RESEARCH-ARTICLE

Participation Role-Driven Engagement Estimation of ASD Individuals in Neurodiverse Group Discussions

KALIN STEFANOV, Monash University, Melbourne, VIC, Australia

YUKIKO I NAKANO, Seikei University, Musashino, Tokyo, Japan

CHISA KOBAYASHI, Seikei University, Musashino, Tokyo, Japan

IBUKI HOSHINA, Seikei University, Musashino, Tokyo, Japan

TATSUYA SAKATO, Seikei University, Musashino, Tokyo, Japan

FUMIO NIHEI, Nippon Telegraph and Telephone Corporation, Tokyo, Japan

View all

Published: 04 November 2024

ICMI '24: INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION
November 4 - 8, 2024
San Jose, Costa Rica

Participation Role-Driven Engagement Estimation of ASD Individuals in Neurodiverse Group Discussions

Kalin Stefanov
Monash University
Australia
Seikei University
Japan
kalin.stefanov@monash.edu

Yukiko I. Nakano
Seikei University
Japan
y.nakano@st.seikei.ac.jp

Chisa Kobayashi
Seikei University
Japan
dm246208@cc.seikei.ac.jp

Ibuki Hoshina
Seikei University
Japan
dm246215@cc.seikei.ac.jp

Tatsuya Sakato
Seikei University
Japan
sakato@st.seikei.ac.jp

Fumio Nihei
NTT Human Informatics Laboratories, NTT Corporation
Japan
fumio.nihei@ntt.com

Chihiro Takayama
NTT Human Informatics Laboratories, NTT Corporation
Japan
chihiro.takayama@ntt.com

Ryo Ishii
NTT Human Informatics Laboratories, NTT Corporation
Japan
ryoct.ishii@ntt.com

Masatsugu Tsujii
Chukyo University
Japan
Seikei University
Japan
masatsugtsujii@gmail.com

Abstract

Adults with autism spectrum disorder (ASD) face difficulties in communicating with neurotypical people in their daily lives and workplaces. In addition, research on modeling communication in neurodiverse groups is scarce. To recognize communication difficulties caused by neurodiversity, we first, collected a multimodal corpus for decision-making discussions in neurodiverse groups that included a person with ASD and two neurotypical participants. For corpus analysis, we investigated eye-gaze and facial expression exchanges between individuals with ASD and neurotypical participants during both listening and speaking. The findings were extended to automatically estimate the engagement of ASD individuals. To capture the effect of contingent behaviors between ASD individuals and neurotypical participants, we developed a transformer-based model that considers the participation role by changing the direction of cross-person attention depending on whether the ASD individual is listening or speaking. The proposed approach yields comparable results to the state-of-the-art for engagement estimation in neurotypical group conversations while accounting for the dynamic nature of behavior influence in face-to-face interactions. The code associated with this study is available at https://github.com/IUI-Lab/switch-attention.

ICMI '24, November 04–08, 2024, San Jose, Costa Rica

ACM ISBN 979-8-4007-0462-8/24/11
https://doi.org/10.1145/3678957.3685721

CCS Concepts

• **Computing methodologies** → Machine learning; • **Human-centered computing** → Collaborative and social computing.

Keywords

Adults with autism spectrum disorder; neurodiverse group communication; engagement estimation; participation role

ACM Reference Format:
Kalin Stefanov, Yukiko I. Nakano, Chisa Kobayashi, Ibuki Hoshina, Tatsuya Sakato, Fumio Nihei, Chihiro Takayama, Ryo Ishii, and Masatsugu Tsujii. 2024. Participation Role-Driven Engagement Estimation of ASD Individuals in Neurodiverse Group Discussions. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '24), November 04–08, 2024, San Jose, Costa Rica.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3678957.3685721

1 Introduction

Autism spectrum disorder (ASD) is a developmental disability resulting from differences in the brain. It is characterized by persistent deficits in social communication and social interaction across multiple contexts, including deficits in social reciprocity, nonverbal communicative behaviors used for social interaction, and skills in developing, maintaining, and understanding relationships [2]. European surveys show that 62% of individuals with autism have never had employment [9], suggesting that adults diagnosed with ASD (henceforth referred to as ASD individuals) often face communication problems in their daily life and the workplace.

However, the term neurodiversity, coined by Singer [36], rejects the idiom of impairment. It attempts to promote an understanding of alternative cognitive styles, while the term neurotypical was introduced as a counterpart to describe non-neurodiverse individuals in society [9]. Under the concept of neurodiversity, ASD individuals

use communication signals differently from neurotypical ones, and this difference is not a deficit. However, in a society where neurotypical individuals form the majority, ASD individuals are required to join conversations dominated by neurotypical individuals and communication signals are perceived from the perspective of neurotypical individuals. This cause communication difficulty for ASD individuals.

As the first step towards better mutual understanding between different neurotypes, this study aims to provide a technology that automatically detects social signal exchanges that may cause a communication problem in neurodiverse group communications. To understand and assess traits of people with ASD, previous studies have analyzed verbal and nonverbal behaviors, including intonation and vocalization [1], eye-gaze behaviors (e.g., mutual gaze, joint attention) [37], and facial expressions [15]. However, existing studies have not investigated group communications involving both ASD individuals and neurotypical persons. Envisioning new research field of neurodiverse group interactions, this study aims to analyze interactions in which ASD individuals communicate with multiple neurotypical persons.

To automatically recognize whether connections between neurodiverse participants are maintained during interactions, we focus on estimating the social engagement of individuals with ASD, defined as any interaction a human has with another human in a social group [26, 35]. Various machine learning methods have been proposed for social engagement estimation in human-human communication and human-robot interaction [6, 7]. Most of the proposed engagement estimation methods for multiparty communication are based on eye-gaze information [27], facial features [19, 39], and audio/visual features [22]. Accordingly, we analyzed neurodiverse group interactions using information from multiple modalities.

We collected a corpus from small group interactions, where one ASD individual and two neurotypical persons discussed to make a decision. We then analyzed the eye-gaze and facial expressions of participants in the collected corpus. In estimating the engagement of ASD individuals, this study employed cross-person attention mechanisms [19]. Our transformer-based architecture leverages multimodal information and captures co-constructed context between neurodiverse participants by switching the direction of the cross-person attention according to the participation role.

The contributions of this study are summarized as follows.

