Enhancing Human-Robot Interaction through Multi-Human Motion Forecasting

Mohammad Samin Yasar and Tariq Iqbal

***Abstract*— Social navigation is a critical element in facilitating robots' transition to human spaces. The task of accurately predicting human motion is non-trivial and is compounded by the variability of human motion and the presence of multiple humans in proximity. To address some of the open challenges in pose prediction to facilitate social navigation, in this work, we present a novel sequence learning algorithm that models past human motion using a flexible discrete latent space. Our algorithm introduces the concept of Vector Quantization for human motion, enabling the learning of a discrete latent space without being restricted by any static prior. In addition, we propose a new objective function that uses the discriminator objective to penalize deviation of predicted motion from the ground-truth. Finally, to explicitly account for interactions among multiple humans, we incorporate a lightweight attention mechanism that conditions per-agent predictions on the prior hidden states of all agents. Our evaluation in multi-agent scenarios suggest the efficacy of our approach over state-of-the-art approaches, resulting in more feasible human poses that align better with the ground-truth.**

I. Introduction

Robots that can operate alongside humans in settings designed for human-centric activities represent a departure from traditional industrial robots, which typically functioned in isolation within enclosed spaces [1]–[3]. Human interactions, such as navigating crowded areas, handing over or exchanging objects, are heavily reliant on the ability to observe and anticipate the actions of others [4]–[8]. In line with this, for robots to function safely and effectively in the presence of humans, they must continually monitor, predict, and adapt to changes in their environment, especially concerning the movements and intentions of nearby humans [9]–[13]. Despite substantial progress in robot perception, enabling them to detect changes and adjust to new environmental conditions [14]–[17], the ability to reliably predict alterations in environmental dynamics remains an ongoing and significant challenge.

The concept of anticipation has received extensive attention within the field of robotics, particularly in the realm of social navigation. The primary goal in this context is to navigate safely in the presence of humans, thus avoiding any potential interference [18]–[22]. Additionally, previous research has delved into anticipatory planning of robot actions based on inferred goals [23]–[28]. However, as robots are expected to interact with humans over prolonged periods, there is a need to anticipate human motion at a higher spatial and temporal granularity [29]–[33]. This involves predicting future human motion conditioned on past motion (Fig. 1), enabling the robot to plan around the human without disrupting their natural flow. However, achieving accurate predictions of human motion remains challenging due to the intricate and socially influenced nature of human behavior [34], [35].

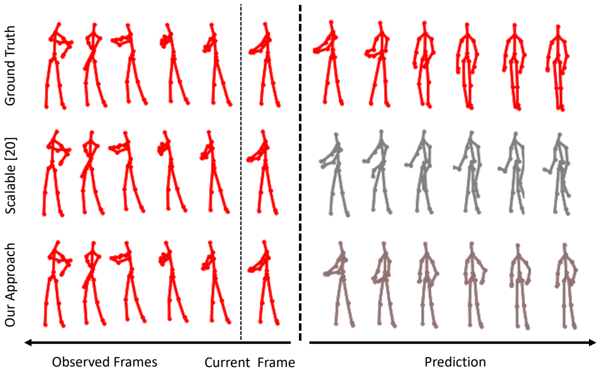

Fig. 1: Qualitative evaluation of the predicted motion of our approaach and the next best performing model, Scalable + Interpretable [34]. Our approach uses a flexible discrete latent space and introduces a discriminator objective, which results in predictions which are closer to the ground-truth.

Previous research has approached the task of predicting future motion as a sequence learning problem to account for the irregular and stochastic nature of human movement [20], [36]–[40]. These methods aim to derive a unified representation from training data that can generalize to test data. In training these networks, the central assumption typically revolves around either learning a distribution capable of fitting a fixed prior [34], [41]–[43] or learning a point estimate based on past observed data, which is then used to predict future human motion [44]–[46]. Learning a static prior introduces an auxiliary objective that acts as a regularizer, which requires careful tuning. On the other hand, learning a point estimate tends to yield less robust representations. Furthermore, majority of prior works have relied only on the reconstruction error for training these networks. Such an objective function may cause the predictions to regress to the mean and, as such, may not be able to capture the spatial and temporal correlations in human motion.

