Seeing Speech and Sound:
Distinguishing and Locating Audio Sources in Visual Scenes

Hyeonggon Ryu*   Seongyu Kim*   Joon Son Chung   Arda Senocak
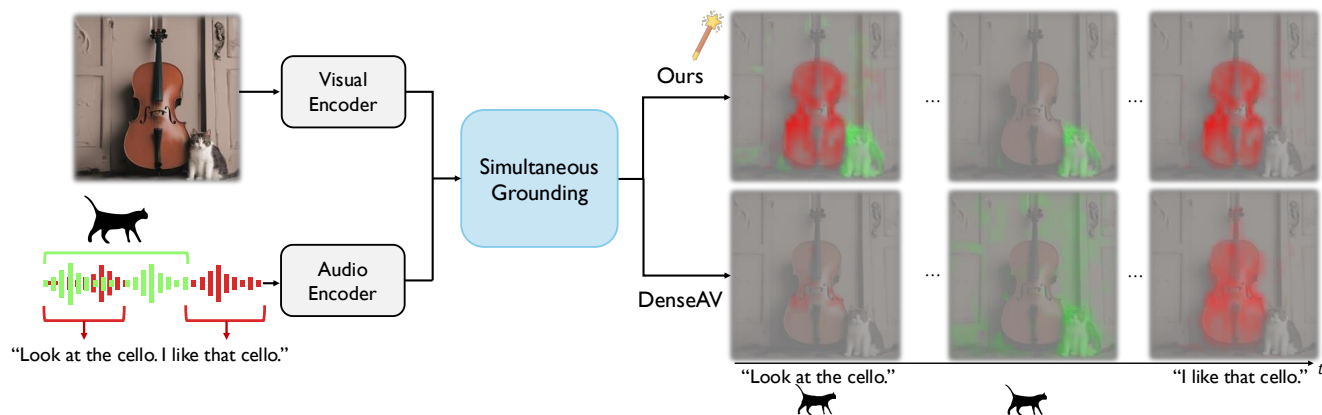
Korea Advanced Institute of Science and Technology

Figure 1. **Simultaneous Audio-Visual Grounding.** We introduce a model for simultaneous grounding of mixed audio types — including *overlapping* spoken language and non-speech sounds — within a visual scene. Our framework takes audio-visual grounding to the next level, overcoming the limitations of existing methods that handle audio types independently or sequentially without overlapping.

Abstract

We present a unified model capable of simultaneously grounding both spoken language and non-speech sounds within a visual scene, addressing key limitations in current audio-visual grounding models. Existing approaches are typically limited to handling either speech or non-speech sounds independently, or at best, together sequentially without mixing. This limitation prevents them from capturing the complexity of real-world audio sources that are often mixed. Our approach introduces a "mix-and-separate" framework with audio-visual alignment objectives that jointly learn correspondence and disentanglement using mixed audio. Through these objectives, our model learns to produce distinct embeddings for each audio type, enabling effective disentanglement and grounding across mixed audio sources. Additionally, we created a new dataset to evaluate simultaneous grounding of mixed audio sources, demonstrating that our model outperforms prior methods. Our approach also achieves comparable or better performance in standard segmentation and cross-modal retrieval tasks, highlighting the benefits of our mix-and-separate approach.

1. Introduction

We live in a world rich with multi-sensory signals. Our interactions with it rely on our natural multimodal perception, drawing on auditory and visual cues as well as spoken language to gain a complete understanding of our surroundings. While we can actively limit or focus our field of view in visual scenes, auditory signals are a mixture of different types of sounds, such as speech, music, and a variety of environmental or object sounds, all entering our ears as a combined mix without any filtering. Despite this complexity, humans can still discern specific types of sounds and associate them with their visual counterparts — for example, the sound of a "lion roaring", someone saying the word "lion", or the song "Nants' Ingonyama"[1] linked with the visual of a *lion*. Remarkably, we do this without any direct supervision.

This natural audio-visual association enables us to locate which object a spoken word refers to or identify which object is producing a sound within a visual scene. Inspired by the human ability to learn these localization skills from natural audio-visual correspondences without explicit supervision, extensive studies [11, 20, 21, 28, 29, 32–36] have

---

*These authors contributed equally to this work.

[1]Nants' Ingonyama is the iconic opening song from The Lion King https://www.youtube.com/watch?v=CF-c1K3WWg4

aimed to replicate this understanding in machine perception. However, most existing methods focus on understanding relationships between images and either spoken utterances or generic sounds, where each approach handles only one type of audio at a time but not both together.

Recently, Hamilton *et al.* [9] proposed a model that visually grounds both spoken language and non-speech sounds within an image scene. This model is trained to disentangle different types of audio, such as environmental sounds or spoken words, and to locate their sources within a visual context, all without supervision. While pioneering, this work has several limitations. First, the model handles either speech or non-speech audio individually, rather than mixed audio sources, limiting its ability to process real-world audio, which often contains overlapping sounds. Second, the model does not achieve effective disentanglement in distinguishing between different types of auditory signals (*e.g.*, speech and generic sounds). Although each of the heads in the audio encoder primarily activates for a specific audio type, these activations function as a mere selection mechanism for each type, rather than effectively controlling the flow of information between types — resulting in information leakage. Consequently, the multi-heads learn features of the opposite audio type, which impairs disentanglement and makes the model vulnerable when confronted with mixed audio inputs.

To address these limitations, we propose a unified method that *simultaneously* grounds both spoken language and non-speech sounds within a visual scene (Figure 1). Our model tackles both challenges using a "mix-and-separate" approach. We introduce a novel mixed audio-visual alignment objective that jointly learns *correspondence* and *disentanglement*, aligning representations of multiple audio types with corresponding visuals in images. Our framework includes shared audio and visual encoders within each modality, facilitating the transfer of audio-visual interactions across learning objectives. Unlike traditional mix-and-separate approaches, we do not generate separated audio; instead, we obtain embeddings from each specialized head of the audio encoder that represent the disentangled sound sources of different types using contrastive learning. These learning objectives enable our framework to process audio mixtures, eliminating the need for multiple audio types to appear sequentially or individually during training and inference. Additionally, this approach improves disentanglement within the multi-head structure.

