International Journal of Social Robotics (2025) 17:3211–3230
https://doi.org/10.1007/s12369-025-01333-3

Robot Behavior Generation for Social Human-Robot Interaction

Esteve Valls Mascaro[1] · Dongheui Lee[1,2]

Received: 15 June 2024 / Revised: 12 July 2025 / Accepted: 2 October 2025 / Published online: 5 November 2025

Abstract
The increasing presence of robots in human workspaces underscores the need for intelligent systems that can understand human behaviors and act accordingly for a natural human-robot interaction (HRI). In this work, we propose a method to generate a robot's behavior for social HRI by integrating both human and robot intentions into the robot's decision-making process. Our system learns appropriate robot behaviors in social scenarios by observing human-human interactions (HHI). Using a transformer-based model, we first capture the dynamics of each individual and then iteratively adapt both human and robot behavior to achieve a successful interaction. By connecting our model with a human-to-robot motion retargeting framework, our system learns how a robot should behave solely from observing human data. To address the disparity between HHI and HRI, we employ several loss functions that encourage our robot to reproduce the social dynamics observed in humans. As a result, our approach outperforms the state-of-the-art in dyadic human motion forecasting prediction in the largest dataset available and obtains high-quality robot behaviors in human-robot interaction scenarios. Finally, we conduct a thorough evaluation of our performance for HHI, and HRI, and implement and test the system in the real-world TIAGo++ robot.

Keywords Human-robot interaction · Imitation learning · Motion forecasting · Deep learning

1 Introduction

In recent years, the coexistence of humans and robots within a shared workspace has become increasingly common, leading to an interest in human-robot interaction. As these entities share physical proximity, robots are compelled to integrate human actions into their decision-making processes. Traditionally, this has been addressed through the design of reactive robotic behaviors to assist humans in achieving a specific goal. However, when it comes to autonomous social interaction between robots and humans, mere prediction of human actions for robot decision-making falls short. Instead, it is desirable that robots infer and understand social norms, individual preferences, and the intentions of the surrounding humans to effectively engage in these interactions. Still, designing robot behaviors that accommodate all these diverse variables presents a significant hurdle. On the contrary, in this paper we propose to learn the social dynamics existing in human-human interactions and translate those learned behaviors into robots. An illustration of our robot's decision-making process is depicted in Fig. 1.

Understanding human behavior is a long-standing challenge in the AI and robotics community, involving the comprehension of complex, context-dependent actions and intentions. In the context of social interactions, the movements of individuals reflect their behavior and intentions. As humans, we predict the future movement and state of a human in the short-term future to optimize for fluent interaction [1]. For instance, when meeting a person, we extend the hand to perform a handshake but adapt our approaching behavior to the observed motion of the other individual, so that both hands meet. In the research community, the task of predicting future human poses based on past observations is known as human motion forecasting. While there has been significant progress in single human motion forecasting

✉ Esteve Valls Mascaro
esteve.valls.mascaro@tuwien.ac.at

Dongheui Lee
dongheui.lee@tuwien.ac.at

[1] Autonomous Systems Lab, Technische Universität Wien (TU Wien), Gusshausstraße 27, Vienna 1040, Austria

[2] Institute of Robotics and Mechatronics, German Aerospace Center (DLR), Münchener Strasse, 82234 Wessling, Germany

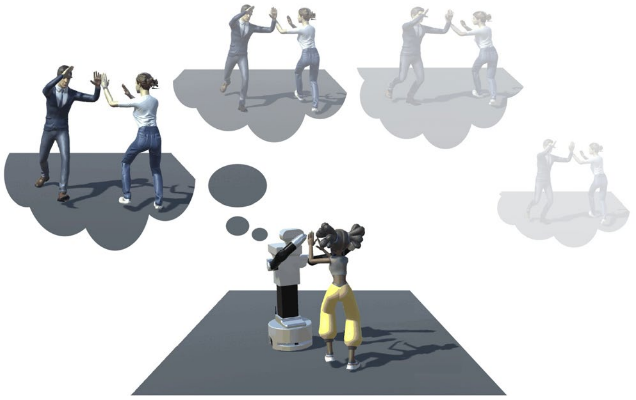


Fig. 1 Illustration of our robot's decision-making process. Our TIAGo++ robot learns to forecast the most natural movements in a human-robot interaction by observing dyadic human interactions. After this learning process, our model can generate high-quality robot behaviors that adapt to the partner's intent. It achieves this by envisioning how human and robot intents can seamlessly blend together through a deep-learning motion forecasting network

[2–9], they primarily focused on modeling the dynamics of a unique skeleton in local representations, without considering their global trajectory. Instead, learning the dependencies between the various individuals in human-human interacting scenarios remains a challenge. Multi-person forecasting [10–15] aims to model the spatial dependencies among the surrounding agents to predict their future movement. Existing studies on multi-person scenarios encode the relationship of multiple humans in scenarios with little or artificially synthesized interactions between the subjects [12–14], or with interactions that are not adequate for robots [15], such as fighting. Instead, we envision scenarios that are more representative of real-world Human-Robot Interactions (HRI) and model highly interactive scenes between humans that are executing a shared action [16], such as handovers, dancing, or greeting.

Even if we can model human dynamics in social settings, transferring this behavior to robots remains a challenge. Previous works [17–19] have focused on a reactive robot behavior, where the robot forecasts human motion and then responds with a set of predefined actions. More recently, [20] showed that by forecasting the surrounding human actions, the robot can anticipate its expected behavior and proactively assist the human, thus making the overall human-robot interaction more fluent. However, all the aforementioned works handcrafted the robot's response to the predicted human intent. In the context of social human-robot interactions, handcrafting the appropriate robot behavior that accounts for the very diverse human actions remains unfeasible. The optimal procedure would be to learn the adequate robot responses similar to multi-person interaction works [10–15], where one of the individuals is a robot. Recently, there have been many efforts in building human-robot interaction motion datasets for manipulation scenarios [21, 22]. However, these datasets are limited in terms of the variability of robot embodiments and action diversity. Additionally, [21, 22] only focused on human motion prediction conditioned on the robot actions but neglected the generation of the robot interaction behavior.

In fact, constructing large-scale human-robot interaction datasets is cumbersome, and the robot behaviors are usually controlled using teleoperation during the recording, limiting the scalability of the framework. Instead, in our previous work [23] we use a human-to-robot motion-retargeting algorithm [24] to unify the human and robot behaviors into a shared latent space. Thus, our ECHO system [23] could learn to generate human-robot interaction behaviors independently of the individual embodiments. Still, by constructing ECHO as a two-step framework and exclusively predicting the robot's behavior in a pre-trained shared latent space, the robot's decoded actions lacked spatial awareness. That is, in a human-robot handshake, the individual robot behavior resembled a proper handshake in local coordinates, but did not adapt to the partner's hand location. Therefore, the human and robot hands did not meet. In this work, we extend ECHO [23] by proposing a contact-aware robot behavior generation that learns in an end-to-end manner the spatio-temporal dependencies between humans and robots, thus encouraging the social dynamics to be met during robot execution. Additionally, to achieve a trade-off between faithfully imitating the style of the human references while preserving the semantics of the interactions, we incorporate a proximity sensor loss, which dynamically adapts where the model should focus during training. Our novel spatial dynamic loss boosts the quality of the generated human-robot interactions. We conduct a thorough quantitative and qualitative evaluation to ablate the benefits of our approach for HRI, and introduce new social metrics to measure the quality of an interaction. Finally, we implemented a real-time framework using ROS to generate such HRI behaviors in a TIAGo++ robot.

In conclusion, we propose an end-to-end learning model that generates human-robot behaviors from purely human-human interaction observation. We adopt the single and dual motion transformer from [23] that decouples the human and robot future movements in the early stages and learns to refine the interaction behavior by considering the overall global movements. However, instead of predicting a pre-trained shared representation among humans and robots, our framework generates robot joint angles to directly control the robot behavior. Later, we perform forward kinematics in the robot to obtain the end-effector position and encourage those to be closer to the reference human motion. Thanks to our adaptive proximity loss, our training encourages the model to achieve high-quality human imitation while following the physical contacts of the reference interactions. Given the evolutionary nature of this work, in this paper, we extend the results from [23] to the human-robot interaction behavior, introducing new qualitative and quantitative experiments for HRI scenarios, better techniques to boost the results in HRI and novel social metrics to evaluate those, and we implement our framework in the real world using the TIAGo++ robot. The contributions of our paper can be summarized as follows:

- An end-to-end deep learning framework to generate robot behaviors in social settings that are aware of the spatio-temporal dependencies in human-human interactions.
- An efficient model that achieves state-of-the-art performance in real-time for social human forecasting and in human-robot collaborative scenarios.
- A novel proximity-aware dynamic loss that weighs the importance of different measurements during training, to achieve the right balance between individual behavior imitation and social interactions.
- The implementation of the system in a real-world robot to generate fluent social behaviors with humans.

2 Related Work

This section is organized as follows. First, we review the literature on human behavior modeling from the motion perspective. Then we focus on the translation of those behaviors to robots using imitation learning. Finally, we introduce various works on Human-Robot Interaction that consider human behavior modeling in their robot behavior response.

2.1 Human Intent in HRI

For robots to interact alongside humans to achieve a shared goal, they need to understand the human partner's intent and incorporate it into their decision-making process, so that both entities are coordinated. Losey et al. [25] identified three key terms essential for human intention understanding in physical HRI: intent information, which refers to how intent is defined; intent measurement, as the modality of the data; and intent interpretation, which involves how to incorporate this data into the robot control.

The intent information and measurements have been defined differently according to the task the intelligent system is performing. Some examples include human trajectory prediction in autonomous driving [26], gaze following to convey human attention to objects [20, 27, 28], action classification for predicting future action plans [29–31], and 3D skeleton movement for understanding human behavior [22, 32–34], synchronizing movements [35], or determining when the robot should provide mutual support [36]. In this work, we focus on predicting the 3D human skeleton during HRI to ensure the robot's behavior is coordinated with the human partner. Unlike previous works that incorporate human intention without considering the robot as part of the interaction, such as [32–34], we involve both humans and robots in the decision-making process. Our system iteratively refines the human's expected behavior based on the robot's intent and vice versa, enhancing the understanding of social dynamics and making the robot's behavior proactive. This results in more fluent robot actions, as they are conditioned on the expected human states and do not need to wait for a human movement to finish before responding.

2.2 Human Behavior Modeling

To achieve robots that coexist with humans within a shared workspace, we first need intelligent systems that can recognize, interpret, and reason about the behavior of the surrounding humans. Human behavior understanding encompasses various aspects, including the prediction of human movements [2, 6, 8, 15, 23, 37, 38], their interactions with the objects of the scene [20, 28], their gaze direction [39–41] and the actions they perform [30, 31, 42]. Given our focus on understanding human behavior for social interactions, this section will specifically review works related to human movement prediction, both in single- and multiple-person scenarios.

The field of human motion forecasting has primarily concentrated on modeling the spatio-temporal patterns inherent in human joints to predict future 3D skeleton information. Sequence-based neural networks, including Recurrent Neural Networks (RNNs) [2, 43], Discrete Cosine Transform (DCT) with Graph Convolutional Networks (GCN) [4, 5], and more recently, attention-based models [6–8] have been extensively employed for this purpose. However, all aforementioned works only consider the spatial dependencies among the different joints of an individual body, overlooking the interactions between individuals involved in a social activity.

In the context of multi-person motion forecasting, it becomes imperative to incorporate global coordinates of individuals and the relationships between them. Initial works [37, 44] focused on predicting the global trajectory of humans in a scene. However, for scenarios involving human-human and human-robot interaction, it is essential to extend the problem to encompass the 3D representation of all joints of a human skeleton. Recent studies have explored various techniques to address these challenges [38]. leveraged context information from images to condition the motion generation [14], decoupled individual and multiple human features using transformers to enhance long-term prediction for groups of people, and [15] focused on modeling dyadic interactions of humans, enhancing the motion forecasting based on others through cross-attention mechanisms. To explicitly capture interactions among joints within the same individual and with others [45], operated on each joint with self-attention, and [10] partitioned the body into parts and operated on the flattened sequence through self-attention. While these strategies facilitate better capturing of spatial relationships between joints within individuals, they increase the complexity of transformers in capturing inter-human dependencies. Recently [11], proposed to reuse DCT and GCN [46] in an autoregressive manner for dyadic interactions. However, such approaches utilizing DCT may produce overly smooth synthesized motions that fail to capture subtle nuances within motions. Moreover, predicting the entire future sequence in one step, as demonstrated in our work, avoids the accumulation of errors over iterations, associated with autoregressive approaches [11, 15], preventing potential collapse in the long term.