To address the above challenges, we present a novel approach that aims to close two critical gaps in human pose forecasting: 1) learning a robust representation of the past motion and 2) improving temporal and spatial correlation in the prediction. Our proposed framework extends the encoder-decoder model by incorporating codebook learning and distribution matching. Leveraging codebooks enables

[*]The authors are with the School of Engineering and Applied Science, University of Virginia, USA. {msy9an,tiqbal}@virginia.edu.

our approach to acquire discrete representations of observed motion data. Additionally, given that pose prediction involves complex data dependencies, conventional mean squared error (MSE)-based objectives become problematic. Therefore, we introduce a novel loss function based on a discriminator to promote temporal and spatial consistency by penalizing predictions that diverge from the ground-truth distribution.

We conducted a comprehensive series of experiments to assess the effectiveness of our approach in various multi-agent pose prediction scenarios using the NTU RGB+D 60 dataset [47]. The results underscore effectiveness of our approach in addressing the open challenges in pose prediction, as it consistently outperformed all evaluated algorithms achieving the lowest prediction error at various temporal horizons, both in quantitative and qualitative terms. For an in-depth discussion of our methodology and extensive experimental analyses, we encourage the reader to refer to the complete version of this paper [48].

II. Related Works

Human trajectory prediction poses a significant challenge, mainly because the policies of the agents involved are not directly observable. Various data-driven approaches have been applied to address the prediction of intricate interactions in social navigation [36], [37], [49], autonomous vehicle operations [38], [39], [50], and human-robot interaction (HRI) scenarios [2], [51]. Alahi et al. [36] introduced Social-LSTM, which employs agent-specific Long Short-Term Memory (LSTM) networks to summarize the past observations of each agent. The hidden states of neighboring LSTMs are interconnected using a social pooling strategy, and this collective information serves as input to the LSTM cell at the subsequent time step. Gupta et al. [37] presented Social GAN, which introduced an efficient pooling mechanism comprising a Multi-Layer Perceptron (MLP) followed by max pooling to address the challenge of human trajectory prediction.

While previous research in trajectory forecasting and pose prediction has made significant strides in advancing the state-of-the-art, the generation of human motion that is both feasible and temporally coherent remains a challenging and open research problem [34], [44], [45], [52]. Furthermore, unlike some other domains of machine intelligence, such as computer vision or machine translation, there is no widely accepted consensus on the optimal framework for capturing the spatial and temporal dynamics of human motion. Although recent approaches have adopted an encoder-decoder framework, the task of learning a robust representation within the encoder that effectively encapsulates past human motion is an ongoing research endeavor. Additionally, while there are advantages to learning a distribution over past observed motion, the identification of a prior that can accurately model this distribution remains a challenging task. This challenge leads to difficulties in optimization, which necessitates the reconciliation of both reconstruction and distribution matching aspects.

III. Problem Formulation

Our objective is to improve the robot's perception by providing it with the capability to forecast the motion of all human in the scene. Human pose prediction is formally described as the task of predicting the future human motion for a certain period, given their past motion.

We assume that the number of agents in the scene, denoted by K, is known beforehand. The input to the model consists of the observed motion of all agents in the scene from time $t = 1$ to τ: $\mathbf{X} = \{X^1, \dots, X^K\} = \{x^{1:K}_1, x^{1:K}_2, \dots, x^{1:K}_\tau\}$. The model aims to predict the future trajectory frames over horizon H: $\mathbf{Y} = \{Y^1, \dots, Y^K\} = \{y^{1:K}_{\tau+1}, y^{1:K}_{\tau+2}, \dots, y^{1:K}_{\tau+H}\}$.

