The Audio-Visual Conversational Graph: From an Egocentric-Exocentric Perspective

Wenqi Jia[1,2,*], Miao Liu[2], Hao Jiang[2], Ishwarya Ananthabhotla[2],
James M. Rehg[3], Vamsi Krishna Ithapu[2], Ruohan Gao[2]
[1]Georgia Institute of Technology, [2]Meta Reality Labs Research, [3]University of Illinois Urbana-Champaign
wenqi.jia@gatech.edu, {miaoliu,haojiang,ishwarya,ithapu,rhgao}@meta.com, jrehg@illinois.edu

Abstract

*In recent years, the thriving development of research related to egocentric videos has provided a unique perspective for the study of conversational interactions, where both the visual and audio signals play a crucial role. While most prior work focus on learning about behaviors that directly involve the camera wearer, we introduce the Ego-Exocentric Conversational Graph Prediction problem, marking the first attempt to infer exocentric conversational interactions from egocentric videos. We propose a unified multi-modal, multi-task framework — Audio-Visual Conversational Attention (*AV-CONV*), for the joint prediction of conversation behaviors—speaking and listening—for both the camera wearer as well as all other social partners present in the egocentric video. Specifically, we customize the self-attention mechanism to model the representations across-time, across-subjects, and across-modalities. To validate our method, we conduct experiments on a challenging egocentric video dataset that includes first-person perspective, multi-speaker, and multi-conversation scenarios. Our results demonstrate the superior performance of our method compared to a series of baselines. We also present detailed ablation studies to assess the contribution of each component in our model. Project page: https://vjwq.github.io/AVConv/*

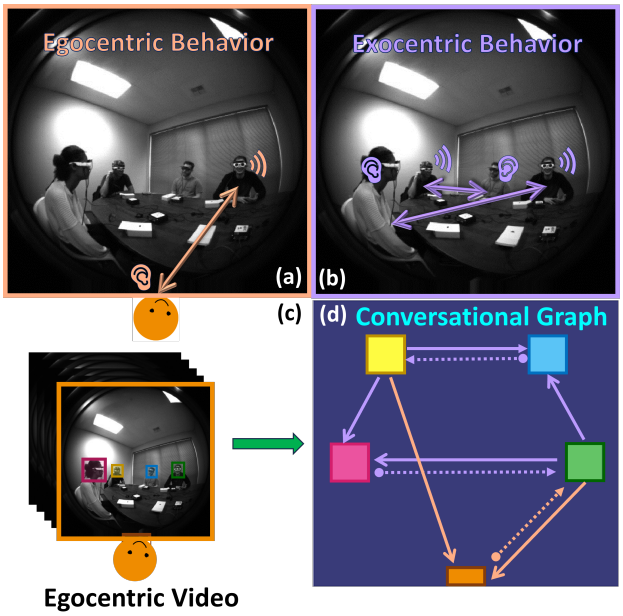

Figure 1. We propose (d) the Ego-Exocentric Conversational Graph Prediction problem that jointly learns (a) the egocentric behaviors—whether the camera wearer is speaking or listening to others), and (b) the exocentric behaviors—whether the other social partners in the scene are speaking or listening to one another, given only the egocentric video input (c).

1. Introduction

With the thriving growth of social media and message-based communication in recent decades, the concept of *Social Graph* [39] has found wide applications, such as on human intention understanding, personalized recommendations [12, 22, 30], etc. Going beyond texts and images, the latest advancement of VR/AR technology and egocentric perception presents a new multimodal data format and also the associated new challenges on constructing graph representation for the multi-participant conversational behaviors in egocentric video streams.

While there has been exciting progress on building large-scale egocentric datasets [7, 13] and models to detect the talking and listening behaviors [18, 29, 34] of the subjects in egocentric videos, existing work mostly focus on analyzing behaviors or actions that directly involve the camera wearer (ego). However, cognitive science studies tell us that we humans also have remarkable ability of understanding other people' belief state during social interactions, often referred as Theory of Mind (ToM) [42]. Subjects involved

* This work was done during an internship at Reality Labs Research.

in a conversation often enact orienting behaviors, such as shifting the position of the head or the focus of the eyes and other overt behaviors, such as hand gestures, synchrony in movement, all of which can be indicators of their explicit (i.e., actively engaged in a conversation) or implicit (i.e., listening discreetly) visual and auditory attention [2, 9, 26].

Motivated by the above, we introduce the concept of *Audio-Visual Conversational Graph*, which describes the conversational behaviors—speaking and listening—for both the camera wearer and all social partners involved in the conversation. As shown in Fig. 1, we present a challenging Ego-Exocentric Conversational Graph Prediction problem. When provided with an egocentric video containing multiple people actively engaged in a conversation, our goal is to create a complete dynamic directed graph that can instantly reflect the conversational behaviors and relationships among all participants. Unlike prior work that focuses on only a single task [18, 34], we simultaneously address multiple closely related tasks in a multi-task setting, and we are the first to explicitly predict the dense conversational behaviors from an exocentric point of view.

We propose the Audio-Visual Conversational Attention (AV-CONV) model that leverages both the multi-channel audio and visual information for analyzing the behaviors and relationships between different social partners. We use a self-attention mechanism tailored to the egocentric conversation setting, and fuse information across-time, across-subjects, and across-modalities. We evaluate our model on a complex multi-speaker, multi-conversation dataset, and obtain an average accuracy of 86.15% on egocentric-related predictions, and an average accuracy of 81.04% on exocentric-related predictions, significantly outperforming the baseline methods.

In summary, we make the following main contributions:

- We introduce the Ego-Exocentric Conversational Graph Prediction problem, the first attempt to explore exocentric conversational interactions from egocentric videos.
- We propose a unified multi-modal, multi-task framework that jointly predicts the interaction states of all social entities captured in the egocentric videos.
- Evaluating our AV-CONV model on a challenging first-person perspective multi-speaker, multi-conversation dataset, we demonstrate the effectiness of our model design compared to baseline methods.

