Egocentric Auditory Attention Localization in Conversations

Fiona Ryan[1,2]*, Hao Jiang[2], Abhinav Shukla[2], James M. Rehg[1,2], Vamsi Krishna Ithapu[2]
Georgia Institute of Technology[1], Meta Reality Labs Research[2]
{fkryan, rehg}@gatech.edu, {haojiang, ithapu}@meta.com, abhinav.shukla.research@gmail.com

Abstract

In a noisy conversation environment such as a dinner party, people often exhibit selective auditory attention, or the ability to focus on a particular speaker while tuning out others. Recognizing who somebody is listening to in a conversation is essential for developing technologies that can understand social behavior and devices that can augment human hearing by amplifying particular sound sources. The computer vision and audio research communities have made great strides towards recognizing sound sources and speakers in scenes. In this work, we take a step further by focusing on the problem of localizing auditory attention targets in egocentric video, or detecting who in a camera wearer's field of view they are listening to. To tackle the new and challenging Selective Auditory Attention Localization problem, we propose an end-to-end deep learning approach that uses egocentric video and multichannel audio to predict the heatmap of the camera wearer's auditory attention. Our approach leverages spatiotemporal audiovisual features and holistic reasoning about the scene to make predictions, and outperforms a set of baselines on a challenging multi-speaker conversation dataset. Project page: https://fkryan.github.io/saal

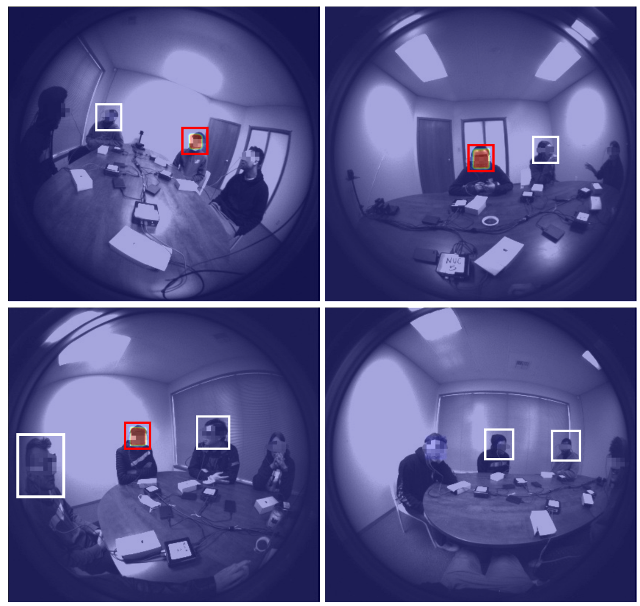

Figure 1. We address the novel task of Selective Auditory Attention Localization in multi-speaker environments: given egocentric video and multichannel audio, we predict which people, if any, the camera wearer is listening to. We show our model's predicted heatmaps, with red bounding boxes denoting ground truth auditory attention targets and white boxes denoting non-attended speakers.

1. Introduction

One of the primary goals of wearable computing devices like augmented reality (AR) glasses is to enhance human perceptual and cognitive capabilities. This includes helping people have natural conversations in settings with high noise level (e.g. coffee shops, restaurants, etc.) by selectively amplifying certain speakers while suppressing noise and the voices of background speakers. This desired effect mirrors selective auditory attention (SAA), or humans' ability to intentionally focus on certain sounds while tuning out others. People exercise SAA in everyday conversational settings; at restaurants people tune out the voices at adjacent tables, and in group social settings, such as dining at a large table or socializing at a party, people often engage in conversations with smaller subsets of people while others converse in close proximity. Being able to determine which speaker(s) a person is selectively listening to is important for developing systems that can aid communication in noisy environments and assist people with hearing loss.

While the computer vision community has made strides towards understanding conversational dynamics with large-scale datasets like Ego4D [37], AVA-ActiveSpeaker [78], VoxConverse [24], and AMI [57], the problem of modeling selective listening behavior has not yet been addressed. In fact, determining auditory attention among competing sound signals has traditionally been approached using neurophysiological sensing [11, 50]. These approaches involve a controlled listening task, where competing audio signals

*Work done during internship at Meta.

are played simultaneously and a listener selectively listens to one. Statistical models are then used to correlate brain activity and the attended sound signal. However, the sensing approach is obtrusive and currently not feasible for use in realistic conversational settings in daily life.

In this work, we approach modeling SAA from an egocentric audiovisual perspective. The egocentric setting provides a compelling lens to study social conversational dynamics, as it captures both the audio and visual stimuli present in a scene and how the camera wearer orients their view in response to them. We hypothesize that the behavior of the wearer that is implicitly captured in the egocentric video and multichannel audio can facilitate the prediction of auditory attention targets. To this end, we introduce and formulate the novel task of **Selective Auditory Attention Localization (SAAL)**, which uses egocentric video and multichannel audio to localize the target of auditory attention in egocentric video. We specifically target multi-speaker conversation scenarios where the camera wearer must selectively attend to certain speaker(s) while tuning out others. This challenging setting is representative of everyday noisy conversation environments and highlights the complexity of modeling auditory attention in naturalistic settings.

We propose a deep audiovisual video network for SAAL that leverages both appearance-based features from the egocentric video stream and spatial audio information from multichannel audio. Our key insight is to extract a spatiotemporal feature representation from each modality and use a transformer on the fused features to reason holistically about the scene. Our contributions are:

- We introduce the novel problem of Selective Auditory Attention Localization (SAAL) as an egocentric multimodal learning task.
- We propose a new architecture for SAAL. Our model extracts spatiotemporal video and audio features and uses a transformer to refine selection of an attention target by reasoning globally about the scene.
- We evaluate our approach on a challenging multi-speaker conversation dataset and demonstrate our model's superiority over intuitive baselines and approaches based on Active Speaker Localization (ASL).
- We conduct thorough experiments to give insight into our multimodal architecture design, effective multichannel audio and visual input representations, and the relationship between SAAL and ASL.

2. Related Work

Computational Models for Auditory Attention Attention is typically categorized into two types: bottom-up, where a stimulus is likely to cause attention due to inherent salient properties, and top-down, where a person intentionally focuses on a certain stimulus. Top-down, or selective auditory attention has been studied by correlating the attended sound signal with the neural response measured by EEG / MEG using statistical models [4,17–19,27,33,38,40,65,67,71,74,89], and more recently deep neural networks [36,58]. This is done in a controlled listening scenario where competing sounds, such as audio books, are played at once and a participant is instructed to listen to a particular sound. Models are used to predict which sound source is attended and/or decode the attended signal. The sources may be combined into a single channel or played in separate ears, as in the dichotic listening task paradigm [20]. Some works have used acoustic simulation methods to reflect more realistic multi-speaker environments [17, 33, 89]. However, these studies have not addressed modeling SAA in in-the-wild social conversation scenarios. Beyond using brain activity to predict auditory attention, Lu et al. develop a head movement-based approach for predicting the attended speaker in a conversation [64]. However, this work focuses on single group conversations where there is typically only one speaker.

