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

Yiming Tang, Abrar Anwar, and Jesse Thomason
University of Southern California

***Abstract*— Understanding social signals in multi-party conversations is important for human-robot interaction and artificial social intelligence. Social signals include body pose, head pose, speech, and context-specific activities like acquiring and taking bites of food when dining. Past work in multi-party interaction tends to build task-specific models for predicting social signals. In this work, we address the challenge of predicting multimodal social signals in multi-party settings in a single model. We introduce M3PT, a causal transformer architecture with modality and temporal blockwise attention masking to simultaneously process multiple social cues across multiple participants and their temporal interactions. We train and evaluate M3PT on the Human-Human Commensality Dataset (HHCD), and demonstrate that using multiple modalities improves bite timing and speaking status prediction. Source code: https://github.com/AbrarAnwar/masked-social-signals/.**

I. Introduction

Human behavior in social interactions is shaped by a continuous dance of multimodal signals [16], [17]. There are dozens of social signals such as gesture, head orientation, gaze, and speech, that humans send during interactions. In shared social settings like dining, where individuals engage in simultaneous verbal and non-verbal communication, these signals do not exist in isolation. Instead, each person's behavior is intertwined with the actions of others. A person's gaze may shift in response to a conversational partner's body language, or they may pause mid-sentence as another prepares to take a bite of food. Research on learning from groups [2], [15] typically focuses on learning individual features such as facial features [26], head pose [35], body pose [28], hands pose [25], end-of-turn prediction [18], eye contact detection [20], averted gaze prediction [21], or bite timing [29] from other interactants. Understanding such social signals is important for affective computing [34], [19], social robotics [22], [33], navigation [8], [12], or robot-assisted feeding [29]. In multi-party settings, co-occurring social signals have been jointly modeled in prior work on rapport [24] and leadership detection [23]. To effectively capture such interactions, models of social behavior must learn from diverse social signals while accounting for inherent temporal dependencies.

In this work, we propose a causal transformer model M3PT (**M**ulti-**M**odal, **M**ulti-**P**arty **T**ransformer), which leverages modality-specific and time-based blockwise attention masking so that a single model can predict and leverage multiple features of social signals. This architecture is capable of leveraging information from multiple modalities — such as body pose, head orientation, gaze direction, speech, and others — while also attending to some length of past history. By processing multimodal features over time, the model can learn the choreography of social signals and predict an individual's behaviors.

We leverage M3PT on the Human-Human Commensality (HHCD) dataset [29] of triadic dining. The dataset was constructed for predicting when someone takes a bite of food, which is called bite timing, and contains various multimodal signals, such as speech, body pose, gaze, food interactions, and more. In this work, we demonstrate that M3PT is able to predict an individual's behavior based on the multimodal cues of others in their group. This task, particularly bite timing, is an important problem as an autonomous feeding system can help people with mobility limitations be fed autonomously with a robot in social settings [3]. We then show the importance of including multiple modalities to predict these social signals, along with ablations on the role of larger temporal contexts and temporal chunking.

II. Multimodal Social Signal Prediction Task

The goal of a social prediction task is to predict the social behaviors of a target person by using social cues from their interlocutors. Formally, at timestep t, an individual i gives off signals $X_i(t)$. $X_i(t)$ contains all the social signals a person gives off such as gaze, body gesture, and speech. In a scenario with n interactants, the social signals of person i are dependent on those from other interactants $j, j \neq i$ from the current and past timesteps in addition to social signals of person i from previous timesteps. Thus the goal of social signal prediction is to learn a function $\mathcal{F}$ that is able to predict the social signals of an individual i at timestep t:

$$X_i(t) = \mathcal{F}(X_i(0:t-1), \{X_j(0:t), \forall j \neq i\}). \quad (1)$$

Learned function $\mathcal{F}$ is meant to predict social signals of arbitrary participants. We consider each timestep t to be an interval of length c as opposed to an absolute point of time. For example, a social signal such as gaze at timestep t is represented as a c-second gaze signal segment. Then, each of these segments can be downsampled to any framerate.

We note that there is a distinction between social signal prediction and social signal forecasting, as the latter involves predicting signals in some time frame in the future, while prediction involves predicting concurrent signals. In this work, we limit our scope to signal prediction, which includes conditioning on concurrent features from other interactants.

