arXiv:2311.12943v4 [cs.RO] 2 Jun 2024

INTERACT: Transformer Models for Human Intent Prediction Conditioned on Robot Actions

Kushal Kedia[1], Atiksh Bhardwaj[1], Prithwish Dan[1], Sanjiban Choudhury[1]

***Abstract*— In collaborative human-robot manipulation, a robot must predict human intents and adapt its actions accordingly to smoothly execute tasks. However, the human's intent in turn depends on actions the robot takes, creating a chicken-or-egg problem. Prior methods ignore such inter-dependency and instead train *marginal* intent prediction models independent of robot actions. This is because training *conditional* models is hard given a lack of paired human-robot interaction datasets.**

Can we instead leverage large-scale human-human interaction data that is more easily accessible? Our key insight is to exploit a correspondence between human and robot actions that enables transfer learning from human-human to human-robot data. We propose a novel architecture, INTERACT, that pre-trains a conditional intent prediction model on large human-human datasets and fine-tunes on a small human-robot dataset. We evaluate on a set of real-world collaborative human-robot manipulation tasks and show that our conditional model improves over various marginal baselines. We also introduce new techniques to tele-operate a 7-DoF robot arm and collect a diverse range of human-robot collaborative manipulation data which we open-source. We release our code and datasets at https://portal-cornell.github.io/interact/.

I. INTRODUCTION

If robots are to work alongside human partners to achieve shared goals, they need models for how to coordinate with humans. Such coordination is dependent on understanding the human partner's intent and predicting how these intents might change in response to the robot's actions [1]. Consider the shared human-robot manipulation task in Fig. 1 where a human and a robot are simultaneously reaching for objects on a shelf. The robot needs to predict the human's intent, i.e., which object they are reaching for, to safely and confidently reach for a different object. However, the human's intent in turn depends on the action the robot takes in the future. This cyclic dependency between human intent and robot actions presents a non-trivial chicken-or-egg problem. We tackle the problem in this paper by training intent prediction models that condition on future robot actions.

There's been a lot of recent focus on intent prediction for collaborative manipulation [2]–[5], including approaches [6] that leverage large-scale human-activity datasets [7], [8]. Nevertheless, these models predominantly operate in a *marginal* framework, without conditioning on future robot actions. Such an approach can yield sub-optimal outcomes; consider again the scenario illustrated in Fig. 1. An unconditioned model may estimate that the human has an equal likelihood of reaching for either object on the shelf. Consequently, the robot may deduce that it is unsafe to proceed with reaching for any object.

[1]Department of Computer Science, Cornell University

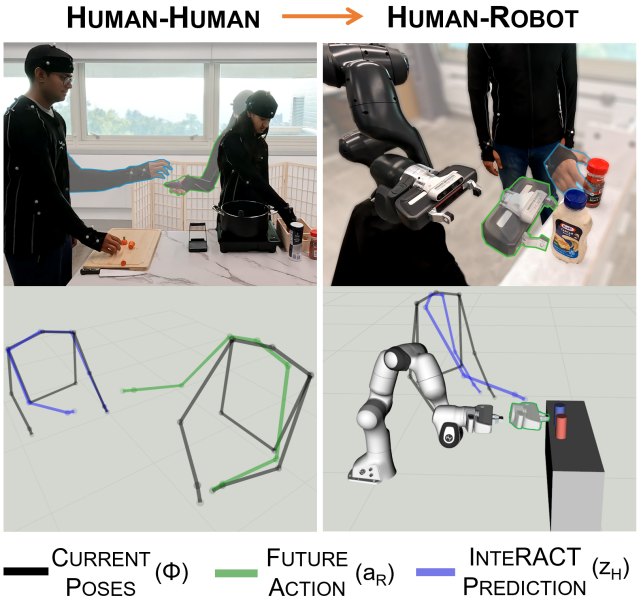

Fig. 1: We present INTERACT, a model that predicts future human intent conditioned on the future robot action. **Left:** When a human passes an object over, INTERACT conditions on the future object handover action of one human and predicts that the other human will move towards it. **Right:** In this human-robot interaction, given the robot's plan to reach for the can on the right, INTERACT predicts the human will reach for the pepper. We transfer a model trained on human-human interactions to human-robot interactions.

Conditional transformer models show promise in overcoming such issues and have been successfully used in self-driving [9]–[13] to model dependencies between road agents and forecast their joint behaviors. Such models require extensive human-generated driving data [14], [15]. However, adapting such methods to the domain of human-robot collaborative manipulation is not straightforward due to a key obstacle: the scarcity of large-scale human-robot interaction datasets for training. Acquiring such datasets, even on a smaller scale, poses its own challenges, given the complexity of teleoperating 7-DoF robot arms. The question then arises: can we capitalize on the readily available, large-scale human-human interaction data?

Our key insight lies in leveraging the correspondence between human and robot actions to facilitate transfer learning from human-human to human-robot interactions. For example, in common manipulation tasks such as object handovers, humans often discern each other's intentions by observing arm and hand movements. We hypothesize that

human reactions to robot arm movements exhibit similar patterns, allowing for the effective transfer of learned models.

