Toward Interactive Sound Source Localization: Better Align Sight and Sound!

Arda Senocak*, Hyeonggon Ryu*, Junsik Kim*, Tae-Hyun Oh, Hanspeter Pfister and Joon Son Chung

Abstract—Recent studies on learning-based sound source localization have primarily focused on localization performance. However, prior work and existing benchmarks often overlook a crucial aspect: cross-modal interaction, which is essential for interactive sound source localization. This interaction is vital for understanding semantically matched or mismatched audio-visual events, such as silent objects or true sound sources among multiple objects. In this work, we comprehensively examine the cross-modal interaction of existing methods, benchmarks, evaluation metrics, and cross-modal understanding tasks. We identify the overlooked points of previous studies and make several contributions to address them. First, we propose a learning framework that incorporates retrieval-based and hand-crafted augmentation techniques, enhancing cross-modal interaction through cross-modal alignment. Second, we introduce new evaluation metrics to accurately and rigorously assess localization methods, focusing on both localization performance and cross-modal interaction. Third, to thoroughly analyze interactive sound source localization, we present a new semi-synthetic benchmark with diverse categorical combinations. Finally, we evaluate both interactive sound source localization and auxiliary cross-modal retrieval tasks, benchmarking competing methods alongside our own. Our new benchmark and evaluation metrics reveal that previous methods struggle with interactive sound source localization tasks, largely due to their limited cross-modal interaction capabilities. Our method, which features enhanced cross-modal alignment, demonstrates superior sound source localization and cross-modal interaction performance . This work provides the most comprehensive analysis of sound source localization to date, with extensive validation of competing methods on both existing and new benchmarks using both new and standard evaluation metrics. Code is available at https://github.com/kaistmm/SSLalignment

Index Terms—Audio-visual learning, sound source localization, self-supervision, multi-modal learning, cross-modal retrieval.

1 Introduction

HUMANS can effortlessly determine the origin of sounds in a scene. We instinctively focus on the direction of the sound and associate the incoming audio-visual signals to comprehend the event. Achieving human-level audio-visual perception has led to extensive research on sound source localization in visual scenes [2], [7], [18], [27], [32], [34]–[37], [45], [50]–[53], [55]. Inspired by the way humans learn from natural audio-visual correspondences without explicit supervision, most studies are based on the fundamental assumption that audio and visual signals are temporally correlated. Under this assumption, the losses in sound source localization tasks are modeled using audio-visual correspondence as a self-supervision signal, implemented through contrastive learning of audio-visual pairs.

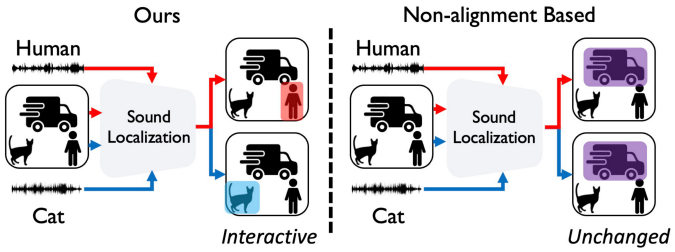


Fig. 1. **Our alignment-based sound source localization method enhances cross-modal interaction capabilities.** The key conceptual difference between prior approaches and our proposed model is that our model accurately follows cross-modal interactions for various given sounds, while competing methods tend to focus on visually dominant areas in a scene, regardless of the accompanying audio.

While these approaches appear to be unsupervised methods, they heavily rely on partial supervision; for instance, using pre-trained vision networks [18], [45], [50], [51], [53], [55] and visual objectness estimators for post-processing [36], [37]. Recent studies [43] have pointed out the visual objectness bias in existing sound source localization benchmarks and some works [36], [37] have explicitly exploited this bias to enhance localization accuracy. These studies demonstrate that a model can achieve high localization accuracy by relying solely on visual signals even without interaction between visual and audio signals, which contradicts the true objective of the sound source localization task. In short, the current evaluation, benchmark and model settings for sound source localization do not adequately capture the audio-visual interaction capability, as illustrated in Figure 1.

In this paper, we comprehensively examine the cross-modal interaction of sound source localization methods by proposing a new benchmark, cross-modal alignment, evaluation metrics, and cross-modal understanding tasks.

- *A. Senocak, H. Ryu, and J.S. Chung are with School of Electrical Engineering, KAIST, Daejeon, Republic of Korea.*
- *T.H. Oh is with Department of Electrical Engineering, POSTECH, Pohang, Republic of Korea.*
- *J.Kim and H.Pfister are with School of Engineering and Applied Sciences at Harvard University, Boston, MA, USA.*
- **These authors contributed equally to this work.*

We first point out that existing sound source localization benchmarks inadequately capture audio-visual interaction, which is essential for sound source localization. This inadequacy arises mainly from two issues. Firstly, most sounding objects in existing benchmarks dominate the scene, allowing the use of objectness priors to solve the problem without proper audio-visual interaction, *i.e.*, hallucination [28]. This approach becomes ineffective if sounding objects are smaller than silent objects or are not visible. Secondly, we need multiple sounding sources and their sounds paired with a single visual data to authentically assess the audio-visual interaction capabilities of methods in a systematic way, but there is no large-scale or reliable dataset suitable for the purpose. Most commonly used large-scale benchmarks involve single sounding source samples, while the few multiple sounding source benchmarks have limited categories and sample sizes. To address these issues, we create a new semi-synthetic sound source localization benchmark that includes diverse categories of objects with varied combinations and background contexts. Each sample in our benchmark contains multiple sounding source objects with corresponding audio, enabling the evaluation of audio-visual interaction by testing the same image with different audio pairs. This allows us to conduct systematic experiments with clear control variables.

Secondly, we revisit the importance of semantic understanding shared across audio and visual modalities. Previous methods [45], [50], [51], [55] induce cross-modal semantic alignment through instance-level cross-modal contrastive learning, *i.e.*, cross-modal instance discrimination between visual and audio features. However, they rely on labels or supervisedly pre-trained encoders [1]. We show that learning better semantic feature alignment is a crucial recipe and enables to achieve higher performance even with training from scratch, compared to the methods exploiting strong pre-trained models.

We also propose two new evaluation metrics to address issues overlooked in previous sound source localization studies. Initially, we observe that while our proposed method outperforms others numerically, there was a discrepancy between the improved numbers and the superior qualitative results. Upon further analysis, we identify that the existing cIoU evaluation metric was a limiting factor, underestimating our model's superior performance, especially in accurately localizing small objects. The current cIoU metric uses a fixed threshold size, which becomes less accurate when the ground truth area differs from the threshold value. To address this, we propose an adaptive version of cIoU to accurately measure localization performance regardless of ground truth sizes. Additionally, we introduce Interactive IoU (IIoU) as a new metric to measure cross-modal interaction capability in multiple sound source scenarios. Unlike single-source scenarios, IIoU considers a method successful only when it predicts all sounding sources in a scene paired with different audio signals.

We further comprehensively benchmark sound source localization methods for cross-modal retrieval tasks to analyze their cross-modal interaction capabilities. This task assesses whether the learned representations can accurately interact between audio and visual modalities, indicating fine-grained audio-visual correspondence essential for genuine sound source localization. Our benchmarking shows the importance of cross-modal interaction, demonstrating that higher sound source localization performance on sound source localization benchmarks does not guarantee higher cross-modal retrieval performance. This finding highlights the need to evaluate sound source localization methods from a more diverse perspective, supporting our contributions of proposing a new benchmark, evaluation metrics, and learning cross-modal alignment.

In short, we extensively benchmark our method and competing methods on diverse sound source localization scenarios using seven benchmarks with new evaluation metrics. Additionally, we evaluate our method on interactive sound source localization and retrieval tasks. Our proposed method performs favorably against recent state-of-the-art approaches in comprehensive experiments, demonstrating its ability to excel in both localization and alignment tasks, closely approaching the ideal scenario.

We summarize the contributions of our work as follows:

- We analyze existing sound source localization benchmarks and identify their inadequacy in evaluating true cross-modal semantic understanding, which may lead to poor performance in interactive sound source localization and cross-modal retrieval tasks.
- We construct a new benchmark specifically designed for the evaluation of interactive sound source localization.
- We propose a novel method that utilizes semantic alignment with multi-views and semantically similar samples to achieve state-of-the-art performance in both sound source localization and cross-modal retrieval.
- We introduce new evaluation metrics to comprehensively analyze the cross-modal interaction capabilities of sound source localization methods.
- We extensively benchmark our method and competing methods on sound source localization, cross-modal retrieval tasks and audio-visual segmentation, providing the most comprehensive analysis of cross-modal interaction capabilities of existing methods to date.

This work builds upon our previous conference paper [54], which explored the importance of cross-modal alignment in sound source localization. This journal submission offers a substantial extension in three key aspects. First, we introduce a new semi-synthetic benchmark for interactive sound source localization, providing a more rigorous and quantitative evaluation framework (Section 4.1). Second, we propose new evaluation metrics designed to accurately measure interactive sound source localization, addressing the limitations of previous methods (Section 4.2). Third, we expand our experimental evaluations significantly, including extensive tests on multiple audio-visual datasets, and detailed comparisons with state-of-the-art methods in both localization and cross-modal interaction tasks (Section 5). These enhancements provide a robust foundation for evaluating cross-modal interactivity in sound source localization.

[1]Typically, an image encoder is pre-trained on ImageNet [15] and an audio encoder is pre-trained on AudioSet [19] in supervised ways.

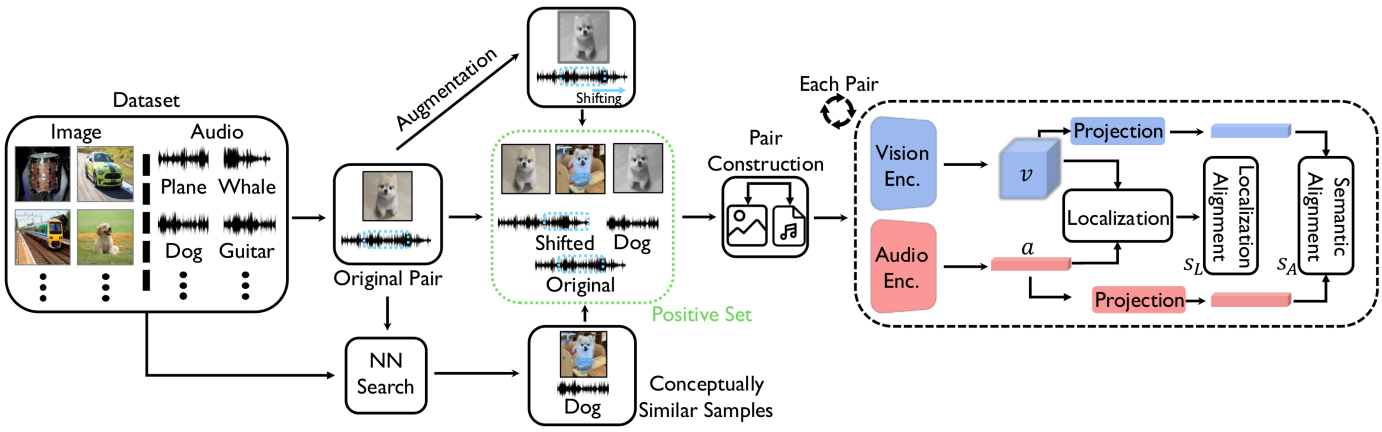


Fig. 2. **Our sound source localization framework.** Our model construct multiple positive pairs with augmentation and Nearest Neighbor Search (Semantically Similar Samples). By using these newly constructed 9 pairs, our model employs spatial localization, $s_L(\cdot)$, and semantic feature alignment, $s_A(\cdot)$, for each pair to learn a better sound source localization ability.

2 Related Work

Sound source localization. Sound source localization in visual scenes has been investigated by exploiting correspondences between audio and visual modalities. The most widely used approach for sound source localization is cross-modal attention [50], [51], [58] with contrastive loss [12], [25], [40]. Later, the attention-based method is improved by intra-frame hard sample mining [7], iterative contrastive learning with pseudo labels [34], feature regularization [35], positive mining [52], negative mining [56], negative free learning [55] with stop-gradient operation [11], momentum encoders [36], or modified contrastive loss [44].

Some sound source localization approaches exploit additional knowledge of objectness, *e.g.*, semantic labels [32], [45], [53] or object prior [37], [62]. Semantic labels are used to pretrain audio and vision encoders with classification loss [32], [53] or refine audio-visual feature alignment [45]. A more explicit way to refine localization output is to use object prior. EZVSL [37] proposes post-processing to combine attention based localization output with a pre-trained visual feature activation map. Similarly, Xuan *et al*. [62] propose to combine off-the-shelf object proposals with attention based sound source localization results. However, post-processing by object prior may generate a false positive output as it is solely based on vision without audio-visual interaction.

Self-supervised representation learning. In a broader categorization, recent sound source localization studies rely on self-supervised multimodal learning. Our work is also relevant to self-supervised audio-visual representation learning, and other multimodal learning studies.

