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.

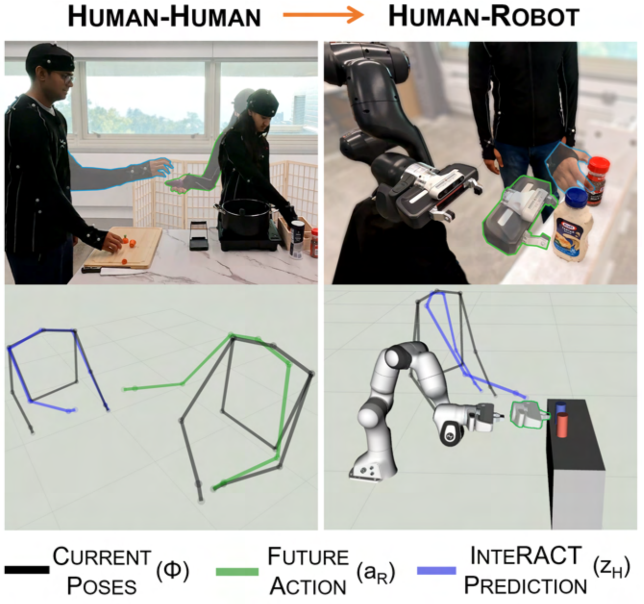

Fig. 1: INTERACT predictions condition on the future action of the other agent. **Left:** (Human-Human object handover) Given the future object handover action of the human, INTERACT predicts that the human will move towards it. **Right:** (Human-Robot cabinet item pick) Given the robot reaches for the 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

[1] Department of Computer Science, Cornell University

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. 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 [16]. 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 II-C). More details included in Section III.

B. Model Architecture: Action-Conditioned Transformer

Encoding the Scene Context. textscInteRACT's model architectue is based on Multi-Range Transformer (MRT) [17]. Both the human history and robot history are passed through linear layers and projected to the same embedding dimension. 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 and the global transformer encoding are concatenated together. Note that prior to any values being passed into the encoders, a Discrete Cosine Transform (DCT) is applied to them, and an Inverse DCT is applied to the final decoder outputs.

Decoding Human-Intent using Action-Conditioning. MRT decodes future human intent by passing an embedding of the last observable human pose 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 embedding as the query. When training on human-human data, the human pose 1s in the future is passed in instead. The future action is passed through a linear layer and projected to the same embedding dimension as the encoded contexts. 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 is first passed through a sequence of linear layers to generate a T-horizon embedding. Finally, a linear layer decodes the embedding vector to the human's joint dimensions.

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

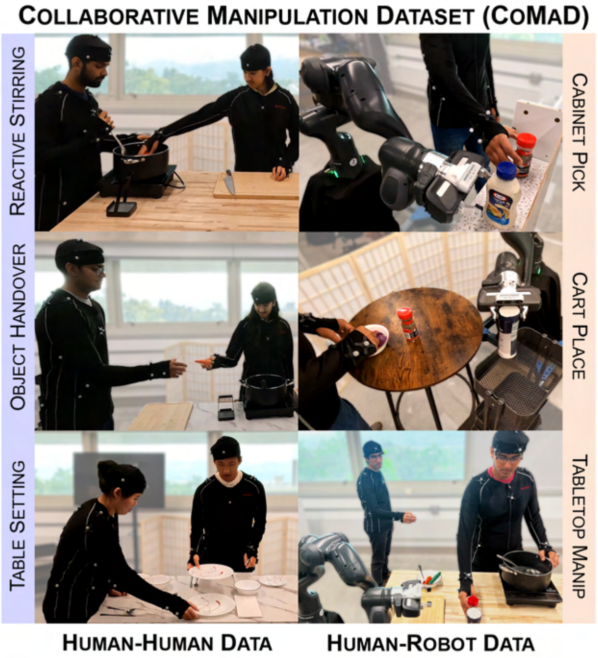

Fig. 2: **Collaborative Manipulation Dataset (CoMaD)** consists of Human-Human and Human-Robot interaction data. We collect data on three different H-H tasks and three different H-R tasks across several subjects. The bottom right image shows our tele-operation setup for paired human-robot data collection.

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 [18] loss for the history embedding vectors as follows:

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

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:

$$L(\theta) = \lambda_p L_{pred}(\theta) + \lambda_h L^{hist}_{align}(\theta_{hist}) + \lambda_f L^{fut}_{align}(\theta_{fut}) \tag{2}$$

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

$$L_{pred}(\theta) = \frac{1}{T} \sum_{t=1}^{T} \left\| \hat{s}^H_t - s^H_t \right\|_2^2 \tag{3}$$

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.

