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

Wenqi Jia[1,2], Miao Liu[4], 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, [4] GenAI, Meta

CVPR SEATTLE, WA JUNE 17-21, 2024

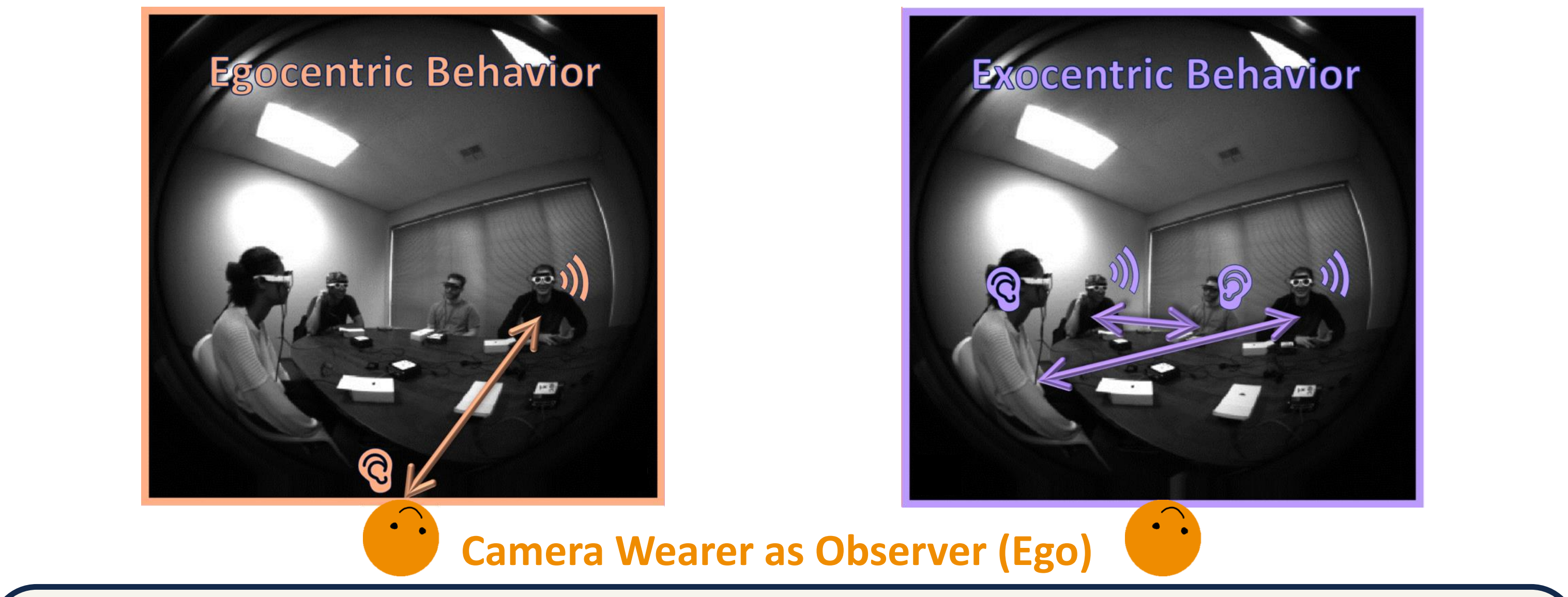
Find more on our project page!