We propose a novel architecture, **INTERACT** (**Inte**nt Prediction via **R**obot **A**ction-**C**onditioned **T**ransformer) that can predict a human's intent based on the robot's planned future action. Our model is trained in two stages. First, we utilize large sources of both single and multi-human interaction data, where our model predicts human intent conditioned on the future action of the other human in the scene (Fig 1). Then, we exploit a low-level correspondence between the human's hand and the robot end-effector to tele-operate a 7-DoF Franka Emika robot arm alongside a human partner. This collected Human-Robot dataset contains human-robot interaction data as well as the corresponding motion data of the human tele-operating the human arm. We utilize this pairing to align human and robot representations for effective transfer learning. Our key contributions are:

1) We introduce a novel transformer-based architecture that conditions on robot actions to predict human intent.
2) We propose a technique to collect a paired human-robot dataset via tele-operation for fine-tuning models with aligned representations and open-source a first dataset of human-robot collaborative manipulation.
3) Our prediction model demonstrates improved human intention prediction on multiple real-world datasets of human-human and human-robot interaction.

II. Related Work

Predicting Human Intent for Navigation. Human intent prediction has been extensively studied in social navigation. Research in this domain has focused on developing better input and output representations for modeling inter-agent interactions [16]–[18]. The multi-modality of human intents can be captured by generative neural network architectures such as Trajectron++ [19]–[22]. Yet, these works are largely independent of robot actions and predict human intents independently for each agent in the scene without enforcing joint future consistency. Attempts have been made to develop joint forecasting and planning frameworks [23], [24]. Our work is inspired by recent progress in the self-driving domain; transformer models have been applied to predict the joint futures of agents in the scene by conditioning the robot's future actions [9]–[13]. While such an approach is feasible in self-driving where large-scale interaction datasets [14], [15] are readily available, it is difficult to collect human-robot collaboration data. We introduce new techniques to collect a dataset of human-robot interactions to fine-tune forecasting models trained on human-human interactions.

Human Pose Prediction. In this work, we represent human intention as a trajectory of human pose predictions, which is a challenging problem due to the wide range of possible human joint movements. Recently, the release of large-scale datasets of human motion [7], [8], has made this problem more tractable, leading to rapid progress in this field. A number of approaches have been proposed to model the spatio-temporal interaction between human joints using Graph Neural Networks and Transformers [25], [26]. However, such approaches are limited to predicting future motion for just one human. Recent works have extended pose prediction from single-person to multi-person settings. SoMoFormer [27] uses a transformer architecture that can accept any joint embedding as a query, allowing the model to learn interactions between all joints in the scene, including from different humans. Multi-Range Transformers (MRT) [28] utilizes a combination of a local encoder to learn temporal dependencies between a single agent's body pose and a global encoder to learn dependencies with other agents in the scene. The small sizes of existing human-human interaction datasets limit these approaches. In this work, we extend existing human-human interaction datasets and additionally collect a dataset of human-robot collaboration.

Predicting Human Intent for Collaborative Manipulation. Human-robot collaboration requires representing future human intents in some form. Many approaches consider the human to be completely static [29]–[32]. Others reason about the future motion of specific human joints such as the wrist, hand, or head [33]–[35]. Similar to social navigation, almost all human intent predictors in the context of collaborative manipulation [2]–[5], [36]–[38] generate marginal predictions independent of robot actions. Mainprice et al. [39] use motion capture data of two humans engaged in a collaborative task within a shared workspace to predict single-arm reaching motions. However, they do not use any human-robot data to align human and robot representations.

Human-Robot Correspondence. Our objective is to collect a dataset of paired human-robot interaction. However, it is challenging to program a reliable robot policy that can operate safely with humans. Besides, we require a method to transfer prediction models trained on human-human data to our collected data. When robot and human morphologies match, as in the case of humanoid robots [40]–[42] or dexterous hands [43], [44], human motion can be directly imitated. While robotic arm joint movements differ from human joints, a lower-dimensional mapping must be created [45]. In this work, we define a correspondence between the human hand and wrist joints and the robot's end effector to be able to control robots for collecting interaction data. We detect the 6-DoF pose of the human hand using an OptiTrack motion capture suit for robust pose detection amidst occlusions in the environment. Then, we map the pose to the robot's end-effector and control it using an IK-based controller. Similar to the human-robot alignment function used by MimicPlay [46], an imitation learning framework from human demonstrations, we use the pairing between the robot's motion and the tele-operating human's motion to transfer our prediction model.

III. Problem Formulation

Marginal Intent Prediction. We begin by modeling the marginal intent prediction problem as predicting the human's intent given the scene context. For simplicity[1], we assume

[1]Our framework can easily be extended to handle multiple humans, a richer context that includes environment information and visual cues, and alternate intent definitions that are a function of the observations.