- We constructed a corpus capturing communication in neurodiverse groups comprising an individual with ASD and two neurotypical individuals.
- We revealed the influence of neurodiversity and participation roles on social signals by analyzing eye-gaze and facial expressions in the collected corpus.
- We developed a social engagement estimation model for ASD individuals by employing a transformer architecture with cross-person attention mechanisms that capture interactions between participants considering the participation role (speaker or listener).

2 Background

2.1 Social Signal Display by ASD Individuals

In autism research, there have been many studies that compared communication behavior of ASD and neurotypical individuals. A survey paper [1] analyzed the prosodic features of ASD and neurotypical individuals and discussed that mean pitch value and pitch range can be considered a reliable feature to distinguish ASD from neurotypical individuals. For eye-gaze behavior, it is known that people with ASD have difficulty in mutual gaze or eye contact during face-to-face communication. For facial expressions, ASD individuals display less frequent facial expressions, and neurotypical individuals have difficulties recognizing facial expressions by ASD individuals and vice versa [15]. As a study focusing on the interaction dynamics between ASD and neurotypical individuals, Warlaumont et al. [42] analyzed naturalistic recordings of child-adult vocal exchanges, and reported that the strongest trend concerned the adult's responses to the child rather than the child's responses to the adult. The length of time before an adult responded to an ASD child's speech was larger in ASD than for neurotypical children.

Based on these findings, past research in social signal processing has attempted to automatically detect people with strong autistic traits using machine learning techniques [17, 32]. However, no other studies have addressed the engagement estimation of ASD individuals in neurodiverse group interactions.

2.2 Engagement Estimation

There is a large number of studies concerned with engagement estimation. Various types of engagement have been discussed, including social, task, and emotional engagement [26, 39, 41]. Some studies of group interactions addressed group-level engagement [12, 27, 33] as well as the engagement of individual participants. Aiming to estimate the attitude of ASD individuals in group interactions, our study addresses the social aspect of engagement (social engagement) while considering ASD individuals (individual engagement). Based on the discussion by Oertel et al. [26], which surveyed various definitions of social engagement [23, 29, 35], our study defined social engagement as "any interaction a human has with another human in a social group".

Various features for engagement estimation have been examined. Some studies reported that eye-gaze or head pose are good predictive features [4, 14, 24, 27]. In child-robot interaction, it was found that social engagement can be modeled as a state consisting of affect and attention components in the context of the interaction [6]. Physiological reactions such as heart rate and electrodermal activity can be used to predict engagement with an agent [8].

Most recent research in engagement estimation employs deep neural networks. CNN and LSTM were used to model facial features extracted from videos [10, 39]. Rudovic et al. [30] proposed a multimodal method for engagement estimation that combines body, face, audio, and autonomic physiology data of children with autism during robot-assisted autism therapy. Later, they extended their study to a deep reinforcement learning architecture [31]. Studies more closely related to ours proposed transformer-based engagement estimation methods. Ma et al. [21] used transformer architecture to assess individual student learning engagement and Lee et al. [19] addressed video-based engagement estimation in group settings using a cross-person attention mechanism to model contingent behavior between pairs of participants. Kim et al. [16] proposed an architecture that combines cross-modal attention [40] and cross-person attention. However, these studies did not consider the participation role in modeling group interactions.

Figure 1: Snapshot of a corpus collection session.

Table 1: Characteristics of the collected corpus.

Total number of groups	11
Total number of conversations	22
Average length of a conversation	14 min 41 s
Total number of utterances (VADs)	12016
Total number of turns	8311
Average duration of an utterance	1.6 s
Average duration of a turn	2.7 s

3 Neurodiverse Group Discussions Corpus

To analyze and model neurodiverse interactions, we first collected data related to group communication between individuals with ASD and neurotypical individuals.

3.1 Procedure

With the cooperation of a nonprofit organization that supports people with ASD, 11 ASD individuals and 5 neurotypical nonprofit organization members (henceforth referred to as supporters) with experience in supporting them participated in data collection. In addition, 22 neurotypical people were recruited from the general public. The average age of individuals with ASD was 29.6 years old (10 male and 1 female), that of supporters was 31.2 years old (4 male and 1 female), and that of the public participants was 31.7 years old (11 male and 11 female).

Every ASD individual participated in two consecutive sessions, namely, one session communicating with two supporters and another session communicating with two public participants. Each discussion group consisted of one ASD individual and two neurotypical participants, and 22 neurodiverse group discussions were conducted: 11 conversations between ASD individuals and supporters and 11 conversations between ASD individuals and the public participants. Some supporters participated in multiple sessions, while each public participant participated in only one session. All the participants were native Japanese speakers.

The discussion task was designed based on the MATRICS corpus [25] comprising three group decision-making tasks. We used two tasks from the MATRICS corpus: guest selection and travel planning. For the guest selection task, given a list of 10 celebrities, the participants were asked to decide the ranked order of celebrities to be invited to an event considering audience attraction. For the travel planning task, the participants were instructed to create a two-day travel plan for foreign friends visiting Japan on vacation. The order of the tasks and conditions of the neurotypical group members (supporter/general public) were randomized.

The study was approved by our ethics review committee. The individuals with ASD understood how to cooperate with data collection, and all participants provided their consent for data collection.

3.2 Data

Video: A snapshot of corpus recording is shown in Figure 1. A webcam and video camera were used to record the face of each participant. The webcam was placed on top of the video camera. Thus, the positions and angles of the two cameras were almost the same. The webcams recorded 1280×720 MP4 videos, and the video cameras recorded 1440×1080 MTS videos. An additional video camera was used to record an overview of the communication. The frame rate of all the cameras was 30 frames per second.

Audio: Each participant wore a pin microphone to record individual speech at a sampling rate of 48000 Hz. Voice activity detection (VAD) program was applied to individual audio recordings to detect speech intervals. The amplitude-level threshold was set to 400. At this threshold, if the silent interval was longer than 200 milliseconds, it was identified as an utterance boundary. A sequence of utterances by the same speaker was identified as a turn.

Language: VAD as an utterance was transcribed using Google ASR. The recognition errors in the automatic transcription were manually corrected.

The basic characteristics of the collected corpus are listed in Table 1. The average length of the 22 conversations was 14 minutes and 41 seconds, and the corpus contained 12016 utterances and 8311 turns. The average duration of the utterance was 1.6 seconds, and the average turn length was 2.7 seconds.

