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

Yiming Tang, Abrar Anwar, Jesse Thomason

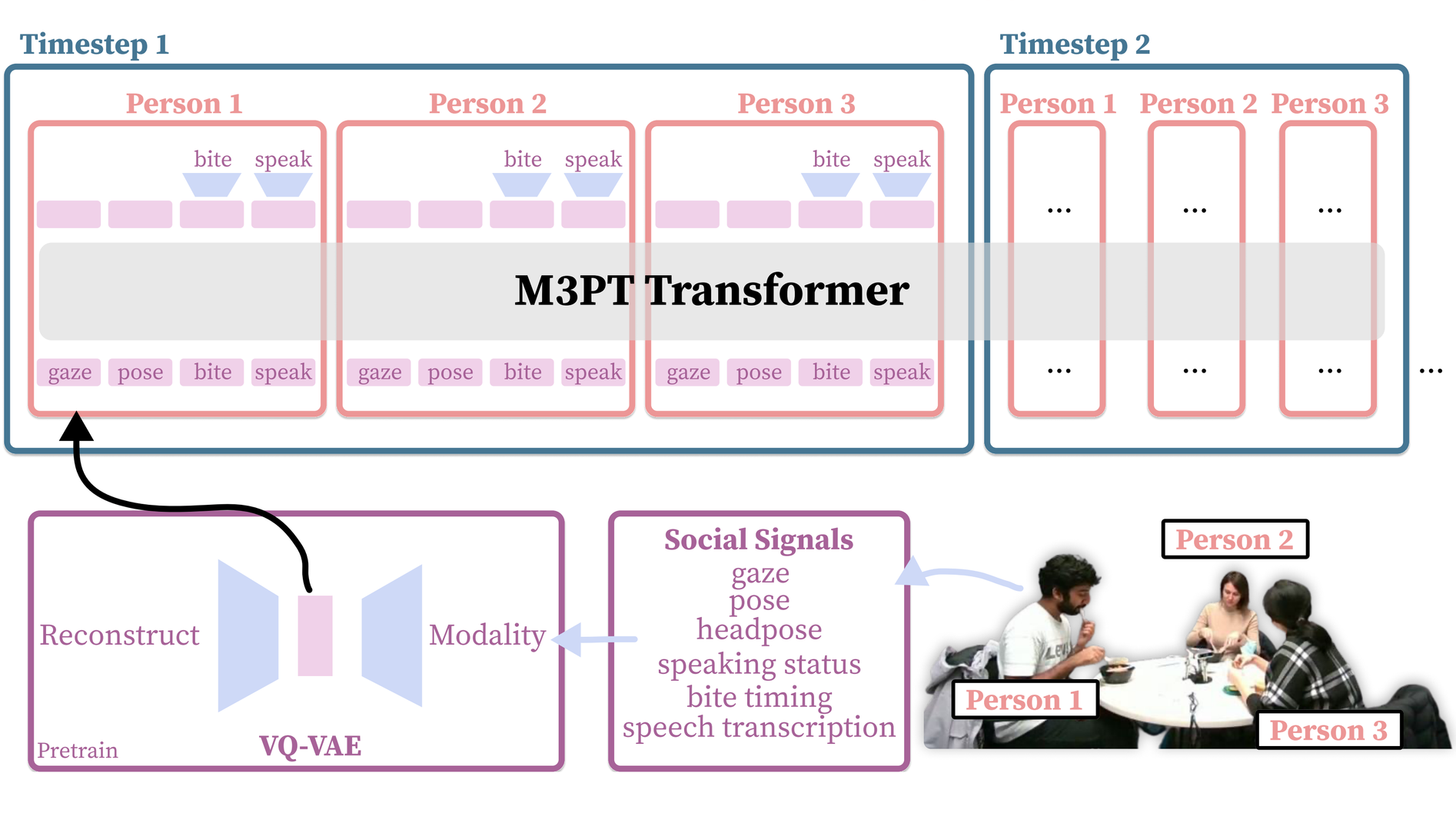
USC

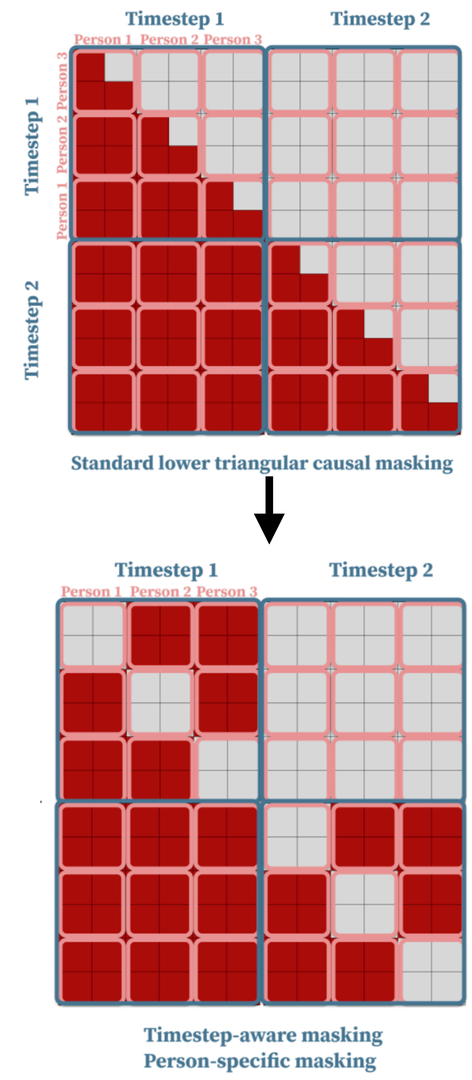
Introduction

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

Method





Pretrained VQ-VAE to discretize social

Modality- and Temporal-specific Masking

Experiments

Main Experiment

Features	F1	Precision	Recall	nMCC
All Features	0.86± 0.06	0.81± 0.13	0.93± 0.04	0.92± 0.03
No Gaze	0.61± 0.35	0.60± 0.33	0.62± 0.37	0.78± 0.19
No Headpose	0.83± 0.17	0.81± 0.26	0.91± 0.00	0.91± 0.08
No Pose	0.89± 0.06	0.84± 0.13	0.97± 0.03	0.94± 0.03
No Word	0.73± 0.30	0.73± 0.36	0.77± 0.19	0.84± 0.18
No Speaker	0.24± 0.06	0.16± 0.04	0.46± 0.12	0.56± 0.02
Bite Only	0.44± 0.35	0.39± 0.36	0.59± 0.29	0.69± 0.19

Bite Timing Prediction Removing any modality degrades performance, especially gaze and speaker, confirming that multiple social signals improves bite timing prediction.

Features	F1	Precision	Recall	nMCC
All Features	0.91± 0.03	0.86± 0.09	0.97± 0.04	0.94± 0.02
No Gaze	0.83± 0.05	0.73± 0.08	0.97± 0.02	0.89± 0.03
No Headpose	0.85± 0.03	0.75± 0.05	1.00± 0.00	0.90± 0.02
No Pose	0.89± 0.01	0.81± 0.03	0.99± 0.00	0.93± 0.01
No Word	0.75± 0.18	0.66± 0.19	0.87± 0.17	0.83± 0.13
