2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Robot Interaction Behavior Generation based on Social Motion Forecasting for Human-Robot Interaction

Esteve Valls Mascaro[1] and Yashuai Yan[1] and Dongheui Lee[1,2]
evm7.github.io/ECHO

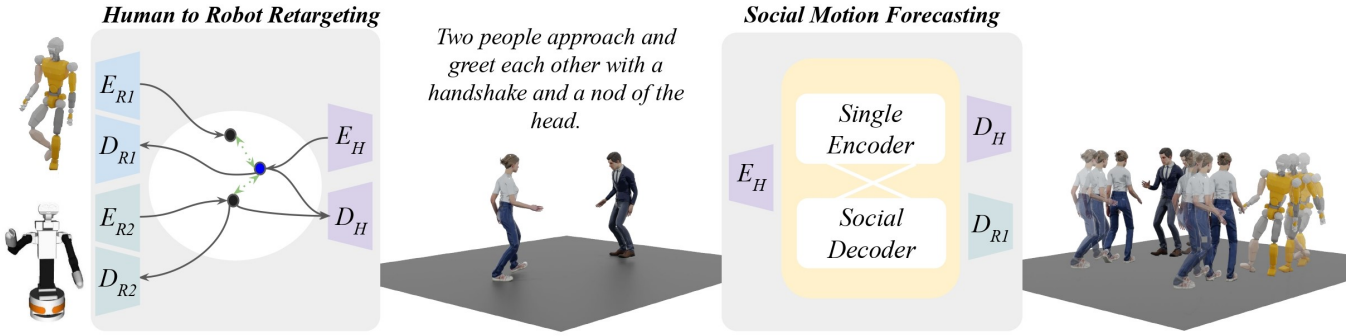

Fig. 1: **Overview of our ECHO framework.** First, we learn how to encode (E) and decode (D) the JVRC-1 robot [1] (in the top left, $R1$) and the TIAGo++ robot (in the bottom left, $R2$) to a latent representation shared with a human (H) while preserving its semantics. Then, we take advantage of this shared space in the social motion forecasting task. Our Single Encoder learns the dynamics of single agents given a textual intention and its past observations. Later, we iteratively refine those motions based on the social context of the surrounding agents using the Social Decoder. Our overall framework can decode the robot's motion in a social environment, closing the gap for natural and accurate Human-Robot Interaction.

***Abstract*— Integrating robots into populated environments is a complex challenge that requires an understanding of human social dynamics. In this work, we propose to model social motion forecasting in a shared human-robot representation space, which facilitates us to synthesize robot motions that interact with humans in social scenarios despite not observing any robot in the motion training. We develop a transformer-based architecture called ECHO, which operates in the aforementioned shared space to predict the future motions of the agents encountered in social scenarios. Contrary to prior works, we reformulate the social motion problem as the refinement of the predicted individual motions based on the surrounding agents, which facilitates the training while allowing for single-motion forecasting when only one human is in the scene. We evaluate our model in multi-person and human-robot motion forecasting tasks and obtain state-of-the-art performance by a large margin while being efficient and performing in real-time. Additionally, our qualitative results showcase the effectiveness of our approach in generating human-robot interaction behaviors that can be controlled via text commands.**

I. INTRODUCTION

As humans, we commonly find ourselves in social scenarios in which we interact and communicate with each other. Through our experience, we learn how to navigate through those dynamic settings by understanding social norms, individual differences, and the intentions of the surrounding people. While robots have made remarkable strides in various fields, integrating them into social environments still remains a complex challenge. In this work, we propose to tackle this problem by first building a shared representation between humans and robots that we use to learn natural dynamics in social scenarios through motion forecasting. An overview of our framework is depicted in Fig. 1.

Human motion forecasting is the task of predicting the human future poses given its past observations. While the community has largely explored individual human motion forecasting [2]–[9], how to effectively model the dependencies in human-human interacting scenarios is still a challenge. When modeling single motions, prior works [2]–[9] only consider local skeleton dynamics and do not predict their global trajectory. However, in multi-person forecasting [10]–[15], there is the need to contextualize the spatial dependencies with the surrounding agents. However, prior works focused on encoding the relationship of multiple humans in scenarios with little or artificially synthesized interactions between the subjects [12]–[14], or with strong interactions that are not adequate for robots [15]. On the contrary, we envision scenarios closer to 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.

To effectively model the social dynamics in human-human interactions, we first ground our model to individual motions. We construct a Transformer-based encoder [17] that forecasts the next human motion using its own past poses. Previous works [10]–[12], [15] fuse the spatial relations of multiple humans in the early stage to build a social motion

[1]Esteve Valls Mascaro and Yashuai Yan and Dongheui Lee are with Autonomous Systems, Technische Universität Wien (TU Wien), Vienna, Austria (e-mail: {esteve.valls.mascaro, yashuai.yan, dongheui.lee}@tuwien.ac.at).
[2]Dongheui Lee is also with the Institute of Robotics and Mechatronics (DLR), German Aerospace Center, Wessling, Germany.

representation. However, we observe that first dealing with individual movements leads to better long-term prediction. Later on, we refine these generated single motions based on the surrounding agents using a cross-attention decoder [17]. By first dealing with individual motions, we ensure that our refinement not only takes into account the current state of the humans in the scene but also the prediction of what other agents and ourselves might do in the future. Our findings observe that understanding other's intentions directly helps to better aggregate all motions into a more natural and human-like social setting. Therefore, we additionally propose to condition the motion synthesis through a text command, that summarizes this overall intention of the social interaction.

