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,

https://github.com/dondongwon/Multipar-T

IJCAI 2023

2023. 6. 8 • ChongQing

Leibniz-Institut für Sozialwissenschaften

Reported by JiaWei Cheng

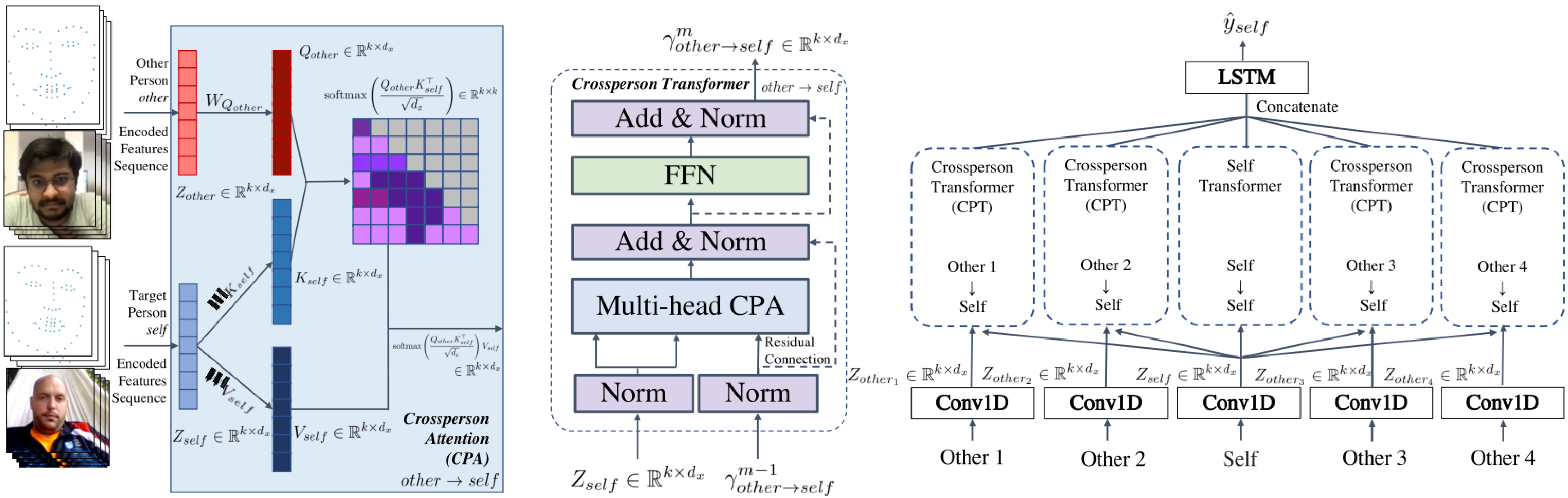
Motivation

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

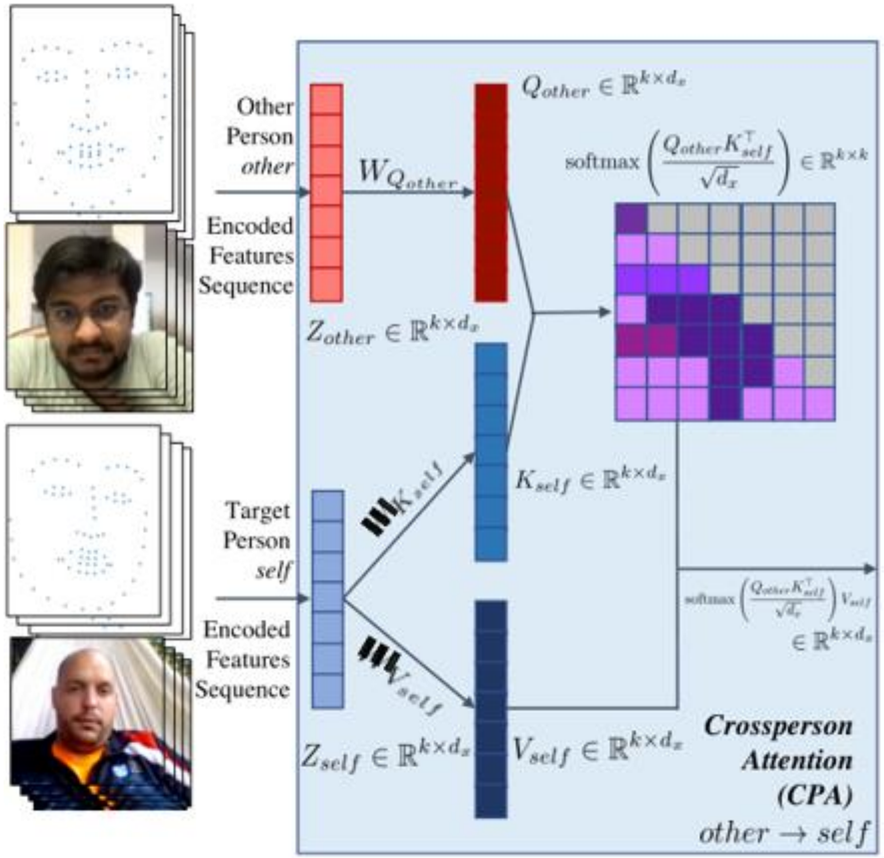
Finally, it must also recognize the subtle interactions that take place between group members as it can provide more insights into what is being communicated.

