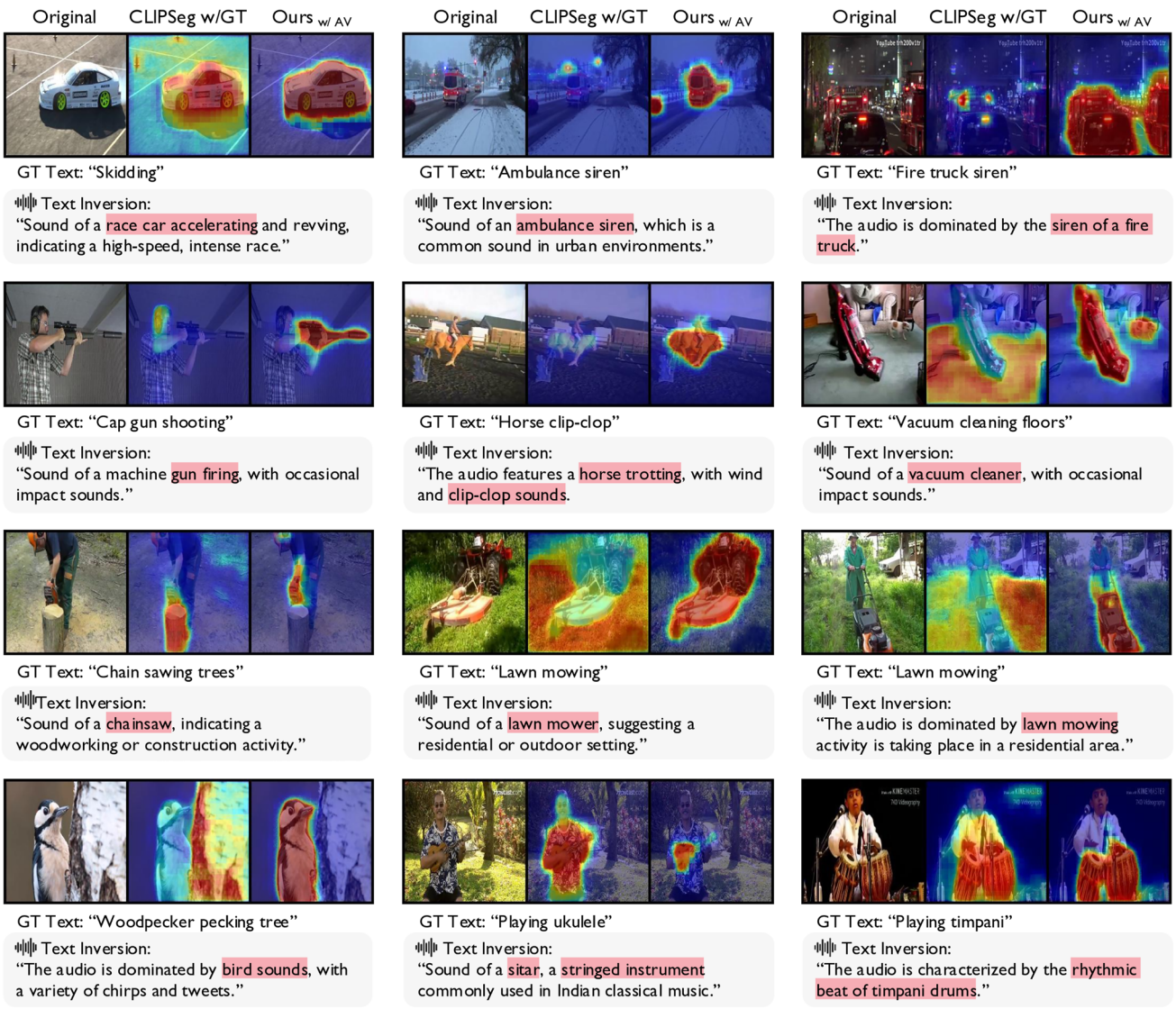


**Fig. 12** Analysis on AudioTokenizer and visual comparison with racle method. However, our model can localize sound sources accurately using only the audio embeddings learned through audio-visual correspondence. This indicates that our AudioTokenizer enables the generation of semantically accurate audio embeddings. We also visualize what our audio embeddings describe using text inversion approach

ysis is based on our *AV Alignment* variant, and all examples are drawn from the VGG-SS dataset.

