Personalization Dynamic Multiscale Graph Neural Networks for Human Motion Prediction

Kin Man Lee
klee863@gatech.edu

Erin Botti (Hedlund)
erin.botti@gatech.edu

Vivek Mallampati
vmallampati6@gatech.edu

Lily (Chunyue) Xue
chunyuexue@gatech.edu

Abstract

When humans and robots collaborate in a proximate setting, the robot needs to be able to accurately predict the human's motion to ensure safety and collaborate effectively. Prior approaches have used Recurrent Neural Networks (RNN) networks to capture time series information on Graph Neural Networks (GNN) to represent the pose as a graph. However, these approaches are one-size-fits-all methods that may not perform well for any specific individual. They do not specifically account for differences in movement that may arise from personal idiosyncrasies, i.e., gender, body type, skill, etc. We hypothesize that accounting for personalization will improve the accuracy of these pose prediction algorithms. In this project, we extend a prior state-of-the-art GNN by adding a personalization network to improve pose estimation. We evaluate our method compared to the GNN baseline with a human motion capture dataset. We show that including a personalization network can help to improve human motion prediction using skeletal joint data. Further, we show that calibrating the personalization embeddings to a specific user can allow for greater improvements in the predictions.

1. Introduction

In human-robot teams where the human and robot are collaborating in close proximity, for safety, the robot accurately must predict where the human will potentially move next. For example, in table tennis doubles, the robot will need to know when it should try to hit the ball and when to move out of the person's way. One method to predict where the human may be is through pose estimation from motion capture data. We would like to be able to predict a person's pose in future time steps based on their current pose and poses in the previous time steps.

Deep learning models such as Graph Neural Networks (GNN), Recurrent Neural Networks (RNN), and Long-Short Term Memory (LSTM) networks have been developed to predict short-term human motion. Li et al. [4] proposed a GNN-based multi-scale network, DMGNN, that can forecast the future pose based on past human-body skeletons. It uses an encoder-decoder framework with three key components: a multi-scale graph computational unit to extract and fuse features, a graph-based gated recurrent unit (GRU) to learn and update the hidden state, and a different operator to capture dynamics information. We will use this method as one of the baselines in our project. A new follow-up framework exists, multi-scale Spatio-temporal GNN(MST-GNN). Compared to the previous model, it considers the temporal information besides a spatial graph [**?**].

In this work, we intend to contribute the following:

1. DMGNN [4] and RNN-based [6] baselines implemented using the BML-MoVi motion capture dataset [5] instead of the Human3.6M skeleton dataset [3].

2. A novel framework with a personalization layer, built on DMGNN [4], that predicts human pose.

3. Results comparing our framework to our baselines in terms of prediction error and computation time.

Researchers who are trying to use pose estimation in human-robot collaboration settings would greatly appreciate the efforts we are putting in. The personalized pose estimation can help to create human-centered robots where the action of robots will adapt according to a specific human partner's behavior and implicit preferences to improve user satisfaction.

2. Related Works

Previous literature explores the use of recurrent neural networks for human prediction. Martinez et al. [6] explored the idea of modifications to the current RNN architectures that results in a simple and scalable RNN architecture that obtains state-of-the-art performance on human motion prediction. Results show that their sequence- to-sequence architecture with residual connections outperformed the baseline when trained on a sample-based loss. We evaluate this RNN method as a potential baseline near the beginning of the project and only select DMGNN since it significantly outperforms RNN at the end.

Our baseline algorithms learn one-size-fits-all prediction approaches that do not account for heterogeneity between people. However, a person's movement may differ based on person-to-person differences, such as gender, body composition, skill, etc. There are some algorithms utilizing personalization to improve prediction accuracy. For example, Antwarg et al. [1] developed an HMM model trained with demographic data and the sequence of actions to predict humans' action intention in a task-level situation. However, their method predicted the action intention or goal instead of the joint positions of the human pose. Our objective concentrates on pose prediction with millisecond time steps not exceeding one second, but Antwarg et al. show the potential of action prediction using the attribute-driven method in our circumstance [1].