3.3 Engagement Annotation

The collected corpus was annotated for social engagement of ASD individuals. Social engagement was rated in terms of the following aspects: attitude of participating in the discussion as a listener (e.g., no reaction/response to other participants' speech, distractive actions such as yawing and stretching) and attitude as a speaker (e.g., keep talking without being conscious of others' reactions, talking to himself/herself, not looking at others, completely looking down during speaking). In addition, if the ASD individual did not appear to be participating in the discussion at all, that attitude was also rated as disengagement.

Based on the abovementioned criteria, three levels of disengagement (0, 1, and 2) were annotated. When the ASD individual displayed disengagement behaviors, disengagement level 2 was assigned as the annotation. When an ASD individual did not display such behavior, a disengagement level of 0 is assigned. A disengagement level of 1 was assigned when a weak indication of these types of behaviors appeared. The disengagement levels of ASD individuals were annotated for all utterances, including their own. Two annotators performed the annotations. They rated the disengagement levels using the ELAN video annotation tool while looking at the overview video with a close view of the upper body of each participant.

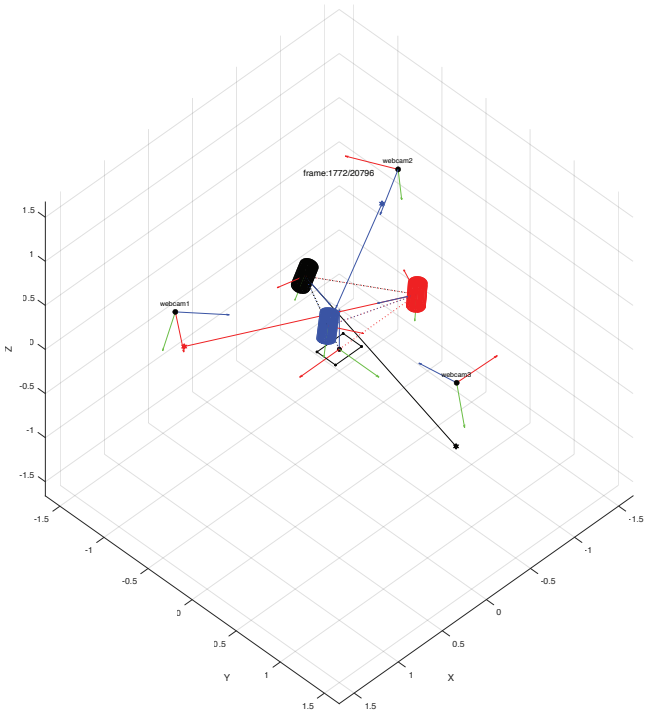

Figure 2: 3D head pose reconstruction for one frame.

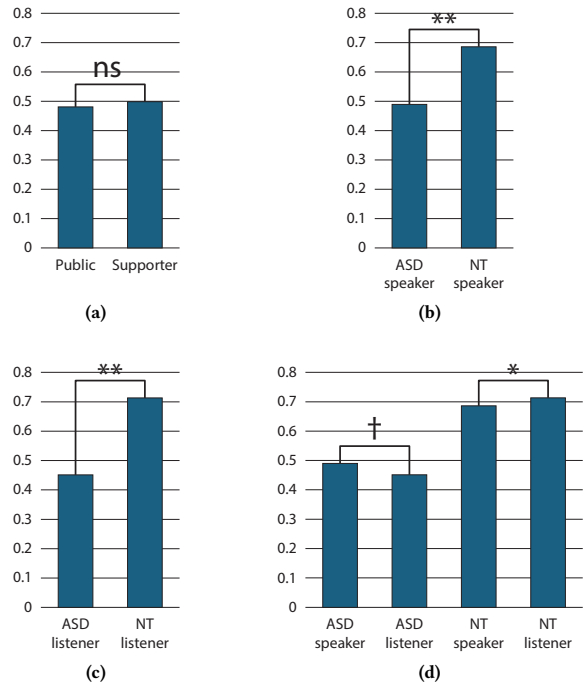

Figure 3: Average gaze ratios of participants.

The two annotators independently labeled a couple of sessions, discussed the samples for which they did not agree, and refined the coding scheme to reach a consensus on the annotation criteria. By merging levels 0 and 1 into "engaged" and labeling level 2 as "disengaged," the inter-rater agreement for the two classes was computed, indicating a substantial agreement between the annotators (Cohen's Kappa = 0.65). To ensure the dataset's reliability, the two annotators individually labeled all samples, and then discussed any disagreements to decide on the final labels.

4 Data Analysis

4.1 Gaze Behaviors

The participants' gaze behavior was analyzed using the head pose generated with OpenFace [3]. Given that information and a camera calibration procedure, the head pose information obtained with OpenFace was brought to a shared 3D space. Figure 2 illustrates the head pose reconstruction for one frame using that technique. Given the unified 3D representation of the head pose of all participants, we employed the discrete head pose encoding technique described in [38] to estimate the visual attention target for each participant.

The gaze ratio during speaking was defined as the proportion of frames in which a participant was looking at others to the total number of frames during speaking. Similarly, the gaze ratio during listening was computed as the proportion of frames in which the target participant gazed at the speaker compared with the number of frames acquired during their listening time. In three-party conversations, a participant can take one of three participation roles, namely, speaker, addressee, and side participant. However, it is difficult to automatically distinguish the addressee and side participant. Sometimes, both participants may be addressees. Therefore, the addressee and side participants were not distinguished, and the two non-speakers were simply considered as listeners in this study.

Figure 3a shows the average gaze ratio that ASD individuals looked at other participants when they were speakers. The ratios of gazing at the other participants were 0.499 and 0.48 for the supporter sessions and the general public sessions, respectively. No significant differences were observed between those sessions. Thus, we treated both supporters and participants recruited from the general public as neurotypical participants and conducted the subsequent analyses.

Figure 3b shows the average proportion of ASD participants gazing at neurotypical participants while speaking (0.490), and that of neurotypical participants (NT speaker) gazing at other participants while speaking (0.686). The result of the t-test was statistically significant ($t = -3.535$, $p < .001$), indicating that ASD individuals were paying significantly less attention to other participants while speaking compared to neurotypical participants.

