PoseTron: Enabling Close-Proximity Human-Robot Collaboration Through Multi-human Motion Prediction

Anonymous Author(s)

ABSTRACT

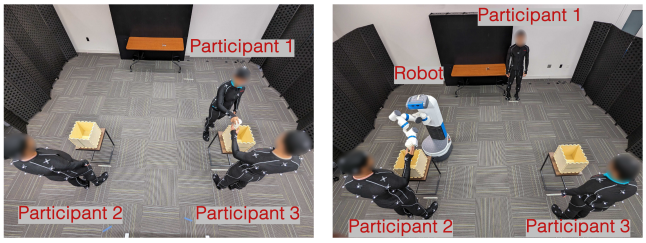
As robots enter human workspaces, there is a crucial need for robots to understand and predict human motion to achieve safe and fluent human-robot collaboration (HRC). However, achieving accurate human motion prediction remains a significant challenge due to the lack of large-scale datasets capturing close-proximity HRC and the lack of efficient, generalizable algorithms that can reliably predict the motion of multiple humans in human-robot teams. To address these challenges, we introduce INTERACT, a comprehensive multimodal dataset comprising 3-D Skeleton data, RGB+D data from two viewpoints, ego-view, eye-tracking, and gaze data of two participants, and robot joint data, covering both human-human and human-robot collaboration in teams. Next, to address the gap in learning algorithms to predict multi-human motion accurately, we propose PoseTron, a novel transformer-based encoder-decoder architecture that can generalize to multiple agents and utilize various data modalities. One of PoseTron's key contributions is the novel conditional attention mechanism in the encoder, enabling efficient extraction and weighing of motion information from all agents to incorporate team dynamics. Additionally, the decoder introduces a novel multimodal attention mechanism, which weights representations from different modalities and the encoder outputs to predict future motion accurately. We extensively evaluated PoseTron by comparing its performance on human-human and human-robot collaboration scenarios from the INTERACT dataset against state-of-the-art multi-agent motion prediction methods. The results suggest that PoseTron outperformed all other methods across all the scenarios and evaluated temporal horizons. Furthermore, we conducted a comprehensive ablation study that underscores the architectural and multimodal design choices. The superior performance of PoseTron provides a promising direction to integrate motion prediction with robot perception and enable safe and effective HRC.

ACM Reference Format:

Anonymous Author(s). 2024. PoseTron: Enabling Close-Proximity Human-Robot Collaboration Through Multi-human Motion Prediction. In *Proceedings of ACM/IEEE International Conference on Human-Robot Interaction (HRI).* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

HRI, 2024, Boulder, Colorado, USA.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn



a) Human-Human Collaboration b) Human-Robot Collaboration

Figure 1: Samples of Close-proximity Human-Human and Human-Robot Collaboration from the INTERACT Dataset.

1 INTRODUCTION

Collaborative robots (cobots) capable of safely operating in close-proximity to humans have the potential to significantly enhance efficiency and productivity across various industries, ranging from manufacturing to fulfillment [49]. At the core of achieving effective and fluent close-proximity human-robot collaboration (HRC) lies the crucial ability of robots to perceive and anticipate human intentions [11, 17, 18, 28, 31, 46, 47, 64]. This imperative need for anticipation and adaptability mirrors the fundamental aspects of human interactions. Tasks such as navigating through crowded environments or exchanging objects heavily rely on our innate capacity to observe and anticipate the actions of others [10, 13, 60, 61]. This anticipatory capability would empower robots to proactively adjust their actions, avoid collisions, and provide valuable assistance to humans in dynamic and often unpredictable environments, similar to how humans employ anticipatory and feedback mechanisms to develop suitable motor behaviors [50, 55].

The concept of anticipation has received extensive attention particularly in the social navigation and collaborative manipulation domains. The primary goal in the former is to navigate safely in the presence of humans, thus avoiding any potential interference [36, 38–40, 49]. On the other hand, anticipating the next human activity would enable robots to contribute to the task proactively, improving efficiency and taking preemptive action to enhance safety [13, 22, 30, 57, 58]. However, with the introduction of cobots, which are expected to engage with humans over extended periods in close-proximity settings, there is a need to anticipate human motion at a higher spatial and temporal granularity [19, 23, 62]. Anticipation in this scenario would entail predicting future human motion conditioned on past motion, representing a shift from the 2-D global position predicted in social navigation or categorical human activities in collaborative manipulation to 3-D skeletal joint positions.

While the concept of anticipation is crucial for cobots, the current state-of-the-art in motion prediction needs to be revised when addressing the specific challenges in close-proximity HRC. These challenges are multifaceted, with one of the primary bottlenecks being the scarcity of comprehensive real-world datasets that feature scenarios involving robots collaborating with one or more

Table 1: Summary of Publicly Available Multimodal Human-Robot Interaction Datasets.

Datasets	Setting	# Agents	Sensor Modalities							Duration of Recordings
			# 3-D Skeletons	RGB	Multi-view	Depth	Multi-view	Ego Data	No. of Person	Robot Joint Positions	(Approximate)
HARMONIC[42]	HRI (Shared Autonomy)	1	1	✔	✗	✔	✗	✔	1	✔	5 hours
MHHRI [8]	HHI and HRI	2	1	✔	✔	✔	✔	✗	N/A	✗	7 hours
MoGaze [26]	HI	1	1	✗	✗	✗	✗	✔	1	✗	3 hours
UE-HRI [4]	HRI	1	1	✔	✔	✔	✗	✗	✗	✗	12 hours
FACT HRC [54]	HRC	1	1	✔	✔	✔	✔	✔	1	✔	20 hours
INTERACT (Ours)	HHC and HRC	3 (HHC) & 4 (HRC)	3	✔	✔	✔	✔	✔	2	✔	9 hours

humans [6, 10, 54]. The availability of such datasets is essential as it is a fundamental requirement for developing and validating algorithms capable of accurately predicting human motion and intention. Furthermore, there is a notable gap in learning algorithms capable of reliably predicting the motion of multiple humans. Existing research on human motion prediction predominantly focuses on dyadic scenarios – involving one human and one robot [5, 54], or, in some cases, excludes robots altogether [26]. This limitation constrains the robot's anticipation capabilities to dyadic collaboration, which may not always reflect the nature of real-world scenarios which may haave multiple humans and/or robots.

To address the aforementioned challenges, we introduce INTERACT, a comprehensive human-human and human-robot collaboration dataset. INTERACT stands out by featuring a large-scale collection of multimodal data encompassing both Human-Human Collaboration (HHC) and Human-Robot Collaboration (HRC). Each collaborative task contains at least three participants, presenting a shift from dyadic to team interaction. INTERACT comprises assembly tasks involving three participants in HHC scenarios and four participants in HRC scenarios, with one of the participants being a robot (see Fig. 1). The dataset comprises 3-D human Skeleton joint positions of 3 participants, RGB and depth data of the workspace and the interaction from two viewpoints, ego-view data from the two human participants, eye-tracking and gaze data from two of the humans, and robot joint data, all synchronized to provide an exhaustive picture of the collaboration task. As part of the data collection effort, we recruited 63 participants, which amounted to 21 groups of three human participants for HHC and three humans and a robot participant for HRC scenarios, amounting to approximately 1 M samples of synchronized multimodal data.