The primary motivation of this paper is to achieve simultaneous grounding of multiple audio types; accordingly, we aim to evaluate our model's ability to perform this task. However, no existing dataset is suitable for this purpose. To address this, we created a new dataset by extending the IS3 dataset [36] — which provides segmentation masks for multiple sound-producing objects with corresponding au-

dio — by adding spoken utterances to each sample. Evaluation on this dataset demonstrates that our model outperforms existing methods in simultaneous sound and speech grounding, a task that requires handling mixed audio inputs. Additionally, by achieving superior performance on standard segmentation and cross-modal retrieval experiments on the datasets from [9], we show the benefits of our self-supervised mix-and-separate approach.

Our main contributions are as follows:
- We propose a method for *simultaneously* grounding both speech and non-speech audios within a visual scene.
- We introduce a novel audio-visual alignment objective that uses *audio mixtures* and jointly learns *correspondence* and *disentanglement* using contrastive learning.
- We provide a new dataset, *Extended-IS3*, for simultaneous sound and speech grounding and demonstrate that our model outperforms existing methods.

2. Related Works

Sound source localization. Audio-visual sound source localization is the task of identifying objects, events, or areas within a visual scene based on corresponding audio cues. In the deep learning era, Senocak *et al.* [32, 33] pioneered this field by introducing a cross-modal attention mechanism with contrastive learning to align audio and visual information. As this research field advanced, diverse techniques have emerged, including hard sample mining [3], multiple-instance contrastive learning [21], negative-free strategies [38], momentum encoders [20], intra-modality similarity learning [39], and the use of multiple positives in contrastive learning [35]. Recently, CLIP has also been incorporated into sound source localization without using text [27]. Most prior methods focus on single-source localization, while another line of research addresses multiple sound source localization with generic sounds [15, 17–19, 23, 30]. In contrast to these models, which primarily localize generic sounds, our work targets both generic sounds and spoken utterances, either simultaneously or individually, with a single model. Our approach addresses the added complexity of mixed generic sounds and spoken language, requiring a deeper understanding of language and its integration with visual scenes.

Visually grounded speech. Visually grounded speech (VGS) aims to associate images with corresponding spoken utterances. Various tasks have been explored in this domain, including cross-modal retrieval [28, 43], multilingual associations across languages using images as an interlingua [11, 24, 25, 31], speech unit or word discovery and keyword spotting [13, 29, 37], and visual grounding [12]. Our approach aligns with the visual grounding work of [12]; however, it differentiates itself from previous work by enabling the grounding of both spoken utterances and generic
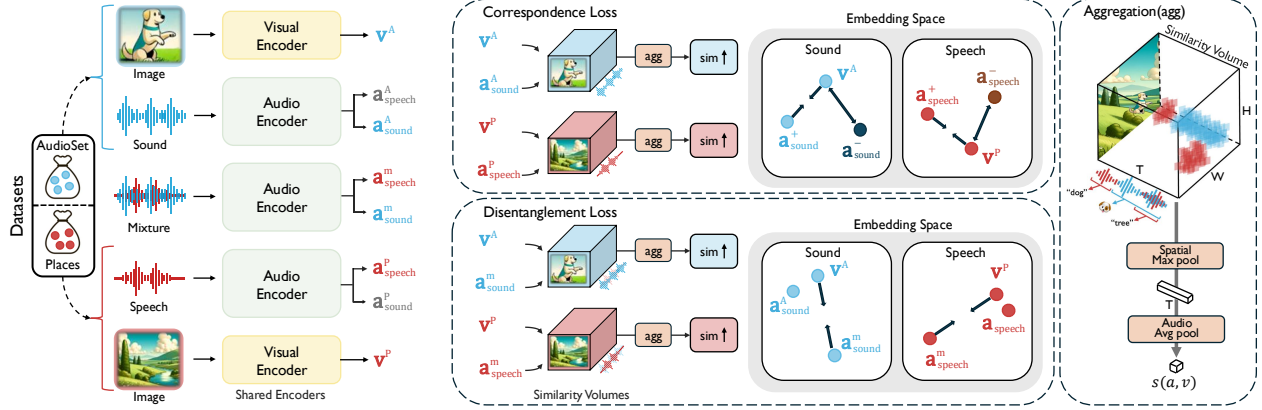
Figure 2. **The pipeline of our framework.** Visual and audio encoders extract features from clean audios, their paired images, and mixed audio inputs. These features are used to compute audio-visual similarities, which are then used in our correspondence and disentanglement losses. The correspondence objective ensures audio-visual matching, while the disentanglement objective enables feature-level separation of mixed audio sources.

sounds within a single model. This broader capability allows our model to handle diverse audio inputs, advancing VGS beyond isolated spoken language associations.

Audio-visual sound source separation. Audio-only sound source separation has a long history, and with the advances in deep learning, audio-visual sound source separation has become an established field. Audio-visual separation methods leverage visual cues to isolate individual sound sources, often employing a mix-and-separate approach. As a pioneering work, Zhao *et al.* [44] introduced a method that uses pixel-level correspondence between visual information and audio to separate sound sources in images. Subsequent approaches have incorporated additional cues, such as scene graphs [2], motion [45], gestures and keypoints [4], to enhance separation performance. The majority of these works focus on general sounds, such as objects or musical instruments, while other lines of research address multi-speaker speech separation [5, 26] and distinguishing on-screen sounds from off-screen sounds [26, 42]. While most prior work focuses exclusively on audio separation, [22, 41] propose models that jointly optimize separation and localization within a unified architecture for generic sounds only. Like previous approaches, our method employs a mix-and-separate strategy; however, rather than generating separated audio, we perform separation at the feature level, yielding features that represent distinct sound sources. Additionally, whereas existing models typically specialize in either speech or specific type of general sounds, our approach introduces a unified model capable of separating and localizing both generic sounds and spoken language, either simultaneously or individually, within a single framework.