Contrastive learning aims to learn representations from large-scale raw data without annotations. Recent representation learning approaches [9], [10], [24], [60] use instance discrimination by contrastive learning [12], [25], [40] as a pretext task with notable advancements in visual recognition tasks. Recently, positive mining by nearest-neighbor search shows the effectiveness in more sharpening learned representations of images [16], [17], [61], videos [23], neural recordings [4], and text-image [33]. In this work, we extend the previous works by incorporating both multi-views and semantically similar samples into audio-visual modalities for cross-modal feature alignment.

A series of audio-visual representation learning studies have shown that audio and visual contents in a video are correlated, therefore a visual representation can be learned by sound prediction [42] or audio representation can be distilled from visual representation [3], [57]. Later, a variety of joint audio-visual representation learning methods are proposed with an assumption that there is a semantic [1], [26], [38], [39] or temporal [13], [14], [31], [41] correspondence between them. However, simply learning sound source localization by audio-visual correspondence with instance discrimination ignores the semantic similarity of audio-visual contents among samples, introducing false negatives or positives. In order to mitigate this issue, clustering [26], sampling [39], weighting [38], and hard mining [31] are proposed. Similarly, in this work, we go beyond instance discrimination by using multiple positive samples to enforce semantic understanding across modalities.

3 Method

3.1 Preliminaries

Contrastive learning learns representation by using positive and negative pairs. Given an encoded query sample $\mathbf{q}$ and its encoded positive pair $\mathbf{k}^+$ and negative pairs $\mathbf{k}$, the loss can be defined as:

$$\mathcal{L} = -\log \frac{\exp(\mathbf{q} \cdot \mathbf{k}^+/\tau)}{\sum_i \exp(\mathbf{q} \cdot \mathbf{k_i}/\tau)} \tag{1}$$

where τ is the temperature parameter.

Cross-modal contrastive learning extends contrastive learning across multiple modalities. In sound source localization, audio-visual correspondence is used to define positive and negative cross-modal pairs. With an audio-visual dataset $\mathcal{D} = \{(v_i, a_i) : i = 1, ..., N\}$ and its encoded features $\mathbf{v}_i = f_v(v_i)$ and $\mathbf{a}_i = f_a(a_i)$, cross-modal contrastive learning loss is defined as:

$$\mathcal{L}_i = -\log \frac{\exp(s(\mathbf{v}_i, \mathbf{a}_i)/\tau)}{\sum_j \exp(s(\mathbf{v}_i, \mathbf{a}_j)/\tau)} \tag{2}$$

where $s(\cdot)$ is a cross-modal similarity function. The cross-modal contrastive loss Eq. (2) can be extended to symmetric form [46] as used in a few previous works [36], [37].

3.2 Cross-Modal Feature Alignment

We consider both spatial localization and semantic feature alignment for sound source localization. To this end, we use two different similarity functions $s_L(\cdot)$ and $s_A(\cdot)$ for contrastive learning (Eq. (2)), $s_L(\cdot)$ for localization and $s_A(\cdot)$ for cross-modal feature alignment. Our sound source localization framework is illustrated in Figure 2.

Recent studies rely on audio-visual spatial correspondence maps to learn sound source localization by contrasting them. Given a spatial visual feature $\mathbf{v} \in \mathbb{R}^{c \times h \times w}$ and audio feature $\mathbf{a} \in \mathbb{R}^{\mathbf{c}}$, audio-visual similarity with a correspondence map can be calculated as follows:

$$s_L(\mathbf{v}, \mathbf{a}) = \sum_{xy \in M} \frac{1}{|M|} \frac{\mathbf{v}^{xy} \cdot \mathbf{a}}{\|\mathbf{v}^{xy}\| \|\mathbf{a}\|} \tag{3}$$

where $\mathbf{v}^{xy}$ is a feature vector at location (x, y), and M is an optional binary mask when an annotation or pseudo-mask [7], [35] is available. Since we assume no supervision for sound source localization, we do not use any mask, therefore, $M = \{(x, y) \mid 0 \leq x < w,\ 0 \leq y < h\}$.

The contrastive loss with localization similarity $s_L(\cdot)$ enforces location dependent alignment giving sparse but strong audio-visual correspondence which enables to perform localization. However, our empirical studies on cross-modal retrieval indicate that strong localization performance does not guarantee semantic understanding. To overcome the low semantic understanding in recent studies, we propose to add instance-level contrastive loss. Instance-level contrasting encapsulates the whole context in a scene, enforcing better audio-visual semantic alignment. Nonetheless, instance-level contrasting may smooth out spatial discriminativeness learned by Eq. (3). Inspired by SimCLR [9], we adopt a projection layer to align audio-visual semantics in a projection space. The projection layer separates the latent space of localization and semantic alignment, thereby preventing the alignment loss smoothing out the spatial discriminativeness. The similarity function for cross-modal feature alignment is defined as follows:

$$s_A(\mathbf{v}, \mathbf{a}) = \frac{p_v(\mathsf{avg}(\mathbf{v})) \cdot p_a(\mathbf{a})}{\|p_v(\mathsf{avg}(\mathbf{v}))\| \|p_a \mathbf{a}\|} \tag{4}$$

where $\mathsf{avg}(\cdot)$ is spatial average pooling, p_v is a projection layer for visual features, and p_a is a projection layer for audio features.

3.3 Extending with Multiple Positive Samples

Typically, contrastive learning contrasts between one positive pair and multiple negative pairs as shown in Eq. (1). In audio-visual learning, by an audio-visual correspondence assumption, an audio-image pair from the same clip is used as a positive pair while negative pairs are sampled from different clips. However, single-instance discrimination may not be sufficient to achieve strong cross-modal alignment. In this section, we expand contrastive learning beyond single instance discrimination by positive set construction and pairing them. To construct a positive set, we incorporate both hand-crafted positive and semantically similar positive samples for each modality. Later, we adjust the contrastive learning to incorporate multiple positive pairs to enforce cross-modal alignment.

Obtaining hand-crafted positive samples. Using randomly augmented samples as positive multi-view pairs are widely adopted in self-supervised representation learning, *i.e.*, instance discrimination. Similarly, we extend a single anchor audio-image pair to multiple positive pairs by applying simple augmentations on image and audio samples separately. While we utilize common image transformations on images, we apply temporal shifting to audios. It is worth noting that sound source localization task learns from the underlying semantic consistency rather than subtle time differences as in videos. Thus, a slight shift in the audio does not alter contextual information significantly. As a result of hand-crafted multi-view positive pair generation, we obtain additional $\mathbf{v}^{aug}$ and $\mathbf{a}^{aug}$ samples.

Obtaining semantically similar positive samples. Apart from manually created augmented views, we additionally expand our positive set with semantically similar samples. The sampling strategy with nearest neighbor search can be performed in a various way, such as on-the-fly sampling [16], [33], [48], [61], sampling by pre-trained encoders [52], or guided sampling [17], [23] using another modality. For selecting our semantically similar samples, we utilize pre-trained encoders. Note that pre-trained encoders trained either with supervised or self-supervised learning are effective in positive sample mining as shown in the experiment section. By employing readily available image and audio encoders, we use the k-nearest neighbor search to sample semantically similar samples in both modalities. In particular, given a pair of image and audio, we compute cosine similarity with all other samples and choose the top-k most similar samples among the training set for each modality. From a set of k samples, we randomly select one sample to obtain semantically similar samples for each modality, $\mathbf{v}^{conc.}$ and $\mathbf{a}^{conc.}$. By utilizing the semantically similar samples as positive samples, our model expands semantic understanding, which allows us to sharpen the expressiveness and boundaries of the representations between positives and negatives.

Pair construction. Once we obtain the semantically similar and hand-crafted positive samples for each modality, we proceed to create 9 distinct audio-visual pairs by pairing $\mathbf{V} = \{\mathbf{v}, \mathbf{v}^{aug}, \mathbf{v}^{conc}\}$ and $\mathbf{A} = \{\mathbf{a}, \mathbf{a}^{aug}, \mathbf{a}^{conc}\}$. This is done to ensure semantic alignment and consistency between them through contrastive learning. It is worth noting that some of these pairs are a combination of hand-crafted and semantically similar samples, which further enhances the feature alignment of our model during training. The negative pairs are randomly paired from the remaining samples in the mini-batch.

3.4 Training

Our loss formulation incorporates both localization and instance-level similarity functions with multiple positive pairs constructed by augmentation and semantically similar sample search. The final loss term is defined as follows:

$$\mathcal{L}_i = -\sum_{\mathbf{v}_i \in \mathbf{V}} \sum_{\mathbf{a}_i \in \mathbf{A}} \left[\log \frac{\exp(s_L(\mathbf{v}_i, \mathbf{a}_i)/\tau)}{\sum_j \exp(s_L(\mathbf{v}_i, \mathbf{a}_j)/\tau)} + \log \frac{\exp(s_A(\mathbf{v}_i, \mathbf{a}_i)/\tau)}{\sum_j \exp(s_A(\mathbf{v}_i, \mathbf{a}_j)/\tau)} \right] \tag{5}$$

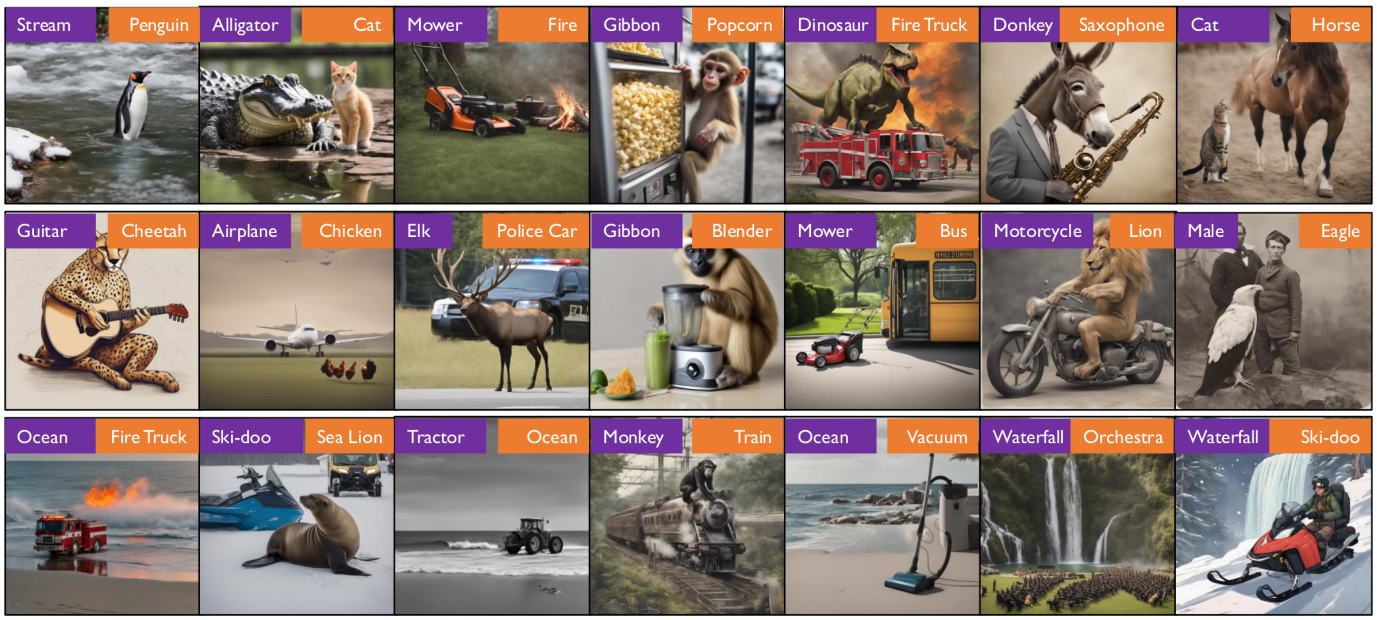


Fig. 3. **IS4 dataset samples.** Each image is generated using the category names indicated in the top left and top right corners. These category names are randomly matched and sourced from the VGG-SS dataset. By leveraging diffusion models, we can create an interactive sound source localization test set with diverse and rare combinations of objects in both realistic and stylized images, such as cartoon or graphic styles.

where $\mathbf{V}$ and $\mathbf{A}$ indicate positive sample sets.

Implementation details. We use two ResNet18 models for both audio and vision encoding. Unlike prior approaches, we do not fine-tune (or use a pre-trained) a visual encoder from ImageNet pre-trained weights. Instead, we train both the audio and vision encoders from scratch. We preprocess images and audios following the previous works [7], [52]. To create multiple pairs, we utilize both NN search and generic augmentation approaches. For NN search, we experiment on two different setups to retrieve k semantically similar samples: (1) For supervisedly pre-trained encoder experiments, We employ ResNet and VGGSound models pre-trained on ImageNet and VGGSound respectively, (2) For self-supervisedly pre-trained encoder experiments, we utilize the CLIP Vision Encoder [46] and Wav2CLIP [59] Audio Encoder. We use k=1000 for the experiments. To perform image augmentations, we follow the augmentations used in SimCLR [9]. For audios, we randomly select time-window shifts in a time axis. The model is trained for 50 epochs with Adam Optimizer and a learning rate of 0.0001. τ is set to 0.07 in contrastive learning.

