Multipar-T: Multiparty-Transformer for Capturing Contingent Behaviors in Group Conversations

Dong Won Lee , **Yubin Kim** , **Rosalind Picard** , **Cynthia Breazeal** , **Hae Won Park**
Massachusetts Institute of Technology
{dongwonl, ybkim95, picard, cynthiab, haewon}@mit.edu,

Abstract

As we move closer to real-world AI systems, AI agents must be able to deal with *multiparty* (group) conversations. Recognizing and interpreting multiparty behaviors is challenging, as the system must recognize individual behavioral cues, deal with the complexity of multiple streams of data from multiple people, and recognize the subtle contingent social exchanges that take place amongst group members. To tackle this challenge, we propose the Multiparty-Transformer (Multipar-T), a transformer model for multiparty behavior modeling. The core component of our proposed approach is the Crossperson Attention, which is specifically designed to detect contingent behavior between pairs of people. We verify the effectiveness of Multipar-T on a publicly available video-based group engagement detection benchmark, where it outperforms state-of-the-art approaches in average F-1 scores by 5.2% and individual class F-1 scores by up to 10.0%. Through qualitative analysis, we show that our Crossperson Attention module is able to discover contingent behavior.

1 Introduction

In order to develop AI agents that can co-exist with people in the real-world, it must be able to understand people's behavior in a multiparty (group) setting, as many common forms of important communicative behavior take place in small group settings. Accurate recognition and interpretation of multiparty behavior enables AI agents to support and facilitate group conversations across many domains, including educational lessons, business meetings, and collaborations at the workplace. Importantly, due to COVID-19 and the proliferation of hybrid work, there is an urgent need of modeling multiparty behavior in online settings, as meetings are transitioning from in-person to online in today's society. Hence, in this work, we focus on online multiparty interaction settings.

Recognizing and interpreting group behaviors is much more challenging than that of individual behaviors. Firstly, the system must perform well in recognizing individual behavioral cues. Secondly, it must do so simultaneously, while keeping track of every individual in the group. Finally, it must also recognize the subtle interactions that take place between group members as it can provide more insights into what is being communicated. Natural human conversations are **interactively contingent**, where people act and react in a coordinated fashion in turns [Kopp, 2010]. Consequently, understanding group behavior in multiparty conversations requires recognizing contingent behaviors between group members.

To address these challenges, we propose the Multiparty-Transformer (Multipar-T), which is able to handle multiple streams of input data for all of the members of the group. At the core of Multipar-T is Crossperson Attention. Instead of using cross attention to discover alignment between two sequences of different modalities (i.e. vision and language) [Tsai *et al.*, 2019] or differing views of a single visual input to learn multi-scale feature representations as in previous approaches [Chen *et al.*, 2021], we propose and show that cross attention can be effectively used to capture contingent behavior between two behavioral sequences across pairs of people; we call this Crossperson Attention (CPA). CPA implicitly searches for how and when one person's current behavior is contingent on another person's past behavior, whereas previous approaches that do not take contingent behaviors into account. Via careful construction of the direction of attention, Crossperson Attention controls the direction of contingent behaviors it captures. Furthermore, applying multiple layers of crossperson attention allows the model to discover relationships between the contingent behaviors with the other parts of the target person's behavioral sequence. We also include a self-transformer to address cases where behaviors are non-contingent and need to rely solely on the target person's behaviors. In summary, given a target person's input behavioral sequence, Multiparty-Transformer applies Crossperson Attention with the behavioral sequences of other members of the group in a pairwise manner to output an embedding that has contextualized information about the rich, social contingent interactions in reference to the target person's behaviors.

In order to measure the effectiveness of our proposed approach, we focus on the important task of engagement prediction in online learning activities. Engagement prediction is a task that requires understanding of contingent behaviors, as many studies demonstrate that the presence and lack of contingent behaviors influence people's engagement [Masek *et al.*, 2021; Xu *et al.*, 2022; Boyd and Rubin, 2006;

Sage and Kindermann, 1999]. Furthermore, engagement detection in group settings is an important problem in developing AI systems that can gauge a group's interest to develop behavior policies and strategies to maximize the group's overall satisfaction with the AI agent's actions. We establish baselines to compare against the proposed model on a publicly available group engagement detection dataset in an online educational setting [Reverdy *et al.*, 2022], where we find that Multipar-T consistently outperforms previous approaches, across all levels of engagement. We provide in-depth ablation studies to show how each specific components in Multipar-T contributes to the performance boost. Furthermore, we empirically show that the Crossperson Attention mechanism is able to discover contingent behaviors across pairs of people, which is especially important with the new EU policies [EUCommission, 2021] requiring explainability in affect recognition models.

Our contributions are summarized as follows: (1) We introduce the Multiparty-Transformer (Multipar-T), a novel transformer model which can handle multi-stream multi-party data in group conversations. (2) The key component is Crossperson Attention, which is the first of its kind to reframe cross attention to discover contingent behavior between group members by controlling the direction of the attention. The output embedding prioritizes parts of the target person $self$'s behavioral sequence that is contingent on another person $other$'s behavior. (3) The inclusion of the self-transformer module further contextualizes the embedding with the target person's own behavior and handles non-contingent behaviors. (4) We run extensive experiments on a important and timely task of multiparty engagement detection in online educational setting and show that Multipar-T significantly outperforms all previous state-of-the-art approaches.

2 Related Works

2.1 Contingent Behavior

Contingent behavior refers to a person's action that takes place as a response to another person's behavior. The term falls under a broader umbrella of "inter-personal coordination" [Bernieri, 1988], which refers to people's behavioral adaptations that happen as a result of social resonance in natural interaction. An example of contingent behavior is mimicry and interactional synchrony. Furthermore, nonconscious contingent behavior acts as a "social glue" to enhance the naturalness and sympathy in conversation [Lakin *et al.*, 2003]. Temporal coordination between people in communication is found in body movements and facial expressions [Bernieri *et al.*, 1994]. With regards to engagement, studies have shown that contingent and reciprocal interaction amongst groups of peers [Sage and Kindermann, 1999], teachers and students [Boyd and Rubin, 2006], caregivers and children [Masek *et al.*, 2021; Chen *et al.*, 2022], children and on-screen characters [Xu *et al.*, 2022], robots and humans [Admoni and Scassellati, 2014; Park *et al.*, 2017; Chen *et al.*, 2020b], influences individual engagement, motivation, and learning.