Overview



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

Method



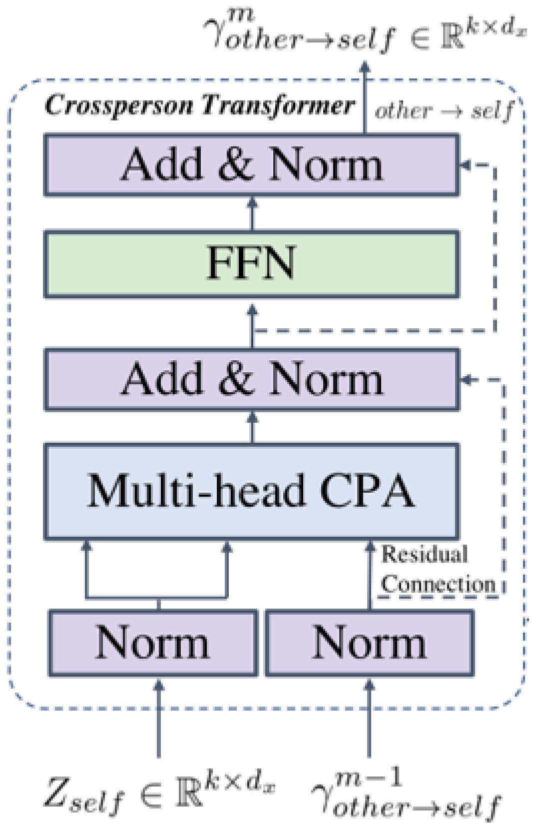
(a) Crossperson Attention

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

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

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

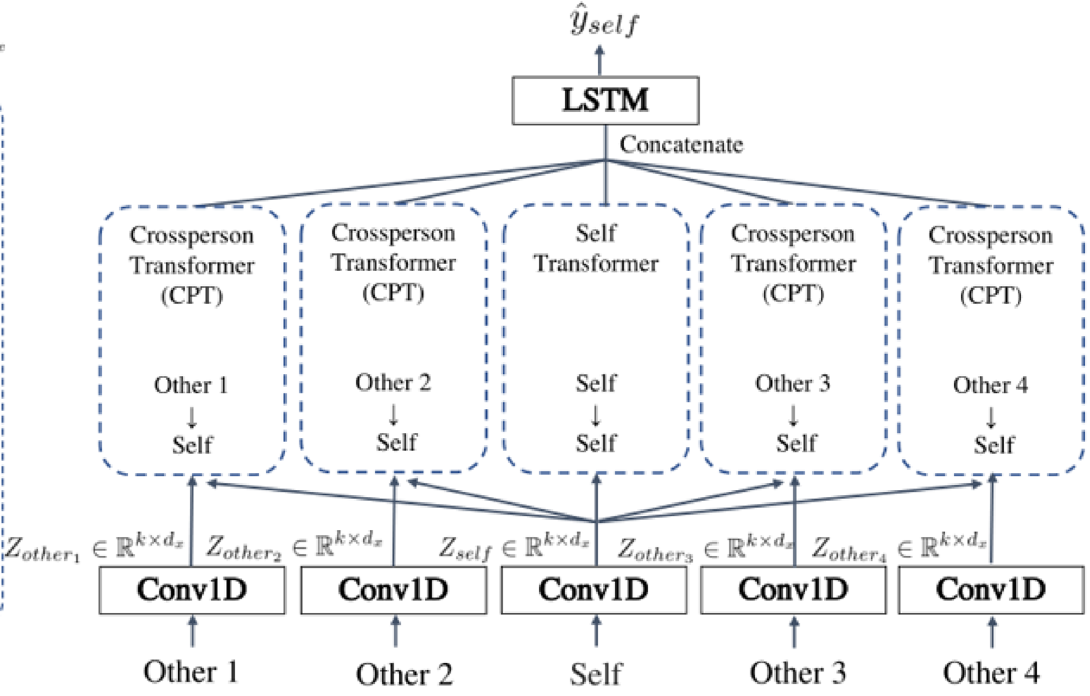
Method



(b) Crossperson Transformer

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

Method



(c) Multiparty-Transformer (MultiPar-T)

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

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

Experiments

	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.

Experiments

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.