We assume that the future human motion of each agent is conditioned on the observed motion of all agents, and predict each frame in an auto-regressive manner. Thus, the multi-agent pose prediction problem is formulated as follows:

$$p_\theta(\hat{\mathbf{Y}}^a) = \prod_{\delta=\tau+1}^{\tau+H} p_\theta(\hat{y}^a_\delta | \hat{y}^a_{\tau:\delta-1}, x^{1:K}_{1:\tau}); \quad \forall a = 1, \dots, K \tag{1}$$

IV. Proposed Method

Here, we present our proposed framework, which has been introduced in detail in [48]. Our approach comprises two primary components: the encoder-decoder architecture with discrete latent representation and the discriminator network (see Fig. 2). The objective of the encoder-decoder architecture is to predict future human motion. The discriminator is tasked to distinguish frames that are coming from the ground-truth distribution, from frames that are not, penalizing the latter. This creates a min-max game between the two networks, which combine to provide more accurate pose prediction. Our framework represents a unified framework for predicting the motion in single, multi-agent, and human-robot collaboration scenarios, with the main difference between the number of encoders and decoders, which scale with the number of agents that require modeling.

Encoder: The encoder seeks to learn a salient representation over the raw input space. The input to the encoder is the past observed motion, represented in skeletal joint position, velocity and acceleration. We use separate encoders to process the position, velocity, and acceleration streams. The goal of each encoder is to learn a spatio-temporal representation of the observed motion data. The operations in the encoder can be formulated as follows:

$$h_{s,t} = Encoder(h_{s,t-1}, x_{s,t}, \phi_s) \tag{2}$$

here s represents position, velocity, or acceleration. Here, $x_{s,t}$ represents the input to the Encoder at time t and will take the value of $x_{pos,t}, x_{vel,t}, x_{acc,t}$ for position, velocity, and acceleration at time t, respectively. $h_{s,t-1}$ represents the past hidden output at time $t-1$ and ϕ_s represents the stream-specific encoder weights. The output from each encoder is passed to a self-attention module [53]. The attention module is tasked to sparsely and adaptively extract the salient features from the three streams.

$$h_t = Concat(h_{pos,t}, h_{vel,t}, h_{acc,t}); \; z_t = Att(h_t, \phi_{att}) \tag{3}$$

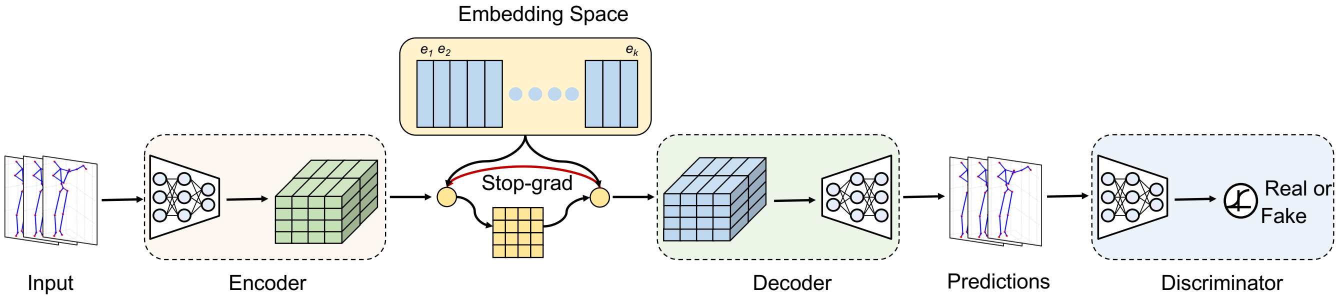

Fig. 2: Vector-Quantized Generative Adversarial Network for Motion Prediction [48]. The primaray objective of the encoder-decoder framework is to generate future human motion, using a discrete latent codebook. This codebook offers a dynamic prior, which adapts through learning. The discriminator is designed to differentiate between the ground-truth distribution and the predictions generated by the decoder.

