IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

HHI-Assist: A Dataset and Benchmark of Human-Human Interaction in Physical Assistance Scenario

Saeed Saadatnejad*, [1], Reyhaneh Hosseininejad*, [1], Jose Barreiros[2], Katherine M. Tsui[2] and Alexandre Alahi[1]

***Abstract*—The increasing labor shortage and aging population underline the need for assistive robots to support human care recipients. To enable safe and responsive assistance, robots require accurate human motion prediction in physical interaction scenarios. However, this remains a challenging task due to the variability of assistive settings and the complexity of coupled dynamics in physical interactions. In this work, we address these challenges through two key contributions: (1) HHI-Assist, a dataset comprising motion capture clips of human-human interactions in assistive tasks; and (2) a conditional Transformer-based denoising diffusion model for predicting the poses of interacting agents. Our model effectively captures the coupled dynamics between caregivers and care receivers, demonstrating improvements over baselines and strong generalization to unseen scenarios. By advancing interaction-aware motion prediction and introducing a new dataset, our work has the potential to significantly enhance robotic assistance policies. The dataset and code are available at: https://sites.google.com/view/hhi-assist/home.**

***Index Terms*—Data Sets for Robot Learning, Physical Human-Robot Interaction, Intention Recognition**

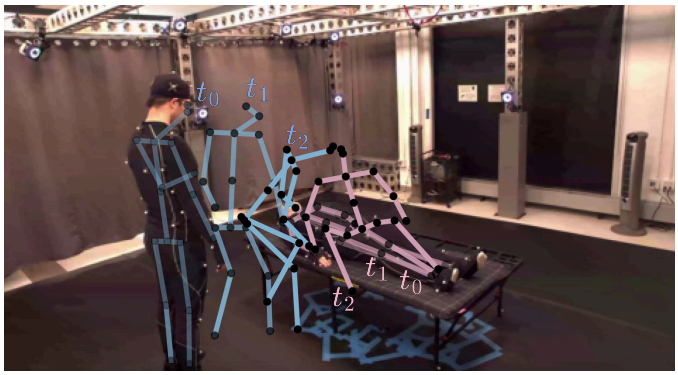

Fig. 1: Predictions of joint positions for the caregiver (blue skeleton) and care receiver (pink skeleton) in a physical assistive task (lay-to-sit transfer), overlaid on a snapshot from a video clip. Future poses are illustrated for three different timesteps. t_0 represents the starting pose, while t_1 and t_2 correspond to two future timesteps with 0.5 seconds of difference.

I. Introduction

THE rising labor shortage and the increasing aging population drive the need for robots capable of assisting humans in need of care [1], [37], [55]. A significant challenge in assisting humans, in contrast to object manipulation, is managing human agency. Ensuring safe physical interactions with humans requires robots to not only manipulate the care receiver's body but also anticipate their movements [21], [34]. Controllers or policies that lack awareness of human motion intention struggle to adapt swiftly to sudden changes in human actions [34], [54]. Hence, the ability to predict human motion during physical interactions is vital for developing effective physically assistive robots. A challenge in motion prediction in these scenarios is the coupled dynamics that emerge from the physical interaction between two agents due to the reciprocal influence of one agent's actions on the other. Take, for example, the task of transferring an elderly person from a laying position to a sitting position (Figure 1). In this setting, the caregiver (CG) approaches the care receiver (CR) and initiates contact, which in turn signals the CR to engage their muscles for the lift. Through contact interaction, the CG senses the amount of effort exerted by the CR and applies sufficient force to support the maneuver. Here, the actions of the CR and CG are interleaved and influence each other.

In this work, we study the problem of motion prediction in physically assistive tasks. Given the scarcity of physical human-robot interaction (HRI) data, we focus on human-human interaction scenarios, where collecting a large dataset is practical using well-established techniques like motion capture. We believe this data and the techniques presented in this work are transferable to physical HRI settings and reserve this for future study.

We collected 908 clips from assistive tasks: sit-to-stand transfer from a chair, lay-to-sit transfer from a bed, lay-to-stand transfer from a bed, and unconstrained movements. Each task includes different sequences of actions or maneuvers. This dataset promises to be a valuable resource for (i) predicting human motion, (ii) data-driven methods that distill control policies for robots, (iii) performance baselines for physical human-robot interaction, and (iv) informing the design of robots that operate in assistive settings.

With this dataset in hand, we aim to predict human movements by predicting future pose sequences based on previously

Manuscript received: December 29, 2024; Revised: March 26, 2025; Accepted: April 17, 2025. This paper was recommended for publication by Editor Angelika Peer upon evaluation of the Associate Editor and Reviewers' comments.

[1]Saeed Saadatnejad, Reyhaneh Hosseininejad and Alexandre Alahi are with VITA laboratory, EPFL, Lausanne, Switzerland (email: {saeed.saadatnejad, reyhaneh.hosseininejad, alexandre.alahi}@epfl.ch).

[2]Jose Barreiros and Katherine M. Tsui are with Toyota Research Institute, Cambridge, MA, USA (email: {jose.barreiros, kate.tsui}@tri.global).

* Equal contribution.

Digital Object Identifier (DOI): see top of this page.

observed poses. This area has attracted considerable interest due to its importance in applications such as autonomous driving [11], human-robot collaboration [12], [51], and robotic navigation [7], [8]. The problem is inherently complex, requiring both spatial and temporal reasoning, and is compounded by the variability of assistance scenarios and human dynamics. In addition to the aforementioned challenges, physical interactions introduce emerging coupled dynamics, which add yet another layer of complexity.

Inspired by the success of Denoising Diffusion Probabilistic Models (DDPMs) [17] in image generation [39] and recently in human pose prediction [44], we extend these methods to motion interaction scenarios in physically assistive tasks. To this end, we propose an interaction-aware denoising diffusion (IDD) model capable of producing realistic and accurate predictions of human poses. To the best of our knowledge, this is the first pose prediction model that considers close contact interactions between agents in physical assistive tasks.