3. Methodology

Our goal is to perform *simultaneous* grounding of both speech and non-speech audio within a visual scene using a single model. To this end, we propose a self-supervised "mix-and-separate" strategy. Given an audio mixture and its corresponding image, the model parses the audio into speech and sound sources, and then localizes them within the image. We frame this task as a representation learning problem, using two contrastive losses—correspondence and disentanglement—which are jointly optimized to transfer complementary information across objectives. Our proposed pipeline is shown in Figure 2.

3.1. Preliminaries

Contrastive learning aims to learn representations by distinguishing between positive and negative pairs. In audio-visual learning, for an encoded audio sample $\mathbf{a}_i = f_a(a_i)$ and its corresponding positive image feature pair $\mathbf{v}_i = f_v(v_i)$, with negative pairs $\mathbf{v}_j$ for $i \neq j$, from a dataset $\mathcal{D} = \{(v_i, a_i) : i = 1, ..., N\}$, where v_i and a_i are image and audio sample, f_v and f_a are image and audio encoder respectively, the loss can be formulated as:

$$\mathcal{L}_i = -\log \frac{\exp(s(\mathbf{a}_i, \mathbf{v}_i)/\tau)}{\sum_j \exp(s(\mathbf{a}_i, \mathbf{v}_j)/\tau)}, \tag{1}$$

where s refers to the cross modal similarity function and τ is a trainable temperature parameter. The contrastive loss Eq. (1) is applied symmetrically by following previous works [9, 21, 40].

Audio and visual encoders with aligner. Given an image v_i with its corresponding audio a_i, our backbone networks extract features for each modality. The vision backbone f_v processes an image frame to extract a spatial feature map, $\mathbf{v_d} \in \mathbb{R}^{C_D \times H \times W}$, while the audio backbone f_a processes the raw waveform to extract an audio feature, $\mathbf{a_h} \in \mathbb{R}^{C_H \times F \times T}$. Each modality encoder has its own aligner network attached at the end of the backbones. These

aligners consist of a Channel-wise LayerNorm block and convolution layers, which reshapes the image spatial feature map to $\mathbf{v} \in \mathbb{R}^{C \times K \times H \times W}$ and the audio feature to $\mathbf{a} \in \mathbb{R}^{C \times K \times F \times T}$. Here, K denotes the number of audio types (set to 2), meaning there is a specific head for each audio type, while F, T, H and W represent the frequency, temporal, and spatial axes, respectively. C_D, C_H and C refers to channel dimension of vision encoder, audio encoder and shared feature space respectively.

Similarity function. To obtain distinct features representing each audio sub-modality, it is essential to use a function that computes the similarity between audio and image with consideration of the audio type – the two embedding spaces for speech and sound. We then build a similarity volume $S(\mathbf{a}, \mathbf{v}) \in \mathbb{R}^{K \times F \times T \times H \times W}$ from the image spatial feature map and the audio feature:

$$S(\mathbf{a}, \mathbf{v}) = \sum_{c=1}^{C} \mathbf{a}[c,k,f,t] \cdot \mathbf{v}[c,k,h,w], \quad (2)$$

where $\mathbf{a}[c,k,f,t]$ refers to the value of $\mathbf{a}$ at location $[c,k,f,t]$ and $\cdot$ denotes scalar multiplication. Thus, each audio type has its own similarity volume with the corresponding image. Aggregation across the head axis can be done in two different ways, either through max pooling or head selection:

$$S_M(\mathbf{a}, \mathbf{v}) = \max_k (S(\mathbf{a}, \mathbf{v})), \quad (3)$$

$$S_N(\mathbf{a}, \mathbf{v}) = S(\mathbf{a}, \mathbf{v})[N], \quad (4)$$

where $[N]$ follows PyTorch slicing to select activations for the *N-th* audio type head. After aggregation along the head axis, spatial max pooling is applied, followed by average pooling across the frequency and temporal axes to define the similarity score $s_M(\mathbf{a}, \mathbf{v})$ or $s_N(\mathbf{a}, \mathbf{v})$ (shown in Figure 2):

$$s_M(\mathbf{a}, \mathbf{v}) = \frac{1}{FT} \sum_{f=1}^{F} \sum_{t=1}^{T} \max_{h,w} (S_M(\mathbf{a}, \mathbf{v})[f,t,h,w]), \quad (5)$$

$$s_N(\mathbf{a}, \mathbf{v}) = \frac{1}{FT} \sum_{f=1}^{F} \sum_{t=1}^{T} \max_{h,w} (S_N(\mathbf{a}, \mathbf{v})[f,t,h,w]), \quad (6)$$

where we use either N or N' for $s_N(\mathbf{a}, \mathbf{v})$, denoting the index of the head specialized in sound and speech respectively.

3.2. Mix-and-Separate

Our method employs a "mix-and-separate" approach with a novel mixed audio-visual alignment objectives that jointly learn both *correspondence* and *disentanglement*, aligning representations of multiple audio types with their corresponding visuals in images.