Motivation

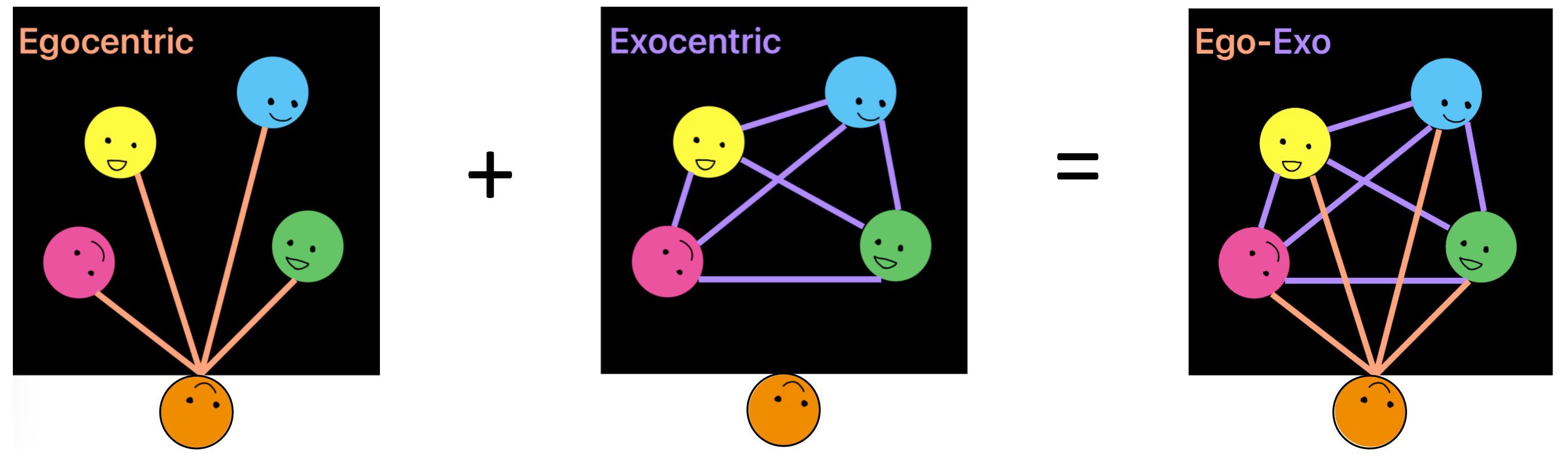
Concurrent conversations are common in life

- Could be noisy and ambiguous
- Capturing social states of participants helps decide which sound source to enhance for whom
- Facilitate effective and efficient communication

Ego-Exo Conversational Graph



Humans can understand both *Egocentric* and *Exocentric* conversational behaviors



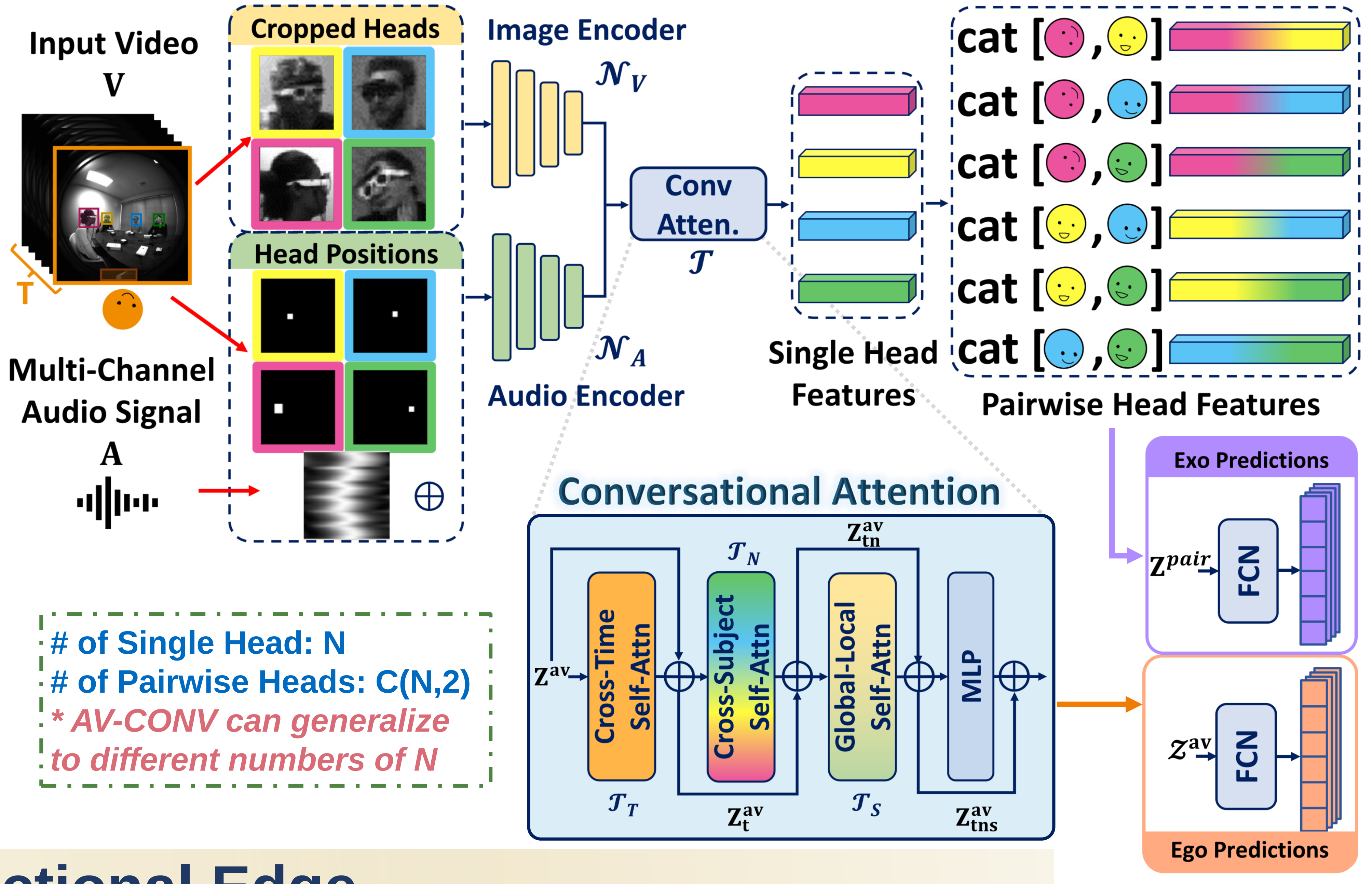
Ego-Exocentric Conversational Graph Prediction