Our model predicts an agent's pose by conditioning not only on its own previous poses but also on the pose of the interacting agent, allowing for dynamic adjustments that reflect the nature of the interaction. An example use case of our model is illustrated in Figure 1, where the predicted joint positions of both the caregiver and the care receiver are shown for three representative timesteps. We validate the effectiveness of our approach through extensive experiments on our dataset.

In summary, our contributions are three-fold:

- We present the HHI-Assist dataset, a collection of motion capture data capturing human-human interaction (HHI) for physical assistance.
- We propose an interaction-aware denoising diffusion (IDD) model that generates realistic and accurate predictions of human poses.
- We perform experiments to evaluate the performance of our model, assess its generalization, and investigate alternative representations for predicting human poses in interactive scenarios.

II. Related Work

While large-scale datasets for single-person motion, such as AMASS [29] and Human3.6M [19], are widely available, datasets capturing human-human interactions remain relatively limited. These datasets can be categorized into the Social Interaction and Close Contact Interaction datasets. Social Interaction datasets primarily focus on non-physical interactions, such as conversations and crowd navigation. For instance, the World-Pose Football dataset [20] and JRDB [31] are utilized mainly for analyzing spatial dynamics and non-contact activities in sports and urban settings. Similarly, 3DPW [52] and EXPI [15] provide insights into multi-person interactions that do not involve direct physical contact, supporting research on weak social interactions [2], [47]. On the other hand, Close Contact datasets [13], [23], [56] capture fine-grained close-contact interactions that involve physical touch or close proximity. Among these, Hi4D [56] and AIR-Act2Act [23] employ markerless motion capture systems and CHI3D [13] uses a semi-marker-based setup in which participants alternate wearing marker suits. Our dataset, HHI-Assist, expands on this category by providing the first marker-based motion capture data specifically designed for physical assistance scenarios. In contrast to existing datasets, HHI-Assist focuses on direct and strong physical interactions, offering a richer resource for studying assistive behaviors.

Predicting future positions of humans at a coarse-grained level, such as predicting center positions [4], [40] or bounding boxes [42], has been extensively studied. Our work, however, focuses on predicting the fine-grained human pose (parameterized as joint positions). Unlike other studies that incorporate additional context, such as action class [6] or extra modalities [41], we limit our focus to the observation pose sequence alone. Our prediction horizon ranges from a few hundred milliseconds to one second, which aligns with typical settings in receding horizon control for high-degree-of-freedom robotic systems [24].

In human pose prediction, early efforts used feed-forward networks [26], and later Recurrent Neural Networks (RNNs) to model the temporal aspects of the task [14], [32]. Later advancements integrated Graph Convolutional Networks (GCNs) to better capture the spatial dependencies of human poses [27], [30]. Subsequently, a unified GCN captured spatio-temporal features [46], and a two-stage GCN refined predictions [28]. More recently, Transformers have demonstrated effectiveness in capturing spatial and temporal dependencies, achieving state-of-the-art performance in human pose prediction. Various Transformer architectures have been explored, including serial spatial and temporal attention blocks [43], parallel spatial and temporal blocks [3], and hybrid models combining both approaches [60]. Our model also leverages the Transformer architecture to capture the spatio-temporal relationship in human pose sequences.

Generative models have lately been utilized to learn the distribution of human motions better. To this end, Generative Adversarial Networks (GANs) [25] and Variational Auto-Encoders (VAEs) [6], [53] have been widely used due to their capability to learn complex data distributions. To promote diversity, various strategies of sampling have been proposed [22], [57]. In addition, uncertainty in human pose prediction has been explored, including approaches that model homoscedastic uncertainty [43], as well as heteroscedastic uncertainty and multimodality through spatial heatmap representations [18].

Building on the success of diffusion models in image generation [39], [58], these models have recently been applied to time-series imputation (i.e., missing value replacement) [48] and human pose prediction [5], [9], [44]. By effectively modeling the data distribution, diffusion models can generate poses that are both realistic and accurate. Our work advances this research direction by introducing a denoising diffusion model specifically tailored for interaction-aware pose prediction. We train and evaluate this model on our dataset and demonstrate its effectiveness in capturing complex interactive dynamics.

III. Dataset

In this section, we present the HHI-Assist dataset.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TABLE I: Details of tasks, maneuvers, unique participant pairs, total takes / demonstrations in the HHI-Assist dataset.

N	Task	Maneuvers	Pairs	Takes
1	Sit-to-stand transfer from a chair	Side Assist	11	173
		Lever	11	160
		Front Assist (Hug)	11	167
2	Lay-to-sit transfer from a bed	Full Assist	12	131
		Side Assist	12	129
		Front Assist (Hug)	13	129
3	Lay-to-stand transfer	Full Assist	1	10
4	Unconstrained	N/A	1	9

A. Data Collection

Data was collected using an Optitrack Motion Capture rig with 20 Optitrack Prime 17W infrared tracking cameras and Motive 3.0 with the skeleton model [35]. This skeleton model includes 21 joints and 20 links, representing an average human. Link dimensions were calibrated to match those of the participants. Support equipment, such as the chair (43.18 L x 44.45 W x 46.99–54.61 H (cm)) and the bed (190.5 L x 99.06 W x 45.72 H (cm)), were also instrumented with markers. Participants wore motion capture suits with 50 reflective markers [36]. Data is saved in files with a BVH format [33] at a frequency of 120 Hz. Data were excluded upon manual inspection when the markers were occluded enough to cause unnatural or infeasible skeleton link overlap behavior (e.g., caregiver's arm crossing through the care receiver's arm).