Correspondence loss is designed to establish shared audio-visual feature spaces. By using clean sound or speech samples as input audio, the correspondence loss encourages the model to build associations between each audio sample and its corresponding image pair. Given the clean audio feature of sound and speech, and the paired image feature map of each audio type, denoted as $\mathbf{a}^{\mathrm{A}}$, $\mathbf{a}^{\mathrm{P}}$, $\mathbf{v}^{\mathrm{A}}$, and $\mathbf{v}^{\mathrm{P}}$, respectively, the correspondence loss is formulated as follows:

$$\mathcal{L}_{cor} = \sum_i -\log \frac{\exp(s_M(\mathbf{a}_i^{\mathrm{A}}, \mathbf{v}_i^{\mathrm{A}})/\tau)}{\sum_j \exp(s_M(\mathbf{a}_i^{\mathrm{A}}, \mathbf{v}_j^{\mathrm{A}})/\tau)}$$
$$+ \sum_i -\log \frac{\exp(s_M(\mathbf{a}_i^{\mathrm{P}}, \mathbf{v}_i^{\mathrm{P}})/\tau)}{\sum_j \exp(s_M(\mathbf{a}_i^{\mathrm{P}}, \mathbf{v}_j^{\mathrm{P}})/\tau)}. \quad (7)$$

We visualize the features in Figure 2 for better understanding:

$$\mathbf{a}_{\mathrm{sound}}^{\mathrm{A}} = \mathbf{a}^{\mathrm{A}}[N], \mathbf{a}_{\mathrm{speech}}^{\mathrm{A}} = \mathbf{a}^{\mathrm{A}}[N'],$$
$$\mathbf{a}_{\mathrm{sound}}^{\mathrm{P}} = \mathbf{a}^{\mathrm{P}}[N], \mathbf{a}_{\mathrm{speech}}^{\mathrm{P}} = \mathbf{a}^{\mathrm{P}}[N']. \quad (8)$$

Disentanglement loss in our model is designed to ensure the disentanglement of audio type heads. To achieve this, we adopt a *mix-and-separate* strategy that uses audio mixtures containing both audio types. The audio encoder generates distinct features for each sub-modality, represented by two separate audio heads that align with features derived from paired image samples for both speech and sound. In this audio-visual feature space, each image feature is already close to its corresponding clean audio feature. The model learns disentanglement by aligning the features from the mixed audio with the corresponding image features and, inherently, with the features from the clean audios as shown in Figure 2. Given the feature of mixed audio that is a combination of sound and speech, $\mathbf{a}_{\mathrm{m}}$, and the paired image feature map from each of sound and speech, $\mathbf{v}_{\mathrm{A}}$, and $\mathbf{v}_{\mathrm{P}}$, the disentanglement loss is formulated as follows:

$$\mathcal{L}_{dis} = \sum_i -\log \frac{\exp(s_N(\mathbf{a}_i^{\mathrm{m}}, \mathbf{v}_i^{\mathrm{A}})/\tau)}{\sum_j \exp(s_N(\mathbf{a}_i^{\mathrm{m}}, \mathbf{v}_j^{\mathrm{A}})/\tau)}$$
$$+ \sum_i -\log \frac{\exp(s_{N'}(\mathbf{a}_i^{\mathrm{m}}, \mathbf{v}_i^{\mathrm{P}})/\tau)}{\sum_j \exp(s_{N'}(\mathbf{a}_i^{\mathrm{m}}, \mathbf{v}_j^{\mathrm{P}})/\tau)}. \quad (9)$$

The features are visualized in Figure 2 for better understanding:

$$\mathbf{a}_{\mathrm{sound}}^{\mathrm{m}} = \mathbf{a}^{\mathrm{m}}[N], \mathbf{a}_{\mathrm{speech}}^{\mathrm{m}} = \mathbf{a}^{\mathrm{m}}[N']. \quad (10)$$

Total loss is formulated as a mixed audio-visual alignment objective to jointly achieve both correspondence and disentanglement, as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{cor} + \mathcal{L}_{dis}. \quad (11)$$

Our proposed model is trained end-to-end together with this training objective.

3.3. Implementation and Training Details

3.3.1. Architecture Details

Image encoder $f_v(\cdot)$**.** We use DINO [1], pre-trained on ImageNet in a self-supervised manner. An aligner module, consisting of a Channel-wise LayerNorm block and convolutional layers, is added to the end of DINO. We apply LoRA [16] to fine-tune only the attention blocks of DINO in our audio-visual end-to-end training objective.

Audio encoder $f_a(\cdot)$**.** We use HuBERT Large [14], pre-trained on Libri-Light in a self-supervised manner. HuBERT is one of the de facto models in speech processing, specifically designed and pre-trained for speech. An aligner module, consisting of a Channel-wise LayerNorm block and convolutional layers, is added to the end of HuBERT. The audio encoder is fully fine-tuned within our audio-visual end-to-end training objective.

3.3.2. Implementation Details

Inputs to our model consist of a single 224 x 224 image and a 10-second audio waveform sampled at 16 kHz. We pre-process images and audio following prior work [9], applying standard image augmentations and trimming or padding audio to 10 seconds. The model is trained for 800K steps on 8 A6000 GPUs with an effective batch size of 64, using a LoRA rank of 8 for image encoder. Additional details are provided in the supplementary material.

Other details. To ensure stable model convergence, we incorporate two additional techniques, following [9]. First, a warm-up stage for the aligners — attached to each modality encoder — is applied, during which the image and audio backbone networks remain frozen, and only the aligners are trained for 3K steps using our correspondence loss objective. Once the warm-up stage ends, the entire model is trained end-to-end as described before. Second, additional regularizer losses are applied to enhance training stability. Please refer to the supp. material and [9] for further details. We emphasize that while these techniques do not significantly impact the model's performance, they contribute to more stable training.

4. Experiments

4.1. Datasets

Training datasets. Our model is trained using the AudioSet [6] and PlacesAudio [10] datasets. AudioSet is a large-scale audio-visual dataset containing ~2M samples of general sounds, while PlacesAudio includes ~400K samples of spoken utterances paired with images. Since our method processes mixture of audios, we combine each PlacesAudio sample with a randomly selected AudioSet

sample, using their respective images as part of the training process (see Figure 2).