2.2 Modeling of Group Interactions

In small group interactions (at least 3 people), each person has their own attributes, and each member of the group communicates with each other via both nonverbal and verbal behaviors [Adams *et al.*, 2006]. Graphical models of explicitly modeling people's interactions have been explored in various tasks such as prediction of group performance [Lin and Lee, 2020], group behavior recognition [Yang *et al.*, 2020], social interaction field modelling [Zhou *et al.*, 2019], and social relation recognition [Li *et al.*, 2020]. The above-mentioned works all involve representing each person's individual features as a node, and their interactions as their edges. There has been work that utilizes attention in modeling 2-person interactions [Curto *et al.*, 2021], however, our work addresses a more complex multiparty interaction setting with a novel set-up of cross attention which captures contingent behavior in all pairwise interactions.

2.3 Engagement Prediction

At a high level, engagement is defined as a state of consciousness where a person is fully immersed in the task at hand [Ren, 2016]. Studies have investigated finer differences between specific types of engagement and shown that engagement is defined to be a multi-dimensional construct, composed of [Fredricks *et al.*, 2004; Silpasuwanchai *et al.*, 2016] behavioral (e.g. [Griffin *et al.*, 2008]), cognitive (e.g. [Corno and Mandinach, 1983]), emotional (e.g. [Park *et al.*, 2012]), and attentional (e.g [Chapman, 1997]) engagement. In our work, we focus on perceived behavioral and emotional engagement. Previous approaches utilize CNN-LSTM models to predict engagement [Del Duchetto *et al.*, 2020; Steinert *et al.*, 2020]. More recently, models that use bootstrapping and ensembling are proposed in BOOT [Wang *et al.*, 2019] and ENS-MODEL [Thong Huynh *et al.*, 2019]. HTMIL uses a Bi-LSTM with multi-scale attention and clip-level and video-level objectives [Ma *et al.*, 2021], and TEMMA [Chen *et al.*, 2020a] utilizes a Resnet-Transfomer model. Unlike our work, previously proposed approaches do not take into account the group setting; they focus on modeling individuals. Closest to our work in multiparty engagement prediction is the work of [Zhang *et al.*, 2022], where they utilize a graph attention network (GAT) to contextualize social interactions between multiple people to estimate engagement in elderly multiparty human-robot settings. To the best of our knowledge, we are the first to utilize a transformer network to model group's behavioral contingencies for engagement prediction.

3 Problem Statement

We formulate the multiparty video-based engagement prediction problem as the following. We are given video clips of groups involved in an online learning activity. We split the videos into N interval clips of k frames. At any arbitrary time t, where t is the exact timestep in which we want to predict each individual's engagement value, we are given the $[t-k, ..., t]$ interval of contextual video information; k is the number of frames we will use as context. Let P be the number of all participants in the video, for a person in the clip $p \in [P]$, their corresponding contextual behavioral features can be viewed as $X_p^t = [x_p^{t-k}, \ldots, x_p^t]$, where $x_p^t \in \mathcal{R}^{\mathcal{F}}$ with dimension size F at the t^{th} frame. For brevity, we will drop the t and assume it is arbitrarily fixed. The input with all of the group's features is a 3-D tensor, $X = [x_1, \ldots, x_P] \in \mathcal{R}^{P \times k \times F}$. For a target person:

$self$, we train a model that takes as input X, which includes the target person's features x_{self} as well as all other members' features $[x_{other}, \forall other \in P \setminus \{self\}]$ to predict the engagment value $\hat{Y}_{self} \in (0,1)^C$.

4 Methods

Here, we describe our proposed **Multipar**ty-**T**ransformer (**Multipar-T**). We utilize the Crossperson Attention (CPA) module to discover the contingencies across time-series sequences of behaviors from pairs of people in the group.

4.1 Crossperson Attention

Given a pair of people, we want to capture their contingent behaviors. We state that target person $self$'s behavior is *contingent on* person $other$'s behavior if person $self$'s behavior was likely to be influenced by person $other$'s behavior ($other \rightarrow self$). We posit that capturing contingent behavior across people can be handled with mechanisms that can capture alignment between sequences. Inspired by the multimodal transformer model [Tsai *et al.*, 2019], which shows one effective method that can automatically align sequences of differing modalities is by using a scaled dot product cross attention [Chen *et al.*, 2021; Vaswani *et al.*, 2017], we propose that it could be applied to pairs of behavioral features, which we call Crossperson Attention (CPA), to automatically discover contingent behavior and subtle social interactions.

Cross attention utilizes query Q, key K, and values V, where the first step is to find the importance of each key with respect to the query. The attention mechanism computes the dot product of the query with each key to obtain a weight for each key. These resulting weights represent the importance of each key to the query. The weights are then used to obtain a weighted sum of the values in the matrix, which is referred to as the context vector. Therefore, for the target person $self$, and another person $other$, we are given their time-aligned encoded visual representations Z_{self}, $Z_{other} \in \mathbb{R}^{k \times d_x}$, where d_x is the dimension size of the visual embeddings. Therefore, we construct the queries, keys, and values as the following: Querys as $Q_{other} = Z_{other} W_{Q_{other}}$, Keys as $K_{self} = Z_{self} W_{K_{self}}$, and Values as $V_{self} = Z_{self} W_{V_{self}}$, where $W_{Q_{other}}, W_{K_{self}}, W_{V_{self}} \in \mathbb{R}^{d_x \times d_x}$ are trainable weight parameters.

$$\begin{aligned}\text{CPA}_{other \rightarrow self}(Z_{other}, Z_{self}) &= \text{softmax}\left(\frac{Q_{other} K_{self}^{\top}}{\sqrt{d_x}}\right) V_{self} \\ &= \text{softmax}\left(\frac{Z_{other} W_{Q_{other}} \left(Z_{self} W_{K_{self}}\right)^{\top}}{\sqrt{d_x}}\right) Z_{self} W_{V_{self}}.\end{aligned} \tag{1}$$

We refer the readers to Figure 1(a) for a visual depiction. In Equation **??**, the scaled softmax produces the attention weight between two people's temporal behavior inputs, which weighs the importance of person $other$'s each behavioral timesteps with respect to the $self$'s behavior. Specifically, the resulting weight is a $k \times k$ matrix, where k is the number of timesteps in the sequence. After the dot poduct with V_{self}, the Crossperson Attention from $other$ to $self$ $\text{CPA}_{other \rightarrow self}(Z_{other}, Z_{self})$ outputs an embedding which has captured the person $self$'s behavior contingent on person $other$'s behaviors. We highlight this is the *reverse* direction of cross attention compared to many previous works [Tsai *et al.*, 2019; Curto *et al.*, 2021], and a crucial distinction in capturing contingent behaviors as we discuss in Section 6.1. Furthermore, Crossperson Attention mechanism is performed with h multiple heads; we define this as $\text{CPA}^{multi}_{other \rightarrow self}$.

$$\begin{aligned}&\text{CPA}^{multi}_{other \rightarrow self}(Z_{other}, Z_{self}) \\ &= \text{Concat}\left(\text{CPA}^{1}_{other \rightarrow self}, \ldots \text{CPA}^{h}_{other \rightarrow self}\right) W^{\text{multi}}\end{aligned} \tag{2}$$

The outputs of each head of CPA are concatenated, then linearly projected with weight matrix: $W^{multi} \in \mathbb{R}^{(h \cdot d_x) \times d_x}$.

