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

Dong Won Lee, Yubin Kim, Rosalind W. Picard, Cynthia Breazeal, Hae Won Park
Massachusetts Institute of Technology
{dongwoni, ybkm95, picard, cynthia, haewon}@mit.edu

Abstract
As we move closer to real-world social 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 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 behaviors.

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 lessons, business meetings, and collaborations at the workplace.

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), a transformer model for multiparty behavior modeling, which is able to handle multiple streams of input data for all of the members of the group, unlike most previous works which focus solely on individual interactions without taking into account the relationship between people’s behaviors. At the core of MultiPar-T is Crossperson Attention (CPA), which is used to attend to the behaviors of the entire group and capture contingent behaviors. 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. 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; Sage and Kindermann, 1999]. Furthermore, engagement detection in group settings is an im-
portant 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 component 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 atención 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 approaches1.

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. It falls under a broader umbrella of “interpersonal 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]. 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 [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 communi-

cates with each other via both nonverbal and verbal behaviors [Adams et al., 2006]. Graphical modelling of 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 dyadic 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 [Fredricks et al., 2004], composed of behavioral (e.g. [Griffin et al., 2008]), cognitive (e.g. [Coro 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 Duccheto 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-Transformer model. Unlike our work, previous 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 $X^p_t = [x^p_{t-k}, . . . , x^p_t]$, where $x^p_{t-k} \in \mathbb{R}^F$ 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 ten-
sor, $X = [x_1, \ldots, x_P] \in \mathbb{R}^{P \times k \times F}$. For a target person: \textit{self}, we train a model that takes as input $X$, which includes the target person’s features $x_{\text{self}}$ as well as all other members’ features $[x_{\text{other}}, \forall \text{other} \in P \setminus \{\text{self}\}]$ to predict the engagement value $Y_{\text{self}} \in (0, 1)^C$.

4 Methods

Here, we describe our proposed \textbf{Multiparty-Transformer (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 \textit{self}’s behavior is contingent on person \textit{other}’s behavior if person \textit{self}’s behavior was likely to be influenced by person \textit{other}’s behavior \textit{(other}→\textit{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 \cite{Tsay2019}, which shows one effective method that can automatically align sequences of differing modalities is by using a scaled dot product cross attention \cite{Chen2017, Vaswani2017}, we propose that it could be applied to pairs of behavioral sequences, 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 \textit{self}, and another person \textit{other}, we are given their time-aligned encoded visual representations $Z_{\text{self}}, Z_{\text{other}} \in \mathbb{R}^{k \times d_z}$, where $d_z$ is the dimension size of the visual embeddings. Therefore, we construct the queries, keys, and values as the following: Queries as $Q_{\text{other}} = Z_{\text{other}} W_{Q_{\text{other}}}$, Keys as $K_{\text{self}} = Z_{\text{self}} W_{K_{\text{self}}}$, and Values as $V_{\text{self}} = Z_{\text{self}} W_{V_{\text{self}}}$, where $W_{Q_{\text{other}}}, W_{K_{\text{self}}}, W_{V_{\text{self}}} \in \mathbb{R}^{d_x \times d_z}$ are trainable weights.

\begin{equation}
\text{CPA}_{\text{other} \rightarrow \text{self}}(Z_{\text{other}}, Z_{\text{self}}) = \text{softmax}\left(\frac{Q_{\text{other}} K_{\text{self}}}{\sqrt{d_k}}\right) V_{\text{self}}
\end{equation}

We refer the readers to Figure 1(a) for a visual depiction. In Equation 4.1, the scaled softmax produces the attention weight between two people’s temporal behavior inputs, which weighs the importance of person \textit{other}’s each behavioral timesteps with respect to the \textit{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 product with $V_{\text{self}}$, the Crossperson Attention from \textit{other} to \textit{self} CPA$_{\text{other} \rightarrow \text{self}}$($Z_{\text{other}}, Z_{\text{self}}$) outputs an embedding which has captured the person \textit{self}’s behavior contingent on person \textit{other}’s behaviors. We highlight this is the reverse direction of cross attention compared to many previous works \cite{Tsay2019, Curto2021}, 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 CPA$_{\text{other} \rightarrow \text{self}}^{\text{multi}}$.

\begin{equation}
\text{CPA}_{\text{other} \rightarrow \text{self}}^{\text{multi}}(Z_{\text{other}}, Z_{\text{self}}) = \text{Concat}(\text{CPA}_{\text{other} \rightarrow \text{self}}^1, \ldots, \text{CPA}_{\text{other} \rightarrow \text{self}}^h) W_{\text{multi}}
\end{equation}

The outputs of each head of CPA are concatenated, then linearly projected with weight matrix: $W_{\text{multi}} \in \mathbb{R}^{h \times d_z}$.