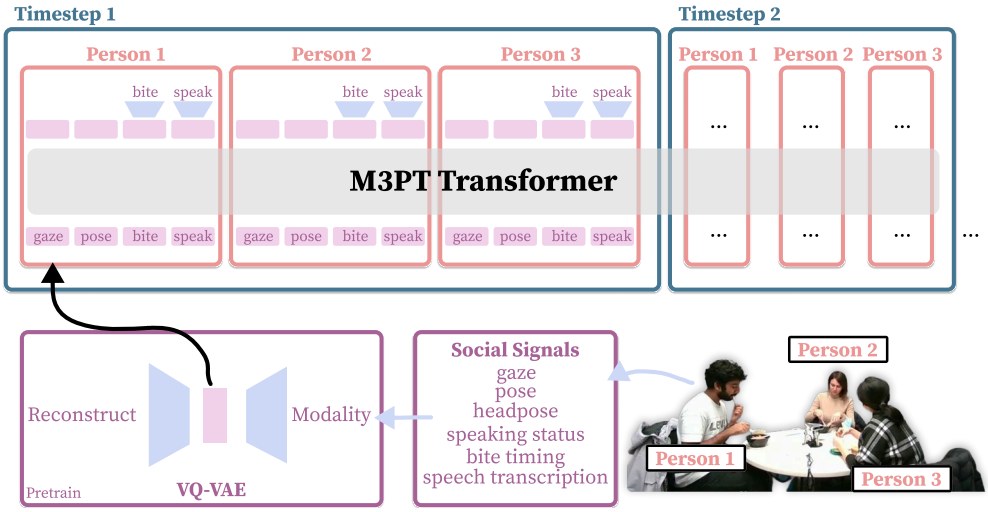

Fig. 1: **Model Architecture.** M3PT consists of a multi-party, multimodal causal transformer that attends to social signal features across people and timesteps. This transformer allows our model to reason about the interactants in multi-party settings over time. We encode these features with a VQ-VAE to tokenize continuous social signals for the transformer. We apply time-specific, person-specific, and modality-specific positional encodings along with a blockwise attenttion masking strategy, as shown in Fig. 2 to allow for the transformer to learn the relationship between these inputs. Then, we reconstruct the discrete social signals, speaking status and bite timing.

III. M3PT (Multi-Modal, Multi-Party Transformer)

We introduce M3PT (**M**ulti-**M**odal, **M**ulti-**P**arty **T**ransformer), a model designed for multi-party social signal prediction. Unlike previous work that focuses on predicting a single social signal with a small time horizon or a single social signal, M3PT simultaneously considers and predicts multiple social signals.

M3PT is designed to predict discrete social signals, but can handle a wide array of multimodal social signals, such as gaze, head pose, pose, and speech transcriptions as inputs. We handle the diverse number of inputs by tokenizing each input modality from a person by pretraining a vector-quantized autoencoder for each signal. Then, to learn the temporal relationship of these social signals across multiple interlocutors, we introduce a person-aware and modality-aware blockwise attention masking approach to attend over these tokens across modality, time, and interlocutor.

A. Learning Modality-Specific Quantized Codebooks

Because we are operating on c-second time segments, it is difficult to learn features for continuous features. We tokenize continuous inputs with a Vector Quantized Variational Autoencoder (VQ-VAE) [32]. Signals like gaze, headpose, or pose can be represented as keypoints, and we use a 1D CNN-based encoder and decoder to learn quantized keypoint representations. We first train these modality-specific VQ-VAEs on the HHCD training data, then freeze them during the training of the M3PT transformer.

To train the VQ-VAE, we take a c-second segment and encode each individual keyframe with a 1D-CNN. If the segment contains m frames, we then have m embeddings to represent the segment. Each keyframe's embedding is then quantized by selecting from a codebook, which provides a discrete representation of the social signal. Given these m quantized embeddings, we aggregate them into a single embedding z with a linear projection step. During the decoding step, we up-project z back into m temporal embeddings, from which we use the codebook to re-select m quantized embeddings. We train the VQ-VAE with standard selection and commitment losses [32]. Since we leverage the quantization step in both the encoding step and the decoding step to learn z, we average the two selection losses and add a reconstruction loss. Thus, the total training objective becomes: $\mathcal{L} = \log p(x|z_q(x)) + 0.5(\mathcal{L}^1_{\text{select}} + \mathcal{L}^2_{\text{select}})$.