Assuming that we are able to model human dynamics in social settings, it is still a challenge to apply this behavior to robots. Previous works [18]–[21] considered a reactive robot behavior where it first forecasts a human motion and then acts accordingly through a set of learned motions. However, we aim to include the robot behavior in the synthesis of the social dynamics. In our previous work [22] we built a shared latent space between humans and robots that encodes poses with similar semantics. We propose to extend [22] so that it can deal with more complex robot kinematics while having a shared representation for multiple robots and a human decoder. By pre-training our human and robot encoders and decoders, we are able to learn the human dynamics in a human-robot shared space, which facilitates a more efficient and natural human-robot interaction generation.

To sum up, we consider the task of social human motion forecasting but operate in a human-robot shared representation space which allows for the synthesis of accurate and real-time human-robot interactions. The contributions of our paper can be summarized as follows:

- A deep-learning architecture that forecasts individual and high-interactive human motions while facilitating their condition on a given interaction.
- The encoding of humans and robots in a shared space to synthesize a human-robot interaction in a social context.
- An efficient model that achieves state-of-the-art performance in real-time for the social human forecasting task as well as in human-robot collaborative scenarios.

II. RELATED WORK

A. Human Motion Forecasting

In the early stages of research, Recurrent Neural Networks (RNNs) [2], [23] were used to understand the time dependencies of human motions and therefore better predict the future human poses. Additionally, [4], [5] adopted the Discrete Cosine Transformations (DCT) with Graph Convolutional Networks (GCN) [24] to better model the spatio-temporal relationships. With the success of attention mechanisms, [6], [7] used a transformer to model the joint relations in both space and time. Despite significant advances in performance and efficiency, all these models are specialized for motion forecasting tasks within a single human and do not take into consideration the social dynamics essential in interactive scenarios.

Modeling multi-person interactions has been a long-standing challenge. Early works [25], [26] tackle the global trajectory prediction of humans in a scene. Later, [27] considered the task of multi-person 3D motion forecasting where the context was used to condition the next movements. Recently, [28] proposed to parallelly leverage individual and multiple human features using transformers to enhance long-term prediction for higher groups of people. On the contrary, [15] focused on modeling dyadic interactions of humans performing extreme actions via cross-attention. [15], [28] used a cross-attention mechanism (CA) to enhance one's motion based on others. However, they consider the CA as the synthesis model, while we only leverage CA for motion refinement and address the synthesis with self-attention (SA). To explicitly capture the interactions among joints between the same individual and with others, [29] operates in each joint with SA, and [10] partitions the body into parts and operates in the flattened sequence through SA. While this strategy facilitates better capturing the spatial relationship between joints in each individual, it increases the complexity of the transformer in capturing inter-human dependencies. Finally, [11] proposes an overall recipe for accurate dyadic motion prediction by reusing DCT and GCN [24] in an autoregressive manner. Contrary to our work, all these previous works adopt DCT to encode the temporal dependencies. We show in our experiments that when using DCT the synthesized motions are too smooth and do not capture the nuances between motions. Additionally, we predict the whole future sequence in one step, which avoids the autoregressive approach from [11], [15] that accumulates error over iterations and collapses in the long term.

B. Motion retargeting in robotics

Human motion retargetting has been widely explored in the animation community [30]–[32]. However, in robotics, it is essential to consider not only the naturality of the motions but also the feasibility and adequate control of the robot [22], [33]–[37]. Early works [34], [35] considered optimization-based approaches to transfer a human pose to robots, which required handcrafted features with limited generalizability. To cope with this issue, deep-learning-based methods [22], [36], [37] learned how to construct a shared latent space to transfer human to robot poses. While [36] and [37] require human annotated or synthetically created human-robot skeleton pairs to learn the retargeting task, [22] proposed to construct a shared representation space that preserves pose semantics without the need of those pairs. In this work, we extend [22] to more complex robots by generating a single latent space from which we can encode or decode different kinematics seemingly. We later use this shared space to predict human-robot interaction motions.

C. Human-Robot Interaction (HRI)

The rapid growth in robotics is leading to their deployment not only in tightly controlled industrial settings but also in more populated and diverse environments. Therefore, there is a need to develop algorithms that understand the