We present qualitative comparison results on the VGG-SS dataset in Fig. 12. These results highlight two key observations: (1) Sound sources cannot always be accurately localized, even when class label text information is available. Audio provides a broader descriptive context in visual scenes, and sound source localization models should focus on learning true audio-visual correspondence. (2) The successful results of our model compared to the Oracle method demonstrates that our AudioTokenizer module effectively encodes the true audio context, enabling proper learning of the audio-visual correspondence objective.

To further support our claim that the audio-driven embeddings effectively capture audio content, we apply a text inversion approach with them to visualize what they represent and how they describe the input. We perform the text inversion experiment using a CLIP-based captioning model. Specifically, we use CAPDEC (Nukrai et al., 2022), an autoregressive model that generates text by conditioning a GPT2 (Radford et al., 2019) decoder on CLIP embeddings. We train CAPDEC on image-text pairs constructed from the VGGSound dataset by generating audio captions via GAMA (Ghosh et al., 2024) and pairing them with the corresponding images. Once trained, the model receives our audio-driven embeddings instead of image embeddings,

allowing us to generate descriptive text from audio to see the semantics encoded by the learned audio representations. Text inversion results are also in Fig. 12. The results demonstrate that our audio-driven embeddings carry accurate and descriptive semantic information.

## 5 Conclusion

In this work, we introduced a self-supervised framework for sound source localization that leverages the multimodal alignment knowledge of foundational models, specifically CLIP. Our AudioTokenizer module transforms audio signals into CLIP-compatible tokens, enabling audio-visual alignment without relying on textual class labels. Through contrastive learning, our model effectively grounds sounding regions and aligns audio and visual representations using audio-visual correspondence. Extensive experiments across diverse tasks – including single-source and multi-source localization, segmentation, interactive localization, and robustness – demonstrate that our method consistently outperforms existing approaches, achieving state-of-the-art performance and strong generalization in zero-shot settings. Additionally, we propose an LLM-guided extension to further enhance alignment through object-aware audio-visual scene understanding during training. Our findings suggest that the core challenge of sound source localization – achieving strong audio-visual alignment – can be addressed by building upon structured multimodal alignment priors offered by large-scale pre-trained foundation models.

**Acknowledgements** This work was partly supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2025-02215122, Development and Demonstration of Lightweight AI Model for Smart Homes; 50%), the Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government [26ZC1100, Development of Spatial Media Technology and Interaction Technology for Convergence of the Real and Virtual World; 40%], and the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2020-II201336, Artificial Intelligence Graduate School Program (UNIST); 10%).

**Funding** Open Access funding enabled and organized by KAIST.

**Data Availability** All data supporting the findings of this study are available online. The VGGSound dataset can be downloaded from https://www.robots.ox.ac.uk/~vgg/data/vggsound/. The SoundNet-Flickr dataset can be downloaded from https://github.com/ardasnck/learning_to_localize_sound_source. The IS4 dataset can be downloaded from https://github.com/kaistmm/SSLalignment. The VPO datasets can be downloaded from https://github.com/cyh-0/CAVP. The AVSBench datasets can be downloaded from http://www.avlbench.opennlplab.cn/download. The DenseAV ADE20K datasets can be downloaded from https://github.com/mhamilton723/DenseAV. The Extended VGG-SS/Flickr datasets can be downloaded from https://github.com/stoneMo/SLAVC. The VGGSound-Instrument datasets can be downloaded from https://web.eecs.umich.edu/~ahowens/mix-localize/. The VGGSound-Duet datasets can be downloaded from https://github.com/stoneMo/AVGN. The AVSBench-Robust datasets can be downloaded from https://github.com/jiali-home/AVSBench-Robust.



## References

Arandjelović, R., & Zisserman, A. (2018). Objects that sound. In: *European Conference on Computer Vision (ECCV)*.

Bhati, S., Villalba, J., Moro-Velazquez, L., Thebaud, T., & Dehak, N. (2023). Segmental speechclip: Utilizing pretrained image-text models for audio-visual learning. In: *INTERSPEECH*.

Cha, J., Mun, J., & Roh, B. (2023). 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 (CVPR)*.

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

Chen, H., Xie, W., Afouras, T., Nagrani, A., Vedaldi, A., & Zisserman, A. (2021). Localizing visual sounds the hard way. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chen, S., Wu, Y., Wang, C., Liu, S., Tompkins, D., Chen, Z., & Wei, F. (2023). BEATs: Audio pre-training with acoustic tokenizers. In: *International Conference on Machine Learning (ICML)*.