B. Transformer Architecture

In a multi-party setting, each participant produces various social signals at each timestep. We use a VQ-VAE to compute discretized tokens for each c-second segment in time. The encoded representations of all social signals from all individuals are proccessed by a causal (left-to-right) transformer designed to capture relationships between individuals and modalities over time. For a modality k for person i over time $z_0^{ik}, \cdots, z_T^{ik}$, we add a cyclic positional encoding over time, an embedding representing the person, and a modality-specific embedding.

We sequence the modalities over time per-person as shown in Fig. 1, where each timestep block contains blocks for the modalities of each individual. However, this blockwise structure is not applicable for the traditional lower triangular attention mask typically used in causal transformers. In particular, a lower triangular causal attention allows for a model to leverage past information to predict a given token;

TABLE I: **Bite Timing Prediction** performance across ablated input modalities. Using all available signals leads to the best performance, and that dropping some signals causes substantial performance loss across all metrics, confirming the value of multimodality in predicting bite timing.

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

however, we want to predict the features of a person i based on features from previous timesteps *and* features from other participants in the current timestep. Thus, we introduce multi-party, multimodal blockwise causal masking.

Blockwise Attention Masking. To enforce temporal structure and modality-specific interactions, we design a blockwise causal masking strategy. This mask consists of a lower triangular matrix that restricts attention to only past and present time steps, preserving the left-to-right structure of the model. Additionally, we modify the mask by creating small blocks along the diagonal, where signals from each individual can attend to both their own previous signals and those of others in the group as shown in Fig. 2. This design ensures that the model can attend not only each person's behavior but also to the potential influence of others' social signals on that behavior.

Right-shifted Residual Connection. In bidirectional encoder architectures like BERT [7], mask tokens prevents residual connections from leaking information to masked positions. Similarly, in unidirectional decoder architectures like GPT [5], predicting the next token inherently avoids

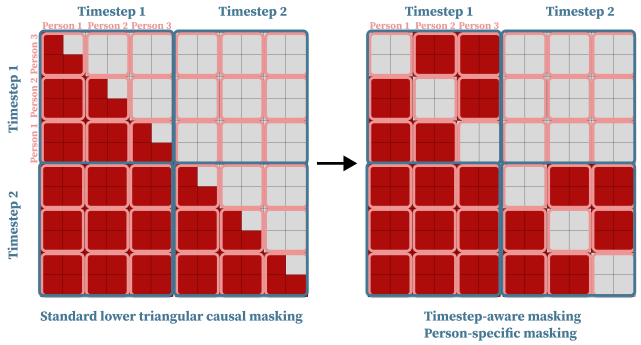

Fig. 2: **Blockwise Attention Masking**: Conventional lower-triangular masking (left) would not be aware of the timestep and person-specific feature chunks present in multi-party, multimodal settings. Our blockwise attention mask (right) masks each person's social signals when they must be predicted, allowing the model to focus solely on capturing interactions from others' social signals.

TABLE II: **Speaking Status Prediction**. We find that the best results for speaking status prediction are achieved when all features are included. We find that word and bite features are most predictive of predicting speaking status.

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

data leakage through residual connections. However, in our case, our blockwise attention mechanism prevents the direct application of conventional residual connection strategies. We found that simply removing residual connections leads to poor model training due to small gradients.

To address this difficulty, we use a right-shifted residual connection. Instead of directly adding residual features to the hidden states, we right-shift all features by one segment before adding them. This shift ensures that each position receives residual information only from the preceding segment, which prevents leakage of the signal to be predicted during inference for evaluation.