Motivated by the recent success of denoising diffusion probabilistic models (DDPMs) [47] for human motion generation [48, 49], several studies have focused on applying DDPMs to multi-person motion generation in social dynamics using diffusion-based models [16], synthesizing the reactive behavior of a human given that of their counterpart [50, 51]. Despite the high motion diversity and fidelity achieved with DDPMs, they often deviate too far from the ground truth compared to deterministic models [48] or become unrealistic within a historical context. Additionally, DDPMs are computationally intensive, requiring significantly more resources and time for inferring a single motion sequence. Due to these limitations, we do not include DDPMs in our comparison, as they are not suitable for real-time robot behavior generation in social human-robot interaction, which is the goal of our work.

2.3 Imitation Learning

The release of large-scale dyadic human motion interaction datasets involving intense contact interactions [15], dancing [52] or very diverse social actions [16] have motivated the extensive research on modeling these social dynamics [11, 14, 15, 38, 45, 50, 51]. However, the available human-robot interaction datasets are limited and focus on manipulation scenarios [21, 22]. To overcome this issue, we propose to translate human-human interactions to human-robot interactions by making use of motion retargeting approaches. This task aims to translate a motion from humans to robots while maintaining a high visual resemblance.

Motion retargeting has been a long-standing challenge in the animation and robotics community, driven by the need to achieve natural human-like movements across different embodiments. Early efforts approached retargeting as an optimization problem, solving inverse kinematics (IK) with specific space-time constraints [53–57]. These methods aimed to preserve end-effector or intermediate joint positions but often struggled to generalize to complex human motions. To address these limitations, learning-based methods reframed human-to-robot motion retargeting as a domain adaptation problem, emphasizing the preservation of visual fidelity between source and target motions. While motion retargeting has been extensively explored to translate human motions to animated characters [58–62], this paper focuses specifically on the retargeting to robotic embodiments.

Learning-based human-to-robot motion retargeting aims to preserve visual resemblance during the imitation process while enabling effective control of real robots. Unlike animation-focused methods that rely on Cartesian space or rotation representations between body limbs [58–62], the goal here is to generate control commands that accurately imitate a human movement. Given the difficulty and time-consuming process of manually building a dataset of human and robot pairs, [63] proposed an automatic pipeline to synthesize human-robot pairs offline. This method involved retrieving the closest synthesized pose from a pre-generated dataset based on a given human pose. However, [63] adopted a nearest neighbor retrieval algorithm that struggled to generate smooth motions.

To address this, [24] proposed to learn a shared representation space between humans and robots, allowing for direct decoding of robot commands from this latent space. In this paper, we use ImitationNet [24] to construct a dataset of noisy human-robot interactions. These interactions are considered noisy, because [24] only ensures visual resemblance in local coordinates with respect to the source human, failing to maintain the spatial dependencies between the generated robot pose and the human that is part of the dyadic interaction. For instance, in a handshaking scenario, the retargeted robot end-effector may not align closely with the reference human hand if the robot's height is lower than that of the human being imitated.

To this end, our work not only teaches the future robot controls to interact with a human in a social setting but also adapts and refines the robot's expected behavior to be as close as possible to the reference human-human interaction during close situations. To address this, we propose proximity-aware spatial losses that consider the similarity between the source human motion and the generated robot behavior, as well as maintaining the spatial dependencies with the counterpart human in the interaction. Our dynamic losses decide which training objective to pursue at each time-step depending on the semantics of the interaction.

3 Methodology

This section presents the task of human-human and human-robot interaction behavior generation based on motion forecasting. First, we formally describe the social forecasting task. Later, we describe the different parts of our proposed architecture, which is illustrated in Fig. 2. Finally, we introduce the differences between human-human and human-robot interaction behavior generation and propose various techniques applied to improve the overall behavior.

3.1 Problem Formulation

Let $\mathbf{S}$ be a dyadic interaction between two agents described by a textual description D. Depending on the embodiment of the agents, our social scenario can be classified as human-human ($\boldsymbol{S}_{HH}$) or human-robot ($\boldsymbol{S}_{HR}$) interaction. We define an interaction $\mathbf{X}^i$ where $i = \{H, R\}$ as a motion sequence composed of T states, such that $\mathbf{X}^{\mathbf{i}} = [\mathbf{x}_0^i, \cdots, \mathbf{x}_T^i]$. We describe an agent state in reference to a map reference system in terms of global coordinate trajectory and local agent

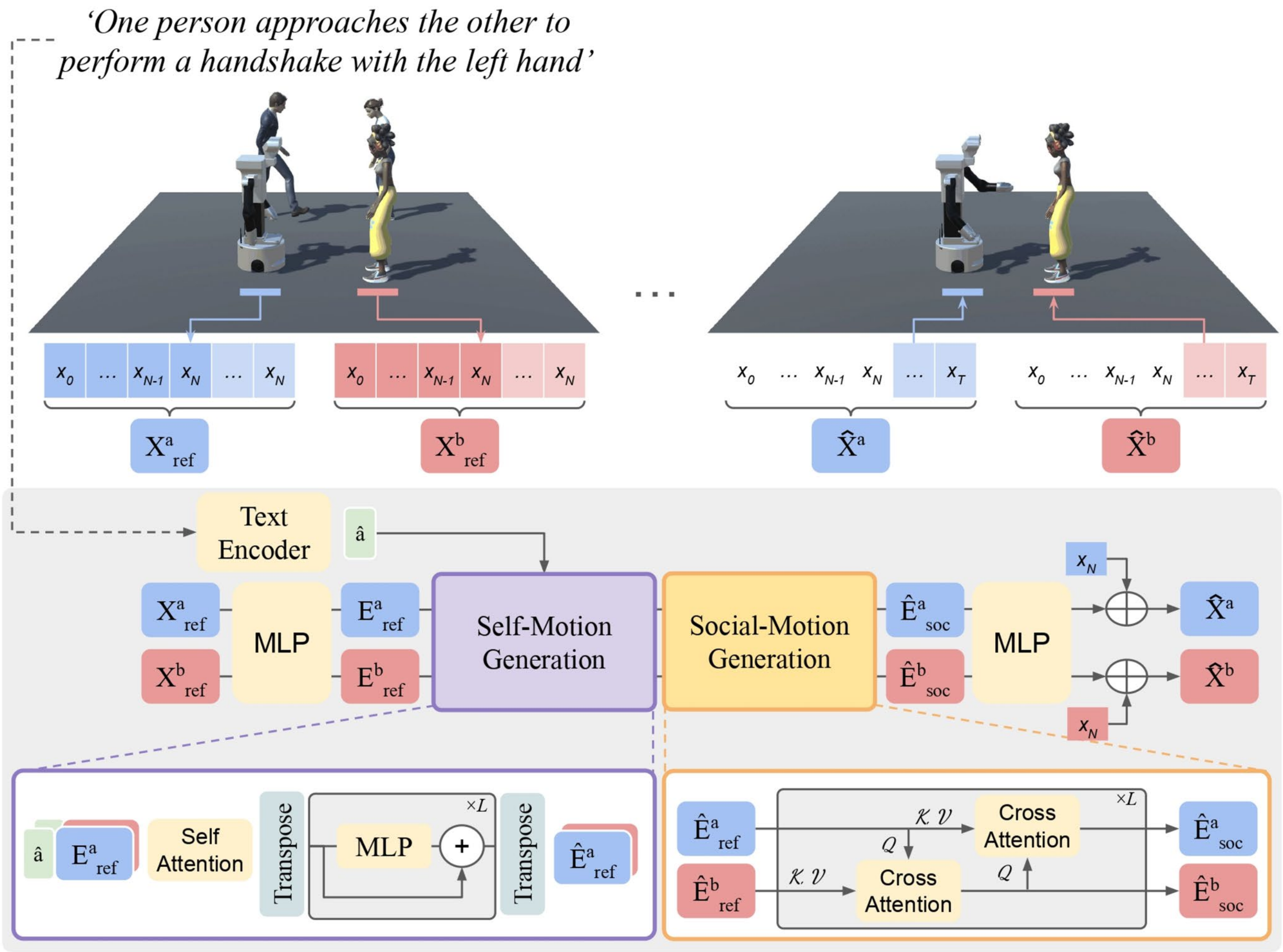


Fig. 2 Overview of our human-robot interaction (HRI) behavior generation. During training, a given human-human interaction is retargeted to a human and robot interaction using ImitationNet [24]. The first step is to construct the appropriate motion sequences $\boldsymbol{X}^a_{ref}$ and $\boldsymbol{X}^b_{ref}$ for the agents a and b that participate in the interaction. $\boldsymbol{X}^i_{ref}$ contains the observed last N motion states for the agent i padded with the repetition of the last pose observed x^i_N, such that $\boldsymbol{X}^i_{ref} = [\boldsymbol{x}^i_0, \cdots \boldsymbol{x}^i_N, \cdots, \boldsymbol{x}^i_N]$, with a total sequence length T. Then, we encode both reference motions using a Multi-Layer Perceptron (MLP), so that $\boldsymbol{E}^i_{ref} \in \mathbb{R}^{T \times D}$. We also encode the textual social description with [49] and obtain $\hat{a}$. We prepend $\hat{a}$ to $\boldsymbol{E}^i_{ref}$ and forward each individual motion to the self-motion generation module, which provides a future motion reference $\hat{\boldsymbol{E}}^i_{ref}$ through a self-attention transformer and a sequence of L temporal MLP layers. To align both single motion references to the other partner in the interaction, we forward $\hat{\boldsymbol{E}}^a_{ref}$ and $\hat{\boldsymbol{E}}^i_{ref} = b$ to the social-motion generation module. There, we iteratively refine the motions from agent a based on b, and vice versa, obtaining $\hat{\boldsymbol{E}}^a_{soc}$ and $\hat{\boldsymbol{E}}^b_{soc}$. Finally, we decode each $\hat{\boldsymbol{E}}^i_{soc}$ and sum it to the last observed motion state $\boldsymbol{x}^i_N$ to simplify the training objective

pose. We consider the global trajectory $\boldsymbol{g}_t^i \in \mathbb{R}^7$ for an agent $i = \{H, R\}$ at time *t* as the *xyz* translation $\boldsymbol{v}_t^i \in \mathbb{R}^3$ and quaternion rotation $\boldsymbol{\Omega}_t^i \in \mathbb{R}^4$ from the reference coordinate system. In addition, we define a human local pose $\boldsymbol{h} \in \mathbb{R}^{J_h \times n}$ with J_h joints, using quaternions (n=4) or *xyz* (n=3) for joint representation. Similarly, a robot pose $\boldsymbol{r} \in \mathbb{R}^{J_r \times s}$ has J_r joints, represented by joint angles (s=1).

The task of social behavior generation is defined as the prediction of the future motion $\boldsymbol{X}_{fut}^i = [\boldsymbol{x}_{N+1}^i, \cdots, \boldsymbol{x}_T^i]$ per all entities in the scenario (either $i = \{H, H\}$ or $i = \{H, R\}$) given their past observations $\boldsymbol{X}_{past}^i = [\boldsymbol{x}_0^i, \cdots, \boldsymbol{x}_N^i]$, where N represents the number of observed motion states in the past. In this paper, we reformulate the forecasting objective so that our network f_θ only learns the displacement of the future states with respect to the last state observed $\mathbf{x}_N$, such that $\boldsymbol{X}_{fut} = f_\theta(\boldsymbol{X}_{past})) + \boldsymbol{X}_N$. This same strategy has been shown to be effective in prior works [7, 8, 64].

3.2 Human and Robot Behavior Generation

Modeling the different and diverse behaviors encountered in a typical social scenario requires understanding the spatio-temporal dependencies of the agent's participant in the interaction. Motivated by the high-quality performance of attention-based models [65] in human motion [6–8, 10, 15, 45], we adopt transformers as the core of our architecture. Instead of considering a single-token autoregressive approach that predicts the next state of each agent in the scene, we aim to forecast the whole interaction at once. Therefore, we pad our observed interactions by repeating the last state observed $T{-}N$ times, so the input sequence has length *T*. We refer to the padded interactions as $\boldsymbol{X}_{ref}^i = [\boldsymbol{x}_0^i, \cdots, \boldsymbol{x}_N^i \cdots, \boldsymbol{x}_N^i]$. Intuitively, $\boldsymbol{X}_{ref}^i$ represents that the agents do not change their state in the future. Our transformer-based architecture learns how to change those reference motion behaviors to achieve a social interaction.