A smaller body of work focuses on modeling bottom-up auditory attention, or the qualities of a sound signal that make it likely to be attended to. Works in this domain construct representations of audio signals and evaluate their ability to reflect human judgment of saliency [30, 49, 51, 52, 56, 85, 86], distinguish between changes in perceptual qualities of sound [48], or perform on a downstream speech recognition task [47]. In this work, we focus on top-down, or selective auditory attention.

Active Speaker Detection & Localization Our task builds upon a rich foundation of work on Active Speaker Detection (ASD) and Active Speaker Localization (ASL). Recognizing speech activity is an established problem in computer vision, with early approaches correlating video and audio streams [25], or using purely visual features [31, 39, 79] to determine if a person is speaking. The AVA-ActiveSpeaker dataset [78] has accelerated deep learning approaches for ASD by providing a large-scale dataset and a benchmark. Methods use deep learning to extract visual features from a head bounding box track and combine them with features from monoaural scene audio to classify the person as speaking and audible, speaking and inaudible, or not speaking in each frame [8–10, 26, 54, 66, 82, 95]. Some approaches additionally explicitly model the relationships between multiple people within a scene to refine classification of speakers [8, 54, 66]. Xiong et al. [93] develop an architecture to jointly perform ASD and speech enhancement.

Other methods approach the task of ASL, which seeks to localize speakers spatially within the scene rather than classifying bounding box tracks [6,15,22,34,35,76,92,94]. Several use multichannel audio to incorporate directional audio information [6, 15, 22, 34, 35, 92, 94]. Recently, Jiang et al. [46] expanded upon these approaches by developing a deep model that also leverages multichannel audio to lo-

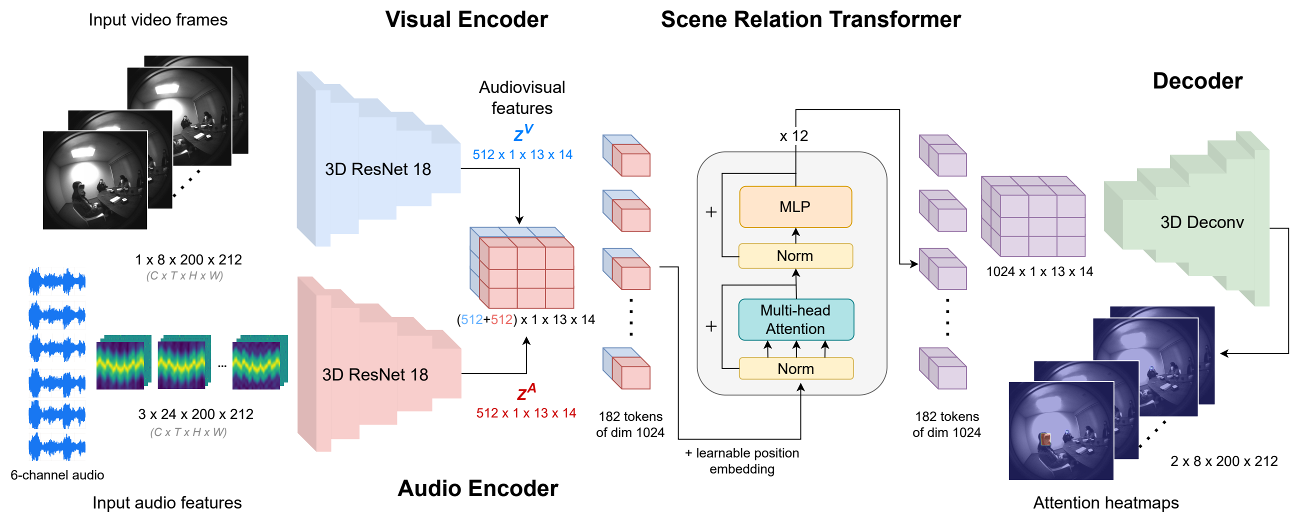

Figure 2. Our architecture for predicting auditory attention heatmaps from egocentric video and multichannel audio: the **Video Encoder** and **Audio Encoder** embed spatiotemporal features from the video and spatial audio modalities respectively. The audio and visual features are fused and passed to the **Scene Relation Transformer**, which models the relationships between different regions of the scene. Finally, the features are passed to the convolutional **Decoder**, which produces the predicted attention heatmaps for the input video frames.

calize speakers in the challenging egocentric setting, which contains motion blur and rapid changes of field of view (FOV), using the EasyCom dataset [28]. Our problem setting is different from ASD and ASL in that we seek to determine which of the speakers the camera wearer is *listening to* using egocentric audiovisual cues. Importantly, active speakers are not necessarily auditorily attended; they can be background people or nearby people engaged in a different conversation. Thus, SAAL demands a novel approach.

Modeling of Conversation Dynamics and Social Interactions SAAL is also related to work on modeling social interactions and conversational dynamics. Prior works have explored detecting social interactions in egocentric video [2, 16, 32]. The recent Ego4D dataset and benchmark suite [37] introduces the "Talking to Me" task for identifying which people are speaking to the camera wearer in egocentric video. This task is closely related to SAAL in that it models dyadic conversational behavior between the camera wearer and people in their FOV. However our task differs by modeling egocentric listening behavior, particularly in the presence of multiple speakers. Researchers have also explored modeling conversational turn taking behaviors [88] as well as predicting listener motion in dyadic conversations [70]. To our knowledge, we are the first to address the problem of modeling auditory attention using egocentric video, wearable audio, or both modalities together.

Egocentric Visual Attention Prediction Our problem also relates to egocentric visual attention prediction, which regresses a heatmap localizing the target of visual attention in egocentric video. Approaches for this task use eye tracking data as ground truth for attention targets and develop models to predict the attention location from video. Researchers have proposed several vision models for this task [5,44,59,61,83], including some that jointly model gaze and action [43,62] or gaze and IMU [84]. Importantly, visual attention and auditory attention are not the same; people will generally not always look at the person they are conversing with. Rather, SAAL can be considered as an auditory variant of egocentric visual attention prediction, and SAAL additionally demands the use of audio cues.

Audiovisual Representation Learning Our work relates to a larger body of research on learning effective audiovisual feature representations [3, 7, 12, 45, 55, 69, 72, 73, 75] and localizing sound sources in video [1, 13, 23, 41, 42, 63, 68, 77, 80, 81]. These methods learn semantic correspondences between visual and audio features, typically using contrastive training paradigms. In contrast to these approaches, we leverage multichannel audio and employ a fusion mechanism that enforces spatial alignment between the visual and spatial audio domains, allowing us to reason globally about the audiovisual content in different regions of the scene.