INTERACT has three novel characteristics that distinguish it from prior datasets such as Mogaze [26], Thor [46], and FACT-HRC[54] (see Tab. 1). First, INTERACT represents close-proximity collaboration in groups of three humans + one robot for the HRC scenarios, which provides a novel and highly interactive scenario for data collection, distinguishing it from other datasets [26, 54] which were limited to dyadic interaction. Second, we introduce a simple but effective setup for data collection where the robot was completely autonomous in close-proximity settings with humans, which is a shift from other datasets that were primarily tele-operated or featured a form of shared autonomy. Finally, to understand the present state of HRC, we collected data on the same task comprising only human participants. This allows us to investigate how HRC compared to HHC and provides a mechanism to systematically analyze how humans behave differently in the presence of a robot collaborator. In addition, it allows the research community to evaluate the generalizability of their algorithms by training them on data from HHC scenarios and testing in HRC scenarios.

To address the gap in learning algorithms that can accurately predict the motion of multiple agents, we propose a novel and efficient transformer architecture [59], PoseTron (*pronounced "pos·i·tron"*). PoseTron employs an encoder-decoder framework where the encoder is tasked to extract spatio-temporal representation in human motion, fuse representations from diverse modalities, and learn the interaction dynamics among all agents. Additionally, we introduce specialized attention modules to capture agent-specific motion patterns. We employ self-attention mechanisms for extracting spatio-temporal features within each agent's motion and conditional attention mechanisms that enable agents to incorporate team dynamics by querying another agent's representations. The encoder output comprises encoded representation from non-skeletal modalities and skeleton representation that encapsulates the complexity of individual human motion and team dynamics.

The output of the encoder, along with the last observed motion, is passed to the decoder. The decoder employs an auto-regressive mechanism for future motion prediction. Similar to the encoder, it incorporates learnable positional embeddings to encode positional features into observed and generated motion sequences. Additionally, the decoder utilizes a conditional attention mechanism to incorporate salient representation from its generated output and the encoder's representations.

We conducted extensive experiments to assess the efficacy of PoseTron by deploying on the INTERACT dataset. Our experiments included evaluating the performance of PoseTron on i) HHC-train, HHC-test, ii) HRC-train, HRC-test, and iii) HHC-train, HRC-test setups. Our results suggest PoseTron consistently outperformed the state-of-the-art approaches over all three evaluation scenarios. Furthermore, we conducted a comprehensive ablation analysis of PoseTron's learning modules and the relevance of multimodal data in the INTERACT dataset. The results validate PoseTron's architectural choices and underscore its ability to leverage complementary information from diverse data sources. The outcomes of these experiments promise to narrow the gap in close-proximity HRC by providing a substantial dataset and valuable insights for advancing state-of-the-art anticipation techniques.

2 RELATED WORK

2.1 Multimodal Datasets in HRI

Multimodal datasets have garnered interest in many different communities, such as human–robot interaction [42, 46, 54], computer vision [14, 29], action recognition [7, 25, 51], and natural language processing [27, 32]. In the context of HRI, capturing and analyzing data from multiple modalities is essential for enabling robots to understand, anticipate, and coexist with humans in diverse environments. Along this line, new datasets have been proposed providing multimodal data for shared-autonomy [42], social navigation [46],

and dyadic HRC [54]. For example, Newman et al. [42] introduced the HARMONIC dataset, which captures multimodal data, including RGB, Gaze, and Robot information, during human-robot collaboration tasks, offering potential applications in intent prediction, mental modeling, and shared autonomy. To investigate the relationship between personality traits and engagement in human-human and human-robot interactions, Celikhutan et al. [8] proposed the MHHRI dataset, which comprises dyadic (Human-Human) and triadic (Human-Human-Robot) interactions. Tian et al. [54] presented the FACT-HRC dataset, which centers around human-robot handover interactions in collaborative environments.

While existing datasets have made strides in addressing various aspects of HRI, a noticeable gap persists in the availability of datasets tailored to close-proximity HRC scenarios involving multiple humans. Furthermore, many of these datasets rely on teleoperation within Wizard-of-Oz setups, which does not accurately represent how humans will naturally interact with an autonomous robotic agent. The challenges of recruiting human participants, physically co-locating robots and humans, and the imperative to uphold human privacy rights further compound the limitations of such datasets. As a result, most datasets feature a limited number of participants, which often fails to capture the complexity and diversity of behaviors encountered in real-world HRI settings.

2.2 Human Motion Prediction

Human motion prediction is widely considered one of the essential parts of robotic intelligence that would enhance robot perception and allow for rapid and high fidelity reactions towards complex environment changes [35, 37, 47, 53]. The notion of prediction has found application in diverse areas within HRI, spanning shared autonomy [42, 44], social navigation [40, 41], and autonomous vehicles [16, 48]. Tang et al. [53] illustrated the relevance of motion prediction in planning, computing conditional probability density for the trajectories of other agents based on a hypothetical rollout of the self-agent. Yasar et al. [63] proposed a multi-agent adversarial auto-encoder approach for predicting future human motion, with the authors using a self-attention mechanism to weigh the different agent representations before predicting future motion. Adeli et al. [1] proposed a social pooling mechanism on top of the seq2seq architecture for predicting multi-agent motion prediction.

Human motion prediction serves as a crucial component in ensuring the safety of HRI. In order to fully enable close-proximity HRC, robot perception must possess the capability to anticipate human motion with greater spatial and temporal precision. While recent advancements in motion prediction have pushed the boundaries of what is possible, they often fall short in terms of their applicability to HRC scenarios. This limitation arises from the fact that these approaches are typically trained on datasets exclusively consisting of primarily single-agent human motion, often in the absence of robots within the workspace.

3 INTERACT: HHC AND HRC DATASET

In this section, we present our first solution for enabling close-proximity HRC: the INTERACT dataset. Our proposed dataset stands out from other datasets in close-proximity HHC and HRC scenarios by providing a large-scale collection of synchronized multimodal data, as illustrated in Fig 1 and summarized in Tab. 2. The dataset includes 3-D Skeletal joint data of human participants, RGB and depth data from two viewpoints in the workspace, ego-view, eye-tracking, and gaze positions data from the two human participants, and robot joint data. The comprehensive dataset can be leveraged for various tasks, including motion prediction, goal prediction, and imitation learning.

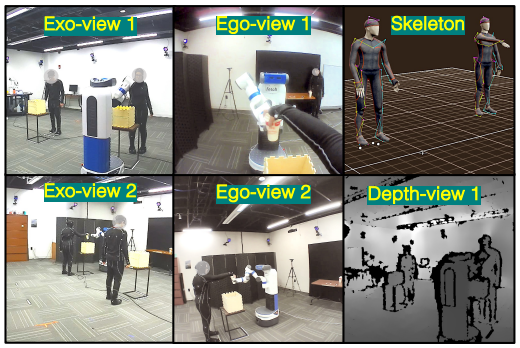


Figure 2: Human-Robot Collaboration samples from the INTERACT dataset. The dataset comprises 3-D skeletons from three participants, and RGB+D Camera views from two perspectives and Ego POV from two Participants.

