M3PT: A Transformer for Multimodal, Multi-Party Social Signal Prediction with Person-aware Blockwise Attention USC

Yiming Tang, Abrar Anwar, Jesse Thomason

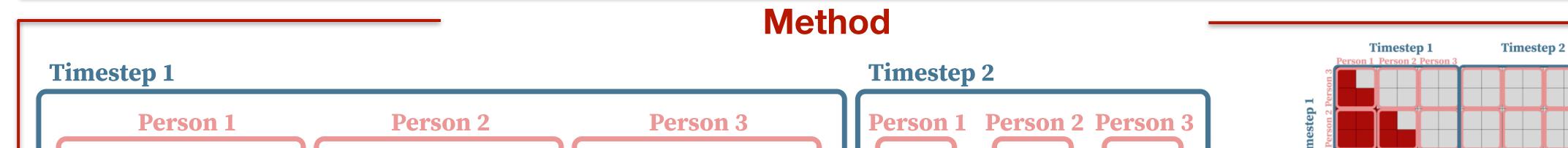
Motivation

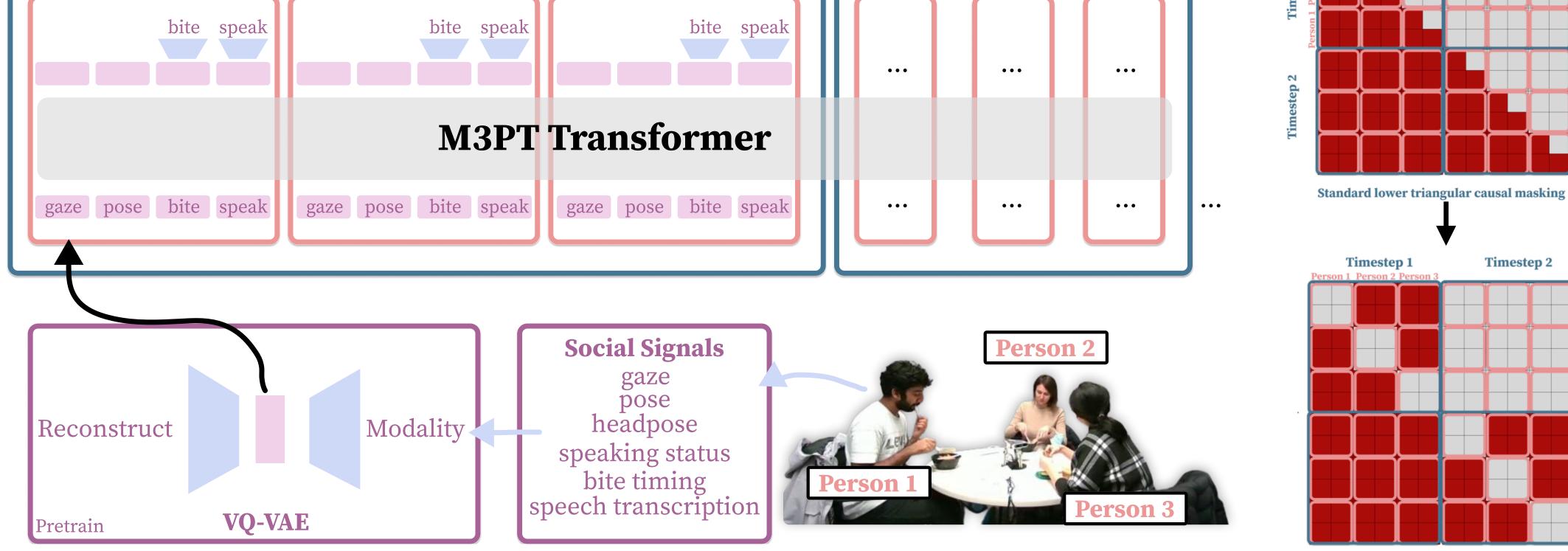
- Language, gestures, and gaze occur simultaneously
- Auto-regressive transformers are causally left-to-right, not helping in modeling concurring signals
- We introduce **M3PT**, a causal transformer architecture with modality and temporal blockwise attention masking
- Human-human Commensensality Dataset (HHCD)
 - 30 triadic session of 90 people eating
 - 18+ hours of video