Figure 3c shows similar analysis results for the listener case. When the ASD participants were listeners, the mean ratio of gazing at other participants was 0.451, compared with 0.713 for neurotypical participants (NT listener). Again, the t-test showed a statistically significant difference between the two groups ($t = -4.904$, $p < .001$), indicating that participants with ASD paid significantly less attention to other participants compared to neurotypical participants, even when they were listeners. Furthermore, Figure 3d left shows a comparison of the percentage of attention paid to others when participants with ASD were speakers (0.490) and listeners (0.451). The difference was marginally significant in the t-test, indicating that ASD individuals tended to show a lower gazing rate when

Table 2: Results of facial expression analysis.

	Gaze from	Gaze to	AU01	AU04	AU06	AU07	AU09	AU10	AU12	AU17	AU20	AU23	AU25	AU26	AU45
(a)	NT speaker	ASD listener	0.362	**0.453***	0.304*	0.647†	0.082**	0.553	0.562**	0.505	**0.129**†	0.133	0.891**	0.582**	0.234**
(b)	NT listener	ASD speaker	0.414	0.397	0.249	0.584	0.061	0.506	0.479	0.535	0.111	0.140	0.720	0.447	0.157
(c)	ASD speaker	NT listener	0.243	0.262	0.354**	0.707**	0.086**	0.391**	0.320**	0.408	0.124	0.149	0.735**	0.443	0.168*
(d)	ASD listener	NT speaker	**0.362***	0.221	0.296	0.527	0.059	0.321	0.268	0.469*	0.126	**0.200***	0.588	0.390	0.134
(e)	NT speaker	NT listener	0.360	0.444	0.325†	0.650*	0.102*	0.549*	0.571**	0.474	0.138	0.175	0.810**	0.598**	0.223†
(f)	NT listener	NT speaker	0.405	0.434	0.271	0.584	0.081	0.463	0.475	0.555**	0.153	0.159	0.636	0.501	0.181

they were listeners than when they were speakers. In contrast, neurotypical participants gazed at others more when they were listeners (0.713) than when they were speakers (0.686) (Figure 3d right), showing a statistically significant difference between these roles ($t = -2.421$, $p < .05$).

The results suggest that even if neurotypical participants as speakers pay attention to ASD individuals, there is a possibility that ASD participants may not look at the speaker and may not provide sufficient feedback, which is necessary for social engagement.

4.2 Facial Expressions

We also analyzed the facial expressions considering a combination of gaze direction and participation role. The results of the analysis are listed in Table 2. For example, the data in row (a) indicate the facial expression of a neurotypical participant as a speaker (NT speaker) gazing toward an ASD individual as a listener (ASD listener), and those in row (b) indicate the facial expression of a neurotypical participant as a listener (NT listener) gazing toward an individual with ASD as a speaker (ASD speaker). By comparing rows (a) and (b), we investigated whether the facial reactions of neurotypical participants while looking at a person with ASD differed depending on their participation role. Comparisons were made for each action unit (AU). In addition, because there were two neurotypical participants per group, we first extracted the data for the combination of neurotypical participant 1 (NT1) and the participant with ASD as well as those of neurotypical participant 2 (NT2) and the participant with ASD, and then merged these data. Similarly, a comparison between rows (c) and (d) allowed us to investigate whether the facial reaction of the participants with ASD while looking toward a neurotypical participant differed depending on their participation role. For comparison with pairs including the individuals with ASD, rows (e) and (f) provide the characteristics of facial reactions occurring between neurotypical participants.

For AU6, AU7, AU9, AU10, AU12, AU25, AU26, and AU45, the speakers' average AU values were larger than those of the listeners in all three combinations (rows (a)-(b), (c)-(d), and (e)-(f)). These AUs were assumed to be speech-induced facial motions expressed regardless of who was looking at whom. By contrast, the comparison between rows (a) and (b) shows that the average values of AU4 and AU20 expressed by NT speakers are larger than those expressed by NT listeners. For AU4, the difference was statistically significant (NT speaker, 0.453; NT listener, 0.397). For AU20, the difference was marginally significant (NT speaker, 0.129; NT listener, 0.111). According to [11], AU4 expresses sadness, fear, and anger, whereas AU20 expresses fear. By comparing rows (c) and (d), ASD individuals expressed AU1 and AU23 towards a neurotypical participant more strongly when they were listeners (ASD listener) than when they were speakers (ASD speaker). The difference was statistically significant for AU1 (ASD speaker, 0.243; ASD listener, 0.362) and AU23 (ASD speaker, 0.149; ASD listener, 0.2). In [11], AU1 expresses fear and surprise, and AU23 expresses anger, suggesting that for pairs including ASD individuals, the intensity of facial features expressing sadness, fear, and anger are different depending on their participation roles.

A comparison between rows (e) and (f) shows no specific facial expressions for the neurotypical pairs. Notably, for AU6, the difference was marginally significant, whereas it was statistically significant for the other combinations (rows (a)–(b) and (c)–(d)). This was because the difference in the AU6 intensity between speakers and listeners became smaller in neurotypical pairs, suggesting that the speaker and listener exchanged smiles with similar intensity. Moreover, in rows (d) and (f), the AU17 value for the listener was larger than that for the speaker. In appraisal theory [13], it was discussed that AU17 is expressed in obstructive situations, suggesting that the listeners expressed their evaluation of the situation when it was difficult to continue discussing or reach a decision.

Overall, we found that the participants' facial expressions differed depending on whether they were speakers or listeners. It was also found that some specific AUs occurred only in the interaction between ASD individuals and neurotypical participants. These results suggest that it is crucial to consider the factors of participation role in creating models to estimate the engagement of ASD individuals based on facial features.

5 Engagement Estimation

5.1 Modalities and Features

Two modalities capturing the nonverbal behavior of the participants were analyzed. Videos captured by front-facing cameras were used to track facial expressions and audio recorded by pin microphones was used for speech prosody analysis.

Advances in self-supervised learning have dramatically improved the quality of speech representations. For each utterance in the corpus, we extracted Distilled Universal Paralinguistic Speech Representations (TRILLsson) [34] features to capture the speech prosody. TRILLsson is a collection of efficient state-of-the-art paralinguistic speech models based on knowledge distillation. We used the 1024-dimensional embedding generated by TRILLsson3.

