The Audio-Visual Conversational Graph: From an Egocentric-Exocentric Perspective (Supplementary Materials)

The supplementary materials consist of:

1. Annotation Details

Here we introduce the details on how we obtain the labels for the edge attributes of the conversational graph in our task. Ideally, we should annotate each individual's intention of speaking and listening behavior per moment. However, such densely annotated per-moment intention labels are not available in the *Egocentric Concurrent Conversations Dataset* we use, and also hard to obtain in practice. The *Egocentric Concurrent Conversations Dataset* pre-defines two two- or three-people conversational groups per session, and all five participants are instructed to engage in conversations with their own group. In this way, each participant selectively listens to others who belong to their same group in their sessions with multiple self-driven concurrent speakers, allowing for close estimation of auditory attention ground truth. They are synchronized between all participants' annotations to construct the **Listening To** ground-truth labels in edge attributes. Unlike the **Listening To** label, we define the **Speaking To** behavior as not selective, as speaking is a spontaneous behavior and the intentions of speakers are covert hence difficult to quantify. Similarly, they are spread into each ego- and exocentric edges to represent the speaking attributes in each conversational edge.

A statistical analysis of annotations in the dataset we used reveals that the ratio of positive to negative "Listening To" labels is approximately 1:2, while the ratio of "Speaking To" labels is roughly 1:1. A RANDOM GUESS baseline on a subset achieves accuracies of 24.17% and 53.75% for Egocentric and Exocentric Average Performances, respectively.

2. AV-CONV Model Architectures

The input egocentric frames $\mathbf{V}$, multi-channel audio signals $\mathbf{A}$, and binary mask $\mathbf{S}$ are all resized to 210×210 for proper alignment. To capture the evolution of the conversational graph through a longer temporal stride setting, each instance input in our experiments consists of 6 frames with a temporal stride of 15, spanning a 90-frame window equivalent to 3 seconds. This results in 15682/6329 (Train/Val) audio-visual samples. Predictions are made all at once on each frame, corresponding to a 0.5-second interval. We provide model details for AV-CONV in Fig. 1.

3. Computational Cost and Scalability

AV-CONV costs 53M parameters and 25.28 GFLOPs, and can generalize to different numbers of faces such as larger groups with more visible heads, though with an increased computational cost. For example, additional analysis shows that our model uses the same amount of GFLOPs if there are six people present in the scene, and we can still train our model on 2 GeForce RTX 4090s with a batch size of 4. It is because in our setting, the nature of handling more faces simply equals enlarging the batch size, thus not resulting in more operations.

4. Ablation Studies with mAP Results

4.1. Ablation Study on Conversational Attention

With the second metric mAP, we observe a similar pattern in Table 1 as it in the main paper. Our final model AV-CONV outperforms the DIRECT CONCAT baseline by an average of 7.18% on almost all egocentric-related tasks (-2.48%, $+6.98\%$, $+14.28\%$, $+11.27\%$) and 9.35% on all exocentric-related tasks ($+7.80\%$, $+5.17\%$, $+15.09\%$, $+10.98\%$).

4.2. Ablation Study on Input Modality

However, patterns in Table 2 are slightly different from those in main paper. While MASK ONLY still marks the best performance among almost all single-modality ablations, omitting either the audio signal (HEAD+MASK) or