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 \cite{Vaswani2017}. Therefore, the individual temporal encoder, given target person \textit{self}’s input $X_{\text{self}}$ is:

\begin{equation}
Z_p = \text{Conv1D}(X_p) + \text{PE}(X_p)
\end{equation}

Conv1D includes a kernel that maps each individual’s features into a common dimension $d_z$.

Behavior Interaction Encoder: Crossperson Transformer & Self Transformer

Crossperson Attention (CPA) is a core component of the $M$-layered Crossperson Transformer (CPT). CPA$_{\text{other} \rightarrow \text{self}}^{\text{multi}}$ refers to the multi-head Crossperson Attention from person \textit{other} to person \textit{self} at the $m$-th layer. Following standard transformer operations \cite{Vaswani2017}, \textit{\gamma}$_{\text{other} \rightarrow \text{self}}^{\text{m}}$ refers to the intermediate output after the Crossperson Attention with residual connections. \textit{\gamma}$_{\text{other} \rightarrow \text{self}}^{\text{m}}$ 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, \textit{\gamma}$_{\text{other} \rightarrow \text{self}}^{\text{0}} = Z_{\text{other}}$. We refer the readers to Figure 1(b) for details.

\begin{equation}
\begin{align*}
\text{\textit{\gamma}$_{\text{other} \rightarrow \text{self}}^{\text{m}}$} & = \text{CPT$_{\text{other} \rightarrow \text{self}}^{\text{m}}$} (\text{\textit{\gamma}$_{\text{other} \rightarrow \text{self}}^{\text{m-1}}$, Z_{\text{self}}) \\
\text{\textit{\gamma}$_{\text{other} \rightarrow \text{self}}^{\text{m}}$} & = \text{CPA$_{\text{other} \rightarrow \text{self}}^{\text{multi}}$} (\text{Norm}(\text{\textit{\gamma}$_{\text{other} \rightarrow \text{self}}^{\text{m-1}}$, Z_{\text{self}}) + \text{Norm}(Z_{\text{self}})) + \text{Norm}(\text{\textit{\gamma}$_{\text{other} \rightarrow \text{self}}^{\text{m-1}}$, Z_{\text{self}})) \\
\text{\textit{\gamma}$_{\text{other} \rightarrow \text{self}}^{\text{0}}$} & = \text{Norm}(\text{FFN}(\text{\textit{\gamma}$_{\text{other} \rightarrow \text{self}}^{\text{m}}$) + \text{\textit{\gamma}$_{\text{other} \rightarrow \text{self}}^{\text{m}}$})
\end{align*}
\end{equation}

With this formulation, CPA$_{\text{other} \rightarrow \text{self}}^{\text{0}}$($Z_{\text{other}}, Z_{\text{self}}$) 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
Figure 1: Full model architectures (a) Diagram of our Crossperson Attention (CPA\(_{other→self}\)) 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’ 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 \(\text{other} \in P \setminus \{self\}\), CPA\(_{other→self}\), 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 CPA\(_{self→self}\)(\(Z_{self}, Z_{self}\)). 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, \(\mathbf{[::]}\) 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 \(Y_{self}\).

\[
\zeta_{self,\text{hidden}}^{n} = \text{LSTM}([\gamma_{1→self}^{M}], \ldots, [\gamma_{P→self}^{M}], \text{hidden}^{n−1})
\]

\[
Y_{self} = \text{FFN}(\zeta_{self,\text{hidden}}^{n}) \quad \text{for } n \in [k]
\]

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. Online interaction settings are well-suited to study contingent behaviors as nonverbal online actions, indication of online presence, demarcation of others’ contributions, and expressions of support affect interactants’ experiences in video-conferencing [Park and Whiting, 2020]. Engagement is focused on off-task/on-task engagement, where the task at hand is led by the instructor. Roomreader consists of 5-person and 4-person settings. In our experiments, we focused on the 5-person setting which is the majority of the dataset (26/30 groups, 106/118 unique participants).