To simultaneously learn the dependencies between the individual local joint coordinates and global trajectory (i.e., if the evolution of the agent poses represents a walking forward behavior, the global trajectory should adapt accordingly), we encode both representations together. That is, we flatten the rotation-based local skeleton, both for human $\boldsymbol{h}_t^i$ or robot $\boldsymbol{r}_t^i$, and concatenate the global trajectory $\boldsymbol{g}_t^i$. We embed this information using a multi-layer perceptron (MLP) per each time-step and agent in the interaction, obtaining $\boldsymbol{E}_{ref}^i \in \mathbb{R}^{T \times D}$, which is a higher-level representation of the human or robot observed behavior.

When aiming to anticipate the expected behavior in a social interaction, we consider that three important factors should be taken into account. First, the expected social interaction description: depending on location, social norm, relationship, etc., two humans might perform a different joint behavior. Second, the evolution of the current status: we cannot abruptly change our global trajectory or the current pose, but we transition smoothly to achieve a certain behavior. In fact, when performing a handshake, our intention drives the arm movement to be extended following the dynamics of the ongoing motion. We name this second term as the 'self intent', which drives one behavior without considering the other person. However, in a social scenario, we refine the end position of the hand movement to match the other participant's hand, so we can perform a proper handshake. This third factor, which we describe as 'social intent', refines the movement from the 'self intent' to accomplish the social interaction, taking into account not only the individual dynamics but also the behavior of the partner. Our goal is to translate those social dynamics into the designed architecture for a higher-quality behavior generation.

First, we encode the social description a using the encoder of a text-to-motion pre-trained model [66], such that $\hat{a} \in \mathbb{R}^D$. Second, we define a self-motion generation module that operates on each single behavior without considering the partner in the interaction. Our self-motion generation follows a traditional strategy of self-attention architecture [65], where we first add a sinusoidal positional embedding to $\boldsymbol{E}_{ref}^i$, append the social description token $\hat{a}$ to the encoded observed behavior such that $\bar{\boldsymbol{E}}_{ref}^i = [\hat{a}, \boldsymbol{E}_{ref}^i] \in \mathbb{R}^{(T+1) \times D}$, and forward $\bar{\boldsymbol{E}}_{ref}^i$ to a self-attention transformer. Then, similar to [8, 9] for single-motion forecasting tasks, we iteratively smooth the output of the self-motion transformer by expanding and compressing the time dimensionality of the output through L temporal MLP layers, obtaining $\hat{\boldsymbol{E}}_{ref}^i \in \mathbb{R}^{\times (T+1) \times D}$.

Until here, we have only modeled each individual's behavior as an independent entity without considering the partner's behavior. Motivated by the aforementioned 'social intent', we propose to refine $\hat{\boldsymbol{E}}_{ref}^a$ conditioned on $\hat{\boldsymbol{E}}_{ref}^b$. Note that *a* and *b* are the two agents in the interaction, which can be depicted as two humans $H1$ and $H2$ or human H and robot R depending on the scenario. Our social-motion generation module uses a series of two cross-attention layers to refine one subject motion based on the other. Our cross-attention mechanism learns how to blend an input **Q**uery (**Q**) based on a conditioning **K**ey (**K**) and **V**alue (**V**). First, we use $\hat{\boldsymbol{E}}_{ref}^a$ as **Q** and $\hat{\boldsymbol{E}}_{ref}^b$ as **K** and **V** for the first cross-attention. The goal is that the resulting motion of the subject a ($\hat{\boldsymbol{E}}_{soc}^a$) has been refined to be compliant with subject b. This step is now repeated in the inverse order, being $\hat{\boldsymbol{E}}_{ref}^b$ as

$\mathbf{Q}$ and $\hat{\boldsymbol{E}}^{a}_{soc}$ as $\mathbf{K}$ and $\mathbf{V}$. This dual cross-attention strategy is repeated *k* times to iteratively enhance each agent's behavior to be in synchrony with the other agent during the social interaction. For an illustrated description of our self- and social-motion generation module refer to Fig. 2.

Finally, given $\hat{\boldsymbol{E}}^{a}_{soc}$ and $\hat{\boldsymbol{E}}^{b}_{soc}$, that represent the behavior of each agent in the interaction, we use an MLP layer to infer each motion behavior at each time-horizon in the future, such that $\hat{\boldsymbol{h}}^{i}_{t}$ or robot $\hat{\boldsymbol{r}}^{i}_{t}$ represent the local pose of an agent and $\hat{\boldsymbol{g}}^{i}_{t}$ indicates the predicted global information.

3.3 Losses

Our task is to optimize the parameters of a deep learning neural network to learn the behaviors of humans in human-human interactions, as well as the robot's behavior when participating in human-robot interactions. For that, we consider the weighted sum of different losses based on the task at hand, which aims to ensure that the generated behaviors are natural and follow the dynamism in typical social interactions.

3.3.1 Movement Losses

Following the aforementioned factors that we considered relevant to encounter in social interaction, we define two losses that encourage the model to learn the 'self intent' and 'social intent' expected from social behavior. First, we enforce the output of the self-motion generation module to generate dynamic movements that approximate the evolution of the individual. Then, we guide the refinement of the motions produced in the social-motion generation module to the ground-truth motion state information for each individual. We denote these two losses as self-loss $\mathcal{L}_{self}$ and social loss $\mathcal{L}_{soc}$, which take the form of a Mean Square Error (MSE) function. Note that D_H represents the decoder used to project the learned motion representations, either from the self- ($\hat{\boldsymbol{E}}^{i}_{ref}$) or social-module ($\hat{\boldsymbol{E}}^{i}_{soc}$), to the respective local and global agent information $\boldsymbol{X}^i$.

$$\mathcal{L}_{self}(\boldsymbol{X}^i) = MSE(D_H(\hat{\boldsymbol{E}}^{i}_{ref}) - \boldsymbol{X}^i) \tag{1}$$

$$\mathcal{L}_{soc}(\boldsymbol{X}^i) = MSE(D_H(\hat{\boldsymbol{E}}^{i}_{soc}) - \boldsymbol{X}^i) \tag{2}$$

3.3.2 Interaction Losses

$\mathcal{L}_{self}$ and $\mathcal{L}_{soc}$ enforce each individual of the interaction to follow their original behavior. However, for a natural social interaction, if one of the individual's behavior shifts, the partner should adapt their movement accordingly. This spatial synchrony is very explicit in physical interactions, such as dancing together, handing over objects, or greeting with a handshake, where any subtle change of movement of one individual affects and should be considered by the other to carry on with the interaction. To enforce learning this behavior, we propose an interaction loss $\mathcal{L}_{inter}$ that minimizes the distance between some correspondent joints from agent *a* to *b* in the interaction, referred to as Distance Matrix (DM). Note that $\mathcal{L}_{inter}$ is applied to the global Cartesian position of the agents' joints with respect to a reference map frame.

In the case of human-human interaction, the human motion is already represented as *xyz* for the local joint position and root coordinates. Thus, we can easily obtain the global information. Moreover, given that both agents participating in the interaction are human, we consider all joints in the calculation of the Distance Matrix. However, for human-robot interaction, we first need to compute forward kinematics to the robot's joint angles $\mathbf{r}$ to obtain the *xyz* positions of each robot joint with respect to the robot's base, $FK(\boldsymbol{r})$. We later compute the transformation matrices from the robot's base to the global map coordinates using the root position $\mathbf{v}^R$ and orientation $\boldsymbol{\Omega}^R$ and obtain the joint Cartesian position in the global coordinate system. Given that humans and robots only share certain joints, we consider only the human hands and the robot's end-effectors for the DM computation. The proposed interaction loss $\mathcal{L}_{inter}$ is shown in Eq. 3. For simplification, we denote as $\mathbf{Y}^i$ the *xyz* global position of all J_i joints of an agent *i*, either human or robot, and $\mathbf{v}^i$ as its root position. Similarly, $\hat{\boldsymbol{Y}}^{i}$, and $\hat{\boldsymbol{v}}^i$ are the predicted global coordinates and root positions.

$$\mathcal{L}_{int}(\boldsymbol{Y}^a, \boldsymbol{Y}^b) = MSE(\mathrm{DM}(\hat{\boldsymbol{Y}}^{a}, \hat{\boldsymbol{Y}}^{b}) - \mathrm{DM}(\boldsymbol{Y}^a, \boldsymbol{Y}^b)) \tag{3}$$

3.3.3 Embodiment Losses

Up to this point, the proposed losses have optimized our model to generate a natural behavior ($\mathcal{L}_{self}$ and $\mathcal{L}_{soc}$) that synchronizes and adapts to the other participant in the interaction. However, there is a need to enforce that the generated embodied poses are feasible and adequate. For that, in the case of human-human interactions, we adopt a bone loss $\mathcal{L}_{bone}$ to reinforce the predicted body joints in *XYZ*-euclidean representations to have consistent bone lengths. We use the kinematic structure of the human to minimize the MSE between the bone lengths of the ground-truth human skeleton and the predicted ones.

As we adopt joint angle representation to describe the robot's joints, we do not encounter that issue in robots. However, while the human motions are obtained from high-quality motion capture systems in real-world interactions, the robot partner behavior used for training is generated by a pre-trained imitation model, named ImitationNet [24]. ImitationNet only focuses on retargeting the style of a human pose to a robot pose but does not consider external constraints in the learning process, such as ensuring that the hand is in a similar global position as the reference human. For example, when considering a reference human pose in global coordinates $\mathbf{Y}^H$ who is extending the arm to perform a handshake, ImitationNet will generate a robot pose $\mathbf{Y}^R$ also with the extended arm. However, given that the height of the TIAGo++ robot is less than the height of a human, the robot hand is in a much lower position. In an ideal situation, we would like the generated behavior to focus less on the local configuration of the robot at the times of close proximity (e.g., hugs, contacts) and more on the relative distance between the human and the robot's limbs. While $\mathcal{L}_{inter}$ enforces this spatial synchrony between the two individuals in the interaction, we consider it necessary to encourage the predicted robot pose to be as close as possible to the reference human pose used in the retargeting within the local coordinate system. Therefore, we propose an imitation loss $\mathcal{L}_{imit}$ that minimizes the distance between the human hands of the reference human motion $\tilde{\boldsymbol{Y}}^H_{ee}$ and the end-effector position obtained from the predicted robot motion $\hat{\boldsymbol{Y}}^R_{ee}$.

$$\mathcal{L}_{imit} = MSE(\tilde{\boldsymbol{X}}^H_{ee} - \hat{\boldsymbol{X}}^R_{ee}) \tag{4}$$

3.3.4 Adaptive Proximity Losses for HRI

Our previously proposed multiple losses have aimed to enhance the model performance according to the different nature of the dataset, either HHI (clean, collected using precise MoCap systems) or HRI (noisy, obtained by retargeting one individual to a robot using ImitationNet [24]). During human-human interactions, the optimization objectives of the neural network is to minimize a total loss compounded by:

$$\begin{aligned}\mathcal{L}_{hhi} = &\lambda_{self} * \mathcal{L}_{self} + \lambda_{soc} * \mathcal{L}_{soc} \\ &+ \lambda_{int} * \mathcal{L}_{int} + \lambda_{bone} * \mathcal{L}_{bone}\end{aligned} \tag{5}$$

Given that all losses operate over the same clean HHI data, the overall $\mathcal{L}_{hhi}$ simply focuses on reconstructing the masked future behaviors by aggregating penalties from different point of view (individual or social conditions, relative distances between the subjects or bone changes).

However, this process remains challenging for HRI, as the movement losses ($\mathcal{L}_{self}$ and $\mathcal{L}_{soc}$) prioritize generating behaviors that closely match ImitationNet's robot reference, while the interaction ($\mathcal{L}_{int}$) and imitation losses ($\mathcal{L}_{imit}$) prioritize preserving the end-effector trajectories from the reference human motion. In the scenario where all aforementioned HRI losses are compounded together, the overall training objective only aligns when the end-effector of the human reference matches the end-effector of the retargeted robot using ImitationNet (i.e., when $\tilde{\boldsymbol{X}}^H_{ee}$ equals $\tilde{\boldsymbol{X}}^R_{ee}$), which is usually not the case if the targeted robot has a different shape than the original human (e.g., lower height, longer arms). Overemphasizing the weight of $\mathcal{L}_{int}$ and $\mathcal{L}_{imit}$ would lead to unnatural robot behaviors that constantly aim to follow the global trajectory of the human hand. For instance, due it the shorter height of the TIAGo++ robot, it has to raise its arms to match the human's during walking, as depicted in the top-left snapshot of Fig. 3.