Participants are healthy individuals from our laboratory who were not involved in the project, aged 21 to 50 years, with an average height of 169 ± 17 cm. Prior experience as a care receiver or caregiver was not required or explicitly considered when recruiting the participants. All participants agreed and signed informed consent and received a non-monetary incentive equivalent to 10 USD/h. Pairs of individuals were drawn from our pool of participants. Each pair was given a live demonstration of the maneuver, acted out by the facilitators, along with a verbal description of the major steps. Participants were then assigned roles (caregiver or care receiver) and instructed to demonstrate the maneuvers. The maneuvers were selected based on demonstrations by an occupational therapist showing various possible ways to assist care receivers with each task. Variations account for differences in care receiver's range of motion, strength, and areas of sensitivity or pain. While the dataset reflects plausible motions used in care scenarios, it does not guarantee that these motions are ergonomically correct or safe for clinical use.

We ensure participant privacy by anonymizing identities and distributing only motion capture data, with all video footage excluded.

B. Dataset Details

Table I summarizes our dataset, which comprises 908 demonstrations spanning four assistive tasks:

1) **Sit-to-stand transfer from a chair:** This task involves transferring the care receiver from a seated position on a chair to a stable standing posture with assistance from the caregiver.
2) **Lay-to-sit transfer from a bed:** This task involves transferring the care receiver from a lying position on a bed to a stable sitting posture, assisted by the caregiver.
3) **Lay-to-stand transfer from a floor:** This task involves the transfer of the care receiver from a supine position on the floor to a standing position with the caregiver's assistance, supported by a bed or chair.
4) **Unconstrained:** In this task, participants were instructed to pantomime or act out various scenarios with or without props. The scenarios include playing with a tennis ball, cooking in the kitchen, performing mirroring exercises (i.e., participants face each other and copy each other's movements without speaking), grooming and dressing tasks, dancing, item handover, ambulation assistance (from a seated position), fighting, and calisthenics/exercise (standing).

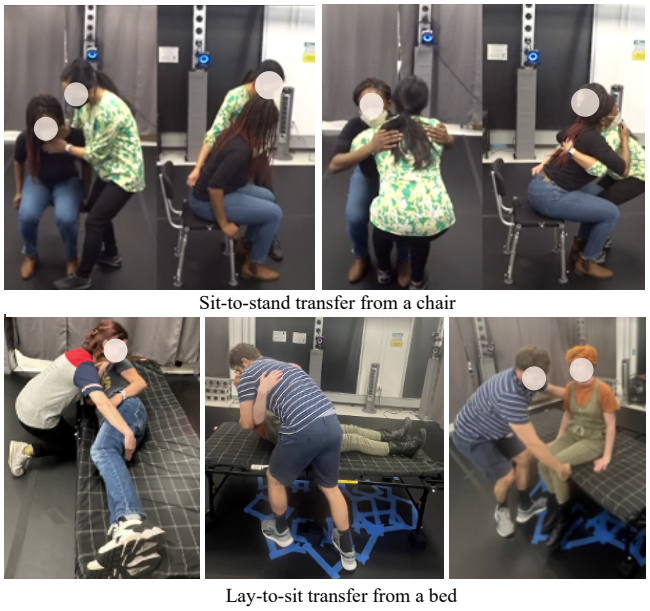

Fig. 2: Examples of physical assistance scenarios from the HHI-Assist dataset. Top: Task 1 (Sit-to-Stand), with Side Assist on the left and Front Assist (Hug) on the right. Bottom: Task 2 (Lay-to-Sit), with Side Assist on the left and Full Assist in the center and on the right.

Tasks 1 and 2 constitute the core of the dataset, with 500 and 389 demonstrations respectively, each performed by 11 to 13 unique participant pairs across three maneuver variations. Figure 2 illustrates sample frames from them, showcasing different maneuver types. On average, sit-to-stand maneuvers lasted 9.4 seconds, while lay-to-sit transfers averaged 18.3 seconds. Task 3 contains 10 demonstrations and is used exclusively for evaluating model generalization. Task 4 includes 9 demonstrations and serves as a source for initialization or data augmentation due to its fundamentally different motions.

The t-SNE visualization [50] of data motion sequences, along with four randomly selected samples, is presented in Figure 3. The plot demonstrates that Task 1 and Task 2 are

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

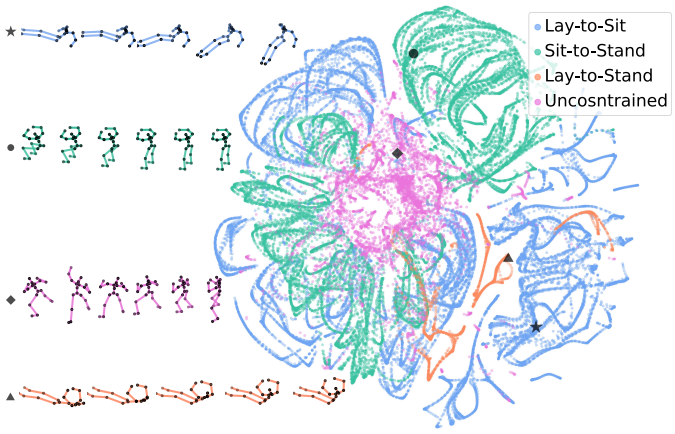

Fig. 3: The t-SNE plot of the HHI-Assist dataset motions showing the separability of the tasks, along with one randomly selected sample from each task.

well-separated in the feature space, while Task 3 shows some overlap with both Task 1 and Task 2. This overlap is influenced by the clipped portion of data, which shares similarities with movements from either Task 1 or Task 2. Additionally, the unconstrained motions (Task 4) are relatively well-separated, which can be attributed to the distinct nature of these motions compared to the constrained tasks.