<table>
<tr><td colspan="2">Visual Input [T, C, H, W]</td><td>Audio Input [T, C, H, W]</td></tr>
<tr><td colspan="2">6 x 1 x 210 x 210</td><td>6 x 3 x 210 x 210</td></tr>
<tr><td>Cropped Heads</td><td>Binary Mask</td><td>Copy x N</td></tr>
<tr><td rowspan="3">6 x 4 x 1 x 210 x 210</td><td>6 x 4 x 1 x 210 x 210</td><td>6 x 4 x 3 x 210 x 210</td></tr>
<tr><td colspan="2">Concatenation</td></tr>
<tr><td colspan="2">6 x 4 x 4 x 210 x 210</td></tr>
<tr><td>ResNet 18 + Projection</td><td colspan="2">ResNet 18 + Projection</td></tr>
<tr><td>6 x 4 x 256 x 14 x 14</td><td colspan="2">6 x 4 x 256 x 14 x 14</td></tr>
<tr><td colspan="3">Concatenation + Flatten [T, N, C, HW]</td></tr>
<tr><td colspan="3">6 x 4 x 512 x 196, Positional Embedding: (6 x 4 x 196, 512)</td></tr>
<tr><td colspan="3">Conversational Attention</td></tr>
<tr><td colspan="3">6 x 4 x 512 x 196</td></tr>
<tr><td>Ego FCN Classifier</td><td colspan="2">Pairwise Fusion [T, M, C, HW]</td></tr>
<tr><td rowspan="3">Preds [6 x 4]</td><td colspan="2">6 x 6 x 1024 x 196</td></tr>
<tr><td colspan="2">Exo FCN Classifier</td></tr>
<tr><td colspan="2">Preds [6 x 6]</td></tr>
</table>

Figure 1. Architecture Details of AV-CONV

the visual signal (AUDIO+MASK) results in a drop in recall on all tasks. This suggests that both audio and visual signals play an important role in the model's performance, and that combining them through multi-modal fusion is crucial for achieving optimal results. Results with using all modality outperforms MASK ONLY by an average of 24.41% on all egocentric-related tasks ($+27.53\%$, $+16.76\%$, $+21.43\%$, $+28.19\%$) and 16.1% on all exocentric-related tasks ($+17.73\%$, $+16.07\%$, $+17.42\%$, $+13.2\%$).

5. More Qualitative Results

In Fig. 2, we provide four additional qualitative results of our model's predictions to further illustrate its performance. In each column, all 6 visualizations come from consecutive input frames of the same validation instance that spans 3 seconds. For each visualization, we present the raw visual input, predictions from our model, and the ground-truth G_{Conv}. For each sequence, the ground-truth G_{Conv} changes 1-3 times, resulting in 2-4 evolutions of the conversational graph. Our prediction is able to accurately capture this very challenging graph evolution behavior. When the graph suddenly changes with a drastic difference (in Fig. 2(d)), our model fails to capture all changes but is still able to capture most of them while producing some wrong guesses.

In Fig. 3, we provide qualitative results from ASL+Layout baseline of Fig. 2(d).

6. Limitation and Future Work.

Our current available dataset does not include any complex group dynamics such as free-standing or walking scenario, or splitting and merging behaviors of conversational groups. With additional efforts on generating annotations, our task can further extend to large-scale dataset like Ego4D. It is also possible to include Natural language processing (NLP) module for context and intention detection in a concurrent conversational setting.