4.2 Multiparty Transformer

In order to successfully address the complex social interactions taking place in a group setting, we must properly represent each person's individual temporal features, address the group social interactions, then take into account the group's temporal nature. We describe in detail the individual components which are designed to tackle these challenges.

Individual Temporal Encoder: Convolutions and Positional Encoding We utilize 1D convolutional layers such that the convolution kernel convolves over the temporal dimension and each timestep in the sequence is contextualized by its surroundings. Furthermore, we further enforce the temporal structure by including the additive positional encoding (PE) used in [Vaswani *et al.*, 2017]. Therefore, the individual temporal encoder, given target person $self$'s input X_{self} is:

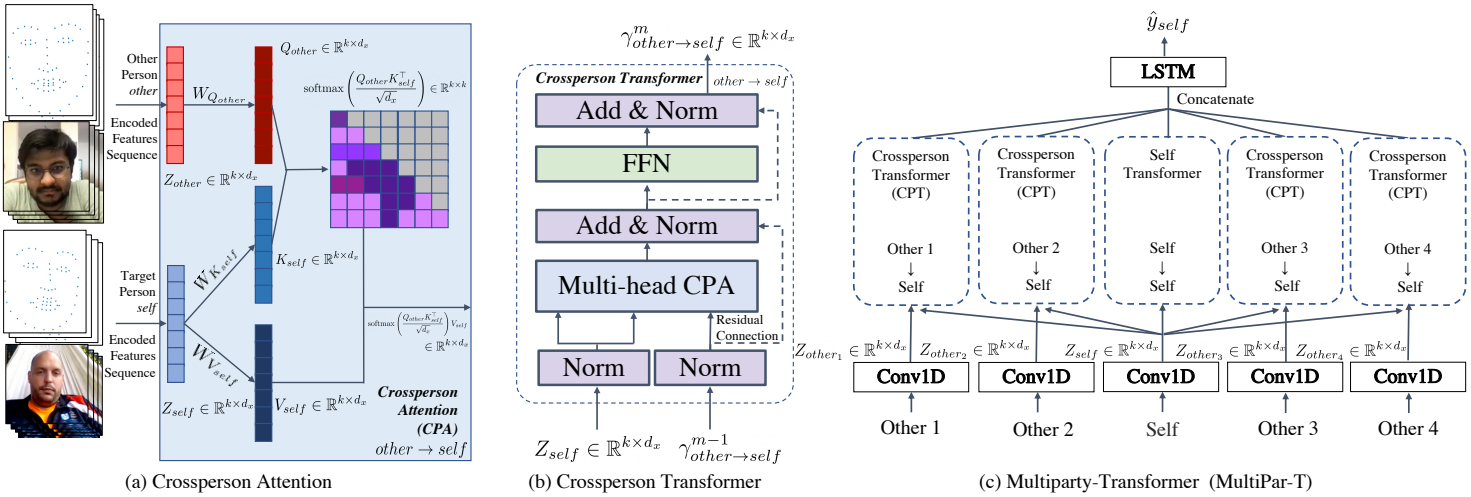
$$Z_p = \text{Conv1D}(X_p) + \text{PE}(X_p) \tag{3}$$

Conv1D includes a kernel that maps each individual's features into a common dimension d_x.

Behavior Interaction Encoder: Crossperson Transformer & Self Transformer Crossperson Attention (CPA) is a core component of the M-layered Crossperson Transformer (CPT). $\text{CPA}^{m,\text{multi}}_{other \rightarrow self}$ refers to the multi-headed Crossperson Attention from person $other$ to person $self$ at the m-th layer. Following standard transformer operations [Vaswani *et al.*, 2017], $\hat{\gamma}^{m}_{other \rightarrow self}$ refers to the intermediate output after the Crossperson Attention with residual connections. $\gamma^{m}_{other \rightarrow self}$ refers to the final output of a cross-person transformer block after feedforward network (FFN) and residual connections. As the input to the first layer, $\gamma^{0}_{other \rightarrow self} = Z_{other}$. We refer the readers to Figure 1(b) for details.

$$\begin{aligned}\gamma^{m}_{other \rightarrow self} &= \text{CPT}^{m}_{other \rightarrow self}(\gamma^{m-1}_{other \rightarrow self}, Z_{self}) \\ \hat{\gamma}^{m}_{other \rightarrow self} &= \text{CPA}^{m,multi}_{other \rightarrow self}(\text{Norm}(\gamma^{m-1}_{other \rightarrow self}), \text{Norm}(Z_{self})) \\ &\quad + \text{Norm}(\gamma^{m-1}_{other \rightarrow self})) \\ \gamma^{m}_{other \rightarrow self} &= \text{Norm}(\text{FFN}(\hat{\gamma}^{m}_{other \rightarrow self}) + \hat{\gamma}^{m}_{other \rightarrow self})\end{aligned} \tag{4}$$

With this formulation, $\text{CPA}^{0}_{other \rightarrow self}(Z_{other}, Z_{self})$ discovers contingent behaviors in the first layer. Then, in the later CPT layers, CPA contextualizes the embedding by discovering correlations on how the contingent behavior is related to different parts of the target person's behaviors. We empirically show that standalone first layer CPT is not



(a) Crossperson Attention (b) Crossperson Transformer (c) Multiparty-Transformer (MultiPar-T)

Figure 1: Full model architectures(a) Diagram of our Crossperson Attention ($\text{CPA}_{other \to self}$) module, which automatically searches for $self$'s behaviors that are contingent on $other$'s behaviors. (b) Diagram of the proposed new Crossperson Transformer. (c) Diagram of the overarching new model: Multiparty-Transformer takes in all other persons's behavioral features, applies Crossperson Attention w.r.t to $self$'s features in the Crossperson Transformer, as well as self-attention in the Self Transformer. Best viewed zoomed in and in color.

enough, and that further contextualization with multiple layers of transformer blocks is useful in Section 6.1.