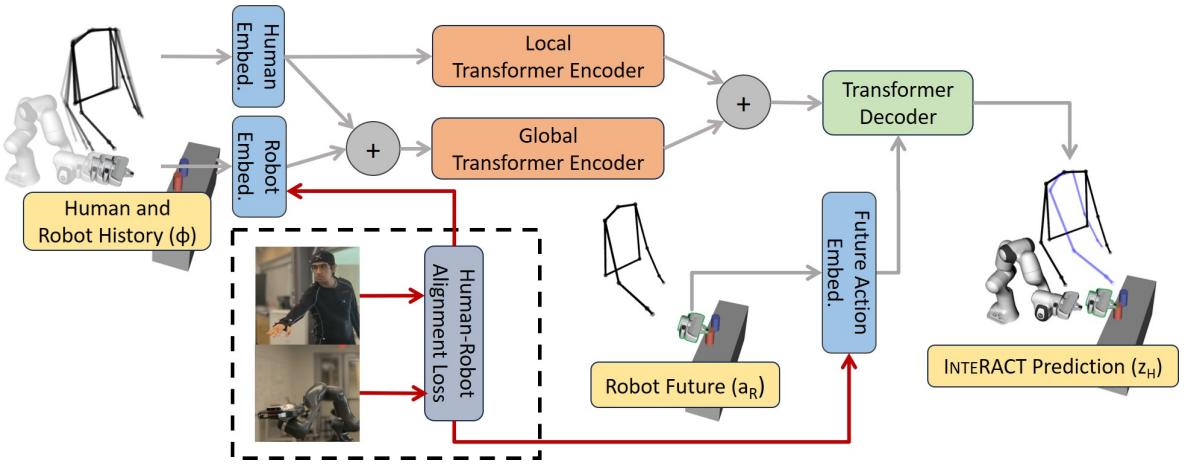

Fig. 2: **INTERACT Model Architecture**. The scene history ϕ is encoded by the local and global transformer encoders. The future action a_R of the robot is passed as a query to the transformer decoder to generate an **action-conditioned** human intent prediction z_H. The robot pose embeddings are aligned with paired human pose embeddings via an **alignment loss**.

there is a single human H interacting with a robot R.

We define the human's ***intent*** z_H as a T-horizon sequence of future poses, i.e., $z_H = \{s^H_1, s^H_2, \ldots, s^H_T\}$, where $s^H_t \in \mathbb{R}^d$ is the d-dimensional human pose at future timestep t. For many tasks, this information is sufficient for the robot to plan it's future actions. We define ***context*** ϕ as salient information in the scene to predict the human intent. We set context as the current and past history of human states $\phi = \{s^H_{-T+1}, s^H_{-T+2}, \ldots, s^H_0\}$. We define ***marginal intent prediction*** model $P_\theta(z_H|\phi)$ as predicting human intent z_H conditioned only on the context ϕ, where θ are parameters of the model. Notably, the predictions are independent of the robot R. We train this model via Maximum Likelihood Estimation (MLE) on observed human motions.

$$\max_\theta \ \mathbb{E}_{\phi, z_H} \log P_\theta(z_H|\phi) \tag{1}$$

Action-Conditioned Intent Predictions. We hypothesize that marginal model $P_\theta(z_H|\phi)$ is insufficient for accurate intent prediction in close-proximity interactions and requires conditioning on robot actions. We define the robot's ***action*** $a_R \in \mathbb{R}^j$ to be its planned goal location denoted as the $j-$ dimensional robot goal pose. For instance in Fig. 1, the current poses (and a 1s history) of both the human and robot represents ϕ, the robot's planned future end-effector position is denoted by a_R, and the future human intent z_H is its future upper body pose. We define an ***action-conditioned intent prediction*** model $P_\theta(z_H|\phi, a_R)$ as predicting human intent z_H conditioned on both the context ϕ and the robot action a_R. We augment the context to also include the past history of robot states, $\phi = \{s^H_{-T+1}, \ldots, s^H_0, s^R_{-T+1}, \ldots, s^R_0\}$, where $s^R_t \in \mathbb{R}^j$ is robot pose at time t. We also train this model via a similar MLE objective:

$$\max_\theta \ \mathbb{E}_{\phi, z_H} \log P_\theta(z_H|\phi, a_R) \tag{2}$$

However, optimizing the objective above poses an important practical challenge - collecting large-scale paired human-robot interaction data is costly. It requires humans to work around robots that are already planning reasonable motions. It is also non-trivial to make use of existing public human-human motion datasets to aid in this task. We address both of these challenges next in our approach.

IV. Approach

We present **INTERACT** (**Inte**nt Prediction via **R**obot **A**ction-**C**onditioned **T**ransformer), a framework for predicting human intent conditioned on future robot actions for collaborative manipulation. At train time, we first pre-train a conditional intent prediction model on human-human interaction data combining publicly available datasets and task specific datasets that we collect. We then fine-tune this model on a small scale human-robot dataset where we predict human intent conditioned on robot actions. Our approach has two main features: (1) an alignment loss between human and robot representations to allow transfer between domains (2) a new tele-operation technique to control a 7-DoF robot arm for paired human-robot interaction.

A. Data: Collecting Paired Human-Robot Interaction

We make use of large-scale single-human activity data (AMASS [7]) as well as extend the human-human dataset in CoMaD [6] as our source of human-human interaction data. In order to transfer our action-conditioned model for collaborative manipulation, we further require a dataset of paired human-robot interactions. However, it is not easy to design a robot policy that can be deployed alongside a human partner. To control a robot arm with natural arm movements, we develop a low-level correspondence between the human and the robot. Specifically, we map the human hand's 3-D position as a translation and use the 3-D rotation from the human wrist joint to the hand joint to generate a 6-D end-effector pose for the robot. We track this end-effector pose using an IK-based joint impedance controller [47]. Our tele-operation system utilizes an Optitrack Motion capture system that detects human joint positions at 120Hz and can track the calculated 6-D end-effector pose in real-time. We collect not only the joint positions of the robot and its human partner

but also the robot-paired joint positions of the tele-operating human. The paired data allows us to align human and robot representations for effective transfer learning (Section IV-C). More details included in Section V.

B. Model Architecture: Action-Conditioned Transformer

Encoding the Scene Context. Fig 2 gives an overview of INTERACT's model architecture, which is based on Multi-Range Transformer (MRT) [28]. Both the human history $\in \mathbb{R}^{T\times d}$ and robot history $\in \mathbb{R}^{T\times j}$ (when training on human-human data, the dimensions of both histories are the same) are passed through linear layers and projected to the same embedding dimension $\in \mathbb{R}^{T\times D}$. The human history is passed through a *local* transformer encoder, whereas the combined human and robot history is passed through a *global* transformer encoder. To form the final scene context encoding, both the local transformer encoding $\in \mathbb{R}^{\times T\times D}$ and the global transformer encoding $\in \mathbb{R}^{2\times T\times D}$ are concatenated together $\in \mathbb{R}^{3\times T\times D}$. Note that prior to any values being passed into the encoders, a Discrete Cosine Transform (DCT) is applied to them, and an Inverse Discrete Cosine Transform (IDCT) is applied to the final decoder outputs. This practice was introduced by [48] to enforce smoothness and periodicity in generated pose outputs.