3. Method

3.1. Problem Definition

At each time t, we predict an auditory attention heatmap $\mathcal{H}_t$, given the raw audiovisual input $\{I_n, a_n | n = 0...t\}$, where I_n is an egocentric video frame at time n and a_n is a multi-channel audio segment aligned with the video frame I_n. Each pixel $\mathcal{H}_t(i,j)$ has a value in (0,1) and indicates the probability of the camera wearer's auditory attention. Video I captures the movement of both the camera wearer and the other people in the scene, as well as information about their head and body poses, facial expression, mouth movement, and overall scene layout. Multichannel audio a, from a microphone array, captures people's speech as well as other spatialized sound sources. We use video from a 180-degree fish-eye camera with resolution 200×212 and audio from a 6-channel microphone array. In this paper, we demonstrate how to use $\{I, a\}$ to infer $\mathcal{H}$, where we construct $\mathcal{H}_{\text{label}}$ by setting all pixels within the head bounding box of an attended speaker in a frame as 1.

3.2. Model

SAAL is a complex task that involves understanding both the camera wearer's behavior and the activity of the people in the scene. We design a deep multimodal model for this task to have 4 key abilities: **(1)** It extracts relevant visual features such as lip movement, the relative positions and orientations of heads, and social body language like when a person looks at the camera wearer. **(2)** It leverages multichannel audio to construct a spatial representation of voice activity occurring around the camera wearer. **(3)** It encodes temporal information like how the wearer moves their head in response to stimuli as well as the movement of people in the scene and pattern of voice activity. **(4)** It reasons globally about the audiovisual content of scene as a whole to select an attention target; while an active speaker detector identifies all speakers in a scene, determining attention demands holistic reasoning about different potential attention targets in a scene to select the most likely one(s).

To this end, our method consists of 4 main components as shown in Figure 2: the *Visual Encoder* and *Audio Encoder* to extract spatiotemporal features from the video and multichannel audio inputs respectively, the *Scene Relation Transformer* to refine attention localization by modeling the relationships between audiovisual features in different spatial regions of the scene, and a convolutional *Decoder* to produce an auditory attention heatmap for each input frame.

Visual Encoder The Visual Encoder $\mathcal{V}$ takes the visual input X^V as a stack of 200×212 frames over a time interval T and extracts appearance-based features about the scene. We construct X^V by cropping the people's heads in the raw grayscale video frames and filling the background with black. This representation leverages 1) our focus on conversational attention, and 2) the effectiveness of face detection methods. The input is a 3D video tensor of head tubes that represents not only how the people in the scene move, how they are laid out in space, where they look, their facial expression, mouth movement, and who they talk to, but also the head movement of the wearer over time. In our experiments, we compare this representation to using the raw image as well as a binary map of the bounding boxes. We find that this representation improves performance by removing background distractions while retaining appearance-based facial features. Additionally, this representation makes the model more generalizable to different scenes. We use a 3D ResNet-18 network [87] as our backbone, which uses 3D convolutions to capture both spatial and temporal features. The result of $\mathcal{V}(X^V)$ is a spatiotemporal visual feature map $\mathcal{Z}^V$ of dimension $512 \times 1 \times 13 \times 14$, where 512 is the channel dimension, 1 is the temporal dimension, and 13 and 14 correspond to the height and width. We use 8 frames as input with temporal stride 3, so each input X^V spans a 24 frame window, which is 0.8 seconds in 30fps video.

Audio Encoder The Audio Encoder $\mathcal{A}$ extracts spatial audio features from the multichannel audio in the input clip to encode information about the conversational speech patterns and spatially localize speakers. We process the audio into 2000 sample $\times$ 6 channel frames corresponding to each frame in the video input clip. We construct input features X^A by calculating the complex spectogram with 201 frequency bins, window length 20, and hop length 10 and stacking these vertically for the 6 channels. We found this performs slightly better than concatenating along the channel dimension. We additionally construct a feature representation of the cross correlation between all pairs of channels as used in Jiang et al. [46] and stack the real and complex parts of the spectogram and the channel correlation features along the channel dimension. We then resize the audio features to 200×212 spatially in order to extract feature maps that align with the visual domain. Due to the importance of capturing dense temporal audio features to predict auditory attention, such as the pattern of speech, we pass all 24 audio frames in the clip to $\mathcal{A}$, which uses another 3D ResNet-18 backbone. The resultant feature map $\mathcal{Z}^A$ is of size $512 \times 3 \times 13 \times 14$. We average pool the temporal dimension to 1 to match the dimension of $\mathcal{Z}^V$.

Scene Relation Transformer From $\mathcal{A}$ and $\mathcal{V}$, we extract visual and spatial audio features maps $\mathcal{Z}^V$ and $\mathcal{Z}^A$, which we concatenate along the channel dimension to form $\mathcal{Z}^{AV} \in 1024 \times 1 \times 13 \times 14$. We choose to concatenate the features along the channel dimension because our approach importantly leverages multichannel audio, which can extract a spatial representation of voice activity in the scene so that the visual and audio domains can be spatially aligned.

We use a transformer $\mathcal{T}$ to further refine this fused multimodal feature map by modeling correlations between the features in different regions of the scene to holistically select a likely attention target. As in the transformer literature [14, 29], we treat each spatial location $Z^{AV}_{1,h,w}$ as a spatiotemporal patch of embedded dimension $D = 512$, and flatten the feature map into a sequence of length $L = 1 \times 13 \times 14 = 182$. A learnable position embedding $E \in \mathbb{R}^{L \times D}$ is added to preserve absolute and relative positional information, which is particularly important for the egocentric setting where the camera wearer orients their view of the scene in response to stimuli. The transformer consists of 12 transformer blocks [90] where each computes self-attention on the audiovisual feature map as *Softmax*$(QK^T/\sqrt{D})V$,

Split	Total frames	Frames with 1+ attention target	Avg. heads per frame	Avg. speakers per frame
Train	1,416,262	881,931	3.05	1.43
Test	571,027	367,536	3.06	1.45

Table 1. Dataset statistics: our evaluation dataset of multi-speaker conversations is representative of challenging conversation environments where several people and speakers are present. Note that we use $\frac{1}{3}$ of these frames for training and evaluation with temporal stride 3.

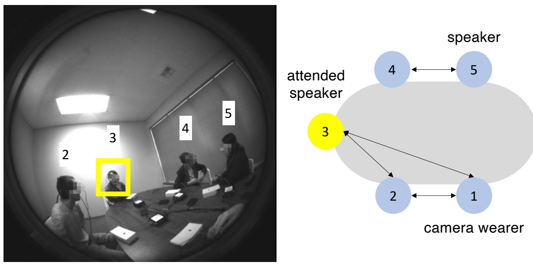

Figure 3. In our multi-speaker evaluation dataset, participants are divided into 2 subgroups that converse simultaneously. A person is labeled as being an auditory attention target if they are both speaking and within the camera wearer's conversation group.

where Q, K, and V are learned linear projections on the multimodal spatiotemporal patches. In this way, the model learns correlations between audiovisual features occurring in different regions of the scene and can suppress or enhance values accordingly to refine attention prediction.

Decoder A decoder consisting of 4 3D tranpose convolutions and 1 final convolutional classifier layer upsamples the resulting feature map $\mathcal{T}(\mathcal{Z}^{AV})$ to produce the attention class heatmap $\mathcal{H}_{\text{pred}}$ of size $2 \times T \times H \times W$. The 2 channels reflect the not attended and attended classes.