where ϕ_{att} represents weights of the attention module. In the self-attention module, we first linearly project the concatenated output h_t into a separate query (Q), key (K), and value (V) embeddings for each head. These embeddings are then used to calculate attention weights using the scaled-dot product softmax (sf) approach. The functions for each head in the multi-head self-attention module are defined as follows:

$$\begin{gathered} Q = h_t W^Q;\ K = h_t W^K; V = h_t W^V \\ Att(Q,K,V) = sf\left(\frac{QK^T}{\sqrt{d_k}}\right)V \end{gathered} \tag{4}$$

where, W^Q, W^K, W^V represent the linear projection weights and $\frac{1}{\sqrt{d_k}}$ is the scaling factor for calculating the attention weights. The output of the attention module is passed to the discrete codebook to obtain the latent space.

Latent codebook: We propose the use of a codebook for calculating the latent space, similar to the Vector Quantization approach in VQ-VAE [54], which has been successfully applied for image synthesis. Compared to previous methods in pose prediction [34], [41], [43], [55], which have relied on variational bottlenecks or discriminators to impose a fixed prior, our proposed method involves the adoption of a flexible prior that evolves dynamically throughout the training process. This approach offers the advantage of not constraining the learning process by regularizing the latent space to adhere to a static distribution, thereby reducing the risk of mode collapse.

We introduce the latent embedding as a codebook, denoted as $e \in \mathbb{R}^{L\times M}$, where L represents the size of the categorical latent space, and M signifies the dimension of each individual categorical embedding vector e_i. To compute the discrete latent space, we employ the encoder's output and determine the nearest neighbors in the shared embedding space e. Consequently, the latent space z can be conceptualized as a posterior categorical distribution, denoted as $q(z|x)$, where the probabilities associated with the categorical vector are one-hot and defined as follows:

$$q(z=k|x) = \begin{cases} 1 & \text{for k} = \text{argmin}_j ||z - e_j||_2, \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

where, $z(x)$ is the output of the encoder network, e_j represents a vector from the codebook e.

Decoder: To predict the current motion, we employ an auto-regressive decoder that relies on the information from previous time steps. This decoder exclusively generates the positions of the skeleton joints. It is fed with two inputs: a discrete embedding vector e_k that encapsulates past motion observations and the most recent predicted frame. The latter is processed by a Keyless attention module [56]. This attention module computes the weights that reflect the relationship between the immediate past output and the latent representation of the observed motion. The operations within the decoder are formulated as follows:

$$\begin{aligned} p_t = Concat(z_t, h_{dec,t-1});\ p_{att,t} = Att(p_t, \phi_{att}) \\ S_t = Decoder(S_{t-1}, p_{att,t}, \phi_{pos}); \end{aligned} \tag{6}$$

where, the latent representation is denoted as z_t, while $h_{dec,t-1}$ represents the previous hidden output of the Decoder. The output of the attention mechanism in the decoder is denoted by $p_{att,t}$, which is passed along with the previous Decoder output S_{t-1} to the Decoder. ϕ_{att} and ϕ_{pos} denote the weights of the attention module and the decoder network, respectively.

Discriminator: The discriminator consists of a separate encoder and is tasked with distinguishing between samples originating from the ground-truth distribution and those generated by the decoder. To achieve this, the discriminator receives two inputs: the ground-truth data denoted as $T_{real} = Y$, and the predicted motion $T_{fake} = \hat{Y}$, classifying them as either real or fake. The inputs are passed through an encoder, similar to Eq .2. The encoder's output is passed to the linear layer to obtain the classification results.