4 Proposed Evaluation Suite

4.1 Interactive Semi-Synthetic Sound Source (IS4) Dataset.

To our knowledge, there is no large-scale test set for interactive sound source localization (See Figure 1). To address this gap, we introduce a new semi-synthetic test set named IS4. By leveraging diffusion models [47], we generate images containing multiple sounding objects. Compared to the manual collection of real-world samples, generating a synthetic test set is more efficient and accurate, ensuring the presence of multiple objects within each scene. One notable advantage of synthetic data is its controllability. We can synthesize any combination of sounding objects to be appeared in the same scene as shown in Figure 3. Additionally, this dataset offers unusual scenes and unique combinations that are rarely found in nature, such as "a donkey playing a saxophone" or "a sea lion on the snow," seamlessly blended into the generated scenes, unlike cutmix approaches [64]. The dataset features both realistic images and those in other styles, such as cartoon or graphic styles, introducing additional challenges for benchmarking interactiveness with bounding box and mask annotations. We provide details on dataset generation, annotation, and statistics below.

Fig. 4. **IS4 annotations.** Our dataset provides both segmentation maps and bounding box information.

Dataset generation. To construct the dataset, we use Stable Diffusion, combining all the category labels in the VGG-SS dataset as pair combinations. We use the following prompts to generate a synthetic image containing two sounding objects:

TABLE 1
Comparison with the existing sound-source localization and audio-visual segmentation benchmarks. Note that our Interactive Semi-Synthetic Sound Source (IS4) has more unique audio-visual instances.

Benchmark	# Data	# Classes	Multi-Obj.	# Instance	BBox	Seg.
Flickr-SoundNet [50]$_{CVPR18}$	250	~50	✗	250	✓	✗
VGG-SS [7]$_{CVPR21}$	5158	220	✗	5158	✓	✗
AVSBench-S4 [68]$_{ECCV22}$	740	23	✗	740	✗	✓
AVSBench-MS3 [68]$_{ECCV22}$	64	23	✓	2120	✗	✓
VPO-SS [63]$_{CVPR24}$	890	21	✗	890	✗	✓
VPO-MS [63]$_{CVPR24}$	1437	21	✓	2164	✗	✓
ADE20K [22]$_{CVPR24}$	106	20	✗	106	✗	✓
Ours	3240	118	✓	6480	✓	✓

$$\text{Prompt} = \{\text{'a photo of a } (c_1) \text{ and } (c_2)\text{'}, \text{'}(c_1) \text{ is next to } (c_2)\text{'}, \text{'}(c_1) \text{ is playing a } (c_2)\text{'}\} \quad (6)$$

where c_1 and c_2 are the object categories. After generating a large selection of images, we conduct human verification to ensure the images contain both categories and are recognizable. Workers are tasked with eliminating any unsuitable images. Following this filtering process, we randomly match the images with audios from the VGGSound dataset based on their category names.

Annotation. After obtaining the images and audio, we use an online annotation tool[2] that features the Segment Anything Model (SAM) [30] for segmentation. This tool allows users to place keypoints via mouse clicks, yielding high-quality segmentation results. Workers annotate the images based on the given categories, resulting in both segmentation maps and bounding box information. Some of the annotation results are visualized in Figure 4.

Statistics. Our newly proposed dataset includes 3240 images, resulting in 6480 unique audio-visual instances (with 2 objects per image) across 118 categories. Since our dataset provides not only bounding boxes but also segmentation maps, it can be easily utilized in the Audio-Visual Segmentation research as well. Although primarily designed for interactive localization, it also supports standard single sound source evaluation by considering each unique instance individually. A comparison with other Audio-Visual Localization and Segmentation benchmarks is summarized in Table 1. It shows that our dataset is larger in terms of unique instances than any other benchmark. Also, IS4 offers six times more categories than current audio-visual segmentation benchmarks.

4.2 Re-Visiting Evaluation Metrics

To evaluate sound source localization performance, Senocak *et al.* [50] proposed Consensus Intersection over Union (cIoU) as the evaluation metric, which has become the current standard. However, we argue that cIoU alone does not comprehensively capture the necessary evaluation criteria for sound source localization methods. In this paper, we propose two new evaluation metrics to address the limitations overlooked in previous studies: Adaptive cIoU and Interactive IoU.

[2]https://www.cvat.ai/

a) Audio-Visual Attention Map | b) Prediction from cIoU | c) Prediction from Adaptive cIoU

Fig. 5. **Qualitative comparison of cIoU and Adaptive cIoU in the area used for quantitative analysis.** (a), (b), and (c) depict the audio-visual attention map results, the predicted area from the perspective of cIoU, and the perspective of Adaptive cIoU, respectively. The gray color signifies the background. The ground truth bounding box is annotated in green. Although the localization area successfully covers the bounding box in (b), the sample cannot be considered correct since the prediction is much larger than the ground truth. However, Adaptive cIoU better evaluates model performance with small ground truth sizes.

cIoU: Sound source localization methods produce per-pixel response indicating whether visual and audio signals correspond. Benchmarks provide either bounding boxes or object masks as ground truth. A suitable threshold must be set to identify the sounding regions from the predicted per-pixel probabilities, which are then compared to the ground truth. However, determining an optimal threshold is challenging and under-explored in the sound source localization field.

When calculating the standard cIoU, the threshold is set to the *top 50% value of* the audio-visual attention map. Then, Intersection over Union (IoU) is calculated between the ground truth bounding box and this localized area. After calculating the IoU of each sample, samples with an IoU higher than 0.5 are counted as correct samples. This metric was introduced with the Flickr-SoundNet benchmark [50], where most sounding regions are large; therefore, heuristically setting a threshold with a relatively large value of top 50% of values in attention works reasonably. However, this heuristic threshold becomes problematic with smaller sounding objects. Since the pixels having attention values larger than the top 50% turn out 50% of area of an image, this predicted region is always larger than the ground truth smaller than 50% of image size, yielding low IoU values. Due to this sensitivity of cIoU to the size of ground truth objects, even with accurate predictions, the evaluation metric considers it as a failure, as shown in Figure 5.

Adaptive cIoU: To address this issue, we propose a new metric named Adaptive cIoU. Instead of using the top 50% pixels, Adaptive cIoU adaptively considers the top B pixels, where B is the area of the ground truth. Adaptive cIoU avoids the challenge of setting an arbitrary threshold and focuses solely on measuring audio-visual correspondence. Our experiments demonstrate that Adaptive cIoU better evaluates model performance with small ground truth sizes, where cIoU fails to do so effectively despite precise predictions. This underscores the importance of determining an appropriate threshold.

Interactive IoU: Interactivity is a core criterion in sound source localization, but it has not been properly evaluated in previous studies. To explicitly measure the interactivity of sound source localization methods, we propose a new evaluation metric named *Interactive IoU (IIoU)*. IIoU evaluates the model's capability to localize all sound sources given

TABLE 2
Quantitative results on the VGG-SS and SoundNet-Flickr test sets. † is the result of the model released on the official project page[3]. "First place" and "second place" results are indicated with bold and underline, respectively.

Train → Test		VGG-Sound → VGG-SS				VGG-Sound → Flickr-SoundNet				Flickr-SoundNet → Flickr-SoundNet			
Method	**Pre. Vis.**	**cIoU**	**w/ Adap.**	**AUC**	**w/ Adap.**	**cIoU**	**w/ Adap.**	**AUC**	**w/ Adap.**	**cIoU**	**w/ Adap.**	**AUC**	**w/ Adap.**
Attention [50]$_{CVPR18}$	✓	18.50	-	30.20	-	66.00	-	55.80	-	66.00	-	55.80	-
CoarseToFine [45]$_{ECCV20}$	✓	29.10	-	34.80	-	-	-	-	-	-	-	-	-
LCBM [53]$_{WACV22}$	✓	32.20	-	36.60	-	-	-	-	-	-	-	-	-
LVS [7]†$_{CVPR21}$	✗	30.30	40.47	36.40	41.75	68.40	73.60	56.40	60.58	71.20	77.20	58.06	62.72
LVS [7]$_{CVPR21}$	✗	34.40	-	38.20	-	71.90	-	58.20	-	69.90	-	57.30	-
HardPos [52]$_{ICASSP22}$	✗	34.60	-	38.00	-	76.80	-	59.20	-	75.20	-	59.70	-
SSPL (w/o PCM) [55]$_{CVPR22}$	✓	27.00	-	34.80	-	73.90	-	60.20	-	69.90	-	58.00	-
SSPL (w/ PCM) [55]$_{CVPR22}$	✓	33.90	-	38.00	-	76.70	-	60.50	-	75.90	-	61.00	-
EZ-VSL (w/o OGL) [37]$_{ECCV22}$	✓	35.96	43.52	38.20	42.41	78.31	80.40	61.74	64.48	72.80	78.00	58.82	63.82
SSL-TIE [35]$_{ACM\ MM22}$	✗	38.63	51.92	39.65	48.06	79.50	84.80	61.20	65.64	81.50	88.80	61.10	65.38
SLAVC (w/o OGL) [36]$_{NeurIPS22}$	✓	37.79	49.41	39.40	45.79	83.60	85.20	-	66.70	-	-	-	-
MarginNCE (w/o OGL) [44]$_{ICASSP23}$	✓	38.25	50.76	39.06	46.39	83.94	88.80	63.20	68.92	84.74	86.00	63.08	68.32
FNAC (w/o OGL) [56]$_{CVPR23}$	✓	39.50	47.00	39.66	43.30	**84.73**	**89.20**	63.76	**69.78**	78.71	84.40	59.33	64.16
Ours													
↳ NN w/ Sup. Enc.	✗	39.94	54.20	40.02	48.18	79.60	86.80	63.44	69.02	85.20	**90.80**	62.20	67.40
↳ NN w/ Self-Sup. Enc.	✗	39.16	53.71	39.70	47.82	79.20	86.00	63.02	68.08	84.80	89.20	**63.84**	**69.22**
↳ NN w/ Sup. Enc.	✓	**41.42**	**57.25**	**40.76**	**49.32**	83.20	88.00	**64.00**	69.36	**86.00**	90.40	63.50	68.40
with OGL:													
LVS (w/ OGL) [7]†$_{CVPR21}$	✗	40.92	57.92	40.69	49.65	79.20	88.40	62.50	68.80	84.40	89.60	63.54	69.48
EZ-VSL (w/ OGL) [37]$_{ECCV22}$	✓	38.85	57.77	39.54	49.00	83.94	88.80	63.60	68.90	83.13	90.80	63.06	69.14
SSL-TIE [35]$_{ACM\ MM22}$	✗	42.47	60.45	41.46	51.29	81.20	90.40	63.88	70.02	81.60	91.20	63.86	69.84
SLAVC (w/ OGL) [36]$_{NeurIPS22}$	✓	39.80	59.08	-	49.73	**86.00**	90.00	-	69.28	-	-	-	-
MarginNCE (w/ OGL) [44]$_{ICASSP23}$	✓	39.78	59.29	40.01	50.23	85.14	91.60	64.55	70.78	**85.54**	91.60	64.27	70.66
FNAC (w/ OGL) [56]$_{CVPR23}$	✓	41.85	58.78	40.80	49.66	85.14	**92.40**	64.30	70.54	83.93	90.80	63.06	68.84
Ours (w/ OGL)													
↳ NN w/ Sup. Enc.	✗	42.53	60.45	41.34	51.04	82.40	91.20	64.60	70.74	84.00	92.80	64.18	70.32
↳ NN w/ Self-Sup. Enc.	✗	42.49	60.11	41.37	51.11	82.80	90.80	64.48	70.70	85.20	92.80	**64.80**	70.82
↳ NN w/ Sup. Enc.	✓	**42.96**	**61.63**	**41.57**	**51.66**	84.40	91.60	**65.14**	**71.70**	84.80	**93.60**	64.70	**70.98**
with Optical Flow:													
HearTheFlow [18]$_{WACV23}$	✓	39.40	54.56	40.00	48.01	84.80	-	64.00	-	86.50	-	63.90	-
HearTheFlow (w/ OGL) [18]$_{WACV23}$	✓	40.24	58.07	40.23	49.28	84.80	-	64.00	-	86.50	-	63.90	-

multiple audios and an image pair. In essence, we utilize existing cIoU or Adaptive cIoU metrics to measure the accuracy of each sound source within an image. A sample is considered successful by IIoU if the model accurately localizes all sound sources. Conversely, if any sound source is incorrectly localized, the entire sample is marked as a failure, regardless of the other sources' accuracy.

5 Experiments

5.1 Experiment Setup

In this section, we discuss all the datasets we use for training and testing.

Training datasets: Our method is trained using the VGGSound-144K [8] and Flickr-SoundNet-144K [50], [51]. VGGSound is an audio-visual dataset containing around ∼200K videos. Flickr-SoundNet-144K set is the subset of Flickr-SoundNet [3].

Testing datasets: After training, we test the sound source localization performance with the datasets below.