7. Demo Video

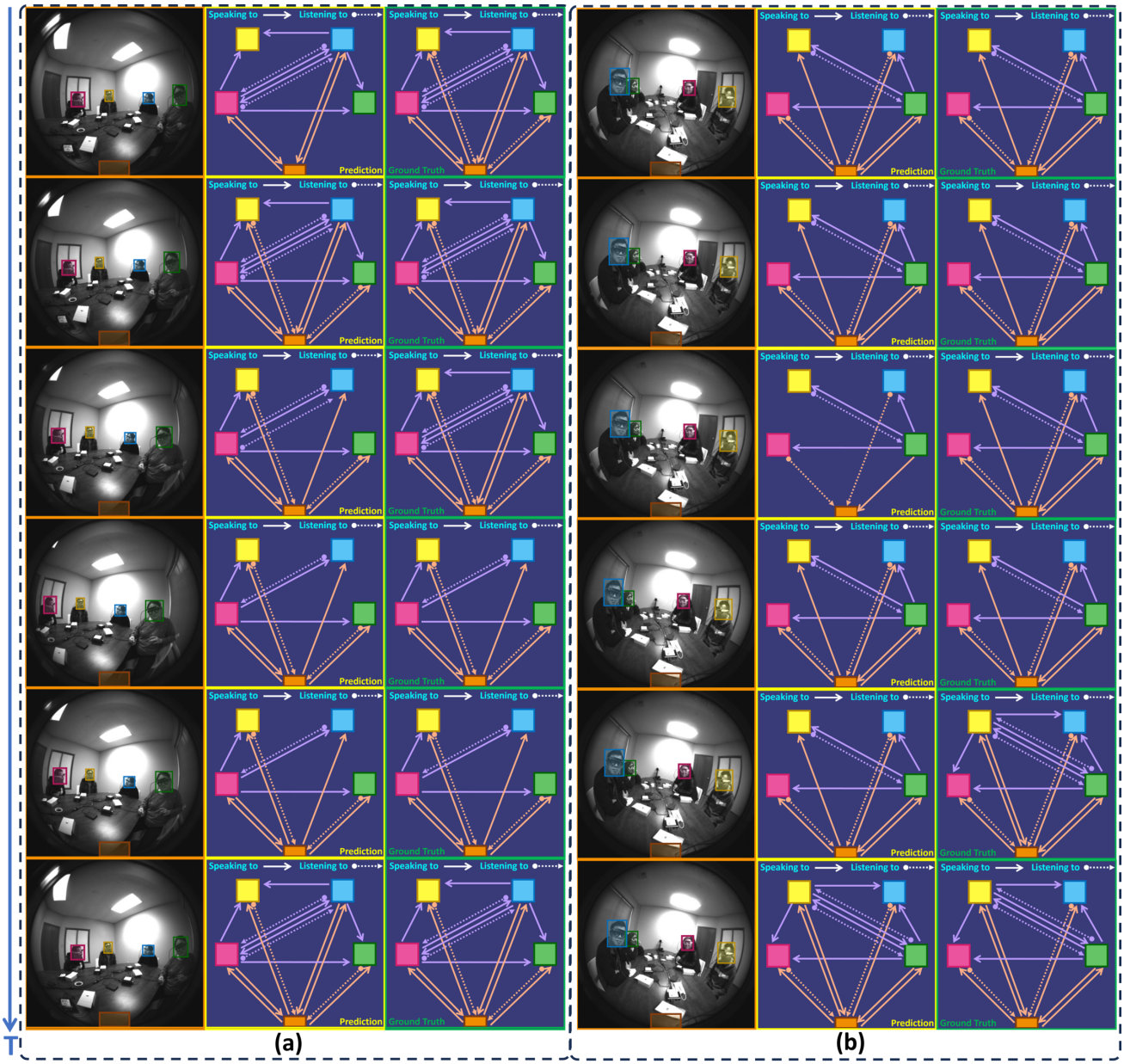
We include videos featuring demonstrations of 5 together with the *Egocentric Concurrent Conversations Dataset* and the code on our AV-CONV project page.

	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	84.56	61.96	46.42	54.21	62.12	64.93	58.19	17.26	18.31	40.08
AV-CONV (T)	83.75	67.78	56.15	57.83	66.38	70.41	**64.50**	21.89	22.06	44.71
AV-CONV (N)	83.24	63.59	52.73	57.35	64.23	66.46	59.25	23.50	25.11	43.58
AV-CONV (S)	83.85	61.16	47.48	54.42	61.73	64.41	56.72	20.05	20.47	40.41
AV-CONV (TN)	84.71	66.80	55.12	56.65	65.80	68.78	63.51	24.83	**30.21**	46.83
AV-CONV (TS)	**84.92**	67.04	54.02	58.46	66.11	70.00	63.32	21.64	23.42	44.60
AV-CONV (NS)	84.41	63.49	53.22	55.10	64.05	66.13	59.31	22.05	22.45	42.49
AV-CONV	82.08	**68.94**	**60.70**	**65.48**	**69.30**	**72.73**	63.36	**32.35**	29.29	**49.43**

Table 1. **Ablation Study on Conversational Attention in mAP**. As described in the main paper, we report the classification mAP results.

	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**

Table 2. **Modality Ablation in mAP**. As described in the main paper, we report the classification mAP results.