Overall objective function: In our approach, there are two distinct modules that are trained in opposition to each other, following the min-max setup in GANs [57]: the encoder-decoder architecture and the discriminator. As such, the overall training procedure can be summarized by the following objective function:

$$\begin{aligned} \min_G \max_D V(G,D) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \\ \mathbb{E}_{z\sim p(z)}[\log(1 - D(G(z)))] \end{aligned} \tag{7}$$

where G represents the encoder-decoder architecture and D denotes the discriminator network. Furthermore, as we use a discrete codebook for representing the latent space, there are additional terms in the objective for the encoder-decoder architecture, which reflects the vector quantization algorithm that is used to learn the discrete representation, along with additional commitment loss to ensure that the encoder commits to an embedding, instead of arbitrarily growing in embedding space. The objective function can be defined as follows:

$$\mathcal{L} = \log p(y|z(x)) + ||z(x) - e||_2^2 + \beta||z_e(x) - sg[e]||_2^2 \tag{8}$$

where, the first term is the reconstruction loss. sg represents a stop-gradient operator that prevents the flow of gradient through the codebook. The second term is the nearest-neighbor embedding loss for selecting the embedding vector from the codebook. The third term is the commitment cost, ensuring the encoder commits to a specific embedding.

V. Experimental Setup

A. Datasets

Here, we present the experimental evaluation of our approach on the NTU RGB+D 60 dataset [47]. We focused on 11 joint actions that involve more than one agent. We followed the cross-subject evaluation scheme and used 20 subjects for training and validation and another 20 for testing. We used the skeleton modality and all the provided joints of each agent for pose prediction, as prior work has shown that using RGB data only provides marginal improvements due to the constrained environmental setup in which the data were collected [46].

B. State-of-the-art methods and baselines

To evaluate the performance of our approach, we compared against several state-of-the-art models: Joint Learning [46], Joint Learning + Social [46], Joint Learning + Social + Context [46], and Scalable + Interpretable [34]. The Joint Learning architecture is based on a sequence-to-sequence architecture and operate under the assumption that agents do not interact, thus predicting the motion of each agent independently. However, the Joint Learning + Social method introduces a permutation-invariant pooling mechanism for aggregating social features across all agents, while the Joint Learning + Social + Context method incorporates an additional spatio-temporal context Convolutional Neural Network (CNN) module to extract RGB features from the scene. The Scalable + Interpretable method presents an encoder-decoder framework with adversarial regularization on the latent space, featuring an attention module for disentangling and extracting multi-agent features. To ensure a fair comparison, we fine-tuned hyperparameters for all the models.

C. Evaluation Metric

We evaluated the performance of all models using the Mean Squared Error (MSE), which is the l_2 distance between the ground-truth and predicted poses at each timestep, averaged over the number of joints and sequence length, similar to prior work [41], [45], [46], [55]. The MSE is calculated as:

$$\mathcal{L}(\mathcal{X}, \hat{\mathcal{X}}) = \frac{1}{H \times K} \sum_{t=1}^{H} \sum_{i=1}^{D} (x_{t,i} - \hat{x}_{t,i})^2 \tag{9}$$

where, H and D are the total number of frame and joints respectively. The MSE jointly encodes global body motion and skeletal movements [46], making it an ideal metric.

TABLE I: MSE (in cm^2) comparison of different multi-agent methods on the NTU-RGBD 60 Dataset (Lower is better).

Approaches	Frames					
	2	4	8	10	13	15
Joint Learning [46]	9.68	15.84	29.88	37.52	49.55	57.93
Joint Learning + Social [46]	9.71	15.97	30.36	38.69	51.68	59.38
Joint Learning + Social + Context [46]	9.78	16.02	30.46	38.39	50.91	59.63
Scalable + Interpretable [34]	9.66	15.66	29.05	36.16	47.20	54.84
Our Approach	**9.65**	**15.48**	**28.57**	**35.64**	**46.71**	**54.39**

VI. Results and Discussion

Results: We report the results of all evaluated models for multi-agent pose prediction scenarios on the NTU-RGB+D 60 [47] datasets. We report the results in distinct frame intervals instead of seconds, similar to [34] to circumvent the problem of frame drops during data collection and subsequent evaluation. We use these frame intervals to evaluate the models' performance across short-term horizons (2 & 4 frames), mid-term horizons (8 & 10 frames), and long-term horizons (13 & 15 frames). The results in Tab. I suggest that our approach outperformed all other methods across all the evaluated horizons.