- **VGG-SS and Flickr-SoundNet-Test.** VGG-SS [7] and Flickr-SoundNet-Test [50] are the de facto benchmarks for the main experiments. These evaluation sets have bounding box annotations of sound sources for ∼5K and 250 samples, respectively.
- **AVSBench.** The AVSBench dataset [68] is introduced to tackle the Audio-Visual Segmentation (AVS) problem, offering pixel-level annotations of sounding objects through segmentation masks. AVSBench includes two primary subsets: the Single-source subset (S4), which includes videos with only one sound-emitting object at a time, and the Multi-source subset (MS3), which features videos where multiple objects can produce sound simultaneously. The dataset's statistics are detailed in Table 1.
- **VPO Benchmark.** A concurrent work [63] introduces a new dataset called the Visual Post-production (VPO) benchmark for the audio-visual segmentation task. The VPO benchmark is created using a combination of images and segmentation masks from the COCO dataset and audio files from VGGSound. The process involves randomly matching COCO segmentation masks with related audio files based on instance labels. The VPO benchmark comprises distinct settings: Single-Source (VPO-SS), which includes samples containing one sounding object, with 890 samples for testing, and Multi-Source (VPO-MS), which includes samples that can contain up to five sounding objects from different classes, with 1437 samples for testing. Although this dataset provides segmentation masks as annotations, we obtain bounding boxes that cover these maps to make this dataset usable for standard sound source localization tasks.
- **DenseAV ADE20K.** A recent work [22] introduced a new dataset for sound-prompted image segmentation by pairing images from a subset of the ADE20K dataset [66] with audio samples from VGGSound. The dataset includes 106 images with 20 ADE20K classes.

5.2 Quantitative Results on Standard and New Benchmarks

Comparison with strong baselines on VGG-SS and Flickr-SoundNet. In this section, we conduct a comparative analysis of our sound source localization method against existing approaches. We carry out our evaluations in two settings, following previous approaches. Firstly, we train our model

[3]We omit the scores of the methods that do not release their pre-trained model weights. SLAVC [36] does not provide AUC scores and Flickr-SoundNet numbers.

on VGGSound-144K and evaluate it on VGG-SS and Flickr-SoundNet test sets. Secondly, we train our model on Flickr-SoundNet-144K and evaluate it on the Flickr-SoundNet test set. It is important to note that all the compared models are trained using the same amount of data. We present our results in Table 2.

We compare our method with various settings against prior approaches on three sound source localization scenarios. The proposed models achieve higher overall performance compared to previous methods, regardless of how the NN search module is trained or whether a pre-trained vision encoder is used. We demonstrate the performance of our model with pre-trained encoders learned through supervised learning (NN w/ Sup. Pre. Enc.) and with models pre-trained through self-supervised learning (NN w/ Self-Sup. Pre. Enc.) *in the NN search module*. The results indicate that using either self-supervised or supervised pre-trained encoders in NN search outperforms competing methods. This shows that our model can utilize any type of pre-trained encoder feature for nearest neighbor search. It is important to note that these pre-trained encoders are not used in the backbone networks of the sound source localization module, but only in the NN search module, as illustrated in Figure 2.

We also compare the performance of our method with and without using a pre-trained vision encoder as the backbone. Unlike most previous works, our method achieves competitive performance without supervisedly trained vision encoders. Using supervised pre-trained models in the backbone violates the definition of fully self-supervised learning, which has been an unfavorable practice in the recent sound source localization field. Therefore, we demonstrate that our method "NN w/ Self-Sup Pre. Enc. without Pre. Vision" operates in a fully self-supervised setting by not exploiting any supervised pre-trained encoders. The results show that our method performs favorably against prior methods even when trained from scratch. However, our method can further improve its performance when fine-tuned from a pre-trained vision encoder.

We also discuss the methods employed by previous studies, such as SSPL [55] which utilizes a sub-module called PCM to reduce the impact of background noise, HTF [18] which utilizes Optical Flow, and EZ-VSL [37] which refines its initial audio-visual localization outcomes through object guidance obtained from an ImageNet pre-trained visual encoder. Our model, on the other hand, and any of its variations do not require any task-specific modules or operations to achieve the state-of-the-art results. This suggests that using additional semantic and multi-view correspondence, as well as feature alignment, provides more varied and robust supervision for better aligned audio and visual features, as opposed to using task-specific approaches.

The quantitative results presented in Table 2 also showcase the performance of previous methods that utilize *object guided refinement (OGL)* to evaluate their final sound source localizations. Our model outperforms or gives comparable results to all previous methods that employ object guidance. Additionally, we acknowledge that the inclusion of OGL results in modest improvements for prior methods, while our method shows less performance improvement. This can be explained by the fact that our model already accurately localizes the sounding objects, thus adding OGL has less impact. Unlike prior methods, we do not use OGL in our architecture for the remainder of this paper, unless directly comparing with OGL-based methods. We believe that using OGL contradicts the motivation of audio-visual sound source localization and hinders to understand true audio-visual alignment capability.

TABLE 3
Sound source localization results. All models are trained on the VGGSound-144K dataset. Results without object guided refinement (OGL) are reported.

	Method	Pre. Vis.	cIoU	w/ Adap.	AUC	w/ Adap.
IS4	LVS [7]$_{CVPR21}$	✗	33.4	39.4	39.0	41.1
	EZ-VSL [37]$_{ECCV22}$	✓	34.2	42.1	39.6	42.7
	SSL-TIE [35]$_{ACM\ MM22}$	✗	38.5	49.3	41.7	46.7
	SLAVC [36]$_{NeurIPS22}$	✓	36.9	45.0	40.2	42.7
	MarginNCE [44]$_{ICASSP23}$	✓	40.6	52.6	42.5	47.7
	FNAC [56]$_{CVPR23}$	✓	39.2	49.5	42.0	46.1
	Ours					
	↳ NN w/ Sup. Enc.	✗	45.1	59.4	43.9	50.9
	↳ NN w/ Self-Sup. Enc.	✗	43.3	56.7	43.0	49.6
	↳ NN w/ Sup. Enc.	✓	**45.7**	**63.1**	**44.1**	**52.4**
VPO-SS	LVS [7]$_{CVPR21}$	✗	28.5	31.8	29.1	32.7
	EZ-VSL [37]$_{ECCV22}$	✓	26.6	30.3	29.0	32.9
	SSL-TIE [35]$_{ACM\ MM22}$	✗	31.7	**39.1**	**30.6**	36.7
	SLAVC [36]$_{NeurIPS22}$	✓	29.1	34.3	30.0	34.2
	MarginNCE [44]$_{ICASSP23}$	✓	**32.1**	35.6	30.3	35.1
	FNAC [56]$_{CVPR23}$	✓	31.5	36.3	30.8	34.8
	Ours					
	↳ NN w/ Sup. Enc.	✗	31.1	38.4	30.4	36.4
	↳ NN w/ Self-Sup. Enc.	✗	29.2	38.4	30.1	**36.8**
	↳ NN w/ Sup. Enc.	✓	30.4	38.7	30.4	36.2
VPO-MS	LVS [7]$_{CVPR21}$	✗	25.0	28.9	27.8	30.9
	EZ-VSL [37]$_{ECCV22}$	✓	25.4	31.0	28.7	32.6
	SSL-TIE [35]$_{ACM\ MM22}$	✗	27.5	35.4	29.4	35.1
	SLAVC [36]$_{NeurIPS22}$	✓	27.1	33.9	29.2	34.0
	MarginNCE [44]$_{ICASSP23}$	✓	28.7	33.5	29.5	34.1
	FNAC [56]$_{CVPR23}$	✓	28.4	34.9	29.8	34.2
	Ours					
	↳ NN w/ Sup. Enc.	✗	28.7	**37.5**	**30.1**	**36.0**
	↳ NN w/ Self-Sup. Enc.	✗	27.4	36.8	29.5	35.5
	↳ NN w/ Sup. Enc.	✓	**29.0**	36.4	29.7	35.2
AVS-Bench S4	LVS [7]$_{CVPR21}$	✗	42.0	51.2	41.0	47.2
	EZ-VSL [37]$_{ECCV22}$	✓	44.9	52.4	41.9	47.5
	SSL-TIE [35]$_{ACM\ MM22}$	✗	47.4	60.8	43.2	53.3
	SLAVC [36]$_{NeurIPS22}$	✓	46.8	58.2	43.2	50.7
	MarginNCE [44]$_{ICASSP23}$	✓	47.7	59.0	43.7	51.8
	FNAC [56]$_{CVPR23}$	✓	48.4	58.5	43.8	50.7
	Ours					
	↳ NN w/ Sup. Enc.	✗	51.7	**68.2**	**45.0**	**56.2**
	↳ NN w/ Self-Sup. Enc.	✗	50.5	66.4	44.3	55.0
	↳ NN w/ Sup. Enc.	✓	**52.1**	67.4	45.0	55.9
ADE20K	LVS [7]$_{CVPR21}$	✗	33.0	36.8	35.0	39.2
	EZ-VSL [37]$_{ECCV22}$	✓	35.8	45.3	36.4	42.5
	SSL-TIE [35]$_{ACM\ MM22}$	✗	38.7	48.1	37.9	44.9
	SLAVC [36]$_{NeurIPS22}$	✓	38.7	45.3	38.5	45.7
	MarginNCE [44]$_{ICASSP23}$	✓	34.9	44.3	37.4	44.8
	FNAC [56]$_{CVPR23}$	✓	38.7	42.5	37.8	43.5
	Ours					
	↳ NN w/ Sup. Enc.	✗	**44.3**	**55.7**	**39.0**	46.9
	↳ NN w/ Self-Sup. Enc.	✗	34.9	50.0	37.5	47.2
	↳ NN w/ Sup. Enc.	✓	40.6	51.9	38.5	**47.4**

Finally, in comparison to HearTheFlow, which utilizes an additional Optical Flow modality, our method outperforms it on the VGGSS test set, and achieves slightly lower performance on the Flickr-SoundNet test set without utilizing any additional modalities, but instead relying on better audio-visual correspondence and alignment.

Comparison on IS4. For the IS4 test set, we provide results in Table 3. Since IS4 contains multiple objects in one image and each image is paired with multiple unique audio clips, we evaluate each unique pair independently. IS4 features various backgrounds, unusual spatial locations of the objects, and different appearances, such as realistic, graphic, or cartoon. Our model achieves state-of-the-art results with a significant margin across every evaluation metric. This indicates that our model not only localizes objects more accurately but is also more robust to large domain gaps due

TABLE 4
Quantitative results on the Extended VGG-SS and Extended Flickr-SoundNet sets. All models are trained with 144K samples from VGG-Sound. Some of the results of the prior approaches are obtained from [36] and denoted with †. Results without object guided refinement (OGL) are reported.

Method	Pre. Vis.	Extended Flickr AP	Extended Flickr max-F1	Extended Flickr LocAcc	Extended VGG-SS AP	Extended VGG-SS max-F1	Extended VGG-SS LocAcc
†CoarseToFine [45]$_{ECCV20}$	✓	0.00	38.20	47.20	0.00	19.80	21.93
†LVS [7]$_{CVPR21}$	✗	9.80	17.90	19.60	5.15	9.90	10.43
†Attention10k [50]$_{CVPR18}$	✓	15.98	24.00	34.16	6.70	13.10	14.04
†DMC [26]$_{CVPR19}$	✓	25.56	41.80	52.80	11.53	20.30	22.63
†DSOL [27]$_{NeurIPS20}$	✓	38.32	49.40	72.91	16.84	25.60	26.87
†OGL [37]$_{ECCV22}$	-	40.20	55.70	77.20	18.73	30.90	36.58
†EZ-VSL [37]$_{ECCV22}$	✓	46.30	54.60	66.40	24.55	30.90	31.58
†SLAVC [36]$_{NeurIPS22}$	✓	51.63	59.10	83.60	32.95	40.00	37.79
MarginNCE [44]$_{ICASSP23}$	✓	57.99	61.80	83.94	30.58	36.80	38.25
FNAC [56]$_{CVPR23}$	✓	50.40	62.30	**84.73**	23.48	33.70	39.50
Ours							
↳ NN w/ Sup. Enc.	✗	64.43	66.90	79.60	**34.73**	40.70	39.94
↳ NN w/ Self-Sup. Enc.	✗	62.67	66.10	79.20	33.09	40.00	39.20
↳ NN w/ Sup. Enc.	✓	**70.09**	**69.80**	83.20	36.81	**42.50**	**41.42**

to its strong cross-modal alignment capability.

Comparison on VPO Benchmark. As aforementioned, we utilize the VPO-SS and VPO-MS datasets by obtaining bounding boxes that cover the segmentation maps provided in these datasets for the standard sound source localization task. Similar to IS4, VPO-MS also contains multiple objects in one image and each image is paired with multiple unique audio clips. We apply the same evaluation process as in previous section. Table 3 shows that our model achieves favorable performance compared to existing methods.

Comparison on AVS-Bench S4. Similar to VPO Benchmark, we obtain bounding boxes from this segmentation dataset and evaluate all the models. Results are in Table 3. Our method outperforms the baseline methods across all experimental settings and evaluation metrics. Considering these results, along with previous comparisons on other datasets, our method consistently delivers better performance.

Comparison on ADE20K. Similar to the VPO Benchmark and AVS-Bench S4, we extract bounding boxes from segmentation masks and evaluate all models accordingly. As shown in the results, our method outperforms all others by a large margin (*+5.60 cIoU and +7.60 cIoU Adap.*). This suggests that establishing strong cross-modal alignment is a key factor in accurately localizing sound sources.