Testing datasets. After training, we test the grounding and retrieval performances with the datasets below.
- **Test sets from DenseAV [9].** For standard audio-visual grounding, we use the ADE20K Speech and Sound dataset from [9], which includes 3,030 image-audio pairs for speech and 106 for sound. For retrieval tasks, we use the PlacesAudio test set and a 1,000-sample subset of the AudioSet validation set, selected randomly following [9].
- **Extended IS3 [36].** The IS3 Dataset is a test set consisting of 3,420 images for evaluating interactive sound source localization, where two objects are present in the visual scene. Each image is paired with two audio samples that represent each visible object. We extend this dataset by generating two individual speech audio samples for each image, each reading the class category of the sounding objects in the scene. This extension provides every image with both the generic sounds of visible objects and their spoken labels. The speech samples are generated using the Google gTTS API. This new dataset enables us to assess whether models can simultaneously ground both audio types within a mixture.

4.2. Baselines

We compare our method against the following baselines. ImageBind [7] is a multimodal foundation model trained on large-scale datasets across several modalities, including text. CAV-MAE [8] is a state-of-the-art model designed for audio-visual learning, achieving strong performance on tasks like audio-visual retrieval. DenseAV [9] is the most recent and closely related work, as discussed earlier. We report results using the released checkpoints of this model from its official page.

4.3. Main Results

In this section, we present results comparing our proposed method with prior and closely related works on simultaneous audio-visual grounding, speech or generic sound prompted visual grounding, and cross-modal retrieval tasks.

4.3.1. Simultaneous Audio-Visual Grounding

The main goal of this paper is *simultaneously* grounding both speech and non-speech audios within a visual scene; therefore, in this section, we aim to evaluate our model's ability to perform this task.

Implementation. In this experiment, we use our *Extended IS3* dataset. Audio input is a mixed audio that is the summation of speech and sound. The evaluation metrics are mean average precision (mAP) and intersection over union (mIoU). Simultaneous semantic segmentation results for each audio type are obtained using the corresponding head.
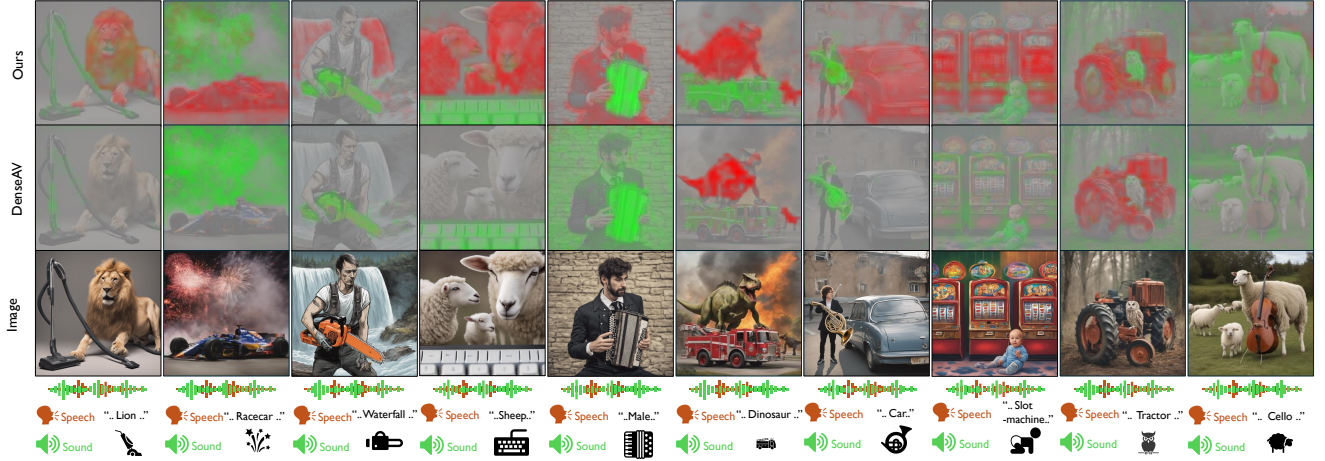
Figure 3. **Qualitative results for simultaneous audio-visual grounding on Extended IS3 dataset.** Our model accurately localizes both overlapping audio types simultaneously within the mixed audio, whereas competing method [9] cannot.

Method	Sim-Sound Seg.		Sim-Speech Seg.	
	mAP	mIoU	mAP	mIoU
Foundation Model ImageBind [7]$_{CVPR23}$	17.9	19.2	24.2	22.6
CAVMAE [8]$_{ICLR23}$	31.0	27.9	15.6	19.0
DenseAV [9]$_{CVPR24}$	30.7	27.0	53.4	44.0
Ours	**43.4**	**34.6**	**58.6**	**47.0**

Table 1. **Simultaneous Audio-Visual segmentation results.**

Method	Clean SemSeg							
	Sound SemSeg				Speech SemSeg			
	Total		Sound		Total		Speech	
	mAP	mIoU	mAP	mIoU	mAP	mIoU	mAP	mIoU
Foundation Model ImageBind [7]$_{CVPR23}$	18.3	18.2	N/A	N/A	20.2	19.8	N/A	N/A
CAVMAE [8]$_{ICLR23}$	21.6	20.7	N/A	N/A	17.6	19.4	N/A	N/A
DenseAV [9]$_{CVPR24}$	30.3	23.0	30.9	23.3	**43.0**	32.8	**47.7**	36.2
Ours	**36.7**	**27.0**	**32.8**	**24.3**	42.1	32.8	46.7	35.4

Table 2. **Audio-Visual semantic segmentation with individual audios.**

Quantitative results. We present our results in Table 1, demonstrating that our model outperforms the recent work [9], DenseAV, by **+12.7 mAP** and **+5.2 mAP** for Sound and Speech segmentation, respectively, in the simultaneous grounding scenario. These results indicate that our proposed model successfully learns to localize distinct audio types simultaneously through our mix-and-separate approach and specific learning objectives.