3.1 Study Apparatus and Implementation

The objective of collecting each of the modalities in INTERACT is to provide the robot with a comprehensive understanding of the interaction and its surrounding environment. For data collection in INTERACT, we utilized the OptiTrack Motion Capture system [20] for collecting 3-D Skeleton data, 2 ZED Cameras from StereoLabs [21] to capture RGB and depth from two different viewpoints of the workspace, and 2 eye-tracking devices by Pupil Labs [15] for collecting ego point of view data, eye-tracking and gaze data. For HRC, we introduce a Fetch Robot [12], which is a mobile manipulator in the shared workspace to collaborate with the other participants.

Prior to data collection, all equipment were meticulously calibrated to ensure accurate data synchronization. For collecting 3-D Skeleton data, participants were equipped with a full-body motion tracking suit by OptiTrack, comprising 41 passive markers. The 3-D Skeleton poses represent a lightweight and accurate source of information for predicting human intent. In addition to 3-D Skeleton data, we incorporated RGB and depth data obtained from two cameras in opposing corners of the workspace, as illustrated in Fig. 2. This setup allowed us to maximize coverage and provide diverse perspectives on the collaboration. To further enhance the prediction of human intent and motion, we equipped two participants with eye-tracking devices, enabling the collection of valuable first-person viewpoints in addition to the third-person perspectives captured by other sensors, with prior work showing the benefit of eye-tracking and gaze information for predicting human intent [26, 42, 54]. Finally, in HRC scenarios, we deployed the Fetch robot and collected robot joint data.

3.2 Human Ethics

Our study protocols were reviewed and approved by the Institutional Review Board. All participants provided informed consent for

participating in the study and having their data recorded as part of a public dataset for research purposes. At the end of the study, participants were compensated with a $30 gift card for approximately 2 hours of their time.

3.3 Participants

A total of 63 adults participated in the study (31.7% female ($n = 20$), 66.67% male ($n = 42$) and 1.58% non-binary ($n = 1$)). The mean age of the participants was 23.65 years ($SD = 3.91$). The participants were predominantly right-handed 84.1% ($n = 53$) and 15.9% left-handed ($n = 10$). Participants also recorded their experience with robots on a Likert scale from "no experience" (1) to "expert-level experience" (5), with the mean experience level at 2.21 ($SD = 1.19$).

3.4 Data Collection Procedure

All the tasks involved three human participants in both HHC and HRC scenarios. This arrangement led to 21 groups from the 63 recruited participants. Each group engaged in 12 collaborative assembly task sessions, with an equal distribution of 6 HHC and 6 HRC sessions. Within each scenario, two variations were introduced – one with obstacles in the workspace and one without.

Pre-Task Survey: Before beginning the study, participants were asked to review consent documents and task instructions. They then filled out a pre-task survey, which collected demographic information and their prior robot experience. Next, all three participants were equipped with Motion Capture suits to collect 3-D Skeleton Pose data. Two participants (Participants 2 and 3) also wore the eye-tracker, which would collect their ego point-of-view, eye-tracking, and gaze data during the task.

The scenarios were counterbalanced, and each group was assigned either HHC or HRC as their initial scenario. After completing all the sessions for their initial scenario, they switched to the other. They participated in six sessions within each scenario, three for each variation. In Variation-1, there were no obstacles in the workspace, whereas in Variation-2, there were obstacles that participants would have to move around. To minimize the learning effect, participants rotated roles after each session. For instance, if a person started as Participant 1 in the first session, they became Participant 2 in the second and Participant 3 in the third session. This rotation applied to all group members in both HHC and HRC scenarios, ensuring balanced role distribution across variations.

Human-Human Collaboration (HHC) Scenario: In this scenario, three human participants collaborated on an assembly task. In each session, Participant 1 transported cups to Workspace 2 and Workspace 3 three times each, while Participants 2 and 3 followed this workflow:

- Received a cup from Participant 1.
- Moved to Workspace-1.
- Extracted Lego pieces and instructions from the cup.
- Assembled Lego pieces as per instructions.
- Repeated steps 1-4 for three times.

The session was considered complete when Participants 2 and 3 assembled the Lego structure as specified in the instructions. We used different Lego structures for different variations to minimize the learning effect. After three sessions, obstacles were introduced to the workspace, or the scenario changed.

Table 2: Summary Statistics of the INTERACT Dataset.

Scenario	Number of Participants	Variation	Average Duration (sec)	Total Timestamps	Total Multimodal Frames (Million)
HHC	3 H	w obstacle	104.5	201 K	1.20 M
		w/o obstacle	127.0	263 K	1.58 M
HRC	3 H + 1 R	w obstacle	140.9	300 K	1.80 M
		w/o obstacle	149.8	306 K	1.84 M
Total	-	-	-	1.07 M	6.42 M

Human-Robot Collaboration (HRC) Scenario: In this scenario, a Fetch robot (as Participant 4) joined three human participants to complete a similar assembly task. The robot received cups from Participant 1 and transported them between Workspaces 2 and 3. Participants 2 and 3 followed a similar workflow to the HHC scenario.

Post-Session Survey: After each session, participants filled out a post-session survey containing questions that covered various aspects: participant-specific questions (e.g., "I needed to observe and anticipate the activities of group member-1/2/3"), group-specific questions (e.g., "Which group member had the greatest impact on the coordination of the group?") and robot-specific assessments, rated on a Likert scale (e.g., "The robot was effective in coordinating the actions with both the group members").

4 MULTI-AGENT MOTION PREDICTION

Our objective is to improve the robot's perception by providing it with the capability to forecast the motion of all human collaborators in the team. Human motion prediction is formally described as the task of estimating the future human pose for a certain period, given their past pose. We will present the problem for single-agent motion prediction and later extend the formulation to multiple humans. We assume access to 3-D skeletal joint positions as the primary data source, along with additional modalities (e.g., RGB). Our notation consistently utilizes superscripts to indicate agents and subscripts to represent time across all formulations.

We begin by considering the scenario of an individual agent, denoted as agent i. The objective here is to predict the future trajectory of this agent's pose, given the observed pose trajectory spanning from time $t = 1$ to τ, represented as $\mathbf{X}^i = \{x_1^i, \ldots, x_\tau^i\}$, and any additional sensor data from other sources, referred to as $D = \{d_1, \ldots, d_\tau\}$. In this context, each pose frame $x_t^i \in \mathbb{R}^N$ represents the skeletal pose in an N-dimensional space. The dimensionality, N, is determined by the number of joints, indicated as J, in the skeleton and the dimension of each joint, with $N = 3 \times J$. The input frame from other sensors, $d_t \in \mathbb{R}^N$, comprises raw data from complementary modalities such as RGB and Gaze data.