2. Related Work

Exocentric Conversation Interaction. Modeling the interactive behavior between people in a conversation has been a challenging problem. The term *F-formations*, first introduced by anthropologist Edward Hall [14] and further defined by Ciolek and Kendon [5], represents the spatial arrangement of individuals in a group during face-to-face communication. In recent years, a series of work from the computer vision community have contributed to this topic [6, 15, 17, 33, 38, 40]. However, these studies focus on monitoring crowd activities from a surveillance camera or overhead camera perspective, and aim to detect a conversation group as a whole but do not examine the individual relationships within it. To the best of our knowledge, [25] is the only work that studies both first- and third-person activities from an egocentric point of view; however, it targets the detection of actions that are not related to conversational interactions, and does not generate predictions that are conditioned on specified individuals. Our proposed task is the first to explore exocentric inter-person interactions from egocentric videos.

Egocentric Conversation Interaction. Reasoning about human social interactions from a first-person perspective has emerged as a prevailing topic in egocentric vision. Earlier works addressed the problem of social behavior classification [35, 46, 47], social saliency prediction [8, 21, 36], and motion estimation of a conversation partner[24, 37, 45]. Our work is most relevant to the Ego4D [37] Social Benchmark Talking to Me (TTM) task, which identifies the conversation partners that are talking to the camera wearer. Xue et al. [44] introduced a task translation method that achieves state-of-the-art performance on multiple Ego4D benchmark tasks, including TTM. Lin et al. [23] proposed fusing visual prediction and audio prediction based on face quality score to address the TTM task. Jiang et al. [18] introduced the task of Active Speaker Localization (ASL), that entails predicting active speakers in an egocentric scene. Recently, Ryan et al. [34] proposed a Selective Auditory Attention Localization (SAAL) problem, aiming to localize the speaker who is the camera wearer's target of auditory attention. Notably, previous methods are merely a sub-task of our proposed problem of constructing the complete conversational graph. Moreover, we propose the first model that explicitly reasons about the subject-level correlations of the visual and audio signals captured by the egocentric camera to infer their conversational behaviors.

Audio-Visual Learning in Egocentric Vision. Limited prior work has tackled audio-visual learning in the egocentric setting. Thanks to the recent large-scale datasets [7, 13] that contain both visual and audio streams, recent inspiring work integrate cues from both modalities for egocentric action recognition [19, 20, 31, 43], sound object localization in egocentric videos [16], and visual representation learning from audible interactions in egocentric videos [28]. Another stream of work studies egocentric audio-visual learning in simulated environments, enabling embodied agents to both see and hear in order to perceive 3D environments, and tackles tasks such as embodied navigation [3, 4, 11],

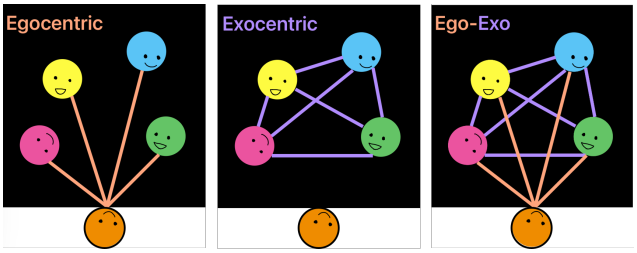

Figure 2. **An illustration of the Conversational Graph.** The left, center, and right figures visualize G_{Ego}, G_{Exo}, and G_{Conv}, respectively. See Sec. 3 for details.

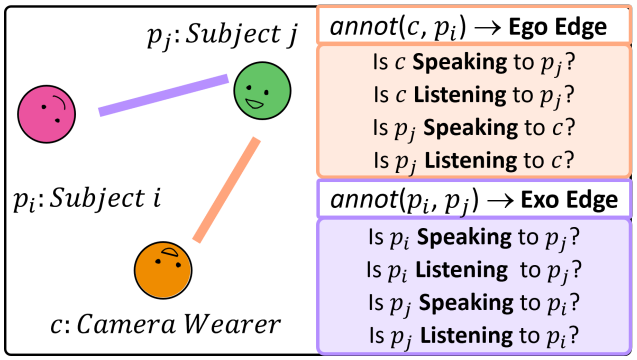

Figure 3. **An example of the edge attributes.** We have binary annotations for each pair of the participants in the conversation, including the camera wear and all other partners.

echolocation [10], and scene mapping [27]. Our work proposes a completely new task, wherein we aim to infer exocentric conversational interactions from egocentric videos leveraging both the visual and audio cues.

3. The Audio-Visual Conversational Graph

We start by formally defining a *Conversational Graph* (Sec. 3.1), and then introduce the dataset and metrics for the proposed task (Sec. 3.2).

3.1. Problem Formulation

Given an egocentric video clip $\mathbf{X} \in \mathbb{R}^{C\times T\times H\times W}$ with T indicating the number of frames, our goal is to generate a dynamic directed graph $\mathbf{G}_{Conv}$ that describes the conversational social interactions– who is looking at and listening to whom, at each moment– of all social partners in $\mathbf{X}$. Formally, we define the *Conversational Graph* $\mathbf{G}_{Conv} = (V, E)$, which consists of two connected components $\mathbf{G}_{Ego}$ and $\mathbf{G}_{Exo}$.

As shown in Fig. 2, $\mathbf{G}_{Ego} = (V_{ego}, E_{ego})$ is a bipartite graph that describes the conversational interactions between the camera wearer and *each* of the other social partners in the scene, while $\mathbf{G}_{Exo} = (V_{exo}, E_{exo})$ is a non-bipartite graph that describes the interactions among all subjects *except* the camera wearer.