Chen, Y., Liu, Y., Wang, H., Liu, F., Wang, C., Frazer, H., & Carneiro, G. (2024). Unraveling instance associations: A closer look for audio-visual segmentation. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dong, H. W., Takahashi, N., Mitsufuji, Y., McAuley, J., & Berg-Kirkpatrick, T. (2022). Clipsep: Learning text-queried sound separation with noisy unlabeled videos. In: *International Conference on Learning Representations (ICLR)*.

Fan, Y., Wu, Y., Lin, Y., & Du, B. (2023). Revisit weakly-supervised audio-visual video parsing from the language perspective. *arXiv preprint* arXiv:2306.00595.

Fedorishin, D., Mohan, D. D., Jawade, B., Setlur, S., & Govindaraju, V. (2023). Hear the flow: Optical flow-based self-supervised visual sound source localization. In: *IEEE Winter Conference on Applications of Computer Vision (WACV)*.

Ghosh, S., Kumar, S., Seth, A., Evuru, C.K.R., Tyagi, U., Sakshi, S., Nieto, O., Duraiswami, R., & Manocha, D. (2024). Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities. *arXiv preprint* arXiv:2406.11768.

Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K. V., Joulin, A., & Misra, I. (2023). Imagebind: One embedding space to bind them all. (pp. 15180–15190) In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Guzhov, A., Raue, F., Hees, J., & Dengel, A. (2022). AudioCLIP: Extending CLIP to image, text and audio. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Hamilton, M., Zisserman, A., Hershey, J. R., & Freeman, W. T. (2024). Separating the chirp from the chat: Self-supervised visual grounding of sound and language. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hu, D., Qian, R., Jiang, M,. Tan, X., Wen, S., Ding, E., Lin, W., & Dou, D. (2020). Discriminative sounding objects localization via self-supervised audiovisual matching. *Advances in Neural Information Processing Systems (NeurIPS)*.

Hu, X., Chen, Z., & Owens, A. (2022). Mix and localize: Localizing sound sources in mixtures. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hu, X., Chen, Z., & Owens, A. (2022). Mix and localize: Localizing sound sources in mixtures. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jang, E., Gu, S., & Poole, B. (2017). Categorical reparameterization with gumbel-softmax. In: *International Conference on Learning Representations (ICLR)*.

Jia, C., Yang, Y., Xia, Y., Chen, Y. T., Parekh, Z., Pham, H., Le, Q., Sung, Y. H., Li, Z., & Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. In: *International Conference on Machine Learning (ICML)*.

Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de Las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W.E. (2023). Mistral 7b. *arXiv preprint* arXiv:2310.06825.

Kim, D., Um, S. J., Lee, S., & Kim, J. U. (2024). Learning to visually localize sound sources from mixtures without prior source knowledge. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li, J., Li, D., Savarese, S., & Hoi, S. (2023). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: *International Conference on Machine Learning (ICML)*.

Li, J., Zhao, W., Huang, Z., Guo, Y., & Tian, Y. (2025). Do audio-visual segmentation models truly segment sounding objects? *arXiv preprint* arXiv:2502.00358.

Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., Chang, K.W., & Gao, J. (2022). Grounded language-image pre-training. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li, S., Tian, Y., & Xu, C. (2021). Space-time memory network for sounding object localization in videos. In: *British Machine Vision Conference (BMVC)*.

Lin, Y. B., Tseng, H. Y., Lee, H. Y., Lin, Y. Y., & Yang, M. H. (2023). *Unsupervised sound localization via iterative contrastive learning*. Computer Vision and Image Understanding (CVIU).

Liu, J., Ju, C., Xie, W., & Zhang, Y. (2022). Exploiting transformation invariance and equivariance for self-supervised sound localisation. In: *ACM MM*.

Lüddecke, T., & Ecker, A. (2022). Image segmentation using text and image prompts. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Mahmud, T., & Marculescu, D. (2023). Ave-clip: Audioclip-based multi-window temporal transformer for audio visual event localization. In: *IEEE Winter Conference on Applications of Computer Vision (WACV)*.

Mahmud, T., Tian, Y., & Marculescu, D. (2024). T-vsl: Text-guided visual sound source localization in mixtures. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Mo, S., & Morgado, P. (2022). A closer look at weakly-supervised audio-visual source localization. *Advances in Neural Information Processing Systems (NeurIPS)*.

Mo, S., & Morgado, P. (2022). Localizing visual sounds the easy way. In: *European Conference on Computer Vision (ECCV)*.

