# Can CLIP Help Visual Sound Localization?

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

[1] Korea Advanced Institute of Science and Technology, South Korea
[2] Electronics and Telecommunications Research Institute, South Korea

## Abstract

*By leveraging large-scale image-text foundation models such as CLIP, our method advances sound source localization without the need for explicit text input. Our approach transforms audio signals into tokens compatible with CLIP, enabling the generation of precise sound localization maps through audio-driven embeddings. By aligning audio-grounded image features with these embeddings, our approach outperforms state-of-the-art methods, demonstrating the effectiveness of leveraging pre-trained image-text models in sound localization tasks.*

## 1. Introduction

The ability to accurately localize the origin of sound is crucial for environmental awareness. Humans and other organisms continuously receive multisensory information, such as auditory and visual inputs, enabling them to comprehend relationships among stimuli, infer the entities or events producing sound, and focus on those generating the auditory cues. One fundamental approach in machine learning to acquire such abilities entails leveraging the natural correspondence between auditory and visual signals, without requiring explicit supervision or annotations. A primary method for achieving this involves aligning auditory-visual representations within a contrastive learning framework, facilitating self-supervised learning.

To enhance sound localization methods, research has proposed leveraging additional prior knowledge, such as visual objectness [4, 5] or object proposal networks [17]. However, these approaches may introduce biases that impede genuine audio-visual alignment. In our study, we prioritize the utilization of robust multimodal alignment knowledge to improve genuine sound source localization. To achieve this, we employ the foundation model, especially CLIP (Contrastive Language-Image Pretraining) [8] model, renowned for its robust representation and alignment capabilities learned directly from large-scale raw text and images. By doing so, we offer broader supervision beyond limited category labels.

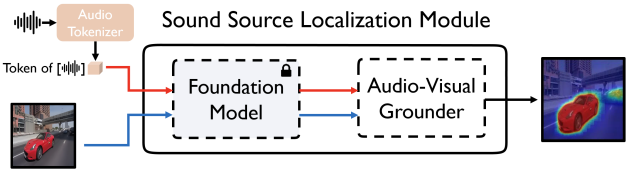


Figure 1. **The proposed sound source localization method based on foundation model.**

Typically, CLIP models utilize text queries instead of audio prompts. We investigated methods to incorporate audio prompts into the CLIP model (as illustrated in Figure 1). To achieve this goal, we introduce a framework that generates contextual embeddings for provided audio inputs by converting audio signals into tokens compatible with CLIP's text encoder, referred to as "audio-driven embeddings." Our primary concept involves harmonizing audio and visual features through autonomous alignment of learning modules. This process integrates audio-driven embeddings to accentuate regions within visual scenes where sounds occur, extract audio-based visual features, and facilitate audio-visual correspondence within a comparative learning framework. The entire proposed architecture is trained with the objective of audio-visual alignment. Through experimentation, we validate the efficacy of our proposed method, achieving superior performance compared to existing approaches and baselines.

## 2. Method

### 2.1. Audio-Driven Embedder

Our objective is to integrate audio inputs into the CLIP text encoder framework without relying on textual prompts. To achieve this, we employ the Audio Tokenizer module,

*These authors contributed equally to this work. This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government, [24ZC1100, The research of the basic media·contents technologies] and the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2023-00259991) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

which converts audio context into text token. Similar to [18], audio features are extracted from a pre-trained audio encoder $E_A$ and projected into the textual token space, creating CLIP-compatible tokens for CLIP text encoder. During training, the audio encoder remains fixed, and only the projection layer is trained end-to-end using the audio-visual alignment objective.

### 2.2. Audio-Visual Grounder

Our audio-visual grounder detects sound regions in input audio-visual pairs, generating masks for further processing. These masks guide the extraction of visual embeddings at both image and feature levels for audio-visual alignment. We utilize a pre-trained CLIP image encoder ($E_{\text{CLIP}_v}$) to encode input images into global and spatial features. Our grounder ($G$) utilizes CLIPSeg, an off-the-shelf CLIP-based segmentation network, to identify sound regions. Notably, while CLIPSeg typically relies on CLIP visual features and text condition for generating segmentation mask, we indicate potential sound regions using our audio-driven embedding ($\mathbf{A}$) instead of text condition. Throughout the training process, both the image encoder and the audio-visual grounder remain fixed.