IV. Evaluation on HHCD Dataset

To evaluate M3PT, we use the Human-Human Commensality Dataset (HHCD), which contains triadic interactions among three participants without mobility limitations during shared meals. The purpose of the dataset is to predict bite timing events in social settings, which can be used to build socially-aware robot-assisted feeding systems for individuals who do have mobility limitations. The dataset contains multimodal social data like gaze, body pose, speech, as well as annotations for bite events, drink interactions, and utensil usage. In this work, we repurpose this dataset towards our task of multi-party, multimodal social signal prediction. We process gaze, headpose, body pose, transcripted speech, speaking status, and bite timing as inputs to M3PT, and we focus on improving the prediction of the two binary social signals of an individual given their past social signals and features of the other two participants: speaking status and bite timing.

HHCD contains 30 unique sessions of triadic social dining. For each session, we sample 36-second sequences with an 18-second rolling window. We then split this 36-second sequence into 12 three-second segments.

M3PT performs binary classification tasks for both speaking status and biting time over each segment. Since each segment is 3-second long, we classify a segment as "speaking" if more than 30% of the frames indicate speaking for a specific user. This is to account for noise in the speaker estimations. A frame is labeled as "biting" if at least one frame within the segment has a bite.

TABLE III: **Does M3PT learn from longer temporal context?**. We maintain the segment size of each token to be 3 seconds, and have n segments $\times$ 3 seconds-per-segment. Functionally, this ablation studies how M3PT makes predictions with longer contexts. We find that as the total context time considered increases, the F1 score begins to fall.

	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

TABLE IV: **Impact of Segment Length on Performance.** We hold the total time to be at a constant 36 seconds, but modify the length of a segment to k seconds and thus the number of segments n to have $n \times k$s. We find that longer segment lengths leads to a collapse in prediction performance in speaking status (S) and bite timing (B) to majority class.

	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

Training objective. Our training objective is to minimize cross entropy losses for bite timing and speaking status prediction. In HHCD, speaking status and biting time are highly imbalanced, as most of the time individuals are neither speaking nor biting. To address this imbalance, we apply inverse class frequency weighting to the loss functions.

Evaluation metrics. HHCD contains 30 triadic sessions; we treat each session as a fold and train on 29 of them and test on 1. We repeat this across 3-folds, training and testing on different sets of sessions. We evaluate our models on classification accuracy, F1 score, precision, recall, and a normalized Matthews correlation coefficient (nMCC). Unlike F1 score, nMCC summarizes the full confusion matrix of true/false positives and negatives, which past work [29] found informative for bite timing.

V. Results

We present the test set performance of M3PT. We report the average performance and standard deviation of M3PT and various ablations over three folds of the HHCD data.

Multimodality improves social signal prediction performance. As shown in Tables I and II, using all the modalities improves performance, both for speaking status and bite timing. For bite timing prediction, we find large distinctions across features. We find that all metrics, especially F1 and nMCC are reduced when gaze, words, or speaking status are removed. This social signals are empirically the most informative for predicting bite timing.

We find similar results indicating that multiple modalities improves speaking status prediction. We find that using no other signals except for speaking status, the M3PT produces nearly random predictions, but using all modalities improves the performance the most. The largest performance degradation occurs when bite timing is removed, indicating that it is an important feature for speaking status prediction. Other features have slight performance losses, however, they are not as strong of a drop-off compared to bite timing.

Temporal context on bite timing. Table III presents an ablation that explores how varying the total time length, while maintaining a constant segment size, affects model performance for speaking status (S) and bite timing (B) tasks. The experiment is conducted using a smaller model to avoid overparameterization. The results show that reducing the temporal context from 36 seconds to lower values actually increases performance, possibly due to task difficulty increasing as data gets noisier. However, we believe that this 36 second length is practical for situations like predicting bite timing, where a system has to run continuously.

Large segment lengths lead to mode collapse. Table IV shows that, as the segment length increases, both speaking status and bite timing prediction accuracy drastically falls across all metrics. This finding is consistent with mode collapse in the tokenization step: the VQ-VAE step is representing more of the temporal information as opposed to the transformer. Our use of 12 segments with a length of 3 seconds strikes a balance between having the transformer represent the temporal and multi-party information and having the VQ encoders represent only the modality information.