Experiments

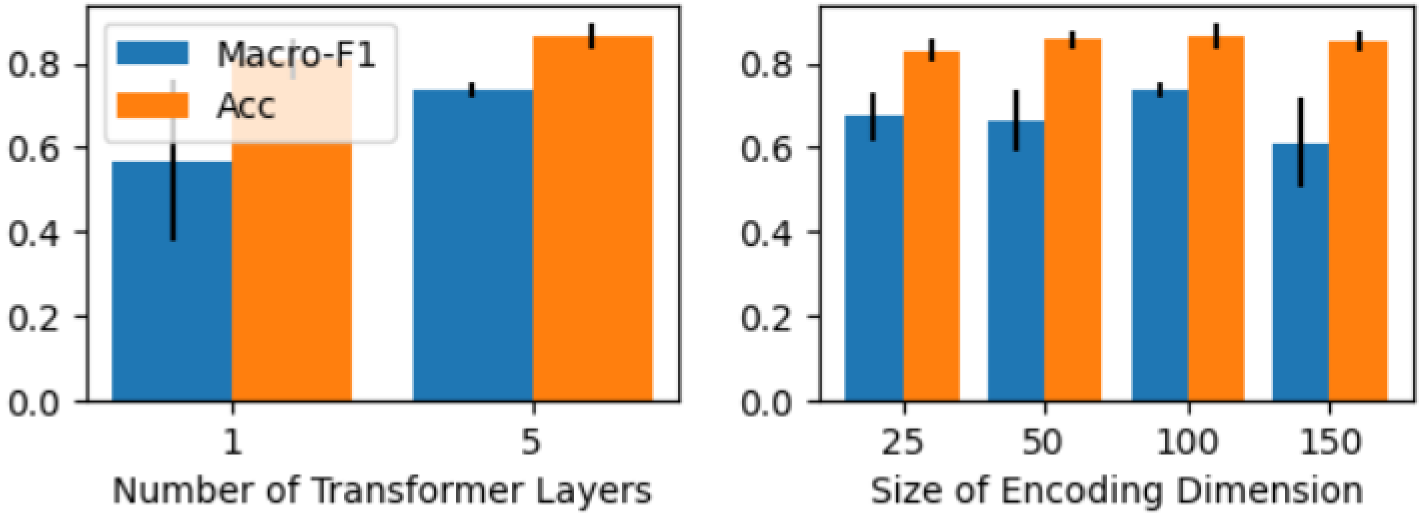


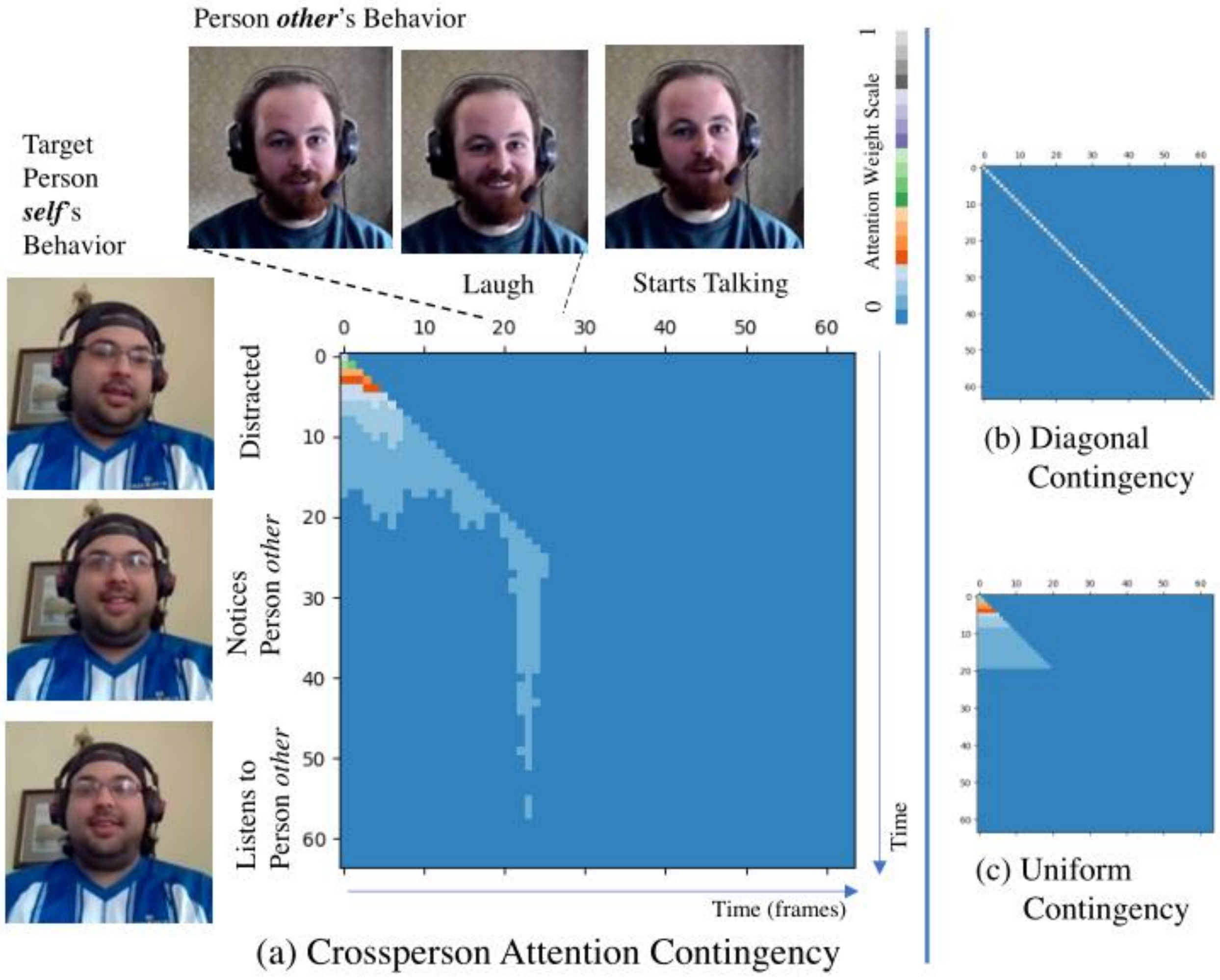
Figure 3: Macro-F1 and Accuracy scores for important hyperparameters for Multipar-T. (Left) Multi-layered transformers and (Right) encoding dimension of $d_x = 100$ boosts performance.

Experiments

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.

Experiments



Thanks!