In addition to all pair-wise Crossperson Attention across all other members of the group $\forall other \in P \setminus \{self\}$, $\text{CPA}^{multi}_{other \to self}$, we also compute self-attention in order to (1) account for how one's earlier behavior correlates with their current behavior and (2) handle cases where there are no contingent behavior information. This is equivalent to performing CPA with an equivalent query, key and value matrices (i.e. given a target person $self$ we perform $\text{CPA}^{multi}_{self \to self}(Z_{self}, Z_{self})$). Its usefulness is tested with ablation studies in Section 6.1.

Temporal Classifier Finally, we concatenate the outputs from the above-mentioned behavior interaction encoder, $[\cdot||\cdot]$ refers to concatenation. The concatenated outputs are fed into an LSTM for k steps to enforce a stronger temporal structure. The resulting output is passed through fully connected layers (FFN) for the final prediction $\hat{Y}_{self}$,

$$\begin{aligned} \zeta^n_{self}, hidden^n &= \text{LSTM}([\gamma^M_{1 \to self}|| \dots ||\gamma^M_{P \to self}], hidden^{n-1}) \\ \hat{Y}_{self} &= \text{FFN}(\zeta^k_{self}) \qquad\qquad for\ n \in [k] \end{aligned} \tag{5}$$

5 Experiments

5.1 Dataset

We utilize the RoomReader [Reverdy *et al.*, 2022] as a benchmark to measure the performance of our proposed method against other baselines. RoomReader [Reverdy *et al.*, 2022] is a corpus of multimodal, multiparty conversational interactions in which participants followed a collaborative online student-tutor scenario designed to elicit spontaneous speech. Engagement is focused on off-task/on-task engagement, where the task at hand is led by the instructor.

Engagement Classification RoomReader provides continuous annotations for engagement, where the engagement

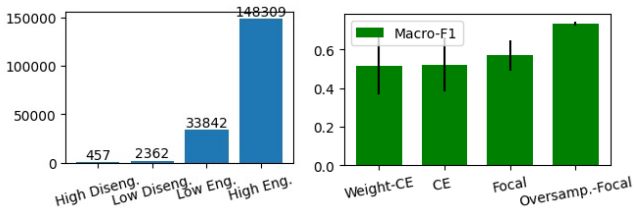


Figure 2: (Left) Distribution of class labels in the dataset. There is a severe class imbalance. (Right) Macro-F1 scores for Multipar-T for each class imbalance strategy, (CE refers to Cross-Entropy), where the combination of oversampling and focal loss performs well.

labels range from $[-2, 2]$. Instead of regression, we define the task as a 4-class classification, where labels between $(1, 2]$ refer to high engagement, $(0, 1]$: low engagement, $(-1, 0]$: low disengagement, $(-2, -1]$: high disengagement. Setting up the task in this way results in more interpretable evaluation metrics than regression losses (such as MAE, or MSE), and allows us to report categorical metrics conditioned on each class. Generally, it is well known that class imbalance is often severe for datasets with engagement labels [Del Duchetto *et al.*, 2020; Dhall *et al.*, 2020; Steinert *et al.*, 2020]. In the RoomReader dataset, 80.2% of the entire dataset consists of highly engaged samples, 18.3%: low engagement, 1.3%: low disengagement, and 0.2%: high disengagement. We refer the readers to Figure 2 for the imbalanced distribution of labels. To counter the effects of class imbalance, we (1) oversample the infrequent class to balance the dataset and (2) train the models with a Focal Loss [Lin *et al.*, 2017] that applies a modulating term to the cross entropy loss in order to focus learning on hard misclassified examples as shown below. y_{ic} refers to the ground truth labels, $\hat{Y}_{ic}$ is the probability prediction, and α is a hyperparameter that weighs how much easy samples should be down-weighted.

	All Engagement Classes			High Dis-Eng.	Low Dis-Eng.	Low Eng.	High Eng.
Model	Accuracy	Weighted F1	Macro F1	F1	F1	F1	F1
ConvLSTM [Del Duchetto *et al.*, 2020]	0.859 ± 0.01	0.857 ± 0.02	0.699 ± 0.05	0.741	0.459 ± 0.22	0.699 ± 0.12	0.907 ± 0.01
OCtCNN-LSTM [Steinert *et al.*, 2020]	0.769 ± 0.08	0.695 ± 0.14	0.410 ± 0.10	0.588	0.119 ± 0.17	0.233 ± 0.33	0.864 ± 0.05
TEMMA [Chen *et al.*, 2020a]	0.823 ± 0.02	0.822 ± 0.02	0.561 ± 0.11	0.286	0.254 ± 0.19	0.621 ± 0.13	0.885 ± 0.01
EnsModel [Thong Huynh *et al.*, 2019]	0.760 ± 0.07	0.675 ± 0.12	0.302 ± 0.03	0	0.000 ± 0.00	0.160 ± 0.23	0.860 ± 0.05
BOOT [Wang *et al.*, 2019]	0.817 ± 0.03	0.822 ± 0.03	0.636 ± 0.09	0.714	0.320 ± 0.24	0.658 ± 0.12	0.873 ± 0.02
HTMIL [Ma *et al.*, 2021]	0.820 ± 0.02	0.818 ± 0.02	0.460 ± 0.05	0	0.000 ± 0.00	0.633 ± 0.12	0.880 ± 0.02
GAT [Zhang *et al.*, 2022]	0.739 ± 0.06	0.631 ± 0.08	0.261 ± 0.03	0	0.000 ± 0.00	0.006 ± 0.01	0.848 ± 0.04
MulT [Tsai *et al.*, 2019]	0.847 ± 0.02	0.845 ± 0.02	0.624 ± 0.12	0.625	0.310 ± 0.25	0.665 ± 0.12	0.901 ± 0.01
Multipar-T (Ours)	$\mathbf{0.888 \pm 0.03}$	$\mathbf{0.887 \pm 0.03}$	$\mathbf{0.751 \pm 0.05}$	**0.800**	$\mathbf{0.559 \pm 0.07}$	$\mathbf{0.759 \pm 0.11}$	$\mathbf{0.927 \pm 0.02}$

Table 1: Results and standard deviations for engagement recognition models for 3 seeds (std dev for High Dis-Eng. not reported due to 2 seeds not having corresponding labels). Despite high accuracy and weighted-F1 scores, many previous baselines fail at infrequent disengagement classes. Multipar-T outperforms other approaches across all metrics.

$$L_{Focal} = -\frac{1}{N}\sum_{i}^{N}\sum_{c}^{C}(1-\hat{Y}_{ic})^{\alpha} y_{ic} \log\left(\hat{Y}_{ic}\right) \quad (6)$$

Combined, we find that the models are able to predict the infrequent classes and overcome the class imbalance problem, which can be seen in Figure 2.