The model's objective is to generate future trajectory frames within a time horizon H, denoted as $\mathbf{Y}^i = \{y_{\tau+1}^i, \ldots, y_{\tau+H}^i\}$. Our primary goal is to acquire the underlying representation that enables the model to accurately predict plausible future human poses, which are denoted as $\hat{\mathbf{Y}}^i = \{\hat{y}_{\tau+1}^i, \ldots, \hat{y}_{\tau+H}^i\}$. We work under the assumption that predicting future human poses relies on past observed and generated poses, and we predict each frame in an autoregressive manner, as described below:

$$p_\theta(\hat{\mathbf{Y}}^i) = \prod_{\delta=\tau+1}^{\tau+H} p_\theta(\hat{y}_\delta^i | \hat{y}_{\tau:\delta-1}^i, x_{1:\tau}^i, d_{1:T}) \tag{1}$$

In the context of multi-agent motion prediction, the input consists of the observed poses of all agents in the scene from time $t = 1$ to τ: $\mathbf{X} = \{X^1, \ldots, X^K\} = \{x_1^{1:K}, x_2^{1:K}, \ldots, x_\tau^{1:K}\}$ and additional

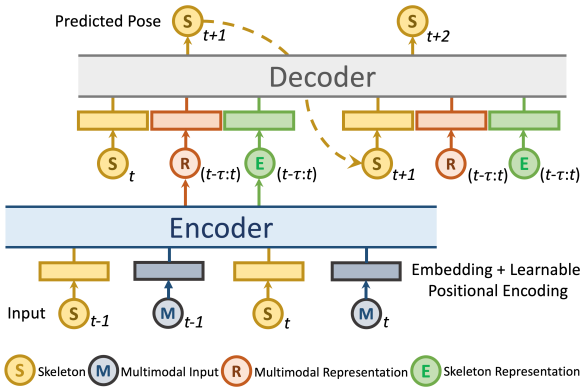


Figure 3: Overall Architecture of PoseTron. PoseTron consists of two modules: Multimodal Pose Encoder and Decoder. The Encoder encodes the motion of all the agents and modalities of all the data streams. The Decoder uses the encoded skeleton representation and the multimodal non-skeletal representation to forecast future pose.

multimodal input: $\mathbf{D} = \{d_1, \ldots, d_\tau\}$. The expected output of the model is the future trajectory frames over a horizon H, which represents the ground truth poses over the horizon $t = \tau + 1$ to $\tau + H$: $\mathbf{Y} = \{Y^1, \ldots, Y^K\} == \{y^{1:K}_{\tau+1}, y^{1:K}_{\tau+2}, \ldots, y^{1:K}_{\tau+H}\}$. Thus, the multi-agent motion prediction problem can be formulated as follows:

$$p_\theta(\hat{\mathbf{Y}}^i) = \prod_{\delta=\tau+1}^{\tau+H} p_\theta(\hat{y}^i_\delta | \hat{y}^i_{\tau:\delta-1}, x^{1:K}_{1:\tau}, d_{1:\tau});\ \forall i = 1, \ldots, K \tag{2}$$

5 POSETRON

We now introduce our proposed framework for multi-agent human motion prediction: PoseTron. PoseTron (see Fig. 3) is a multimodal sequence learning architecture that aims to accurately predict the future poses of all humans, irrespective of their number or their collaborative scenario (HHC/HRC). PoseTron comprises two specialized modules: the encoder (Sect. 5.1), which aims to encode the motion of all the agents and modalities of all the data streams, and the decoder (Sect. 5.2) which uses the encoded representation to forecast future human pose.

5.1 Multimodal Pose Encoder

The input to the Encoder is observed motion of all the agents, comprising agent-specific skeletal input $\mathbf{X} = \{X^1, \ldots, X^K\}$ and non-skeletal input $\mathbf{D} = \{d_1, \ldots, d_\tau\}$, as depicted in Fig. 3. For the skeleton sequence spanning T timesteps, we extend the 3-D joint position with time derivatives: velocity and acceleration. Thus, the original input, comprised of T tokens, is extended to $3T$ tokens.

5.1.1 *Input Embedding.*

Skeleton modalities: We separately encode the skeletal input for each agent, $X^i_{input} \in \mathbf{X}$. The input sequence is first passed through an embedding layer to convert pose information into $d-$dimensional vectors. Next, we add positional encoding to each input frame. This is required as we are not using recurrent neural architectures, instead relying on a simple feedforward architecture, following the transformer implementation [59], which lacks the inherent notion of token order or position. We use a learnable positional encoding instead of the fixed sinusoidal positional encoding of the transformer architecture [59]. The operations can be formulated as follows:

$$\begin{aligned} X^i_{\text{token}} &= \text{E}(X^i_{input}); X^i_{\text{positional}} = \text{PE}(X^i_{\text{token}}) \\ X^i_{\text{embed}} &= X^i_{\text{token}} + X^i_{\text{positional}} \end{aligned} \tag{3}$$

Here, E represents the token embedding function that maps input tokens X^i_{input} to token embeddings X^i_{token}. PE represents the learnable positional embedding function that is required to inject information about the relative or absolute position of the tokens in the sequence. The operations in Eq. 3 are repeated for velocity and acceleration input.

Non-Skeleton modalities: For vision modalities such as RGB, the encoding process involves leveraging a feature extractor to obtain representation over a time horizon, followed by a temporal encoding of these features. We use a pre-trained SwinTransformer [33] architecture to extract features. This allows us to reduce the training footprint and leverage existing architectures for extracting rich representations. We pass the extracted representations, which have a dimension of $\mathbb{R}^{\mathbb{T}\times\mathbb{K}}$ with K and T being the feature dimension and timesteps, respectively, through the input embedding layers, using the same operations as Eq. 3 to add positional encoding. The overall operations are summarized as follows:

$$\begin{aligned} X_{features,\, m,\, t} &= \text{FE}(X_{m,\, t}) \\ X_{embed,\, m,\, t} &= \text{InputEmbeeding}(X_{features,\, m,\, t}) \end{aligned} \tag{4}$$

Here, m represents the modality, which can be one of RGB, Gaze, or any other available modality, and $X_{m,t}$ is the raw input of the modality. FE represents a pre-trained feature extractor, which is used to extract representations $X_{features,m,t}$. The extracted representations are then passed to the input embedding function, previously defined in Eq. 3.

5.1.2 *Multi-Head Self-Attention.*

The self-attention module is crucial in establishing temporal connections among individual skeleton embeddings. These skeletal embeddings undergo a self-attention process, enabling our framework to assess the significance of various tokens within the input sequence while handling each token. In this process, every position in the input sequence is linked to a weighted sum of all positions, including itself. These weights are determined dynamically based on the similarity between positions. The mechanism used to calculate a weighed representation for each position is as follows:

$$\begin{aligned} Q = X_{embed}W^Q;\ K = X_{embed}W^K; V = X_{embed}W^V \\ \text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \end{aligned} \tag{5}$$

Here, Q represents the query matrix representing the queries for each token, K represents the key matrix denoting the keys for each token, and V represents the value matrix denoting the values for each token. For each token, we calculate the attention scores over itself and all other tokens using the softmax function. 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.