IV. Method

In this section, we begin by outlining the problem formulation and notations. Next, we detail the architecture of our model. Finally, we discuss the diffusion process in the training and inference of our Interaction-aware Denoising Diffusion (IDD) model.

A. Problem Formulation and Notations

We denote the observed pose sequence of a subject $s \in \{\text{CG}, \text{CR}\}$, where CG stands for the caregiver and GR for care receiver, as

$$X_s = [p_{s_{-O+1}}, p_{s_{-O+2}}, \ldots, p_{s_0}] \in \mathbb{R}^{O \times J \times 3},$$

where J is the number of joints, O is the number of observed timesteps, and each joint pose p is represented as a 3D point in Cartesian space. The corresponding future pose sequence is

$$Y_s = [p_{s_1}, \ldots, p_{s_F}] \in \mathbb{R}^{F \times J \times 3},$$

where F is the number of timesteps to be predicted.

In a high-level view, the model conditions on the concatenated observed sequences $[X_{\text{CG}}, X_{\text{CR}}]$ and predicts future poses $[Y_{\text{CG}}, Y_{\text{CR}}]$.

B. Model Architecture

We extend the Transformer model from previous work [44] to address pose prediction in interactive scenarios. The model, as shown in Figure 4, takes joint positions from two interacting agents and processes them through a 1D convolution layer and a series of residual transformer blocks. Each block applies temporal and spatial multi-head attention to capture the spatio-temporal dynamics of the sequences both between and within agents. The resulting embeddings from these blocks are summed and decoded with a 1D convolution layer to predict future poses. To effectively handle multivariate time series, each residual layer uses a two-dimensional attention mechanism: a temporal transformer layer learns dependencies across time, while a feature transformer layer models relationships among features.

C. The Diffusion Process

Each diffusion process has a noising and denoising part. In the noising part, random Gaussian noise is gradually added to the clean data, transforming it into a pure Gaussian distribution with zero mean and identity covariance after T steps. Mathematically, this can be expressed as:

$$x^t = \sqrt{\alpha_t} x^0 + \sqrt{1 - \alpha_t}\epsilon,$$

where x^t is the noisy data at step t, x^0 is the original clean data, α_t is a time-dependent scaling factor, and $\epsilon \sim \mathcal{N}(0, I)$ is the Gaussian noise. In the denoising part, the model learns to denoise x^T and retrieve the original clean data x^0.

In this work, we focus on the conditional diffusion process where a model can take the extra condition for a more accurate prediction. We tailor this process for our interaction-aware model that receives the observations of both agents as the conditions and is aimed to learn two conditional distributions

$$p(\tilde{Y}^t_{CG} | X_{CG}, X_{CR}) \quad \text{and} \quad p(\tilde{Y}^t_{CR} | X_{CG}, X_{CR}).$$

During training, the model takes a sample of the embedded noisy pose sequences at a diffusion step t by

$$\tilde{Y}^t_s = \sqrt{\alpha_t} Y_s + \sqrt{1 - \alpha_t}\epsilon,$$

as input and learns to predict the added noise at that step ϵ_θ. The loss function can be written as:

$$\mathcal{L} = \mathbb{E}_{Y_s, \epsilon, t} \| \epsilon - \epsilon_\theta(\tilde{Y}^t_s, t \mid X_{CG}, X_{CR}) \|^2_2,$$

where Y_s refers to clean data and θ indicates learnable parameters of the model.

During inference, we begin with a sample drawn from a Gaussian distribution and iteratively denoise it using the trained network, conditioned on the interacting observation. This step-by-step process produces a noise sequence for CG and CR. In the final step of the diffusion process, the model generates the pose sequences by subtracting the last estimated noise from the last noisy input resulting in the predicted pose sequence $\tilde{Y}^0_{CG}$ and $\tilde{Y}^0_{CR}$. In summary, the network's noise estimations define a trajectory that progressively transforms the initial noise into the pose distribution, guided by the interacting observations. In our experiments, we set the diffusion steps $T = 50$.

V. Experiments

A. Dataset and Metric

We use the first two tasks of the HHI-Assist dataset (sit-to-stand transfer from a chair and lay-to-sit transfer from

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

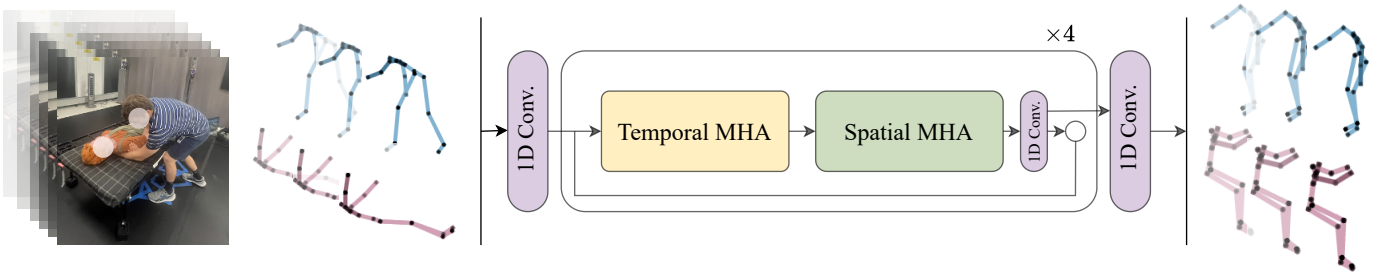

Fig. 4: IDD takes the pose sequences of both CG (blue skeleton) and CR (pink skeleton) as input, predicting their corresponding future pose sequences, shown here for three timesteps. The architecture consists of two 1D convolution layers—one serving as an embedding layer for the inputs and another as the decoder—and four Transformer Blocks with skip connections. Each Transformer Block applies cascaded temporal and spatial multi-head attention to effectively capture spatio-temporal dependencies, followed by a 1D convolution layer.

a bed) for training and testing and consider the third task (lay-to-stand) only for evaluating the generalization of the models. We use data from the fourth task (unconstrained) for the initialization of the models. After downsampling the video clips to a frame rate of 24 fps, each sequence in the dataset contains $O = 24$ observation timesteps and $F = 24$ prediction timesteps, corresponding to 1 second of observed and 1 second of future motion data. The dataset is divided into a training set with $44.8k$ sample sequences, a validation set with $3.5k$ sequences, and a test set with $8.7k$ sequences, with no participant overlap between the test set and the others.