3.3. Implementation & Model Training

We train end-to-end with pixel-wise cross entropy loss. We initialize $\mathcal{V}$ and $\mathcal{A}$ with pretrained weights on Kinetics 400 [21]. The transformer consists of 12 transformer blocks using multi-head self-attention with 8 heads [90]. We use the Adam optimizer [53] and learning rate 1e-4. The training procedure converges quickly in less than 10 epochs.

4. Experimental Results

4.1. Evaluation Dataset & Criteria

To our knowledge there is no existing dataset with associated labels suited to our task. While egocentric AV datasets like Ego4D [37] and EasyCom [28] capture egocentric conversations, they have few situations with multiple people speaking at the same time, with the result that SAAL reduces to ASL. To evaluate our approach, we collected a dataset of complex multi-speaker conversations, where groups converse in close proximity. The dataset consists of 50 participants conversing in groups of 5, where each person wears a headset with a camera and microphone array. In total, there are ∼20 hours of egocentric video.

In each 10-minute recording, a group of 5 people sits at a table in a meeting room and is partitioned into two conversational groups. Participants are instructed to only listen to and engage in conversations with their own group. In this way, we design a naturalistic social version of traditional dichotic listening tasks, where participants intentionally focus on certain speakers while tuning out others. By splitting the group into smaller groups, we simulate conversational layouts that may occur at a large dinner table, coffee shop, restaurant, or party, where groups of people converse in close proximity with others. As illustrated in Figure 3, we construct ground truth auditory attention labels by determining that the camera wearer is listening to a given person if they are both speaking and in the wearer's conversation group. The dataset was collected under an approved IRB.

Each participant wears a headset that includes an Intel SLAM camera and a 6-microphone array. We record the fish-eye grayscale video and the audio from the 6 microphones simultaneously for all the wearers in each session. We generate speaker activity and head bounding box annotations automatically by using person tracking and wearer speech detection algorithms, and leveraging synchronization between each person's egocentric audio and video in each session. We obtain active speaker labels for each person by using their own headset's microphone array to determine if they are speaking, using a deep model adapted from [46] trained on synthetic data for the headset's microphone array configuration. We then use a 3D tracking-by-detection method that uses camera pose from the SLAM camera to annotate head bounding boxes and identities for each person in each egocentric FOV. From these person tracking results and speaker activity labels for each person, we have a set of visible head bounding boxes for each egocentric frame along with a label for each box indicating if that person is speaking. We then use the identity tracking and known conversation group assignments to label if each person is auditorily attended to by the camera wearer.

Due to the wide 180-degree FOV of the camera, the auditorily attended person is almost always in the FOV, so we constrain our modeling approach to this scenario. Additionally, by design of the dataset, there are typically several people and often multiple speakers within FOV, as shown in Table 1. Along with the inherent challenges of motion blur and rapid head movements in the egocentric modality, this makes our dataset a rich means for studying auditory attention in complex and naturalistic conversation environments.

We evaluate the result of auditory attention localization using mean average precision (mAP) on the task of classifying each person in the FOV as being a target of auditory attention or not. From the output heatmap $\mathcal{H}_{\text{pred}}$, we calculate the auditory attention score for each person by taking the maximum value of $Softmax(\mathcal{H}_{\text{pred}})$ in channel 1 (the attended class) in the region of their head bounding box. We split our data and use 70% data for training and 30% for testing, with no overlapping participants between splits.

4.2. Competing methods

Because SAAL is a new problem, there are no previous methods that can be used directly to solve this problem. We therefore compare our approach to several naive baseline methods as well as a set of baselines that adapt Jiang et al.'s multichannel audiovisual ASL architecture (MAVASL) [46] for our task. The methods we compare are the following:

(1) Our proposed method with different audio and visual input representations: In our experiments, we label variations of our method as "Ours - [visual] & [audio]", where [visual] is the visual input representation type and [audio] is the audio input representation type. The visual input representations we consider are **Image** (the raw image), **Bbox** (a binary map of the bounding boxes), and **Heads** (the cropped heads from the raw image on a black background). The audio representations are **Channel Corr** (the channel correlation features as described in Section 3.2), **Channel Corr + Spectogram** (the channel correlation features concatenated with the real and complex parts of the multichannel spectogram), and **ASL** (active speaker localization output maps from MAVASL). We report two different training strategies for using MAVASL to generate ASL maps for our dataset: $\textbf{ASL}_{\textbf{synthetic}}$, which trains the ASL model on synthetic data created from VTCK [91] and EasyCom, and $\textbf{ASL}_{\textbf{real}}$, which directly trains the ASL model on our dataset using the active speaker labels. In this way, we compare using audio features constructed from the raw multichannel audio with using ASL maps generated from an existing model, both with and without the advantage of tuning to our dataset.

(2) Naive baselines using ground truth ASL labels and heuristic strategies: We first report the SAAL mAP from simply selecting all active speakers as being attended using the ground truth ASL labels (Perfect ASL). We also report 2 baselines that use the center prior, which is the intuition that the attended person is likely near the center of the FOV because people often face the speaker they are paying attention to. Similar center prior baselines are used in the egocentric visual attention literature. We construct 2 center prior baselines: CP–I finds the head bounding box nearest to the center of FOV and selects it as the attention target if they are speaking, and otherwise predicts that there is no attention in the frame. CP–II selects the speaking person closest to the center of the FOV as the target of attention. We also construct 2 baselines that select the "closest" speaker based on the size of their bounding box. We can reason that the person with the largest head bounding box area is likely the closest to the camera wearer, and if speaking, likely appears the loudest. We consider 2 baselines based on this strategy: CS–I selects the "closest" person only if they are speaking, and CS–II selects the "closest" speaker. We note that because these methods depend on the dataset's ground truth ASL labels, they are not realistic results, and are not a fair comparison to our method. However, we report them to demonstrate that Perfect ASL and heuristic strategies perform poorly on our evaluation dataset. This clearly implies that SAAL is a complicated task that demands a novel approach, and that our evaluation dataset is rich enough to demonstrate the complexity of this task.

(4) Networks adapted from MAVASL: The first adaption (MAVASL-I) is to directly use MAVASL to localize active speakers and use this as the auditory attention heatmap. The second adaptation (MAVASL-II) directly trains the network for the task of SAAL instead of ASL. The third adaptation (MAVASL-III) is similar to (MAVASL-II) but uses a multi-frame video input (3-frame) to the AV network to further incorporate temporal information. MAVASL consists of an Audio Network, which predicts a coarse spatial voice activity map from multichannel audio, and an AV Network, which uses this voice activity map and the image to predict a final ASL map. For MAVASL-II and III, we achieve best results by supervising the audio network in MAVASL using the ASL labels from our dataset and the AV network with the auditory attention labels. We use **Heads** as the visual input representation. Because MAVASL is a frame-based network, we additionally apply temporal smoothing with a window size of 10 frames as in the original method.