Data Preprocessing For the input features, we utilize the normalized eye gaze direction, location of the head, location of 3D landmarks, and facial action units extracted via OpenFace [Baltrusaitis *et al.*, 2018]. In addition, we extract frame-wise image features from the penultimate layer of Resnet-50 [He *et al.*, 2016]. The two features are concatenated per timestep to be used as input. The input feature dimension size per timestep is $F = 2183$. For each label at timestep t, we use 8 seconds worth of video context information, where the frame rate is 8 fps. We utilize $k = 64$ frames as input. We apply a sliding window with an interval of 1 second between each sample. In total, we have 184970 samples.

5.2 Baseline Models

We compare our proposed model with a family of baselines in engagement prediction, as well as action recognition. We run the newest versions of these models and report their scores on a unified benchmark. We compare Multipar-T to ConvLSTM [Del Duchetto *et al.*, 2020], OCtCNN-LSTM [Steinert *et al.*, 2020], TEMMA [Chen *et al.*, 2020a], BOOT [Wang *et al.*, 2019] and ENS-MODEL [Thong Huynh *et al.*, 2019], GAT [Zhang *et al.*, 2022], and MulT [Tsai *et al.*, 2019]. For action recognition models, we compare our method with TimeSformer [Bertasius *et al.*, 2021], SlowFast [Feichtenhofer *et al.*, 2019] and I3D [Wang *et al.*, 2018].

5.3 Implementation Details

We train our models on 2 NVIDIA GeForce GTX 1080 Ti with a batch size of 64 for 20 epochs. We use the AdamW [Loshchilov and Hutter, 2017] optimizer with an initial learning rate of 0.0001 with a scheduler that decays the learning rate by 0.1 every 5 epochs. We train on 16 groups' data, validate on 3 groups, and test on 1 group for 3 seeds. The model is exposed to different totally held-out subsets of groups for cross-validation. Our code can be found in the Supplementary, and will be shared on a public github repository with camera ready. Multipar-T can be used in real-time, where the inference time only takes 0.0981 ± 0.0029 seconds.

6 Results & Discussion

In this section, we discuss the quantitative and qualitative results of our experiments. We compare our approach Multipar-T with state-of-the-art baselines. Then, we discuss the importance of the attention modules and the effects of their directions. Finally, we qualitatively demonstrate that Crossperson Attention has learned to recognize contingent behaviors.

6.1 Quantitative Results

Following previous works [Del Duchetto *et al.*, 2020; Dhall *et al.*, 2020; Steinert *et al.*, 2020], we report accuracy and weighted-F1, which is the weighted mean of all per-class F1 scores considering each class's support in the data. Most importantly, we report the macro-F1, i.e., the unweighted mean of per-class F1. A high macro-F1 score demonstrates that the model performs well across all engagement classes regardless of its frequency in the dataset.

Comparisons Against State-of-the-Art In Table 1, state-of-the-art engagement prediction models are compared to Multipar-T. Multipar-T outperforms all previous state-of-the-art approaches in accuracy by 2.9%, weighted F1 by 3.0%, and Macro-F1 by 5.2%. Again, Macro-F1 is the most informative metric, as predicting the most infrequent class, i.e., disengagement, is the most challenging component of this problem. For a closer look into how Multipar-T performs for individual levels of engagement, we refer the readers to the right of Table 1, where we compare each model's performance for each specific engagement class. Although other baselines result in comparable high accuracy and weighted-F1 scores, they fail to predict the infrequent disengagement class. Multipar-T has significant performance gains (10% increase against best performing baseline) in the most challenging task of accurately predicting high and low disengagement, which consists of only 2% of the entire dataset. Moreover, comparing with the other group behavior encoding method, GAT [Zhang *et al.*, 2022], we find that Multipar-T outperforms across all metrics, highlighting that our method is a more effective method of capturing group behavior. We also compare with an adaptation of the multimodal transformer (MulT) [Tsai *et al.*, 2019] where we replace differing modality inputs with differing person's behavioral sequences, which is equivalent to Multipar-T w/o Self Transformer and reversed attention direction $CPA_{self \rightarrow other}$ in

Attention Direction	Ablation	All Classes			High Dis-Eng.	Low Dis-Eng.	Low Eng.	High Eng.
		Accuracy	Weighted F1	Macro F1	Binary F1	Binary F1	Binary F1	Binary F1
	Multipar-T w/o Crossperson Transformer	0.847 + 0.0154	0.844 + 0.14	0.661 + 0.018	0.588	0.433 + 0.1	0.66 + 0.12	0.901 + 0.01
$CPA_{self \to other}$	Multipar-T w/o Self Transformer	0.847 + 0.0167	0.845 + 0.021	0.624 + 0.12	0.625	0.31 + 0.25	0.665 + 0.12	0.901 + 0.01
	Multipar-T	0.865 + 0.03	0.862 + 0.036	0.735 + 0.02	**0.769**	**0.587 + 0.12**	0.698 + 0.15	0.912 + 0.02
$CPA_{other \to self}$	Multipar-T w/o Self Transformer	**0.883 + 0.02**	**0.884 + 0.024**	**0.75 + 0.04**	**0.769**	**0.555 + 0.11**	**0.762 + 0.08**	**0.923 + 0.02**
	Multipar-T	**0.883 + 0.02**	**0.885 + 0.02**	**0.75 + 0.06**	0.714	0.557 + 0.19	**0.766 + 0.08**	**0.923 + 0.02**

Table 2: Ablation results for Self Tranformer and Crossperson Transformer mechanisms. Attending to $other$'s and own $self$ behaviors boosts performance. We refer the readers to Figure 1. Multipar-T w/o Crossperson Transformer refers to the ablation of all pairwise Crossperson Transformers with only the Self Transformer remaining. Multipar-T w/o Self Transformer refers the ablation of the Self Transformer and utilizing the pairwise Crossperson Transformers. Results with different directions of Crossperson Attention are displayed, where $CPA_{other \to self}$ performs well generally and $CPA_{self \to other}$ performs well for disengaged instances.

Table 2. The inclusion of the self-transformer and the configuration of the attention directions is a key component in modeling human multiparty behavior, different from multimodality alignment, as we demonstrate in ablation studies in the next sections.

Importance of Crossperson Attention (CPA) and Self-Attention In Table 2, we present results where we ablate Crossperson Attention and Self-Attention from our models. We see that ablating Crossperson Attention leads to a significant drop in performance metrics, especially Macro-F1. The model struggles at harder, low-data disengagement instances. Therefore, the inclusion of Crossperson Attention, which allows the model to attend to how others are behaving in the group provides more context information for the model to differentiate harder cases. We also find that the ablation of self-attention leads to significant drops in performance metrics, as self-attention provides more information regarding one's own behavior, which is especially important when there are no contingent behaviors.