the *first* to explore Exocentric conversational interactions from Egocentric videos

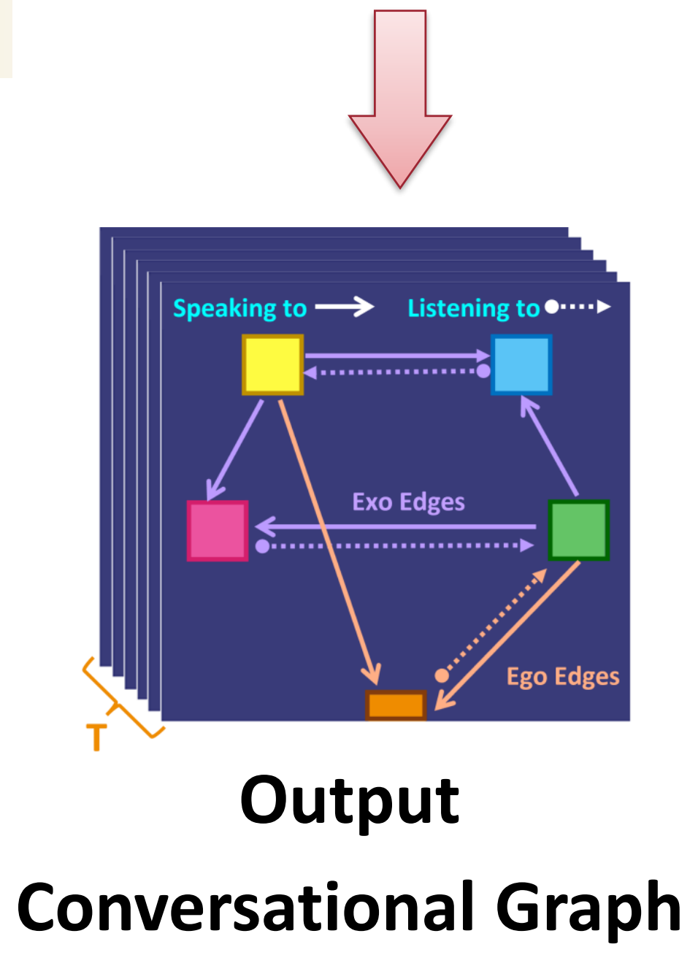
✓ Jointly modeling talking *and* listening behaviors