More formally, we define the nodes, edges, and edge attributes of $\mathbf{G}_{Conv}$ as follows:

- **Nodes:** $V = V_{ego} + V_{exo}$, with:
  - $V_{ego} = \{c\}$, where c denotes camera wearer.
  - $V_{exo} = \{p_1, p_2, \ldots, p_N\}$, where N denotes the number of other social partners besides the camera wearer.
- **Edges:** $E = E_{ego} + E_{exo}$, in which we have:
  - $E_{ego} = \{c \rightarrow p_i,\ p_i \rightarrow c \mid \text{for all } p_i \in V_{exo}\}$
  - $E_{exo} = \{p_i \rightarrow p_j,\ p_j \rightarrow p_i \mid \text{for all } p_i,\ p_j \in V_{exo} \text{ with } i \neq j \text{ and } i < j\}$
- **Edge Attributes:** For each pair of nodes, we aim to determine (1) whether the two subjects involved are actively engaged in **Speaking To** (S) each other and (2) whether they are actively **Listening To** (L) each other during their social interaction. Therefore, we define the following four types of binary attributes for each pair of nodes:

$$e^S_{c\rightarrow p_i} = \begin{cases} 1 & \text{if } c \text{ is speaking to } p_i \\ 0 & \text{otherwise} \end{cases}$$

$$e^L_{c\rightarrow p_i} = \begin{cases} 1 & \text{if } c \text{ is listening to } p_i \\ 0 & \text{otherwise} \end{cases}$$

$$e^S_{p_i\rightarrow c} = \begin{cases} 1 & \text{if } p_i \text{ is speaking to } c \\ 0 & \text{otherwise} \end{cases}$$

$$e^L_{p_i\rightarrow c} = \begin{cases} 1 & \text{if } p_i \text{ is listening to } c \\ 0 & \text{otherwise} \end{cases}$$

Similarly, for edges $e_{p_i\rightarrow p_j}$, $e_{p_j\rightarrow p_i}$ in E_{exo}, we define the same set of binary attributes for each pair of nodes. An intuitive illustration is shown in Fig. 3 for better understanding. These edge attributes fully characterize the conversational interactions among all subjects involved in the scene, and they are independent from each other, because anyone can be speaking to or listening to one another regardless the behaviors of others.

3.2. Dataset and Annotations

The study of directional conversation-related social behaviors has been explored in the *Talking to Me* task from the Ego4D [13] Social Benchmarks. However, that task focuses on the social behaviors performed by the other social partners towards the camera wearer, and does not provide annotations between the social partners. Besides, it does not provide any listening-related labels. Therefore, we cannot use it for our proposed task.

Instead, we make use of the recently introduced *Egocentric Concurrent Conversations Dataset* [34], which contains a total of 50 participants, evenly distributed across 10∼30-minute data collection sessions, with each session comprising groups of five individuals. Each individual wears a

headset with an Intel SLAM camera and an array of six microphones during the sessions, resulting in ∼20 hours of egocentric videos in total. This means in each session with five people in total, each of them serves as the camera wearer in their egocentric video, and the maximum number of the other social partners captured in the egocentric video frames is four.

To re-purpose the dataset for our problem of computing the audio-visual conversational graph, we generate the ground-truth labels for 4 egocentric interactions, and 6 exocentric interactions in each recorded video leveraging existing annotations. The dataset provides head bounding boxes for all visible faces captured in the egocentric frames, speaking activity annotations of the camera wearer, and the auditory attention target of the camera wearer (as defined and used in their proposed task). Note that these annotations are egocentric-oriented and we need to further synchronize the information collected from different camera wearers within the same data collection session to generate annotations for our exocentric tasks. See the supplementary for details on how we generate the ground-truth annotations.

4. The Audio-Visual Conversational Attention Model

Predicting the complete conversational interactions of all subjects in the aforementioned conversation graph purely from egocentric videos is nontrivial. The complexity arises from the fact that the conversational dynamics among a social group involve a rich set of signals, including the visual appearance of all social partners, the directional sounds they make, and their spatial locations. Moreover, solving the task requires reasoning of the correlations among these diverse signals across all subjects within the social group.

In the following, we first explain how we extract audio-visual representations for analyzing conversational behaviors (Sec. 4.1). Then, we present our Audio-Visual Conversational Attention (AV-CONV) model that disentangles the correlations of the representations from different subjects with an attention mechanism (Sec. 4.2). Finally, we introduce the classifiers used to predict the edge attributes (Sec. 4.3).

4.1. Audio-Visual Feature Extraction

Below we introduce the representations we use for the location of all speakers, their visual appearance, and the multi-channel audio.

Speaker Location. To efficiently capture the location information of each individual, we use the head bounding boxes annotations as mentioned in Sec. 3.2 to crop all visible heads in the video frames. Similar to [34], we use binary masks at the locations of the cropped heads, denoted as $\mathbf{S} \in \mathbb{R}^{N\times 1\times T\times h\times w}$, to preserve the spatial positions for all social partners.

Head Feature Extraction. Given an egocentric video clip $\mathbf{V} \in \mathbb{R}^{3\times T\times H\times W}$, a head image tube $\mathbf{H} \in \mathbb{R}^{N\times 3\times T\times h\times w}$ can be generated based on the speaker location information, where $N = 4$[1]. The Image Encoder $\mathcal{N}_V$ takes the image tube and outputs a downsampled (by a factor of s) feature map, followed by a projection layer that reduces the dimension of the feature map. We have $\mathbf{Z}^v \in \mathbb{R}^{N\times C\times T\times \frac{H}{s}\times \frac{W}{s}}$, where $C = 256$ is the number of channels. Since each frame may capture a maximum of four individuals, we assume that there are four people present in every frame and use zero paddings to represent those who are not visible in the frame.

Audio Feature Extraction. We follow [18, 34] to concatenate channel correlation features with the real and complex part of the multi-channel spectrogram as our audio input. The edge attributes defined in our task are closely related to the location and the vocal activity of each individual. However, the multi-channel audio input only provides the global information for the entire scene. To obtain the representation of the vocal activity for each social partner in each frame, we sequentially concatenate each person's binary mask $\mathbf{S}$ to the audio input to generate a video-length location-aware multi-channel audio signal $\mathbf{A} \in \mathbb{R}^{N\times C\times T\times H\times W}$ as the input to the Audio Encoder $\mathcal{N}_A$. It outputs a downsampled feature map with reduced dimension $\mathbf{Z}^a \in \mathbb{R}^{N\times C\times T\times \frac{H}{s}\times \frac{W}{s}}$, with the same size as $\mathbf{Z}^v$.

After obtaining the visual feature $\mathbf{Z}^v$ and the audio feature $\mathbf{Z}^a$ for each subject in the frame, we concatenate them along the channel dimension to generate the Audio-Visual feature $\mathbf{Z}^{av} \in \mathbb{R}^{N\times D\times T\times \frac{H}{s}\times \frac{W}{s}}$ with $D = 512$ being the number of channels.