**Engagement Classification** RoomReader provides continuous annotations for engagement, where the 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]. 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 to focus learning on hard misclassified examples. $y_{i,c}$ refers to the ground truth labels, $\hat{y}_{i,c}$ is the probability prediction, and $\alpha$ is a hyperparameter that weighs how much easy samples should be down-weighted.

$$L_{\text{Focal}} = -\frac{1}{N} \sum_{i} \sum_{c} (1 - \hat{y}_{i,c})^\alpha y_{i,c} \log (\hat{y}_{i,c}) \quad (6)$$

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 Ducchietto et al., 2020], OICCNN-LSTM [Steinert et al., 2020], TEMMA [Chen et al., 2020a], EnsModel [Thong Huynh et al., 2019], BOOT [Wang et al., 2019], HTMLIL [Ma et al., 2021], GAT [Zhang et al., 2022], and MultiT [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. MultiPar-T can be used in real-time, where the inference time only takes $0.0981 \pm 0.0029$ 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 Ducchietto 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.

### Table 1: Results and standard deviations for engagement recognition models for 3 seeds (standard deviation for High Dis-Engagement not reported due to 2 seeds without 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.

<table>
<thead>
<tr>
<th>Model</th>
<th>All Engagement Classes</th>
<th>High Dis-Eng.</th>
<th>Low Dis-Eng.</th>
<th>Low Eng.</th>
<th>High Eng.</th>
<th># Params</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy ↑</td>
<td>Weighted F1 ↑</td>
<td>Macro F1 ↑</td>
<td>F1 ↑</td>
<td>F1 ↑</td>
<td>F1 ↑</td>
</tr>
<tr>
<td>ConvLSTM [Del Ducchietto et al., 2020]</td>
<td>0.859 ± 0.01</td>
<td>0.857 ± 0.02</td>
<td>0.699 ± 0.05</td>
<td>0.741 ± 0.09</td>
<td>0.459 ± 0.22</td>
<td>0.699 ± 0.12</td>
</tr>
<tr>
<td>OICCNN-LSTM [Steinert et al., 2020]</td>
<td>0.769 ± 0.08</td>
<td>0.695 ± 0.14</td>
<td>0.410 ± 0.10</td>
<td>0.588 ± 0.11</td>
<td>0.19 ± 0.17</td>
<td>0.233 ± 0.33</td>
</tr>
<tr>
<td>TEMMA [Chen et al., 2020a]</td>
<td>0.823 ± 0.05</td>
<td>0.822 ± 0.02</td>
<td>0.561 ± 0.11</td>
<td>0.286 ± 0.24</td>
<td>0.254 ± 0.19</td>
<td>0.621 ± 0.13</td>
</tr>
<tr>
<td>EnsModel [Thong Huynh et al., 2019]</td>
<td>0.760 ± 0.07</td>
<td>0.675 ± 0.12</td>
<td>0.302 ± 0.03</td>
<td>0.000 ± 0.00</td>
<td>0.000 ± 0.00</td>
<td>0.160 ± 0.23</td>
</tr>
<tr>
<td>BOOT [Wang et al., 2019]</td>
<td>0.817 ± 0.03</td>
<td>0.822 ± 0.03</td>
<td>0.636 ± 0.09</td>
<td>0.714 ± 0.32</td>
<td>0.24 ± 0.12</td>
<td>0.658 ± 0.12</td>
</tr>
<tr>
<td>HTMLIL [Ma et al., 2021]</td>
<td>0.820 ± 0.02</td>
<td>0.818 ± 0.02</td>
<td>0.460 ± 0.05</td>
<td>0.000 ± 0.00</td>
<td>0.000 ± 0.00</td>
<td>0.633 ± 0.12</td>
</tr>
<tr>
<td>GAT [Zhang et al., 2022]</td>
<td>0.739 ± 0.06</td>
<td>0.631 ± 0.08</td>
<td>0.261 ± 0.03</td>
<td>0.000 ± 0.00</td>
<td>0.000 ± 0.00</td>
<td>0.006 ± 0.01</td>
</tr>
<tr>
<td>MultiT [Tsai et al., 2019]</td>
<td>0.847 ± 0.02</td>
<td>0.845 ± 0.02</td>
<td>0.624 ± 0.12</td>
<td>0.625 ± 0.31</td>
<td>0.25 ± 0.12</td>
<td>0.665 ± 0.12</td>
</tr>
<tr>
<td>MultiPar-T (Ours)</td>
<td>0.888 ± 0.03</td>
<td>0.887 ± 0.03</td>
<td>0.751 ± 0.05</td>
<td>0.800 ± 0.02</td>
<td>0.559 ± 0.07</td>
<td>0.759 ± 0.11</td>
</tr>
</tbody>
</table>

