A Appendix for ManiCast: Collaborative Manipulation with Cost-Aware Human Forecasting

We provide the proof for lemma 1 and include additional details about the MPC planner, the collaborative manipulation tasks, our dataset, and model implementation.

A.1 Proof for Lemma 1

Proof. We first analyze the performance of a model trained on $P(\phi)$.

Let's assume that a model trained with MLE loss on $P(\phi)$ bounds the average L1 distance between the ground truth distribution $P(\xi_H|\phi)$ and the learned distribution $P_\theta(\xi_H|\phi)$ on $P(\phi)$ by ε, i.e.

$$\sum_{\phi} P(\phi) \sum_{\xi_H} |P(\xi_H|\phi) - P_\theta(\xi_H|\phi)| \leq \varepsilon$$

Then, for any given ξ_R be the robot trajectory ξ_R, the final loss $\ell(\theta)$, i.e. the expected cost difference of ξ_R due to the ground truth distribution and the forecast model can be expressed as:

$$\begin{aligned}
\ell(\theta) =& \sum_{\phi} P(\phi) \left(\left| \sum_{\xi_H} C(\xi_R|\xi_H) P(\xi_H|\phi) - \sum_{\xi_H} C(\xi_R|\xi_H) P_\theta(\xi_H|\phi) \right| \right) \\
=& \sum_{\phi} P(\phi) \left(\left| \sum_{\xi_H} C(\xi_R|\xi_H) \left(P(\xi_H|\phi) - P_\theta(\xi_H|\phi) \right) \right| \right) \\
\leq& \sum_{\phi} P(\phi) \left(\|C(\xi_R|\xi_H, \phi)\|_\infty \sum_{\xi_H} |P(\xi_H|\phi) - P_\theta(\xi_H|\phi)| \right) && \text{(Holder's ineq.)} \\
\leq& \sum_{\phi} P(\phi) \left(C_{\max}(\phi) \sum_{\xi_H} |P(\xi_H|\phi) - P_\theta(\xi_H|\phi)| \right) \\
\leq& \max_{\phi} C_{\max}(\phi) \sum_{\phi} P(\phi) \left(\sum_{\xi_H} |P(\xi_H|\phi) - P_\theta(\xi_H|\phi)| \right) \\
\leq& C_{\max} \varepsilon
\end{aligned}$$

where $C_{\max}(\phi) = \|C(\xi_R|\xi_H, \phi)\|_\infty$ is the maximum cost of a robot trajectory given a context, $C_{\max} = \max_\phi C_{\max}(\phi)$ is the maximum cost across all context. $C_{\max}$ can be high in general, resulting in an inflated bound for the model above.

Now let's assume we train a model to minimize loss on the new distribution $Q(\phi) = 0.5P(\phi) + 0.5P_T(\phi)$ and get the following bound

$$\sum_{\phi} Q(\phi) \sum_{\xi_H} |P(\xi_H|\phi) - P_\theta(\xi_H|\phi)| \leq \varepsilon$$

Then the loss can be expressed as:

$$\begin{aligned}
\ell(\theta) =& \sum_{\phi} P(\phi) \left(\left| \sum_{\xi_H} C(\xi_R|\xi_H) P(\xi_H|\phi) - \sum_{\xi_H} C(\xi_R|\xi_H) P_\theta(\xi_H|\phi) \right| \right) \\
\leq& \sum_{\phi} P(\phi) \left(C_{\max}(\phi) \sum_{\xi_H} |P(\xi_H|\phi) - P_\theta(\xi_H|\phi)| \right) \\
\leq& \sum_{\phi} Q(\phi) \frac{P(\phi) C_{\max}(\phi)}{Q(\phi)} \sum_{\xi_H} |P(\xi_H|\phi) - P_\theta(\xi_H|\phi)| \\
\leq& \max_{\phi} \frac{P(\phi) C_{\max}(\phi)}{Q(\phi)} \sum_{\phi} Q(\phi) \sum_{\xi_H} |P(\xi_H|\phi) - P_\theta(\xi_H|\phi)|
\end{aligned}$$