Speaking to
Listening to
Prediction
Ground Truth
T
(a)
(b)

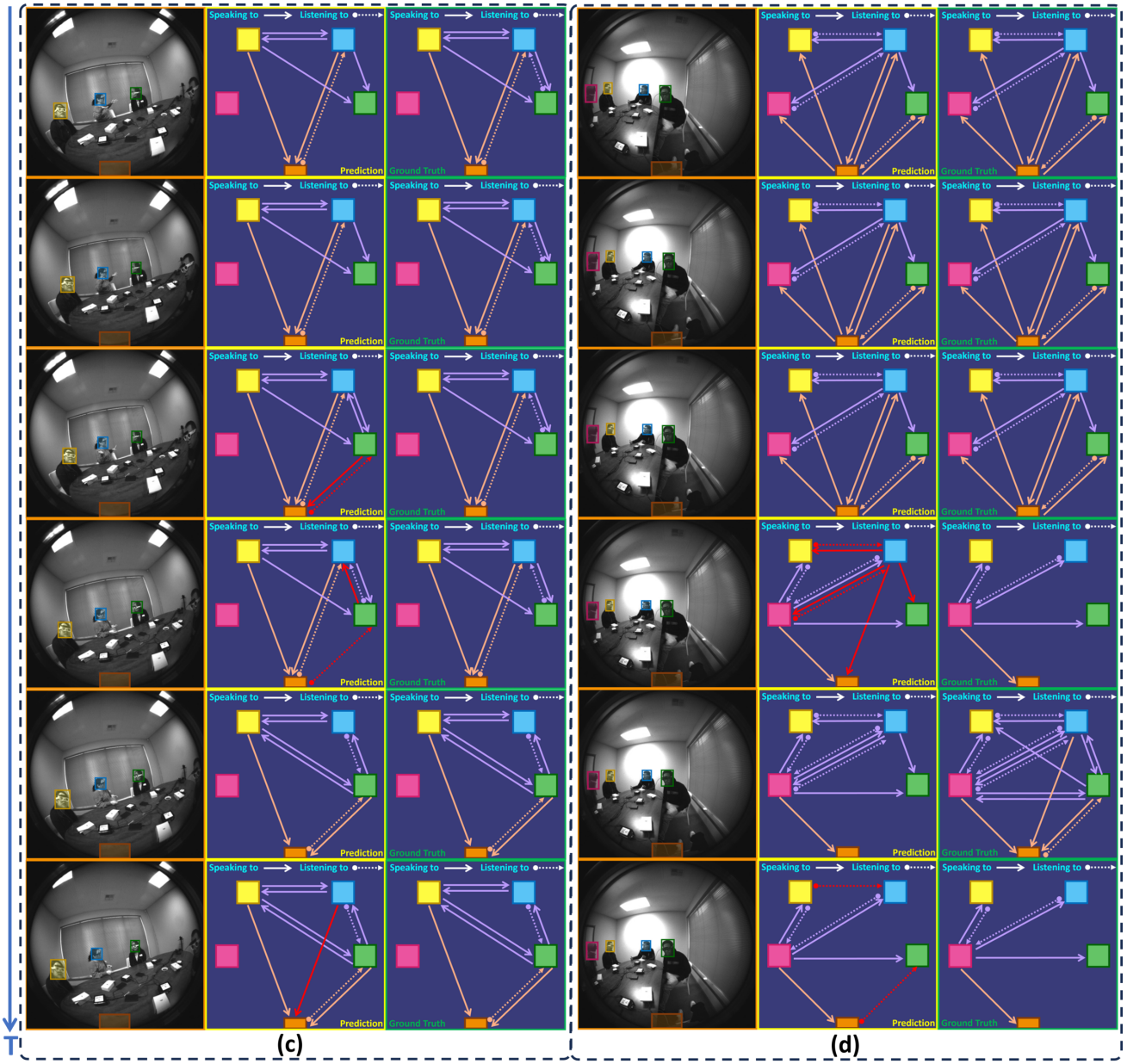


Figure 2. Visualization of the Ego-Exocentric Conversational Graph from our model prediction.