Qualitative results. To further evaluate our method, we visualize and compare our simultaneous audio-visual segmentation results with recent prior work. The visual samples in Figure 3 again demonstrate that our model accurately grounds both spoken words and sound-producing objects, even when speech and generic sounds overlap.

4.3.2. Individual Audio-Visual Grounding

Here, we validate the effectiveness of our proposed method in the standard semantic segmentation task as in [9], using two types of audio individually: speech and generic sound. The task here is that, for a given audio input, the model generates attention maps for the *meaning* of the spoken word or the *object* that produces the sound.

Implementation. We follow the evaluation protocol in DenseAV [9], using the same test sets as in [9] (Section 4.1) and employing mAP and mIoU as evaluation metrics. Our

model is evaluated with two aggregation approaches across the heads: *total* and *specialized head*. The *total* approach uses element-wise max pooling across all heads, while the *specialized head* approach selects the head specialized for the target audio type. For speech grounding, we average activations over the duration of the "[object]" in the prompt "A picture of a(n) [object]", using word timing information from the speech.

Results on sound prompted visual grounding. In this experiment, we conduct a comparative analysis of our proposed method using only generic sound. We present our results in Table 2. Our method achieves improvements of **+6.4 mAP** and **+1.9 mAP** in the *Total* and *Sound Head* evaluations, respectively, compared to previous work [9]. This is particularly important given that the audio backbone network is HuBERT [14], one of the de facto models in speech processing, specifically designed and pre-trained for speech. Adapting HuBERT to better understand generic sounds demonstrates the effectiveness of our approach.

Results on speech prompted visual grounding. Similar to the previous experiment, we evaluate our model using individual speech only, with results shown in Table 2. These results indicate that our method performs approximately similar to Hamilton *et al.* [9], though slightly lower, on the speech-prompted segmentation task. This outcome aligns

Method	Clean Retrieval R@10							
	AudioSet				Places			
	Total		Sound		Total		Speech	
	I→A	A→I	I→A	A→I	I→A	A→I	I→A	A→I
Foundation Model								
ImageBind [7]_{CVPR23}	66.8	65.0	N/A	N/A	0.9	1.1	N/A	N/A
CAVMAE [8]_{ICLR23}	43.1	31.4	N/A	N/A	1.0	1.0	N/A	N/A
DenseAV [9]_{CVPR24}	41.8	43.6	46.7	46.5	**92.7**	92.4	93.8	93.9
Ours	**45.5**	**46.6**	**51.2**	**50.0**	92.0	**93.1**	**94.0**	**94.0**

Table 3. **Cross-modal retrieval task on Places and AudioSet.**

with our expectations because our approach enhances the model's understanding of non-speech sounds, allowing the sound head to specialize more effectively in non-speech audio without compromising the speech head's ability to accurately interpret spoken language.

4.3.3. Cross-Modal Retrieval

After demonstrating our method's capabilities in both simultaneous and standard audio-visual grounding tasks, we also evaluate its performance on cross-modal retrieval.

Implementation. This experiment follows the same evaluation protocol of [9]. We assess cross-modal retrieval using the PlacesAudio and AudioSet test sets as described in Section 4.1, using cross-modal retrieval accuracy@10 as the evaluation metric. For this experiment, the total aggregation approach involves summing across the heads, as in [9], different than individual audio-visual grounding experiments.

Results. Learning audio-visual associations is crucial for sound source localization. Therefore, models that perform well in sound source localization are also expected to do well in cross-modal retrieval [36], as they learn how objects look and sound, or how objects appear when referenced by spoken words. Table 3 shows the results for cross-modal retrieval tasks. As expected, our method significantly outperforms DenseAV in cross-modal retrieval on the AudioSet for both Audio→Image and Image→Audio tasks, achieving substantial improvements in both *Total* and *Sound Head* evaluations. On the Places dataset, our method performs almost identically to DenseAV across all experiments and head evaluations, following the same trend observed in the standard audio-visual segmentation task (Section 4.3.2). We note that our model is outperformed by ImageBind [7] in cross-modal retrieval on AudioSet. This result is expected, given that ImageBind is a foundational model trained on large-scale datasets spanning multiple modalities, including text. However, ImageBind performs poorly on the Places dataset, whereas our method shows strong performance, underscoring its ability to effectively handle both types of audio.

4.3.4. Standard Experiments with Mixed Audio

Inspired by Mo *et al*. [20], we added sound or spoken utterances from non-visible (*i.e.*, off-screen) objects to the DenseAV datasets used in our experiments in Section 4.3.2

Method	Mixed SemSeg							
	Sound SemSeg				Speech SemSeg			
	Total		Sound		Total		Speech	
	mAP	mIoU	mAP	mIoU	mAP	mIoU	mAP	mIoU
Foundation Model								
ImageBind [7]_{CVPR23}	18.8	18.5	N/A	N/A	19.3	19.5	N/A	N/A
CAVMAE [8]_{ICLR23}	19.6	18.9	N/A	N/A	18.6	19.9	N/A	N/A
DenseAV [9]_{CVPR24}	19.6	19.0	18.7	18.6	30.5	24.2	33.0	25.6
Ours	**22.3**	**20.2**	**28.4**	**22.9**	**39.3**	**29.5**	**41.6**	**31.2**

Table 4. **Standard Audio-Visual semantic segmentation with mixture of audios.**

Method	Mixed Retrieval R@10							
	AudioSet				Places			
	Total		Sound		Total		Speech	
	I→A	A→I	I→A	A→I	I→A	A→I	I→A	A→I
Foundation Model								
ImageBind [7]_{CVPR23}	35.5	35.0	N/A	N/A	1.20	0.70	N/A	N/A
CAVMAE [8]_{ICLR23}	**19.1**	8.9	N/A	N/A	0.8	1.3	N/A	N/A
DenseAV [9]_{CVPR24}	12.3	10.4	14.4	13.4	49.6	54.2	50.6	58.7
Ours	17.8	**20.3**	**44.3**	**42.6**	**79.0**	**82.8**	**87.3**	**86.2**