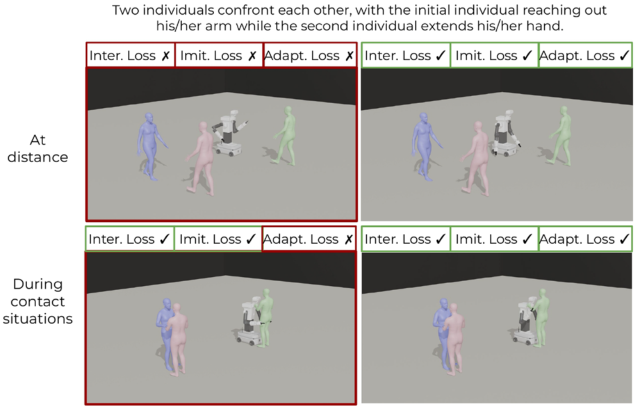

Two individuals confront each other, with the initial individual reaching out his/her arm while the second individual extends his/her hand.
Inter. Loss ✗
Imit. Loss ✗
Adapt. Loss ✗
Inter. Loss ✓
Imit. Loss ✓
Adapt. Loss ✓
At distance
Inter. Loss ✓
Imit. Loss ✓
Adapt. Loss ✗
Inter. Loss ✓
Imit. Loss ✓
Adapt. Loss ✓
During contact situations

Fig. 3 Qualitative effect of different losses over the generated behavior of the TIAGo++ robot. When human and robot are at a certain distance, the robot should focus on following the reference human motion (in blue), encouraging the ImitationNet reconstruction loss. On the contrary, during close situations, the robot should focus on imitating the reference trajectory, ensuring that the robot arms are in contact to the partner's elbows. We achieve this trade-off through the use of a adaptive loss

To resolve the aforementioned misalignment in training objectives, we introduce a novel loss function that adaptively adjusts the importance of the individual and interactive losses at each time-step of the interaction. Specifically, when subjects are at a certain distance, we prioritize following the ImitationNet reference; when they are close, we emphasize following the reference trajectory. Rather than defining thresholds based on the relative distance between the subjects' base, we introduce a new method to detect physical contact or close proximity. This method works as follows:

1. **Detecting physical contact in human-human interaction.** We model the human body with simple collision meshes: limbs and torso as cylinders, and hands as spheres. We manually set the radius for each mesh (as shown in Fig. 4). When the meshes of two subjects collide, we assign a binary value indicating contact. Additionally, we manually define a proximity margin,

m, to detect when the meshes are close, defining proximity sensors $S_m(H^A, H^B)$, where $S_m(H^A_{ref}, H^B_{ref})$ returns 1 if the distance between any pair of human A to human B is less than m and 0 otherwise. For example, in a handshake, when $m=0$, the sensors indicate contact, and when $m=0.05$, they indicate proximity within 5 cm.

2. **Detect physical contact in the human-robot interaction.** Similarly, we define collision meshes for the robot's body parts (arms, torso, and end-effector) and apply the same method as for HHI to detect collisions in HRI.
3. **Design an adaptive loss**.The proximity information from HHIs is used to adjust the weighting of interaction losses. The weight factor, w_{pr}, is calculated as: $w_{pr} = 0.5 * S_{m=0}(H^A_{ref}, H^B_{ref}) + 0.5 * S_{m=0.05}(H^A_{ref}, H^B_{ref})$. On that regard, our final loss is computed as shown in Eq. 6, where the interaction and imitation losses are only encouraged during close proximity (i.e., $w_{pr} = 1$ during contact and $w_{pr} = 0.5$ when very close). Notice that the values of $m=0.0$ and $m=0.05$ have been manually predefined to account for contact situations and very close proximity levels.

$$\begin{aligned}\mathcal{L}_{hri} =& w_{pr} * (\lambda_{int} * \mathcal{L}_{int} + \lambda_{imit} * \mathcal{L}_{imit}) + \\ & (1 - w_{pr}) * (\lambda_{self} * \mathcal{L}_{self} + \lambda_{soc} * \mathcal{L}_{soc})\end{aligned} \tag{6}$$

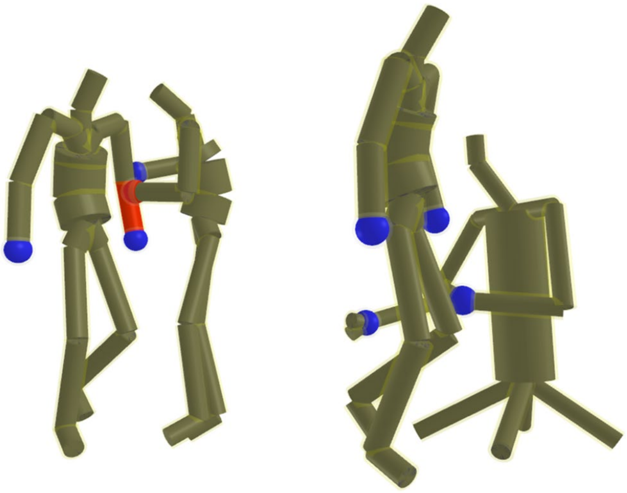


Fig. 4 Illustration of a physical contact in a human-human interaction (left) alongside the reference human-robot interaction (right) from ImitationNet. The original behavior depicts a human who has helped another person to stand-up by providing balance by grabbing the partner's arm. However, ImitationNet cannot preserve the contact between the left end-effector of the robot and the human left arm, which emphasizes the need of refining the reference robot motion during close situations between both agents

4 Experiments

4.1 Datasets

4.1.1 InterGen Dataset

InterGen [16] is the largest 3D human-human motion dataset, featuring 6022 dyadic interactions described by 16,756 natural language annotations. The dataset includes physical interactions recorded by professionals (e.g., dancing, boxing) and everyday social activities (e.g., handover, greeting, communication). Although we are more interested in the latter interaction types for learning social HRI behaviors for robots, we train our models on the full dataset to increase the robustness of our model to diverse interactions and offer a fair benchmark for future works that only focus on HHI. We define the forecasting task as predicting the motion of each individual entity in the next 1.5 seconds given an observation of 0.5 seconds. A human is described by 22 joints comprising both legs, arms, torso, and head.

4.1.2 Human-Robot Interaction Dataset

Due to the lack of a proper social dataset for human-robot interaction, we make use of a pre-trained human-to-robot motion retargeting model [24] to transfer the human motions to a TIAGo++ humanoid robot. ImitationNet [24] converts a human local pose represented as quaternions to a robot pose in joint angles. Still, ImitationNet only generates the robot joint angles and cannot predict the prismatic trunk height, the robot's head orientation, or transfer the root position. For those, we decide to copy the human's head orientation to the robot as well as the root trajectory. Regarding the prismatic trunk, we set the default trunk height to 1 meter. Note that in our previous approach [23] we only control the robot arms. Overall, we use the inverse kinematics from [67] to extract the local pose of the human reference as well as the root orientation (in quaternions) and translation (in Cartesian space). ImitationNet is used to convert those local human poses to robot joint angles, which can then be transformed into the robot's Cartesian position through Forward Kinematics and the corresponding transformation matrices obtained from the original human. Finally, we define the robot behavior generation task as predicting the robot's motion in the next 1.0 seconds given an observation of 0.5 seconds.

4.1.3 CHICO Dataset

Although transferring social behaviors is the goal of this work, we also evaluated our framework in a Human-Robot Collaboration (HRC) scenario involving a shared manipulation task. CHICO [21] contains a single operator in a smart factory environment performing seven assembly tasks together with a Kuka LBR robot. The 3D motions of both the human and the robot are recorded. In this case, the task is to predict the operator's motion intent (e.g., object pick and place, surface polishing, hammering, or object lifting) in the HRC scenario, while also considering the robot's behavior. We follow the standard evaluation and predict the next 1000 ms given 400 ms of past observations.

4.2 Metrics

Given the various nature of our evaluation, we consider different metrics per each task. For the human-human and human-robot behavior generation tasks, we follow the standard metrics in multi-person motion forecasting [10, 12]. Given that the standard evaluations focus on *xyz* positions of the human joints, we convert human and robot motions to the global coordinate system in the Cartesian space, as indicated in Section 3.3.2.

4.2.1 Metrics in HHI

JPE. We compute the Joint Position Error (JPE) to measure the error (in millimeters) of each joint position in a given future time step with respect to the map coordinate frame. Note that our evaluation for human-human interaction considers both humans as part of the equation (*subj*=2), but in human-robot interaction, we only evaluate the robot's behavior (*subj*=1).

$$\mathrm{JPE}(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = \frac{1}{subj}\sum_{i=1}^{subj}\frac{1}{J_i}\sum_{j=1}^{J}||X_j^i - \hat{X}_j^i||^2, \qquad (7)$$

AJPE. Aligned JPE (AJPE) only considers the local position error with respect to the root position (i.e., the pelvis of the human or the base footprint of our TIAGo++ robot), allowing us to evaluate non-physical interactions such as waving or communicating, where precise prediction of the root position is less important.

$$\mathrm{AJPE}(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = \mathrm{JPE}(\boldsymbol{Y} - v, \hat{\boldsymbol{Y}} - \hat{v}). \qquad (8)$$

FDE. While AJPE and JPE focus on the agent joint information, we compute the Final Displacement Error (FDE) to evaluate the global trajectory of each individual behavior, where v_t^i and $\hat{v}_t^i$ are the estimated and ground truth root position of the final pose at the *t*-th predicted timestamp for each agent *i*.

$$\mathrm{FDE}(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = ||v_t - \hat{v}_t^i||^2 \qquad (9)$$

MPJPE. Following the standard evaluation adopted in the CHICO dataset [21], we use the Mean Per Joint Position Error (MPJPE) to evaluate human motion forecasting.

4.2.2 Metrics in HRI

While AJPE and JPE allow for precise evaluation in human-human interaction generation, they are not sufficient in the case of robot behavior generation due to the lack of a clean dataset. In Section 3.3.3, we pointed out the limitations of the human-to-robot retargeting models, which enable us to transfer human behavior to robots locally but do not preserve the relative distances existing in the human-human partners. In our HRI scenario, AJPE and JPE only account for the alignment to the ImitationNet reference, which is important during proximity situations, where individual style is the most important. However, they do not measure the alignment of the HRI based on the original HHI. To better measure this social synchrony between humans and robots, we propose two new metrics.

To do this, we detect the ground-truth proximity between the original human-human interaction $(S_m(H^A_{ref}, H^B_{ref}))$ and compare it with the generated human-robot behaviors $(S_m(H^A_{gen}, R^B_{gen}))$ using the F1-Score. We define F1 Contact as the F1-Score when using no margin (*m*=0.0) and F1-Proximity when using a small margin (*m*=0.05).

Finally, we also evaluate the gaze behavior of the robot compared to the reference human. We consider that transferring a natural gaze behavior to the robot is very important in social interaction, as it helps the human partner to better understand the robot's intent. We measure the difference in head rotation as the distance between the normalized quaternions of the reference human head and the retargeted robot head.

4.3 Implementation Details

All models were trained on a single GPU for 100 epochs using an exponential decay scheduler and AdamW as an optimizer, with a 5-epoch warm-up. We observed that the evaluation code used in [23] only considered the human pelvis in the global coordinate system, instead of all joints. This affected the measurements of the JPE and AJPE across all baselines. We corrected the issue and re-evaluated all metrics.

4.4 Quantitative Evaluation

Motivated by the goal of designing a robust and accurate model to generate high-quality robot behaviors for HRI, we first conduct an extensive evaluation of the proposed model using ground-truth data and existing benchmark datasets. Therefore, we first evaluate our model in the InterGen dataset [16] for Human-Human Interaction, and in the CHICO dataset [21] for Human-Robot Collaboration, and perform a thorough ablation study to investigate the benefits of all proposed approaches. Later, we evaluate our best model for human-robot interaction.