Discussion: Our approach consistently outperformed all evaluated models across all intervals, demonstrating superior representation learning and sequence modeling for multi-agent scenarios. One key factor in its success is its ability to model interaction dynamics within the decoder by calculating attention weights across all agents' hidden states. Unlike previous methods, our approach explicitly conditions predictions on both the past hidden states and latent states of all agents. Additionally, the use of a discriminator loss encourages more accurate trajectory generation. Our framework's Keyless Attention mechanism in the decoder effectively models interactions without adding computational complexity compared to traditional self-attention mechanisms.

For robots to co-exist with humans, they need to anticipate the pose and trajectory of their human counterparts. In this work, we introduce a novel mechanism that focused on predicting pose, which investigated human motion at a greater spatial granularity compared to trajectory forecasting. To tackle the issue of learning a robust representation of past observed poses, we proposed the use of vector quantization to learn a discrete latent space, with no restrictions of a static prior. Additionally, we proposed using the discriminator loss to compliment the MSE objective to improve the accuracy of pose prediction, As both pose prediction and trajectory forecasting have similar sequence learning challenges, our findings can generalize to trajectory forecasting and social navigation. Moreover, in close-proximity collaboration where anticipating human pose is crucial, our approach can be incorporated in robot's perception, to allow accurate planning.

REFERENCES

[1] M. Knudsen and J. Kaivo-Oja, "Collaborative robots: Frontiers of current literature," *Journal of Intelligent Systems: Theory and Applications*, vol. 3, no. 2, pp. 13–20, 2020.

[2] V. V. Unhelkar, P. A. Lasota, Q. Tyroller, R.-D. Buhai, L. Marceau, B. Deml, and J. A. Shah, "Human-aware robotic assistant for collaborative assembly: Integrating human motion prediction with planning in time," *IEEE RA-L*, 2018.

[3] T. Iqbal and L. D. Riek, "A Method for Automatic Detection of Psychomotor Entrainment," *IEEE T-AC*, 2016.

[4] Z. Wang, K. Mülling, M. P. Deisenroth, H. Ben Amor, D. Vogt, B. Schölkopf, and J. Peters, "Probabilistic movement modeling for intention inference in human–robot interaction," *IJRR*, 2013.

[5] A. M. Williams, P. Ward, J. M. Knowles, and N. J. Smeeton, "Anticipation skill in a real-world task: measurement, training, and transfer in tennis." *Journal of Experimental Psychology: Applied*, 2002.

[6] 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, pp. 42–52.

[7] M. Fiore, A. Clodic, and R. Alami, "On planning and task achievement modalities for human-robot collaboration," in *Experimental Robotics: The 14th International Symposium on Experimental Robotics*. Springer, 2016, pp. 293–306.

[8] A. Clodic, E. Pacherie, R. Alami, and R. Chatila, "Key elements for human-robot joint action," *Sociality and normativity for robots: philosophical inquiries into human-robot interactions*, pp. 159–177, 2017.

[9] M. M. Islam and T. Iqbal, "MuMu: Cooperative multitask learning-based guided multimodal fusion," in *AAAI*, 2022.

[10] C. Finn, A. Rajeswaran, S. Kakade, and S. Levine, "Online meta-learning," in *ICML*, 2019.

[11] M. S. Yasar and T. Iqbal, "Robots that can anticipate and learn in human-robot teams," in *ACM/IEEE HRI*, 2022.

[12] A. Rudenko, L. Palmieri, M. Herman, K. M. Kitani, D. M. Gavrila, and K. O. Arras, "Human motion trajectory prediction: A survey," *IJRR*, 2020.

[13] M. S. Yasar and T. Iqbal, "Coral: Continual representation learning for overcoming catastrophic forgetting," *AAMAS*, 2023.