For each utterance in the corpus, we also extracted Masked Autoencoder for Facial Video Representation Learning (MARLIN) [5] features to capture facial expressions. MARLIN is a self-supervised

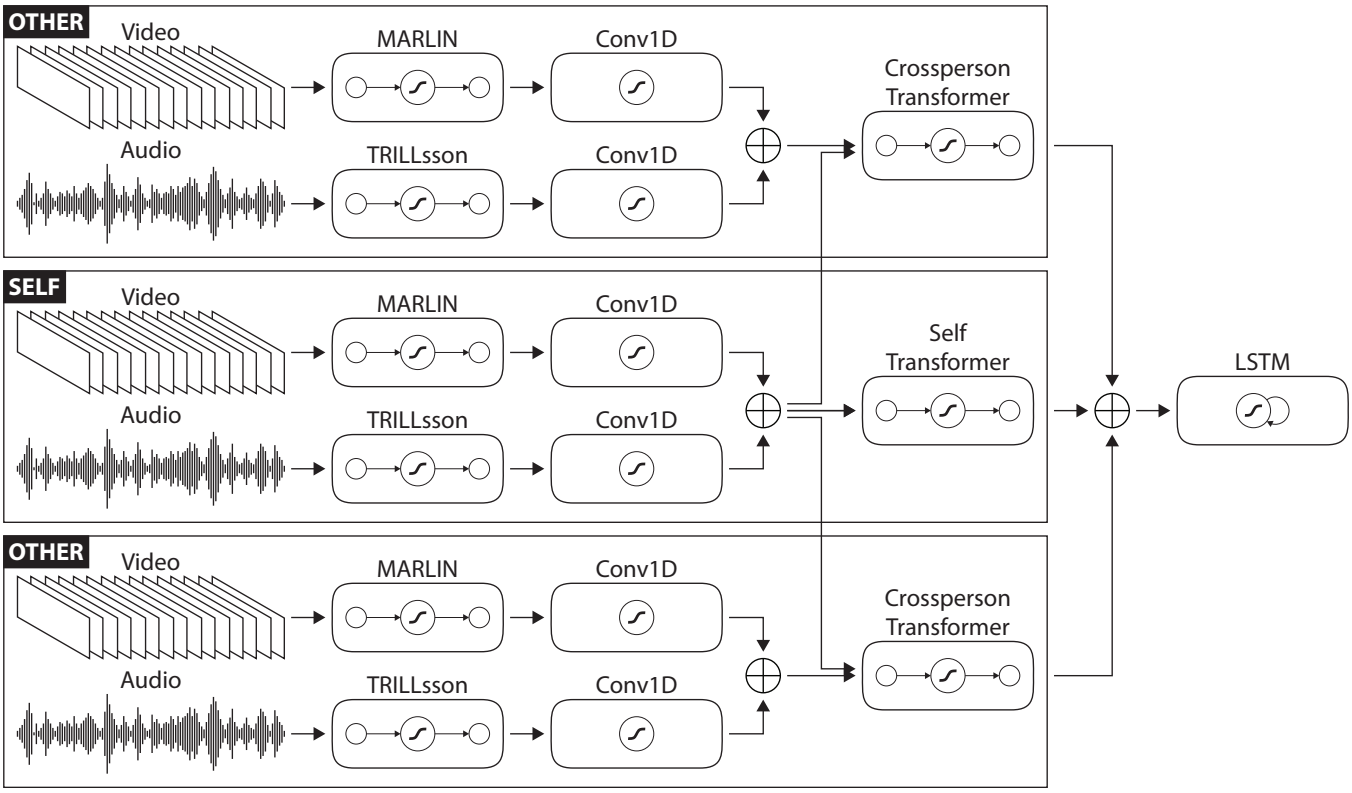

Figure 4: Three different modeling approaches for engagement estimation were considered in this study: 1) *self-attention* uses only the middle data stream in the architecture, 2) *cross-attention* uses all data streams and implements (other → self) cross-attention, and 3) *switch-attention* uses all data streams and implements (other → self) cross-attention when self is speaking and (self → other) cross-attention when self is listening. ⊕ denotes concatenation.

approach to learn state-of-the-art universal facial representations from videos that can transfer across a variety of facial analysis tasks. We used the 768-dimensional embedding generated by MARLIN.

5.2 Modeling Approaches

We formulated the engagement estimation as a binary classification task. Following the state-of-the-art in engagement estimation, we developed 3 transformer-based approaches for modeling the behavior of ASD individuals in neurodiverse group discussions: 1) *self-attention*, 2) *cross-attention* and 3) *switch-attention* (Figure 4). Self-attention accounts for how one's earlier behavior correlates with their current behavior without taking into account the behavior of the other participants. Cross-attention is based on a recently proposed state-of-the-art method for modeling engagement in group conversations [19] that uses cross-attention to capture contingent behavior across pairs of people. This method implements the idea that a target person's self behavior is contingent on the other's person behavior if the person's self behavior was likely to be influenced by the other's behavior (other → self). Switch-attention builds upon the cross-attention above by implementing the idea that behavior influence is a two-way process. This approach implements switching of the cross-attention direction driven by the voice activity of the person. If the target person is speaking, similar to the above formulation, the self's behavior is likely to be influenced by the other's behavior (other → self). However, if the target person is listening, the self's behavior is likely to influence the other's behavior (self → other).

6 Experiments

We conducted subject-dependent and subject-independent experiments with the corpus described in Section 3, the 3 modeling approaches described in Section 5, and 3 different context lengths (number of previous utterances). The goal of the subject-dependent experiment is to test to what extent the modeling can estimate the level of engagement when data from the same participants are considered. We used a 10-fold cross-validation procedure to train and evaluate the models. During training, we held out ≈ 20% of the train data as a validation set. The goal of the subject-independent experiment is to test to what extent the modeling approaches can generalize to unseen data and participants. In this experiment, we trained the models using a leave-one-out cross-validation procedure. During training, we used 8 groups as a train set, 2 groups as a validation set, and the remaining 1 group as a test set.

All models were trained with Adam [18] optimizer with default parameters ($\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$) and focal loss function [20]. The models were trained for 100 epochs

Table 3: Engagement estimation results under different conditions. Subject-dependent models are evaluated with 10-fold cross-validation. The numbers correspond to the mean (standard deviation) over folds.