4.4.1 Human-Human Interaction

We train and evaluate our model, along with state-of-the-art baselines for multi-person motion forecasting, using the InterGen dataset under identical training configurations to ensure a fair comparison. As shown in Table 1, our framework consistently outperforms all baselines across various metrics. We denote *Zero Velocity* as the repetition of the last pose observed, which acts as the simplest baseline for our evaluation. Additionally, [6] is a single-person motion forecasting model that treats each participant in the dyadic interaction as independent. We use [6] in the comparison to showcase the strong benefits of considering the human partner when modeling social interactions. In contrast, models such as [10–12, 15] are specifically designed for multi-person motion forecasting. Table 1 indicates that while autoregressive approaches like those in [11, 15] perform well for short-term predictions, they struggle with capturing long-term dependencies. Furthermore, our model predicts the entire motion sequence in one shot, resulting in significantly faster inference, which will be crucial later for the robot behavior generation in real HRI. Fig. 5 illustrates the performance of our model in human-human interaction scenarios.

4.4.2 Human-Robot Collaboration

To further evaluate the robustness of our approach to different datasets, we train and evaluate our model in the CHICO dataset [21] for the human motion forecasting conditioned on the observed robot and observed human motion. Following the original work [21], we report the MPJPE for the short-term (400 ms) and long-term (1000 ms) horizons. Table 2 showcases that our model outperforms previous baselines, especially in long-term forecasting.

4.4.3 Ablation Study

We conduct a systematic evaluation of different variations of our framework to showcase the benefits of those approaches, which are presented in Table 3.

First, we assess the benefit of using text as an additional modality to guide the human-human interaction generation.

Table 1 Evaluation of our model in the InterGen dataset for the human-human interaction forecasting task [16]. We indicate with bold the best result and with an underscore the second-best result across each different metric, where a lower metric is better

	JPE (mm) ↓				APJE (mm) ↓				FDE (mm)↓			
seconds	**0.20**	**0.50**	**1.00**	**1.50**	**0.20**	**0.50**	**1.00**	**1.50**	**0.20**	**0.50**	**1.00**	**1.50**
Zero Velocity	48.75	126.69	234.45	324.36	29.11	70.54	112.22	132.59	39.93	106.35	205.30	292.73
HisRepIt [6]	55.99	103.85	176.61	235.55	37.48	64.76	93.95	119.96	38.92	78.68	147.26	202.80
SocialTGCN [12]	25.94	59.93	119.53	185.00	20.69	47.13	82.67	106.85	16.24	37.42	84.00	146.06
TBIFormer [10]	26.55	66.27	135.47	205.67	18.90	48.62	88.44	112.74	19.51	46.12	100.39	166.85
TwoBody [11]	**18.24**	47.60	106.19	174.08	**14.57**	37.66	72.30	97.27	11.95	29.45	75.63	140.70
ExPI [15]	19.01	56.69	127.01	203.70	15.30	44.41	85.66	113.83	13.38	37.60	93.15	166.10
Ours	18.82	**42.41**	**71.32**	**108.57**	15.15	**33.40**	**50.22**	**66.65**	**11.61**	**25.70**	**48.15**	**80.92**

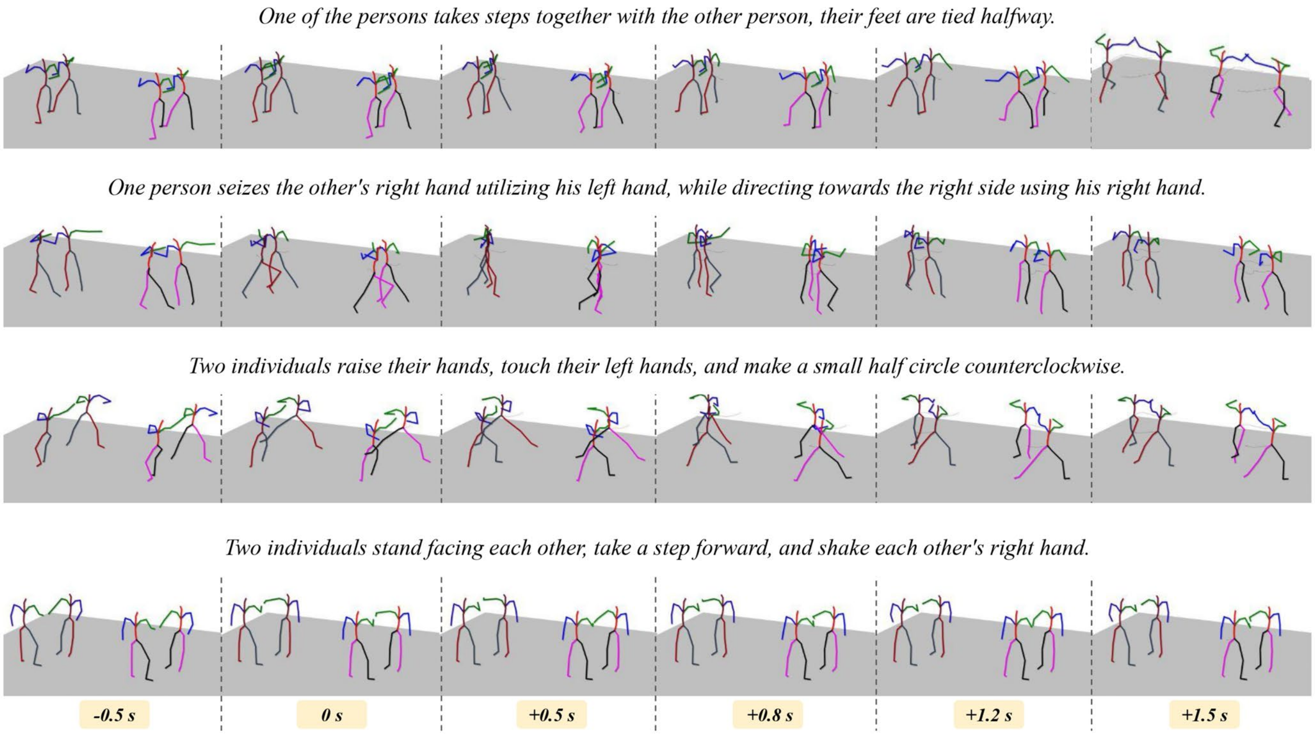

One of the persons takes steps together with the other person, their feet are tied halfway.
One person seizes the other's right hand utilizing his left hand, while directing towards the right side using his right hand.
Two individuals raise their hands, touch their left hands, and make a small half circle counterclockwise.
Two individuals stand facing each other, take a step forward, and shake each other's right hand.
-0.5 s
0 s
+0.5 s
+0.8 s
+1.2 s
+1.5 s

Fig. 5 Qualitative results for Human-Human Interaction in the InterGen [16] dataset. Each scenario shows the ground-truth human pair (left) and the predicted (right) per each time horizon

Table 2 Quantitative evaluation of the short (400 ms) and long-term (1000 ms) motion forecasting in the CHICO dataset [21] reported in MPJPE. Here, bold indicates the best result and underscores the second-best result

milliseconds (ms)	400	1000
Zero Velocity	162.0	282.0
HisRepIt [6]	54.6	91.6
MSR-GCN [4]	54.1	90.7
STS-GCN [68]	53.0	87.4
SeS-GCN [21]	<u>48.8</u>	<u>85.3</u>
Ours	**47.1**	**80.5**

We assess the performance without this feature for a fair comparison with the baseline models that do not consider the text guidance in their architecture. Still, our model outperforms previous baselines by large margins, mostly on the long-term prediction.

Second, we evaluate four different variations of our architecture, such as removing learning only the motion displacement ('w/o Baseline'), adopting the Discrete Cosine Transform ('w/ DCT'), not using a set of temporal MLP layers in the self-motion generator module ('w/o TempMLP'), and not using the iterative refinement in the social motion generation module ('w/o Iterative Refinement'). As mentioned in Sect. 3.2, we simplify the training objective of our model by just learning the displacement of the motions with respect to the last state observed per each individual. Here, 'w/o Baseline' refers to predicting the full motion directly, instead of learning only the displacement, which significantly degrades performance in the long term. Then, we also analyze the use of DCT to encode the motions in the frequency domain as proposed in [11]. While prior works [11] have shown that DCT aids models trained on smaller datasets by enhancing generalization, we observe that when using DCT on large datasets such as InterGen, our model struggles to capture the details in the interaction. The use of TempMLP has also been adopted by prior works in the motion forecasting field, such as [8–10]. Table 3 shows benefits of using TempMLP in the short-term predictions. Finally, we evaluate the iterative refinement proposed in the social-motion generation module by comparing interleaved cross-attention with a sequential refinement approach that first updates one individual and then the other. As expected, interleaving the refinement helps the model incorporate the other partner's intention into the behavior generation of one partner, which improves the stability of long-term forecasting.

Thirdly, we evaluate the use of our individual losses $\mathcal{L}_{ind}$, which help the model in the short term. These results are aligned with the motivation proposed in Section 3.2, as $\mathcal{L}_{ind}$ encourages a better 'self-intent', which drives the human motion in the short-term as it accounts only for one's dynamics. However, it slightly reduces the importance of social motion, which is key for better long-term performance. Adequately weighting both losses helps to optimize the human-human interaction behavior, as shown in Table 3.

4.4.4 Human-Robot Interaction

We adopt the best model configuration from the human-human interaction benchmark and train it for robot behavior generation in social human-robot interactions. We report the results in Table 4, which clearly demonstrates the effectiveness of using the proposed imitation ($\mathcal{L}_{imit}$) and interaction ($\mathcal{L}_{int}$) losses with the adaptive weighting to enhance both the robot individual style (JPE, AJPE, FDE) and social metrics.

Notice that simply using the imitation ($\mathcal{L}_{imit}$) and interaction ($\mathcal{L}_{int}$) losses degrades the quality of the motion style. The reason for this issue was already discussed in Sect. 3.3.4, where a conflicting optimization objective during training causes the model to optimize for opposite behaviors: either following the noisy robot reference or the ground-truth human hand trajectory. Our results showcase that using our adaptive weighting factor w_{ph} presented in Sect. 3.3.4 leads to an improvement across most of the style metrics and all social metrics. We believe that using the $\mathcal{L}_{inter}$ and $\mathcal{L}_{imit}$ in Cartesian space (contrary to $\mathcal{L}_{self}$ and $\mathcal{L}_{social}$ that are in a mix of joint-space and base transformations, as obtained from ImitationNet) provides additional guidance to the model to better preserve the individual style of the robot's behavior and generate a more natural interaction with the human partner. The last row of Table 4 shows how our model improves upon the reference behaviors from ImitationNet as it achieves more accurate contacts and closer proximity.

Additionally, we also evaluate the benefit of adopting the pre-trained ImitationNet [24] as the encoder and decoder for the robot's local joints in our model. Indeed, the only difference lies in whether we train the MLP encoder and decoder layers from scratch or freeze them with weights pre-trained for the human-to-robot retargeting task. We observe a slight improvement in social metrics compared to models that do not use ImitationNet. We believe that the shared latent space from ImitationNet is more continuous and helps overcome the existing coupling effect in joint-based representation, where small changes in root joints drastically affect the Cartesian position of the end-effector. Given that ImitationNet is trained with contrastive losses (pulling visually similar poses together, and pushing different poses away), working on that representation enhances the possibility to manipulate certain joints to achieve an end-effector position without affecting the motion style.