Decoding Human-Intent using Action-Conditioning. MRT decodes future human intent by passing an embedding of the last observable human pose $\in \mathbb{R}^{1\times d}$ as a query to a Transformer Decoder. In this work, we offset the entire scene around the last human observable pose (and add this offset back into the final predictions). Instead of the last observable human pose, we pass in the robot's future action $a_R \in \mathbb{R}^{1\times j}$ embedding as the query. When training on human-human data, the human pose 1s in the future $a_H \in \mathbb{R}^{1\times d}$ is passed in instead. The future action is passed through a linear layer and projected to the same embedding dimension as the encoded contexts $\in \mathbb{R}^{1\times D}$. This future action embedding is passed in as the query to the transformer decoder. The scene context encoding vector forms the key and value for the transformer decoder. The decoder output $\in \mathbb{R}^{1\times D}$ is first passed through a sequence of linear layers to generate a T-horizon embedding $\in \mathbb{R}^{T\times D}$. Finally, a linear layer decodes the embedding vector to the human's joint dimensions $\in \mathbb{R}^{T\times J}$. Note that the only change in our architecture from MRT is the query to the transformer decoder.

C. Aligning Human and Robot Representations

Representation Mismatch. As mentioned in the previous section, the robot and human have different joint dimensions. Besides, they represent different morphologies. In our transformer model, they are projected into D-dimensional embeddings via different linear layers. We wish to align the embeddings from human and robot motion into the same embedding space. For this purpose, we utilize the paired data stored during tele-operation while collecting human-robot data. For each robot pose, $s_R \in R^j$, we have a corresponding human body pose $s_H \in R^d$. We create a dataset D_{HR} from the paired human and robot poses and use it for aligning human-robot representations.

Alignment Loss. To transfer our model from human-human to human-robot data, the learned human and robot embeddings need to be aligned. We leverage the dataset D_{HR} of paired human and robot poses for this purpose. Specifically, for our transformer model parameterized by θ, we wish to align the robot history embedding layers, parameterized by θ^H_{hist} and θ^R_{hist}, where the former is utilized to embed human-history when training on human-human interaction data and the latter is used with human-robot data. Concretely, we employ a simple cosine similarity [49] loss for the history embedding vectors as follows:

$$L^{hist}_{align}(\theta_{hist}) = \sum_{s_R,s_H}^{D_{HR}} \Big[1 - S_C(f_{\theta^R_{hist}}(s_R), f_{\theta^H_{hist}}(s_H))\Big] \tag{3}$$

where S_C is the cosine similarity metric between two embedding vectors. Similarly, we also align the future-action embedding layers, parameterized by θ^H_{fut} and θ^R_{fut}.

Overall Loss Equation. Our complete loss function is therefore the following:

$$L(\theta) = \lambda_p L_{pred}(\theta) + \lambda_h L^{hist}_{align}(\theta_{hist}) + \lambda_f L^{fut}_{align}(\theta_{fut}) \tag{4}$$

where L_{pred} is the prediction loss (MPJPE) on the forecasts

$$L_{pred}(\theta) = \frac{1}{T}\sum_{t=1}^{T} \left\| \hat{s}^H_t - s^H_t \right\|^2_2 \tag{5}$$

Here, $\hat{s}^H_t$, $\hat{s}^H_t$ are the predicted and ground truth human poses respectively. λ_p, λ_h, and λ_f are loss coefficients (set to 1, 0.1, and 0.1 respectively). Note that there are two separate alignment loss terms, one to indicate the alignment of history motion and one for the alignment of future poses.

V. EXPERIMENTS

A. Collaborative Manipulation Dataset (CoMaD)

Previous multi-pose prediction methods [27], [28] train and evaluate on small-scale datasets. For example, the human-human interaction split of CMU-Mocap [50] consists of just 55 episodes of two human interactions with an average length of 3 seconds, totaling 6 minutes of human motion compared to 40 hours of single-human motion in AMASS. In fact, these methods train their models by augmenting existing datasets with synthetic single-person activity.

In this paper, we extend the **Collaborative Manipulation Dataset (CoMaD)** [6]. The human-human interaction dataset (Fig 3.) now includes 8 diverse subjects performing 3 different kitchen tasks with a total of 270 episodes (average 30s length), totaling more than 4 hours of human motion. Further, we introduce the human-robot dataset consisting of 217 episodes of interaction collected via tele-operation of a 7-DoF Franka-Emika Research 3 robot with a human partner (Section IV-A). Episodes of each task are divided into train, validation, and test splits in an 8:1:1 ratio.