### 2.3. Audio-Visual Alignment

After acquiring sound region masks through our audio-visual grounding module, our approach proceeds to extract visual embeddings from these masked regions, operating at both the image and feature levels. These extracted embeddings are then aligned with the audio-driven embedding $\mathbf{A}$ to fulfill the audio-visual alignment objective. To facilitate this alignment process, we utilize contrastive learning losses. Fundamentally, our model is designed to optimize the alignment between visual and audio features within the CLIP space.

## 3. Experiments

**Datasets.** Our method is trained on the VGG-Sound dataset [1], comprising roughly 200K videos without explicit annotations. Sound localization evaluation is conducted on VGG-SS [2] and SoundNet-Flickr-Test [9, 10], with about 5K and 250 annotated samples, respectively. Further evaluations include AVSBench [19], which provides binary segmentation maps categorizing audio-visually related pixels into Single-source (S4) and Multi-source (MS3) subsets. Extended VGG-SS/SoundNet-Flickr [5] datasets are also utilized to explore scenarios with non-visible sound sources."

### 3.1. Quantitative Results

**Baseline.** In addition to existing methods, we compare our approach with several closely-related baselines derived from different components of our overall architecture:

- **CLIPSeg w/ GT Text:** This serves as an oracle method where ground truth sound event class labels of test samples are used as text conditions for grounder, CLIPSeg.
- **CLIPSeg w/ WAV2CLIP Text:** Utilizing WAV2CLIP, relevant text (class label) is retrieved for a given audio, which is then used for grounder, CLIPSeg.
- **CLIPSeg - Sup. AudioTokenizer:** Here, the AudioTokenizer module is trained in a supervised manner using ground truth text labels, and the resulting audio-driven embeddings are employed with grounder, CLIPSeg.
- **WAV2CLIP and AudioCLIP:** These models use pre-trained CLIP embeddings for zero-shot sound localization. A CLIP-like object detector extracts region proposals from images and selects the top-1 proposal based on cosine similarity between visual and audio features.

| Method | VGG-SS cIoU ↑ | VGG-SS AUC ↑ | SoundNet-Flickr cIoU ↑ | SoundNet-Flickr AUC ↑ |
|---|---|---|---|---|
| Attention [9]$_{\text{CVPR18}}$ | 18.50 | 30.20 | 66.00 | 55.80 |
| CoarseToFine [7]$_{\text{ECCV20}}$ | 29.10 | 34.80 | - | - |
| LCBM [12]$_{\text{WACV22}}$ | 32.20 | 36.60 | - | - |
| LVS [2]$_{\text{CVPR21}}$ | 34.40 | 38.20 | 71.90 | 58.20 |
| HardPos [11]$_{\text{ICASSP22}}$ | 34.60 | 38.00 | 76.80 | 59.20 |
| SSPL [14]$_{\text{CVPR22}}$ | 33.90 | 38.00 | 76.70 | 60.50 |
| EZ-VSL (w/o OGL) [4]$_{\text{ECCV22}}$ | 35.96 | 38.20 | 78.31 | 61.74 |
| EZ-VSL (w/ OGL) [4]$_{\text{ECCV22}}$ | 38.85 | 39.54 | 83.94 | 63.60 |
| SLAVC (w/o OGL) [5]$_{\text{NeurIPS22}}$ | 37.79 | 39.40 | 83.60 | - |
| SLAVC (w/ OGL) [5]$_{\text{NeurIPS22}}$ | 39.80 | - | **86.00** | - |
| MarginNCE (w/o OGL) [6]$_{\text{ICASSP23}}$ | 38.25 | 39.06 | 83.94 | 63.20 |
| MarginNCE (w/ OGL) [6]$_{\text{ICASSP23}}$ | 39.78 | 40.01 | 85.14 | 64.55 |
| FNAC (w/o OGL) [15]$_{\text{CVPR23}}$ | 39.50 | 39.66 | 84.73 | 63.76 |
| FNAC (w/ OGL) [15]$_{\text{CVPR23}}$ | 41.85 | 40.80 | 85.14 | 64.30 |
| Alignment (w/o OGL) [13]$_{\text{ICCV23}}$ | 39.94 | 40.02 | 79.60 | 63.44 |
| Alignment (w/ OGL) [13]$_{\text{ICCV23}}$ | 42.64 | 41.48 | 82.40 | 64.60 |
| *Baselines:* | | | | |
| WAV2CLIP [16]$_{\text{ICASSP22}}$ | 37.71 | 39.93 | 26.00 | 29.60 |
| AudioCLIP [3]$_{\text{ICASSP22}}$ | 44.15 | 46.23 | 47.20 | 45.22 |
| CLIPSeg w/ GT Text (***Oracle***) | **49.50** | **48.62** | - | - |
| CLIPSeg w/ WAV2CLIP Text | 24.84 | 26.01 | 37.20 | 32.14 |
| CLIPSeg Sup. AudioTokenizer | 49.09 | 45.75 | 68.00 | 54.96 |
| **Ours (w/o OGL)** | <u>49.46</u> | <u>46.32</u> | 80.80 | **64.62** |

Table 1. **Comparison of Localization Performance on VGG-SS and SoundNet-Flickr**. Each model is trained using 144K paired image-audio samples from VGG-Sound.