We evaluate accuracy using the Mean Per Joint Position Error (MPJPE) measured in millimeters (mm) per timestep, as well as the overall average MPJPE across all timesteps. This metric averages the Euclidean distance between each predicted keypoint and the ground truth pose for all joints and is calculated after aligning the base joint (pelvis).

B. Baselines

In addition to our Interaction-aware Denoising Diffusion (IDD) model, we implemented the following learning-based and non-learning-based baselines:

1) SiMLPe [16]: A competitive multi-layer perceptron (MLP) network designed for pose prediction.
2) TCD [44]: A denoising diffusion model for single human pose prediction.
3) DSTFormer [60]: A dual-stream spatio-temporal Transformer encoder for learning efficient representations of human poses, adapted here for pose prediction.
4) *Zero-Vel*: A baseline that predicts future poses by assuming no movement, outputting the last observed pose for all future timesteps.
5) *Constant-Vel*: A baseline that assumes constant velocity, predicting future poses by calculating it from the last two observed poses and extrapolating forward.

C. Implementation details

Our model was trained using the Adam optimizer with an initial learning rate of 1e-3, which was reduced by a factor of 0.1 after 75% and 90% of the 50 total epochs. Training was conducted on a single NVIDIA GeForce RTX 3090 GPU (24GB VRAM) and took approximately one day to complete.

D. Results

We present the quantitative results of our model in comparison to the baselines for predicting the poses of both the caregiver and the care receiver at various prediction horizons in Table II. We first observe that predicting the pose of the care receiver is generally easier for all models, as the care receiver exhibits less movement compared to the caregiver. Second, we observe that the interaction-aware model (IDD) significantly outperforms the interaction-unaware models and other baselines for both caregiver and care receiver, highlighting the importance of capturing interaction dynamics and demonstrating its effectiveness.

Figure 5 (a) shows a qualitative comparison of predictions from IDD and the baselines, along with the observation input and ground truth future poses. The models take as input the pose skeleton from t_{-23} to t_0 and are expected to predict future poses close to the ground truth from t_1 to t_{24}. The results reveal that IDD effectively captures the data distribution, producing pose predictions that are realistic, close to the ground truth, and outperform the baselines. We have also provided a 3D visualization of two example frames in Figure 5 (b), generated using the Drake simulator [49].

E. Discussions

1) Generalization: Here, we assess the generalizability of our model to unseen state distribution not included in the training dataset. Specifically, we examine whether our model trained on Task 1 and Task 2, can effectively generalize to Task 3. When evaluated on Task 3, IDD achieves an average MPJPE of 89.3 mm and a long-term MPJPE error at 1000 ms of 154.8 mm for CG and an average MPJPE of 62.5 mm and a long-term MPJPE error at 1000 ms of 113.8 mm for CR. It demonstrates the model's ability to adapt to novel tasks, though improving performance under such distribution shifts remains a direction for future work.

2) Delayed Coupled Dynamics: Modeling the coupled dynamics between agents can be challenging in certain scenarios. Here, we investigate the limit of information transfer between agents, i.e., to determine how much predictive power can be gained if we have prior knowledge of the other person's motion. In this experiment, without collecting new data, we

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TABLE II: Quantitative comparison of pose predictions from the baselines and our model on the HHI-Assist dataset for both caregiver and care receiver in terms of MPJPE (mm) at different prediction horizons. Best performing values across each column are shown in bold.

	Caregiver						Care receiver					
Model	85 ms	330 ms	580 ms	750 ms	1000 ms	average	85 ms	330 ms	580 ms	750 ms	1000 ms	average
Constant-Vel	**5.2**	40.5	89.3	124.1	176.6	80.6	**5.4**	33.7	72.8	100.9	143.9	66.0
Zero-Vel	14.5	64.0	92.9	109.3	123.8	72.7	10.8	39.0	62.8	76.7	95.4	54.7
DSTFormer [60]	8.5	32.7	62.7	81.2	105.0	54.8	8.3	24.4	44.1	57.3	77.9	39.9
siMLPe [16]	6.9	30.9	58.9	76.8	100.2	51.7	6.5	23.1	43.0	56.0	75.7	38.5
TCD [44]	6.9	31.8	60.2	78.1	100.1	52.0	6.4	23.9	42.8	55.5	73.9	37.5
IDD	6.1	**29.8**	**58.6**	**74.6**	**94.0**	**50.4**	5.5	**21.7**	**39.6**	**49.5**	**63.2**	**34.3**