Mo, S., & Tian, Y. (2023). Audio-visual grouping network for sound localization from mixtures. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Nukrai, D., Mokady, R., & Globerson, A. (2022). Text-only training for image captioning using noise-injected clip. In: *Empirical Methods in Natural Language Processing (EMNLP)*.

Oya, T., Iwase, S., Natsume, R., Itazuri, T., Yamaguchi, S., & Morishima, S. (2020). Do we need sound for sound source localization? In: *Asia Conference on Computer Vision (ACCV)*.

Park, S., Senocak, A., & Chung, J. S. (2023). MarginNCE: Robust sound localization with a negative margin. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Park, S., Senocak, A., & Chung, J. S. (2024). Can clip help sound source localization? In: *IEEE Winter Conference on Applications of Computer Vision (WACV)*.

Qian, R., Hu, D., Dinkel, H., Wu, M., Xu, N., & Lin, W. (2020). Multiple sound sources localization from coarse to fine. In: *European Conference on Computer Vision (ECCV)*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., & Clark, J. (2021). Learning transferable visual models from natural language supervision. In: *International Conference on Machine Learning (ICML)*.

Ryu, H., Kim, S., Chung, J.S., & Senocak, A. (2025). Seeing speech and sound: Distinguishing and locating audios in visual scenes. *arXiv preprint* arXiv:2503.18880.

Senocak, A., Oh, T. H., Kim, J., Yang, M. H., & Kweon, I. S. (2018). Learning to localize sound source in visual scenes. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Senocak, A., Oh, T. H., Kim, J., Yang, M. H., & Kweon, I. S. (2021). Learning to localize sound sources in visual scenes: Analysis and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 43*(5), 1605–1619.

Senocak, A., Ryu, H., Kim, J., & Kweon, I. S. (2022). Learning sound localization better from semantically similar samples. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Senocak, A., Ryu, H., Kim, J., & Kweon, I. S. (2022). Less can be more: Sound source localization with a classification model. In: *IEEE Winter Conference on Applications of Computer Vision (WACV)*.

Senocak, A., Ryu, H., Kim, J., Oh, T. H., Pfister, H., & Chung, J. S. (2023). Sound source localization is all about cross-modal alignment. In: *IEEE International Conference on Computer Vision (ICCV)*.

Senocak, A., Ryu, H., Kim, J., Oh, T. H., Pfister, H., & Chung, J. S. (2025). Toward interactive sound source localization. *Better align sight and sound! IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

Song, Z., Wang, Y., Fan, J., Tan, T., & Zhang, Z. (2022). Self-supervised predictive learning: A negative-free method for sound source localization in visual scenes. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sun, W., Zhang, J., Wang, J., Liu, Z., Zhong, Y., Feng, T., Guo, Y., Zhang, Y., & Barnes, N. (2023). Learning audio-visual source localization via false negative aware contrastive learning. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sung-Bin, K., Hyun-Bin, O., Lee, J., Senocak, A., Chung, J. S., & Oh, T. H. (2025). Avhbench: A cross-modal hallucination benchmark for audio-visual large language models. In: *International Conference on Learning Representations (ICLR)*.

Tan, R., Ray, A., Burns, A., Plummer, B. A., Salamon, J., Nieto, O., Russell, B., & Saenko, K. (2023). Language-guided audio-visual source separation via trimodal consistency. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wu, H. H., Seetharaman, P., Kumar, K., & Bello, J. P. (2022). Wav2CLIP: Learning robust audio representations from CLIP. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Wu, Y., Chen, K., Zhang, T., Hui, Y., Berg-Kirkpatrick, T., & Dubnov, S. (2023). Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Xie, J., Hou, X., Ye, K., & Shen, L. (2022). Clims: Cross language image matching for weakly supervised semantic segmentation. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xuan, H., Wu, Z., Yang, J., Yan, Y., & Alameda-Pineda, X. (2022). A proposal-based paradigm for self-supervised sound source localization in videos. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yariv, G., Gat, I., Wolf, L., Adi, Y., & Schwartz, I. (2023). AudioToken: Adaptation of text-conditioned diffusion models for audio-to-image generation. In: *INTERSPEECH*.

Zhou, J., Wang, J., Zhang, J., Sun, W., Zhang, J., Birchfield, S., Guo, D., Kong, L., Wang, M., & Zhong, Y. (2022). Audio-visual segmentation. In: *European Conference on Computer Vision (ECCV)*.