We pass the token embeddings X^i_{embed} through the attention mechanism (Attention) to obtain X^i_{att}. We repeat the aforementioned operations in Eqs. 3, 5 for each agents, thus obtaining X^i_{att};

$\forall i = 1, \ldots, K$. The operations can be represented as follows:

$$X^i_{att} = \text{Self-Attention}(X^i_{embed}, X^i_{embed}, X^i_{embed}) \tag{6}$$

5.1.3 Multi-Head Cross-Attention. Having computed the attention weights over skeleton tokens for each agent, the next task is to incorporate team dynamics by learning the association between tokens of different agents. To achieve this, we compute cross-attention scores for each agent-specific token X^i_{att} in the following manner: For a given agent, denoted as i, we treat their tokens as queries and compute key and value matrices for the remaining two agents, j and k. We then determine attention weights and calculate their averages to derive the ultimate token representation for each agent. These operations can be concisely summarized as follows:

$$\begin{aligned} X^{i,j}_{cross-att} &= \text{Cross-Attention}(X^i_{att}, X^j_{att}, X^j_{att}), \\ X^{i,k}_{cross-att} &= \text{Cross-Attention}(X^i_{att}, X^k_{att}, X^k_{att}), \\ X^i_{cross-att} &= \text{Mean}(X^{i,j}, X^{i,k}). \end{aligned} \tag{7}$$

5.2 Multimodal Pose Decoder

The input to the decoder is the agent-specific representation of the past motion *and* the multimodal representation from other non-skeleton modalities (see Fig. 3). Unlike prior multimodal approaches, which fuse representations from different modalities at the encoder [9, 45, 56], we choose to fuse the multimodal representations at the decoder. Fusing the representation at the decoder allows the decoder to leverage the context and dependencies between different modalities, which could lead to a more accurate generation. The decoder is auto-regressive, meaning it predicts future poses one at a time, taking into account the previously generated poses. We use the same decoder to generate the agent-specific poses separately.

5.2.1 Input Embedding. Similar to the encoder, the decoder has a separate input embedding and positional encoding. However, it must accommodate variable input sizes based on the size of the generated poses. In the initial decoding step, we pass the last observed pose along with the encoded skeletal and non-skeletal multimodal representations. We add each generated pose to the decoder's input for each subsequent step while keeping the encoded representations constant. The operations are similar to Eq. 3, and is summarized below:

$$X^i_{dec-embed,t} = \text{InputEmbedding}(X^i_{dec-input,t}) \tag{8}$$

5.2.2 Multimodal Attention. The embedding from the decoder input, denoted as X^i_{dec}, along with the output of the encoder representation, is passed to the encoder-decoder attention module, $X^i_{cross-att}$ and $X_{embed,\,m,\,t}$. We use a similar attention mechanism as previously mentioned in Eq. 5. Here, the query is the decoder input, and the value and key are the output of the encoder, $X^i_{cross-att}$ and $X_{embed,\,m,\,t}$. The operations can be summarized as follows:

$$\begin{aligned} X^i_{encoded} &= \text{Concat}(X^i_{cross-att}, X_{embed,\,m,\,t}) \\ X^i_{decoder-att} &= \text{Decoder-Attention}(X^i_{dec-embed,t}, X^i_{encoded}, X^i_{encoded}) \end{aligned} \tag{9}$$

5.2.3 Output Embedding. The output of the decoder attention module is finally passed through linear layers to generate the output pose. The operations can be formulated as follows:

$$X^i_{output,t} = \text{OE}(X^i_{decoder-att}) \tag{10}$$

Here, $X^i_{output,t}$ represents the generated future pose at time t of agent i. OE represents the output embedding, which is a linear projection of the decoder attention output to the pose space.

6 EXPERIMENTS

In this section, we present our experimental details and results. We introduce the dataset and evaluation metric in Sect. 6.1, the implementation details in Sect. 6.2, and the results and discussion in Sect. 6.3.

6.1 Experimental Setup and Evaluation Metric

6.1.1 INTERACT Dataset. The proposed INTERACT dataset comprises 252 sessions of multimodal close-proximity collaboration data, which are evenly split into 126 episodes of HHC and 126 episodes of HRC scenarios, providing a large-scale dataset of multimodal and multi-agent interaction. For all the evaluation scenarios, we adopt a cross-group evaluation strategy, where we train and test on separate groups. The training set comprises all the even-numbered groups from 1 to 21, and the testing set comprises all the odd-numbered groups from 1 to 21 (recall that the dataset contains 21 groups in total). We propose three evaluation setups:

- HHC-Train, HHC-Test: In this setup, we trained and tested all the evaluated approaches on only Human-Human Collaboration data, using the aforementioned train and test sets. This approach allows us to establish how these approaches perform for multi-agent human motion prediction.
- HRC-Train, HRC-Test: In this setup, we trained and tested all the evaluated approaches on only Human-Robot Collaboration data. The purpose of this setup is to investigate the extent to which performance is influenced by the presence of a robot.
- HHC Train, HRC Test: In this setup, we use the train set of the HHC setup and the test set of the HRC setup. Here, we investigate how models trained on HHC generalize to HRC.

6.1.2 Evaluation metric. We report the Mean Per Joint Position Error (MPJPE) on 3D joint coordinates, a widely used metric for evaluating pose prediction performance [1, 2, 37, 63]. Since our dataset includes multiple agents, we also compute the Per Agent-MPJPE (PA-MPJPE) by averaging this evaluation metric across all agents. PA-MPJPE quantifies the average $L2$-Norm differences between each agent's predictions and ground truth. For all evaluated models, the input and output sequences have 25 timesteps.

6.2 Implementation Details

In all experiments, we employ 3-D Skeletons and RGB data from two camera views as input, with the output being all agents' future 3-D Skeleton poses. While PoseTron adopts a modular architecture capable of accommodating the additional modalities, we restricted to using only 3-D Skeleton and RGB data in this work. We will explore the other modalities as part of future work. The feature dimension for Skeleton joints is 3×51 for each agent, while the RGB data from the two ZED cameras is pre-processed to dimensions of $3 \times 224 \times 224$ for each view. Both input and output sequences have a length of 25.

Table 3: PA-MPJPE of different multi-agent motion prediction methods for the evaluation setup: HHC Train, HHC Test.

Approaches	5	10	15	20	25
Joint Learning [1]	5.74	7.68	9.33	10.85	12.30
Joint Learning + Social [1]	8.29	9.61	10.87	12.06	13.21
MA-AAE [63]	3.27	5.06	6.75	8.39	10.01
PoseTron (Skeleton Only)	3.35	5.04	6.51	7.83	9.03
PoseTron (Multimodal)	**3.21**	**4.86**	**6.34**	**7.65**	**8.84**

All the experiments were conducted using PyTorch [43] 2.0.1 running on an NVIDIA A100 GPU. For all the evaluated methods, we utilized a batch size of 256 and fine-tuned hyperparameters for optimal results. For extracting RGB features, we use a pre-trained SwinTransformer [33]. For PoseTron, we configured the encoder and decoder to have 8 attention heads while keeping the feedforward layer dimension fixed at 256. We used two stacks of encoder and decoder. For training PoseTron, we use the AdamW [24] optimizer with cosine annealing and warm restarts [34] with an initial learning rate of 0.001. We trained each evaluated approach for a maximum of 150 epochs, with a training time of approx. 3 hours.

6.3 Results and Discussion

In this section, we compare our approach, PoseTron, with three state-of-the-art multi-agent motion prediction approaches: Joint Learning [1], Joint Learning + Social [1] and Multi-Agent Adversarial Auto-encoder (MA-AAE) [63]. Joint Learning and Joint Learning + Social represent sequence2sequence [37, 52] approaches for motion prediction using pooling mechanisms to obtain joint representations over all the agents. On the other hand, MA-AAE uses a multi-agent adversarial auto-encoder with a self-attention mechanism to obtain interaction dynamics between multiple agents. We report the PA-MPJPE at distinct frame intervals, 5, 10, 15, 20, and 25, to evaluate model performances over different horizons.