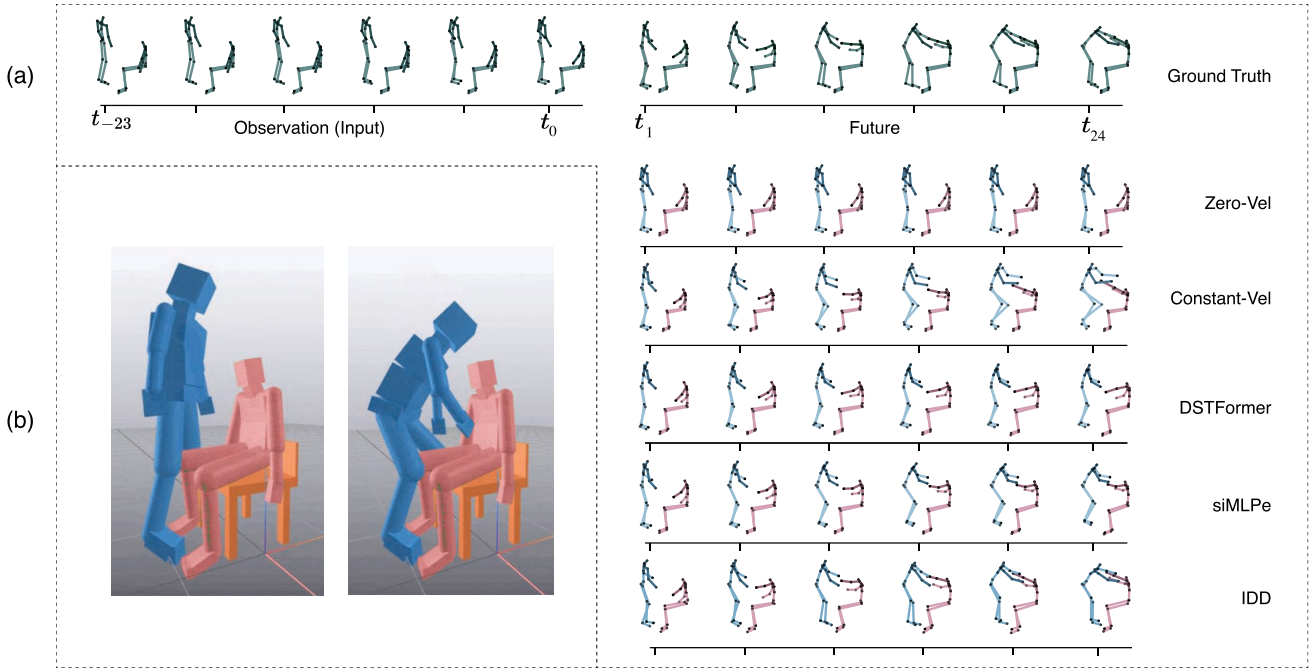

Fig. 5: (a) Qualitative comparison of pose predictions from the baselines and our model on the HHI-Assist dataset shown every four timesteps for both caregiver (blue skeleton) and care receiver (pink skeleton) put together. The models take as input observation the pose skeleton from t_{-23} to t_0 (depicted in the top row in green skeleton) and are supposed to predict future poses close to the ground truth for t_1 to t_{24}. We can observe that IDD predicts poses more accurately. (b) Visualization of two example frames of assisting scenarios on the Drake simulator.

simulate a scenario where one agent waits for 0.5 seconds before responding to the other's movement. Our modified model, called "Delayed IDD", takes as input the observation sequence of CG/CR from $t = 0$ to $t = 1$ and a delayed observation of CR/CG from $t = 0.5$ to $t = 1.5$, and is supposed to predict $t = 1$ to $t = 2$ of CG/CR. This allows each agent to provide a better movement, simplifying the prediction task by reducing the immediate complexity of mutual interaction. Delayed IDD achieves an average MPJPE of 47.6 mm and a long-term MPJPE at 1000 ms of 89.9 mm for CG, and 32.5 mm and 60.0 mm, respectively, for CR. The enhanced prediction accuracy for both agents indicate that the Delayed IDD model can better anticipate each agent's subsequent movements by decoupling their interactions to some extent.

3) Other Representations: In addition to joint positions, we study the impact of using joint angles as an alternative representation. Joint angles describe the relative orientation between connected body links, offering a potentially more compact and invariant representation of human motion. They can be particularly advantageous when maintaining fixed link lengths is important. However, discontinuities in some angular representations, such as Euler angles and quaternions, make them difficult to learn [45], [59]. In this work, we focus on rotation matrices, though they are not the most compact representation. We leave exploration of alternative angle-based representations for future work.

To compare representations, we trained our model twice—once using joint positions and once using joint angles—and evaluated their prediction accuracies. As shown in Table III, the model trained with joint positions achieved better performance in terms of MPJPE, , likely due to better alignment with the loss and evaluation metric. In contrast, the model trained with joint angle representation showed around 7% higher average MPJPE but ensured consistent link lengths throughout the predictions. Note that by applying the forward kinematics, we verified that the predicted rotation matrices are

TABLE III: Comparing the performance of IDD given joint position and joint angle representations in terms of MPJPE (mm) at 1000 ms and average across all prediction horizons.

Representation	Caregiver 1000 ms	Caregiver average	Care receiver 1000 ms	Care receiver average
Joint Position	94.0	50.4	63.2	34.3
Joint Angle	99.9	54.0	67.6	36.8

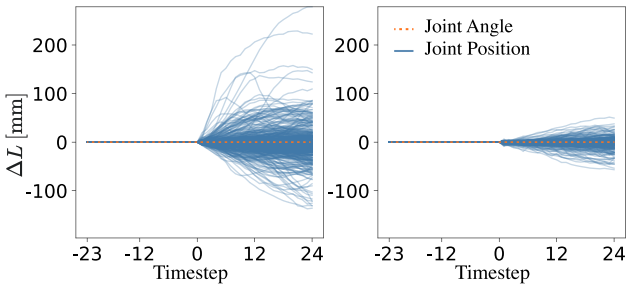

Fig. 6: Comparison of link length variations in predictions using joint position and joint angle representations. The plots show changes in link lengths across different timesteps for 500 randomly selected motions. The left plot refers to the forearm, and the right plot refers to the neck. Predicting joint positions results in link length variations, whereas predicting joint angles preserves consistent link lengths.

members of the SO(3) manifold, i.e., valid rotation matrices. For the joint position-based model, the mean absolute and standard deviation of link length changes across the test set were 7.1 mm and 14.84 mm, respectively. Figure 6 shows length variation of two representative links across all timesteps (both observed and predicted) for 500 random examples. These results confirm that joint positions lead to variations in link lengths, with some samples exhibiting significant changes, whereas joint angles preserve consistent link lengths.