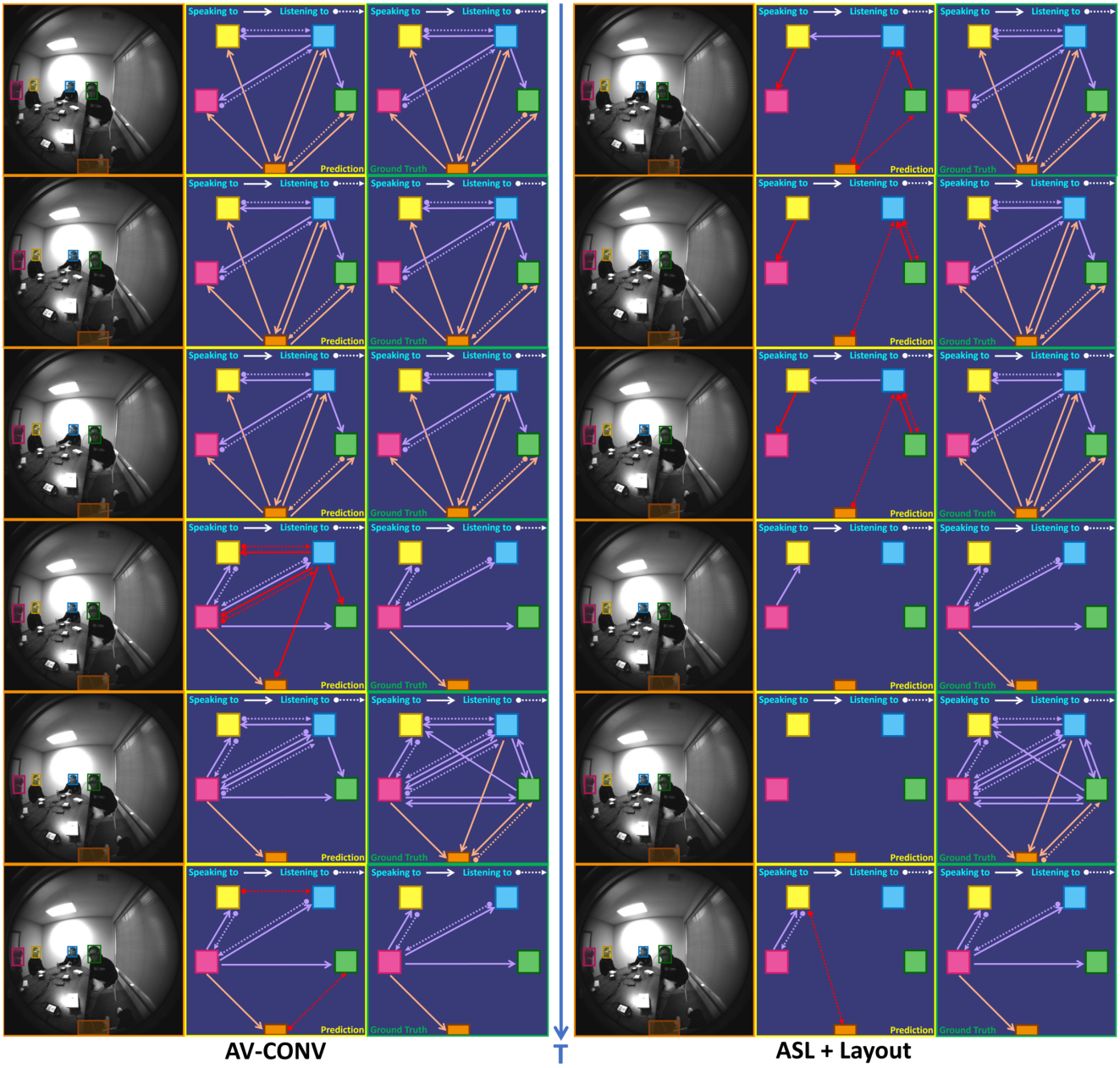


Figure 3. For (d) in Fig. 2, we additionally provide visualization with the prediction results using the ASL+Layout baseline to demonstrate the superiority of our AV-CONV model.