Evaluation	Subject-dependent								
Context	3 utterances			4 utterances			5 utterances		
Metric	Accuracy↑	Weighted F1↑	Macro F1↑	Accuracy↑	Weighted F1↑	Macro F1↑	Accuracy↑	Weighted F1↑	Macro F1↑
Self-attention	84.4 (1.2)	84.0 (1.4)	68.0 (3.3)	84.1 (1.5)	83.7 (1.5)	67.1 (3.0)	84.1 (1.3)	83.6 (1.7)	67.0 (4.1)
Cross-attention	**84.9 (1.2)**	**84.6 (1.6)**	69.2 (3.7)	**84.3 (1.5)**	83.6 (1.9)	66.6 (2.2)	83.8 (0.8)	83.5 (1.1)	67.1 (3.3)
Switch-attention	84.4 (1.3)	84.3 (1.8)	**69.2 (3.2)**	84.0 (1.3)	**84.0 (1.6)**	**68.6 (2.8)**	**84.9 (1.2)**	**84.3 (1.8)**	**68.0 (4.7)**

Table 4: Engagement estimation results under different conditions. Subject-independent models are evaluated with leave-one-out cross-validation. The numbers correspond to the mean (standard deviation) over folds.

Evaluation	Subject-independent								
Context	3 utterances			4 utterances			5 utterances		
Metric	Accuracy↑	Weighted F1↑	Macro F1↑	Accuracy↑	Weighted F1↑	Macro F1↑	Accuracy↑	Weighted F1↑	Macro F1↑
Self-attention	79.6 (13.2)	76.5 (17.3)	47.5 (4.3)	**79.8 (14.4)**	76.8 (17.9)	**49.0 (5.7)**	78.1 (13.8)	75.2 (18.2)	46.3 (4.5)
Cross-attention	80.4 (14.4)	77.5 (18.0)	48.6 (4.6)	79.6 (14.4)	**76.9 (18.0)**	48.0 (4.2)	**80.1 (13.6)**	**76.9 (16.7)**	**47.4 (3.2)**
Switch-attention	**81.1 (14.3)**	**78.1 (17.4)**	**49.1 (4.4)**	76.5 (15.3)	74.7 (17.6)	46.9 (4.5)	77.5 (15.0)	74.9 (18.3)	47.2 (5.7)

using the labeled utterances and the models' states achieving the best validation performance were selected for evaluation on the test set. The models are evaluated on the test set in terms of accuracy, weighted F1, and macro F1 between the estimations and the labels. All models were implemented in PyTorch [28].

7 Results

7.1 Subject-dependent

We have summarized the numerical results of the subject-dependent experiment in Table 3. The mean and standard deviation of the performance over 10 folds (i.e., 10-fold cross-validation) are reported for different context windows and evaluation metrics. The best overall results for all modeling approaches are reached for context windows of size 3. From those, the *cross-attention* modeling approach achieves the best performance for two of the considered metrics. Specifically, 84.9 and 84.6 for accuracy and weighted F1, respectively. The proposed *switch-attention* reaches the best performance in terms of macro F1, 69.2. Given the small differences and variations in performance across evaluations, none of the considered modeling approaches emerged as a leading candidate for modeling the engagement of ASD individuals in neurodiverse group discussions.

7.2 Subject-independent

The numerical results of the subject-independent experiment are summarized in Table 4 by reporting the mean and standard deviation of the performance over 11 folds (i.e., leave-one-out cross-validation) for different context windows and evaluation metrics. The best overall results for all modeling approaches are again reached for context windows of size 3. From those, the proposed *switch-attention* modeling approach achieves the best performance for all considered metrics. Specifically, 81.1, 78.1 and 49.1 for accuracy, weighted F1 and macro F1, respectively. Similarly to the results of the subject-dependent experiment, we observed small differences and variations in performance across evaluations of the considered modeling approaches. This experiment validated that none of the modeling approaches emerges as a leading candidate for modeling the engagement of ASD individuals in neurodiverse group discussions.

8 Discussion

Our experiments show that engagement estimation of ASD individuals in neurodiverse group discussions is a challenging task. The state-of-the-art modeling approach (i.e., *cross-attention* in our experiments) proposed in [19] was reported to reach 88.8, 88.7, and 75.1 in accuracy, weighted F1 and macro F1, respectively, on a 4-class engagement estimation for interactions involving only neurotypical participants. However, we observed a significant drop in the performance of this approach when applied to the 2-class engagement estimation of ASD individuals in neurodiverse group discussions. One reason for this could be that there are significant individual differences between ASD individuals and their nonverbal signals used for social engagement prediction. A similar observation was made by Rudovic et al. [30]. They discussed that many individuals with autism have unusually diverse styles of expressing their affective-cognitive states. To tackle the heterogeneity in ASD individuals, they proposed to create a personalized model for each ASD individual to automatically estimate engagement during robot-assisted autism therapy.

In this study, we proposed the *switch-attention* approach that builds upon the cross-attention [19] by implementing the idea that behavior influence is a two-way process. The results show that this strategy does not produce significantly different performance

compared to cross-attention. Nevertheless, we believe that the proposed switch-attention more naturally represents the dynamics of face-to-face interactions, where behavior contingency is a dynamic two-way interaction. A natural extension of the proposed switch-attention is the use of more nuanced information for the participation roles, taking into account the addressee and side participant (the two non-speakers were considered as listeners).

9 Conclusions and Future Work

In this study, aiming to model the interaction between ASD and neurotypical individuals, we analyzed neurodiverse group interactions. First, we collected decision-making group discussions by neurodiverse groups and constructed a multimodal corpus. Using the collected corpus, we analyzed the gaze direction of the participants and revealed that ASD individuals showed less gaze behavior, especially as listeners. Analysis of facial expressions found that some facial expressions occurred only between ASD participants and neurotypical participants, and facial expressions were different depending on the participation role, i.e., whether the participant was a speaker or a listener.

To capture the effect of contingent behaviors between ASD individuals and neurotypical participants, we developed a transformer-based model that considered the participation role by changing the direction of cross-person attention depending on whether the ASD individual listening or speaking. The proposed approach achieves comparable results to the state-of-the-art in engagement estimation of group conversations, while accounting for the dynamic nature of behavior influence in face-to-face interactions.

Future work includes the extension of our neurodiverse group discussions corpus and the development of more advanced engagement estimation methods for ASD individuals.

Acknowledgments

We would like to acknowledge JSPS KAKENHI Grant Number JP24K02984 and JST Moonshot R&D Grant Number JPMJMS2011 for the support of this work.