Chen et al. [2] describe an unsupervised method, InfoGAN, that learns embeddings to distinguish heterogeneous elements in datasets by deriving a lower bound on mutual information. A similar method has been applied to time-series data to learn different people's driving styles [7]. We plan to use a similar approach to learn a personalized embedding to distinguish heterogeneity from person-to-person differences for pose prediction.

Our baseline algorithms originally used a pose dataset consisting of human skeleton data [3]. The Archive of Motion Capture As Surface Shapes (AMASS) dataset [5] (described further in Section 4) includes motion capture data that details the human pose as a surface. In this dataset, differences between body shapes, gender, etc., are captured. We adapt the baseline algorithm that performs better to work with this more expressive dataset. From our experiments, DMGNN outperformed the RNN-based method. We built upon DMGNN by adding a personalization network that simultaneously learns personalized embedding and poses estimation. We hypothesize that the more expressive dataset may improve baseline performance and will improve our ability to generate informative personalized embeddings with our personalization network.

3. Technical Approach

In this section, we first describe the baseline algorithms we are investigating. Then we detail our proposed novel personalization algorithm that builds on the baselines. Lastly, we list the metrics that we will use to evaluate our algorithm compared to the baselines.

3.1. Baselines

RNN The first baseline method we found was a prediction algorithm using recurrent neural networks [6]. Based on the standard RNN model, three modifications are designed for human motion prediction. It focuses on short-term prediction. To avoid the weakness of standard RNN, it introduces the realistic error in training time without any noise schedule. The authors propose a sequence-to-sequence architecture with residual connections to improve performance and reduce training time. The architecture consists of Gated Recurrent Units (GRUs) to form the encoder and decoder networks. It is trained with a sampling-based loss and models first-order motion derivatives instead of absolute joint angles. To apply for the multi-action task, authors find that a supervised variant by concatenating one-hot vectors can enhance the inputs and lead to the best quantitative result.

DMGNN The second baseline algorithm we evaluated was dynamic multi-scale graph neural network (DMGNN) [4]. DMGNN utilizes a graph-based encoder-decoder architecture to predict the human skeletal joint positions given past joint position data. Based on the joint position data, the human body is represented by a *multiscale* graph, consisting of *single-scale* graphs, where nodes connected represent body parts of the same scale and *cross-scale* graphs, which are bipartite graphs connecting body parts across different scales. The DMGNN encoder is composed of a sequence of multi-scale graph computational units (MGCU), each performing spatial-temporal graph convolutions at a single scale and a set of MLP operations to convert features and perform feature fusion at a cross-scale. For the decoder, a graph-based GRU is introduced for predicting future poses. Each GRU cell takes the current hidden state, and the online skeleton poses data as input to generate the state for the following frame.

3.2. Our Method

We propose a novel framework that adapts DMGNN [4] to add a personalization network. Our framework is depicted in Figure 1. There are two inputs to the framework, a personalized embedding, $w^{(p)}$, and a set of states, $s_{t-n:t}^{(p)}$, representing the person's pose data from n timesteps before t to time t. We adopt the DMGNN encoder to use the pose data from a more expressive dataset [5]. The personalized embedding, $w^{(p)}$, is initialized based upon the prior $\hat{w}^{(p)} \sim \mathcal{N}(0, 1)$. $w^{(p)}$ is also learned by the framework and encapsulates the heterogeneity of a person, p. We hypothesize that $w^{(p)}$ will capture information about a person's body shape and style of movement.

The DMGNN encoder converts the time series and spatial information of the pose information, $s_{t-n:t}^{(p)}$, into an encoding $H_t^{(p)}$. This encoding, $H_t^{(p)}$, is fed into the DMGNN