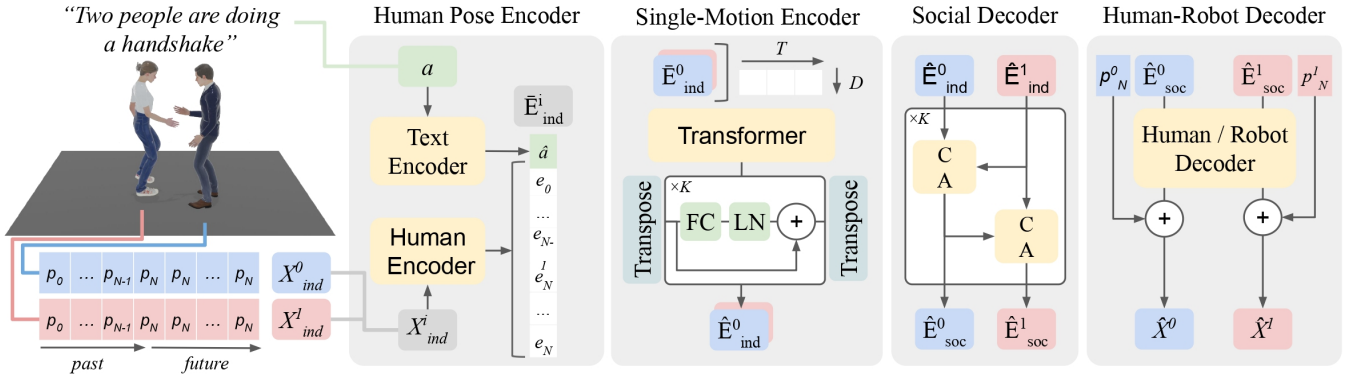

Fig. 2: **Overview of our ECHO architecture**. Our model first focuses on synthesizing individual human motions. First, we pad the observed motion $[\mathbf{p}_t^i, \cdots, \mathbf{p}_N^i]$ for the i-th human by repeating the current pose $\mathbf{p}_N^i$ and obtain $\mathbf{X}_{ind}^i$. As our model is conditioned on the social interaction type a and $\mathbf{X}_{ind}^i$, we encode them both and concatenated them to build $\bar{\mathbf{E}}_{ind}^i$. Then, we forecast our individual motions through a self-attention transformer followed by a Temporal MLP with k layers, such that we obtain a single-motion representation $\hat{\mathbf{E}}_{ind}^i$. As we are considering a social scenario, we iteratively refine the motions per human 0 given the human 1 using cross-attention, and vice versa, obtaining $\hat{\mathbf{E}}_{soc}^0$ and $\hat{\mathbf{E}}_{soc}^1$. This refinement is repeated K times. Finally, we decode each $\hat{\mathbf{E}}_{soc}^i$ and sum the last observed pose $\mathbf{p}_N^i$ to make the model invariant to global translations.

human's intention and assist them in the task at hand [19]–[21], [38]–[40]. Early works [38], [39], [41] proposed to recognize this human motion intent through Hidden Markov Models (HMM) and used learned robot primitives for simple HRI. More recently, [42] focused on the collaborative transportation task through human-robot motion synchronization. However, those works only tackled specific HRI scenarios with close tasks. The success of deep learning opens the doors to properly modeling the dynamic behaviors in HRI [43]. For instance, [44]–[46] proposed to generate natural gestures during the speech, which is essential for embodied conversational robots. While there are clear advances in human motion synthesis conditioned on text [47], [48], it is still a challenge to generate robot behavior that follows the social dynamics established by humans and generalizes to new scenarios. Our work focuses on that problem and proposes to generate a socially compliant full-body robot behavior while preserving the interaction with other humans and generalizing to a high number of interactive scenarios.

III. METHODOLOGY

In this section, we present our ECHO framework for the social motion forecasting task. A visual illustration of our architecture is depicted in Fig. 2.

A. Problem Formulation

Let $\mathbf{p}_t^i = [p_{t,1}^i, \cdots, p_{t,J}^i] \in \mathbb{R}^{J \times n}$ be the pose of a i-th human at time t composed by J joints and $\mathbf{X}^\mathbf{i} = [\mathbf{p}_0^i, \cdots, \mathbf{p}_T^i] \in \mathbb{R}^{T \times J \times n}$ a human motion. Each joint $p_{t,j}^i$ is represented as the standard Euclidean xyz-position for humans ($n = 3$) and joint angle for robot control ($n = 1$). Given a social scenario $\mathbf{S} = [\mathbf{X}^0, \cdots, \mathbf{X}^H]$ with H humans at time N where $0 \leq N \leq T$, the task of social motion forecasting is defined as the prediction of the subsequent motion $\mathbf{X}_{fut}^i = [\mathbf{p}_{N+1}^i, \cdots, \mathbf{p}_T^i]$ per all humans i given their past observations $\mathbf{X}_{past}^i = [\mathbf{p}_0^i, \cdots, \mathbf{p}_N^i]$. In this work, we only consider dyadic situations ($H = 2$). Additionally, we use a global text command a which summarizes the interaction happening in $\mathbf{S}$.

B. Social Motion Forecasting

1) Motion Forecasting as Refinement: Similar to prior works [7], [8], [49], we reformulate the forecasting task by padding the observed motion $\mathbf{X}_{past}^i \in \mathbb{R}^{N \times J \times n}$ to the whole motion dimension T by repeating the last observed pose $\mathbf{X}_{ref}^i = \mathbf{p}_N^i$, obtaining $\mathbf{X}_{ind}^i \in \mathbb{R}^{T \times J \times n}$. Then, our ECHO network learns a function f_θ to refine the $\mathbf{X}_{ind}^i$ from the reference pose $\mathbf{X}_{ref}^i$, such that $\mathbf{X}_{fut}^i = f_\theta(\mathbf{X}_{ind}^i) + \mathbf{X}_{ref}^i$.

2) Pose Encoder: Given $\mathbf{X}_{ind} \in \mathbb{R}^{H \times T \times J \times n}$, we initially flatten the joint parameters and encode each body pose $\mathbf{p}_t^i$ in a D-representation space $\mathbf{E} = [\mathbf{E}_{ind}^0, \cdots, \mathbf{E}_{ind}^H] \in \mathbb{R}^{H \times T \times D}$ using a multi-layer perceptron (MLP). In the case of the robot motion generation, we reuse the pre-trained encoder that we will describe in Section III-D.

3) Single-Motion Encoder: Our single-motion encoder operates on each individual human independently. We use a transformer model [17] to forecast individual human motions conditioned on the observed poses $\mathbf{E}_{ind}^i$ and the overall intention of the interaction a. For that, we initially add a sinusoidal positional embedding to $\mathbf{E}_{ind}^i$ to embed the temporal evolution of the motion. Then, we extract the semantic features of the social intention a using [50], such that $\hat{a} \in \mathbb{R}^D$. We construct the individual motion representation $\bar{\mathbf{E}}_{ind}^i = [\hat{a}, \mathbf{E}_{ind}^i] \in \mathbb{R}^{(T+1) \times D}$ and pass it to a self-attention transformer. Then, inspired by [8], [9] in single-motion forecasting, we adopt k temporal MLP layers to iteratively smooth the output of the individual transformer to $\hat{\mathbf{E}}_{ind}^i \in \mathbb{R}^{\times (T+1) \times D}$ by expanding and compressing the time dimensionality of the output.