4.2. Conversational Attention

We propose *Conversational Attention* $\mathcal{T} = \{\mathcal{T}_T, \mathcal{T}_N, \mathcal{T}_S\}$ that applies a self-attention [41] mechanism to the extracted audio-visual representations. This shares a similar spirit to [1], where they use space-time attention for video understanding. Here we use attention to analyze the conversational interactions across time, subjects, and modalities. We flatten the spatial-temporal dimension of $\mathbf{Z}^{av}$ and obtain $L = T \times N \times S$ tokens, with $S = \frac{H}{s} \times \frac{W}{s}$ and an embedding dimension of D. We add a linear learnable positional embedding $E \in \mathbb{R}^{L\times D}$ to help preserve positional information for each dimension.

Cross-Time Attention. Egocentric vision inherently involves drastic scene changes; this makes it crucial to consider information from both current and neighboring frames for a richer temporal context. $\mathcal{T}_T$ takes in token $\mathbf{Z}^{av} \in \mathbb{R}^{L\times D}$ and applies self-attention to each patch over the tem-

[1]Our model is general in terms of the number of participants N, and we use 4 due to the nature of the dataset we use as described in Sec. 3.2.

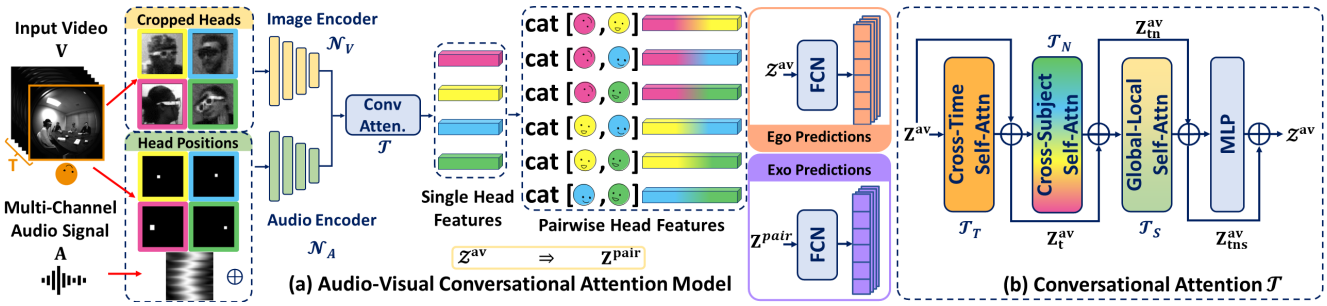

Figure 4. **Model Architecture Overview**: Our model takes multiple egocentric frames and multi-channel audio signals. (a) For each frame, the faces of social partners are cropped to serve as raw visual input, while their corresponding head positions are concatenated with audio inputs to generate positional audio signals. Both visual and audio signals are encoded by two separate ResNet18 Backbones and are concatenated to produce Audio-Visual features for each cropped head. (b) After obtaining temporal Audio-Visual feature tubes of video length, they are flattened into a token to be fed into the Conversational Attention Module to produce augmented Single Head Feature feature $\mathcal{Z}^{av}$. Egocentric Classifiers directly take them to predict Egocentric Edge Attributes, and pairs of these features are arbitrarily combined to generate pairwise audio-visual features to predict Exocentric Edge Attributes.

poral dimension. It outputs a token $\mathbf{Z}_t^{av}$ that aggregates information across time.

Cross-Subject Attention. Our task demands a simultaneous prediction of diverse conversational behaviors for all subjects, thus making it essential to allow representations of each subject to attend to each other for understanding the interpersonal dynamics. $\mathcal{T}_N$ takes in token $\mathbf{Z}_t^{av}$ and applies self-attention to each person's spatiotemporal patch feature over the subject dimension. In this way, we enable the interactions of the audio-visual spatiotemporal features of each individual, thereby increasing the distinction in predictions across different subjects. It outputs token $\mathbf{Z}_{tn}^{av}$.

Global-Local Attention. As depicted in Section 4.1, the head feature is derived from each individual captured in the frame, while the audio feature encompasses both global information about the entire scene and local positional information of the faces. We also apply cross-modality Global-Local self-attention to compute the correlations between the global audio-positional tokens and the local head appearance tokens, enhancing the awareness of the entire scene for each feature and the semantic alignment. $\mathcal{T}_S$ takes in token $\mathbf{Z}_{tn}^{av}$ and compares all patches with each other.

A residual connection is used to aggregate output from each self-attention layer, and the final feature passes through another fully-connected layer. Overall, with a given audio-visual input, we denote the augmented feature $\mathbf{Z}_{tns}^{av}$ as:

$$\mathbf{Z}_{tns}^{av} = \mathcal{T}[\mathcal{N}_V(V) \bigoplus \mathcal{N}_A(A)]. \quad (1)$$

For simplicity, we use the symbol $\mathcal{Z}^{av}$ to refer to $\mathbf{Z}_{tns}^{av}$ in the following text.

4.3. Conversation Edge Attributes Classification

We predict the edge attributes for the egocentric graph G_{Ego} directly based on individual-specific feature $\mathcal{Z}^{av}$ for the corresponding subject in the frame. For edges in the exocentric graph G_{Exo}, we need to pairwise fuse features for any two subjects of interest to predict the corresponding edge attributes. Next, we introduce how we fuse the pairwise feature for the predictions of the exocentric graph, and our prediction head.

Pairwise Feature Fusion. Since $\mathcal{Z}^{av}$ is individual-specific, we need to perform pairwise fusion to obtain Audio-Visual features for any pair of individuals in the video to predict edge attribute for G_{Exo}. Given the maximum possible number of individuals is four in the dataset, we can obtain M ($C_2^4 = 6$) combinations of pairwise features per frame. We directly concatenate the features for the two subjects in each pair to generate the pairwise features $\mathbf{Z}^{pair} \in \mathbb{R}^{M \times P \times T \times \frac{H}{s} \times \frac{W}{s}}$.

Prediction Head. We directly use the refined individual-specific feature $\mathcal{Z}^{av}$ in classifiers for predicting edge attributes of G_{Ego}, and use $\mathbf{Z}^{pair}$ for predicting those for G_{Exo}. As shown in Section 3.1, we use separate classifiers for each type of binary edge attribute. Therefore, we have a total of eight classifiers to predict the complete set of edge attributes. Each of the classifiers contains a pooling layer and a fully connected layer for predicting logits. We train all eight classifiers together with cross-entropy loss and encourage the model to predict the right attribute and to identify the extra background category.