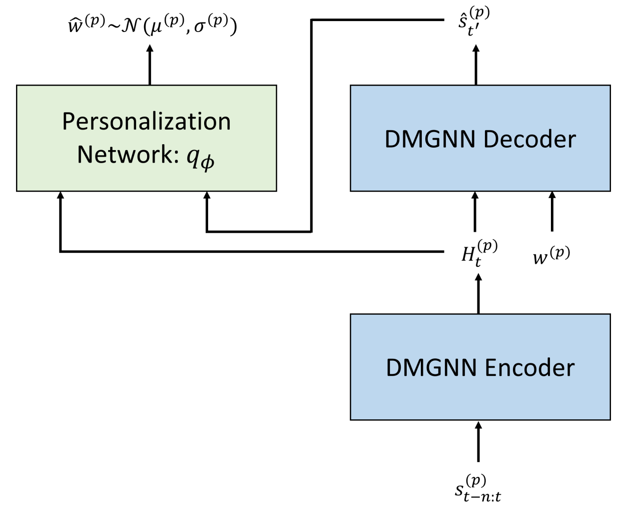

$\hat{w}^{(p)} \sim \mathcal{N}(\mu^{(p)}, \sigma^{(p)})$
$\hat{s}_{t'}^{(p)}$
Personalization Network: q_ϕ
DMGNN Decoder
$H_t^{(p)}$
$w^{(p)}$
DMGNN Encoder
$s_{t-n:t}^{(p)}$

Figure 1. The DMGNN encoder [4] takes in a set of states, $s_{t-n:t}^{(p)}$, which is a series of body positions from n timesteps before t to current timestep t for person p. The DMGNN encoder outputs an encoding, $H_t^{(p)}$, that is fed into the DMGNN decoder and our personalization network, q_ϕ. We modify the DMGNN decoder to have additional input, $w^{(p)}$, which represents the personalized embedding for person p. The DMGNN decoder outputs the prediction for the future state, $\hat{s}_{t'}^{(p)}$, which is also fed into the network q_ϕ. The q_ϕ network learns an approximation of the posterior distribution for the personalized embedding, $w^{(p)}$.

decoder and our personalization network, q_ϕ. We modify the DMGNN decoder and add $w^{(p)}$ as input. The DMGNN decoder maps $H_t^{(p)}$ and the embedding, $w^{(p)}$, to a prediction for the future pose, $\hat{s}_{t'}^{(p)}$, for person p at future time t'. The personalization network, q_ϕ, maps the encoding, $H_t^{(p)}$, and pose prediction, $\hat{s}_{t'}^{(p)}$, to an approximation of the posterior distribution of the personalized embedding, $w^{(p)}$. We sample from the approximate posterior to get an estimate of the person's embedding, $w^{(p)}$.

Our framework works by maximizing the mutual information between the pose prediction $\hat{s}_{t'}^{(p)}$, the learned embedding, $w^{(p)}$, and the encoded input state $H_t^{(p)}$. The uncertainty of our personalized embedding, $w^{(p)}$, decreases as we have more informative future pose prediction. However, in order to maximize mutual information, we need an intractable posterior distribution, $P\left(w^{(p)}|H_t^{(p)}, \hat{s}_{t'}^{(p)}\right)$. Therefore, we use variational inference and the variational lower bound derived in Chen et al. [2].

The DMGNN network uses an l_1 norm loss between predicted pose $\hat{s}_{t'}^{(p)}$ and ground truth $s_{t'}^{(p)}$. Our personalization network will use a mean squared error (MSE) loss between $\hat{w}^{(p)}$ and $w^{(p)}$. This is equivalent to maximizing the log-likelihood of the posterior. We will sum up both of these losses and backpropagate the loss through the network.

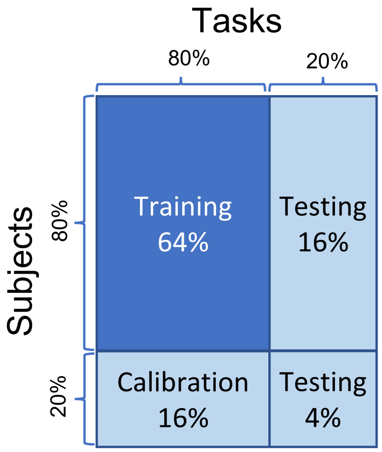


Figure 2. The data is split into training and evaluation sets. We train the data on 80% of subject data and 80% of task data. Then we evaluate each method with two subsets of the data: known subjects completing unseen tasks and new subjects completing unseen tasks. We use the calibration split to first learn the personalized embedding for the new subjects.