Table 5. **Cross-modal retrieval task on Places and AudioSet with mixture of audios.**

and Section 4.3.3. This addition creates a noisy dataset with mixed audio types. In this scenario, we can observe the robustness of the models to different audio types, as well as the degree of disentanglement—that is, how effectively each head is specialized only for its corresponding type of audio-visual association. Note that this setup differs from simultaneous localization experiments in two ways: (1) In the simultaneous setup, both audio types in the mixture have visual counterparts in the image, whereas here, the added audio of the opposite type is off-screen. (2) In simultaneous audio-visual segmentation, the model attempts to localize both audio types at once. In this setup, however, the model is tasked with localizing or retrieving information using only the target audio type, while ignoring any off-screen audio that lacks a visual counterpart.

Implementation. We repeat the experiments from Section 4.3.2 and Section 4.3.3 — *audio-visual segmentation* and *cross-modal retrieval* - using identical implementation details, except that we employ a mixture of audio types as described above.

Results. Experimental results are presented in Table 4 and Table 5. As shown, our method significantly outperforms DenseAV across both tasks and evaluation approaches within each task. While the performance of both models declines with mixed audio inputs, the drop in performance is much smaller for our method (compare Table 2 vs. Table 4 and Table 3 vs. Table 5). We hypothesize that this performance gap and the smaller drop ratios with our method, compared to clean audio inputs, are due to two key factors: (1) Our model is more robust when processing mixed audio inputs due to the mix-and-separate approach and our joint learning objectives. (2) DenseAV's special-

Method	Retrieval R@10							
	Clean				Mixed			
	Places		AudioSet		Places		AudioSet	
	Sound		Speech		Sound		Speech	
	I→A	A→I	I→A	A→I	I→A	A→I	I→A	A→I
DenseAV	13.5	31.6	**3.4**	**6.9**	4.4	12.6	1.8	1.5
Ours	**2.3**	**10.5**	3.7	8.0	**1.4**	**1.4**	1.6	1.3

Table 6. **Measuring disentanglement.** Each head is tasked with retrieval by using the opposite audio type to reveal retained information within each head. A lower number is better performance.

	AudioSet			Places		
Method	**I→A**	**A→I**	**Method**	**I→A**	**A→I**
L_{dis}	30.9	32.0	L_{cor}	23.7	42.1
Ours	**51.2**	**50.0**	**Ours**	**2.3**	**10.5**
(a)			(b)		

Table 7. **Ablation results for training losses.**

ized heads may retain some information from the opposite audio type, which causes disturbances in mixed audio scenarios, while our heads are more effectively disentangled and specialized for their intended audio type. Also, CAV-MAE performs better in the *Total Head* evaluation for I→A in AudioSet, but our specialized *Sound Head* outperforms it significantly (44.3 vs. 19.1) and should be considered the target for this task, as Total Head is an alternative evaluation approach.

4.3.5. Disentanglement Assessment

The success of the model using different heads for the target task is highly dependent on how well each head is specialized for its respective type, in other words, the level of disentanglement. Hamilton *et al.* [9] evaluate disentanglement in two ways. First, they measure whether a head's average activation strength can reliably indicate if a given sample predominantly contains "language" or "sound" - *Pred.Dis*. Second, they assess the frequency with which the "sound" head incorrectly activates when only the "language" head should be active, and vice versa - *Act.Dis*. We followed the evaluation protocols from [9] exactly. The results for Pred.Dis. are 99.92% for ours vs 99.91% for DenseAV, and for Act.Dis., 86.90% for ours vs. 84.48% for DenseAV. As seen, the performances are almost identical, as this approach merely acts as a selection mechanism for each type. However, depending on the level of disentanglement, each head may still retain some information from the opposite type of audio (*i.e.*, information leakage), which could hinder the model's performance. To assess how much information each head still retains, and indirectly evaluate the disentanglement ability, we perform retrieval tasks using the opposite audio-type heads.

Results. As shown in Table 6, our method generally yields lower performance compared to DenseAV, which suggests that the heads produced by our method contain less information about the opposite type of audio. Additionally, we observe that our proposed approach has improved the specialization of the sound head more than the speech head, likely because the audio backbone network is already optimized for speech. Overall, these results demonstrate the effectiveness of our approach for improving disentanglement.

4.4. Ablation Results

We conduct experiments to verify that each of our audio-visual alignment objectives — correspondence and disentanglement — functions as intended. Using the "Sound Head" of our model for retrieval tasks throughout this ablation study, we first assess the effect of omitting the correspondence loss by measuring retrieval performance on AudioSet, which targets the same audio type. As shown in Table 7(a), omitting the correspondence loss leads to a ~20% drop in performance, confirming that this loss effectively enables audio-visual matching necessary for cross-modal retrieval. In the second experiment, we examine the impact of omitting the disentanglement loss by evaluating "Sound Head" performance on the Places dataset, which targets the opposite audio type. Here, we expect poor disentanglement to cause the Sound Head to resemble the Speech Head, yielding high retrieval performance for speech. Results in Table 7(b) show a ~22% performance increase for the "Sound Head", indicating it retains speech information due to inadequate disentanglement. In summary, these ablation studies confirm that our learning objectives perform as designed.

5. Conclusion

In this work, we introduce the first unified model for simultaneous grounding of mixed audio types — including spoken language and non-speech sounds — within a visual scene. Using a novel "mix-and-separate" framework with joint correspondence and disentanglement objectives, our approach learns distinct representations for each audio type, addressing key limitations of prior models limited to handling single or sequential audio sources. This framework enables our model to produce accurate, distinct embeddings for each audio type, achieving robust disentanglement and grounding even with overlapping mixed audio. We also present a new dataset for evaluating simultaneous audio-visual grounding with mixed audio inputs, where our model outperforms previous methods across this and other segmentation and retrieval tasks. In summary, our contributions establish a new paradigm for audio-visual grounding in multi-source environments, advancing machine perception capabilities for real-world auditory-visual interactions.