30 triadic sessions of 90 people eating. 18+ hours of video







Timestep-aware masking Person-specific masking

Pretrained VQ-VAE to discretize social

Modality- and Temporal-specific Masking

					Experin	nents							
	Features	F1	Precision	Recall	nMCC]	Features	F 1	Precisi	on Recall	nMCC	•
	All Features	$0.86\pm$ 0.06	0.81± 0.13	$0.93\pm$ 0.04	$0.92\pm$ 0.03			All Featur	res 0.91 ± 0	$0.03 0.86 \pm 0$	$09 0.97 \pm 0.04$	0.94± 0.02	
Main Experiment	No Gaze No Headpose No Pose No Word	$\begin{array}{c} 0.61 \pm \ 0.35 \\ 0.83 \pm \ 0.17 \\ 0.89 \pm \ 0.06 \\ 0.73 \pm \ 0.30 \end{array}$	0.60 ± 0.33 0.81 ± 0.26 0.84 ± 0.13 0.73 ± 0.36	$\begin{array}{c} 0.62 \pm \ 0.37 \\ 0.91 \pm \ 0.00 \\ 0.97 \pm \ 0.03 \\ 0.77 \pm \ 0.19 \end{array}$	$\begin{array}{c} 0.78 \pm \ 0.19 \\ 0.91 \pm \ 0.08 \\ 0.94 \pm \ 0.03 \\ 0.84 \pm \ 0.18 \end{array}$]]]	No Gaze No Headr No Pose No Word	0.89 ± 0.000	$\begin{array}{ccc} 0.03 & 0.75 \pm \ 0 \\ 0.01 & 0.81 \pm \ 0 \\ 0.18 & 0.66 \pm \ 0 \end{array}$	$\begin{array}{cccc} 05 & 1.00 \pm 0.00 \\ 03 & 0.99 \pm 0.00 \\ 19 & 0.87 \pm 0.17 \end{array}$	$\begin{array}{c} 0.90 \pm \ 0.02 \\ 0.93 \pm \ 0.01 \\ 0.83 \pm \ 0.13 \end{array}$	
LAPEIIIIEIII	No Speaker Bite Only	0.24 ± 0.06 0.44 ± 0.35	0.16 ± 0.04 0.39 ± 0.36	0.46 ± 0.12 0.59 ± 0.29	$\frac{0.56 \pm 0.02}{0.69 \pm 0.19}$			No Bite Speaker C	0.77 ± 0 Only 0.89 ± 0				
	Bite Timing Prediction Removing any modality degrades performance, especially gaze and speaker, confirming that multiple social signals improves bite timing prediction.						Speaking Status Prediction : Using all features yields the best performance, with word and bite signals being the most informative for speaking status prediction.						
	Task	F1 Score	Precision	Recall	nMCC	_		Task	F1 Score	Precision	Recall	nMCC	