Human-Human Data. The length of each episode ranges from 20 to 40 seconds. The dataset consists of three tasks:

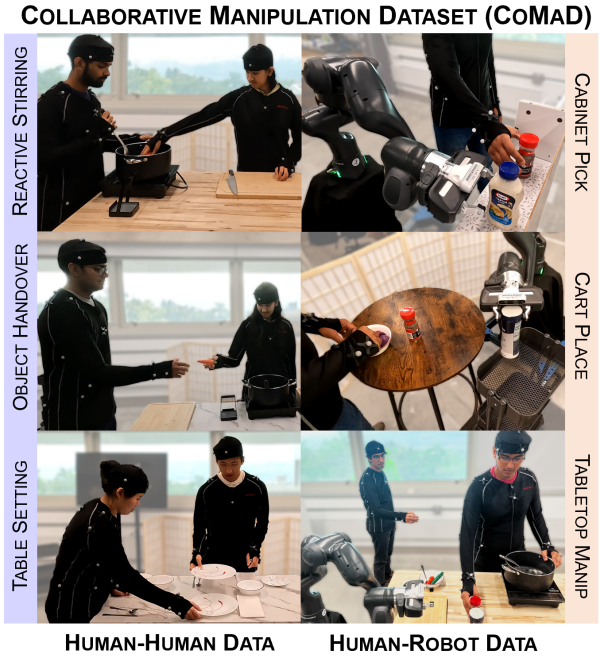

Fig. 3: **Collaborative Manipulation Dataset (CoMaD)** consists of Human-Human and Human-Robot interaction data. We collect data on three different H-H tasks and three different H-R tasks across several subjects. The bottom right image shows our tele-operation setup for paired human-robot data collection.

(1) TABLE SETTING (70 episodes): Two humans manipulate table items, avoiding collisions between them. (2) OBJECT HANDOVER (106 episodes): One human asks for objects, and the other human moves in to hand the object over. (3) REACTIVE STIRRING (94 episodes): One human stirs a pot and reacts to the other human pouring vegetables into it.

Human-Robot Data. The length of each episode ranges from 3 to 15 seconds. The dataset consists of three tasks: (1) CABINET PICK (135 episodes): The robot reaches for one of two objects on the cabinet, and the human responds accordingly. If the robot reaches for the object close to the human, they wait, and if the robot reaches for the object away from the human, both reach for their respective objects. (2) CART PLACE (55 episodes): There is a table and cart setup between the robot and the human. The robot moves an object from the table to the cart, and the human picks up an object from the cart to use. If the human moves first, the robot must wait for the human, and if the robot moves first, the human must wait for the robot. (3) TABLETOP MANIPULATION (27 episodes): A table has two objects on it, with both the human and robot reaching for one of them. The human waits for the robot when it moves in. Similarly, the robot must wait if the human comes in the way.

B. Experimental Setup

Large Human-Activity Databases. We created synthetic two-human data using AMASS [7] and pre-trained the model using the synthetic data and CMU-Mocap [50] data. We use the human-human interaction data in CMU-Mocap without adding any synthetic humans.

Baselines (H-H). MARGINAL [6] uses one human's history to predict intent, whereas MARGINAL (+ HIST) [28] also uses the other human's history. Both are pre-trained on synthetic AMASS data and fine-tuned on H-H data. ONLY FINETUNED is only trained on a smaller amount of H-H data. Our method, INTERACT uses both humans' histories and conditions on the other human's future action.

Baselines (H-R). MARGINAL takes the corresponding H-H model above and fine-tunes on H-R data, whereas ONLY FINETUNED is only trained on H-R data. INTERACT takes our H-H model and fine-tunes on H-R data, replacing the second human's encoding with the robot. INTERACT + ALIGN further incorporates the robot alignment loss (Eq 3).

Implementational Details. We utilize a 1s motion history input to generate a 1s forecast (represented over 15 timesteps). We consider the human pose dimension $d = 27$, which includes 9 upper body 3-D joint positions (upper back, shoulders, elbows, wrists, hands), and the robot pose dimension $j = 6$, which includes two 3-D points on the robot's end-effector corresponding to the human's hand and wrist. We report the Final Displacement Error (FDE), which is the average distance between the predicted joint positions and ground truth joint positions at the end of 1s.

C. Results and Analysis

***O1.* Conditioning on actions improves intent prediction in both human-human and human-robot interactions.** Fig.4 and Fig.6 both show that INTERACT models outperform any MARGINAL models without information about the intent of the other agent in the scene. MARGINAL models produce higher FDE on all three H-H and H-R tasks compared to conditional models. This can be seen qualitatively in Fig.5 where the INTERACT intent predictions anticipate a handover due to knowledge about the planned action of the other human in the scene. Similar trends follow in H-R tasks such as CABINET PICK demonstrated in Fig 7 where conflict arises as a human and robot simultaneously reach for objects. If the robot reaches for the object on the right, we know the human intends to pick the object on the left.

***O2.* Human-Robot Alignment loss helps improve prediction performance.** Fig.6 shows that adding alignment loss (INTERACT + ALIGN) reduces FDE in predicting future human poses. This supports our hypothesis that aligning representations helps in transfer learning from H-H data.

***O3.* Pre-training models on human-human interactions is critical for transfer learning.** Fig.6 shows that ONLY FINETUNED trained only on H-R data performs significantly worse than other MARGINAL and INTERACT that are also trained on H-H data. It yields notably higher FDE across all joints in all three H-R tasks we evaluate on.

***O4.* Pre-training on synthetic human-human activity data helps learn general human motion dynamics.** Fig.4 shows that ONLY FINETUNED produces higher FDE than models pre-trained on synthetic AMASS data despite the synthetic data lacking real H-H interactions. This leads us to believe that large-scale single-human data can be leveraged even in the multi-human setting.