6.3.1 Human-Human Collaboration Scenarios.

Results: In Table 3, we compare the performance of our method, PoseTron against state-of-the-art multi-agent motion prediction methods on the HHC Train, HHC-Test setup. We use the same training and testing strategy for all the evaluated methods. As can be observed in Table 3, PoseTron (Skeleton Only) and PoseTron (Multimodal) strongly outperformed all the state-of-the-art approaches on every frame interval, attaining the lowest PA-MPJPE.

Discussion: The two variants of PoseTron outperformed all the state-of-the-art approaches, which underlines the architectural improvements over the other evaluated methods. While all the evaluated approaches use recurrent neural networks for their sequence learning backbone, PoseTron adopts the transformer approach, which allows it to consider all the frames at the encoder and decoder. This allows PoseTron to extract salient representations from all the available frames. PoseTron also differs in its approach to modeling the interaction between multiple agents. While Joint Learning + Social [1] uses social pooling, an approach that has also been used in social navigation [3, 16], and MA-AAE [63] uses the self-attention mechanism, PoseTron uses the conditional attention at the encoder where for a given query agent, it can separately attend and weigh over the different motion frames of the other agents. The combination of the conditioning and the feedforward architecture, which allows each frame token to attend to all other tokens, enables superior representation learning at the encoder.

Table 4: PA-MPJPE of different multi-agent motion prediction methods for the evaluation setup: HRC Train, HRC Test.

Method	5	10	15	20	25
Joint Learning [1]	6.08	7.72	9.05	10.24	11.36
Joint Learning + Social [1]	7.37	8.73	9.95	11.06	13.26
MA-AAE [63]	3.34	5.08	6.77	8.44	10.03
PoseTron (Skeleton Only)	2.90	4.46	5.89	7.22	8.38
PoseTron (Multimodal)	**2.71**	**4.22**	**5.51**	**6.64**	**7.65**

6.3.2 Human-Robot Collaboration Scenarios.

Results: In Table 4, we report the performance of PoseTron against state-of-the-art multi-agent motion prediction methods on the HRC evaluation setup. Similar to the HHC evaluation setup, we were consistent with the training and testing strategy for all the evaluated methods. The two variants of our approach PoseTron (i.e., Skeleton and Multimodal) again outperformed the state-of-the-art evaluated methods consistently over all the frame intervals.

Discussion: The results presented in Table 4 emphasize the superior performance of PoseTron when compared to the other evaluated methods. Similar to the evaluation for HHC in Table 3, we observed that both variants of PoseTron achieved superior performance. In addition to the encoder operations that provide PoseTron with superior representation and sequence modeling capabilities, it distinguishes itself from other approaches through its unique decoder strategies, which further enhance its performance. While all the approaches follow an auto-regressive approach, PoseTron stands out as the only one capable of attending to all generated and past frames. This capability is achieved without increasing complexity, as it reuses the attention mechanism within the decoder. The key and value matrices are iteratively updated as the number of generated poses increases. Moreover, PoseTron (Multimodal) leverages multimodal representations in the decoder, enabling it to query multimodal features during the generation of future poses. The combination of these operations contributes to its improved performance over PoseTron (Skeleton Only) and other approaches.

6.3.3 Training on HHC, Testing on HRC.

Results: In Table 5, we present PoseTron's performance against state-of-the-art multi-agent motion prediction methods, with all methods being trained on the HHC training set and tested on the HRC test set, ensuring no group overlap. This allows us to assess the generalizability of existing approaches exclusively trained on HHC to HRC scenarios. The test set is the same as in Table 4 for direct comparison. Both PoseTron variants consistently outperform state-of-the-art methods across all frame intervals, demonstrating superior generalizability over the two scenarios.

Discussion: The results presented in Table 5 provide a strong indication of PoseTron's generalizability compared to other approaches. PoseTron's generalizability can be attributed to the attention mechanisms in both the encoder and decoder, which allows PoseTron to efficiently utilize the multiple streams of agent and multimodal data in the context of PoseTron(Multimodal). Compared to Table 4, we observed some interesting trends. Firstly, all the approaches had a performance drop when training on HHC and testing on HRC, compared to training and testing on HRC. As the test is the same,

Table 5: PPA-MPJPE of different multi-agent motion prediction methods for the evaluation setup: HHC Train, HRC Test.

Method	5	10	15	20	25
Joint Learning [1]	6.29	8.19	9.78	11.23	12.57
Joint Learning + Social [1]	8.77	10.33	11.85	13.32	14.73
MA-AAE [63]	3.46	5.33	7.15	8.92	10.64
PoseTron (Skeleton Only)	3.69	5.53	7.12	8.52	9.76
PoseTron (Multimodal)	**2.82**	**4.61**	**6.25**	**7.80**	**9.25**

and only the training data is different, this provides the strongest signal on the importance of HRC data.

6.3.4 Ablation Results. In this section, we compare the performance of PoseTron with ablated versions of itself, firstly at an architectural level where we remove the learnable positional embedding: PoseTron w/o L.P.E (Multiple RGB View), and use the default non-linearity of the original transformer [59], instead of SwishGLU which was used in PoseTron. Next, for the same ablated version, we further remove the RGB modalities, instead using only Skeleton: PoseTron w/o L.P.E (Skeleton Only). Finally, we ablate the modalities, first keeping one RGB view: PoseTron (One RGB View) and then using only Skeletons: PoseTron (Skeleton Only).
Results: We report the results of all the ablation experiments in Table 6, where we trained and tested on the HRC scenario. The results suggest that PoseTron using Multiple RGB View and Learnable Positional Embedding attained the best performance. The next best performing architecture was PoseTron with one RGB view: PoseTron (One RGB View). This was followed by PoseTron w/o L.P.E (Multiple RGB View), with PoseTron w/o L.P.E (Skeleton Only) performing worst of all the ablated versions. **Discussion:** The results in Table 6 validate the architectural decisions made in designing PoseTron. Notably, PoseTron (Multiple RGB View) demonstrated superior performance, underscoring the advantages of incorporating multiple camera views alongside 3-D Skeletons. Following closely in performance was PoseTron (One RGB View), emphasizing two key insights: i) The utilization of additional modalities can enhance performance, and ii) PoseTron's cross-attention mechanism in the decoder effectively leverages multimodal features to construct a comprehensive representation.

The next best performing architecture was PoseTron w/o L.P.E (One RGB View), which features a variant of PoseTron that uses fixed positional encoding [59]. In addition, we ablate the SwishGLU activation function and replace it with ReLU. As observed in Table 6, the removal of these architectural details resulted in a performance drop over all the horizons, justifying the design choices in PoseTron.

Finally, the two Skeleton Only variants had higher prediction errors compared to the multimodal variants. Here again, PoseTron (Skeleton Only) with learnable positional embedding and SwishGLU activation outperformed PoseTron (Skeleton Only) without these features. This further emphasizes the effectiveness of incorporating multimodal information into the architecture.