	$2 \times 3s \frac{S}{B}$	$\begin{array}{c} 0.99 \pm \ 0.00 \\ 0.99 \pm \ 0.00 \end{array}$	$\begin{array}{c} 0.99 \pm \ 0.00 \\ 0.99 \pm \ 0.01 \end{array}$	1.00 ± 0.00 1.00 ± 0.00	$\begin{array}{c} 0.99 \pm \ 0.00 \\ 0.99 \pm \ 0.00 \end{array}$	_	2×1	8s B B	$\begin{array}{c} 0.00 \pm \ 0.00 \\ 0.64 \pm \ 0.04 \end{array}$	$\begin{array}{c} 0.00 \pm \ 0.00 \\ 0.47 \pm \ 0.05 \end{array}$	$\begin{array}{c} 0.00 \pm \ 0.00 \\ 1.00 \pm \ 0.00 \end{array}$	$\begin{array}{c} 0.00 \pm \ 0.00 \\ 0.64 \pm \ 0.00 \end{array}$	
Ablation	$3 \times 3s \frac{S}{B}$	$\begin{array}{c} 0.99 \pm \ 0.00 \\ 0.97 \pm \ 0.02 \end{array}$	$\begin{array}{c} 0.99 \pm \ 0.00 \\ 0.94 \pm \ 0.04 \end{array}$	1.00 ± 0.00 1.00 ± 0.00	$\begin{array}{c} 0.99 \pm \ 0.00 \\ 0.98 \pm \ 0.01 \end{array}$	_	4×9	os S B	$\begin{array}{c} 0.12 {\pm}~0.18 \\ 0.43 {\pm}~0.08 \end{array}$	$\begin{array}{c} 0.08 \pm \ 0.12 \\ 0.27 \pm \ 0.06 \end{array}$	$\begin{array}{c} 0.25 \pm \ 0.35 \\ 1.00 \pm \ 0.00 \end{array}$	$\begin{array}{c} 0.51 {\pm}~0.00 \\ 0.52 {\pm}~0.05 \end{array}$	
	$6 \times 3s \frac{S}{B}$	1.00 ± 0.00 1.00 ± 0.00	1.00 ± 0.00 1.00 ± 0.00	$1.00\pm 0.00 \\ 1.00\pm 0.00$	$1.00\pm 0.00 \\ 1.00\pm 0.00$	_	6×6	os S B	$\begin{array}{c} 0.23 \pm \ 0.17 \\ 0.32 \pm \ 0.08 \end{array}$	$\begin{array}{c} 0.15 \pm \ 0.11 \\ 0.19 \pm \ 0.06 \end{array}$	$\begin{array}{c} 0.50 \pm \ 0.37 \\ 0.99 \pm \ 0.00 \end{array}$	$\begin{array}{c} 0.50 {\pm}~ 0.01 \\ 0.52 {\pm}~ 0.03 \end{array}$	
	$12 \times 3s \frac{S}{B}$	$\begin{array}{c} 0.91 \pm \ 0.03 \\ 0.86 \pm \ 0.06 \end{array}$	$\begin{array}{c} 0.86 \pm \ 0.09 \\ 0.81 \pm \ 0.13 \end{array}$	$\begin{array}{c} 0.97 \pm \ 0.04 \\ 0.93 \pm \ 0.04 \end{array}$	$\begin{array}{c} 0.94 {\pm}~0.02 \\ 0.92 {\pm}~0.03 \end{array}$	_	12×	3s B B	$\begin{array}{c} 0.91 {\pm}~0.03 \\ 0.86 {\pm}~0.06 \end{array}$	$\begin{array}{c} 0.86 {\pm} \ 0.09 \\ 0.81 {\pm} \ 0.13 \end{array}$	$\begin{array}{c} 0.97 \pm \ 0.04 \\ 0.93 \pm \ 0.04 \end{array}$	$\begin{array}{c} 0.94 {\pm}~0.02 \\ 0.92 {\pm}~0.03 \end{array}$	
	Increase	ed tempora	al context ir	ntroduces r	noise	ī	Long	ger seg	ment lengt	hs severely	v cause mod	e collapse	