**Comparison on standard benchmarks.** In Table 1 and Table 2, we present a comparative analysis of our self-supervised method for localizing sound sources against existing methodologies and baselines. Evaluation on the VGG-SS and SoundNet-Flickr test sets demonstrates notable performance enhancements, highlighting the significance of leveraging CLIP's multimodal alignment knowledge. Despite the absence of explicit text input, our AudioTokenizer proficiently encodes audio context, facilitating adept learning of audio-visual correspondence. Comparative assessments against robust baselines reveal that our approach outperforms or achieves comparable performance. Furthermore, our self-supervised strategy surpasses previ-

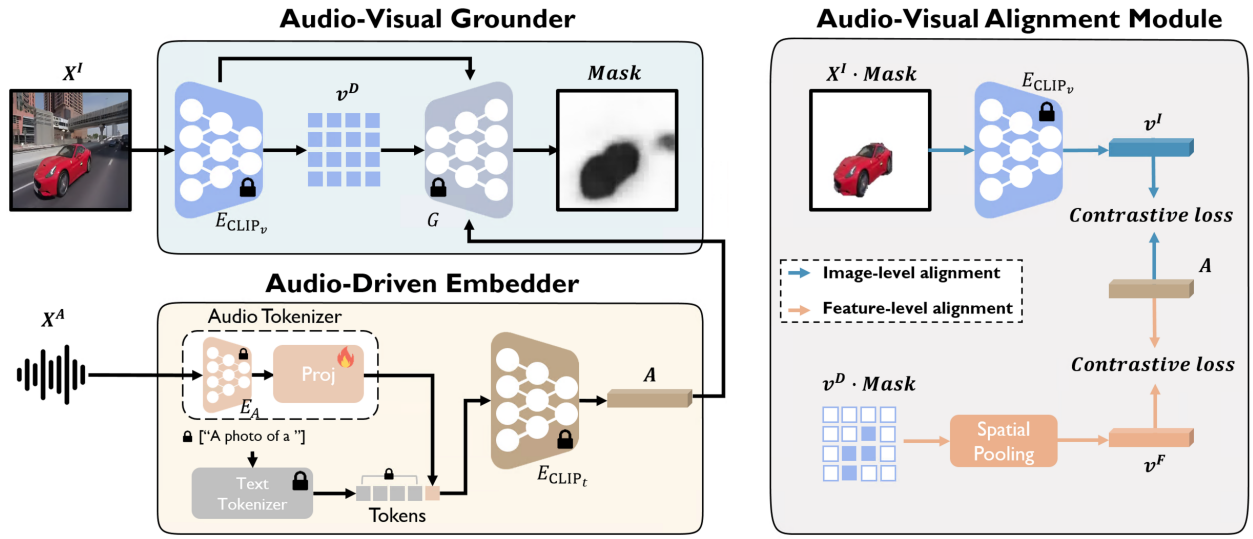


Figure 2. **Overview of our sound source localization framework.** Our method utilizes audio-visual pairs, where audio signals are transformed into CLIP-compatible tokens through the Audio Tokenizer module to produce the audio-driven embedding, $\mathbf{A}$. This embedding emphasizes regions with sound activity in the Audio-Visual Grounder module. Subsequently, the Audio-Visual Alignment module utilizes sounding area masks to extract audio-grounded visual features at both image and feature levels ($\boldsymbol{v}^I$ and $\boldsymbol{v}^F$). These features are then aligned with audio features using contrastive learning.