4.3. Comparison results

As shown in Table 2, our proposed method outperforms the baselines on mAP, including baselines that leverage ground truth active speaker labels. We observe that the Perfect ASL and naive baselines perform poorly on our dataset, illustrating that reducing the problem of SAAL to ASL is not viable in complex and challenging conversation environments with multiple speakers. The highest mAP achieved from adapting Jiang et al.'s ASL architecture is MAVASL-II, which obtains 75.20% mAP. However, our best approach achieves a gain of 7.74% mAP over this baseline, demonstrating that the task of SAAL demands a novel architecture design. We qualitatively compare our results against these baselines in Figure 4.

From our experiments investigating different combinations of audio and visual input representations, we observe that the combination of **Heads** and **Channel Corr + Spec-**

Method	mAP (%)
Perfect ASL*	47.99
CP–I*	63.55
CP–II*	51.48
CS–I*	53.86
CS–II*	49.47
MAVASL–I	59.11
MAVASL–II	75.20
MAVASL–III	72.90
Ours–Bbox & $\text{ASL}_{\text{synthetic}}$	75.93
Ours–Bbox & ASL_{real}	74.97
Ours–Bbox & Channel Corr	80.41
Ours–Bbox & Channel Corr + Spectogram	80.31
Ours–Image & $\text{ASL}_{\text{synthetic}}$	72.20
Ours–Image & ASL_{real}	70.04
Ours–Image & Channel Corr	76.52
Ours–Image & Channel Corr + Spectogram	76.95
Ours–Heads & $\text{ASL}_{\text{synthetic}}$	76.72
Ours–Heads & ASL_{real}	77.11
Ours–Heads & Channel Corr	82.35
Ours–Heads & Channel Corr + Spectogram	**82.94**

Table 2. Comparison results on the multi-speaker conversation dataset. (*) denotes methods that use ground truth ASL, which is not given to our model.

Method	mAP (%)
Visual only	56.81
Visual only + Transformer	57.70
Audio only	76.30
Audio only + Transformer	75.84
Audiovisual	80.10
$\text{Audiovisual}_{\text{1-channel}}$ + Transformer	71.47
$\text{Audiovisual}_{\text{2-channel}}$ + Transformer	76.73
$\text{Audiovisual}_{\text{4-channel}}$ + Transformer	80.99
Audiovisual + Transformer	**82.94**

Table 3. Model ablation: All studies use Heads as the visual input and Channel Corr + Spectogram as the audio input.

Method	mAP (%)
Audiovisual + Transformer (no spatial alignment)	80.57
Audiovisual + Transformer (spatial alignment)	**82.94**

Table 4. Spatial modality alignment in the transformer.

togram produces the best results. In fact, we observe that **Bbox**, the binary bounding box location map, performs better as an input representation than the raw image. We can interpret this result as showing the importance of relative position and sizes of heads to the problem of SAAL. **Channel Corr** and **Channel Corr + Spectogram**, which contain information from the raw multichannel audio signal, generally perform better as the spatial audio input representation than the pre-computed ASL maps, both real and synthetically trained. We hypothesize that this is because the role of audio features in determining auditory attention is greater than just localizing active speakers; SAAL additionally involves correlating different speech signals with the camera wearer's egocentric movement, using the wearer's own voice activity as an indicator of conversational turn taking behavior (e.g. whether they are speaking or listening), and reasoning about the relative volume of different speech signals. Notably, the pre-computed ASL maps have the advantage of being spatially aligned with the visual modality, which the spectograms and channel correlation features do not have. The performance gains from using **Channel Corr** and **Channel Corr + Spectogram** indicate that our Audio Encoder is able to learn this alignment and construct a spatial representation of audio activity from these features.

4.4. Analysis

Model ablation We further investigate the contribution of each component of our best model to its overall performance in Table 3. We observe that the visual modality alone is much weaker than the spatial audio modality, and that clearly audio plays an important role. We also ablate the number of audio channels used, observing that removing the spatial component of the audio ($\text{Audiovisual}_{\text{1-channel}}$ + Transformer) significantly reduces performance of our overall model. We see that spatial audio is an important cue for this task, but using slightly less channels can still achieve effective performance. Adding the transformer to the audiovisual features results in a gain of 2.84 mAP, demonstrating that the transformer improves performance by refining the embedded audiovisual feature maps through modeling relationships across spatial regions of the scene.

Results on Unseen Environment We additionally test our model's ability to generalize to unseen environments by evaluating our best model on a subset of data collected in a room with different acoustic and visual properties. The data follows the same structure as the main evaluation dataset, with 5 participants split into 2 conversation subgroups, and is ∼49 minutes. Further details are provided in the supplement. On this unseen environment, our best model achieves 80.43% mAP, demonstrating scene generalizability.

Spatial Modality Alignment in Self-Attention A key feature of our approach is leveraging multichannel audio to learn a spatial feature representation of voice activity in the scene, which we align with the visual representation by concatenating along the channel dimension. In our transformer design, therefore, each spatiotemporal patch of $\mathcal{Z}^{AV}$ contains the visual and audio features for that region. We validate this choice by designing an additional experiment where the modalities are not fused before the transformer, like in VisualBERT [60]. Rather than concatenating $\mathcal{Z}^{A}$ and $\mathcal{Z}^{V}$ before flattening them into a sequence of tokens to pass to $\mathcal{T}$, we separately create a sequence of 182 tokens each

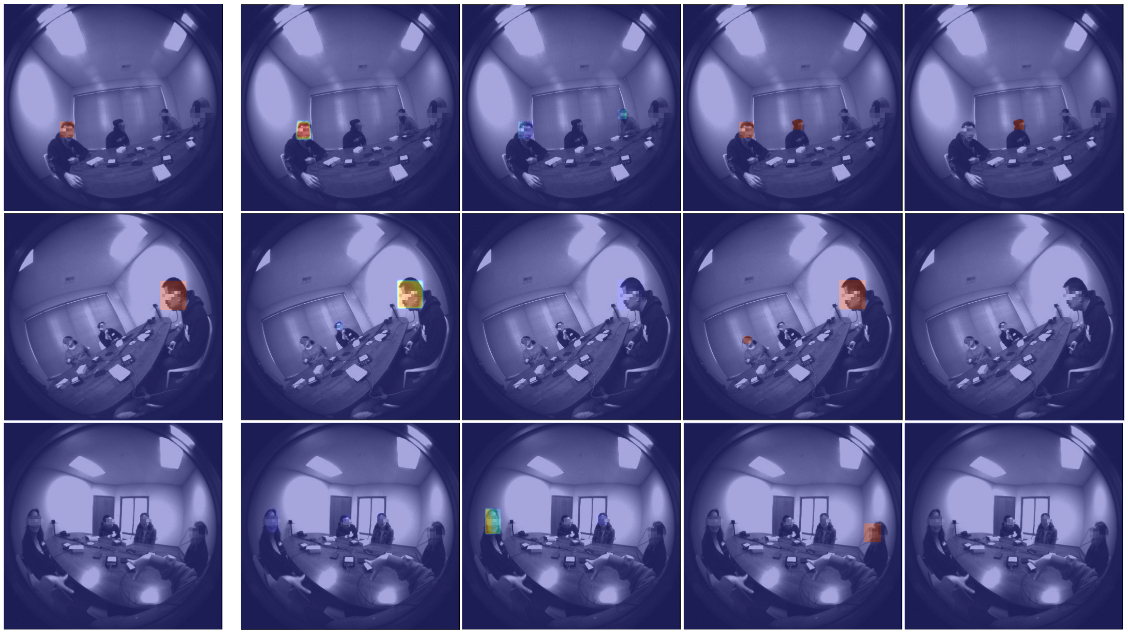

Figure 4. We compare our output heatmaps to competing methods on challenging cases with multiple speakers, the attended speaker being far from the center, or no auditory attention. Our method performs well in these scenarios and more consistently identifies who is and isn't attended than MAVASL-II. Perfect ASL is prone to false positives by selecting all speakers and CPI is prone to false negatives by only considering the person closest to the center, demonstrating that ASL is not enough to solve SAAL. (**Best viewed in color**).