In infrequent disengagement classes. MultiPar-T outperforms other approaches across all metrics.

6.2 Quantitative Results

Following previous works [Del Ducchietto 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 state-of-the-art approaches in accuracy by 2.9%, weighted F1 by 3.0%, and Macro-F1 by 5.2%. 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 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 behaviors. We also compare with an adaptation of the multimodal transformer (MuT) [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 CPA direction self → other in 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 multimodal alignment, as we demonstrate via ablation studies in the following 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 experiment with the directions of the Crossperson Attention mechanism. In Table 2, CPA: other → self refers to our 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 → 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 with other → self and self → other. We find that MultiPar-T with our formulation of cross attention, CPA: other → 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 → other results in better performance. This shows that how the self impacts others could be an important predictor when predicting if the target person is disengaged.

**Encoding Size & Transformer Layers** In Figure 3, we first display results for ablations on varying number transformer 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 leads to a significant improvement in Macro-F1. Secondly, we display results for varying sizes of the embedding dimensions per timestep d_e. We find that the optimal encoding dimension to encode behavior per timestep setting is d_e = 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 is computationally much more expensive, due to the fact that the models are 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, one seed takes ~ 150 minutes for a raw video-based models, compared to ~ 20 minutes for an engagement prediction models. For a fair comparison, we train MultiPar-T with the same training settings and find that MultiPar-T performs better than other architectures given the same training conditions for scarcer disengagement samples.
Person other’s Behavior

Target Person self’s Behavior

(a) Crossperson Attention Contingency

(b) Diagonal Contingency

(c) Uniform Contingency

Figure 4: (a) Multiparty-Transformer Cross-person Attention weights from \( t = 120 \) 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.

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 a larger number of datasets and benchmarks consisting of different contexts including in-person settings, various relationships, diverse cultures, and a wider range of age and number 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.

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 [Wiegrefe 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 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 why the model made this specific prediction for this specific timestep. The attention weights visualized in Figure 4(c) Diagonal Contingency, is a diagonal attention weight matrix, which indicates that Person self’s past behavior is uniformly related to Person self’s current behavior. 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.

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, 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.

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy ↑</th>
<th>Weighted F1 ↑</th>
<th>Macro F1 ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>I3D [Wang et al., 2018]</td>
<td>0.751 ± 0.07</td>
<td>0.658 ± 0.08</td>
<td>0.254 ± 0.05</td>
</tr>
<tr>
<td>TimeSformer [Rettasius et al., 2021]</td>
<td>0.806 ± 0.03</td>
<td>0.752 ± 0.05</td>
<td>0.337 ± 0.14</td>
</tr>
<tr>
<td>SlowFast [Feichtenhofer et al., 2019]</td>
<td>0.718 ± 0.11</td>
<td>0.628 ± 0.12</td>
<td>0.232 ± 0.02</td>
</tr>
<tr>
<td>MultiPar-T (Ours)</td>
<td>0.828 ± 0.02</td>
<td>0.823 ± 0.02</td>
<td>0.466 ± 0.06</td>
</tr>
</tbody>
</table>

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.
Acknowledgments

This work was supported by the IITP grant funded by the Korean Ministry of Science and ICT (No.2020-0-00842, Development of Cloud Robot Intelligence for Continual Adaptation to User Reactions in Real Service Environments). Yubin Kim was supported by Korean Ministry of Trade, Industry, and Energy (MOTIE), under Human Resource Development Program for Industrial Innovation (Global, P0017311) supervised by the Korea Institute for Advancement of Technology.

References