| Method | Extended VGG-SS | | | Extended Flickr | | | AVSBench S4 | | AVSBench MS3 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | AP ↑ | max-F1 ↑ | LocAcc ↑ | AP ↑ | max-F1 ↑ | LocAcc ↑ | mIoU ↑ | F-Score ↑ | mIoU ↑ | F-Score ↑ |
| SLAVC (w/o OGL) [5]$_{\text{NeurIPS22}}$ | 32.95 | 40.00 | 37.79 | 51.63 | 59.10 | 83.60 | 28.10 | 34.60 | 24.37 | 25.56 |
| MarginNCE (w/o OGL) [6]$_{\text{ICASSP23}}$ | 30.58 | 36.80 | 38.25 | 57.99 | 61.80 | 83.94 | 33.27 | 45.33 | 27.31 | 31.56 |
| FNAC (w/o OGL) [15]$_{\text{CVPR23}}$ | 23.48 | 33.70 | 39.50 | 50.40 | 62.30 | **84.73** | 27.15 | 31.40 | 21.98 | 22.50 |
| Alignment (w/o OGL) [13]$_{\text{ICCV23}}$ | 34.73 | 40.70 | 39.94 | 64.43 | 66.90 | 79.60 | 29.60 | 35.90 | - | - |
| *Baselines:* | | | | | | | | | | |
| WAV2CLIP [16]$_{\text{ICASSP22}}$ | 26.67 | 33.00 | 37.71 | 20.99 | 24.80 | 29.60 | 28.70 | 35.35 | 25.09 | 23.84 |
| CLIPSeg w/ GT Text (***Oracle***) | - | - | - | - | - | - | 51.32 | 58.02 | **50.93** | **55.41** |
| CLIPSeg w/ WAV2CLIP Text | - | - | - | - | - | - | 26.52 | 30.60 | 30.82 | 29.97 |
| AudioCLIP [3]$_{\text{ICASSP22}}$ | 23.79 | 32.80 | 44.15 | 34.00 | 38.80 | 45.22 | 36.57 | 42.15 | 27.06 | 26.48 |
| CLIPSeg Sup. AudioTokenizer | 34.96 | 41.00 | 49.09 | 55.14 | 57.00 | 68.00 | 49.82 | 56.43 | 42.57 | 46.72 |
| **Ours (w/o OGL)** | **40.79** | **49.10** | **49.46** | **76.07** | **73.20** | 80.80 | **59.76** | **69.03** | 41.08 | 46.67 |

Table 2. **Comparison of Localization Performance on Extended VGG-SS/Flickr-SoundNet and AVSBench.** Each model is trained using 144K paired image-audio samples from VGG-Sound. Extended VGG-Sound/Flickr-SoundNet does not provide appropriate text labels for non-sounding pairs.

ous techniques trained on image-audio pairs, underscoring the efficacy of audio-based embedding modules in acquiring robust audio-visual alignments.

### 3.2. Qualitative Results

**Comparison with existing methodologies.** Figure 3 presents a comparative analysis between our proposed approach and recent prior research. The visual examples highlight the superior accuracy and granularity of our localization results compared to alternative methods. Notably, our model demonstrates robust performance in identifying small-sized sound sources across diverse test scenarios, outperforming recent methodologies in terms of precision. Furthermore, our approach excels in distinguishing and isolating multiple sound sources, a capability not effectively realized by other methodologies, which often aggregate sound sources within a single, expansive region.

## 4. Conclusion

In our study, we explore the use of large-scale pretrained image-text models like CLIP for sound source localization. Our goal is to leverage CLIP's multimodal alignment capabilities without relying on text inputs, achieved through self-supervised audio-visual correspondence. This involves converting audio signals into CLIP-compatible tokens and using the resulting audio-driven embeddings for audio-visual grounding. Through contrastive learning, our method enables self-supervised audio-visual alignment. We demonstrate the superior performance of our model over existing methods in coarse-grained sound source localization and fine-grained segmentation tasks, even when compared