						-					
	Features	F1	Precision	Recall	nMCC		Features	F1	Precisio	n Recall	nMCC
	All Features	$0.86\pm$ 0.06	0.81 ± 0.13	$0.93\pm$ 0.04	0.92 ± 0.03	_	All Featur	res 0.91 ± 0	.03 $0.86\pm 0.$	$0.900.97 \pm 0.04$	4 0.94 ± 0.02
	No Gaze	0.61± 0.35	0.60 ± 0.33	0.62 ± 0.37	0.78± 0.19	_	No Gaze	0.83 ± 0	$0.05 0.73 \pm 0.00$	0.97 ± 0.02	$2 0.89 \pm 0.03$
	No Headpose	$0.83\pm$ 0.17	0.81 ± 0.26	$0.91\pm$ 0.00	$0.91\pm$ 0.08		No Headp	bose 0.85 ± 0	.03 $0.75 \pm 0.$	1.00 ± 0.00	0.90 ± 0.02
Main	No Pose	$0.89\pm$ 0.06	0.84 ± 0.13	0.97 ± 0.03	0.94 ± 0.03		No Pose	0.89 ± 0	$.01 0.81 \pm 0.$	0.099 ± 0.00	0.93 ± 0.01
	No Word	$0.73\pm$ 0.30	0.73 ± 0.36	$0.77\pm$ 0.19	0.84 ± 0.18		No Word	0.75 ± 0	.18 $0.66 \pm 0.$	19 $0.87 \pm 0.1^{\circ}$	7 0.83 ± 0.13
Experiment	No Speaker	$0.24 \pm \ 0.06$	$0.16\pm$ 0.04	$0.46\pm$ 0.12	$0.56\pm$ 0.02	_	No Bite	0.77 ± 0	$.16 0.67 \pm 0.$	$16 0.90 \pm 0.12$	2 0.84 ± 0.11
	Bite Only	$0.44\pm$ 0.35	0.39 ± 0.36	0.59 ± 0.29	0.69± 0.19	_	Speaker C	Only 0.89 ± 0	$.15 0.88 \pm 0.$	$16 0.90 \pm 0.13$	0.92 ± 0.10
	Bite Timing	Prediction	Removing a	any modality	y degrades	Spe	aking S	tatus Predi	ction : Using	g all features	yields the b
				nd snaakor	, confirming tha	ŀ	nerform	ance with	vord and bi	te signals be	ing the mos
	perform	ance, espec	Jally gaze a	Ind speaker,	, comming tha	L		ianoo, with t			
	•	•		•	•	L	•			0	9
	•	•		•	g prediction.	L	•	tive for spea		0	
	multiple	social sign	als improve	s bite timing	prediction.	L	informa	tive for spea	aking status	prediction.	
	•	•		•	•		•			0	nMCC
	Task	social signa F1 Score	als improves Precision	s bite timing Recall	nMCC		informa Task	tive for spea F1 Score	aking status Precision	prediction. Recall	nMCC
	Task	social signation $F1$ Score 0.99 ± 0.00	als improves Precision 0.99 ± 0.00	s bite timing Recall 1.00± 0.00	nMCC 0.99 ± 0.00		informa Task	tive for spea F1 Score 0.00 ± 0.00	hking status Precision 0.00 ± 0.00	prediction. Recall 0.00 ± 0.00	nMCC 0.00± 0.00
	Task 2×3s B B	social signation $F1$ Score 0.99 ± 0.00 0.99 ± 0.00	als improves Precision 0.99 ± 0.00 0.99 ± 0.01	s bite timing Recall 1.00 ± 0.00 1.00 ± 0.00	nMCC 0.99 ± 0.00 0.99 ± 0.00		informa Task 18s B B	tive for spea F1 Score 0.00 ± 0.00 0.64 ± 0.04	hking status Precision 0.00 ± 0.00 0.47 ± 0.05	prediction. Recall 0.00 ± 0.00 1.00 ± 0.00	$nMCC \\ 0.00 \pm 0.00 \\ 0.64 \pm 0.00$
	$\frac{1}{2 \times 3s} = \frac{S}{S}$	social signation $F1$ Score 0.99 ± 0.00 0.99 ± 0.00 0.99 ± 0.00	als improves Precision 0.99 ± 0.00 0.99 ± 0.01 0.99 ± 0.00	Recall 1.00 ± 0.00 1.00 ± 0.00 1.00 ± 0.00	nMCC 0.99 ± 0.00 0.99 ± 0.00 0.99 ± 0.00	2×	informa Task 18s B 18s B	tive for spea F1 Score 0.00 ± 0.00 0.64 ± 0.04 0.12 ± 0.18	hking status Precision 0.00 ± 0.00 0.47 ± 0.05 0.08 ± 0.12	prediction. Recall 0.00 ± 0.00 1.00 ± 0.00 0.25 ± 0.35	$nMCC \\ 0.00 \pm 0.00 \\ 0.64 \pm 0.00 \\ 0.51 \pm 0.00$