For $Q(\phi) = 0.5P(\phi) + 0.5P_T(\phi)$, we need to bound the ratio

$$\max_{\phi} \frac{P(\phi)C_{\max}(\phi)}{Q(\phi)} = \max_{\phi} \frac{P(\phi)C_{\max}(\phi)}{0.5P(\phi) + 0.5P_T(\phi)}$$

There are two cases to consider:

Case 1: $C_{\max}(\phi) \leq \delta$. Then $P_T(\phi) = 0$, and the ratio is bounded by

$$\max_{\phi} \frac{P(\phi)C_{\max}(\phi)}{0.5P(\phi) + 0.5P_T(\phi)} \leq \frac{P(\phi)\delta}{0.5P(\phi)} \leq 2\delta$$

Case 2: $C_{\max}(\phi) \geq \delta$. Then ratio is maximized when $C_{\max}(\phi)$ is maximized for $\phi = \phi^*$

$$\max_{\phi} \frac{P(\phi)C_{\max}(\phi)}{0.5P(\phi) + 0.5P_T(\phi)} \leq \frac{P(\phi^*)C_{\max}(\phi^*)}{0.5P_T(\phi^*)} \leq \frac{P(\phi^*)C_{\max} \sum_{\phi} P(\phi)\mathbb{I}(C_{\max}(\phi) \geq \delta)}{0.5P(\phi^*)}$$
$$\leq 2C_{\max}\mathbb{E}_{P(\phi)}\left[\mathbb{I}(C_{\max}(\phi) \geq \delta)\right]$$

Combining these cases, we can bound the ratio $\max_{\phi} \frac{P(\phi)C_{\max}(\phi)}{Q(\phi)}$ as

$$\max_{\phi} \frac{P(\phi)C_{\max}(\phi)}{Q(\phi)} \leq 2\max(\delta, C_{\max}\mathbb{E}_{P(\phi)}\left[\mathbb{I}(C_{\max}(\phi) \geq \delta)\right])$$

The ratio above can be no worse than $C_{\max}$ by a factor of 2, and can be much smaller based on the choice of δ. Intuitively setting δ to be very high makes the transition probability $P_T(\phi)$ peaky driving down the second term, while making δ to be small makes the transition probability close to the original distribution, driving down the first term.

□

A.2 MPC Planner Details

We use the open-sourced STORM codebase[1] to implement sampling-based model-predictive control on a 7-DOF Franka Research 3 robot arm. At every timestep, the planner samples robot trajectories and evaluates the cost function with MANICAST forecasts. The robot executes the first action from the lowest-cost plan and updates its sampling distribution for the next timestep using the MPPI [1] algorithm. The manipulation components of the cost function independent of the human remain unchanged. We additionally introduce a collaborative task-specific cost component ($T(\xi_R|\hat{\xi}_H)$) that depends on the future human trajectory. The cost function optimized by the planner is laid out in Eq.1. Self-collisions are checked by training the jointNERF model introduced by Bhardwaj et al. [2].

$$C(\xi_R|\hat{\xi}_H) = \alpha_s\hat{C}_{stop}(\xi_R) + \alpha_j\hat{C}_{joint}(\xi_R) + \alpha_m\hat{C}_{manip}(\xi_R) + \alpha_c\hat{C}_{coll}(\xi_R) + \boldsymbol{\alpha_t}\mathbf{T}(\boldsymbol{\xi}_{\mathbf{R}}|\hat{\boldsymbol{\xi}}_{\mathbf{H}}) \tag{1}$$

A.3 Tasks for Collaborative Manipulation

We describe three collaborative manipulation tasks that focus on house-hold cooking activities.