4) Multiple motion forecasting: Contrary to prior works [10], [15], [28], [29] that merge multi-person features in the early stage of their architecture, we consider a late-refinement strategy to better preserve the details in a motion.

For that, we adopt a series of two cross-attention layers to refine one subject motion based on the others. 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 $\hat{\mathbf{E}}^0_{ind}$ as **Q** and $\hat{\mathbf{E}}^1_{ind}$ as **K** and **V** for the first cross-attention. The goal is that the resulting motion of the subject *0* ($\hat{\mathbf{E}}^0_{soc}$) has been refined to be compliant with subject *1*. This step is now repeated in the inverse order, being $\hat{\mathbf{E}}^1_{ind}$ as **Q** and $\hat{\mathbf{E}}^0_{soc}$ as **K** and **V**. This dual CA is repeated k times to iteratively enhance each motion to be in synchrony with the other subject in the scene.

5) Pose Decoder: Given $\hat{\mathbf{E}}^0_{soc}$ and $\hat{\mathbf{E}}^1_{soc}$, we decode the representation space to the human or robot pose using an MLP layer. In the case of the robot motion generation, we reuse the pre-trained decoder described in Section III-D.

C. Losses

We consider a weighted sum of four different Mean Square Error (MSE) losses to ensure the naturality and dynamism of our generated motion. Here we formulate $\mathbf{X^i}$ and $\hat{\mathbf{X}}^\mathbf{i}$ as the predicted and ground-truth motion for the i-th human.

1) Single ($\mathscr{L}_{ind}$) and Social ($\mathscr{L}_{soc}$) Skeleton Losses: First, we enforce the single-motion encoder to generate plausible motions using $\mathscr{L}_{ind}$. Then, we ensure a proper motion refinement through $\mathscr{L}_{soc}$. $\mathscr{L}_{ind}$ and $\mathscr{L}_{soc}$ are shown in Eq. 1 and Eq. 2 respectively, where D_H represents the MLP-based human decoder.

$$\mathscr{L}_{ind}(\mathbf{X}^i) = MSE(D_H(\hat{\mathbf{E}}^i_{ind}) - \mathbf{X}^i) \tag{1}$$

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

2) Interaction loss ($\mathscr{L}_{int}$): To ensure that both agents are in spatial synchrony during the interaction, we force the distance between all joints from agent 0 to 1, referred to as Distance Matrix (DM) to be coherent between the predicted and ground-truth motion. We formulate $\mathscr{L}_{int}$ in Eq. 3.

$$\mathscr{L}_{int}(\mathbf{X}^0, \mathbf{X}^1) = MSE(\mathrm{DM}(\hat{\mathbf{X}}^0, \hat{\mathbf{X}}^1) - \mathrm{DM}(\mathbf{X}^0, \mathbf{X}^1)) \tag{3}$$

3) Bone loss ($\mathscr{L}_{bone}$): Given the *xyz*-euclidean representations of the body joints, we use $\mathscr{L}_{bone}$ to ensure predicting human poses that are consistent with the bone lengths. We compute the bone lengths from the predictive body joints and ensure their consistency with the real skeleton using MSE.

D. Human-Robot Retargetting

The human-to-robot retargeting task aims to find a function f that maps a human pose to a semantically close robot pose ($f : \mathbf{p}_{human} \longmapsto \mathbf{p}_{robot}$). We adopt the strategy from [22] but extend it to construct a meaningful latent space that represents various robots. First, we observe that by considering local joint rotations our model can more effectively capture the nuances in more complex kinematic structures. Second, we learn a unique latent space between various robots and a human, which forces the representation to be more meaningful and close to the semantics of different kinematics. Fig. 1 presents a simple scheme of our extended proposal for the human to robot retargeting, where all encoders and decoders are MLP layers. We follow similar losses as [22] for the retargeting process, but only consider $\mathscr{L}_{ind}$ and $\mathscr{L}_{soc}$ for the robot agent, as the output from the decoder from [22] is directly joint angles.

IV. EXPERIMENTS

A. Datasets and Metrics

1) InterGen dataset: InterGen [16] is the largest 3D Human motion dataset encompassing 6022 interactions of two people, accompanied by 16756 natural language annotations. The dataset contains both daily (e.g. handover, greeting, communications) and more professional (e.g. dancing, boxing) interactions. We adopt the skeleton-based configuration to describe a human with 22 joints in the *xyz*-euclidean representation. The forecasting task aims to predict the motion of the next 1.5 sec given an observation of 0.5 sec.

2) Robot retargetting collection: We randomly sample robot joint angles from the Tiago++ robot and the JVRC-1 [1] humanoid robot to train our human-to-robot imitation network.

3) CHICO dataset: CHICO [51] is the only available 3D motion dataset for Human-Robot Collaboration (HRC). The dataset contains a single operator in a smart factory environment performing seven assembly tasks together with a Kuka LBR robot. The goal is to predict the operator motion in the HRC task. We follow the standard evaluation and predict the next 1000 ms given 400 ms in the past.