6.4 Overall Discussion

The experiments provide several key insights in the context of motion prediction in HRC. One of the consistent themes across all the evaluation setups is the superior performance of PoseTron compared to state-of-the-art approaches in the field. One of the key distinguishing factors of PoseTron is its encoding mechanism, whereby it uses learnable positional encoding to add the notion of sequence, unlike the other recurrent approaches. This allows PoseTron to exploit the full context of the input and the generated sequence for its prediction. Furthermore, PoseTron introduces a novel mechanism to model the interaction among multiple agents through conditional attention. This enables individualized attention to different motion frames of other agents. Combined with the self-attention mechanism, this design choice leads to superior representation learning at the encoder.

Table 6: Ablation Study: HRC Train, HRC Test.

Approaches	5	10	15	20	25
PoseTron w/o L.P.E (Skeleton Only)	3.11	4.76	6.17	7.46	8.64
PoseTron w/o L.P.E (One RGB View)	3.00	4.56	5.92	7.15	8.23
PoseTron (Skeleton Only)	2.90	4.46	5.89	7.22	8.38
PoseTron (One RGB View)	2.85	4.40	5.75	6.97	8.05
PoseTron (Multiple RGB View)	**2.71**	**4.22**	**5.51**	**6.64**	**7.65**

Another contributing factor to PoseTron's superior performance is its decoder strategy. PoseTron stands out from other approaches by attending to all generated and past frames. Additionally, it incorporates multimodal representations into the decoder, enabling it to query multimodal features when generating future poses. This enhanced approach, as demonstrated in all experiments (Tables 3, 4, 5) and the ablation study (Table 6), leads to more accurate pose predictions.

The experiments also highlight the significance of the data source in training motion prediction models. As observed in Tabs. 4, 5, the performance dropped when the models were trained in one scenario and tested in another scenario. This emphasizes the need for more specialized datasets catering to the HRC setups. Notably, even in this setup, we observe the PoseTron's strong generalizability, as it attained the best performance. This generalizability is crucial in real-world applications where the ability to adapt to different scenarios is essential. While PoseTron and all other approaches were trained on an NVIDIA A100 GPU for training efficiency, we successfully ran PoseTron on a consumer-grade GPU: the NVIDIA RTX 2080Ti. This provides a pathway for our future work, which will focus on deploying PoseTron in real-time human-robot collaboration scenarios as part of the robot's perception stack.

7 CONCLUSION

In this work, we aimed to address two of the open challenges to enabling close-proximity human-robot collaboration by presenting INTERACT, a comprehensive multimodal dataset featuring Human-Human and Human-Robot Collaboration scenarios, and PoseTron, a novel transformer-based framework for multi-agent motion prediction. With INTERACT, we have laid the foundation for developing robust algorithms to facilitate close-proximity HRC. Furthermore, PoseTron's encoder-decoder framework, featuring novel conditional and multimodal attention mechanisms, has demonstrated remarkable performance gains over existing methods in various HHC and HRC scenarios. Our extensive experiments and insights offer a promising direction to enable safe and fluent close-proximity human-robot collaboration. As part of our future work, we will incorporate PoseTron to the robot's perception stack and evaluate its efficacy in real-world HRC scenarios.

REFERENCES

[1] Vida Adeli, Ehsan Adeli, Ian Reid, Juan Carlos Niebles, and S Hamid Rezatofighi. 2020. Socially and Contextually Aware Human Motion and Pose Forecasting. *IEEE RA-L* (2020).

[2] Emre Aksan, Manuel Kaufmann, and Otmar Hilliges. 2019. Structured prediction helps 3d human motion modelling. In *IEEE ICCV*.

[3] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. 2016. Social lstm: Human trajectory prediction in crowded spaces. In *IEEE CVPR*.

[4] Atef Ben-Youssef, Chloé Clavel, Slim Essid, Miriam Bilac, Marine Chamoux, and Angelica Lim. 2017. UE-HRI: a new dataset for the study of user engagement in spontaneous human-robot interactions. In *Proceedings of the 19th ACM international conference on multimodal interaction*. 464–472.

[5] Judith Bütepage, Ali Ghadirzadeh, Özge Öztimur Karadag, Mårten Björkman, and Danica Kragic. 2020. Imitating by Generating: Deep Generative Models for Imitation of Interactive Tasks. *Frontiers in Robotics and AI* (2020).

[6] Judith Bütepage, Hedvig Kjellström, and Danica Kragic. 2018. Anticipating many futures: Online human motion prediction and generation for human-robot interaction. In *IEEE ICRA*.

[7] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6299–6308.

[8] Oya Celiktutan, Efstratios Skordos, and Hatice Gunes. 2017. Multimodal human-human-robot interactions (mhhri) dataset for studying personality and engagement. *IEEE Transactions on Affective Computing* 10, 4 (2017), 484–497.

[9] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*. Springer, 104–120.

[10] Aurélie Clodic, Elisabeth Pacherie, Rachid Alami, and Raja Chatila. 2017. Key elements for human-robot joint action. *Sociality and normativity for robots: philosophical inquiries into human-robot interactions* (2017), 159–177.

[11] Shirine El Zaatari, Mohamed Marei, Weidong Li, and Zahid Usman. 2019. Cobot programming for collaborative industrial tasks: An overview. *Robotics and Autonomous Systems* 116 (2019), 162–180.

[12] Inc Fetch Robotics. 2023. Fetch Robotics. https://fetchrobotics.com/. [Online; accessed 29 September 2023].

[13] Michelangelo Fiore, Aurélie Clodic, and Rachid Alami. 2016. On planning and task achievement modalities for human-robot collaboration. In *Experimental Robotics: The 14th International Symposium on Experimental Robotics*. Springer, 293–306.

[14] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. 2013. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research* 32, 11 (2013), 1231–1237.

[15] Pupil Labs GmbH. 2023. Pupil Labs. https://pupil-labs.com/. [Online; accessed 29 September 2023].

[16] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. 2018. Social gan: Socially acceptable trajectories with generative adversarial networks. In *IEEE CVPR*.

[17] Sami Haddadin and Elizabeth Croft. 2016. Physical human–robot interaction. *Springer handbook of robotics* (2016), 1835–1874.

[18] Sami Haddadin, Alessandro De Luca, and Alin Albu-Schäffer. 2017. Robot collisions: A survey on detection, isolation, and identification. *IEEE Transactions on Robotics* 33, 6 (2017), 1292–1312.

[19] Guy Hoffman. 2019. Evaluating fluency in human–robot collaboration. *IEEE THMS* (2019).

[20] NaturalPoint Inc. 2023. OptiTrack. https://optitrack.com/. [Online; accessed 29 September 2023].

[21] Stereolabs Inc. 2023. Stereolabs. https://www.stereolabs.com/. [Online; accessed 29 September 2023].

[22] T. Iqbal, S. Li, C. Fourie, B. Hayes, and J. A. Shah. 2019. Fast Online Segmentation of Activities from Partial Trajectories. In *ICRA*.

[23] T. Iqbal, S. Rack, and L. D. Riek. 2016. Movement Coordination in Human-Robot Teams: A Dynamical Systems Approach. *IEEE T-RO* (2016).

[24] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. *CoRR* (2015).