Figure 5. Decoupling SAAL & ASL: Without the responsibility for learning ASL, our model learns to segment the people within the camera wearer's conversation group.

from $\mathcal{Z}^V$ and $\mathcal{Z}^A$ and pass these 364 total tokens of dimension 512 to $\mathcal{T}$. We concatenate the modalities afterwards. In this case, the self-attention mechanism is left to discover alignment between the modalities itself by comparing all tokens across both modalities. We observe a decrease in performance, indicating that our choice to explicitly align the visual and audio modalities spatially contributes to our model's ability to use both modalities together successfully. This result suggests utility in incorporating inductive bias about the spatial alignment of visual and multichannel audio modalities into token design for transformer architectures.

Decoupling SAAL & ASL By nature of our dataset design, we train our model to predict auditory attention as the intersection of active speakers and people in the camera wearer's conversation group. To investigate the relationship between SAAL and ASL, we train a model by modulating our model's heatmap for the attended class with the ground truth ASL labels before calculating loss and performing backpropagation. The loss is calculated as $\mathcal{L}_{CE}(\mathcal{H}_{\text{pred}} * \text{ASL}_{\text{label}}, \mathcal{H}_{\text{label}})$. In this way we eliminate the responsibility for ASL from our model. The model achieves 92.67% mAP on SAAL, however, this is a purely theoretical upper bound that relies on perfect ASL. We visualize $\mathcal{H}_{\text{pred}}$ for our model before it is modulated by $\text{ASL}_{\text{label}}$ in Figure 5. We observe that our model learns to more generally segment the camera wearer's conversation group regardless of speaking activity, and likely relies on $\text{ASL}_{\text{label}}$ to constrain the heatmaps to the exact bounding box region.

5. Conclusion

We introduce the novel problem of Selective Auditory Attention Localization (SAAL) and demonstrate that a deep model is able to perform this task on a challenging multi-speaker conversation dataset. We propose a multimodal modeling approach that uses egocentric video and multichannel audio to predict a person's target of auditory attention, combining spatial and temporal features from both modalities and using a transformer to reason holistically about the scene. We believe our work is an important step towards selective sound source enhancement and AR devices that can help people converse in noisy environments.

Acknowledgments: We thank Jacob Donley and Christi Miller for facilitating data collection, and Calvin Murdock and Ishwarya Ananthabhotla for valuable discussions.

References

[1] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *European Conference on Computer Vision*, pages 208–224. Springer, 2020. 3

[2] Maedeh Aghaei, Mariella Dimiccoli, and Petia Radeva. With whom do i interact? detecting social interactions in egocentric photo-streams. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 2959–2964. IEEE, 2016. 3

[3] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34:24206–24221, 2021. 3

[4] Sahar Akram, Alessandro Presacco, Jonathan Z Simon, Shihab A Shamma, and Behtash Babadi. Robust decoding of selective auditory attention from meg in a competing-speaker environment via state-space modeling. *NeuroImage*, 124:906–917, 2016. 2

[5] Mohammad Al-Naser, Shoaib Ahmed Siddiqui, Hiroki Ohashi, Sheraz Ahmed, Nakamura Katsuyki, Sato Takuto, and Andreas Dengel. Ogaze: Gaze prediction in egocentric videos for attentional object selection. In *2019 Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8. IEEE, 2019. 3

[6] Xavier Alameda-Pineda, Vasil Khalidov, Radu Horaud, and Florence Forbes. Finding audio-visual events in informal social gatherings. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 247–254, 2011. 2

[7] Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. *Advances in Neural Information Processing Systems*, 33:25–37, 2020. 3

[8] Juan Leon Alcazar, Fabian Caba, Long Mai, Federico Perazzi, Joon-Young Lee, Pablo Arbelaez, and Bernard Ghanem. Active speakers in context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[9] Juan León Alcázar, Fabian Caba, Ali K Thabet, and Bernard Ghanem. Maas: Multi-modal assignation for active speaker detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 265–274, 2021. 2

[10] Juan Leon Alcazar, Moritz Cordes, Chen Zhao, and Bernard Ghanem. End-to-end active speaker detection. *arXiv preprint arXiv:2203.14250*, 2022. 2

[11] Emina Alickovic, Thomas Lunner, Fredrik Gustafsson, and Lennart Ljung. A tutorial on auditory attention identification methods. *Frontiers in neuroscience*, page 153, 2019. 1

[12] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. *Advances in Neural Information Processing Systems*, 33:9758–9770, 2020. 3

[13] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 3

[14] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6836–6846, 2021. 4

[15] Yutong Ban, Xavier Alameda-Pineda, Laurent Girin, and Radu Horaud. Variational bayesian inference for audio-visual tracking of multiple speakers. *IEEE transactions on pattern analysis and machine intelligence*, 43(5):1761–1776, 2019. 2

[16] Sophia Bano, Tamas Suveges, Jianguo Zhang, and Stephen J Mckenna. Multimodal egocentric analysis of focused interactions. *IEEE Access*, 6:37493–37505, 2018. 3

[17] Adam Bednar and Edmund C Lalor. Where is the cocktail party? decoding locations of attended and unattended moving sound sources using eeg. *NeuroImage*, 205:116283, 2020. 2

[18] Wouter Biesmans, Neetha Das, Tom Francart, and Alexander Bertrand. Auditory-inspired speech envelope extraction methods for improved eeg-based auditory attention detection in a cocktail party scenario. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(5):402–412, 2016. 2

[19] Martin G Bleichner, Bojana Mirkovic, and Stefan Debener. Identifying auditory attention with ear-eeg: ceegrid versus high-density cap-eeg comparison. *Journal of neural engineering*, 13(6):066004, 2016. 2

[20] Donald Eric Broadbent. Successive responses to simultaneous stimuli. *Quarterly Journal of Experimental Psychology*, 8(4):145–152, 1956. 2

[21] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 5

[22] Punarjay Chakravarty, Sayeh Mirzaei, Tinne Tuytelaars, and Hugo Van hamme. Who's speaking? audio-supervised classification of active speakers in video. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 87–90, 2015. 2