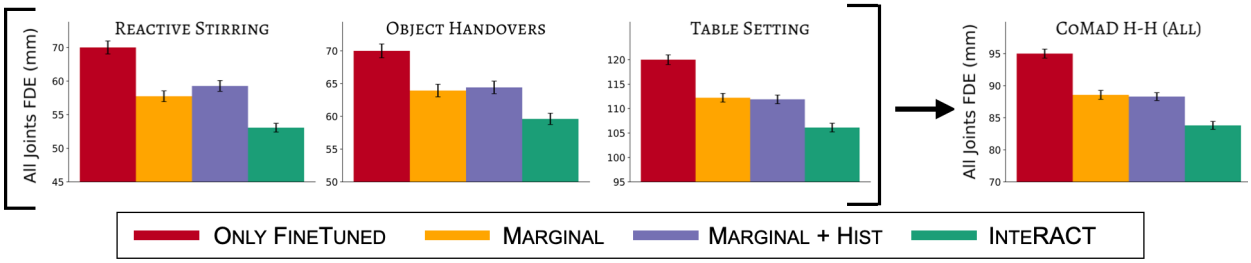

Fig. 4: All Joints Final Displacement Error (FDE) across all tasks in CoMaD H-H. **InteRACT predictions have lowest FDE.**

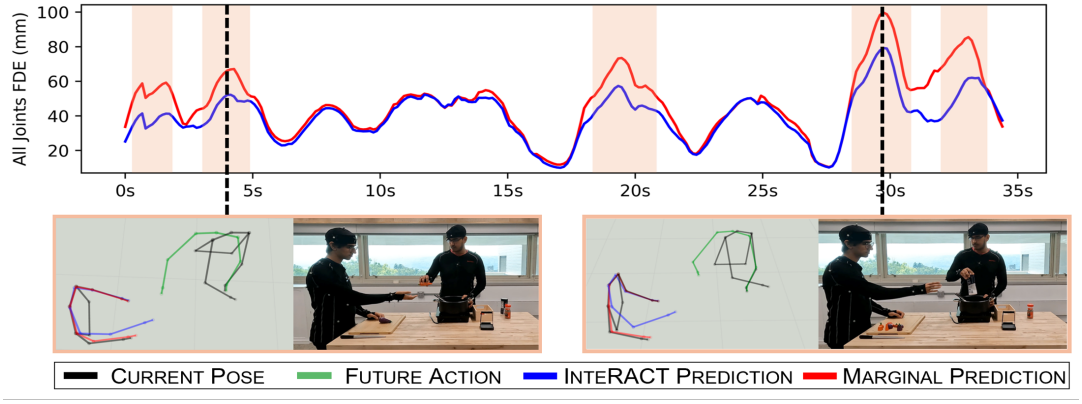

Fig. 5: **Top:** Final Displacement Error (FDE) of all joints over time in a test-set episode of object handover. Highlighted windows indicate all object handovers in the episode, where we observe higher errors for Marginal. **Bottom:** Visualizations of the predictions when the error is at its peak (1s pre-RGB image) show InteRACT anticipates the other's human action and moves towards the handover location.

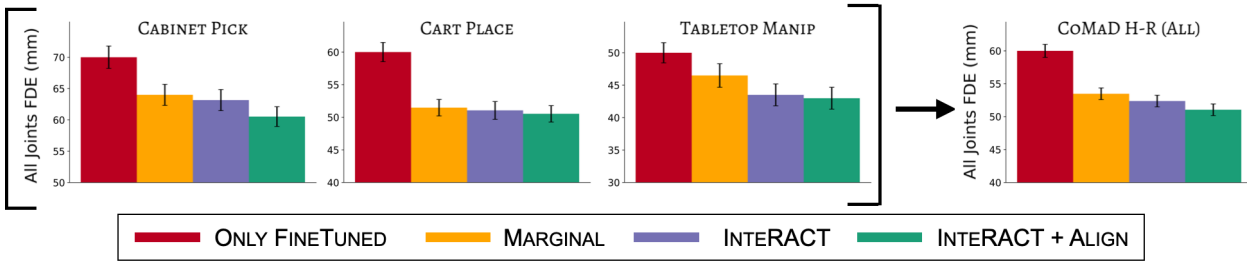

Fig. 6: Final Displacement Error (FDE) on all joints per and across all tasks in CoMaD H-R. InteRACT variants perform better than other models, with reductions in FDE across tasks with human-robot representation alignment.

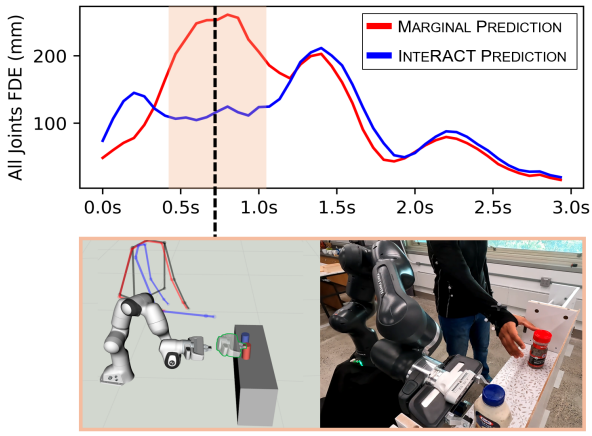

Fig. 7: Comparing Final Displacement Error (FDE) between InteRACT and Marginal predictions in a test-set CoMaD H-R Cabinet Pick episode. InteRACT produces more accurate predictions when the planned robot action is picking up a specific item, indicating the other item is free to pick.