No Bite	0.77± 0.16	0.67± 0.16	0.90± 0.12	0.84± 0.11
Speaker Only	0.89± 0.15	0.88± 0.16	0.90± 0.13	0.92± 0.10

Speaking Status Prediction: Using all features yields the best performance, with word and bite signals being the most informative for speaking status prediction.

Ablation

	Task	F1 Score	Precision	Recall	nMCC
2×3s	S	0.99± 0.00	0.99± 0.00	1.00± 0.00	0.99± 0.00
	B	0.99± 0.00	0.99± 0.01	1.00± 0.00	0.99± 0.00
3×3s	S	0.99± 0.00	0.99± 0.00	1.00± 0.00	0.99± 0.00
	B	0.97± 0.02	0.94± 0.04	1.00± 0.00	0.98± 0.01
6×3s	S	1.00± 0.00	1.00± 0.00	1.00± 0.00	1.00± 0.00
	B	1.00± 0.00	1.00± 0.00	1.00± 0.00	1.00± 0.00
12×3s	S	0.91± 0.03	0.86± 0.09	0.97± 0.04	0.94± 0.02
	B	0.86± 0.06	0.81± 0.13	0.93± 0.04	0.92± 0.03

Increased temporal context introduces noise

	Task	F1 Score	Precision	Recall	nMCC
2×18s	S	0.00± 0.00	0.00± 0.00	0.00± 0.00	0.00± 0.00
	B	0.64± 0.04	0.47± 0.05	1.00± 0.00	0.64± 0.00
4×9s	S	0.12± 0.18	0.08± 0.12	0.25± 0.35	0.51± 0.00
	B	0.43± 0.08	0.27± 0.06	1.00± 0.00	0.52± 0.05
6×6s	S	0.23± 0.17	0.15± 0.11	0.50± 0.37	0.50± 0.01
	B	0.32± 0.08	0.19± 0.06	0.99± 0.00	0.52± 0.03
12×3s	S	0.91± 0.03	0.86± 0.09	0.97± 0.04	0.94± 0.02
	B	0.86± 0.06	0.81± 0.13	0.93± 0.04	0.92± 0.03

Longer segment lengths severely cause mode collapse