[23] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16867–16876, 2021. 3

[24] Joon Son Chung, Jaesung Huh, Arsha Nagrani, Triantafyllos Afouras, and Andrew Zisserman. Spot the conversation: speaker diarisation in the wild. *arXiv preprint arXiv:2007.01216*, 2020. 1

[25] Ross Cutler and Larry Davis. Look who's talking: Speaker detection using video and audio correlation. In *2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No. 00TH8532)*, volume 3, pages 1589–1592. IEEE, 2000. 2

[26] Gourav Datta, Tyler Etchart, Vivek Yadav, Varsha Hedau, Pradeep Natarajan, and Shih-Fu Chang. Asd-transformer: Efficient active speaker detection using self and multimodal transformers. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4568–4572. IEEE, 2022. 2

[27] Nai Ding and Jonathan Z Simon. Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences*, 109(29):11854–11859, 2012. 2

[28] Jacob Donley, Vladimir Tourbabin, Jung-Suk Lee, Mark Broyles, Hao Jiang, Jie Shen, Maja Pantic, Vamsi Krishna Ithapu, and Ravish Mehra. Easycom: An augmented reality dataset to support algorithms for easy communication in noisy environments. *arXiv preprint arXiv:2107.04174*, 2021. 3, 5

[29] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[30] Varinthira Duangudom and David V Anderson. Using auditory saliency to understand complex auditory scenes. In *2007 15th European Signal Processing Conference*, pages 1206–1210. IEEE, 2007. 2

[31] Mark Everingham, Josef Sivic, and Andrew Zisserman. Taking the bite out of automated naming of characters in tv video. *Image and Vision Computing*, 27(5):545–559, 2009. 2

[32] Aliricza Fathi, Jessica K Hodgins, and James M Rehg. Social interactions: A first-person perspective. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1226–1233. IEEE, 2012. 3

[33] Søren Asp Fuglsang, Torsten Dau, and Jens Hjortkjær. Noise-robust cortical tracking of attended speech in real-world acoustic scenes. *Neuroimage*, 156:435–444, 2017. 2

[34] Daniel Gatica-Perez, Guillaume Lathoud, Jean-Marc Odobez, and Iain McCowan. Audiovisual probabilistic tracking of multiple speakers in meetings. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(2):601–616, 2007. 2

[35] Israel D Gebru, Xavier Alameda-Pineda, Radu Horaud, and Florence Forbes. Audio-visual speaker localization via weighted clustering. In *2014 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2014. 2

[36] Masoud Geravanchizadeh and Hossein Roushan. Dynamic selective auditory attention detection using rnn and reinforcement learning. *Scientific Reports*, 11(1):1–11, 2021. 2

[37] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Q. Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh K. Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Z. Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Christian Fuegen, Abrham Gebreselasie, Cristina González, James M. Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Yu Heng Khoo, Jáchym Kolár, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran K. Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Yunyi Zhu, Pablo Arbeláez, David J. Crandall, Dima Damen, Giovanni Maria Farinella, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard A. Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18973–18990, 2022. 1, 3, 5

[38] Marzieh Haghighi, Mohammad Moghadamfalahi, Murat Akcakaya, and Deniz Erdogmus. Eeg-assisted modulation of sound sources in the auditory scene. *Biomedical signal processing and control*, 39:263–270, 2018. 2

[39] Fasih Haider and Samer Al Moubayed. Towards speaker detection using lips movements for humanmachine multiparty dialogue. In *The XXVth Swedish Phonetics Conference (FONETIK)*, pages 117–120. Citeseer, 2012. 2

[40] Cort Horton, Ramesh Srinivasan, and Michael D'Zmura. Envelope responses in single-trial eeg indicate attended speaker in a 'cocktail party'. *Journal of neural engineering*, 11(4):046015, 2014. 2

[41] Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9248–9257, 2019. 3

[42] Di Hu, Yake Wei, Rui Qian, Weiyao Lin, Ruihua Song, and Ji-Rong Wen. Class-aware sounding objects localization via audiovisual correspondence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3

[43] Yifei Huang, Minjie Cai, Zhenqiang Li, Feng Lu, and Yoichi Sato. Mutual context network for jointly estimating egocentric gaze and action. *IEEE Transactions on Image Processing*, 29:7795–7806, 2020. 3

[44] Yifei Huang, Minjie Cai, Zhenqiang Li, and Yoichi Sato. Predicting gaze in egocentric video by learning task-dependent attention transition. In *Proceedings of the European conference on computer vision (ECCV)*, pages 754–769, 2018. 3

[45] Aren Jansen, Daniel P. W. Ellis, Shawn Hershey, R. Channing Moore, Manoj Plakal, Ashok Popat, and Rif A. Saurous. Coincidence, categorization, and consolidation: Learning to recognize sounds with minimal supervision. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 121–125, 2020. 3

[46] Hao Jiang, Calvin Murdock, and Vamsi Krishna Ithapu. Egocentric deep multi-channel audio-visual active speaker lo-

calization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10544–10552, 2022. 2, 4, 5, 6

[47] Ozlem Kalinli and Shrikanth S Narayanan. A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech. In *INTERSPEECH*, pages 1941–1944, 2007. 2

[48] Emine Merve Kaya and Mounya Elhilali. A temporal saliency map for modeling auditory attention. In *2012 46th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE, 2012. 2

[49] Emine Merve Kaya and Mounya Elhilali. Investigating bottom-up auditory attention. *Frontiers in human neuroscience*, 8:327, 2014. 2

[50] Emine Merve Kaya and Mounya Elhilali. Modelling auditory attention. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1714):20160101, 2017. 1

[51] Christoph Kayser, Christopher I Petkov, Michael Lippert, and Nikos K Logothetis. Mechanisms for allocating auditory attention: an auditory saliency map. *Current biology*, 15(21):1943–1947, 2005. 2

[52] Kyungtae Kim, Kai-Hsiang Lin, Dirk B Walther, Mark A Hasegawa-Johnson, and Tomas S Huang. Automatic detection of auditory salience with optimized linear filters derived from human annotation. *Pattern Recognition Letters*, 38:78–85, 2014. 2

[53] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[54] Okan Köpüklü, Maja Taseska, and Gerhard Rigoll. How to design a three-stage architecture for audio-visual active speaker detection in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1193–1203, 2021. 2

[55] 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

[56] Sandeep Reddy Kothinti, Nicholas Huang, and Mounya Elhilali. Auditory salience using natural scenes: An online study. *The Journal of the Acoustical Society of America*, 150(4):2952–2966, 2021. 2

[57] Wessel Kraaij, Thomas Hain, Mike Lincoln, and Wilfried Post. The ami meeting corpus. 2005. 1

[58] Ivine Kuruvila, Jan Muncke, Eghart Fischer, and Ulrich Hoppe. Extracting the auditory attention in a dual-speaker scenario from eeg using a joint cnn-lstm model. *Frontiers in Physiology*, 12, 2021. 2

[59] Bolin Lai, Miao Liu, Fiona Ryan, and James Rehg. In the eye of transformer: Global-local correlation for egocentric gaze estimation. *arXiv preprint arXiv:2208.04464*, 2022. 3