✓ Jointly modeling Egocentric *and* Exocentric behaviors as graph

Method



# of Single Head: N
# of Pairwise Heads: C(N,2)
* AV-CONV can generalize to different numbers of N

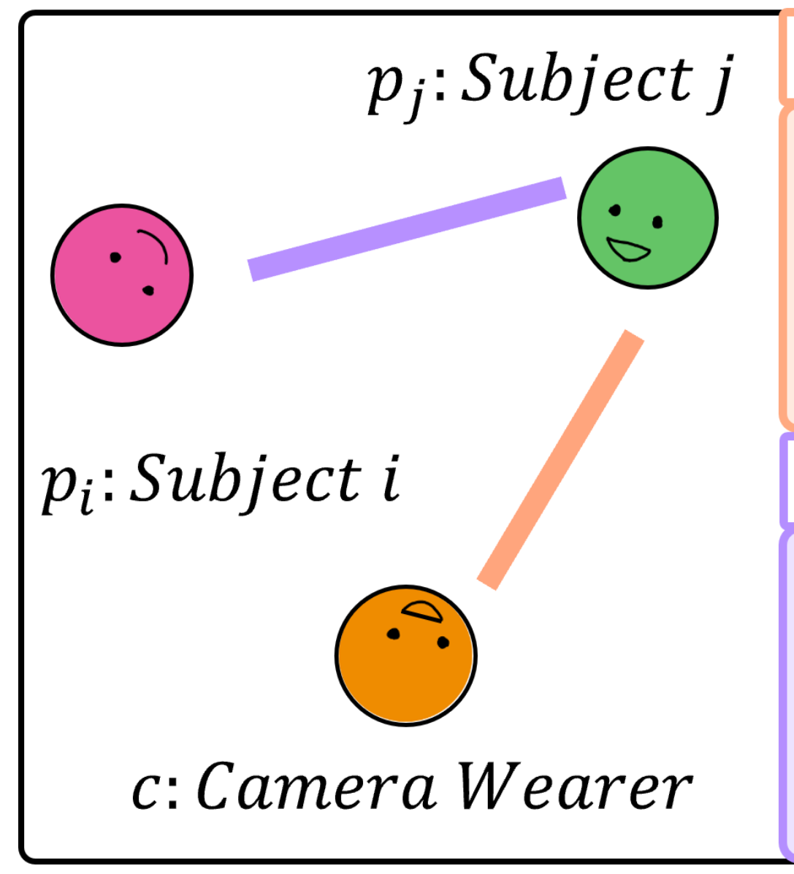


Output Conversational Graph

Ego-Exo Directional Edge

- For each pair of nodes (c, p_j) or (p_i, p_j), we aim to determine:
  - If they are **Speaking To** (*S*) each other
  - If they are **Listening To** (*L*) each other
- Results in four attributes:
  - For each Egocentric Edge:
  - For each Exocentric Edge:

$e^S_{c\to p_i}$	$e^S_{p_i\to c}$	$e^L_{c\to p_i}$	$e^L_{p_i\to c}$
$e^S_{p_i\to p_j}$	$e^S_{p_j\to p_i}$	$e^L_{p_i\to p_j}$	$e^L_{p_j\to p_i}$



$annot(c, p_i) \to$ **Ego Edge**

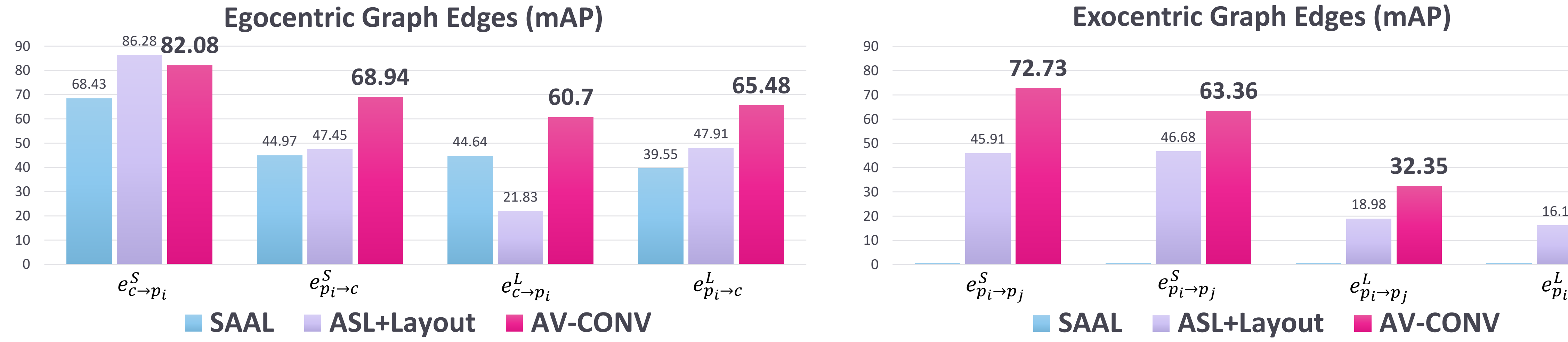
Is c **Speaking** to p_j?
Is c **Listening** to p_j?
Is p_j **Speaking** to c?
Is p_j **Listening** to c?