Importance of the Direction of CPA In predicting a target person $self$'s engagement value, we hypothesized that the $self$'s behavior contingent on $other$'s behavior is an important predictor of engagement and disengagement. On the other hand, we hypothesized that the $other$'s behavior contingent on the $self$'s behavior would not be an important predictor. To test these hypotheses, we carefully experiment with the directions of the Crossperson Attention mechanism. In Table 2, CPA$_{other \to self}$ refers to our set up of the attention direction, performing Crossperson Attention where the query corresponds to behaviors of *other* persons in the group, and the key and value correspond to the behavior of the target, *self*. The resulting embedding contains information about the $self$'s behavior which is contingent on $other$'s. Conversely, CPA$_{self \to other}$, outputs an embedding with the $other$'s behavior contingent on the $self$'s behavior. This is similar to the cross attention set-up in [Tsai *et al.*, 2019; Curto *et al.*, 2021].

We refer the readers to the results for Multipar-T in CPA$_{other \to self}$ and CPA$_{self \to other}$. We find that Multipar-T with our formulation of cross attention, CPA$_{other \to self}$, performs significantly better, which indicates that it is important to explicitly set the direction of cross attention such that the output embedding prioritizes parts of target's behavioral sequence that is contingent on another person's behavior. Interestingly, when predicting low disengagement, Multipar-T with CPA$_{self \to other}$ results in better performance. This shows that how the $self$ impacts $others$ could be a important predictor when predicting if the target person is disengaged.

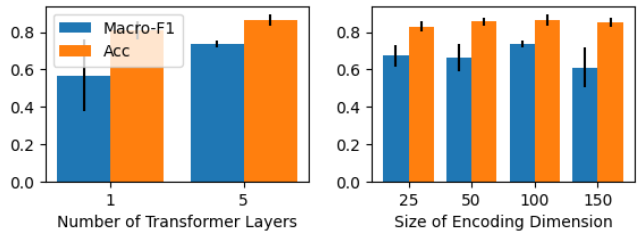


Figure 3: Macro-F1 and Accuracy scores for important hyperparameters for Multipar-T. (Left) Multi-layered transformers and (Right) encoding dimension of $d_x = 100$ boosts performance.

Encoding Size & Transformer Layers In Figure 3, we first display results for ablations on varying number transformer encoder layers M. The first layer of CPT encodes the $self$'s behavior contingent on $other$. The later layers further contextualizes the contingent $self$'s behavior with its own behavior again, allowing it to attend to other parts of its own behavioral sequence. We find that having multiple layers of transformer encoders leads to a significant improvement in Macro-F1. Secondly, we also display results after varying the size of the embedding dimension per timestep d_x, which is an important hyperparameter that controls the expressivity of the model. We find that the optimal encoding dimension to encode behavior per timestep setting is $d_x = 100$.

Comparisons against Action Recognition Models Following recent approaches in utilizing action recognition models in engagement prediction [Ai *et al.*, 2022; Kim *et al.*, 2022], we compare Multipar-T to activity recognition models in Table 3. Training state-of-the-art action recognition models are computationally much more expensive than engagement prediction models, due to the fact that the model is trained on a time-series of raw pixels end-to-end. Therefore, instead of applying an 8 seconds window with 1 second interval, we apply an interval of 8 seconds. Even with a modified set-up, there is a large discrepancy between the training time between these two classes of models, one seed takes $\sim$ 150 minutes for a raw video-based model, compared to $\sim$ 20 minutes for an engagement prediction model. Nonetheless, for a fair comparison, we train our proposed model with the same training settings. We find that the training of raw video models end-to-end yields poor performance specifically for scarce labels in disengagement. Multipar-T performs better than other architectures given the same training conditions.

Model	Accuracy	Weighted F1	Macro F1
I3D [Wang *et al.*, 2018]	0.751 ± 0.07	0.658 ± 0.08	0.254 ± 0.05
TimeSformer [Bertasius *et al.*, 2021]	0.806 ± 0.03	0.752 ± 0.05	0.337 ± 0.14
SlowFast [Feichtenhofer *et al.*, 2019]	0.718 ± 0.11	0.628 ± 0.12	0.232 ± 0.02
Multipar-T (Ours)	$\mathbf{0.828 \pm 0.02}$	$\mathbf{0.823 \pm 0.02}$	$\mathbf{0.466 \pm 0.06}$

Table 3: Raw video-based action recognition models and Multipar-T trained with less computationally heavy training set-up. Results and standard deviation are reported for 3 seeds. We see the limitations of training end-to-end raw video-based models.

6.2 Qualitative Analyses

Given the upcoming EU regulations [EUCommission, 2021] requiring explainability for any affect-related AI, an important facet of our work is that the resulting attention weights can be used as a way [Wiegreffe and Pinter, 2019] to explain why the model made this specific prediction for this specific timestep. Given an AI agent which utilizes Multipar-T as its backbone engagement detection module, if a person inquires the agent why it thought that they were disengaged at a point in time, the AI agent could examine is Crossperson Attention weights and provide some rationale behind its prediction based on discovered contingent behaviors.

To demonstrate how Crossperson Attention (CPA) could be used to explain model predictions, in Figure 4, we visualize the attention weights from the first layer of the Crossperson Attention between Person $self$, (left) and Person $other$, (top) and show that it is able to capture contingent behavior between pairs of individuals' behaviors. The x and y axes refer to each person's behavior aligned to timesteps. We first provide attention weights that demonstrate the lack of contingency for comparison. The attention weights visualized in Figure 4(b) Diagonal Contingency, is a diagonal attention weight matrix, which indicates that Person $self$ and Person $other$'s behavior are only related at the exact same time steps. The attention weights visualized in Figure 4(c), Uniform Contingency, is the default behavior if we assume that all of Person $other$'s past behavior is uniformly related to Person $self$'s current behavior, demonstrated by the uniform color across each row, which indicates that the attention weights are uniformly distributed across the available timesteps. The upper triangular matrix is masked to encode the natural assumption that $other$'s future behavior shouldn't affect $self$'s current behavior.

Figure 4(a) shows the learnt Crossperson Attention weights, which indicates that Person $self$'s behavior in timesteps 16–60 is contingent on Person $other$'s behavior in timestep 20−25. We find that, after inspecting the video at the aligned timesteps, that Person $other$ laughs during timestep $20 - 25$ and starts to talk afterwards. Person $self$ was initially distracted, but after they see Person $other$ laughing at timestep $20 - 25$, they look at Person $other$ and starts listening. Hence, we find that Crossperson Attention has discovered meaningful contingent behavior between two people. We kindly refer the reader to the supplementary for more examples.