5. Experiments

We validate our approach by comparing it with a series of baselines, and present both quantitative comparisons and qualitative visualizations.

5.1. Implementation Details

Both the visual and audio encoder are adapted from the ResNet-18 backbone pre-trained on ImageNet-1K, which is composed of four convolutional blocks. The AV-CONV

Method	Egocentric Graph				Exocentric Graph			
	$e^S_{c\to p_i}$	$e^S_{p_i\to c}$	$e^L_{c\to p_i}$	$e^L_{p_i\to c}$	$e^S_{p_i\to p_j}$	$e^S_{p_j\to p_i}$	$e^L_{p_i\to p_j}$	$e^L_{p_j\to p_i}$
SAAL (Acc)	86.23	67.10	86.48	86.99	/	/	/	/
ASL+Layout (Acc)	13.71	55.53	83.48	77.35	65.96	43.63	64.49	79.04
AV-CONV (Acc)	**90.02**	**75.94**	**87.80**	**90.63**	**86.15**	**75.89**	**75.91**	**85.75**
SAAL (mAP)	68.43	44.97	44.64	39.55	/	/	/	/
ASL+Layout (mAP)	86.28	47.45	21.83	47.91	45.91	46.68	18.98	16.15
AV-CONV (mAP)	**82.08**	**68.94**	**60.70**	**65.48**	**72.73**	**63.36**	**32.35**	**29.29**

Table 1. **Comparing to Prior Work.** There are no pre-exisiting baselines directly comparible for our proposed new task. We devise two baseline methods with components adapted from prior work: SAAL [34] and ASL [18]+Layout. We report both accuracy and mAP.

	Egocentric Graph					Exocentric Graph				
	$e^S_{c\to p_i}$	$e^S_{p_i\to c}$	$e^L_{c\to p_i}$	$e^L_{p_i\to c}$	Ego Avg	$e^S_{p_i\to p_j}$	$e^S_{p_j\to p_i}$	$e^L_{p_i\to p_j}$	$e^L_{p_j\to p_i}$	Exo Avg
DIRECT CONCAT	88.69	68.65	83.83	86.85	82.00	67.60	69.53	85.07	83.52	76.43
AV-CONV (T)	89.49	73.60	86.88	87.43	84.35	72.97	74.57	85.13	84.36	79.26
AV-CONV (N)	88.62	68.83	85.12	88.30	82.72	68.39	69.34	85.72	84.36	76.95
AV-CONV (S)	88.62	69.38	85.11	87.47	82.65	68.11	70.23	85.09	83.96	76.85
AV-CONV (TN)	89.58	75.04	87.05	88.89	85.14	74.59	75.12	85.91	85.44	80.27
AV-CONV (TS)	89.36	75.12	86.57	88.29	84.84	75.42	74.76	85.23	84.25	79.92
AV-CONV (NS)	89.73	74.23	87.52	88.81	85.07	73.60	74.37	86.53	85.51	80.00
AV-CONV	**90.02**	**75.94**	**87.80**	**90.63**	**86.15**	**75.89**	**75.91**	**86.61**	**85.75**	**81.04**

Table 2. **Ablation Study on Conversational Attention**. To explore the separate impact of cross-time attention, cross-subject attention, and Gloabl-Local attention, we exhaustively explore different combinations of the components in our conversational attention model. We report the classification accuracy and see Supp. for the mAP results.

model is composed of two 8-head self-attention blocks [41], and we use similar implementations as in [1]. The model is implemented in PyTorch, and trained for around 9 epochs using an Adam optimizer with a learning rate of 1e-4.

5.2. Baselines

Since we are the first to address the challenging problem of conversational graph, there are no pre-existing baselines directly comparable. We compare with two baselines adapted from prior work [18, 34], and a series of ablated versions of our method.

- **SAAL** [34]: SAAL was originally designed to predict the camera wearer's auditory attention from an egocentric point of view, corresponding to $e^L_{c\to p_i}$ in our problem setting. To extend its applicability to our broader egocentric-related tasks, we adapt our annotations to their setting and add extra prediction layers to its decoder for the other egocentric tasks we tackle.
- **ASL [18] + Layout**: This is a heuristics baseline that combines 3D person layout estimation and the active speaker localization (ASL) to infer the ego and exo conversational interactions. Using single view depth estimation [32] and a 3D head pose regression model, we predict the participants' 3D head locations and facing directions in the camera wearer's frame. We consider the interaction probability of two subjects proportional to the angle between the vector from one person to another and the the facing direction of the person. We multiply the interaction probability and the voice activity probability to generate the final predictions for each edge attribute.
- **DIRECT CONCAT**: We exclude the entire Conversational Attention $\mathcal{T}$ component for this baseline. The feature representation $\mathbf{Z}^{av}$ is obtained by directly concatenating $\mathbf{Z}^v$ and $\mathbf{Z}^a$ without additional augmentation.
- **AV-CONV (T, N, S)**: To explore the separate impact of Cross-Time attention (T), Cross-Subject attention (N), and Global-Local attention (S), we use only $\mathcal{T}_T$, $\mathcal{T}_N$, and $\mathcal{T}_S$, respectively from our model for feature fusion while keeping all other settings the same.
- **AV-CONV (TN, TS, NS)**: This is the same as the previous three baselines except that we use two types of attention for aggregating the features. In particular, AV-CONV (TS) uses similar spatiotemporal attention mechanism commonly employed in various video understanding work [1, 34].
- **HEAD/AUDIO/MASK ONLY**: To explore the impact of different input modalities, we retain only the head image input (Head), the multi-channel audio signal (Audio), or the positional binary masks (Mask) for predicting the edge attributes of the conversational graph.
- **HEAD+MASK / AUDIO+MASK**: This is the same the previous three baselines except that we discard either the head image or the audio input.