References

[1] Seyedeh Zahra Asghari, Sajjad Farashi, Saeid Bashirian, and Ensiyeh Jenabi. 2021. Distinctive prosodic features of people with autism spectrum disorder: a systematic review and meta-analysis study. *Scientific Reports* 11, Article 23093 (2021). https://doi.org/10.1038/s41598-021-02487-6

[2] American Psychiatric Association. 2013. *Diagnostic and statistical manual of mental disorders : DSM-5. — 5th ed.*

[3] Tadas Baltrusaitis, Peter Robinson, and Louis Philippe Morency. 2016. OpenFace: An open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016*. IEEE, 1–10.

[4] Dan Bohus and Eric Horvitz. 2009. Models for Multiparty Engagement in Open-World Dialog. In *Proceedings of the SIGDIAL 2009 Conference*, Patrick Healey, Roberto Pieraccini, Donna Byron, Steve Young, and Matthew Purver (Eds.). Association for Computational Linguistics, London, UK, 225–234. https://aclanthology.org/W09-3933

[5] Zhixi Cai, Shreya Ghosh, Kalin Stefanov, Abhinav Dhall, Jianfei Cai, Hamid Rezatofighi, Reza Haffari, and Munawar Hayat. 2023. MARLIN: Masked Autoencoder for facial video Representation LearnINg. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1493–1504. https://doi.org/10.1109/CVPR52729.2023.00150

[6] Ginevra Castellano, Iolanda Leite, André Pereira, Carlos Martinho, Ana Paiva, and Peter W. Mcowan. 2014. Context-Sensitive Affect Recognition for a Robotic Game Companion. *ACM Trans. Interact. Intell. Syst.* 4, 2, Article 10 (jun 2014), 25 pages. https://doi.org/10.1145/2622615

[7] Ginevra Castellano, André Pereira, Iolanda Leite, Ana Paiva, and Peter W. McOwan. 2009. Detecting user engagement with a robot companion using task and social interaction-based features. In *Proceedings of the 2009 International Conference on Multimodal Interfaces* (Cambridge, Massachusetts, USA) *(ICMI-MLMI '09)*. Association for Computing Machinery, New York, NY, USA, 119–126. https://doi.org/10.1145/1647314.1647336

[8] A. Choi, C. D. Melo, W. Woo, and J. Gratch. 2012. Affective engagement to emotional facial expressions of embodied social agents in a decision-making game. *Comput. Anim. VirtualWorlds* 23 (2012). https://doi.org/10.1002/cav.1458

[9] Nick S. Dalton. 2013. Neurodiversity HCI. *interactions* 20, 2 (March 2013), 72–75. https://oro.open.ac.uk/37774/

[10] Soumia Dermouche and Catherine Pelachaud. 2019. Engagement Modeling in Dyadic Interaction. In *2019 International Conference on Multimodal Interaction* (Suzhou, China) *(ICMI '19)*. Association for Computing Machinery, New York, NY, USA, 440–445. https://doi.org/10.1145/3340555.3353765

[11] Shichuan Du, Yong Tao, and Aleix M. Martinez. 2014. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences* 111, 15 (2014), E1454–E1462. https://doi.org/10.1073/pnas.1322355111 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.1322355111

[12] D. Gatica-Perez, L. McCowan, Dong Zhang, and S. Bengio. 2005. Detecting group interest-level in meetings. In *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, Vol. 1. I/489–I/492 Vol. 1. https://doi.org/10.1109/ICASSP.2005.1415157

[13] Kornelia Gentsch, Didier Grandjean, and Klaus Scherer. 2015. Appraisals Generate Specific Configurations of Facial Muscle Movements in a Gambling Task: Evidence for the Component Process Model of Emotion. *PLoS ONE* 10, 8 (2015), e0135837. https://doi.org/10.1371/journal.pone.0135837

[14] Ryo Ishii, Yukiko I. Nakano, and Toyoaki Nishida. 2013. Gaze awareness in conversational agents: Estimating a user's conversational engagement from eye gaze. *ACM Trans. Interact. Intell. Syst.* 3, 2, Article 11 (aug 2013), 25 pages. https://doi.org/10.1145/2499474.2499480

[15] Connor Tom Keating and Jennifer Louise Cook. 2020. Facial Expression Production and Recognition in Autism Spectrum Disorders: A Shifting Landscape. *Child and Adolescent Psychiatric Clinics of North America* 29, 3 (2020), 557–571. https://doi.org/10.1016/j.chc.2020.02.006 Autism Spectrum Disorder Across the Lifespan: Part II.

[16] Yubin Kim, Dong Won Lee, Paul Pu Liang, Sharifa Alghowinem, Cynthia Breazeal, and Hae Won Park. 2023. HIINT: Historical, Intra- and Inter- personal Dynamics Modeling with Cross-person Memory Transformer *(ICMI '23)*. Association for Computing Machinery, New York, NY, USA, 314–325. https://doi.org/10.1145/3577190.3614122

[17] Young-Kyung Kim, Rimita Lahiri, Md. Nasir, So Hyun Kim, Somer Bishop, Catherine Lord, and Shrikanth S. Narayanan. 2021. Analyzing Short Term Dynamic Speech Features for Understanding Behavioral Traits of Children with Autism Spectrum Disorder. In *Proc. Interspeech 2021*. 2916–2920. https://doi.org/10.21437/Interspeech.2021-2111

[18] D. P. Kingma and J. Ba. 2014. Adam: a method for stochastic optimization. *Computing Research Repository* abs/1412.6980 (2014).

[19] Dong Won Lee, Yubin Kim, Rosalind Picard, Cynthia Breazeal, and Hae Won Park. 2023. Multipar-T: Multiparty-Transformer for Capturing Contingent Behaviors in Group Conversations. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23)*.

[20] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2020. Focal Loss for Dense Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 2 (2020), 318–327. https://doi.org/10.1109/TPAMI.2018.2858826

[21] Jiayao Ma, Xinbo Jiang, Songhua Xu, and Xueying Qin. 2021. Hierarchical Temporal Multi-Instance Learning for Video-based Student Learning Engagement Assessment. In *IJCAI 2021*. 2782–2789.