4) Metrics: We use the same evaluation procedure from prior works in multi-person motion forecasting [10], [12]. Inspired by the Mean Per Joint Position Error (MPJPE) used in single motion forecasting, we adopt Joint Position Error (JPE) to measure millimeter error per global joint position in a given time in the future, and Aligned JPE (AJPE) to only consider the local position error in respect to the root. Equations 4 and 5 represent the JPE and AJPE metrics, where p_r and $\hat{p}_r$ are the estimated and ground-truth root positions of the human body.

$$\mathrm{JPE}(\mathbf{X}, \hat{\mathbf{X}}) = \frac{1}{H \times J} \sum_{i=1}^{H} \sum_{j=1}^{J} ||X^i_j - \hat{X}^i_j||^2, \tag{4}$$

$$\mathrm{AJPE}(\mathbf{X}, \hat{\mathbf{X}}) = \mathrm{JPE}(\mathbf{X} - p_r, \hat{\mathbf{X}} - \hat{p}_r), \tag{5}$$

Additionally, we adopt the Final Displacement Error (FDE) to evaluate the global trajectory of each individual, where $p_{r,t}$ and $\hat{p}_{r,t}$ are the estimated and ground truth root position of the final pose at t-th predicted timestamp.

$$\mathrm{FDE}(\mathbf{X}, \hat{\mathbf{X}}) = ||p_{r,t} - \hat{p}_{r,t}||^2, \tag{6}$$

B. Quantitative Evaluation

1) Social Motion Forecasting: Our model is tested in the InterGen dataset for dyadic motion forecasting in social settings and outperforms on average all state-of-the-art models in multi-person motion forecasting, as shown in Table I. All models reported in the paper have been trained for the InterGen dataset with the same training configuration for a fair comparison. We consider *Zero Velocity* as the repetition

	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	30.92	75.37	121.55	145.89	53.28	135.25	239.80	323.89	39.93	106.35	205.30	292.73
HisRepIt [6]	39.25	68.34	100.64	120.18	56.25	106.69	179.04	234.06	38.92	78.67	147.26	202.8
SocialTGCN [12]	21.56	49.13	86.65	113.65	28.77	65.60	125.48	187.88	15.59	37.38	84.05	146.24
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
ExPI [15]	19.01	56.69	127.01	203.70	**15.30**	**44.41**	85.96	113.83	13.38	37.60	93.15	166.10
TwoBody [11]	**14.90**	38.45	75.37	103.43	20.52	51.29	111.06	177.42	11.97	29.18	75.34	140.09
ECHO (ours)	15.57	**34.37**	**52.11**	**70.15**	20.22	45.01	**73.68**	**110.04**	**11.37**	**25.37**	**48.85**	**80.81**

TABLE I: Performance comparison of ECHO model in InterGen dataset [16]. All results have been trained specifically for the dataset. A lower score is better. Here. bold indicates the best result and underscores the second-best result.

TABLE II: Quantitative evaluation of the short (400ms) and long-term (1000ms) motion forecasting in the CHICO dataset [51] 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
MRS-GCN [4]	54.1	90.7
STS-GCN [52]	53.0	87.4
SeS-GCN [51]	48.8	85.3
ECHO (ours)	**47.1**	**80.5**

of the last pose observed, which acts as the simplest baseline for our evaluation. Additionally, [6] is only focused on single human motion forecasting, which shows the benefit of considering the scene context to refine a given individual motion. On the contrary, [10]–[12], [15] are multi-person motion forecasting models. Table I demonstrates that the use of auto-regressive approaches [11], [15] facilitates a better prediction in the short-term, but fails to capture the long-term dependencies of a model. Moreover, as our model predicts the whole motion in one shot, the inference is much faster, which is crucial for real-world HRI. Our model outperforms the other baselines in all metrics on average. We showcase the performance of our ECHO model in Fig. 4 for the social motion forecasting with human decoding. All models were trained in a single GPU for 150 epochs using an exponential decay scheduler and AdamW as an optimizer, with 5 epochs warmup.

2) Motion Forecasting in Human-Robot Collaboration: Additionally, we train and evaluate our model in the CHICO dataset [51] for the single motion forecasting conditioned by the robot motion and the observed human motion. Due to the different number of joints between the human operator and the robot, we use different MLPs to encode each agent. Similar to the original work [51], we only consider the MPJPE metrics for the short-term (400ms) and long-term (1000ms) horizons. Our results, reported in Table II, show that our ECHO model outperforms previous baselines, mostly in long-term forecasting.

3) Human to Robot Motion: We train the human-to-robot retargeting for only the TIAGo++ robot as [22] and also use our new approach for various robots. Our results have lower reconstruction error in the joint angle (0.005 versus 0.009) while additionally decoding into the JVRC-1 robot. Additionally, we showcase in Figure 3 the qualitative evaluation to demonstrate the effectiveness of our overall ECHO framework in decoding HRI.

Two individuals raise their hands, touch their left hands, and make a small half circle counterclockwise.

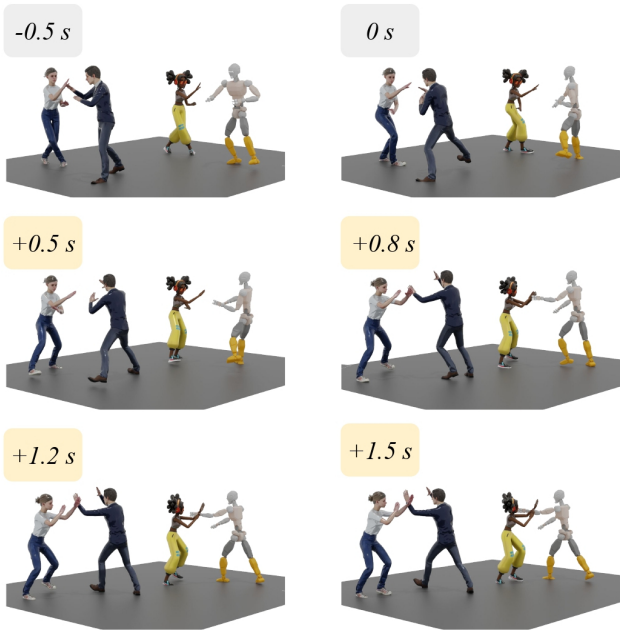

Fig. 3: **Social motion forecasting for Human-Robot Interaction**. Human-Human pair represents the ground truth, while the human-robot pair represents the forecasted human-robot interaction.