Extended Flickr and VGG-SS datasets. The prior study [36] points out that the current sound source localization benchmarks overlook false positive detection. It is because the evaluation samples always contain at least a sounding object in a scene; thus, they cannot capture false positive outputs, *e.g.*, silent objects or off-screen sounds. To analyze false positive detection, Mo and Morgado [36] extended the benchmarks with non-audible, non-visible, and mismatched audio-visual samples. The expectation is that a sound source localization model should not localize any objects when audio-visual semantics do not match. The experiment with the extended datasets in Table 4 shows that our method performs favorably against state-of-the-art competitors. Our method performs better than the competing methods in false positive detection measured by **AP** and **max-F1**, while other methods [36], [44], [56] achieve better localization performance on Extended Flickr-SoundNet. Since false positive detection requires cross-modal interaction, our method shows strong performance in this task.

TABLE 5
Summary of retrieval recall scores for all models. All of the models are trained on VGGSound 144K data and retrieval is performed on entire VGG-SS dataset, containing ∼5K samples.

Model	Pre. Vis.	A → I R@1	A → I R@5	A → I R@10	I → A R@1	I → A R@5	I → A R@10
LVS [7]$_{CVPR21}$	✗	3.87	12.35	20.73	4.90	14.29	21.37
EZ-VSL [37]$_{ECCV22}$	✓	2.62	7.91	12.59	4.12	14.07	22.47
SSL-TIE [35]$_{ACM\ MM22}$	✗	10.29	30.68	43.76	12.76	29.58	39.72
SLAVC [36]$_{NeurIPS22}$	✓	4.77	13.08	19.10	6.12	21.16	32.12
MarginNCE [44]$_{ICASSP23}$	✓	4.49	16.43	24.75	6.64	21.80	33.32
FNAC [56]$_{CVPR23}$	✓	1.33	7.02	10.01	2.02	8.34	14.84
Ours Backbone							
↳ NN w/ Sup. Enc.	✗	16.47	36.99	49.00	20.09	42.38	53.66
↳ NN w/ Self-Sup. Enc.	✗	14.31	37.81	49.17	18.00	38.39	49.02
↳ NN w/ Sup. Enc.	✓	19.16	40.11	51.66	23.91	46.83	59.05
Ours Projected							
↳ NN w/ Sup. Enc.	✗	22.14	46.66	57.37	25.50	48.87	58.95
↳ NN w/ Self-Sup. Enc.	✗	20.06	43.93	54.91	23.93	46.40	57.27
↳ NN w/ Sup. Enc.	✓	**22.72**	**48.40**	**58.52**	**29.43**	**52.85**	**62.90**

5.3 Qualitative Results

In this section, we visualize and compare our sound source localization results with the recent prior works on six benchmarks. The visualized samples in Figure 6 show that localized regions of the proposed method are more compact and accurately aligns with the sounding objects than the other methods. For instance, small size keyboard is localized accurately compared to the recent methods in the first column of the fifth row.

5.4 Cross-Modal Retrieval

As we discussed earlier, audio-visual correspondence is the most essential aspect for genuine sound source localization. Thus, any sound source localization model should ensure cross-modal semantic understanding. In most previous work, this aspect has been overlooked in evaluating benchmarks by solely focusing on sound source localization performance. Considering the visual biases in existing sound source localization benchmarks, additional tasks are necessary to inspect the models. To explore this, we propose a cross-modal retrieval task as an auxiliary evaluation task to assess cross-modal interactivity of sound source localization methods. The expectation is that models which perform well at the sound source localization task should also show good performance in this task, as these models learn how objects look and sound. We evaluate sound source localization models on the VGG-SS dataset for cross-modal retrieval. Our proposed model outputs representations from each modality in two different ways (See Figure 2). One way is the features from the backbone encoders directly, and the other is the projected features that we apply a projection layer to the backbone features (explained in Section 3.2). We compute the retrieval scores in two different setups according to the feature type that is used. We report *Recall @1, @5 and @10* for cross-modal query-retrieval in Table 5.

Results with backbone features. Given a query modality feature from the backbone features, we compute its distance to the other modality backbone features in the retrieval pool. Note that the backbone features are used for all the competing methods as well. As shown in Table 5, our

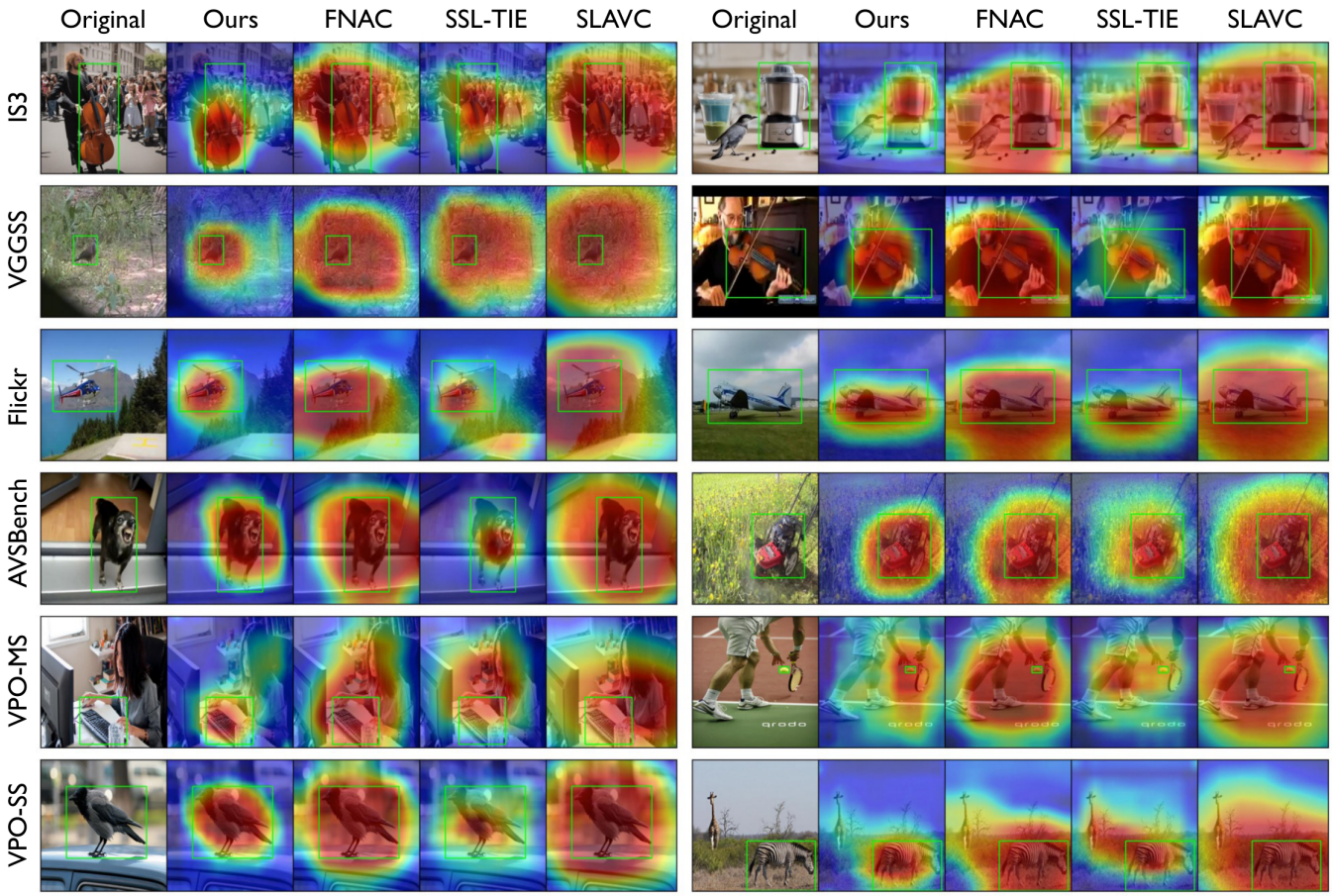


Fig. 6. **Qualitative sound source localization results.**

method clearly outperforms other state-of-the-art methods. One interesting observation is that FNAC and MarginNCE perform notably worse in cross-modal retrieval tasks, despite their high and comparable performance to ours in standard benchmarks for sound source localization (See Section 5.2 and Table 2). This finding indicates that high performance in sound source localization according to existing standard benchmarks does not necessarily translate to better audio-visual semantic understanding. Thus, it is essential to additionally evaluate sound source localization methods on cross-modal understanding tasks. Another observation is the significant performance gap between our method and the strongest competitor, SSL-TIE [35], which is more prominent in cross-modal retrieval than in sound source localization. This discrepancy is due to the robust cross-modal feature alignment achieved by our method, which is overlooked in sound source localization benchmarks.

Results with projected features. As illustrated in Figure 2, the backbone feature is directly used for localization, while the projected feature is used for semantic alignment. It is intuitive to consider that the projected feature contains more semantic information, making it more suitable for cross-modal retrieval tasks. To verify this, we explore retrieval performance using projected features. The results presented in Table 5 indicate that using projected features substantially improves retrieval performance compared to backbone features across various settings. This improvement also highlights a clear distinction between existing sound source localization methods and our approach.

Results with using pre-trained vision encoder. We also provide the retrieval results of our model, which is trained with pre-trained vision encoders in Table 5. We observe that retrieval performance is further improved in both the backbone and projected features settings. Notably, as expected, Image-to-Audio retrieval shows a more significant performance improvement. Despite this, our method, when trained in a fully self-supervised manner without the pre-trained vision encoder, still outperforms competing methods in cross-modal retrieval.

Compositional image retrieval. Given an image and a semantic target condition from different modalities, Compositional Image Retrieval retrieves the target images from the database. This task requires understanding the semantic coupling between the given image content and the condition from the other modality. Compositional Image Retrieval has recently attracted considerable attention [5], [6], [20], [29], [49], with the main trend being retrieval with textual conditions. Similarly, here, we aim to demonstrate the compositional ability of our model with audio conditions. We use multimodal embedding space arithmetic for compositional image retrieval. We start by extracting a visual feature ($\mathbf{v}$) and an audio feature ($\mathbf{a}$) from an image and audio, respectively. Then, we interpolate between these two features in the latent space to obtain a multimodal composed feature, $\mathbf{z}^{\mathbf{new}} = \lambda\mathbf{v} + (1-\lambda)\mathbf{a}$, where the interpolation coefficient (λ)

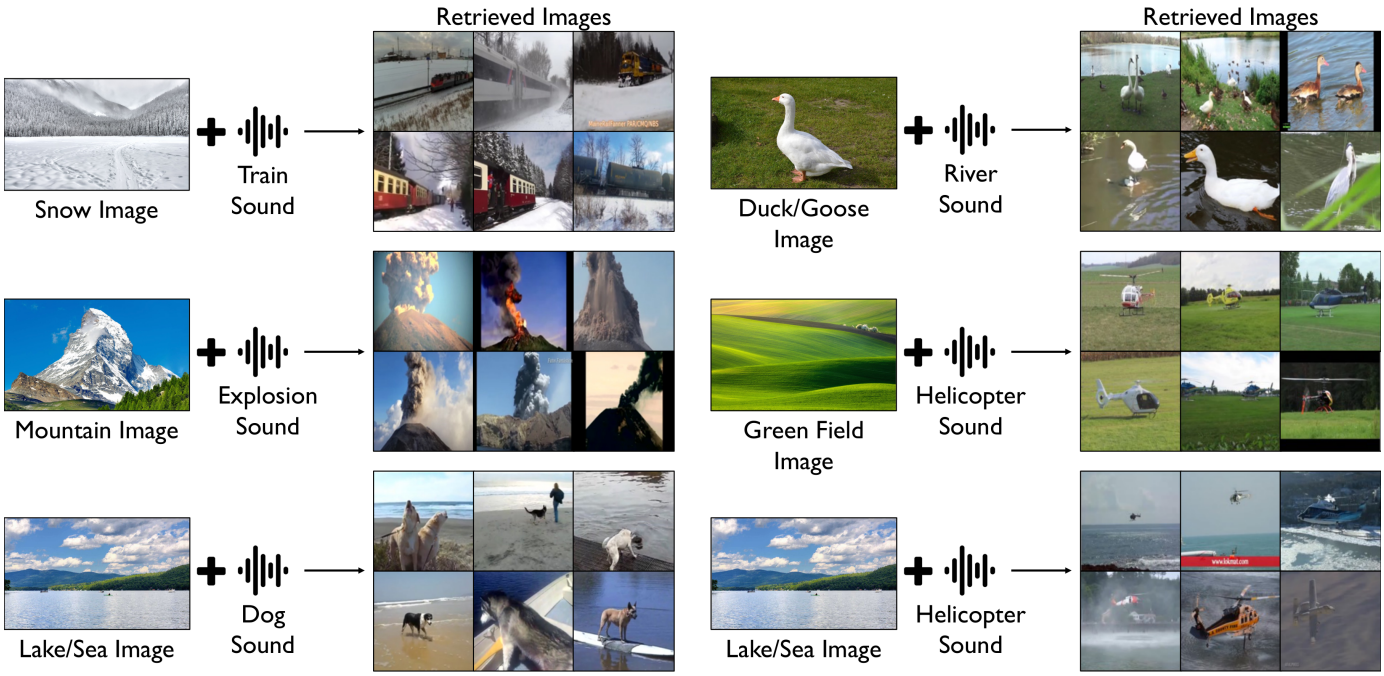


Fig. 7. **Compositional image retrieval.** Our method retrieves the desired images based on the given image and audio. We use simple multimodal embedding space arithmetic for compositional image retrieval. Due to the strong cross-modal alignment, our method achieves meaningful results in compositional image retrieval with straightforward vector arithmetic.

