Supplementary Materials: Multi-task Gaze Communication Understanding

Anonymous Author(s)
Submission Id: 311

1 Overview

The supplementary material complements Section 5.6 of the main paper, which details the ablation studies.

2 Ablation Study on VACATION Dataset

Baselines Following the settings on the GP-static++ dataset, we use the following baselines: (i) *Ours w/o. ITA*: We removed the Interaction Aware Module. (ii) *Ours w/o. GF*: We removed the gaze target estimation branch. (iii) *Ours Atomic-MLP*: The face tokens outputted by the Interaction-Aware Module were fed into an MLP for atomic-level classification instead of utilizing the text encoder. (iv) *Ours w/o. Event*: We removed the event-level prediction during atomic-level prediction. (v) *Ours w/o. Atomic*: We removed the atomic-level prediction head during event-level prediction.

Results The ablation results are shown in tables 1 and 2. In terms of the metrics of Avg. Acc., except the Ours w/o. GF, the performance of other ablation baselines show a clear decrease which is consistent with the results shown on the GP-Static++ dataset. The performance of the Ours w/o. GF shows a slight increase. However, it should be noted that the VACATION dataset does not have detailed gaze target estimation annotation. The gaze target estimation module only performs when the interactive objects are annotated. However, the annotated interactive objects are limited in the dataset, which may hinder the model's performance.

In terms of precision and F1-score for atomic-level prediction, all baseline models fail to accurately classify the refer and follow gaze behaviours. This suggests that these classes depend heavily on contextual information from other gaze behaviours—an aspect that the Interaction-Aware Module (ITA) effectively captures. Additionally, refer and follow gaze rely strongly on gaze target information, as these behaviours typically involve interactive objects.

Furthermore, event-level features play a crucial role in the classification of refer and follow. Since gaze-following is a higher-order phenomenon spanning multiple atomic behaviours, a model without event-level context struggles to distinguish these classes. Finally, replacing the text encoder with an MLP significantly weakens the model's ability to leverage event-level information, likely causing it to overfit dominant classes while failing to recognize minority behaviours like refer and follow.

References

[1] Lifeng Fan, Wenguan Wang, Siyuan Huang, Xinyu Tang, and Song-Chun Zhu. 2019. Understanding human gaze communication by spatio-temporal graph reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5724–5733.

Table 1: Quantitative results of atomic-level gaze communication classification on VACATION.

Task	Automic-level Gaze Comminication													
	single		mutual		avert		refer		follow		share		Avg. Acc.	
metrics	$\mathcal{P}$(%)↑	$\mathcal{F}$(%)↑	$\mathcal{P}$(%)↑	$\mathcal{F}$(%)↑	$\mathcal{P}$(%)↑	$\mathcal{F}$(%)↑	$\mathcal{P}$(%)↑	$\mathcal{F}$(%)↑	$\mathcal{P}$(%)↑	$\mathcal{F}$(%)↑	$\mathcal{P}$(%)↑	$\mathcal{F}$(%)↑	top-1(%)↑	top-2(%)↑
Ours	**70.02**	**73.58**	65.43	61.71	43.26	31.88	40.00	10.00	**40.00**	**23.53**	38.24	39.93	**63.36**	**78.02**
Fan [1]	22.10	26.17	**98.68**	**98.60**	59.20	74.20	**56.90**	**53.16**	32.83	18.05	**61.51**	46.61	55.02	76.45
Chance [1]	16.50	16.45	16.42	16.65	16.65	16.51	16.07	16.06	16.80	16.74	16.20	16.25	16.44	-
CNN [1]	21.30	27.89	15.99	14.48	47.81	50.82	0.00	0.00	19.21	23.10	11.70	2.80	23.05	40.32
CNN-LSTM [1]	22.10	11.78	18.55	16.37	64.24	59.57	13.69	18.55	22.70	29.13	17.18	3.61	24.65	45.50
CNN-SVM [1]	19.92	23.63	28.46	38.30	**68.53**	**76.07**	15.15	6.32	23.28	16.87	40.76	**49.24**	36.23	-
CNN-RF [1]	53.12	57.98	20.78	0.24	0.00	0.00	51.88	27.31	15.90	19.39	35.56	44.42	37.68	-
Ours w/o. ITA	68.53	72.85	64.99	56.12	15.76	17.24	0.00	0.00	0.00	0.00	38.73	34.55	59.33	79.52
Ours w/o. GF	65.02	74.23	62.61	57.83	58.82	6.21	0.00	0.00	0.00	0.00	58.63	40.40	**63.93**	85.52
Ours w/o. Event	64.57	73.13	56.31	51.30	23.44	13.86	0.00	0.00	0.00	0.00	58.43	38.01	61.36	76.33
Ours Atomic-MLP	67.75	71.27	56.97	60.41	51.35	33.55	0.00	0.00	0.00	0.00	39.28	33.19	61.48	78.10

Table 2: Quantitative results of event-level gaze communication classification on VACATION.

Task	Event-level Gaze Communication											
	No-Comm.		Mutual Gaze		Gaze Aversion		Gaze Following		Joint Attention		Avg. Acc.	
Metrics	$\mathcal{P}$(%)↑	$\mathcal{F}$(%)↑	$\mathcal{P}$(%)↑	$\mathcal{F}$(%)↑	$\mathcal{P}$(%)↑	$\mathcal{F}$(%)↑	$\mathcal{P}$(%)↑	$\mathcal{F}$(%)↑	$\mathcal{P}$(%)↑	$\mathcal{F}$(%)↑	top-1(%)↑	top-2(%)↑
Ours	57.1	58.8	**52.9**	**59.0**	**40.0**	21.1	**33.3**	30.1	48.0	**49.0**	51.0	66.7
Fan-w. GT [1]	**91.4**	**72.7**	14.5	32.3	18.5	**45.5**	20.0	**66.7**	**62.2**	30.8	**55.9**	**79.4**
Fan-w/o. GT [1]	50.7	49.3	16.7	21.0	8.2	11.3	6.2	7.7	60.9	40.0	37.1	65.5
Ours w/o. ITA	48.9	57.5	1.00	44.4	22.2	17.4	54.8	58.6	50.0	40.0	49.5	65.7
Ours w/o. GF	53.5	60.5	28.6	28.6	0.00	0.00	54.8	58.6	60.0	54.6	51.4	64.8
Ours w/o. Atomic	51.2	56.8	16.7	15.4	28.6	19.1	52.8	60.3	66.7	51.3	50.5	62.9
Ours Atomic-MLP	47.4	50.7	75.0	54.6	25.0	11.1	48.7	56.3	40.9	39.1	46.7	67.6