VI. Discussion and Limitations

In this work, we present InteRACT, a novel architecture that predicts human intentions by **conditioning on future robot actions**. We also expand the Collaborative Manipulation Dataset (CoMaD) with a novel **paired human-robot dataset** collected by tele-operation allowing us to effectively **align** a model trained on human-human data to human-robot interactions. In the future, we aim to demonstrate the performance of InteRACT in online planning scenarios. By reasoning about how actions can influence human intent, robots can be more confident in their plans.

Limitations. There are notable limitations to our work that we highlight in this section. Robot safety in close proximity interactions is extremely important, and collisions can be a concern in the case of errors in human intent prediction. Safety mechanisms [51] studied extensively should be used to help target these potential issues. While we collect data across several subjects, we are limited to certain environments per task. Our goal is to collect data in a distribution

that represents a few different modes of motion that are common in human-robot interactions, and plan to expand the dataset in the future to cover a wider distribution.

VII. Acknowledgements

This work was partially funded by NSF RI (#2312956).

References

[1] A. D. Dragan, "Robot planning with mathematical models of human state and action," *arXiv preprint arXiv:1705.04226*, 2017.

[2] L. Gui, K. Zhang, Y.-X. Wang, X. Liang, J. M. F. Moura, and M. M. Veloso, "Teaching robots to predict human motion," *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 562–567, 2018.

[3] J. Laplaza, A. Pumarola, F. Moreno-Noguer, and A. Sanfeliu, "Attention deep learning based model for predicting the 3d human body pose using the robot human handover phases," in *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 2021, pp. 161–166.

[4] 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.

[5] J. Laplaza, F. Moreno-Noguer, and A. Sanfeliu, "Context attention: Human motion prediction using context information and deep learning attention models," in *ROBOT2022: Fifth Iberian Robotics Conference: Advances in Robotics, Volume 1*. Springer, 2022, pp. 102–112.

[6] K. Kedia, P. Dan, A. Bhardwaj, and S. Choudhury, "Manicast: Collaborative manipulation with cost-aware human forecasting," in *7th Annual Conference on Robot Learning*, 2023. [Online]. Available: https://openreview.net/forum?id=rxlokRzNWRq

[7] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black, "AMASS: Archive of motion capture as surface shapes," in *International Conference on Computer Vision*, Oct. 2019, pp. 5442–5451.

[8] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 7, pp. 1325–1339, 2013.

[9] J. Ngiam, V. Vasudevan, B. Caine, Z. Zhang, H.-T. L. Chiang, J. Ling, R. Roelofs, A. Bewley, C. Liu, A. Venugopal, D. J. Weiss, B. Sapp, Z. Chen, and J. Shlens, "Scene transformer: A unified architecture for predicting future trajectories of multiple agents," in *International Conference on Learning Representations*, 2022.

[10] Z. Huang, H. Liu, J. Wu, and C. Lv, "Conditional predictive behavior planning with inverse reinforcement learning for human-like autonomous driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, pp. 7244–7258, 2022.

[11] Z. Huang, H. Liu, J. Wu, W. Huang, and C. Lv, "Learning interaction-aware motion prediction model for decision-making in autonomous driving," *ArXiv*, vol. abs/2302.03939, 2023.

[12] H. Song, W. Ding, Y. Chen, S. Shen, M. Y. Wang, and Q. Chen, "Pip: Planning-informed trajectory prediction for autonomous driving," in *European Conference on Computer Vision*, 2020.

[13] E. V. Tolstaya, R. Mahjourian, C. Downey, B. Varadarajan, B. Sapp, and D. Anguelov, "Identifying driver interactions via conditional behavior prediction," *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3473–3479, 2021.

[14] S. M. Ettinger, S. Cheng, B. Caine, C. Liu, H. Zhao, S. Pradhan, Y. Chai, B. Sapp, C. Qi, Y. Zhou, Z. Yang, A. Chouard, P. Sun, J. Ngiam, V. Vasudevan, A. McCauley, J. Shlens, and D. Anguelov, "Large scale interactive motion forecasting for autonomous driving : The waymo open motion dataset," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9690–9699, 2021.

[15] W. Zhan, L. Sun, D. Wang, H. Shi, A. Clausse, M. Naumann, J. Kümmerle, H. Königshof, C. Stiller, A. de La Fortelle, and M. Tomizuka, "Interaction dataset: An international, adversarial and cooperative motion dataset in interactive driving scenarios with semantic maps," *ArXiv*, vol. abs/1910.03088, 2019.

[16] C. Chen, Y. Liu, S. Kreiss, and A. Alahi, "Crowd-robot interaction: Crowd-aware robot navigation with attention-based deep reinforcement learning," in *2019 international conference on robotics and automation (ICRA)*. IEEE, 2019, pp. 6015–6022.

[17] H. Chen, S. Feng, Y. Zhao, C. Liu, and P. A. Vela, "Safe hierarchical navigation in crowded dynamic uncertain environments," in *2022 IEEE 61st Conference on Decision and Control (CDC)*, 2022, pp. 1174–1181.

[18] A. Monti, A. Bertugli, S. Calderara, and R. Cucchiara, "Dag-net: Double attentive graph neural network for trajectory forecasting," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 2551–2558.

[19] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone, "Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*. Springer, 2020, pp. 683–700.

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

[21] S. Poddar, C. Mavrogiannis, and S. S. Srinivasa, "From crowd motion prediction to robot navigation in crowds," *arXiv preprint arXiv:2303.01424*, 2023.

[22] P. Kothari and A. Alahi, "Safety-compliant generative adversarial networks for human trajectory forecasting," *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[23] R. Tian, L. Sun, A. V. Bajcsy, M. Tomizuka, and A. D. Dragan, "Safety assurances for human-robot interaction via confidence-aware game-theoretic human models," *2022 International Conference on Robotics and Automation (ICRA)*, pp. 11 229–11 235, 2021.