[25] Quan Kong, Ziming Wu, Ziwei Deng, Martin Klinkigt, Bin Tong, and Tomokazu Murakami. 2019. MMAct: A Large-Scale Dataset for Cross Modal Human Action Understanding. In *Proceedings of the IEEE International Conference on Computer Vision*. 8658–8667.

[26] Philipp Kratzer, Simon Bihlmaier, Niteesh Balachandra Midlagajni, Rohit Prakash, Marc Toussaint, and Jim Mainprice. 2020. Mogaze: A dataset of full-body motions that includes workspace geometry and eye-gaze. *IEEE Robotics and Automation Letters* 6, 2 (2020), 367–373.

[27] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision* 123 (2017), 32–73.

[28] Thibault Kruse, Amit Kumar Pandey, Rachid Alami, and Alexandra Kirsch. 2013. Human-aware robot navigation: A survey. *Robotics and Autonomous Systems* 61, 12 (2013), 1726–1743.

[29] Arsenii Kuznetsov, Pavel Shvechikov, Alexander Grishin, and Dmitry Vetrov. 2020. Controlling overestimation bias with truncated mixture of continuous distributional quantile critics. In *International Conference on Machine Learning*. PMLR, 5556–5566.

[30] Przemyslaw A Lasota, Gregory F Rossano, and Julie A Shah. [n. d.]. Toward safe close-proximity human-robot interaction with standard industrial robots. In *2014 IEEE (CASE)*. 339–344.

[31] Séverin Lemaignan, Mathieu Warnier, E Akin Sisbot, Aurélie Clodic, and Rachid Alami. 2017. Artificial cognition for social human–robot interaction: An implementation. *Artificial Intelligence* 247 (2017), 45–69.

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 740–755.

[33] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*. 10012–10022.

[34] Ilya Loshchilov and Frank Hutter. 2017. SGDR: Stochastic Gradient Descent with Warm Restarts. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Skq89Scxx

[35] Kedi Lyu, Haipeng Chen, Zhenguang Liu, Beiqi Zhang, and Ruili Wang. 2022. 3D human motion prediction: A survey. *Neurocomputing* 489 (2022), 345–365.

[36] Jim Mainprice and Dmitry Berenson. 2013. Human-robot collaborative manipulation planning using early prediction of human motion. In *IROS*. IEEE.

[37] Julieta Martinez, Michael J Black, and Javier Romero. 2017. On human motion prediction using recurrent neural networks. In *IEEE CVPR*.

[38] Christoforos Mavrogiannis, Krishna Balasubramanian, Sriyash Poddar, Anush Gandra, and Siddhartha S Srinivasa. 2022. Winding Through: Crowd Navigation via Topological Invariance. *IEEE Robotics and Automation Letters* 8, 1 (2022), 121–128.

[39] Christoforos I Mavrogiannis, Valts Blukis, and Ross A Knepper. 2017. 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, 6817–6824.

[40] Christoforos I Mavrogiannis and Ross A Knepper. 2020. Decentralized multi-agent navigation planning with braids. In *Algorithmic foundations of robotics XII*. Springer, 880–895.

[41] Christoforos I Mavrogiannis, Wil B Thomason, and Ross A Knepper. 2018. Social momentum: A framework for legible navigation in dynamic multi-agent environments. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. 361–369.

[42] Benjamin A Newman, Reuben M Aronson, Siddhartha S Srinivasa, Kris Kitani, and Henny Admoni. 2022. HARMONIC: A multimodal dataset of assistive human–robot collaboration. *The International Journal of Robotics Research* 41, 1 (2022), 3–11.

[43] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. (2017).

[44] Calvin Z Qiao, Maram Sakr, Katharina Muelling, and Henny Admoni. 2021. Learning from demonstration for real-time user goal prediction and shared assistive control. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 3270–3275.

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[46] Andrey Rudenko, Tomasz P Kucner, Chittaranjan S Swaminathan, Ravi T Chadalavada, Kai O Arras, and Achim J Lilienthal. 2020. Thör: Human-robot navigation data collection and accurate motion trajectories dataset. *IEEE Robotics and Automation Letters* 5, 2 (2020), 676–682.

[47] Andrey Rudenko, Luigi Palmieri, Michael Herman, Kris M Kitani, Dariu M Gavrila, and Kai O Arras. 2020. Human motion trajectory prediction: A survey. *IJRR* (2020).

[48] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. 2020. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *European Conference on Computer Vision*. Springer, 683–700.

[49] Lindsay Sanneman, Christopher Fourie, Julie A Shah, et al. 2021. The state of industrial robotics: Emerging technologies, challenges, and key research directions. *Foundations and Trends® in Robotics* (2021).

[50] Reza Shadmehr and Ferdinando A Mussa-Ivaldi. 1994. Adaptive representation of dynamics during learning of a motor task. *Journal of neuroscience* 14, 5 (1994), 3208–3224.

[51] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. 2016. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *IEEE CVPR*.

[52] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems* 27 (2014).

[53] Yongyi Tang, Lin Ma, Wei Liu, and Wei-Shi Zheng. 2018. Long-term human motion prediction by modeling motion context and enhancing motion dynamic. In *IJCAI*.

[54] Leimin Tian, Kerry He, Shiyu Xu, Akansel Cosgun, and Dana Kulic. 2023. Crafting with a Robot Assistant: Use Social Cues to Inform Adaptive Handovers in Human-Robot Collaboration. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 252–260.

[55] Emanuel Todorov and Michael I Jordan. 2002. Optimal feedback control as a theory of motor coordination. *Nature neuroscience* 5, 11 (2002), 1226–1235.

[56] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, Vol. 2019. NIH Public Access, 6558.

[57] Vaibhav V Unhelkar, Shen Li, and Julie A Shah. 2020. Decision-making for bidirectional communication in sequential human-robot collaborative tasks. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 329–341.

[58] Vaibhav V Unhelkar and Julie A Shah. 2018. Learning models of sequential decision-making without complete state specification using bayesian nonparametric inference and active querying. (2018).

[59] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

[60] Zhikun Wang, Katharina Mülling, Marc Peter Deisenroth, Heni Ben Amor, David Vogt, Bernhard Schölkopf, and Jan Peters. 2013. Probabilistic movement modeling for intention inference in human–robot interaction. *IJRR* (2013).

[61] A Mark Williams, Paul Ward, John M Knowles, and Nicholas J Smeeton. 2002. Anticipation skill in a real-world task: measurement, training, and transfer in tennis. *Journal of Experimental Psychology: Applied* (2002).

[62] Mohammad Samin Yasar and Tariq Iqbal. 2021. Improving human motion prediction through continual learning. *ACM/IEEE Int. Conf. on Human-Robot Interaction (HRI), LEAP-HRI Workshop* (2021).

[63] Mohammad Samin Yasar and Tariq Iqbal. 2021. A Scalable Approach to Predict Multi-Agent Motion for Human-Robot Collaboration. In *IEEE RA-L*.

[64] Xinyan Yu, Marius Hoggenmueller, and Martin Tomitsch. 2023. Your Way Or My Way: Improving Human-Robot Co-Navigation Through Robot Intent and Pedestrian Prediction Visualisations. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 211–221.