$annot(p_i, p_j) \to$ **Exo Edge**

Is p_i **Speaking** to p_j?
Is p_i **Listening** to p_j?
Is p_j **Speaking** to p_i?
Is p_j **Listening** to p_i?

Experiments and Results

- **Dataset:** Egocentric Concurrent Conversations Dataset (15,682/6,329 Train/test)
- **Baselines:** 1. SAAL (Ryan 2023)
  2. Active Speaker Localization (Jiang 2022) + 3D person layout estimation



Our AV-CONV consistently outperforms both baselines across all subtasks

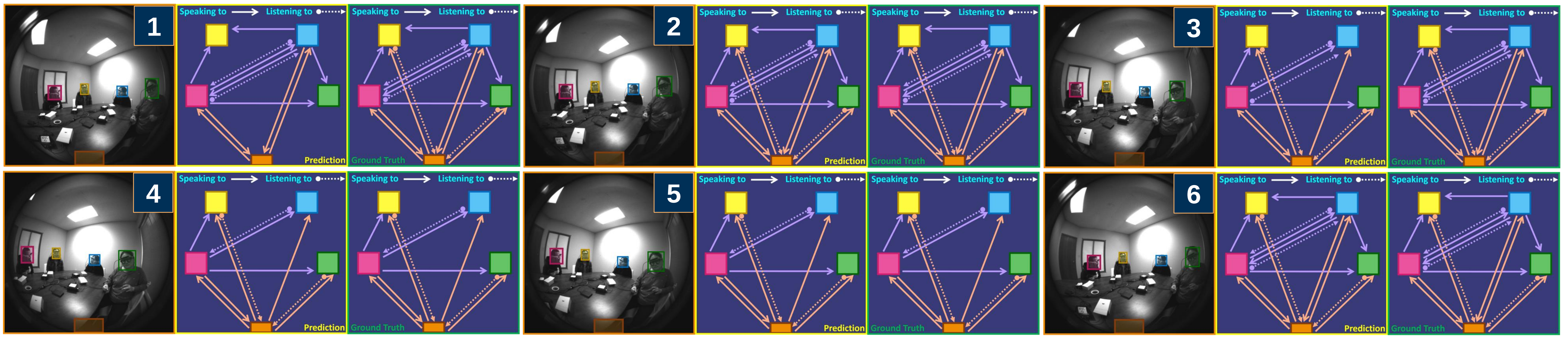
subject-specified → Head Only
global → Audio Only
global-local, subject-specified → Mask Only

	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	51.20	51.65	37.19	29.38	42.36	54.52	48.12	16.48	17.33	34.11
Audio Only	**84.32**	53.43	22.94	24.26	46.24	51.63	43.89	14.17	15.58	31.32
Mask Only	54.55	52.18	39.27	33.54	44.89	55.00	47.29	14.93	16.09	33.33
Head+Mask	47.84	50.28	35.80	22.38	39.08	52.85	45.90	14.83	15.89	32.37
Audio+Mask	45.83	47.40	22.83	21.31	34.34	50.40	43.86	14.76	15.95	31.24
AV-Conv	82.08	**68.94**	**60.70**	**65.48**	**69.30**	**72.73**	**63.36**	**32.35**	**29.29**	**49.43**

The combination of subject-specified visual cues, global audio information, and spatial context through the positional mask is essential for accurate prediction

Visualization

- **Conversational Dynamics:** 6 frames with a temporal stride of 15, ~3 seconds



- **Future Work:** 1) extend our framework to other social behavior; 2) study more complex social relationships such as conversation groups' mobility

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