[14] T. Iqbal and L. D. Riek, "Human robot teaming: Approaches from joint action and dynamical systems," *Humanoid Robotics: A Reference, Springer*, 2017.

[15] M. M. Islam and T. Iqbal, "Multi-gat: A graphical attention-based hierarchical multimodal representation learning approach for human activity recognition," in *IEEE RA-L*, 2021.

[16] M. M. Islam, M. S. Yasar, and T. Iqbal, "Maven: A memory augmented recurrent approach for multimodal fusion," *IEEE Trans. Multimedia*, 2022.

[17] M. M. Islam and T. Iqbal, "Hamlet: A hierarchical multimodal attention-based human activity recognition algorithm," in *IROS*, 2020.

[18] J. Mainprice and D. Berenson, "Human-robot collaborative manipulation planning using early prediction of human motion," in *IROS*. IEEE, 2013.

[19] L. Sanneman, C. Fourie, J. A. Shah *et al.*, "The state of industrial robotics: Emerging technologies, challenges, and key research directions," *Foundations and Trends® in Robotics*, 2021.

[20] C. I. Mavrogiannis and R. A. Knepper, "Decentralized multi-agent navigation planning with braids," in *Algorithmic foundations of robotics XII*. Springer, 2020, pp. 880–895.

[21] C. Mavrogiannis, K. Balasubramanian, S. Poddar, A. Gandra, and S. S. Srinivasa, "Winding through: Crowd navigation via topological invariance," *IEEE Robotics and Automation Letters*, vol. 8, no. 1, pp. 121–128, 2022.

[22] C. I. Mavrogiannis, V. Blukis, and R. A. Knepper, "Socially competent navigation planning by deep learning of multi-agent path topologies," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 6817–6824.

[23] G. Hoffman and C. Breazeal, "Effects of anticipatory action on human-robot teamwork efficiency, fluency, and perception of team," in *ACM/IEEE HRI*, 2007, pp. 1–8.

[24] T. Iqbal, S. Li, C. Fourie, B. Hayes, and J. A. Shah, "Fast online segmentation of activities from partial trajectories," in *ICRA*, 2019.

[25] P. A. Lasota, G. F. Rossano, and J. A. Shah, "Toward safe close-proximity human-robot interaction with standard industrial robots," in *2014 IEEE (CASE)*, pp. 339–344.

[26] R. Freedman and S. Zilberstein, "Integration of planning with recognition for responsive interaction using classical planners," in *AAAI*, 2017.

[27] E. Renaudo, S. Devin, B. Girard, R. Chatila, R. Alami, M. Khamassi, and A. Clodic, "Learning to interact with humans using goal-directed and habitual behaviors," in *Ro-Man 2015, Workshop on Learning for Human-Robot Collaboration*, 2015.

[28] A. Pokle, R. Martín-Martín, P. Goebel, V. Chow, H. M. Ewald, J. Yang, Z. Wang, A. Sadeghian, D. Sadigh, S. Savarese *et al.*, "Deep local trajectory replanning and control for robot navigation," in *2019 international conference on robotics and automation (ICRA)*. IEEE, 2019, pp. 5815–5822.

[29] M. S. Yasar and T. Iqbal, "Improving human motion prediction through continual learning," *ACM/IEEE Int. Conf. on Human-Robot Interaction (HRI), LEAP-HRI Workshop*, 2021.

[30] G. Hoffman, "Evaluating fluency in human–robot collaboration," *IEEE THMS*, 2019.

[31] T. Iqbal and L. D. Riek, "Coordination dynamics in multi-human multi-robot teams," *IEEE RA-L*, 2017.

[32] T. Iqbal, S. Rack, and L. D. Riek, "Movement coordination in human-robot teams: A dynamical systems approach," *IEEE T-RO*, 2016.