Ablation	Task 2×3s B B	social signation $F1$ Score 0.99 ± 0.00 0.99 ± 0.00	als improves Precision 0.99 ± 0.00 0.99 ± 0.01	s bite timing Recall 1.00 ± 0.00 1.00 ± 0.00	nMCC 0.99 ± 0.00 0.99 ± 0.00	2×	informa Task 18s B B	tive for spea F1 Score 0.00 ± 0.00 0.64 ± 0.04	hking status Precision 0.00 ± 0.00 0.47 ± 0.05	prediction. Recall 0.00 ± 0.00 1.00 ± 0.00	$nMCC \\ 0.00 \pm 0.00 \\ 0.64 \pm 0.00$
Ablation	$ \begin{array}{c} \text{multiple}\\ \hline Task\\ 2 \times 3s \begin{array}{c}S\\B\\\hline 3 \times 3s \begin{array}{c}S\\B\\\hline \end{array}\\ \end{array} $	social signation $F1$ Score 0.99 ± 0.00 0.99 ± 0.00 0.99 ± 0.00	als improves Precision 0.99 ± 0.00 0.99 ± 0.01 0.99 ± 0.00	Recall 1.00 ± 0.00 1.00 ± 0.00 1.00 ± 0.00	nMCC 0.99 ± 0.00 0.99 ± 0.00 0.99 ± 0.00	2× 4×	informa Task 18s B 18s B 18s B Stanting	tive for spea F1 Score 0.00 ± 0.00 0.64 ± 0.04 0.12 ± 0.18 0.43 ± 0.08	Precision 0.00 ± 0.00 0.47 ± 0.05 0.08 ± 0.12 0.27 ± 0.06	o Recall 0.00 ± 0.00 1.00 ± 0.00 0.25 ± 0.35 1.00 ± 0.00	$\begin{array}{c} \mathbf{nMCC} \\ 0.00 \pm \ 0.00 \\ 0.64 \pm \ 0.00 \\ 0.51 \pm \ 0.00 \\ 0.52 \pm \ 0.05 \end{array}$
Ablation	$ \begin{array}{c} \text{multiple} \\ \hline Task \\ 2 \times 3s \begin{array}{c} S \\ B \\ \hline 3 \times 3s \begin{array}{c} S \\ B \\ \end{array} $	Social signation $F1$ Score 0.99 ± 0.00 0.99 ± 0.00 0.99 ± 0.00 0.99 ± 0.00 0.97 ± 0.02	als improves Precision 0.99 ± 0.00 0.99 ± 0.01 0.99 ± 0.00 0.99 ± 0.00 0.94 ± 0.04	Recall 1.00 ± 0.00	nMCC 0.99 ± 0.00 0.99 ± 0.00 0.99 ± 0.00 0.99 ± 0.00 0.99 ± 0.00 0.98 ± 0.01	2× 4×	informa Task 18s B 18s B	tive for spea F1 Score 0.00 ± 0.00 0.64 ± 0.04 0.12 ± 0.18	hking status Precision 0.00 ± 0.00 0.47 ± 0.05 0.08 ± 0.12	prediction. Recall 0.00 ± 0.00 1.00 ± 0.00 0.25 ± 0.35	$nMCC \\ 0.00 \pm 0.00 \\ 0.64 \pm 0.00 \\ 0.51 \pm 0.00$
Ablation	$ \begin{array}{c} \text{Frank}\\ \hline \text{Task}\\ 2 \times 3s \begin{array}{c}S\\B\\\hline 3 \times 3s \begin{array}{c}S\\B\\\hline \end{array} \\ 6 \times 3s \begin{array}{c}S\\S\\\end{array} \end{array} $	Social signation of the second secon	als improves Precision 0.99 ± 0.00 0.99 ± 0.01 0.99 ± 0.00 0.99 ± 0.00 0.94 ± 0.04 1.00 ± 0.00	Recall 1.00 ± 0.00	nMCC 0.99 ± 0.00 1.00 ± 0.00	2× 4× 6×	informa Task 18s B 18s B 18s B 18s B 18s B 18s S B	tive for spea F1 Score 0.00 ± 0.00 0.64 ± 0.04 0.12 ± 0.18 0.43 ± 0.08 0.23 ± 0.17	Precision 0.00 ± 0.00 0.47 ± 0.05 0.08 ± 0.12 0.27 ± 0.06 0.15 ± 0.11	prediction. Recall 0.00 ± 0.00 1.00 ± 0.00 0.25 ± 0.35 1.00 ± 0.00 0.50 ± 0.37	$\begin{array}{c} \mathbf{nMCC} \\ 0.00 \pm \ 0.00 \\ 0.64 \pm \ 0.00 \\ 0.51 \pm \ 0.00 \\ 0.52 \pm \ 0.05 \\ 0.50 \pm \ 0.01 \end{array}$

Exporimonto