Finally, Table 4 shows that all models are able to follow the reference gaze behavior with high precision, with

Table 3 Ablation study of our ECHO model for the social motion forecasting task in the InterGen dataset [16]

	JPE (mm) ↓			APJE (mm) ↓			FDE (mm) ↓		
seconds	**0.20**	**0.50**	**1.00**	**1.50**	**0.20**	**0.50**	**1.00**	**1.50**	**0.20**	**0.50**	**1.00**	**1.50**
w/o Text	22,84	45,40	73,31	110,13	18,86	36,49	52,94	69,69	13,24	26,77	48,33	80,46
w/o Baseline	21,94	48,15	82,48	124,35	17,80	38,24	58,85	76,46	13,57	29,12	55,13	93,10
w/ DCT	21,01	44,44	72,69	109,67	17,10	35,27	51,87	68,65	13,05	26,83	48,70	80,94
w/o TempMLP	20,76	44,20	72,29	109,49	17,28	35,48	51,86	68,64	13,04	26,82	48,70	80,94
w/o Iterative Refinement	19,04	42,90	72,37	110,16	15,42	33,94	51,20	67,55	11,62	25,74	48,62	82,16
w/o $\mathcal{L}_{ind}$	20,11	45,03	72,97	**107,89**	16,69	35,96	52,17	67,90	11,84	26,64	48,55	**79,20**
Ours	**18,82**	**42,41**	**71,32**	108,57	**15,15**	**33,40**	**50,22**	**66,65**	**11,61**	**25,70**	**48,15**	80,92

Table 4 Evaluation of the robot behavior generation in the Human-Robot Interaction scenario. We use the InterGen dataset [16] and transfer a human behavior to a robot behavior using ImitationNet [24]. We evaluate the impact of our imitation ($\mathcal{L}_{imit}$) and interaction ($\mathcal{L}_{int}$) losses, with and without the adaptive weighting factor w_{ph}, as well as the use of the pre-trained ImitationNet module as the encoder and decoder for the robot local motions. The last row indicates the F1 contact and proximity when using the reference motions from ImitationNet

$\mathcal{L}_{imit}$	$\mathcal{L}_{int}$	Imit.Net [24]	JPE* (mm) ↓			APJE* (mm) ↓			FDE* (mm) ↓			Social Metrics		
			0.20	**0.50**	**1.00**	**0.20**	**0.50**	**1.00**	**0.20**	**0.50**	**1.00**	**Head Rot.↓**	**F1 Contact↑**	**F1 Proximity↑**
0	0	×	**33.21**	79.25	156.14	**24.68**	57.27	102.42	17.05	43.50	97.79	**0.0017**	0.20	0.34
1	0	×	91.40	190.43	225.02	83.64	103.90	137.35	17.78	44.49	99.77	0.0018	0.30	0.36
1	1	×	90.52	191.76	224.98	83.73	124.98	139.14	18.98	45.52	101.59	0.0018	0.32	0.35
0	0	✓	33.82	81.12	158.80	24.28	58.02	103.76	18.73	45.63	99.76	**0.0017**	0.20	0.32
1	0	✓	137,73	170.11	232.90	128.61	146.74	176.22	20.25	49.58	109.30	0.0019	0.24	0.38
1	1	✓	160.16	194.88	267.58	136.01	161.68	203.77	43.96	67.56	129.09	0.0021	0.36	0.37
w_{ph}	w_{ph}	✓	41.48	**76.71**	**124.47**	34.23	**56.76**	**76.32**	**16.32**	**41.70**	**87.58**	**0.0017**	**0.55**	**0.66**
													0.21	0.34

slight improvement when no additional social or embodiment losses are applied. This is expected as removing $\mathcal{L}_{int}$ and $\mathcal{L}_{imit}$ causes all the learning to focus on the self $\mathcal{L}_{self}$ and social $\mathcal{L}_{soc}$ losses, which are responsible for optimizing the prediction of the robot head orientation.

Additionally, we showcase three generated human-robot interaction behaviors in different contexts: an imitation game (Fig. 6), a physical contact scenario (Fig. 7), and an actor-reactor interaction (Fig. 8). In all images, the blue and pink human characters represent the ground-truth human-human interaction, while the green character and the robot reflect the behaviors generated by our framework. The textual descriptions conditioning the generated behavior are shown above each image sequence. Notably, the robot successfully establishes physical contact in Fig. 7 by raising its arm toward the partner's shoulder. Similarly, Fig. 8 demonstrates how our generative model captures the spatio-temporal dependencies inherent in social interactions, as the robot waits for the partner's wave before responding with the appropriate gesture.

5 Real-World Experiments

To further evaluate the robustness of our model and to demonstrate the possibility of controlling a robot in real-time, we implemented an end-to-end pipeline to generate the commands for a TIAGo++ robot for social human-robot interactions.

Our pipeline is presented in Fig. 9. We make use of YOLOv9 object detector [69] to detect the humans from the robot's onboard camera, and forward the cropped human bounding box to HybridIK [70] for the human pose estimation. With the aforementioned pipeline we are able to obtain human poses at approximately 15 FPS , which is closer to the 15 FPS rate used when training our model in the InterGen dataset. By aligning the depth image from the robot's sensor and the detected 2D poses, we are able to obtain the human pose in the robot's optical frame and transform those coordinates to the robot base. Then, we construct the human and robot motion sequences as proposed in Sect. 3 and generate the appropriate robot's joint angles, root position, and rotation, as well as head movement. Given that the robot's behavior was completely autonomous and no safety measures were implemented to avoid obstacles in the scene, and because we observed unstable behaviors when both the base and the robot arms were moving at the same time, we chose not to control the robot's base in the real world. The results, depicted in Figs. 10 and 11, showcase the control of the real TIAGo++ robot for human-robot interactions. In particular, Fig. 10 depicts the robot performing a social behavior where it waves back to a human. Notice that we projected

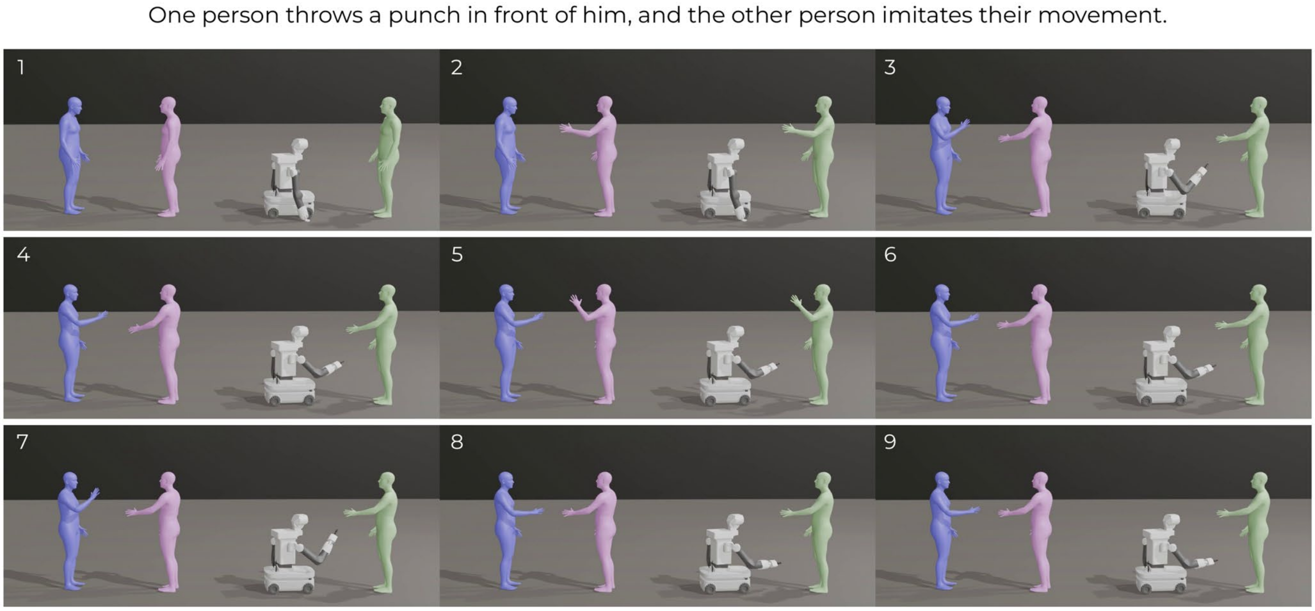

One person throws a punch in front of him, and the other person imitates their movement.

Fig. 6 Imitation game. This scenario depicts the ability of our generative model to understand simple games such as repeating other movements. (blue human imitates red person). We believe that using the shared latent space of ImitationNet helps reducing the gap between both embodiments, which helps the robot understand what the word 'imitation' entails

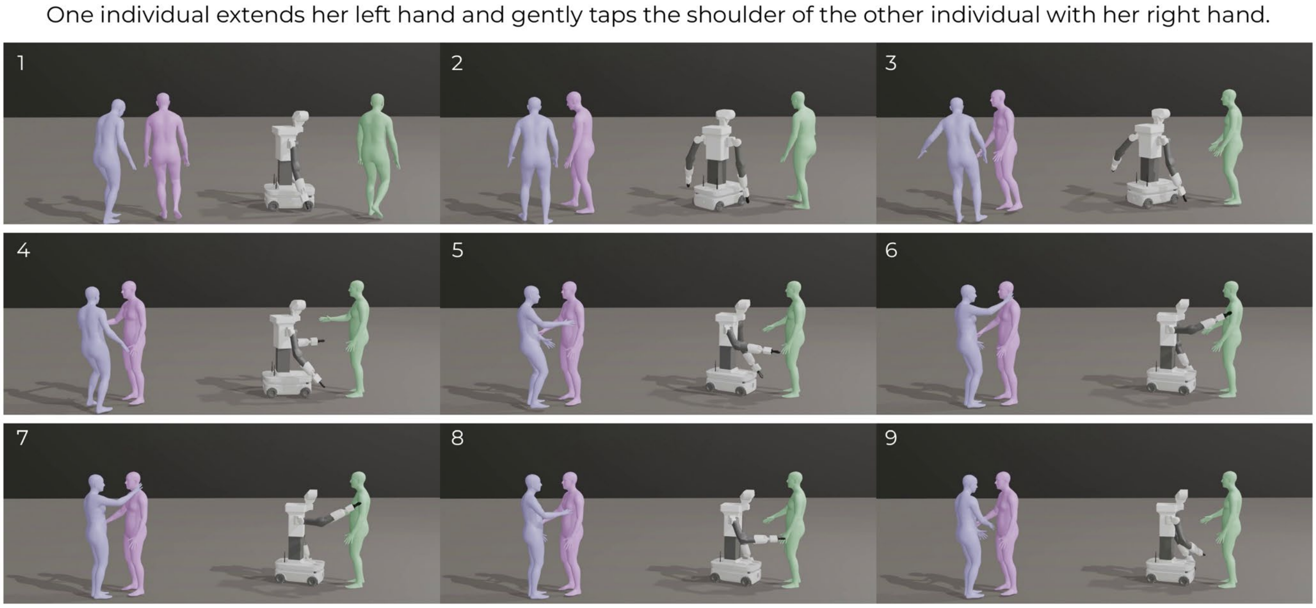

One individual extends her left hand and gently taps the shoulder of the other individual with her right hand.

Fig. 7 Physical contact. Thanks to the proposed losses, the robot is able to extend the arms to contact the human partner's shoulder during this interaction

the estimated human pose in the background at each time of the snapshot, which showcases the inaccuracies of the fed data to the model. Similarly, in Fig. 11 we evaluated a handshake between a human and a robot. However, given that we are using the on-board camera of the TIAGo++ for pose estimation, we observe that if the human was too close to the robot, the pose estimation was incorrect or even not detecting the person and as a result, the robot was not behaving properly. Therefore, we simulated a handshake from a distance. Still, we can observe that the TIAGo++ robot adapts its end-effector position to the human's hand height, which demonstrates that it captured the importance of ensuring contact with the human hand during a prompted 'handshake' interaction.

One individual stands facing another individual and lifts both hands, greeting him with a wave. At the same time, the second person also raises both hands and reciprocates the gesture.