[33] T. Iqbal and L. D. Riek, "Temporal anticipation and adaptation methods for fluent human-robot teaming," in *IEEE ICRA*, 2021.

[34] M. S. Yasar and T. Iqbal, "A scalable approach to predict multi-agent motion for human-robot collaboration," in *IEEE RA-L*, 2021.

[35] T. Iqbal, M. J. Gonzales, and L. D. Riek, "Joint action perception to enable fluent human-robot teamwork," in *2015 IEEE ROMAN*.

[36] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social lstm: Human trajectory prediction in crowded spaces," in *IEEE CVPR*, 2016.

[37] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social gan: Socially acceptable trajectories with generative adversarial networks," in *IEEE CVPR*, 2018.

[38] E. Schmerling, K. Leung, W. Vollprecht, and M. Pavone, "Multimodal probabilistic model-based planning for human-robot interaction," in *IEEE ICRA*, 2018.

[39] S. H. Park, G. Lee, M. Bhat, J. Seo, M. Kang, J. Francis, A. R. Jadhav, P. P. Liang, and L.-P. Morency, "Diverse and admissible trajectory forecasting through multimodal context understanding," *ECCV*, 2020.

[40] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone, "Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data," in *European Conference on Computer Vision*. Springer, 2020, pp. 683–700.

[41] J. Bütepage, H. Kjellström, and D. Kragic, "Anticipating many futures: Online human motion prediction and generation for human-robot interaction," in *IEEE ICRA*, 2018.

[42] J. Bütepage, A. Ghadirzadeh, Ö. Ö. Karadag, M. Björkman, and D. Kragic, "Imitating by generating: Deep generative models for imitation of interactive tasks," *Frontiers in Robotics and AI*, 2020.

[43] S. Toyer, A. Cherian, T. Han, and S. Gould, "Human pose forecasting via deep markov models," in *International DICTA*, 2017.

[44] J. Martinez, M. J. Black, and J. Romero, "On human motion prediction using recurrent neural networks," in *IEEE CVPR*, 2017.

[45] E. Aksan, M. Kaufmann, and O. Hilliges, "Structured prediction helps 3d human motion modelling," in *IEEE ICCV*, 2019.

[46] V. Adeli, E. Adeli, I. Reid, J. C. Niebles, and S. H. Rezatofighi, "Socially and contextually aware human motion and pose forecasting," *IEEE RA-L*, 2020.

[47] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *IEEE CVPR*, 2016.

[48] M. S. Yasar and T. Iqbal, "Vader: Vector-quantized generative adversarial network for motion prediction," *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023.

[49] N. Lee, W. Choi, P. Vernaza, C. B. Choy, P. H. Torr, and M. Chandraker, "Desire: Distant future prediction in dynamic scenes with interacting agents," in *IEEE CVPR*, 2017.

[50] N. Rhinehart, R. McAllister, K. Kitani, and S. Levine, "Precog: Prediction conditioned on goals in visual multi-agent settings," in *IEEE ICCV*, 2019.

[51] G. J. Maeda, G. Neumann, M. Ewerton, R. Lioutikov, O. Kroemer, and J. Peters, "Probabilistic movement primitives for coordination of multiple human–robot collaborative tasks," *Autonomous Robots*, 2017.

[52] W. Mao, M. Liu, and M. Salzmann, “History repeats itself: Human motion prediction via motion attention,” in *ECCV*, 2020.

[53] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.

[54] A. Van Den Oord, O. Vinyals *et al.*, “Neural discrete representation learning,” *Advances in neural information processing systems*, 2017.

[55] J. Butepage, M. J. Black, D. Kragic, and H. Kjellstrom, “Deep representation learning for human motion prediction and classification,” in *IEEE CVPR*, 2017.

[56] X. Long, C. Gan, G. Melo, X. Liu, Y. Li, F. Li, and S. Wen, “Multimodal keyless attention fusion for video classification,” in *AAAI*, 2018.

[57] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial networks,” *Communications of the ACM*, 2020.