Reactive Stirring: In this cooking task, the human and robot share a common workspace. While the robot arm is performing a stirring motion, the human may add vegetables into the pot. The robot arm preemptively predicts the arrival of the human arm and retracts back to give the human arm sufficient space to reach into the pot. The task-specific component of the cost function is:

$$T(\xi_R|\hat{\xi}_H) = \sum_{t=1}^{T} \mathbb{1}\left[D(\hat{s}_t^H, s^{pot}) \leq \varepsilon\right] \|s_t^R - s^{rest}\| + \mathbb{1}\left[D(\hat{s}_t^H, s^{pot}) > \varepsilon\right] \|s_t^R - \xi_{stir}^t\| \tag{2}$$

[1] https://github.com/NVlabs/storm

The cost function checks whether the human's position ($\hat{s}_t^H$) is close to the pot's position (s^{pot}) and decides whether to move to a pre-defined resting position (s^{rest}) or to continue stirring in a circular trajectory (ξ_{stir}) starting from the current state of the robot (s_0^R). A cost-aware forecasting model for this task should be able to predict the arrival and departure of the human ahead of time.

Human-Robot Handovers: Handovers of objects are an important task in the kitchen. When a human is handing over an object, a robot arm should move towards the intended handover location. The task-specific component can be described as:

$$T(\xi_R|\hat{\xi}_H) = \sum_{t=1}^{T} \mathbb{1}\left[IsObjectInHand(s_0^H)\right]\hat{C}_{pose}\left(X_t^{ee}, GraspPose(X_0^{ee}, \hat{X}_T^{H_{wrist}})\right) \tag{3}$$

Similar to prior work [3], the robot motion is initiated when the human arm has picked up the handover object. The robot's end-effector (X_t^{ee}) moves towards a grasp location that is computed using the final wrist position of the human ($\hat{X}_T^{H_{wrist}}$). The orientation of the grasp pose is calculated by drawing a straight line from the current end-effector position (X_t^{ee}) to the grasp location.

Collaborative Table Setting: Movements on top of a table in the presence of a human in the workspace are a common collaborative manipulation task. Motion planners should not only avoid collision in the current timestep but also be able to forecast future motion and preemptively avoid collisions with the human body. The cost function is simply given by:

$$T(\xi_R|\hat{\xi}_H) = \sum_{t=1}^{T} \hat{C}_{pose}\left(X_t^{ee}, X_t^G\right) + \beta \hat{C}_{coll}\left(s_t^R, s_t^H\right) \tag{4}$$

Here, β is the relative weight given to the collision avoidance component compared to the goal-reaching component. Collisions are checked between the human body and robot arm by representing them as a pack of sphere and cuboid rigid bodies.

A.4 Collaborative Manipulation Dataset (CoMaD)

Similar to a real-world collaborative activity, in much of the episode, both humans perform their respective cooking tasks in isolation. Episodes of reactive stirring and handovers contain 3-5 close-proximity interactions, each of which are short (4-5 seconds) compared to the length of the overall episode (30-60 seconds). Often, these interactions are initiated by verbal requests or subtle facial gestures. Collaborative table setting consists almost entirely of close-proximity fast human arm movements. We collect an RGB visual view of the scene containing audio along with motion capture data of both humans' upper bodies. We also annotate transition windows for interactions in each episode.

A.5 Model Implementational Details

We train our forecasting models using the STS-GCN [4] architecture on an upper body skeleton consisting of 7 joints (Wrists, Elbows, Shoulders, and Upper Back). The last 0.4 seconds (10 timesteps) of motion is input to the models and the next 1 second (25 timesteps) of motion is predicted. We pretrain for 50 epochs on AMASS (1 hour) and finetune on CoMaD for 50 epochs (5 minutes). We divided the episodes in CoMaD into train, validation, and test sets (8:1:1).

References

[1] G. Williams, N. Wagener, B. Goldfain, P. Drews, J. M. Rehg, B. Boots, and E. A. Theodorou. Information theoretic mpc for model-based reinforcement learning. *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1714–1721, 2017.

[2] M. Bhardwaj, B. Sundaralingam, A. Mousavian, N. D. Ratliff, D. Fox, F. Ramos, and B. Boots. Fast joint space model-predictive control for reactive manipulation. In *Conference on Robot Learning*, 2021.

[3] 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)*, pages 6956–6962. IEEE, 2022.

[4] 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)*, pages 11189–11198, 2021.