VI. Conclusions and Future Works

In this work, we presented the HHI-Assist dataset, a collection of caregiver and care receiver demonstrations in physical assistance scenarios. We proposed an interaction-aware denoising diffusion model for predicting human poses, achieving significant improvements in prediction accuracy by accounting for the dynamic interactions between caregivers and care receivers. We hope our work paves the way for more advanced interaction-aware motion prediction, ultimately enhancing downstream tasks such as robotic assistance policies and improving care for individuals in need of support.

Future work will focus on exploring alternative input representations and addressing various sources of uncertainty in the task. We also plan to investigate the model's potential for robot control, such as integrating predictions into receding horizon controllers (e.g., model predictive controllers) or incorporating them into the observation space for learned policies. Beyond pose prediction, the HHI-Assist dataset enables new research avenues, particularly in policy learning for physical human-robot interaction. The dataset can be retargeted to robot kinematics, providing valuable training data for behavior cloning techniques such as Diffusion Policy [10]. It can also serve as a source of style data for reinforcement learning methods, such as Adversarial Motion Priors [38]. Lastly, the dataset allows for deriving performance metrics for assistive tasks, including evaluations of robot performance in terms of contact sequences, force ranges, and kinematic limits.

Acknowledgments

The authors would like to thank Eric Dusel, Bisi Chikwendu, and Andrew Silva of Toyota Research Institute, and Saged Bounekhel of EPFL. E.D. and B.C. assisted with data collection, A.S. provided valuable feedback, and S.B. contributed to preliminary experiments.

References

[1] Daron Acemoglu and Pascual Restrepo. Demographics and automation. *The Review of Economic Studies*, 89(1):1–44, 06 2021. 1

[2] Vida Adeli, Ehsan Adeli, Ian Reid, Juan Carlos Niebles, and Hamid Rezatofighi. Socially and contextually aware human motion and pose forecasting. *IEEE Robotics and Automation Letters (RA-L)*, 2020. 2

[3] Emre Aksan, Manuel Kaufmann, Peng Cao, and Otmar Hilliges. A spatio-temporal transformer for 3d human motion prediction. In *International Conference on 3D Vision*. IEEE, 2021. 2

[4] Mohammadhossein Bahari, Saeed Saadatnejad, Amirhossein Askari Farsangi, Seyed-Mohsen Moosavi-Dezfooli, and Alexandre Alahi. Certified human trajectory prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2

[5] German Barquero, Sergio Escalera, and Cristina Palmero. Belfusion: Latent diffusion for behavior-driven human motion prediction. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2

[6] Yujun Cai, Yiwei Wang, Yiheng Zhu, Tat-Jen Cham, Jianfei Cai, Junsong Yuan, Jun Liu, Chuanxia Zheng, Sijie Yan, Henghui Ding, et al. A unified 3d human motion synthesis model via conditional variational auto-encoder. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[7] Changan Chen, Sha Hu, Payam Nikdel, Greg Mori, and Manolis Savva. Relational graph learning for crowd navigation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020. 2

[8] Changan Chen, Yuejiang Liu, Sven Kreiss, and Alexandre Alahi. Crowd-robot interaction: Crowd-aware robot navigation with attention-based deep reinforcement learning. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2019. 2

[9] Ling-Hao Chen, Jiawei Zhang, Yewen Li, Yiren Pang, Xiaobo Xia, and Tongliang Liu. Humanmac: Masked motion completion for human motion prediction. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2

[10] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems*, 2023. 7

[11] Xiaoxiao Du, Ram Vasudevan, and Matthew Johnson-Roberson. Bio-lstm: A biomechanically inspired recurrent neural network for 3-d pedestrian pose and gait prediction. *IEEE Robotics and Automation Letters (RA-L)*, 2019. 2

[12] Nuno Ferreira Duarte, Mirko Raković, Jovica Tasevski, Moreno Ignazio Coco, Aude Billard, and José Santos-Victor. Action anticipation: Reading the intentions of humans and robots. *IEEE Robotics and Automation Letters (RA-L)*, 2018. 2

[13] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Reconstructing three-dimensional models of interacting humans. *CoRR*, 2023. 2

[14] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2015. 2

[15] Wen Guo, Xiaoyu Bie, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Multi-person extreme motion prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[16] Wen Guo, Yuming Du, Xi Shen, Vincent Lepetit, Alameda-Pineda Xavier, and Moreno-Noguer Francesc. Back to mlp: A simple baseline for human motion prediction. *arXiv:2207.01567*, 2022. 5, 6

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arxiv:2006.11239*, 2020. 2

[18] Reyhaneh Hosseininejad, Megh Shukla, Saeed Saadatnejad, Mathieu Salzmann, and Alexandre Alahi. Motionmap: Representing multimodality in human pose forecasting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2

[19] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 2014. 2

[20] Tianjian Jiang, Johsan Billingham, Sebastian Müksch, Juan Zarate, Nicolas Evans, Martin Oswald, Marc Pollefeys, Otmar Hilliges, Manuel Kaufmann, and Jie Song. Worldpose: A world cup dataset for global 3d human pose estimation. *European Conference on Computer Vision (ECCV)*, 2024. 2

[21] Jie Kang, Kai Jia, Fang Xu, Fengshan Zou, Yanan Zhang, and Hengle Ren. Real-time human motion estimation for human robot collaboration. In *International Conference on CYBER Technology in Automation, Control, and Intelligent Systems*. IEEE, 2018. 1