Model modifications While the BML-Movi dataset has been converted to the same input format as the H3.6M dataset for DMGNN, the model needs to be modified to adapt to new actions and additional joints introduced in the BML-MoVi dataset. We modified the preprocessing layers of the model to consider the 21 new actions and increased number of data points in the pose. The MGCU layers in the DMGNN encoder scale the body model based on hard-coded joints and the number of data points. Due to the more expressive data in the BML-Movi dataset, we modified these layers to account for the increase in data that is passed through the network.

3.3. Evaluation and Metrics

As discussed in Section 4, the dataset includes data from a set of people completing a set of tasks. We will split the data, using an 80/20% split, into training and evaluation sets based on people and functions (Fig. 2). Our method and each baseline will be trained on the same training data. We will evaluate each technique based on how well it generalizes to unseen tasks and unseen people completing neglected tasks.

We will compare the baseline algorithms to our proposed method on the metrics mean angle error and computation time. Mean angle error (MAE) is an accuracy measure for the output prediction, comparing the average error between the predicted next state and the actual next state. We will evaluate the computation time for the evaluation step to determine if adding the personalization layer significantly slows down prediction time performance.

Action	Sitting Down				Ball Throwing				Running around				Looking Around				Arm Up			
Predict Time (ms)	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
DMGNN	.088	.164	.410	.530	.113	.241	.494	.648	.110	.247	.548	.731	1.61	1.65	2.47	2.48	.036	.080	.195	.244
P-DMGNN (w/o calibration)	.068	.149	.388	.475	.112	.231	1.17	.674	.063	.135	.255	.311	.140	.224	.398	.481	.125	.232	.496	.675
P-DMGNN (ours)	.054	.113	.258	.330	.064	.120	.219	.266	.093	.164	.282	.323	.110	.209	.700	.915	.060	.152	.392	.440

Table 1. MAE scores for five actions within the BML-MoVi dataset

4. Data

AMASS dataset: AMASS is an extensive human motion database unifying different optical marker-based motion capture datasets by representing them within a common framework and parameterization. AMASS is readily helpful for animation, visualization, and generating training data for deep learning [5].

We select one set of experiments from this AMASS database called BML-MoVi. Biomotion Lab of York University has captured the human motion that can be applied in human pose estimation and tracking, human motion prediction and synthesis, action recognition, and gait analysis [5]. AMASS-approved MoVi collection consists of eighty-nine subjects with a total of 1864 different motions captured, with a runtime of 174.39 minutes. AMASS has each frame of the person as a collection of global root orientation, body poses, and finger articulation poses (33 joints in total). The AMASS dataset clearly distinguishes hand poses, making it detailed in visualization. Using an AMASS dataset, we can implement the same algorithm and methodology on another set of actions or test personalization on various tasks.

Human3.6M dataset: The Human3.6M dataset [3] consists of skeleton pose data for seven subjects, 15 classes of actions, and 32 joints per subject.

We chose to use the AMASS dataset due to the increased expressivity of the data and the higher number of subjects in the dataset. Since we are exploring how personalized embeddings can capture heterogeneity and better inform pose prediction, we preferred a dataset with more variation in terms of subjects. Additionally, the AMASS dataset have extended motion capture data that creates a mesh that includes information about the person's body shape and composition that is more expressive and individualized than the skeleton data that we plan to use, allowing for future work that can show the potential of personalization with even more expressive data.