[24] K. Kedia, P. Dan, and S. Choudhury, "A game-theoretic framework for joint forecasting and planning," *ArXiv*, vol. abs/2308.06137, 2023.

[25] W. Mao, M. Liu, and M. Salzmann, "History repeats itself: Human motion prediction via motion attention," in *European Conference on Computer Vision*, 2020.

[26] T. Sofianos, A. Sampieri, L. Franco, and F. Galasso, "Space-time-separable graph convolutional network for pose forecasting," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11 189–11 198, 2021.

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

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

[29] W. Yang, B. Sundaralingam, C. Paxton, I. Akinola, Y.-W. Chao, M. Cakmak, and D. Fox, "Model predictive control for fluid human-to-robot handovers," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 6956–6962.

[30] E. A. Sisbot and R. Alami, "A human-aware manipulation planner," *IEEE Transactions on Robotics*, vol. 28, no. 5, pp. 1045–1057, 2012.

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

[32] H. Liu and L. Wang, "Collision-free human-robot collaboration based on context awareness," *Robotics and Computer-Integrated Manufacturing*, vol. 67, p. 101997, 2021.

[33] S. Scheele, P. Howell, and H. Ravichandar, "Fast anticipatory motion planning for close-proximity human-robot interaction," *arXiv preprint arXiv:2305.11978*, 2023.

[34] H. Ling, G. Liu, L. Zhu, B. Huang, F. Lu, H. Wu, G. Tian, and Z. Ji, "Motion planning combines human motion prediction for human-robot cooperation," in *2022 12th International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER)*. IEEE, 2022, pp. 672–677.

[35] V. Unhelkar, P. A. Lasota, Q. Tyroller, R.-D. Buhai, L. Marceau, B. Deml, and J. A. Shah, "Human-aware robotic assistant for collaborative assembly: Integrating human motion prediction with planning in time," *IEEE Robotics and Automation Letters*, vol. 3, pp. 2394–2401, 2018.

[36] G. Hoffman, T. Bhattacharjee, and S. Nikolaidis, "Inferring human intent and predicting human action in human–robot collaboration," *Annual Review of Control, Robotics, and Autonomous Systems*, 2023.

[37] V. Prasad, D. Koert, R. M. Stock-Homburg, J. Peters, and G. Chalvatzaki, "Mild: Multimodal interactive latent dynamics for learning human-robot interaction," *2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids)*, pp. 472–479, 2022.

[38] Q. Li, G. Chalvatzaki, J. Peters, and Y. Wang, "Directed acyclic graph neural network for human motion prediction," *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3197–3204, 2021.

[39] J. Mainprice, R. Hayne, and D. Berenson, "Goal set inverse optimal control and iterative replanning for predicting human reaching motions in shared workspaces," *IEEE Transactions on Robotics*, vol. 32, no. 4, pp. 897–908, 2016.

[40] E. Valls Mascaro, Y. Yan, and D. Lee, "Robot interaction behavior generation based on social motion forecasting for human-robot interaction," *2024 IEEE International Conference on Robotics and Automation (ICRA*, 2024.

[41] N. S. Pollard, J. K. Hodgins, M. Riley, and C. G. Atkeson, "Adapting human motion for the control of a humanoid robot," *Proceedings 2002 IEEE International Conference on Robotics and Automation (Cat. No.02CH37292)*, vol. 2, pp. 1390–1397 vol.2, 2002.

[42] X. B. Peng, A. Kanazawa, J. Malik, P. Abbeel, and S. Levine, "Sfv: Reinforcement learning of physical skills from videos," *ACM Trans. Graph.*, vol. 37, p. 178, 2018.

[43] A. Handa, K. V. Wyk, W. Yang, J. Liang, Y.-W. Chao, Q. Wan, S. Birchfield, N. D. Ratliff, and D. Fox, "Dexpilot: Vision-based teleoperation of dexterous robotic hand-arm system," *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9164–9170, 2019.

[44] G. Garcia-Hernando, E. Johns, and T.-K. Kim, "Physics-based dexterous manipulations with estimated hand poses and residual reinforcement learning," *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 9561–9568, 2020.

[45] H. Bharadhwaj, A. Gupta, S. Tulsiani, and V. Kumar, "Zero-shot robot manipulation from passive human videos," *ArXiv*, vol. abs/2302.02011, 2023.

[46] C. Wang, L. Fan, J. Sun, R. Zhang, L. Fei-Fei, D. Xu, Y. Zhu, and A. Anandkumar, "Mimicplay: Long-horizon imitation learning by watching human play," *arXiv preprint arXiv:2302.12422*, 2023.

[47] K. Zhang, M. Sharma, J. Liang, and O. Kroemer, "A modular robotic arm control stack for research: Franka-interface and frankapy," *ArXiv*, vol. abs/2011.02398, 2020.

[48] W. Mao, M. Liu, M. Salzmann, and H. Li, "Learning trajectory dependencies for human motion prediction," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9489–9497.

[49] Y. Aytar, C. Vondrick, and A. Torralba, "See, hear, and read: Deep aligned representations," *arXiv preprint arXiv:1706.00932*, 2017.

[50] [Online]. Available: http://mocap.cs.cmu.edu/

[51] P. A. Lasota, T. Fong, J. A. Shah *et al.*, "A survey of methods for safe human-robot interaction," *Foundations and Trends® in Robotics*, vol. 5, no. 4, pp. 261–349, 2017.