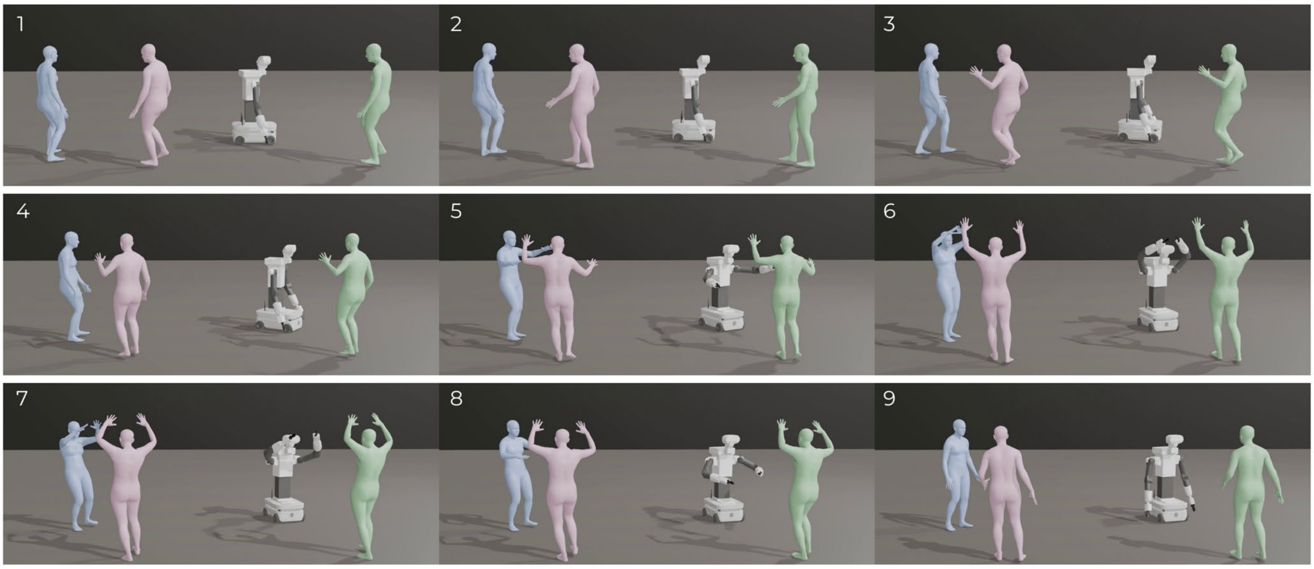


Fig. 8 Actor-reactor situation. During this task, the robot is able to generate a fast response to the partner's waving behavior, showcasing the importance of forecasting the counterpart motion (e.g., anticipating the waving while the partner is only starting to raise the arms) to accurately decide when to generate the reciprocative gesture

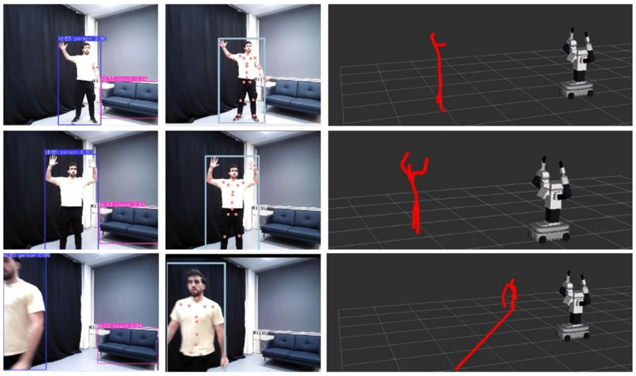


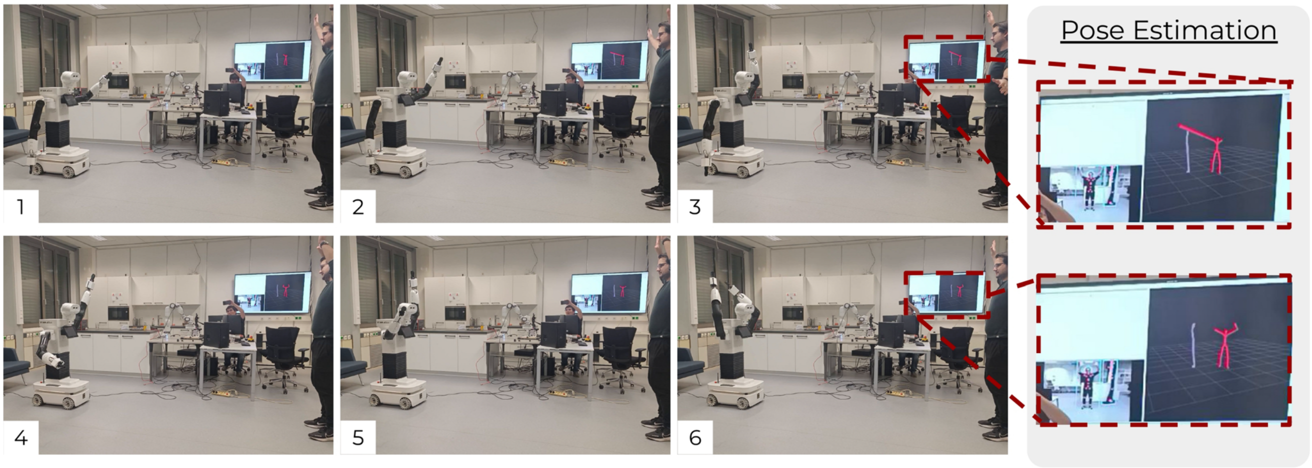
Fig. 9 End-to-end pipeline of our real-world robot control for social human-robot interactions. Here we do not showcase any robot behavior, but the overall human capture system. First, we use YOLOv9 [69] and HybridIK [70] for human detection and pose estimation. Then, we transform all the coordinates to the robot's coordinate frame and generate the robot behavior accordingly

6 Limitations and Future Work

Despite the generated robot behavior depicted in the real-world HRI aligning with the prompted text, it is important to mention that the quality of those behaviors is lower compared to the high-quality interactions when tested in the InterGen dataset. We believe this sim-to-real gap is mainly caused by:

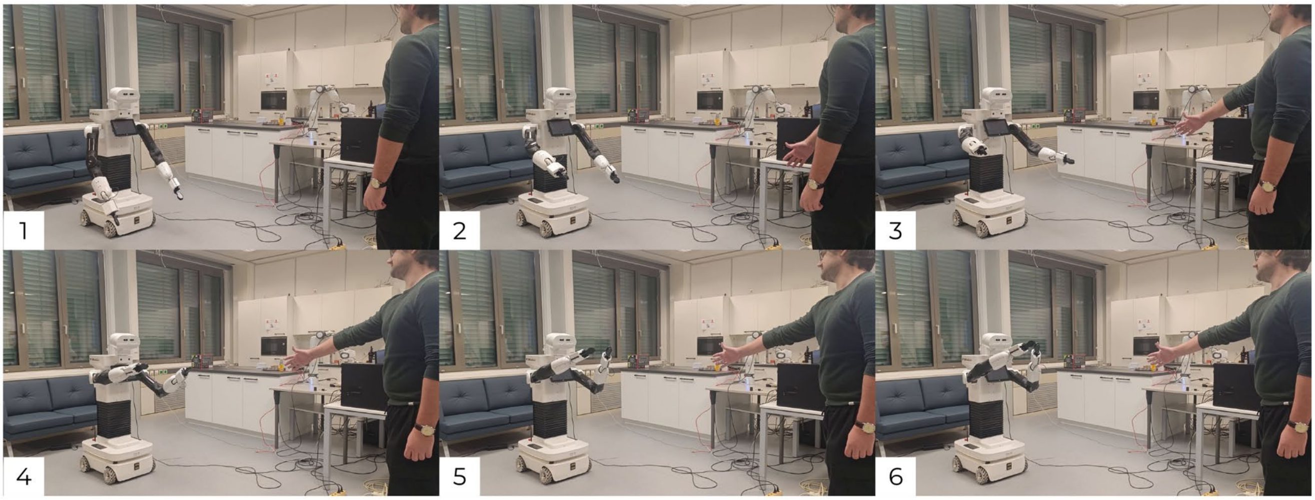
- Inaccuracies in the pose estimations, as can be observed in Figs. 9 and 10. This issue accentuates when the human is positioned very close from the on-board TIAGo++ camera, causing the pose estimator to fail since it cannot detect the entire body. Developing a more reliable pose estimation for very close situations could mitigate this sim-to-real gap. Similarly, adapting our generative architecture by incorporating potential masking in the human partner observation, as [7, 8] , could enhance the robustness to outliers and occlusions in the perception system. We leave this for future work.
- Communication delays that exist between the commanded robot joints and the actual robot response, which make hard the implementation on the TIAGo++ robot without faster controllers.

Additionally, our current model relies on the contextualization of the interaction through a textual description, which describes expected behaviors as the 'waving' depicted in Figs. 8 and 10. We believe that including textual conditions to the generative model is beneficial as it instructs and conditions the expected behaviors of the robots, making them suitable to operate in pre-defined scenarios or roles. For instance, if a service robot is at a conference center greeting and assisting the public, the robot operator could directly specify expected actions as 'shake hands', 'bow', or 'greet'. Similarly, it allows one to coordinate speech with actions assuming a VLM is processing visual inputs and commanding the motions to our framework. Finally, and even simpler, the model could just be prompted to follow partner's commands, such as 'let's dance bachata' or 'give me a hug'.



One individual stands facing another individual and lifts both hands, greeting him with a wave. At the same time, the second person also raises both hands and reciprocates the gesture.

Fig. 10 Real-world waving interaction between a human and the TIAGo++ robot. Despite the pose estimation loses track of the human, our method is robust enough to generate reasonable interaction behaviors and wave back to the human



One individual extends a handshake with their right hand, while the other reciprocates with their own.

Fig. 11 Real-world waving interaction of a human simulating a handshake with the TIAGo++ robot. Since the pose estimation uses the on-board TIAGo++ camera, the handshake motion is performed from distance

Exploring better methods to flexibly decide when to prompt and how to prompt the model remains an underexplored research path, which would enhance even more the autonomy and contextual alignment of the robots' behaviors in populated environments.

7 Conclusions

Motivated by the dynamic and adaptable nature of human behaviors observed in human-human interactions (HHI), this paper introduces a transformer-based framework that predicts the human intent and robot behaviors concurrently to enhance a natural human-robot interaction (HRI). First, we make use of a human-to-robot motion retargeting system to learn robot behaviors from human data. Then, we adopt an iterative refinement process that learns to adapt both human and robot motions to each other's intent under a social interaction. As a result, our model outperforms the state-of-the-art when forecasting dyadic human motion within the largest dataset available and predicting the human intent for a Human-Robot Collaboration task. By leveraging cues from HHI as references during training and with the inclusion of a novel dynamic loss adaptation based on the agents proximity, we ensure that robot-generated behaviors are aligned with social interaction norms. We conduct a comprehensive ablation study to systematically validate the efficacy of our approach in both HHI and HRI contexts and propose novel metrics to gauge the dynamics of social interactions. Finally, we evaluate our framework through qualitative HRI experiments on simulated and real-world TIAGo++ robots, thereby paving the way for autonomous social robots capable of navigating and working alongside humans.

Acknowledgements This work is funded by Marie Sklodowska-Curie Action Horizon 2020 (Grant agreement No. 955778) for the project 'Personalized Robotics as Service Oriented Applications' (PERSEO).

Author Contributions Not applicable.

Funding Open access funding provided by TU Wien (TUW). This work is funded by Marie Sklodowska-Curie Action Horizon 2020 (Grant agreement No. 955778) for the project 'Personalized Robotics as Service Oriented Applications' (PERSEO).

Data Availability Not applicable.

Code Availability Not applicable.

Declarations

Ethics Approval and Consent to Participate Not applicable.

Consent for Publication Not applicable.

Conflict of Interest The authors have no conflicts of interest to declare that are relevant to the content of this article.