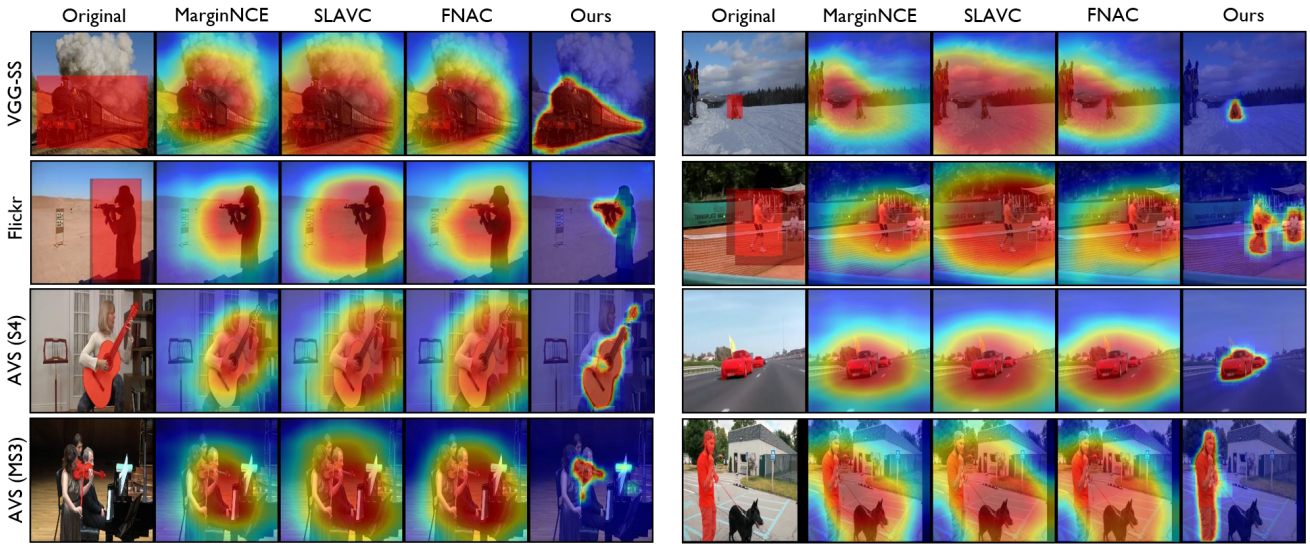


Figure 3. **Comparision of visualization results across VGG-SS, SoundNet-Flickr, and AVSBench datasets with previous methods.**

to fully supervised or text-queried baselines. This highlights the potential of structured multimodal alignment from large-scale pretrained image-text models in sound source localization.

## References


[1] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP*, 2020. 2

[2] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *CVPR*, 2021. 2

[3] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. AudioCLIP: Extending CLIP to image, text and audio. In *ICASSP*, 2022. 2, 3

[4] Shentong Mo and Pedro Morgado. Localizing visual sounds the easy way. In *ECCV*, 2022. 1, 2

[5] Shentong Mo and Pedro Morgado. A closer look at weakly-supervised audio-visual source localization. *NeurIPS*, 2022. 1, 2, 3

[6] Sooyoung Park, Arda Senocak, and Joon Son Chung. MarginNCE: Robust sound localization with a negative margin. In *ICASSP*, 2023. 2, 3

[7] Rui Qian, Di Hu, Heinrich Dinkel, Mengyue Wu, Ning Xu, and Weiyao Lin. Multiple sound sources localization from coarse to fine. In *ECCV*, 2020. 2

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Int. Conf. on Mach. Learn.*, 2021. 1

[9] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *CVPR*, 2018. 2

[10] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound sources in visual scenes: Analysis and applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(5):1605–1619, 2021. 2

[11] Arda Senocak, Hyeonggon Ryu, Junsik Kim, and In So Kweon. Learning sound localization better from semantically similar samples. In *ICASSP*, 2022. 2

[12] Arda Senocak, Hyeonggon Ryu, Junsik Kim, and In So Kweon. Less can be more: Sound source localization with a classification model. In *IEEE Winter Conf. on Appli. of Comput. Vis.*, 2022. 2

[13] Arda Senocak, Hyeonggon Ryu, Junsik Kim, Tae-Hyun Oh, Hanspeter Pfister, and Joon Son Chung. Sound source localization is all about cross-modal alignment. In *ICCV*, 2023. 2, 3

[14] Zengjie Song, Yuxi Wang, Junsong Fan, Tieniu Tan, and Zhaoxiang Zhang. Self-supervised predictive learning: A negative-free method for sound source localization in visual scenes. In *CVPR*, 2022. 2

[15] Weixuan Sun, Jiayi Zhang, Jianyuan Wang, Zheyuan Liu, Yiran Zhong, Tianpeng Feng, Yandong Guo, Yanhao Zhang, and Nick Barnes. Learning audio-visual source localization via false negative aware contrastive learning. In *CVPR*, 2023. 2, 3

[16] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2CLIP: Learning robust audio representations from CLIP. In *ICASSP*, 2022. 2, 3

[17] Hanyu Xuan, Zhiliang Wu, Jian Yang, Yan Yan, and Xavier Alameda-Pineda. A proposal-based paradigm for self-supervised sound source localization in videos. In *CVPR*, 2022. 1

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

[19] Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio-visual segmentation. In *ECCV*, 2022. 2