[60] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 7

[61] Yin Li, Alireza Fathi, and James M Rehg. Learning to predict gaze in egocentric video. In *Proceedings of the IEEE international conference on computer vision*, pages 3216–3223, 2013. 3

[62] Yin Li, Miao Liu, and James Rehg. In the eye of the beholder: Gaze and actions in first person video. *IEEE transactions on pattern analysis and machine intelligence*, 2021. 3

[63] Xian Liu, Rui Qian, Hang Zhou, Di Hu, Weiyao Lin, Ziwei Liu, Bolei Zhou, and Xiaowei Zhou. Visual sound localization in the wild by cross-modal interference erasing. *arXiv preprint arXiv:2202.06406*, 2, 2022. 3

[64] Hao Lu and W Owen Brimijoin. Sound source selection based on head movements in natural group conversation. *Trends in Hearing*, 26:23312165221097789, 2022. 2

[65] Nima Mesgarani and Edward F Chang. Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, 485(7397):233–236, 2012. 2

[66] Kyle Min, Sourya Roy, Subarna Tripathi, Tanaya Guha, and Somdeb Majumdar. Learning long-term spatial-temporal graphs for active speaker detection. *arXiv preprint arXiv:2207.07783*, 2022. 2

[67] Bojana Mirkovic, Stefan Debener, Manuela Jaeger, and Maarten De Vos. Decoding the attended speech stream with multi-channel eeg: implications for online, daily-life applications. *Journal of neural engineering*, 12(4):046007, 2015. 2

[68] Shentong Mo and Pedro Morgado. Localizing visual sounds the easy way. *arXiv preprint arXiv:2203.09324*, 2022. 3

[69] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, 34:14200–14213, 2021. 3

[70] Evonne Ng, Hanbyul Joo, Liwen Hu, Hao Li, Trevor Darrell, Angjoo Kanazawa, and Shiry Ginosar. Learning to listen: Modeling non-deterministic dyadic facial motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20395–20405, 2022. 3

[71] James A O'Sullivan, Alan J Power, Nima Mesgarani, Siddharth Rajaram, John J Foxe, Barbara G Shinn-Cunningham, Malcolm Slaney, Shihab A Shamma, and Edmund C Lalor. Attentional selection in a cocktail party environment can be decoded from single-trial eeg. *Cerebral cortex*, 25(7):1697–1706, 2015. 2

[72] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018. 3

[73] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *European conference on computer vision*, pages 801–816. Springer, 2016. 3

[74] James O'Sullivan, Zhuo Chen, Jose Herrero, Guy M McKhann, Sameer A Sheth, Ashesh D Mehta, and Nima Mesgarani. Neural decoding of attentional selection in multi-speaker environments without access to clean sources. *Journal of neural engineering*, 14(5):056001, 2017. 2

[75] Mandela Patrick, Yuki M Asano, Polina Kuznetsova, Ruth Fong, Joao F Henriques, Geoffrey Zweig, and Andrea

Vedaldi. Multi-modal self-supervision from generalized data transformations. *arXiv preprint arXiv:2003.04298*, 2020. 3
[76] Jie Pu, Yannis Panagakis, and Maja Pantic. Active speaker detection and localization in videos using low-rank and kernelized sparsity. *IEEE Signal Processing Letters*, 27:865–869, 2020. 2
[77] 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*, pages 292–308. Springer, 2020. 3
[78] Joseph Roth, Sourish Chaudhuri, Ondrej Klejch, Radhika Marvin, Andrew Gallagher, Liat Kaver, Sharadh Ramaswamy, Arkadiusz Stopczynski, Cordelia Schmid, Zhonghua Xi, et al. Ava active speaker: An audio-visual dataset for active speaker detection. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4492–4496. IEEE, 2020. 1, 2
[79] Kate Saenko, Karen Livescu, Michael Siracusa, Kevin Wilson, James Glass, and Trevor Darrell. Visual speech recognition with loosely synchronized feature streams. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pages 1424–1431. IEEE, 2005. 2
[80] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4358–4366, 2018. 3
[81] Arda Senocak, Hyeonggon Ryu, Junsik Kim, and In So Kweon. Learning sound localization better from semantically similar samples. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4863–4867. IEEE, 2022. 3
[82] Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou, and Haizhou Li. Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3927–3935, 2021. 2
[83] Hamed Rezazadegan Tavakoli, Esa Rahtu, Juho Kannala, and Ali Borji. Digging deeper into egocentric gaze prediction. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 273–282. IEEE, 2019. 3
[84] Sanket Kumar Thakur, Cigdem Beyan, Pietro Morerio, and Alessio Del Bue. Predicting gaze from egocentric social interaction videos and imu data. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 717–722, 2021. 3
[85] Francesco Tordini, Albert S Bregman, and Jeremy R Cooperstock. The loud bird doesn't (always) get the worm: Why computational salience also needs brightness and tempo. Georgia Institute of Technology, 2015. 2
[86] Francesco Tordini, Albert S Bregman, Jeremy R Cooperstock, Anupryia Ankolekar, and Thomas Sandholm. Toward an improved model of auditory saliency. Georgia Institute of Technology, 2013. 2
[87] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 4
[88] Thanh-Dat Truong, Chi Nhan Duong, Hoang Anh Pham, Bhiksha Raj, Ngan Le, Khoa Luu, et al. The right to talk: An audio-visual transformer approach. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1105–1114, 2021. 3
[89] Simon Van Eyndhoven, Tom Francart, and Alexander Bertrand. Eeg-informed attended speaker extraction from recorded speech mixtures with application in neuro-steered hearing prostheses. *IEEE Transactions on Biomedical Engineering*, 64(5):1045–1056, 2016. 2
[90] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4, 5
[91] Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit. *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2017. 6
[92] Julio Wissing, Benedikt Boenninghoff, Dorothea Kolossa, Tsubasa Ochiai, Marc Delcroix, Keisuke Kinoshita, Tomohiro Nakatani, Shoko Araki, and Christopher Schymura. Data fusion for audiovisual speaker localization: Extending dynamic stream weights to the spatial domain. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4705–4709. IEEE, 2021. 2
[93] Jun Xiong, Yu Zhou, Peng Zhang, Lei Xie, Wei Huang, and Yufei Zha. Look&listen: Multi-modal correlation learning for active speaker detection and speech enhancement. *ArXiv*, abs/2203.02216, 2022. 2
[94] Cha Zhang, Pei Yin, Yong Rui, Ross Cutler, Paul Viola, Xinding Sun, Nelson Pinto, and Zhengyou Zhang. Boosting-based multimodal speaker detection for distributed meeting videos. *IEEE Transactions on Multimedia*, 10(8):1541–1552, 2008. 2
[95] Yuan-Hang Zhang, Jingyun Xiao, Shuang Yang, and Shiguang Shan. Multi-task learning for audio-visual active speaker detection. *The ActivityNet Large-Scale Activity Recognition Challenge*, pages 1–4, 2019. 2