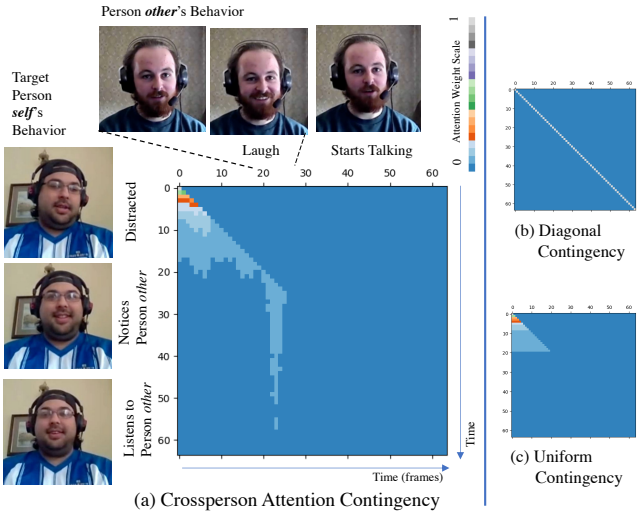


Figure 4: (a) Multiparty-Transformer Cross-person Attention weights from $t = 120s$ for group $S01$. Multipar-T has discovered that $self$'s behavior (smile and listen) from timestep 16 – 60 is contingent on $other$'s behavior in timestep 20 – 25 (laughter). (b) Diagonal Contingency: Crossperson attention weights with the assumption that person $self$ and $other$'s behavior are only related at the exact same timesteps. (c) Uniform Contingency: Default behavior of Crossperson attention weights; i.e. all of person $self$'s past behaviors are uniformly related to person $other$'s current behavior.

7 Conclusion

In this work, we study the challenging task of modelling human behaviors in a multiparty setting. We proposed a new Transformer-based model for multiparty behavior modeling, Multipar-T. At the core of Multipar-T is Crossperson Attention, which is designed to capture contingent behaviors. We compare Multipar-T against previous approaches on a timely and challenging task of engagement prediction in online meetings and provide in-depth analysis and ablation studies. We see significant gains in performance (up to 10%) compared to previous approaches, where we find that controlling the direction of Crossperson Attention, including multiple layers of transformer blocks, and self attention blocks is crucial. We also demonstrate qualitatively that our model is able to find contingent behaviors. Our Multipar-T is a novel approach to modeling contingent behaviors in multiparty conversation, a crucial problem in developing AI agents that can communicate with groups of people. We publicly share our code to enable research in multiparty interactions.

Limitations and Future Work Our evaluation benchmark Roomreader [Reverdy *et al.*, 2022] is collected from online lessons, which is one specific context. Future works should test the generalizability of models in different situations and scenarios including in-person settings, various relationships, diverse cultures, and a wider range of age of participants. We believe that our proposed method would generalize across different group settings, environment, and tasks, as supported by literature reviewed in Section 2.1 that contingent behaviors play an important role in various interaction settings. In addition, here we present results on the 5-person setting, as the dataset offered the most amount of data for this specific size. Follow-up work should focus on developing approaches that can perform well with a variable number of people. Using a unified encoder for all individual behavior as well as a unified cross-person transformer for every pairwise behavior

is promising. This would alleviate the need to train an N-person-specific model. In this work, we purely rely on the visual modality. Contingent behavior exists in language and acoustics as well. Using a multimodal input which includes these modalities should be investigated.

References

[Adams *et al.*, 2006] Katherine L Adams, Gloria J Galanes, and John K Brilhart. *Communicating in groups: Applications and skills. Chapter 4: Using Verbal and Nonverbal Messages in a Group.* McGraw-Hill Boston, 2006.

[Admoni and Scassellati, 2014] Henny Admoni and Brian Scassellati. Data-driven model of nonverbal behavior for socially assistive human-robot interactions. In *Proc. of the 16th int. conf. on multimodal interaction (ICMI)*, 2014.

[Ai *et al.*, 2022] Xusheng Ai, Victor S Sheng, and Chunhua Li. Class-attention video transformer for engagement intensity prediction. *arXiv:2208.07216*, 2022.

[Baltrusaitis *et al.*, 2018] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *Proc. of the 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 59–66, 2018.

[Bernieri *et al.*, 1994] Frank J Bernieri, Janet M Davis, Robert Rosenthal, and C Raymond Knee. Interactional synchrony and rapport: Measuring synchrony in displays devoid of sound and facial affect. *Personality and social psychology bulletin*, 20(3):303–311, 1994.

[Bernieri, 1988] Frank J Bernieri. *Coordinated movement in human interaction: Synchrony, posture similarity, and rapport*. Harvard University, 1988.

[Bertasius *et al.*, 2021] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding?, 2021.

[Boyd and Rubin, 2006] Maureen Boyd and Don Rubin. How contingent questioning promotes extended student talk: A function of display questions. *Journal of Literacy Research*, 38(2):141–169, 2006.

[Chapman, 1997] Peter McFaul Chapman. *Models of engagement: Intrinsically motivated interaction with multimedia learning software*. PhD thesis, University of Waterloo, 1997.

[Chen *et al.*, 2020a] Haifeng Chen, Dongmei Jiang, and Hichem Sahli. Transformer encoder with multi-modal multi-head attention for continuous affect recognition. *IEEE Transactions on Multimedia*, 23:4171–4183, 2020.

[Chen *et al.*, 2020b] Huili Chen, Hae Won Park, and Cynthia Breazeal. Teaching and learning with children: Impact of reciprocal peer learning with a social robot on children's learning and emotive engagement. *Computers & Education*, 150:103836, 2020.

[Chen *et al.*, 2021] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 357–366, 2021.

[Chen *et al.*, 2022] Huili Chen, Sharifa Mohammed Alghowinem, Soo Jung Jang, Cynthia Breazeal, and Hae Won Park. Dyadic affect in parent-child multi-modal interaction: Introducing the dami-p2c dataset and its preliminary analysis. *IEEE Transactions on Affective Computing*, 2022.

[Corno and Mandinach, 1983] Lyn Corno and Ellen B Mandinach. The role of cognitive engagement in classroom learning and motivation. *Educational psychologist*, 18(2):88–108, 1983.

[Curto *et al.*, 2021] David Curto, Albert Clapés, Javier Selva, Sorina Smeureanu, Julio Junior, CS Jacques, David Gallardo-Pujol, Georgina Guilera, David Leiva, Thomas B Moeslund, et al. Dyadformer: A multi-modal transformer for long-range modeling of dyadic interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2177–2188, 2021.