[22] Hee Jae Kim and Eshed Ohn-Bar. Motion diversification networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[23] Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, and Jaehong Kim. Air-act2act: Human–human interaction dataset for teaching non-verbal social behaviors to robots. *International Journal of Robotics Research*, 40(4-5):691–697, 2021. 2

[24] Jonas Koenemann, Andrea Del Prete, Yuval Tassa, Emanuel Todorov, Olivier Stasse, Maren Bennewitz, and Nicolas Mansard. Whole-body model-predictive control applied to the hrp-2 humanoid. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015. 2

[25] Jogendra Nath Kundu, Maharshi Gor, and R Venkatesh Babu. Bihmp-gan: Bidirectional 3d human motion prediction gan. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8553–8560, 2019. 2

[26] Chen Li, Zhen Zhang, Wee Sun Lee, and Gim Hee Lee. Convolutional sequence to sequence model for human dynamics. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[27] Zhenguang Liu, Pengxiang Su, Shuang Wu, Xuanjing Shen, Haipeng Chen, Yanbin Hao, and Meng Wang. Motion prediction using trajectory cues. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[28] Tiezheng Ma, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Progressively generating better initial guesses towards next stages for high-quality human motion prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[29] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of motion capture as surface shapes. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2

[30] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2

[31] Roberto Martin-Martin, Mihir Patel, Hamid Rezatofighi, Abhijeet Shenoi, JunYoung Gwak, Eric Frankel, Amir Sadeghian, and Silvio Savarese. Jrdb: A dataset and benchmark of egocentric robot visual perception of humans in built environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 45(6):6748–6765, 2021. 2

[32] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[33] Maddock Meredith, Steve Maddock, et al. Motion capture file formats explained. *Department of Computer Science, University of Sheffield*, 211:241–244, 2001. 3

[34] Stanley Mugisha, Vamsi Krishna Guda, Christine Chevallereau, Damien Chablat, and Matteo Zoppi. Motion prediction with gaussian processes for safe human-robot interaction in virtual environments. *IEEE Access*, 2024. 1

[35] OptiTrack. Prime 17w. https://optitrack.com/cameras/prime-17w/. Accessed: 2024-06-23. 3

[36] OptiTrack. Skeleton marker set: Core (50), 2023. 3

[37] Srikanta Padhan, Avilash Mohapatra, Senthil Kumar Ramasamy, and Sanjana Agrawal. Artificial intelligence (ai) and robotics in elderly healthcare: enabling independence and quality of life. *Cureus*, 15(8), 2023. 1

[38] Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. Amp: Adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics (ToG)*, 40(4):1–20, 2021. 7

[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[40] Saeed Saadatnejad, Mohammadhossein Bahari, Pedram Khorsandi, Mohammad Saneian, Seyed-Mohsen Moosavi-Dezfooli, and Alexandre Alahi. Are socially-aware trajectory prediction models really socially-aware? *Transportation Research Part C: Emerging Technologies*, 2022. 2

[41] Saeed Saadatnejad, Yang Gao, Kaouther Messaoud, and Alexandre Alahi. Social-transmotion: Promptable human trajectory prediction. In *International Conference on Learning Representations (ICLR)*, 2024. 2

[42] Saeed Saadatnejad, Yi Zhou Ju, and Alexandre Alahi. Pedestrian 3d bounding box prediction. In *Symposium of the European Association for Research in Transportation*, 2022. 2

[43] Saeed Saadatnejad, Mehrshad Mirmohammadi, Matin Daghyani, Parham Saremi, Yashar Zoroofchi Benisi, Amirhossein Alimohammadi, Zahra Tehraninasab, Taylor Mordan, and Alexandre Alahi. Toward reliable human pose forecasting with uncertainty. *IEEE Robotics and Automation Letters (RA-L)*, 2024. 2

[44] Saeed Saadatnejad, Ali Rasekh, Mohammadreza Mofayezi, Yasamin Medghalchi, Sara Rajabzadeh, Taylor Mordan, and Alexandre Alahi. A generic diffusion-based approach for 3d human pose prediction in the wild. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2023. 2, 4, 5, 6

[45] Ashutosh Saxena, Justin Driemeyer, and Andrew Y Ng. Learning 3-d object orientation from images. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2009. 6

[46] Theodoros Sofianos, Alessio Sampieri, Luca Franco, and Fabio Galasso. Space-time-separable graph convolutional network for pose forecasting. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11209–11218, 2021. 2

[47] Julian Tanke, Linguang Zhang, Amy Zhao, Chengcheng Tang, Yujun Cai, Lezi Wang, Po-Chen Wu, Juergen Gall, and Cem Keskin. Social diffusion: Long-term multiple human motion anticipation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2

[48] Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

[49] Russ Tedrake and the Drake Development Team. Drake: Model-based design and verification for robotics, 2019. 5

[50] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research (JMLR)*, 2008. 3

[51] Lorenzo Vianello, Jean-Baptiste Mouret, Eloise Dalin, Alexis Aubry, and Serena Ivaldi. Human posture prediction during physical human-robot interaction. *IEEE Robotics and Automation Letters (RA-L)*, 2021. 2

[52] Timo von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)*. Springer, 2018. 2

[53] Jacob Walker, Kenneth Marino, Abhinav Gupta, and Martial Hebert. The pose knows: Video forecasting by generating pose futures. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017. 2

[54] Dominik Widmann and Yiannis Karayiannidis. Human motion prediction in human-robot handovers based on dynamic movement primitives. In *European Control Conference (ECC)*. IEEE, 2018. 1

[55] Tammy Worth. Are robots the solution to the crisis in older-person care? *Nature*, 2024. 1

[56] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Jie Song, and Otmar Hilliges. Hi4d: 4d instance segmentation of close human interaction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[57] Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *European Conference on Computer Vision (ECCV)*. Springer, 2020. 2

[58] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2

[59] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 6

[60] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2, 5, 6