6. Acknowledgment

References

[1] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 5

[2] Moitreya Chatterjee, Jonathan Le Roux, Narendra Ahuja, and Anoop Cherian. Visual scene graphs for audio source separation. In *ICCV*, 2021. 3

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

[4] Chuang Gan, Deng Huang, Hang Zhao, Joshua B Tenenbaum, and Antonio Torralba. Music gesture for visual sound separation. In *CVPR*, 2020. 3

[5] Ruohan Gao and Kristen Grauman. Visualvoice: Audio-visual speech separation with cross-modal consistency. In *CVPR*, 2021. 3

[6] Jort F Gemmeke, Daniel PW 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 *ICASSP*, 2017. 5

[7] 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 *CVPR*, 2023. 5, 6, 7

[8] Yuan Gong, Andrew Rouditchenko, Alexander H. Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James R. Glass. Contrastive audio-visual masked autoencoder. In *ICLR*, 2023. 5, 6, 7

[9] 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 *CVPR*, 2024. 2, 3, 5, 6, 7, 8, 1, 4

[10] David Harwath, Antonio Torralba, and James Glass. Unsupervised learning of spoken language with visual context. In *NeurIPS*, 2016. 5

[11] David Harwath, Galen Chuang, and James Glass. Vision as an interlingua: Learning multilingual semantic embeddings of untranscribed speech. In *ICASSP*, 2018. 1, 2

[12] David Harwath, Adria Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass. Jointly discovering visual objects and spoken words from raw sensory input. In *ECCV*, 2018. 2

[13] David Harwath, Wei-Ning Hsu, and James Glass. Learning hierarchical discrete linguistic units from visually-grounded speech. In *ICLR*, 2020. 2

[14] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. In *IEEE/ACM TASLP*, 2021. 5, 6

[15] 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. *NeurIPS*, 2020. 2

[16] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 5

[17] Xixi Hu, Ziyang Chen, and Andrew Owens. Mix and localize: Localizing sound sources in mixtures. In *CVPR*, 2022. 2

[18] Dongjin Kim, Sung Jin Um, Sangmin Lee, and Jung Uk Kim. Learning to visually localize sound sources from mixtures without prior source knowledge. In *CVPR*, 2024.

[19] Tanvir Mahmud, Yapeng Tian, and Diana Marculescu. T-vsl: Text-guided visual sound source localization in mixtures. In *CVPR*, 2024. 2

[20] Shentong Mo and Pedro Morgado. A closer look at weakly-supervised audio-visual source localization. In *NeurIPS*, 2022. 1, 2, 7

[21] Shentong Mo and Pedro Morgado. Localizing visual sounds the easy way. In *ECCV*, 2022. 1, 2, 3

[22] Shentong Mo and Pedro Morgado. A unified audio-visual learning framework for localization, separation, and recognition. In *ICML*, 2023. 3

[23] Shentong Mo and Yapeng Tian. Audio-visual grouping network for sound localization from mixtures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

[24] Yasunori Ohishi, Akisato Kimura, Takahito Kawanishi, Kunio Kashino, David Harwath, and James Glass. Pair expansion for learning multilingual semantic embeddings using disjoint visually-grounded speech audio datasets. In *INTERSPEECH*, 2020. 2

[25] Yasunori Ohishi, Akisato Kimura, Takahito Kawanishi, Kunio Kashino, David Harwath, and James Glass. Trilingual semantic embeddings of visually grounded speech with self-attention mechanisms. In *ICASSP*, 2020. 2

[26] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *ECCV*, 2018. 3

[27] Sooyoung Park, Arda Senocak, and Joon Son Chung. Can clip help sound source localization? In *WACV*, 2024. 2

[28] Puyuan Peng and David Harwath. Fast-slow transformer for visually grounding speech. In *ICASSP*, 2022. 1, 2

[29] Puyuan Peng and David Harwath. Word discovery in visually grounded, self-supervised speech models. In *INTERSPEECH*, 2022. 1, 2

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

[31] 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 *ICASSP*, 2023. 2

[32] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *CVPR*, 2018. 1, 2

[33] 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 TPAMI*, 2021. 2

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

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

[36] Arda Senocak, Hyeonggon Ryu, Junsik Kim, Tae-Hyun Oh, Hanspeter Pfister, and Joon Son Chung. Aligning sight and sound: Advanced sound source localization through audio-visual alignment. *arXiv preprint arXiv:2407.13676*, 2024. 1, 2, 5, 7

[37] Yi-Jen Shih, Hsuan-Fu Wang, Heng-Jui Chang, Layne Berry, Hung-yi Lee, and David Harwath. SpeechCLIP: Integrating speech with pre-trained vision and language model. In *SLT*, 2023. 2

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

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

[40] 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 *CVPR*, 2023. 3

[41] Yapeng Tian, Di Hu, and Chenliang Xu. Cyclic co-learning of sounding object visual grounding and sound separation. In *CVPR*, 2021. 3

[42] Efthymios Tzinis, Scott Wisdom, Aren Jansen, Shawn Hershey, Tal Remez, Daniel PW Ellis, and John R Hershey. Into the wild with audioscope: Unsupervised audio-visual separation of on-screen sounds. In *ICLR*, 2020. 3

[43] Jongbhin Woo, Hyeonggon Ryu, Arda Senocak, and Joon Son Chung. Speech guided masked image modeling for visually grounded speech. In *ICASSP*, 2024. 2

[44] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *ECCV*, 2018. 3

[45] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *ICCV*, 2019. 3