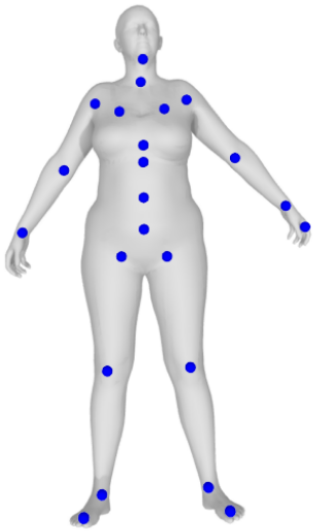
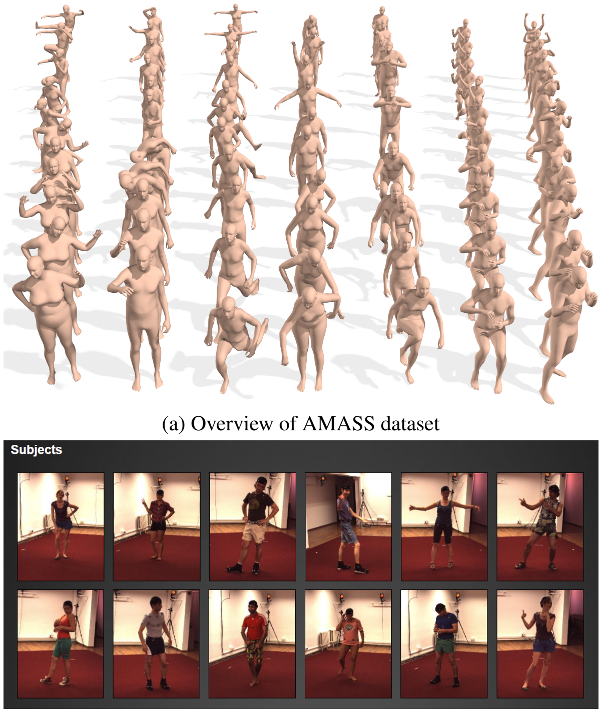


Figure 3. **AMASS Dataset** input after pre-processing. This is an overlayed image of both the model's rendering and the joint positions(blue dots) indicating the model position at a given timestamp.

Data preprocessing: To evaluate DMGNN on the BML-MoVi dataset, we convert the dataset into the same input format expected in DMGNN. The joint position data we downloaded was preprocessed to a text file format from Numpy arrays. We have also visualized the data to understand the model better. As part of processing data, we have also created scripts and classes to make the data split as



(a) Overview of AMASS dataset

(b) Overview of Human3.6M dataset

Figure 4. Datasets and their subjects

mentioned in Fig 2.

5. Results

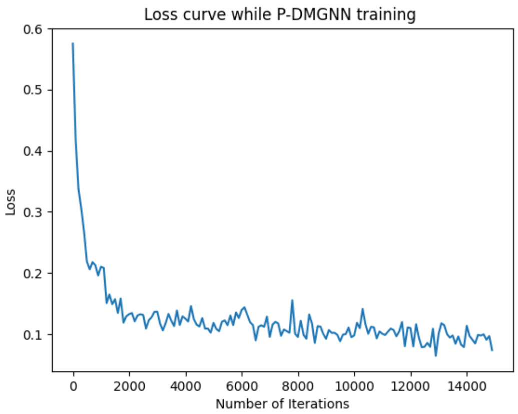
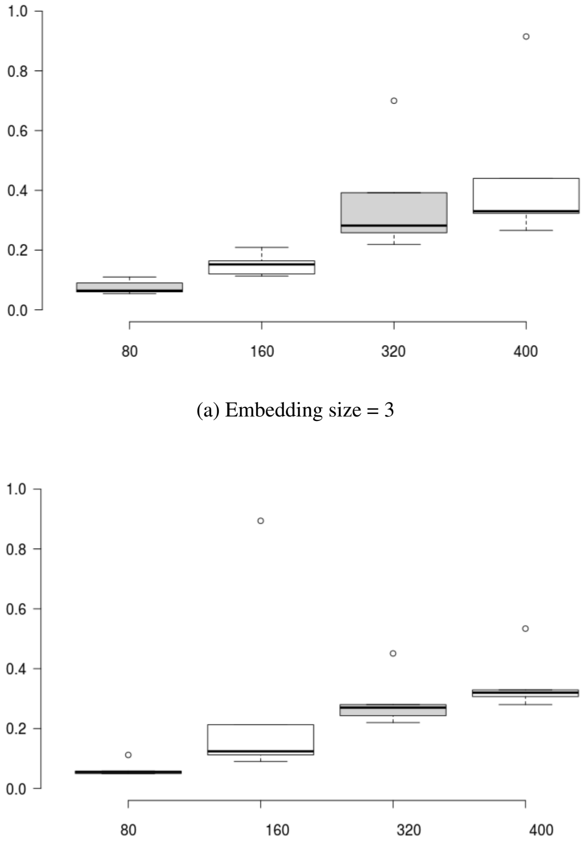


Figure 5. Loss curve in the training process of P-DMGNN

All the following results are based on the AMASS Dataset.