TABLE 6
Cross-modal alignment analysis.

Method	Pre. Vision	Magnitude ↓	Alignment ↑
EZ-VSL [37]$_{\text{ECCV22}}$	✓	1.19 ± 0.05	0.1465
SSL-TIE [35]$_{\text{ACM MM22}}$	✗	1.02 ± 0.08	0.2134
SLAVC [36]$_{\text{NeurIPS22}}$	✓	1.15 ± 0.15	0.2214
MarginNCE [44]$_{\text{ICASSP23}}$	✓	1.08 ± 0.08	0.2423
FNAC [56]$_{\text{CVPR23}}$	✓	1.26 ± 0.03	0.0293
Ours	✗	**0.94 ± 0.17**	**0.5419**

varies across different examples. This new feature is used to retrieve the image. Our strong cross-modal alignment and shared embedding space allow us to obtain meaningful results in compositional image retrieval with simple vector arithmetic. All the qualitative results are shown in Figure 7.

5.5 Cross-Modal Alignment Analysis

In the previous section, we utilize cross-modal retrieval as an auxiliary task to measure the cross-modal alignment of our model together with all the existing methods. In this section, we further analyze the cross-modal alignment with the intuition that the embeddings of a matched image-audio pair should be close. For this analysis, we use the metrics of *alignment* and *magnitude* from [21] and [65], respectively. While *magnitude* measures the gap (distance) between the modalities, *alignment* measures the closeness of the representations of the positive pairs using cosine similarity.

The results of this analysis are shown in Table 6. All the results are obtained from the training set samples. Consistent with the cross-modal retrieval, our model demonstrates superior cross-modal alignment compared to the other existing methods in both evaluation metric.

TABLE 7
Interactive sound source localization results. All models are trained on VGGSound-144K dataset.Object guided refinement (OGL) is not used.

	Method	Pre. Vis.	IIoU	w/ Adap.	IAUC	w/ Adap.
IS4	LVS [7]$_{\text{CVPR21}}$	✗	6.5	11.2	26.0	25.3
	EZ-VSL [37]$_{\text{ECCV22}}$	✓	7.4	13.0	26.4	26.6
	SSL-TIE [35]$_{\text{ACM MM22}}$	✗	9.4	19.0	28.4	31.5
	SLAVC [36]$_{\text{NeurIPS22}}$	✓	7.5	14.5	26.3	25.5
	MarginNCE [44]$_{\text{ICASSP23}}$	✓	11.5	23.7	29.4	32.5
	FNAC [56]$_{\text{CVPR23}}$	✓	11.5	22.4	28.9	31.0
	Ours					
	↳ NN w/ Sup. Enc.	✗	14.8	31.4	31.1	37.4
	↳ NN w/ Self-Sup. Enc.	✗	13.2	27.3	30.2	35.5
	↳ NN w/ Sup. Enc.	✓	**15.8**	**37.6**	**31.4**	**39.5**
VPO-MS	LVS [7]$_{\text{CVPR21}}$	✗	21.2	24.2	24.8	27.0
	EZ-VSL [37]$_{\text{ECCV22}}$	✓	20.9	25.4	25.4	28.3
	SSL-TIE [35]$_{\text{ACM MM22}}$	✗	23.8	30.6	26.4	31.0
	SLAVC [36]$_{\text{NeurIPS22}}$	✓	22.4	28.4	25.8	29.3
	MarginNCE [44]$_{\text{ICASSP23}}$	✓	24.9	28.1	26.4	29.9
	FNAC [56]$_{\text{CVPR23}}$	✓	24.9	29.7	26.8	30.1
	Ours					
	↳ NN w/ Sup. Enc.	✗	24.4	**31.9**	**26.8**	**31.7**
	↳ NN w/ Self-Sup. Enc.	✗	23.5	31.6	26.3	31.3
	↳ NN w/ Sup. Enc.	✓	**24.9**	30.5	26.6	30.9

5.6 Interactive Sound Source Localization

Up to this point, we have discussed the importance of cross-modal semantic understanding. Consequently, we propose two auxiliary tasks for sound source localization methods. The first task is cross-modal retrieval, as outlined in the previous section. The second task is interactive sound source localization. Effective sound source localization methods should be capable of identifying objects correlated with the sound. In other words, the localized area in the image should shift to a different region when the same image is paired with another sound present in the scene (See Figure 1). To evaluate the effectiveness of interactive sound source localization methods, we use the IS4 and VPO-MS datasets. All models are trained on the VGGSound-

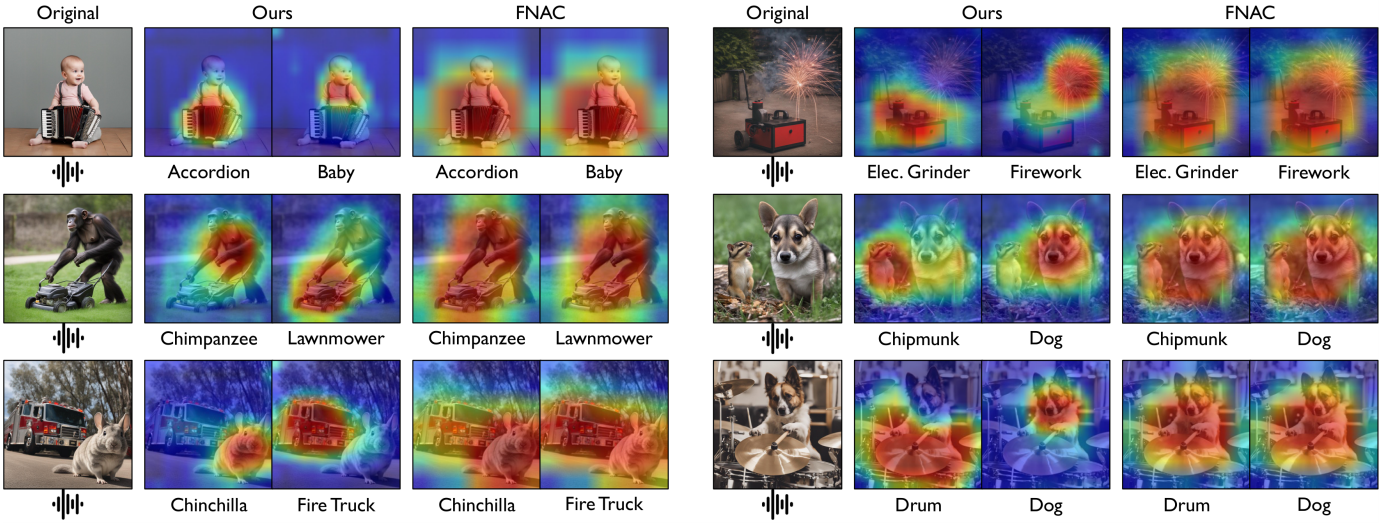


Fig. 8. **Qualitative results for interactive sound source localization on IS4 dataset.** Our model correctly follows the cross-modal interaction for various given sounds, while competing methods always focus on the visually dominant object or area in a scene regardless of the given sound.

TABLE 8
Ablation studies on our proposed method to see the impact of each main component.

	Semantic	Multi-View	Feature Alignment	cIoU	AUC
(A)	✓	✓	✓	**39.94**	**40.02**
(B)	✓	✓	✗	39.10	39.44
(C)	✓	✗	✓	38.75	39.34
(D)	✓	✗	✗	38.24	38.90
(E)	✗	✓	✓	38.30	39.38
(F)	✗	✓	✗	37.72	39.19
(G)	✗	✗	✓	34.93	37.94
(H)	✗	✗	✗	34.22	37.67

144K dataset. We present our results in Table 7 with IIoU metric. Our method consistently outperforms the baselines. The IS4 dataset is specifically designed for the interactive localization task. As the numbers suggest, our method demonstrates a substantial performance gap compared to existing methods across every setting and evaluation metric. Notably, it outperforms by +15.3% in Adaptive IIoU.

Qualitative results. We demonstrate the interactivity of our method across modalities in Figure 8. Genuine sound source localization should be able to identify objects correlated with the sound. We compare our method with the recent state-of-the-art method FNAC [56]. The examples show that our proposed method can localize different objects depending on the context of the sounds, while the competing method cannot, as it always attends to a visually dominant object in a scene.

5.7 Ablation Results

We conduct a series of experiments in order to verify our design choices and make further analysis. To save computational time and resources, we primarily perform ablation studies by training our model on VGGSound-144K with NN Search w/ Supervised Pre. Encoders setup and evaluating it on VGG-SS.

Impact of semantic and multi-view invariance. To understand the impact of each type of invariance (consistency), we analyze the performance of our model trained with different types of invariance, as shown in Table 8. As the comparisons of (C *vs.* E) and (D *vs.* F) reveal, using semantically similar samples (semantic invariance) yields better performance (+0.45% and +0.5% on cIoU, respectively) compared to augmented multi-view invariance. Furthermore, as the comparisons of (A *vs.* C) and (A *vs.* E) depict, the combination of these two types of invariance complements each other and further enhances the model's performance. Integrating these two different types of consistency elements provides additional supervision, invariance, and alignment, leading to a more robust representation and improves sound source localization performance.

TABLE 9
Impact of different modalities.

Type of Positives	Modality	cIoU	AUC
Semantic	Vision (✓), Audio (✓)	38.75	39.34
	Vision (✗), Audio (✓)	37.42	38.73
	Vision (✓), Audio (✗)	37.01	38.44
Hand-crafted	Vision (✓), Audio (✓)	37.72	39.19
	Vision (✗), Audio (✓)	34.05	37.33
	Vision (✓), Audio (✗)	37.57	38.91

Impact of modality on positive samples. Our formulation incorporates additional positive samples from both modalities. To understand the contribution of each modality through these additional positive samples, we trained our model by disabling the additional samples from each modality one at a time in two settings: 1) semantically similar samples, and 2) hand-crafted augmented samples (multi-view). Both settings include a feature alignment loss. In Table 9, we compare the sound source localization accuracy of our model with variations in the source modality of the additional positive samples. The results indicate that each modality contributes to the final performance in both settings. However, we observe that the absence of hand-crafted positive samples from the vision modality significantly impacts the performance, while less so from the audio modality case. Conversely, discarding the semantically similar samples from the audio modality has a greater impact than the vision modality.

Impact of feature alignment. We perform controlled exper-

TABLE 10
Varying k in conceptually similar sample selection.

k in k-NN	10	30	100	500	1000
cIoU	38.80	38.82	39.46	39.90	**39.94**
AUC	39.51	39.67	39.93	40.00	**40.02**

TABLE 11
Comparison of different sampling method baselines. Note that in this experimental setting, we only use semantically similar samples without multi-view samples and feature alignment to see the direct impact of the different methods.

	Sampling Methods	cIoU	AUC
(q)	Ours (Random in top-1000)	**38.24**	**38.90**
(w)	Fixed Same Sample (k=1)	35.80	38.16
(e)	Identical (Anchor itself)	34.25	37.63
(r)	No Semantically Similar Sample	34.22	37.67

iments to verify the effect of the feature alignment strategy, and the results are presented in Table 8. Comparing the performance of the proposed model with and without feature alignment, (A *vs.* B), highlights the importance of this strategy to boost the performance. Further, examining the results of experiments (C *vs.* D) and (E *vs.* F) reveals that feature alignment provides additional gains irrespective of the consistency types. These findings indicate that global feature-based alignment helps the optimization of audio-visual correspondence.

Impact of k in semantically similar sample selection. Selecting an appropriate k value for sampling nearest neighbors is crucial. If this value is set too high, it may result in noisy samples that could disrupt the learning phase. Conversely, if the value is set too low, only very similar samples to the anchor will be provided, limiting semantic invariance. Nevertheless, compared to Table 8 (E), we observe performance gains throughout the range of k values used in the ablation study, as shown in Table 10. The results indicate that an optimal choice is k=1000. However, setting k to smaller values still provides benefits over not using semantically similar samples. An optimal k value balances semantic similarity and sufficient diversity.

Impact of sampling strategy. Our proposed approach selects a random sample from a set of k samples to obtain a semantically similar sample. In this ablation, we additionally consider two special cases to analyze the effects of different sampling methods: 1) always selecting the same sample, *i.e.*, k=1, and 2) selecting the query itself (anchor itself) as the semantically similar sample. We conduct this ablation study in the setting of using only semantically similar samples (without multi-view and feature alignment) to observe the direct impact. This setting is identical to (D) in Table 8. As the results in Table 11 demonstrate, always selecting the fixed sample (w) leads to an improvement over the no semantic samples setup (r), but falls behind the proposed approach (q) due to limited diversity and semantic invariance. Additionally, using multiple positive samples that are identical to the anchor (e) has no impact (e *vs.* r) as expected. This indicates that our model's performance improvement is not due to multiple losses but rather due to obtaining semantically similar samples in a diverse but semantically consistent manner.