C. Ablation Study

This section provides a systematic assessment of the approaches proposed. The results of the ablation study are presented in Table III.

1) Discrete Cosine Transformation (DCT): Contrary to prior models, we decided not to adopt the DCT to avoid over-smooth motion generations. We observed that DCT is beneficial for models when they are not trained in large-scale datasets, as it facilitates generalization but fails to capture the nuances of different motions.

	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/ DCT	17.88	36.52	54.01	72.25	23.40	47.98	75.78	111.63	13.22	27.05	48.80	81.10
w/o TempMLP	17.77	36.59	53.95	72.09	22.63	47.44	75.26	111.07	12.17	26.09	47.94	80.24
w/o Text	19.64	37.39	55.13	73.33	24.53	48.86	76.56	112.43	12.93	26.78	48.56	80.83
w/o Baseline	18.41	39.55	61.35	80.69	23.96	51.91	85.79	126.21	13.60	29.12	55.13	93.09
w/o $\mathscr{L}_{ind}$	16.99	36.97	54.10	71.19	21.57	47.95	75.52	**109.33**	11.95	26.72	48.42	**79.09**
w/o Iterative Refinement	**15.46**	34.47	52.71	70.68	**20.17**	45.25	74.56	111,19	11.39	25.50	48.29	81.74
ECHO (ours)	15.57	**34.37**	**52.11**	**70.15**	20.22	**45.01**	**73.68**	110.04	**11.37**	**25.37**	**47.85**	80.81

TABLE III: Ablation study of our ECHO model for the social motion forecasting task in the InterGen dataset [16].

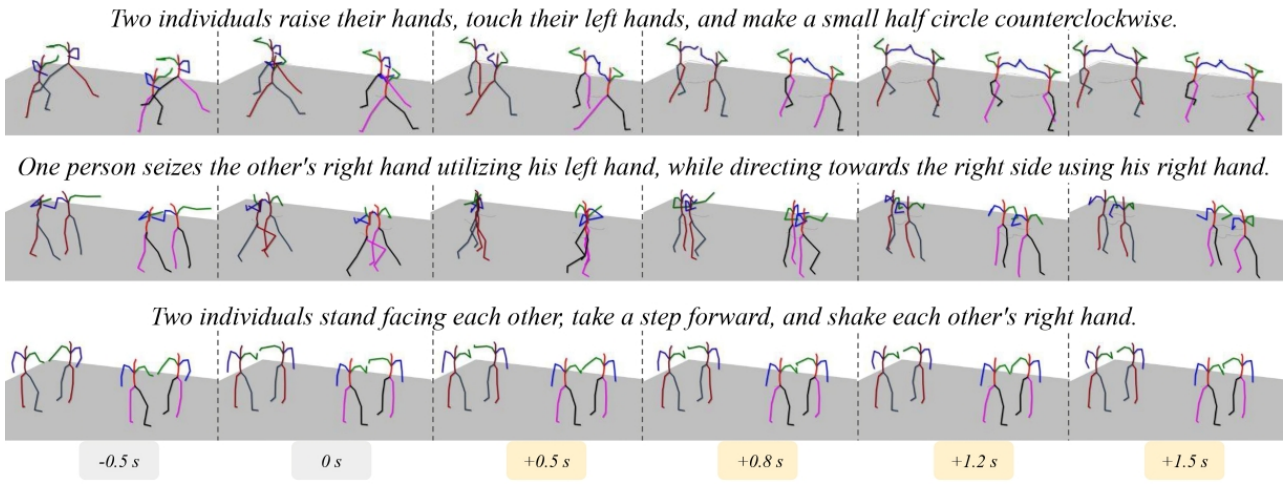

Fig. 4: **Qualitative results for social motion forecasting in the InterGen [16] dataset**. Each scenario shows the ground-truth human pair (left) and the predicted (right) per each time horizon.

2) Text conditioning: While our works benefit from text to condition the motion forecasting to a known social goal, we also asses the performance without this feature for a fair comparison with the baseline models. We observe that using text improves the results, mostly in short-term horizons, but do not suppose a large improvement. Our ECHO model still outperforms by large margins previous baselines despite not using text as a condition.

3) Temporal Multi-Layer Perceptron (TempMLP): The use of TempMLP has also been adopted by prior works in the motion forecasting field, such as [8]–[10]. We observe that TempMLP improves the flow in the short-term horizons as it smoothes out the transition from observed to predicted poses. However, it causes the model to use fixed-length inputs, which tends to reduce the long-term overall performance.

4) Baseline approach: Thanks to our refinement strategy, our ECHO model only needs to learn the variation of the future poses with respect to the last pose observed. Our results reinforce the benefit of this strategy.

5) Individual loss ($\mathscr{L}_{ind}$): The use of $\mathscr{L}_{ind}$ benefit the model in the short-term, as each human focus more on its own motion. However, it slightly reduces the importance of social motion, which is key for better long-term performance.

6) Iterative Refinement: . We assess the iterative refinement of the individual motions by only considering one dual CA ($k = 1$). For a fair evaluation, we extend the number of layers of each CA to k, so the model has the same number of parameters. We observe our iterative refinement improves the forecasting in the long term.