VI. Conclusion and Limitations

In this work, we presented M3PT, a causal transformer able to predict an individual's bite timing and speaking status in multi-party dining based on temporal- and person-aware social signals. We investigated our design decision for 3-second temporal segments, and found that the larger segments cause the transformer to learn less of the temporal structure. We also presented results on the impact of increased temporal context for predicting btie timing and speaking status. M3PT is a first step in building models for multi-party, multimodal social signal processing.

If in robot setting, we need to predict continuous social signals in addition to discrete signals. In preliminary experiments, we found that predicting continuous signals like body pose directly was not feasible. The reconstructions were often poor as transformer models are often better at predicting discretized inputs. When trained to predict discrete pose tokens constructed by the VQ-VAE, the M3PT performance was still lacking, though it is unclear whether the behavior of one's interlocutorrs is a sufficient signal to predict what body pose one will take next.

VII. Ethical Impact Statement.

M3PT was designed to utilize social signals in multi-party settings. Although this work had used pre-collected data from others' work, the implication of a model such as M3PT has potential uses in tracking users and potentially manipulating social scenarios. Additionally, if such a method can process social signals from others who do not know they are expressing such information, consent of the utilization of such data becomes important. Otherwise, such a social signal tracking system could be used for unknowingly manipulating non-consenting participants. Also, the original application of bite timing is to use this method for helping people with mobility limitations. Though this work does not explicitly consider this setting, we must be careful in introducing bias in training such models, as they would be used in sensitive situations with people with mobility limitations. For example, body pose prediction may introduce bias for systems that run alongside people with mobility limitations, and these kinds of considerations need to be made carefully.

To mitigate such risks, the utilization of M3PT can require explicit consenting procedures or only track users who have previously consented to be observed or can rely only on anonymized data such as body pose that were processed through anonymized pipelines. However, we believe that the benefits of such a system are still useful. For bite timing prediction, past work [3] has shown that people with mobility limitations who use robot-assisted feeding systems prefer such autonomous systems as they are able to eat food with their friends and families.

References

[1] C. Ahuja, D. W. Lee, R. Ishii, and L.-P. Morency. No gestures left behind: Learning relationships between spoken language and freeform gestures. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020.

[2] X. Alameda-Pineda, J. Staiano, R. Subramanian, L. Batrinca, E. Ricci, B. Lepri, O. Lanz, and N. Sebe. Salsa: A novel dataset for multimodal group behavior analysis. *IEEE transactions on pattern analysis and machine intelligence*, 2015.

[3] T. Bhattacharjee, E. K. Gordon, R. Scalise, M. E. Cabrera, A. Caspi, M. Cakmak, and S. S. Srinivasa. Is more autonomy always better? exploring preferences of users with mobility impairments in robot-assisted feeding. In *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction*, 2020.

[4] C. Birmingham, K. Stefanov, and M. J. Mataric. Group-level focus of visual attention for improved next speaker prediction. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2021.

[5] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners, 2020.

[6] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields, 2019.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[8] A. Dieberger, P. Dourish, K. Höök, P. Resnick, and A. Wexelblat. Social navigation: Techniques for building more usable systems. *interactions*, 7(6):36–45, 2000.

[9] A. Fathi, J. K. Hodgins, and J. M. Rehg. Social interactions: A first-person perspective. In *Conference on Computer Vision and Pattern Recognition*, 2012.

[10] T. Fischer, H. J. Chang, and Y. Demiris. Rt-gene: Real-time eye gaze estimation in natural environments. In *Proceedings of the European conference on computer vision (ECCV)*, 2018.

[11] G. Gordon, S. Spaulding, J. K. Westlund, J. J. Lee, L. Plummer, M. Martinez, M. Das, and C. Breazeal. Affective personalization of a social robot tutor for children's second language skills. In *Proceedings of the AAAI conference on artificial intelligence*, 2016.

[12] B. Holman, A. Anwar, A. Singh, M. Tec, J. Hart, and P. Stone. Watch where you're going! gaze and head orientation as predictors for social robot navigation. In *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021.

[13] D.-A. Huang and K. M. Kitani. Action-reaction: Forecasting the dynamics of human interaction. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13*. Springer, 2014.