5.3. Quantitative Results

Table 1, 2, 3 show our results for comparing with the prior methods, ablation study on our conversational attention de-

	Egocentric Graph					Exocentric Graph				
	$e^S_{c\to p_i}$	$e^S_{p_i\to c}$	$e^L_{c\to p_i}$	$e^L_{p_i\to c}$	Ego Avg	$e^S_{p_i\to p_j}$	$e^S_{p_j\to p_i}$	$e^L_{p_i\to p_j}$	$e^L_{p_j\to p_i}$	Exo Avg
HEAD ONLY	63.18	57.76	79.34	80.39	70.17	57.51	58.43	84.58	82.98	70.81
AUDIO ONLY	88.57	59.34	77.12	76.97	75.50	32.47	33.60	67.43	71.84	51.34
MASK ONLY	63.33	58.32	81.03	80.21	70.72	57.57	58.75	84.96	84.09	71.19
HEAD+MASK	64.45	59.18	80.86	80.92	71.31	59.66	59.17	84.50	81.59	71.23
AUDIO+MASK	89.20	75.29	**87.89**	87.74	85.03	75.67	74.74	84.64	84.06	79.78
AV-CONV	**90.02**	**75.94**	87.80	**90.63**	**86.15**	**75.89**	**75.91**	**86.61**	**85.75**	**81.04**

Table 3. **Modality Ablation**. As described in 4.1, our input signals consist of three components: head images, multi-channel audio, and binary position masks. In this ablation study, we explore different choices of input signals, and assess their relative impact on the final performance. We report the classification accuracy and see Supp. for the mAP results.

sign, ablation study on input modalities, respectively. In all three tables, the best results are highlighted with **boldface**. Since the prediction of each edge attribute is a binary classification problem. We report the accuracy and mean Average Precision (mAP) for each edge attribute classification. Additionally, we present the average performance for egocentric and egocentric tasks. Namely, $e^S_{c\to p_i}$, $e^S_{p_i\to c}$, $e^L_{c\to p_i}$, $e^L_{p_i\to c}$ of $\mathbf{G}_{Ego}$ represent the conversational interaction between the camera wear and other social partner, while $e^S_{p_j\to p_i}$, $e^S_{p_i\to p_j}$, $e^L_{p_j\to p_i}$, $e^L_{p_i\to p_j}$ of $\mathbf{G}_{Exo}$ represent the pair-wise conversational interaction between arbitrary two social partners.

Comparing to Prior Work. As shown in Table 1, our AV-CONV model consistently outperforms both the SAAL and ASL+Layout baselines across all sub-tasks. For SAAL, we outperform it by an average of approximately 4.45% on egocentric-related tasks ($+3.99\%$, $+8.84\%$, $+1.32\%$, $+3.64\%$, respectively), demonstrating the benefit of learning these closely related sub-tasks jointly. We observe that the sub-task with the smallest margin is $e^L_{c\to p_i}$, and this is because SAAL is tailored to address this sub-task, thus making the model design and the fine-tuned model parameters well suited for this task. Compared with the heuristic baseline ASL+Layout, our model outperform it by a large margin for both the egocentric and exocentric edges. This further demonstrates that this is a challenging task, which can not be easily solved just by leveraging the people layout and the results from active speaker localization.

Ablation Study on Conversational Attention. In Table 2, we investigate how each component and their combinations in the Conversational Attention module $\mathcal{T}$ contribute to the overall performance. We can see that DIRECT CONCAT leads to the worst performance across almost all tasks, particularly on identifying the exocentric speakers. Using either the Cross-Time attention, Cross-Subject attention, or the Gloabl-Local attention all positively contribute to the final performance. Noticeably, attention across time leads to the largest gain, suggesting the importance of aggregating information from nearby frames to more reliably detect speech activities.

Our final model leverages all three types of attention mechanisms to build the conversational attention block, resulting in a more comprehensive understanding of conversational interactions in egocentric videos. It outperforms the DIRECT CONCAT approach by an average of approximately 4.15% on egocentric-related tasks ($+1.33\%$, $+7.29\%$, $+3.97\%$, $+3.78\%$, respectively) and 4.61% on exocentric-related tasks ($+8.29\%$, $+6.38\%$, $+1.54\%$, $+2.23\%$, respectively). The Cross-Time attention enables the model to capture temporal dependencies and aggregate information from adjacent frames, enhancing its ability to detect voice activities over time. The Cross-Subject attention contributes by effectively comparing and distinguishing features from different individuals, particularly those related to social partners' relative positions in the frame. The Global-Local attention focuses on features extracted from the appearance of input signals, capturing orientations of heads, movement of lips, and facial expressions. The model achieves a synergistic effect, benefiting from the complementary nature of each attention type.

Ablation Study on Input Modality. As illustrated in Sec. 4.1, our input modalities consist of three components: 1) *Heads images* cropped from the egocentric frames, which are subject-specified. 2) *Audio input* containing global cross-correlation features of multi-channel audio from the egocentric video. 3) *Positional Binary Mask*, specifying each individual's location in the frame and serving as an intermediate global-local, subject-specified representation.

In Table 3, we show the results of evaluating the model's performance by selectively excluding one or more modalities to discern their contributions. As expected, using only the head images leads to significant performance degradation on all speaking-related tasks, as the reasoning of the speaking behaviors needs audio. On the other hand, using only audio produces good results on Ego_S while failing on the other tasks, implying the necessity of local features that relate to conversational partners. Using only the positional mask has similar performance as using only the head image, but it infers the potential relationships among social partners based on the abstract representation of their head locations in the entire scene. Interestingly, it outperforms