[Del Duchetto *et al.*, 2020] Francesco Del Duchetto, Paul Baxter, and Marc Hanheide. Are you still with me? continuous engagement assessment from a robot's point of view. *Frontiers in Robotics and AI*, 7:116, 2020.

[Dhall *et al.*, 2020] Abhinav Dhall, Garima Sharma, Roland Goecke, and Tom Gedeon. Emotiw 2020: Driver gaze, group emotion, student engagement and physiological signal based challenges. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, 2020.

[EUCommission, 2021] EUCommission. Proposal for a Regulation of the European Parliament and of the Council, Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, SEC (2021) 167 final, COM (2021) 2006 final, 2021.

[Feichtenhofer *et al.*, 2019] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE international conference on computer vision*, 2019.

[Fredricks *et al.*, 2004] Jennifer A Fredricks, Phyllis C Blumenfeld, and Alison H Paris. School engagement: Potential of the concept, state of the evidence. *Review of educational research*, 74(1):59–109, 2004.

[Griffin *et al.*, 2008] Mark A Griffin, Sharon K Parker, and Andrew Neal. Is behavioral engagement a distinct and useful construct? *Industrial and Organizational Psychology*, 1(1):48–51, 2008.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Kim *et al.*, 2022] Yubin Kim, Huili Chen, Sharifa Alghowinem, Cynthia Breazeal, and Hae Won Park. Joint engagement classification using video augmentation techniques for multi-person human-robot interaction. *arXiv:2212.14128*, 2022.

[Kopp, 2010] Stefan Kopp. Social resonance and embodied coordination in face-to-face conversation with artificial interlocutors. *Speech Communication*, 52(6):587–597, 2010.

[Lakin *et al.*, 2003] Jessica L Lakin, Valerie E Jefferis, Clara Michelle Cheng, and Tanya L Chartrand. The chameleon effect as social glue: Evidence for the evolutionary significance of nonconscious mimicry. *Journal of nonverbal behavior*, 27(3):145–162, 2003.

[Li *et al.*, 2020] Wanhua Li, Yueqi Duan, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Graph-based social relation reasoning. In *European Conference on Computer Vision*, pages 18–34. Springer, 2020.

[Lin and Lee, 2020] Yun-Shao Lin and Chi-Chun Lee. Predicting performance outcome with a conversational graph convolutional network for small group interactions. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[Lin *et al.*, 2017] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[Loshchilov and Hutter, 2017] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[Ma *et al.*, 2021] Jiayao Ma, Xinbo Jiang, Songhua Xu, and Xueying Qin. Hierarchical temporal multi-instance learning for video-based student learning engagement assessment. In *IJCAI*, pages 2782–2789, 2021.

[Masek *et al.*, 2021] Lillian R Masek, Brianna TM McMillan, Sarah J Paterson, Catherine S Tamis-LeMonda, Roberta Michnick Golinkoff, and Kathy Hirsh-Pasek. Where language meets attention: How contingent interactions promote learning. *Developmental Review*, 60:100961, 2021.

[Park *et al.*, 2012] Sira Park, Susan D Holloway, Amanda Arendtsz, Janine Bempechat, and Jin Li. What makes students engaged in learning? a time-use study of within-and between-individual predictors of emotional engagement in low-performing high schools. *Journal of youth and adolescence*, 41(3):390–401, 2012.

[Park *et al.*, 2017] Hae Won Park, Mirko Gelsomini, Jin Joo Lee, and Cynthia Breazeal. Telling stories to robots: The effect of backchanneling on a child's storytelling. In *Proc. of the ACM/IEEE Int. Conf. on Human-Robot Interaction (HRI)*, 2017.

[Ren, 2016] Xiangshi Ren. Rethinking the relationship between humans and computers. *Computer*, 49(8):104–108, 2016.

[Reverdy *et al.*, 2022] Justine Reverdy, Sam O'Connor Russell, Louise Duquenne, Diego Garaialde, Benjamin R Cowan, and Naomi Harte. Roomreader: A multimodal corpus of online multiparty conversational interactions. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2517–2527, 2022.

[Sage and Kindermann, 1999] Nicole A Sage and Thomas A Kindermann. Peer networks, behavior contingencies, and children's engagement in the classroom. *Merrill-Palmer Quarterly (1982-)*, pages 143–171, 1999.

[Silpasuwanchai *et al.*, 2016] Chaklam Silpasuwanchai, Xiaojuan Ma, Hiroaki Shigemasu, and Xiangshi Ren. Developing a comprehensive engagement framework of gamification for reflective learning. In *Proc. of the ACM Conf. on Designing Interactive Systems*, pages 459–472, 2016.

[Steinert *et al.*, 2020] Lars Steinert, Felix Putze, Dennis Küster, and Tanja Schultz. Towards engagement recognition of people with dementia in care settings. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 558–565, 2020.

[Thong Huynh *et al.*, 2019] Van Thong Huynh, Soo-Hyung Kim, Guee-Sang Lee, and Hyung-Jeong Yang. Engagement intensity prediction withfacial behavior features. In *2019 International Conference on Multimodal Interaction*, pages 567–571, 2019.

[Tsai *et al.*, 2019] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558, 2019.

[Vaswani *et al.*, 2017] 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.

[Wang *et al.*, 2018] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. *CVPR*, 2018.

[Wang *et al.*, 2019] Kai Wang, Jianfei Yang, Da Guo, Kaipeng Zhang, Xiaojiang Peng, and Yu Qiao. Bootstrap model ensemble and rank loss for engagement intensity regression. In *2019 International Conference on Multimodal Interaction*, pages 551–556, 2019.

[Wiegreffe and Pinter, 2019] Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. *arXiv preprint arXiv:1908.04626*, 2019.

[Xu *et al.*, 2022] Ying Xu, Valery Vigil, Andres S Bustamante, and Mark Warschauer. Contingent interaction with a television character promotes children's science learning and engagement. *Journal of Applied Developmental Psychology*, 81:101439, 2022.

[Yang *et al.*, 2020] Fangkai Yang, Wenjie Yin, Tetsunari Inamura, Mårten Björkman, and Christopher Peters. Group behavior recognition using attention-and graph-based neural networks. In *ECAI 2020*, pages 1626–1633. 2020.

[Zhang *et al.*, 2022] Zhijie Zhang, Jianmin Zheng, and Nadia Magnenat Thalmann. Engagement estimation of the elderly from wild multiparty human–robot interaction. *Computer Animation and Virtual Worlds*, page e2120, 2022.

[Zhou *et al.*, 2019] Chen Zhou, Ming Han, Qi Liang, Yi-Fei Hu, and Shu-Guang Kuai. A social interaction field model accurately identifies static and dynamic social groupings. *Nature human behaviour*, 3(8):847–855, 2019.