For prediction accuracy evaluation, we compared the MAE score and the average computation time cost for both our baseline DMGNN and two versions of P-DGMNN - with and without personalization calibration. Poses for 21 different classes (16 for training and 5 for testing) of actions were predicted 400ms into the future. Examples of four different action's MAE scores for each model are in Table 1 (lower is better). Both models were evaluated on the same machine after training for a fair comparison of computation time. The computation time to generate a predicted pose 400ms into the future for each model is in Table 2. Here we drop the computation time of the first action while testing since it includes the time of model loading.



(a) Embedding size = 3

(b) Embedding size = 10

Figure 6. MAE Scores of P-DMGNN with different embedding sizes

From the results in Table 1, we show that our method, P-DMGNN, with and without personalized calibration, outperforms the baseline in most cases for the five tested actions. This reinforces our previous hypothesis that personalization can increase the accuracy of human motion prediction. Additionally, more than half of the predictions generated by P-DMGNN with calibration have lower MAE scores than P-DMGNN without calibration, indicating that calibration can help the embedding layers better capture personalized features and improve the performance to an

Architecture	Action	**Baseline**	**P-DMGNN w/o calibration**	**P-DMGNN**
Running Time (ms)	18	36.230	36.453	36.387
	19	37.557	35.97	36.825
	20	33.360	38.359	37.427
	21	32.798	36.493	38.520

Table 2. Running time when evaluating networks for each action in the BML-MoVi dataset

extent.

As for computation time, we find that our model incurs a small runtime penalty on average in the predictions as shown in Table 2, just below 17%. Combined with MAE results, we believe it is a worthy trade-off between computation time and prediction accuracy.

To better understand the embedding layer parameters, we trained. P-DMGNN with different embedding sizes, 3 and 10 (3 is the default setting we used). The average MAE scores for all actions with different prediction times are shown in Figure 6. The upper figure uses a size of 3, and the lower one uses a size of 10. Interestingly, we do not find major differences between the performances, which indicates two possibilities: first, the current embedding size (3) we use can learn most personalized features well enough, or second, we need to try a much larger embedding size to observe a larger gap in performance.

6. Conclusion

Through our experiments, we have shown that including personalization can help to improve human motion prediction when using a GNN for human motion prediction. Additionally, we showed that calibrating to a specific subject can allow for greater accuracy improvements in human motion prediction. However, we note that this was not the case across all actions and hypothesize that the joint data we used may not have been expressive enough to create good personalized embeddings for those specific actions. One possible idea to extend our current work is to explore using datasets with higher fidelity data (e.g., increased joint count or mesh data) as it can potentially yield greater accuracy improvements, as this provides more data for the personalization network to capture latent traits in the subjects.

References

[1] Liat Antwarg, Lior Rokach, and Bracha Shapira. Attribute-driven hidden markov model trees for intention prediction. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(6):1103–1119, 2012. 2

[2] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 2180–2188, Red Hook, NY, USA, 2016. Curran Associates Inc. 2, 3

[3] 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*, 36(7):1325–1339, jul 2014. 1, 2, 4

[4] Maosen Li, Siheng Chen, Yangheng Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian. Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2, 3

[5] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of motion capture as surface shapes. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2019. 1, 2, 4

[6] Julieta Martinez, Michael Black, and Javier Romero. On human motion prediction using recurrent neural networks. pages 4674–4683, 07 2017. 1, 2

[7] Mariah L. Schrum, Erin Hedlund-Botti, Nina Moorman, and Matthew C. Gombolay. Mind meld: Personalized meta-learning for robot-centric imitation learning. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '22, page 157–165. IEEE Press, 2022. 2