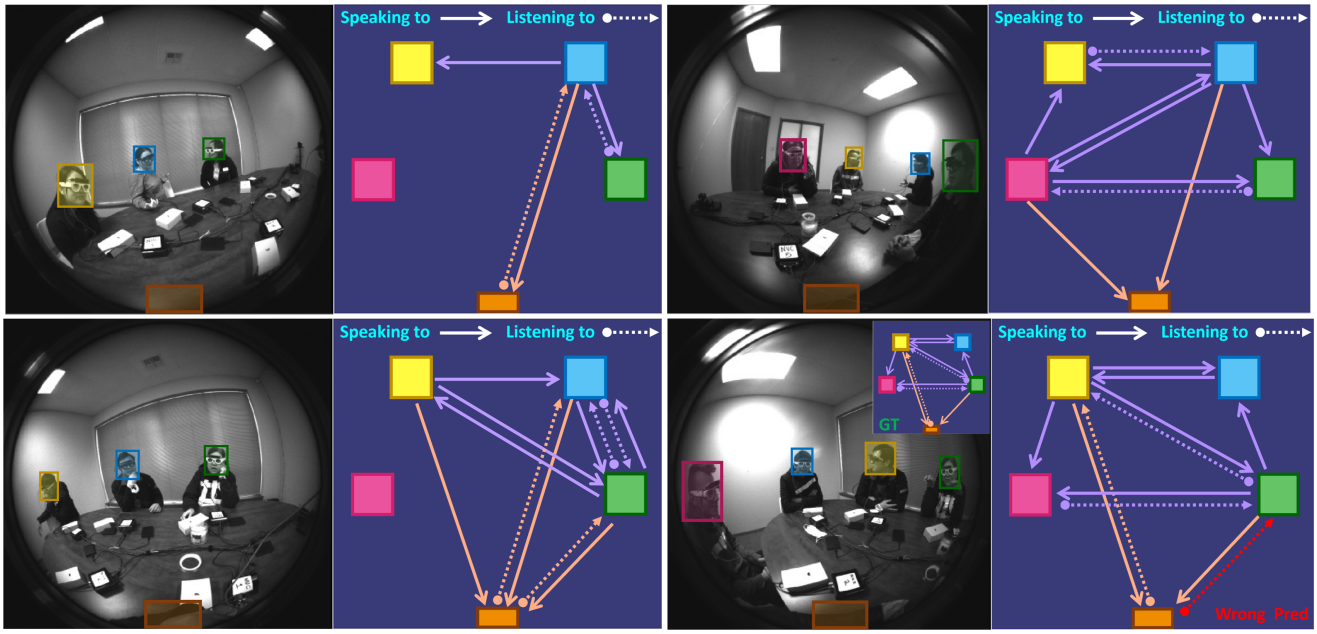

Figure 5. **Visualization of the Ego-Exocentric Conversational Graph from our model prediction.** We show three successful cases and one failure case in the bottom right. For the last failure example, we also overlay the ground truth of the conversational graph on the top right corner of the video frame as reference.

HEAD ONLY on listening sub-tasks, probably because they rely more on the position information of people in the space.

When omitting either the audio signal (HEAD+MASK) or the visual signal (AUDIO+MASK), we find that the multi-channel audio is more useful when combined with the positional mask, which also agrees with findings in prior work [34]. This audio-visual modality input provides both global audio activity information and a local abstracted visual prior of all social partners' positions from an egocentric point of view, including interpersonal relative location relationships. It outperforms HEADS+MASK by an average of approximately 13.72% on egocentric-related tasks (+24.75%, +16.11%, +7.03%, +6.82%, respectively) and 8.55% on exocentric-related tasks (+16.01%, +15.57%, +0.14%, +2.47%, respectively). The main contribution comes from speaking-related tasks. When adding all three modalities as our input, we achieve the best performance.

The above results underscores the synergistic contribution of each modality to the overall effectiveness of the proposed model in predicting egocentric social interactions. The combination of subject-specified visual cues, audio information, and spatial context through the positional mask prove to be essential for achieving comprehensive and accurate predictions.

5.4. Visualization

In Fig. 5, we visualize some examples from the *Egocentric Concurrent Conversations Dataset* and show the prediction results of the conversational graph from our model. We can see that it is a very challenging problem, as the subjects in the video frames are all visually similar and exhibit complex conversational interactions—there can be multiple people speaking and listening at the same time. Despite the challenges and diverse people layout, our model makes accurate predictions of the complete conversation behaviors for both the camera wearer as well as all other social partners present in the scene from just the egocentric videos, demonstrating the effectiveness of our AV-CONV model.

The bottom right corner shows a typical failure case, where our model mistakenly predicts that the camera wearer is listening to the subject in green. We suspect it's because the subjects in blue and green are both speaking at the same time and they are physically very close to each other, so that the model finds hard to tell who is speaking. It would be interesting future work to explore better audio representation to capture more high-resolution spatial information from the multi-channel audio.

6. Conclusion

We presented the Audio-Visual Conversational Graph Prediction problem to infer exocentric conversational interactions from egocentric videos. With our unified multi-modal, multi-task framework AV-CONV, we have demonstrated the superiority of our proposed method on a challenging multi-speaker, multi-conversation egocentric dataset. Our work marks a new research direction to learn exocentric so-

cial interactions from egocentric videos. As future work, we plan to extend our framework to jointly solve other social behavior tasks such as gaze prediction, and study more complex social relationships such as conversation groups.

References

[1] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, page 4, 2021. 4, 6

[2] W Owen Brimijoin, David McShefferty, and Michael A Akeroyd. Auditory and visual orienting responses in listeners with and without hearing-impairment. *The Journal of the Acoustical Society of America*, 127(6), 2010. 2

[3] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 17–36. Springer, 2020. 2

[4] Changan Chen, Sagnik Majumder, Ziad Al-Halah, Ruohan Gao, Santhosh Kumar Ramakrishnan, and Kristen Grauman. Learning to set waypoints for audio-visual navigation. *arXiv preprint arXiv:2008.09622*, 2020. 2

[5] T Matthew Ciolek and Adam Kendon. Environment and the spatial arrangement of conversational encounters. *Sociological Inquiry*, 50(3-4):237–271, 1980. 2

[6] Marco Cristani, Loris Bazzani, Giulia Paggetti, Andrea Fossati, Diego Tosato, Alessio Del Bue, Gloria Menegaz, and Vittorio Murino. Social interaction discovery by statistical analysis of f-formations. In *BMVC*, pages 10–5244, 2011. 2

[7] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018. 1, 2

[8] Alircza Fathi, Jessica K Hodgins, and James M Rehg. Social interactions: A first-person perspective. In *CVPR*, 2012. 2

[9] Alexandra Frischen, Andrew P Bayliss, and Steven P Tipper. Gaze cueing of attention: visual attention, social cognition, and individual differences. *Psychological bulletin*, 133(4): 694, 2007. 2

[10] Ruohan Gao, Changan Chen, Ziad Al-Halah, Carl Schissler, and Kristen Grauman. Visualechoes: Spatial image representation learning through echolocation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 658–676. Springer, 2020. 3