TABLE 12
Impact of pre-trained encoders from different modalities. All models are trained on the VGGSound 144K dataset. SSL Enc. refers to the backbone encoders used in the Sound Source Localization module. ✓ and ✗ indicate whether the encoders are initialized with pre-trained weights or not.

Ours w/ NN Setup	SSL Enc.	cIoU	AUC
NN w/ Sup. Enc.	Vision (✗), Audio (✗)	39.94	40.02
	Vision (✓), Audio (✗)	41.42	40.76
	Vision (✗), Audio (✓)	40.30	40.23
	Vision (✓), Audio (✓)	**42.06**	**41.08**
NN w/ Self-Sup. Enc.	Vision (✗), Audio (✗)	39.16	39.70
	Vision (✓), Audio (✗)	40.19	40.44
	Vision (✗), Audio (✓)	39.68	40.22
	Vision (✓), Audio (✓)	**40.99**	**40.71**

TABLE 13
Impact of additional intra-modality feature alignment. All models are trained with 144K samples from VGG-Sound and tested on VGG-SS and SoundNet-Flickr.

		VGG-SS				Flickr-SoundNet			
Method	Pre. Vis.	cIoU	w/ Adap.	AUC	w/ Adap.	cIoU	w/ Adap.	AUC	w/ Adap.
Ours	✗	39.94	54.20	40.02	48.18	79.60	86.80	**63.44**	69.02
w/ L_{intra}	✗	**40.45**	**56.50**	**40.44**	**49.29**	**80.80**	**88.00**	63.24	**69.16**

Impact of pre-trained encoders from different modalities. Throughout the paper, we follow the common practice of training models with and without vision encoders initialized with ImageNet-pretrained weights, while the audio encoder is always trained from scratch. Importantly, both encoders are trained using our self-supervised learning. As expected, initializing the vision encoder with pretrained weights leads to improved performance compared to training from scratch. To further investigate the impact of pre-trained backbones in the sound source localization module, we also incorporate a pre-trained audio encoder and evaluate all possible combinations with the pre-trained vision encoder. The results are shown in Table 12. As mentioned in the implementation details, our audio encoder is a ResNet-18 model; therefore, we initialize it with weights from a ResNet-18 audio encoder pre-trained on VGGSound (200K samples), as proposed in [8]. The results show that using pre-trained encoders (initializing the weights) in either modality leads to better performance than training from scratch, with the best results achieved when both encoders are pre-trained.

Impact of additional intra-modality feature alignment. Our method employs cross-modal feature alignment to incorporate global context for enhanced audio-visual semantic alignment. As previously described, our positive set includes multiple samples from the same modality (see Figure 2). In this ablation study, we investigate whether adding intra-modality feature alignment, in addition to cross-modal feature alignment, further improves sound source localization performance. We train our model with intra-modality feature alignment, and the results are shown in Table 13. The results indicate that intra-modality alignment brings additional performance improvements of 0.5% and 0.4% cIoU and 2.3% and 1.2% cIoU Adap. on the VGG-SS and Flickr-SoundNet datasets, respectively, by providing extra regularization in the shared embedding space. We present this setup as an ablation study and do not adopt it as the default proposed method for simplicity.

TABLE 14
Audio-visual segmentation results on AVS Bench S4 and MS3 datasets. All models are trained on the VGGSound 144K dataset. Object guided refinement (OGL) is not used.

	Method	Pre. Vis.	mIoU	w/ Adap.	F-Score	w/ Adap.
AVS-Bench S4	LVS [7]$_{CVPR21}$	✗	27.0	30.5	33.4	42.4
	EZ-VSL [37]$_{ECCV22}$	✓	27.7	30.7	34.1	42.8
	SSL-TIE [35]$_{ACM\ MM22}$	✗	28.9	38.9	35.2	52.5
	SLAVC [36]$_{NeurIPS22}$	✓	28.0	32.8	34.4	45.5
	MarginNCE [44]$_{ICASSP23}$	✗	28.9	35.4	35.3	48.6
	FNAC [56]$_{CVPR23}$	✓	28.8	33.0	35.3	45.6
	Ours					
	↳ NN w/ Sup. Enc.	✗	**30.1**	**40.6**	36.3	**54.3**
	↳ NN w/ Self-Sup. Enc.	✗	29.5	39.5	35.8	53.2
	↳ NN w/ Sup. Enc.	✓	30.1	39.2	**36.3**	53.0
AVS-Bench MS3	LVS [7]$_{CVPR21}$	✗	22.8	26.8	25.1	28.9
	EZ-VSL [37]$_{ECCV22}$	✓	22.6	27.8	25.0	30.9
	SSL-TIE [35]$_{ACM\ MM22}$	✗	23.5	**32.7**	25.9	**37.8**
	SLAVC [36]$_{NeurIPS22}$	✓	22.1	26.1	24.3	28.5
	MarginNCE [44]$_{ICASSP23}$	✓	23.1	30.1	25.5	35.4
	FNAC [56]$_{CVPR23}$	✓	23.2	30.4	25.5	34.2
	Ours					
	↳ NN w/ Sup. Enc.	✗	**23.7**	30.9	26.1	35.1
	↳ NN w/ Self-Sup. Enc.	✗	23.6	31.5	25.9	35.9
	↳ NN w/ Sup. Enc.	✓	23.7	31.4	**26.2**	35.9

TABLE 15
Audio-visual segmentation results. All models are trained on VGGSound-144K dataset. Object guided refinement (OGL) is not used.

	Method	Pre. Vis.	cIoU	w/ Adap.	AUC	w/ Adap.	mIoU	F-Score
IS4	LVS [7]$_{CVPR21}$	✗	6.3	11.1	23.9	24.2	23.8	29.7
	EZ-VSL [37]$_{ECCV22}$	✓	7.2	13.4	24.5	26.4	24.5	30.3
	SSL-TIE [35]$_{ACM\ MM22}$	✗	9.2	20.7	26.0	31.8	26.0	32.1
	SLAVC [36]$_{NeurIPS22}$	✓	7.1	15.1	24.4	26.2	24.3	30.1
	MarginNCE [44]$_{ICASSP23}$	✓	9.2	18.5	26.1	30.8	26.1	31.9
	FNAC [56]$_{CVPR23}$	✓	7.3	14.7	25.3	27.5	25.3	31.1
	Ours							
	↳ NN w/ Sup. Enc.	✗	9.6	25.4	27.0	35.4	27.0	32.9
	↳ NN w/ Self-Sup. Enc.	✗	9.5	24.4	26.7	34.8	26.7	32.5
	↳ NN w/ Sup. Enc.	✓	**10.6**	**28.5**	**27.3**	**36.6**	**27.3**	**33.1**
VPO-SS	LVS [7]$_{CVPR21}$	✗	12.7	14.6	20.8	21.4	20.3	25.5
	EZ-VSL [37]$_{ECCV22}$	✓	9.6	12.2	20.4	22.1	20.0	25.3
	SSL-TIE [35]$_{ACMMM23}$	✗	12.8	20.4	21.4	26.3	21.0	26.4
	SLAVC [36]$_{NeurIPS22}$	✓	12.0	15.3	21.1	22.1	20.6	25.8
	MarginNCE [44]$_{ICASSP23}$	✓	11.3	14.8	21.3	23.4	20.8	26.1
	FNAC [56]$_{CVPR23}$	✓	12.1	15.7	21.5	23.2	21.1	26.3
	Ours							
	↳ NN w/ Sup. Enc.	✗	**13.3**	**20.7**	**21.6**	26.4	**21.2**	**26.5**
	↳ NN w/ Self-Sup. Enc.	✗	12.7	20.2	21.4	**26.5**	21.0	26.3
	↳ NN w/ Sup. Enc.	✓	12.4	17.6	21.5	24.9	21.0	26.3
VPO-MS	LVS [7]$_{CVPR21}$	✗	8.2	10.9	18.3	18.8	17.8	22.7
	EZ-VSL [37]$_{ECCV22}$	✓	9.4	11.8	18.9	20.6	18.5	23.4
	SSL-TIE [35]$_{ACM\ MM22}$	✗	10.8	19.2	19.5	24.3	19.1	24.0
	SLAVC [36]$_{NeurIPS22}$	✓	10.3	14.8	19.2	21.7	18.7	23.6
	MarginNCE [44]$_{ICASSP23}$	✓	10.4	15.0	19.6	22.2	19.2	24.1
	FNAC [56]$_{CVPR23}$	✓	9.4	13.8	19.5	21.3	19.1	24.0
	Ours							
	↳ NN w/ Sup. Enc.	✗	11.4	19.8	**20.1**	**25.2**	**19.7**	**24.6**
	↳ NN w/ Self-Sup. Enc.	✗	10.6	**19.9**	19.6	24.9	19.2	24.1
	↳ NN w/ Sup. Enc.	✓	**11.7**	18.7	19.9	24.0	19.5	24.3
ADE20K	LVS [7]$_{CVPR21}$	✗	5.7	10.4	21.4	19.7	22.1	27.2
	EZ-VSL [37]$_{ECCV22}$	✓	4.7	11.3	21.6	21.2	22.3	27.3
	SSL-TIE [35]$_{ACM\ MM22}$	✗	6.6	11.3	22.4	25.1	23.6	28.7
	SLAVC [36]$_{NeurIPS22}$	✓	6.6	9.4	23.3	24.8	24.2	29.0
	MarginNCE [44]$_{ICASSP23}$	✓	**9.4**	13.2	22.9	23.9	24.1	28.9
	FNAC [56]$_{CVPR23}$	✓	5.7	7.5	22.9	21.8	23.9	28.7
	Ours							
	↳ NN w/ Sup. Enc.	✗	7.5	**18.9**	**24.0**	27.3	**25.0**	**30.0**
	↳ NN w/ Self-Sup. Enc.	✗	5.7	17.0	23.5	**28.5**	24.4	29.6
	↳ NN w/ Sup. Enc.	✓	8.5	14.2	23.5	27.3	24.6	29.5

5.8 Application: Audio-Visual Segmentation

Although the primary focus of this work is not audio-visual segmentation, we can still assess whether our model can precisely localize sound sources from a segmentation perspective. To this end, we conduct additional experiments using audio-visual segmentation datasets in a zero-shot setting, where our models and the competing models are all trained on the unlabeled VGGSound-144K dataset and evaluated directly on the datasets below without any further fine-tuning (zero-shot setting).

AVSBench [67], [68]. We first compare our method with others using the AVSBench benchmark, the most popular audio-visual segmentation benchmark. For a fair comparison, we only utilize some of the self-supervised sound source localization methods mentioned previously. Following the evaluation method and source code of [67], [68], we use mIoU and F-Score as the main metrics. Our results, presented in Table 14, demonstrate that our method generally achieves higher performance in both single (S4) and multiple sound source (MS3) scenarios.

IS4 Dataset. Since the IS4 dataset also provides segmentation masks, we evaluate our model on this dataset from a segmentation perspective as well. Following the evaluation protocol of AVSBench [67], [68], each unique pair is considered independently for evaluation (as described in Section 5.2). In this dataset, we additionally use cIoU and Adaptive cIoU metrics as well. The results are presented in Table 15. Our proposed method shows superior performance in every evaluation metric.

VPO Benchmark. We assess the segmentation performance of our model on the VPO benchmarks. We follow the same evaluation settings and metrics as used with the IS4 dataset. The results are in Table 15. Consistent with all other experiments throughout this paper, our method demonstrates superior performance across all evaluation metrics.

DenseAV ADE20K. As a final analysis, we evaluate our method on the DenseAV sound-prompted image segmentation dataset and compare it with other approaches (Table 15) following the IS4 and VPO protocols. Consistent with previous results, our method also outperforms existing approaches on this dataset.

All of the experiments in this section verify the superiority of our method, even in the audio-visual segmentation task, which requires more accurate localization ability.

6 Conclusion

In this paper, we conduct an in-depth analysis of cross-modal interactions in existing methods, benchmarks, evaluation metrics, and cross-modal understanding tasks, highlighting the limitations of current benchmarks and metrics in evaluating cross-modal interactivity. Our analysis further reveals the shortcomings of existing methods in interactive sound source localization. To address these limitations, we propose a comprehensive new benchmark, evaluation metric, and sound source localization method designed to evaluate and achieve strong cross-modal interactivity. To enforce strong cross-modal interactivity while maintaining localization capability, we propose semantic alignment with multi-views of audio-visual pairs in a simple yet effective manner. We extensively evaluate our method and competing methods on sound source localization, including single sound source, multiple sound source, and cross-dataset scenarios. Furthermore, we benchmark our method and competing methods on cross-modal retrieval, interactive sound source localization and audio-visual segmentation tasks to comprehensively analyze and evaluate cross-modal interactivity and localization performance. The extensive experiments demonstrate the importance of our new benchmark and evaluation metric, validating the effectiveness

of our method across various tasks and settings. We hope this comprehensive study, including the new benchmark, evaluation setting, and our proposed method, will serve as a valuable reference for future studies in sound source localization.