[14] R. K. Jenamani, D. Stabile, Z. Liu, A. Anwar, K. Dimitropoulou, and T. Bhattacharjee. Feel the bite: Robot-assisted inside-mouth bite transfer using robust mouth perception and physical interaction-aware control. In *International Conference on Human-Robot Interaction (HRI)*, 2024.

[15] H. Joo, T. Simon, M. Cikara, and Y. Sheikh. Towards social artificial intelligence: Nonverbal social signal prediction in a triadic interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[16] A. Kendon. Movement coordination in social interaction: Some examples described. *Acta psychologica*, 1970.

[17] M. LaFrance. Nonverbal synchrony and rapport: Analysis by the cross-lag panel technique. *Social Psychology Quarterly*, 1979.

[18] M.-C. Lee and Z. Deng. Online multimodal end-of-turn prediction for three-party conversations. In *Proceedings of the 26th International Conference on Multimodal Interaction*, 2024.

[19] L. Mathur, R. Adolphs, and M. J. Matarić. Towards intercultural affect recognition: Audio-visual affect recognition in the wild across six cultures. In *International Conference on Automatic Face and Gesture Recognition (FG)*, 2023.

[20] P. Müller, M. X. Huang, X. Zhang, and A. Bulling. Robust eye contact detection in natural multi-person interactions using gaze and speaking behaviour. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, ETRA '18, New York, NY, USA, 2018. Association for Computing Machinery.

[21] P. Müller, E. Sood, and A. Bulling. Anticipating averted gaze in dyadic interactions. In *ACM Symposium on Eye Tracking Research and Applications*, ETRA '20 Full Papers, New York, NY, USA, 2020. Association for Computing Machinery.

[22] M. Murray, N. Walker, A. Nanavati, P. Alves-Oliveira, N. Filippov, A. Sauppe, B. Mutlu, and M. Cakmak. Learning backchanneling behaviors for a social robot via data augmentation from human-human conversations. In *Conference on Robot Learning (CoRL)*, 2022.

[23] P. Müller and A. Bulling. Emergent leadership detection across datasets, 2019.

[24] P. Müller, M. X. Huang, and A. Bulling. Detecting low rapport during natural interactions in small groups from non-verbal behaviour. In *23rd International Conference on Intelligent User Interfaces*, IUI'18, page 153–164. ACM, Mar. 2018.

[25] E. Ng, S. Ginosar, T. Darrell, and H. Joo. Body2hands: Learning to infer 3d hands from conversational gesture body dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[26] E. Ng, H. Joo, L. Hu, H. Li, T. Darrell, A. Kanazawa, and S. Ginosar. Learning to listen: Modeling non-deterministic dyadic facial motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[27] E. Ng, S. Subramanian, D. Klein, A. Kanazawa, T. Darrell, and S. Ginosar. Can language models learn to listen? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[28] E. Ng, D. Xiang, H. Joo, and K. Grauman. You2me: Inferring body pose in egocentric video via first and second person interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[29] J. Ondras, A. Anwar, T. Wu, F. Bu, M. Jung, J. J. Ortiz, and T. Bhattacharjee. Human-robot commensality: Bite timing prediction for robot-assisted feeding in groups. In *Conference on Robot Learning (CoRL)*, 2022.

[30] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision, 2022.

[31] O. Rudovic, J. Lee, M. Dai, B. Schuller, and R. W. Picard. Personalized machine learning for robot perception of affect and engagement

in autism therapy. *Science Robotics*, 2018.
[32] A. van den Oord, O. Vinyals, and K. Kavukcuoglu. Neural discrete representation learning, 2018.
[33] M. Vázquez, E. J. Carter, B. McDorman, J. Forlizzi, A. Steinfeld, and S. E. Hudson. Towards robot autonomy in group conversations: Understanding the effects of body orientation and gaze. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 2017.
[34] Y. Wang, W. Song, W. Tao, A. Liotta, D. Yang, X. Li, S. Gao, Y. Sun, W. Ge, W. Zhang, et al. A systematic review on affective computing: Emotion models, databases, and recent advances. *Information Fusion*, 2022.
[35] M. Zhou, Y. Bai, W. Zhang, T. Yao, T. Zhao, and T. Mei. Responsive listening head generation: a benchmark dataset and baseline. In *European Conference on Computer Vision*. Springer, 2022.