V. CONCLUSIONS

In this paper, we propose a two-step framework that first learns how humans behave in social scenarios to generate a natural Human-Robot Interaction (HRI). First, we build a unique representation space shared between humans and various robot skeletons that preserves the semantics of the poses while facilitating its pose retargeting. Then, we develop a single-to-social transformer architecture, called ECHO, that learns how humans behave in social scenarios through the motion forecasting task. ECHO understands the current scene and synthesizes a natural and meaningful robot motion to interact with a human. Our results support the approaches taken for our framework, which outperforms the state-of-the-art by large margins in the largest dyadic human motion dataset available, as well as in the field of motion forecasting for Human-Robot Collaborative tasks. In conclusion, our approach can decode a compliant robot motion in a social environment, leading to a more natural and accurate Human-Robot Interaction.

ACKNOWLEDGMENT

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

REFERENCES

[1] M. Okugawa, K. Oogane, M. Shimizu, Y. Ohtsubo, T. Kimura, T. Takahashi, and S. Tadokoro, "Proposal of inspection and rescue tasks for tunnel disasters — task development of japan virtual robotics challenge," in *2015 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, pp. 1–2, 2015.

[2] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik, "Recurrent network models for human dynamics," in *IEEE international conference on computer vision*, pp. 4346–4354, 2015.

[3] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena, "Structural-rnn: Deep learning on spatio-temporal graphs," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5308–5317, 2016.

[4] L. Dang, Y. Nie, C. Long, Q. Zhang, and G. Li, "Msr-gcn: Multi-scale residual graph convolution networks for human motion prediction," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11467–11476, October 2021.

[5] W. Mao, M. Liu, M. Salzmann, and H. Li, "Learning trajectory dependencies for human motion prediction," in *International Conference on Computer Vision (ICCV)*, pp. 9489–9497, 2019.

[6] W. Mao, M. Liu, and M. Salzmann, "History repeats itself: Human motion prediction via motion attention," in *European Conference on Computer Vision (ECCV)*, pp. 474–489, 2020.

[7] E. Valls Mascaro, S. Ma, H. Ahn, and D. Lee, "Robust human motion forcasting using transformer-based model," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 10674–10680, 2022.

[8] E. V. Mascaro, H. Ahn, and D. Lee, "A unified masked autoencoder with patchified skeletons for motion synthesis," *arXiv preprint arXiv:2308.07301*, 2023.

[9] W. Guo, Y. Du, X. Shen, V. Lepetit, X. Alameda-Pineda, and F. Moreno-Noguer, "Back to mlp: A simple baseline for human motion prediction," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4809–4819, January 2023.

[10] X. Peng, S. Mao, and Z. Wu, "Trajectory-aware body interaction transformer for multi-person pose forecasting," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17121–17130, June 2023.

[11] M. R. U. Rahman, L. Scofano, E. De Matteis, A. Flaborea, A. Sampieri, and F. Galasso, "Best practices for 2-body pose forecasting," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[12] X. Peng, X. Zhou, Y. Luo, H. Wen, and Z. Wu, "The mi-motion dataset and benchmark for 3d multi-person motion prediction," 2023. arXiv preprint arXiv:2306.13566, 2023.

[13] T. von Marcard, R. Henschel, M. Black, B. Rosenhahn, and G. Pons-Moll, "Recovering accurate 3d human pose in the wild using imus and a moving camera," in *European Conference on Computer Vision (ECCV)*, sep 2018.

[14] J. Wang, H. Xu, M. Narasimhan, and X. Wang, "Multi-person 3d motion prediction with multi-range transformers," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[15] W. Guo, X. Bie, X. Alameda-Pineda, and F. Moreno–Noguer, "Multi-person extreme motion prediction," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13043–13054, 2022.

[16] H. Liang, W. Zhang, W. Li, J. Yu, and L. Xu, "Intergen: Diffusion-based multi-human motion generation under complex interactions," *arXiv preprint arXiv:2304.05684*, 2023.

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

[18] T. Kopp, M. Baumgartner, and S. Kinkel, "Success factors for introducing industrial human-robot interaction in practice: an empirically driven framework," *The International Journal of Advanced Manufacturing Technology*, vol. 112, 01 2021.

[19] M. Chen, S. Nikolaidis, H. Soh, D. Hsu, and S. Srinivasa, "Trust-aware decision making for human-robot collaboration: Model learning and planning," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 9, no. 2, pp. 1–23, 2020.

[20] J. Zhang, H. Liu, Q. Chang, L. Wang, and R. X. Gao, "Recurrent neural network for motion trajectory prediction in human-robot collaborative assembly," *CIRP Annals*, vol. 69, no. 1, pp. 9–12, 2020.

[21] E. V. Mascaro, D. Sliwowski, and D. Lee, "HOI4ABOT: Human-object interaction anticipation for human intention reading assistive roBOTs," in *7th Annual Conference on Robot Learning*, 2023.

[22] Y. Yan, E. V. Mascaro, and D. Lee, "Unsupervised human-to-robot motion retargeting via expressive latent space," 2023. arXiv preprint arXiv:2309.05310, 2023.

[23] J. Martinez, M. J. Black, and J. Romero, "On human motion prediction using recurrent neural networks," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2891–2900, 2017.

[24] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[25] A. Alahi, V. Ramanathan, and L. Fei-Fei, "Socially-aware large-scale crowd forecasting," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[26] J. Amirian, J.-B. Hayet, and J. Pettré, "Social ways: Learning multi-modal distributions of pedestrian trajectories with gans," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[27] V. Adeli, E. Adeli, I. Reid, J. C. Niebles, and H. Rezatofighi, "Socially and contextually aware human motion and pose forecasting," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 6033–6040, 2020.