7 Acknowledgment

A. Senocak, H. Ryu, and J.S. Chung were supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00212845, Multimodal Speech Processing for Human-Computer Interaction). T.-H. Oh was partially supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2024-00457882, National AI Research Lab Project) and the KAIST Cross-Generation Collaborative Lab Project. J.Kim and H. Pfister were partially supported by NIH grant R01HD104969.

References

[1] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *IEEE International Conference on Computer Vision*, 2017. 3
[2] Relja Arandjelović and Andrew Zisserman. Objects that sound. In *European Conference on Computer Vision*, 2018. 1
[3] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. *Advances in Neural Information Processing Systems, NeurIPS*, 2016. 3, 7
[4] Mehdi Azabou, Mohammad Gheshlaghi Azar, Ran Liu, Chi-Heng Lin, Erik C Johnson, Kiran Bhaskaran-Nair, Max Dabagia, Bernardo Avila-Pires, Lindsey Kitchell, Keith B Hengen, et al. Mine your own view: Self-supervised learning through across-sample prediction. *arXiv preprint arXiv:2102.10106*, 2021. 3
[5] Yang Bai, Xinxing Xu, Yong Liu, Salman Khan, Fahad Khan, Wangmeng Zuo, Rick Siow Mong Goh, and Chun-Mei Feng. Sentence-level prompts benefit composed image retrieval. *arXiv preprint arXiv:2310.05473*, 2023. 10
[6] Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. In *IEEE International Conference on Computer Vision*, 2023. 10
[7] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 1, 3, 4, 5, 6, 7, 8, 9, 11, 14
[8] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2020. 7, 13
[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 2020. 3, 4, 5
[10] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 3
[11] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 3
[12] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2005. 3
[13] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asia Conference on Computer Vision*, pages 251–263. Springer, 2017. 3
[14] Soo-Whan Chung, Hong Goo Kang, and Joon Son Chung. Seeing voices and hearing voices: learning discriminative embeddings using cross-modal self-supervision. In *INTERSPEECH*, 2020. 3
[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 2
[16] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *IEEE International Conference on Computer Vision*, 2021. 3, 4
[17] Mohamed El Banani, Karan Desai, and Justin Johnson. Learning Visual Representations via Language-Guided Sampling. In *CVPR*, 2022. 3, 4
[18] Dennis Fedorishin, Deen Dayal Mohan, Bhavin Jawade, Srirangaraj Setlur, and Venu Govindaraju. Hear the flow: Optical flow-based self-supervised visual sound source localization. In *IEEE Winter Conf. on Applications of Computer Vision (WACV)*, 2023. 1, 7, 8
[19] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2017. 2
[20] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 10
[21] Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan Rossi, Vishwa Vinay, and Aditya Grover. Cyclip: Cyclic contrastive language-image pretraining. In *Advances in Neural Information Processing Systems, NeurIPS*, 2022. 11
[22] Mark Hamilton, Andrew Zisserman, John R Hershey, and William T Freeman. Separating the" chirp" from the" chat": Self-supervised visual grounding of sound and language. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 6, 7
[23] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. In *Advances in Neural Information Processing Systems, NeurIPS*, 2020. 3, 4
[24] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 3
[25] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *International workshop on similarity-based pattern recognition*, 2015. 3
[26] Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 3, 9
[27] Di Hu, Rui Qian, Minyue Jiang, Xiao Tan, Shilei Wen, Errui Ding, Weiyao Lin, and Dejing Dou. Discriminative sounding objects localization via self-supervised audiovisual matching. *Advances in Neural Information Processing Systems, NeurIPS*, 2020. 1, 9
[28] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023. 2
[29] Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. Vision-by-language for training-free compositional image retrieval. In *International Conference on Learning Representations*, 2024. 10
[30] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *IEEE International Conference on Computer Vision*, 2023. 6
[31] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. *Advances in Neural Information Processing Systems*, 31, 2018. 3
[32] Sizhe Li, Yapeng Tian, and Chenliang Xu. Space-time memory network for sounding object localization in videos. In *British Machine Vision Conference*, 2021. 1, 3
[33] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *International Conference on Learning Representations*, 2022. 3, 4
[34] Yan-Bo Lin, Hung-Yu Tseng, Hsin-Ying Lee, Yen-Yu Lin, and Ming-Hsuan Yang. Unsupervised sound localization via iterative contrastive learning. *arXiv preprint arXiv:2104.00315*, 2021. 1, 3
[35] Jinxiang Liu, Chen Ju, Weidi Xie, and Ya Zhang. Exploiting transformation invariance and equivariance for self-supervised sound localisation. In *ACM International Conference on Multimedia*, 2022. 1, 3, 4, 7, 8, 9, 10, 11, 14
[36] Shentong Mo and Pedro Morgado. A closer look at weakly-supervised audio-visual source localization. *Advances in Neural Information Processing Systems, NeurIPS*, 2022. 1, 3, 7, 8, 9, 11, 14

[37] Shentong Mo and Pedro Morgado. Localizing visual sounds the easy way. In *European Conference on Computer Vision*, 2022. 1, 3, 7, 8, 9, 11, 14

[38] Pedro Morgado, Ishan Misra, and Nuno Vasconcelos. Robust audio-visual instance discrimination. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 3

[39] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 3

[40] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[41] Andrew Owens and Alexei A. Efros. Audio-visual scene analysis with self-supervised multisensory features. In *European Conference on Computer Vision*, 2018. 3

[42] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 801–816. Springer, 2016. 3

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

[44] Sooyoung Park, Arda Senocak, and Joon Son Chung. MarginNCE: Robust sound localization with a negative margin. In *arXiv preprint arXiv:2211.01966*, 2022. 3, 7, 8, 9, 11, 14

[45] Rui Qian, Di Hu, Heinrich Dinkel, Mengyue Wu, Ning Xu, and Weiyao Lin. Multiple sound sources localization from coarse to fine. In *European Conference on Computer Vision*, 2020. 1, 2, 3, 7, 9

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 3, 5

[47] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 5

[48] Hyeonggon Ryu, Arda Senocak, In So Kweon, and Joon Son Chung. Hindi as a second language: Improving visually grounded speech with semantically similar samples. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2023. 4

[49] Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 10

[50] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2, 3, 6, 7, 9

[51] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound sources in visual scenes: Analysis and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5):1605–1619, 2021. 1, 2, 3, 7

[52] Arda Senocak, Hyeonggon Ryu, Junsik Kim, and In So Kweon. Learning sound localization better from semantically similar samples. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2022. 1, 3, 4, 5, 7

[53] Arda Senocak, Hyeonggon Ryu, Junsik Kim, and In So Kweon. Less can be more: Sound source localization with a classification model. In *IEEE Winter Conf. on Applications of Computer Vision (WACV)*, 2022. 1, 3, 7

[54] Arda Senocak, Hyeonggon Ryu, Junsik Kim, Tae-Hyun Oh, Hanspeter Pfister, and Joon Son Chung. Sound source localization is all about cross-modal alignment. In *IEEE International Conference on Computer Vision*, 2023. 2

[55] Zengjie Song, Yuxi Wang, Junsong Fan, Tieniu Tan, and Zhaoxiang Zhang. Self-supervised predictive learning: A negative-free method for sound source localization in visual scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2, 3, 7, 8

[56] Weixuan Sun, Jiayi Zhang, Jianyuan Wang, Zheyuan Liu, Yiran Zhong, Tianpeng Feng, Yandong Guo, Yanhao Zhang, and Nick Barnes. Learning audio-visual source localization via false negative aware contrastive learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 3, 7, 8, 9, 11, 12, 14

[57] Kim Sung-Bin, Arda Senocak, Hyunwoo Ha, Andrew Owens, and Tae-Hyun Oh. Sound to visual scene generation by audio-to-visual latent alignment. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 3

[58] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *European Conference on Computer Vision*, 2018. 3

[59] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2CLIP: Learning robust audio representations from clip. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2022. 5

[60] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 3

[61] Haohang Xu, Xiaopeng Zhang, Hao Li, Lingxi Xie, Wenrui Dai, Hongkai Xiong, and Qi Tian. Seed the views: Hierarchical semantic alignment for contrastive representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3, 4

[62] Hanyu Xuan, Zhiliang Wu, Jian Yang, Yan Yan, and Xavier Alameda-Pineda. A proposal-based paradigm for self-supervised sound source localization in videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1029–1038, 2022. 3

[63] Chen Yuanhong, Liu Yuyuan, Wang Hu, Liu Fengbei, Wang Chong, and Carneiro Gustavo. Unraveling instance associations: A closer look for audio-visual segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 6, 7

[64] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 5

[65] Yuhui Zhang, Jeff Z HaoChen, Shih-Cheng Huang, Kuan-Chieh Wang, James Zou, and Serena Yeung. Diagnosing and rectifying vision models using language. In *International Conference on Learning Representations*, 2023. 11

[66] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 7

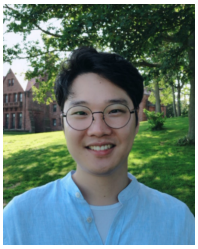
[67] Jinxing Zhou, Xuyang Shen, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, et al. Audio-visual segmentation with semantics. *arXiv preprint arXiv:2301.13190*, 2023. 14

[68] Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio-visual segmentation. In *European Conference on Computer Vision*, 2022. 6, 7, 14

Arda Senocak is a Postdoctoral Research Associate in School of Electrical Engineering, KAIST. His research interest centered around the multi-modal learning. Specifically, he is interested in building a machine perception model capable of learning from visual, auditory, and other sensory information to engage with its surroundings effectively. He completed his doctoral and master's degrees at Electrical Engineering Department of KAIST, and also received his bachelor's degree in Computer Science from KAIST. He is awarded Qualcomm Innovation Award, Samsung Human-Tech Paper Award, Google Conference Travel Grants, and Outstanding Reviewer in ICCV 2023.

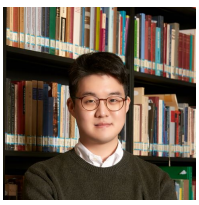
Hyeonggon Ryu is currently pursuing the Ph.D. degree and received the M.S. degree in 2023 with Electrical Engineering at KAIST, South Korea. He received the B.E. degree in Electronic Engineering from Hanyang University, South Korea in 2021. His research interests include computer vision, and Audio-Visual learning.

Junsik Kim received the BS, MS, and Ph.D. degrees in Electrical Engineering Department, KAIST, South Korea, in 2013, 2015, and 2020 respectively. He is currently a postdoctoral researcher in the School of Engineering and Applied Sciences with the Harvard University. Before joining Harvard, he was a postdoctoral researcher with KAIST. His research interests include multimodal learning and data-efficient training. He was a research intern at Hikvision Research America, Santa Clara, in 2018. He received the Qualcomm Innovation Award and ICLR'21 Outstanding Reviewer.



Tae-Hyun Oh is an assistant professor with Electrical Engineering (adjunct with Graduate School of AI and Dept. of Creative IT Convergence) at POSTECH, South Korea. He is also a research director at OpenLab, POSCO-RIST, South Korea. He received the B.E. degree (First class honors) in Computer Engineering from Kwang-Woon University, South Korea in 2010, and the M.S. and Ph.D. degrees in Electrical Engineering from KAIST, South Korea in 2012 and 2017, respectively. Before joining POSTECH, he was a postdoctoral associate at MIT CSAIL, Cambridge, MA, US, and was with Facebook AI Research, Cambridge, MA, US. He was a research intern at Microsoft Research in 2014 and 2016. He serves as an associate editor for the Visual Computer journal. He was a recipient of Microsoft Research Asia fellowship, Samsung HumanTech thesis gold award, Qualcomm Innovation awards, top research achievement awards from KAIST, and CVPR'20 and ICLR'22 outstanding reviewer awards.

Hanspeter Pfister is the An Wang Professor of Computer Science in the School of Engineering and Applied Sciences. His research in visual computing lies at the intersection of computer vision, visualization, and computer graphics. Pfister has a Ph.D. in Computer Science from Stony Brook University, New York, and an M.Sc. in Electrical Engineering from ETH Zurich, Switzerland. Before joining Harvard, he worked for over a decade at Mitsubishi Electric Research Laboratories as Associate Director and Senior Research Scientist. Pfister was elected as an ACM Fellow in 2019 and an IEEE Fellow in 2023. He received the 2010 IEEE Visualization Technical Achievement Award, the 2009 IEEE Meritorious Service Award, and the 2009 Petra T. Shattuck Excellence in Teaching Award. Pfister is a member of the ACM SIGGRAPH Academy and the IEEE Visualization Academy. He was a director of the IEEE Visualization and Graphics Technical Committee and the ACM SIGGRAPH Executive Committee.



Joon Son Chung is an assistant professor at Korea Advanced Institute of Science and Technology, where he is directing research in speech processing, computer vision and machine learning. He received the D.Phil. in Engineering Science from the University of Oxford.