References

1. Thornton MA, Weaverdyck ME, Tamir DI (2018) The social brain automatically predicts others' future mental states. J Neurosci 39:140–148
2. Fragkiadaki K, Levine S, Felsen P, Malik J (2015) Recurrent network models for human dynamics. In IEEE International Conference on Computer Vision, pp 4346–4354
3. Jain A, Zamir AR, Savarese S, Saxena A (2016) Structural-rnn: deep learning on spatio-temporal graphs. In Conference on Computer Vision and Pattern Recognition (CVPR), pp 5308–5317
4. Dang L, Nie Y, Long C, Zhang Q, Li G (2021) Msr-gcn: multi-scale residual graph convolution networks for human motion prediction. In IEEE/CVF International Conference on Computer Vision (ICCV), pp 11467–11476
5. Mao W, Liu M, Salzmann M, Li H (2019) Learning trajectory dependencies for human motion prediction. In International Conference on Computer Vision (ICCV), pp 9489–9497
6. Mao W, Liu M, Salzmann M (2020) History repeats itself: human motion prediction via motion attention. In European Conference on Computer Vision (ECCV), pp 474–489
7. Valls Mascaro E, Ma S, Ahn H, Lee D (2022) Robust human motion forcasting using transformer-based model. In 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
8. Mascaro EV, Ahn H, Lee D (2023). A unified masked autoencoder with patchified skeletons for motion synthesis. arXiv preprint arXiv:2308.07301
9. Guo W, Du Y, Shen X, Lepetit V, Alameda-Pineda X, Moreno-Noguer F (2023) Back to mlp: a simple baseline for human motion prediction. In IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp 4809–4819
10. Peng X, Mao S, Wu Z (2023) Trajectory-aware body interaction transformer for multi-person pose forecasting. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 17121–17130
11. Rahman MRU, Scofano L, De Matteis E, Flaborea A, Sampieri A, Galasso F (2023) Best practices for 2-body pose forecasting. In IEEE/CVF Conference on Computer Vision and Pattern Recognition
12. Peng X, Zhou X, Luo Y, Wen H, Wu Z (2023). The MI-Motion dataset and benchmark for 3D multi-person motion prediction. arXiv preprint arXiv:2306.13566, 2023
13. Marcard T, Henschel R, Black M, Rosenhahn B, Pons-Moll G (2018) Recovering accurate 3d human pose in the wild using imus and a moving camera. In European Conference on Computer Vision (ECCV)
14. Wang J, Xu H, Narasimhan M, Wang X (2021) Multi-person 3d motion prediction with multi-range transformers. In Proceedings of the 35th International Conference on Neural Information Processing Systems. NIP'S 21. Curran Associates Inc, Red Hook, NY, USA
15. Guo W, Bie X, Alameda-Pineda X, Moreno–Noguer F (2022) Multi-person extreme motion prediction. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
16. Liang H, Zhang W, Li W, Yu J, Xu L (2024) Intergen: diffusion-based multi-human motion generation under complex interactions. Int J Comput Vision 132(9):3463–3483
17. Kopp T, Baumgartner M, Kinkel S (2021) Success factors for introducing industrial human-robot interaction in practice: an empirically driven framework. Int J Adv Manuf Technol 112
18. Chen M, Nikolaidis S, Soh H, Hsu D, Srinivasa S (2020) Trust-aware decision making for human-robot collaboration: model learning and planning. ACM Trans Hum Rob Interact (THRI) 9(2):1–23
19. (2020) Recurrent neural network for motion trajectory prediction in human-robot collaborative assembly. CIRP Ann 69(1):9–12
20. Mascaro EV, Sliwowski D, Lee D (2023) HOI4ABOT: human-object interaction anticipation for human intention reading assistive roBots. In 7th Annual Conference on Robot Learning
21. Sampieri A, Melendugno GMD, Avogaro A, Cunico F, Setti F, Skenderi G, Cristani M, Galasso F (2022) Pose forecasting in industrial human-robot collaboration. In European Conference on Computer Vision, Springer pp 51–69

22. Kedia K, Bhardwaj A, Dan P, Choudhury S (2024) Interact: transformer models for human intent prediction conditioned on robot actions. In IEEE International Conference on Robotics and Automation
23. Valls Mascaro E, Yan Y, Lee D (2024) Robot interaction behavior generation based on social motion forecasting for human-robot interaction. In 2024 IEEE International Conference on Robotics and Automation (ICRA
24. Yan Y, Mascaro EV, Lee D (2023). Unsupervised human-to-robot motion retargeting via expressive latent space. arXiv preprint arXiv:2309.05310, 2023
25. Losey DP, McDonald CG, Battaglia E, O'Malley MK (2018) A review of intent detection, arbitration, and communication aspects of shared control for physical human–robot interaction. Appl Mech Rev 70(1):010804
26. Fang J, Wang F, Xue J, Chua T-S (2024) Behavioral intention prediction in driving scenes: a survey. In IEEE Transactions on Intelligent Transportation Systems, pp 1–22
27. Huang C-M, Mutlu B (2016) Anticipatory robot control for efficient human-robot collaboration. In 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp 83–90
28. Ni Z, Valls Mascaró E, Ahn H, Lee D (2023) Human–object interaction prediction in videos through gaze following. Comput Vision Image Underst 233:103741
29. Schydlo P, Rakovic M, Jamone L, Santos-Victor J (2018) Anticipation in human-robot cooperation: a recurrent neural network approach for multiple action sequences prediction. In 2018 IEEE International Conference on Robotics and Automation (ICRA), IEEE pp 5909–5914
30. Zatsarynna O, Gall J (2023) Action anticipation with goal consistency. In 2023 IEEE International Conference on Image Processing (ICIP), pp 1630–1634
31. Mascaro EV, Ahn H, Lee D (2023) Intention-conditioned long-term human egocentric action anticipation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp 6048–6057
32. Nakamura Y, Takano W, Yamane K (2007) Mimetic communication theory for humanoid robots interacting with humans. In: Robotics research. Springer, Berlin, Heidelberg, pp 128–139
33. Lee D, Ott C, Nakamura Y (2010) Mimetic communication model with compliant physical contact in human—humanoid interaction. Int J Robot Res 29(13):1684–1704
34. Medina Hernández J, Lawitzky M, Mörtl A, Lee D, Hirche S (2011) An experience-driven robotic assistant acquiring human knowledge to improve haptic cooperation 2416–2422
35. Yang L, Li Y, Huang D (2018) Motion synchronization in human-robot co-transport without force sensing. In 2018 37th Chinese Control Conference (CCC), pp 5369–5374
36. Wang Z, Peer A, Buss M (2009) An hmm approach to realistic haptic human-robot interaction. In World Haptics 2009 - Third Joint EuroHaptics Conference and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems, pp 374–379
37. Alahi A, Ramanathan V, Fei-Fei L (2014) Socially-aware large-scale crowd forecasting. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
38. Adeli V, Adeli E, Reid I, Niebles JC, Rezatofighi H (2020) Socially and contextually aware human motion and pose forecasting. IEEE Robot Autom Lett 5(4)
39. Baldauf D, Deubel H (2010) Attentional landscapes in reaching and grasping. Vision Res 50(11):999–1013
40. Belardinelli A (2023) Gaze-based intention estimation: principles, methodologies, and applications in hri. ACM Trans Hum Rob Interact
41. Belardinelli A, Kondapally AR, Ruiken D, Tanneberg D, Watabe T (2022) Intention estimation from gaze and motion features for human-robot shared-control object manipulation. In 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE pp 9806–9813
42. Zhao Q, Wang S, Zhang C, Fu C, Do MQ, Agarwal N, Lee K, Sun C (2024) Antgpt: can large language models help long-term action anticipation from videos? ICLR
43. Martinez J, Black MJ, Romero J (2017) On human motion prediction using recurrent neural networks. In Conference on Computer Vision and Pattern Recognition (CVPR), pp 2891–2900
44. Amirian J, Hayet J-B, Pettré J (2019) Social ways: learning multi-modal distributions of pedestrian trajectories with gans. In IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops
45. Vendrow E, Kumar S, Adeli E, Rezatofighi H (2022). SoMoFormer: multi-person pose forecasting with transformers. arXiv preprint arXiv:2208.14023, 2022
46. Kipf TN, Welling M (2016). Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907
47. Sohl-Dickstein J, Weiss E, Maheswaranathan N, Ganguli S (2015) Deep unsupervised learning using nonequilibrium thermodynamics. In International Conference on Machine Learning, PMLR pp 2256–2265
48. Ahn H, Mascaro EV, Lee D (2023) Can we use diffusion probabilistic models for 3d motion prediction? In 2023 IEEE International Conference on Robotics and Automation (ICRA), IEEE pp 9837–9843
49. Tevet G, Raab S, Gordon B, Shafir Y, Cohen-Or D, Bermano AH (2023) Human motion diffusion model. In The Eleventh International Conference on Learning Representations
50. Xu L, Zhou Y, Yan Y, Jin X, Zhu W, Rao F, Yang X, Zeng W (2024) Regennet: towards human action-reaction synthesis. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
51. Ghosh A, Dabral R, Golyanik V, Theobalt C, Slusallek P (2023). Remos: reactive 3d motion synthesis for two-person interactions. arXiv preprint arXiv:2311.17057
52. Siyao L, Gu T, Yang Z, Lin Z, Liu Z, Ding H, Yang L, Loy CC (2024) Duolando: follower GPT with off-policy reinforcement learning for dance accompaniment. In The Twelfth International Conference on Learning Representations
53. Gleicher M (1998) Retargetting motion to new characters. In Proceedings of the 25th annual conference on Computer graphics and interactive techniques
54. Koenemann J, Burget F, Bennewitz M (2014) Real-time imitation of human whole-body motions by humanoids
55. Devanne M, Nguyen SM (2017) Multi-level motion analysis for physical exercises assessment in kinaesthetic rehabilitation
56. Darvish K, Tirupachuri Y, Romualdi G, Rapetti L, Ferigo D, Chavez FJA, Pucci D (2019) Whole-body geometric retargeting for humanoid robots. In 2019 IEEE-RAS 19th International Conference on Humanoid Robots (Humanoids), IEEE pp 679–686
57. Penco L, Clement B, Moduano V, Mingo Hoffman E, Nava G, Pucci D, Tsagarakis N, Mourert J, Ivaldi S (2018) Robust real-time whole-body motion retargeting from human to humanoid 425–432
58. Delhaisse B, Esteban D, Rozo L, Caldwell D (2017) Transfer learning of shared latent spaces between robots with similar kinematic structure
59. Villegas R, Yang J, Ceylan D, Lee H (2018) Neural kinematic networks for unsupervised motion retargetting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 8639–8648
60. Lim J, Chang H, Choi J (2019) Pmnet: learning of disentangled pose and movement for unsupervised motion retargeting.

In Proceedings of the 30th British Machine Vision Conference (BMVC 2019). British Machine Vision Association, BMVA, Cardiff, UK

61. Aberman K, Li P, Lischinski D, Sorkine-Hornung O, Cohen-Or D, Chen B (2020) Skeleton-aware networks for deep motion retargeting. ACM Trans Graph (TOG) 39(4):62–1
62. Zhang J, Weng J, Kang D, Zhao F, Huang S, Zhe X, Bao L, Shan Y, Wang J, Tu Z (2023) Skinned motion retargeting with residual perception of motion semantics & geometry. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
63. Choi S, Song MJ, Ahn H, Kim J (2021) Self-supervised motion retargeting with safety guarantee. In 2021 IEEE International Conference on Robotics and Automation (ICRA), IEEE pp 8097–8103
64. Oreshkin BN, Valkanas A, Harvey FG, Ménard L-S, Bocquelet F, Coates MJ (2023) Motion in-betweening via deep delta-interpolator. In IEEE Transactions on Visualization and Computer Graphics
65. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst (NeurIPS) 30
66. Petrovich M, Black MJ, Varol G (2023) TMR: text-to-motion retrieval using contrastive 3D human motion synthesis. In International Conference on Computer Vision (ICCV)
67. Guo C, Zou S, Zuo X, Wang S, Ji W, Li X, Cheng L (2022) Generating diverse and natural 3d human motions from text. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 5152–5161
68. Sofianos T, Sampieri A, Franco L, Galasso F (2021) Space-time-separable graph convolutional network for pose forecasting. In IEEE/CVF International Conference on Computer Vision, pp 11209–11218
69. Wang C-Y, Yeh I-H, Liao H-YM (2024). Yolov9: learning what you want to learn using programmable gradient information. arXiv preprint arXiv:2402.13616
70. Li J, Xu C, Chen Z, Bian S, Yang L, Lu C (2021) Hybrik: a hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 3383–3393

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Esteve Valls Mascaro received a B.S. and M.S degree in telecommunications engineering from Universitat Politecnica de Catalunya (UPC) in 2020 and 2021, respectively, while he worked in several companies as a computer vision engineer. He joined the Human-Centered Assistive Robotics Group at the Technical University of Munich (TUM),Munich, Germany in 2021, and later the Autonomous Systems Lab at Technische Universität Wien (TU Wien), Vienna, Austria, to work towards his Ph.D. degree. His research is focused on understanding the human intention through AI for a better human-robot interaction.

Dongheui Lee is a Full Professor of Autonomous Systems at TU Wien since 2022. She is also leading the Human-Centered Assistive Robotics group at the German Aerospace Center (DLR), Institute of Robotics and Mechatronics, since 2017. Her research interests include human motion understanding, human-robot interaction, machine learning in robotics, and assistive robotics. Prior to her appointment at TU Wien, she was an Assistant Professor and Associate Professor at the Technical University of Munich (TUM), a Project Assistant Professor at the University of Tokyo, and a research scientist at the Korea Institute of Science and Technology (KIST). She obtained a PhD degree from the Department of Mechano-Informatics, University of Tokyo in Japan. She was awarded a Carl von Linde Fellowship at the TUM Institute for Advanced Study and a Helmholtz professorship prize. She has served as Senior Editor and a founding member of IEEE Robotics and Automation Letters (RA-L) and Associate Editor for the IEEE Transactions on Robotics.