[28] J. Wang, H. Xu, M. Narasimhan, and X. Wang, "Multi-person 3d motion prediction with multi-range transformers," in *Advances in Neural Information Processing Systems* (M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds.), vol. 34, Curran Associates, Inc., 2021.

[29] E. Vendrow, S. Kumar, E. Adeli, and H. Rezatofighi, "Somoformer: Multi-person pose forecasting with transformers," 2022. arXiv preprint arXiv:2208.14023, 2022.

[30] R. Villegas, D. Ceylan, A. Hertzmann, J. Yang, and J. Saito, "Contact-aware retargeting of skinned motion," 2021. arXiv preprint arXiv:2109.07431, 2021.

[31] J. Zhang, J. Weng, D. Kang, F. Zhao, S. Huang, X. Zhe, L. Bao, Y. Shan, J. Wang, and Z. Tu, "Skinned motion retargeting with residual perception of motion semantics & geometry," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[32] K. Aberman, P. Li, D. Lischinski, O. Sorkine-Hornung, D. Cohen-Or, and B. Chen, "Skeleton-aware networks for deep motion retargeting," *ACM Transactions on Graphics*, vol. 39, no. 4, 2020.

[33] C. Ott, D. Lee, and Y. Nakamura, "Motion capture based human motion recognition and imitation by direct marker control," in *Humanoids 2008 - 8th IEEE-RAS International Conference on Humanoid Robots*, 2008.

[34] W. Gomes, V. Radhakrishnan, L. Penco, V. Modugno, J.-B. Mouret, and S. Ivaldi, "Humanoid whole-body movement optimization from retargeted human motions," in *2019 IEEE-RAS 19th International Conference on Humanoid Robots (Humanoids)*, 2019.

[35] S. Choi and J. Kim, "Towards a natural motion generator: a pipeline to control a humanoid based on motion data," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019.

[36] S. Choi, M. Pan, and J. Kim, "Nonparametric motion retargeting for humanoid robots on shared latent space," 07 2020.

[37] S. Choi, M. J. Song, H. Ahn, and J. Kim, "Self-supervised motion retargeting with safety guarantee," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8097–8103, IEEE, 2021.

[38] D. Lee, C. Ott, and Y. Nakamura, "Mimetic communication model with compliant physical contact in human—humanoid interaction," *The International Journal of Robotics Research*, vol. 29, no. 13, pp. 1684–1704, 2010.

[39] J. Medina Hernández, M. Lawitzky, A. Mörtl, D. Lee, and S. Hirche, "An experience-driven robotic assistant acquiring human knowledge to improve haptic cooperation," pp. 2416–2422, 09 2011.

[40] A. Jevtić, A. Flores Valle, G. Alenyà, G. Chance, P. Caleb-Solly, S. Dogramadzi, and C. Torras, "Personalized robot assistant for support in dressing," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 11, no. 3, pp. 363–374, 2019.

[41] Y. Nakamura, W. Takano, and K. Yamane, "Mimetic communication theory for humanoid robots interacting with humans," in *Robotics Research* (S. Thrun, R. Brooks, and H. Durrant-Whyte, eds.), (Berlin, Heidelberg), pp. 128–139, Springer Berlin Heidelberg, 2007.

[42] L. Yang, Y. Li, and D. Huang, "Motion synchronization in human-robot co-transport without force sensing," in *2018 37th Chinese Control Conference (CCC)*, pp. 5369–5374, 2018.

[43] L. Vianello, L. Penco, W. Gomes, Y. You, S. M. Anzalone, P. Maurice, V. Thomas, and S. Ivaldi, "Human-humanoid interaction and cooperation: a review," *Current Robotics Reports*, vol. 2, no. 4, pp. 441–454, 2021.

[44] P. J. Yazdian, M. Chen, and A. Lim, "Gesture2vec: Clustering gestures using representation learning methods for co-speech gesture generation," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3100–3107, 2022.

[45] T. Ao, Z. Zhang, and L. Liu, "Gesturediffuclip: Gesture diffusion model with clip latents," *ACM Trans. Graph.*

[46] L. Zhu, X. Liu, X. Liu, R. Qian, Z. Liu, and L. Yu, "Taming diffusion models for audio-driven co-speech gesture generation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10544–10553, June 2023.

[47] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-or, and A. H. Bermano, "Human motion diffusion model," in *The Eleventh International Conference on Learning Representations*, 2023.

[48] Y. Yuan, J. Song, U. Iqbal, A. Vahdat, and J. Kautz, "Physdiff: Physics-guided human motion diffusion model," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[49] B. N. Oreshkin, A. Valkanas, F. G. Harvey, L.-S. Ménard, F. Bocquelet, and M. J. Coates, "Motion in-betweening via deep delta-interpolator," *IEEE Transactions on Visualization and Computer Graphics*, 2023.

[50] M. Petrovich, M. J. Black, and G. Varol, "TMR: Text-to-motion retrieval using contrastive 3D human motion synthesis," in *International Conference on Computer Vision (ICCV)*, 2023.

[51] A. Sampieri, G. M. D. di Melendugno, A. Avogaro, F. Cunico, F. Setti, G. Skenderi, M. Cristani, and F. Galasso, "Pose forecasting in industrial human-robot collaboration," in *European Conference on Computer Vision*, pp. 51–69, Springer, 2022.

[52] T. Sofianos, A. Sampieri, L. Franco, and F. Galasso, "Space-time-separable graph convolutional network for pose forecasting," in *IEEE/CVF International Conference on Computer Vision*, pp. 11209–11218, 2021.