III. Experiments

A. Collaborative Manipulation Dataset (CoMaD)

In this paper, we extend the **Collaborative Manipulation Dataset (CoMaD)** [6]. The human-human interaction dataset (Fig 2.) 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 II-A). Episodes of each task are divided into train, validation, and test splits in an 8:1:1 ratio.

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 [19] 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) [17] 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 1).

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 3 and Fig 4 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 H-R tasks such as Cabinet Pick. Fig 5 shows a scenario 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.

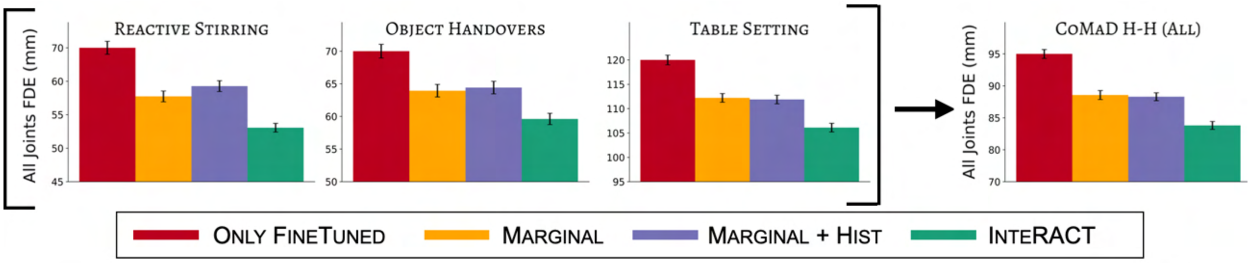

Fig. 3: All Joints Final Displacement Error (FDE) across all tasks in CoMaD H-H. **InteRACT predictions have lowest FDE**.

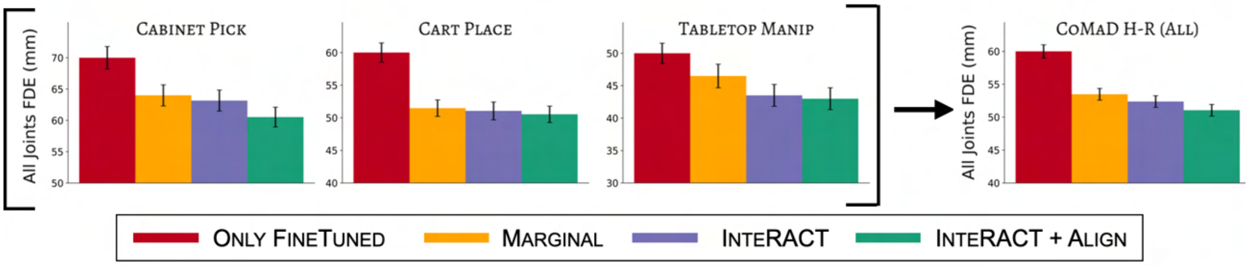

Fig. 4: Final Displacement Error (FDE) on all joints per and across all tasks in CoMaD H-R. InteRACT variants perform better than other models, with reductions in FDE across tasks with human-robot representation alignment.

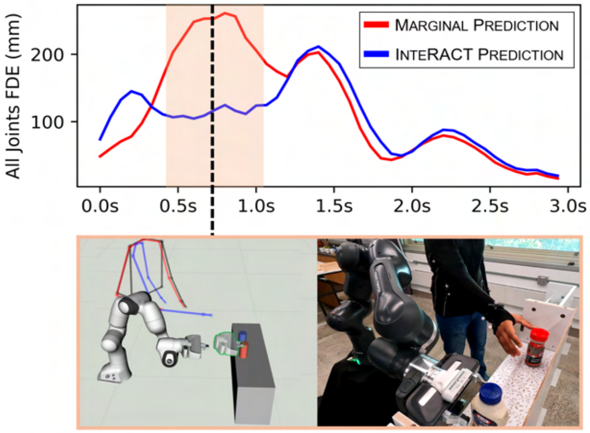

Fig. 5: Comparing Final Displacement Error (FDE) between InteRACT and Marginal predictions in a test-set CoMaD H-R Cabinet Pick episode. InteRACT produces more accurate predictions when the planned robot action is picking up a specific item, indicating the other item is free to pick.

***O2.* Human-Robot Alignment loss helps improve prediction performance.** Fig.4 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.4 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.3 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.

IV. 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 [20] 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.

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] 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.

[17] 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.

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

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

[20] 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.