[22] L. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang. 2005. Automatic analysis of multimodal group actions in meetings. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27, 3 (2005), 305–317. https://doi.org/10.1109/TPAMI.2005.49

[23] Lilia Moshkina, Susan Trickett, and J. Gregory Trafton. 2014. Social engagement in public places: a tale of one robot. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction* (Bielefeld, Germany) *(HRI '14)*. Association for Computing Machinery, New York, NY, USA, 382–389. https://doi.org/10.1145/2559636.2559678

[24] Yukiko I. Nakano and Ryo Ishii. 2010. Estimating user's engagement from eye-gaze behaviors in human-agent conversations. In *Proceedings of the 15th International Conference on Intelligent User Interfaces* (Hong Kong, China) *(IUI '10)*. Association for Computing Machinery, New York, NY, USA, 139–148. https://doi.org/10.1145/1719970.1719990

[25] Fumio Nihei, Yukiko I. Nakano, Yuki Hayashi, Hung-Hsuan Hung, and Shogo Okada. 2014. Predicting Influential Statements in Group Discussions using Speech and Head Motion Information. In *Proceedings of the 16th International Conference on Multimodal Interaction* (Istanbul, Turkey) *(ICMI '14)*. Association for Computing Machinery, New York, NY, USA, 136–143. https://doi.org/10.1145/2663204.2663248

[26] Catharine Oertel, Ginevra Castellano, Mohamed Chetouani, Jauwairia Nasir, Mohammad Obaid, Catherine Pelachaud, and Christopher Peters. 2020. Engagement in Human-Agent Interaction: An Overview. *Frontiers in Robotics and AI* 7 (2020). https://doi.org/10.3389/frobt.2020.00092

[27] Catharine Oertel and Giampiero Salvi. 2013. A gaze-based method for relating group involvement to individual engagement in multimodal multiparty dialogue. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction* (Sydney, Australia) *(ICMI '13)*. Association for Computing Machinery, New York, NY, USA, 99–106. https://doi.org/10.1145/2522848.2522865

[28] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *Autodiff Workshop, NIPS*.

[29] Isabella Poggi. 2007. *Mind, hands, face and body : a goal and belief view of multimodal communication.* Weidler.

[30] Ognjen Rudovic, Jaeryoung Lee, Miles Dai, Bjorn Schuller, and Rosalind W. Picard. 2018. Personalized machine learning for robot perception of affect and engagement in autism therapy. *Science Robotics* 3, 19 (2018), eaao6760. https://doi.org/10.1126/scirobotics.aao6760 arXiv:https://www.science.org/doi/pdf/10.1126/scirobotics.aao6760

[31] Ognjen Rudovic, Meiru Zhang, Bjorn Schuller, and Rosalind Picard. 2019. Multimodal Active Learning From Human Data: A Deep Reinforcement Learning Approach. In *2019 International Conference on Multimodal Interaction* (Suzhou, China) *(ICMI '19)*. Association for Computing Machinery, New York, NY, USA, 6–15. https://doi.org/10.1145/3340555.3353742

[32] Takeshi Saga, Hiroki Tanaka, and Satoshi Nakamura. 2023. Computational analyses of linguistic features with schizophrenic and autistic traits along with formal thought disorders. In *Proceedings of the 25th International Conference on Multimodal Interaction (ICMI '23)*. Association for Computing Machinery, New York, NY, USA, 119–124. https://doi.org/10.1145/3577190.3614132

[33] Hanan Salam, Oya Çeliktutan, Isabelle Hupont, Hatice Gunes, and Mohamed Chetouani. 2017. Fully Automatic Analysis of Engagement and Its Relationship to Personality in Human-Robot Interactions. *IEEE Access* 5 (2017), 705–721. https://doi.org/10.1109/ACCESS.2016.2614525

[34] Joel Shor and Subhashini Venugopalan. 2022. TRILLsson: Distilled Universal Paralinguistic Speech Representations. In *Proc. Interspeech 2022*. 356–360. https://doi.org/10.21437/Interspeech.2022-118

[35] Candace L. Sidner, Cory D. Kidd, Christopher Lee, and Neal Lesh. 2004. Where to look: a study of human-robot engagement. In *Proceedings of the 9th International Conference on Intelligent User Interfaces* (Funchal, Madeira, Portugal) *(IUI '04)*. Association for Computing Machinery, New York, NY, USA, 78–84. https://doi.org/10.1145/964442.964458

[36] Judy Singer. 1999. Why can't you be normal for once in your life? From a problem with no name to the emergence of a new category of difference. *Disability Discourse* (1999), 59–70.

[37] Sudha M. Srinivasan, Inge-Marie Eigsti, Linda Neelly, and Anjana N. Bhat. 2016. The effects of embodied rhythm and robotic interventions on the spontaneous and responsive social attention patterns of children with autism spectrum disorder (ASD): A pilot randomized controlled trial. *Autism Spectrum Disorders* 27 (2016), 54–72. https://doi.org/10.1016/j.rasd.2016.01.004

[38] Kalin Stefanov, Giampiero Salvi, Dimosthenis Kontogiorgos, Hedvig Kjellström, and Jonas Beskow. 2019. Modeling of Human Visual Attention in Multiparty Open-World Dialogues. *J. Hum.-Robot Interact.* 8, 2, Article 8 (jun 2019), 21 pages. https://doi.org/10.1145/3323231

[39] Lars Steinert, Felix Putze, Dennis Küster, and Tanja Schultz. 2020. Towards Engagement Recognition of People with Dementia in Care Settings. In *Proceedings of the 2020 International Conference on Multimodal Interaction* (Virtual Event, Netherlands) *(ICMI '20)*. Association for Computing Machinery, New York, NY, USA, 558–565. https://doi.org/10.1145/3382507.3418856

[40] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal Transformer for Unaligned Multimodal Language Sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 6558–6569. https://doi.org/10.18653/v1/P19-1656

[41] Bethany C. Wangelin, Margaret M. Bradley, Anna Kastner, and Peter J. Lang. 2012. Affective engagement for facial expressions and emotional scenes: The influence of social anxiety. *Biological Psychology* 91, 1 (2012), 103–110. https://doi.org/10.1016/j.biopsycho.2012.05.002

[42] Anne S. Warlaumont, D. Kimbrough Oller, Rick Dale, Jeffrey A. Richards, Jill Gilkerson, and Dongxin Xu. 2010. Vocal Interaction Dynamics of Children with and Without Autism. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, S. Ohlsson and R. Catrambone (Eds.). Cognitive Science Society.