[11] Ruohan Gao, Hao Li, Gokul Dharan, Zhuzhu Wang, Chengshu Li, Fei Xia, Silvio Savarese, Li Fei-Fei, and Jiajun Wu. Sonicverse: A multisensory simulation platform for training household agents that see and hear. In *ICRA*, 2023. 2

[12] Mark Granovetter. The strength of weak ties: A network theory revisited. *Sociological theory*, pages 201–233, 1983. 1

[13] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 1, 2, 3

[14] Edward Twitchell Hall. *The hidden dimension*. Anchor, 1966. 2

[15] Hooman Hedayati, Daniel Szafir, and Sean Andrist. Recognizing f-formations in the open world. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 558–559. IEEE, 2019. 2

[16] Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Egocentric audio-visual object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22910–22921, 2023. 2

[17] Hayley Hung and Ben Kröse. Detecting f-formations as dominant sets. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 231–238, 2011. 2

[18] Hao Jiang, Calvin Murdock, and Vamsi Krishna Ithapu. Egocentric deep multi-channel audio-visual active speaker localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10544–10552, 2022. 1, 2, 4, 6

[19] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5492–5501, 2019. 2

[20] Evangelos Kazakos, Jaesung Huh, Arsha Nagrani, Andrew Zisserman, and Dima Damen. With a little help from my temporal context: Multimodal egocentric action recognition. In *BMVC*, 2021. 2

[21] Bolin Lai, Miao Liu, Fiona Ryan, and James Rehg. In the eye of transformer: Global-local correlation for egocentric gaze estimation. *British Machine Vision Conference*, 2022. 2

[22] Jure Leskovec and Julian Mcauley. Learning to discover social circles in ego networks. *Advances in neural information processing systems*, 25, 2012. 1

[23] Hsi-Che Lin, Chien-Yi Wang, Min-Hung Chen, Szu-Wei Fu, and Yu-Chiang Frank Wang. Quavf: Quality-aware audio-visual fusion for ego4d talking to me challenge. *arXiv preprint arXiv:2306.17404*, 2023. 2

[24] Miao Liu, Dexin Yang, Yan Zhang, Zhaopeng Cui, James M Rehg, and Siyu Tang. 4d human body capture from egocentric video via 3d scene grounding. In *2021 international conference on 3D vision (3DV)*, pages 930–939. IEEE, 2021. 2

[25] Tianshan Liu, Rui Zhao, Wenqi Jia, Kin-Man Lam, and Jun Kong. Holistic-guided disentangled learning with cross-video semantics mining for concurrent first-person and third-person activity recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 2

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

[27] Sagnik Majumder, Hao Jiang, Pierre Moulon, Ethan Henderson, Paul Calamia, Kristen Grauman, and Vamsi Krishna Ithapu. Chat2map: Efficient scene mapping from multi-ego conversations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10554–10564, 2023. 3

[28] Himangi Mittal, Pedro Morgado, Unnat Jain, and Abhinav Gupta. Learning state-aware visual representations from audible interactions. *Advances in Neural Information Processing Systems*, 35:23765–23779, 2022. 2

[29] Curtis Northcutt, Shengxin Zha, Steven Lovegrove, and Richard Newcombe. Egocom: A multi-person multi-modal egocentric communications dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1

[30] Brea L Perry, Bernice A Pescosolido, and Stephen P Borgatti. *Egocentric network analysis: Foundations, methods, and models*. Cambridge university press, 2018. 1

[31] Mirco Planamente, Chiara Plizzari, Emanuele Alberti, and Barbara Caputo. Cross-domain first person audio-visual action recognition through relative norm alignment. *arXiv preprint arXiv:2106.01689*, 2021. 2

[32] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020. 6

[33] Elisa Ricci, Jagannadan Varadarajan, Ramanathan Subramanian, Samuel Rota Bulo, Narendra Ahuja, and Oswald Lanz. Uncovering interactions and interactors: Joint estimation of head, body orientation and f-formations from surveillance videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4660–4668, 2015. 2

[34] Fiona Ryan, Hao Jiang, Abhinav Shukla, James M Rehg, and Vamsi Krishna Ithapu. Egocentric auditory attention localization in conversations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14663–14674, 2023. 1, 2, 3, 4, 6, 8

[35] Michael S Ryoo and Larry Matthies. First-person activity recognition: What are they doing to me? In *CVPR*, 2013. 2

[36] Hyun Soo Park and Jianbo Shi. Social saliency prediction. In *CVPR*, 2015. 2

[37] Hyun Soo Park, Jyh-Jing Hwang, Yedong Niu, and Jianbo Shi. Egocentric future localization. In *CVPR*, 2016. 2

[38] Stephanie Tan, David MJ Tax, and Hayley Hung. Conversation group detection with spatio-temporal context. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pages 170–180, 2022. 2

[39] Johan Ugander, Brian Karrer, Lars Backstrom, and Cameron Marlow. The anatomy of the facebook social graph. *arXiv preprint arXiv:1111.4503*, 2011. 1

[40] Sebastiano Vascon, Eyasu Z Mequanint, Marco Cristani, Hayley Hung, Marcello Pelillo, and Vittorio Murino. Detecting conversational groups in images and sequences: A robust game-theoretic approach. *Computer Vision and Image Understanding*, 143:11–24, 2016. 2

[41] 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, 6

[42] Henry M Wellman. *The child's theory of mind.* The MIT Press, 1992. 1

[43] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020. 2

[44] Zihui Xue, Yale Song, Kristen Grauman, and Lorenzo Torresani. Egocentric video task translation@ ego4d challenge 2022. *arXiv preprint arXiv:2302.01891*, 2023. 2

[45] Takuma Yagi, Karttikeya Mangalam, Ryo Yonetani, and Yoichi Sato. Future person localization in first-person videos. In *CVPR*, 2018. 2

[46] Zhefan Ye, Yin Li, Yun Liu, Chanel Bridges, Agata Rozga, and James M Rehg. Detecting bids for eye contact using a wearable camera. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8. IEEE, 2015. 2

[47] Ryo Yonetani, Kris M Kitani, and Yoichi Sato. Recognizing micro-actions and reactions from paired egocentric videos. In *CVPR*, 2016. 2