# Viewpoint-aware Legible Motion Planning with Imitation and Reinforcement Learning

Khai Nguyen
*Mechanical Engineering*
*Carnegie Mellon University*
Pittsburgh, PA
xuankhan@andrew.cmu.edu

Yi-Hung Chiu
*Mechanical Engineering*
*Carnegie Mellon University*
Pittsburgh, PA
yihungc@andrew.cmu.edu

Prakrit Tyagi
*Mechanical Engineering*
*Carnegie Mellon University*
Pittsburgh, PA
prakritt@andrew.cmu.edu

Saurav Kambil
*Mechanical Engineering*
*Carnegie Mellon University*
Pittsburgh, PA
skambil@andrew.cmu.edu

***Abstract*—Legibility is a crucial concept in terms of efficiency and trust in assistive robotics and human-robot collaboration, where the robot communicates its objectives through its actions in an understandable and predictable manner. Traditional motion planning techniques encounter various issues, including high computational latency, ambiguous objectives, and intensive tuning efforts. To overcome these challenges, we propose a universal planning architecture for learned legible behaviors using reinforcement learning and imitation learning. Furthermore, we introduce a novel planning model that considers human's viewpoint to generate adaptive motions that more effectively express intents. The effectiveness of our frameworks is validated through goal-reaching manipulation tasks conducted using the xArm6 robot in both simulated environments and real-world settings. Human-based evaluations indicate that our trained agent outperforms expert demonstrations by 15%. Our implementation can be accessed at https://github.com/BernieChiu557/xarm6-RL.**

***Index Terms*—reinforcement learning, imitation learning, legible motion planning.**

## I. Introduction

In the field of robotics, the concept of legible motion planning emerges as a cornerstone for enhancing human-robot interaction and shared autonomy [1]–[3]. Legible motion planning refers to the design and implementation of robotic movements that are easily interpretable by human observers, allowing for intuitive anticipation of the robot's future actions. This aspect of robotics is critical in applications where robots and humans work in close proximity, ranging from collaborative manufacturing environments to assistive robots in healthcare settings. The primary goal of legible motion planning is to foster trust and cooperation between humans and robots, ensuring that robotic actions are not only efficient but also predictable and understandable to humans.

The significance of making robotic actions easily interpretable to humans has been a pivotal advancement in robotics. This focus on legibility has led to the development of frameworks for evaluating how predictable and understandable a robot's movements are to people, which is critical for effective human-robot collaboration. Building on these foundational ideas, the research community has explored various methods to enhance the clarity of robot motions. These methods range from altering the paths robots take, adjusting the timing of their movements, to introducing more pronounced gestures that clearly communicate their intended actions. Such efforts aim to ensure that robots can operate in close proximity to humans, not only performing tasks efficiently but also in a manner that humans can easily understand and predict, thus facilitating smoother interactions and cooperation between humans and robots [1].

While advancements in robot legibility have greatly enhanced human-robot interaction, certain limitations persist that challenge the effectiveness and application of these concepts. One such limitation is the complexity of designing universally comprehensible cues that accurately convey a robot's intentions across diverse human populations and settings. The effectiveness of modified trajectories, timing adjustments, and exaggerated movements in signaling intentions can vary significantly based on the environment and the individual's interpretation. Additionally, ensuring these legibility cues do not mislead or confuse users remains a challenge, especially as robots become more prevalent in shared and public spaces. The quest for effective communication between robots and humans, thus, necessitates continuous refinement and testing of legibility strategies to address these concerns [1], [4].

On the other hand, learning-based approaches become increasingly popular in robotics planning and control since they offer a variety of advantages over traditional methods in terms of scalability, robustness, and efficiency. Firstly, deep learning models, as universal approximators, can represent complex systems and behaviors, which enables vision-based input or even end-to-end design [4]. Next, planning via neural network inference exhibits negligible latency [5], especially on GPUs, compared to computationally intensive optimization or search-based methods.

In the pursuit of legible motion planning, two prominent learning-based methods have emerged: reinforcement learning (RL) and imitation learning (IL). Reinforcement learning, characterized by a trial-and-error approach, allows robots to learn optimal actions through interactions with their environment, guided by a reward system that incentivizes legible behaviors. On the other hand, imitation learning focuses on teaching robots by example, where the machine learns legible motions by mimicking human demonstrations. Both methods offer unique advantages and challenges in the quest for legibility, with RL providing the flexibility of autonomous

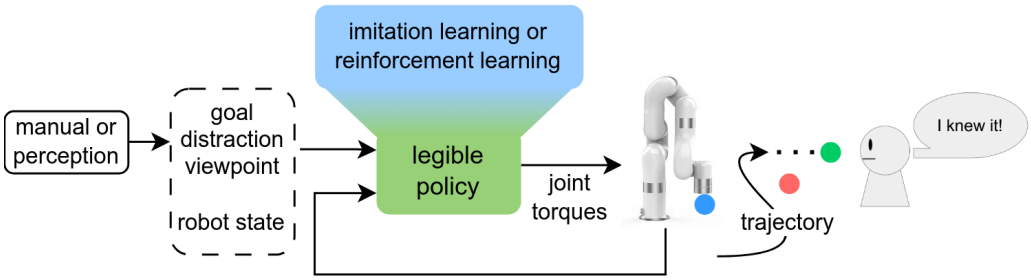



Fig. 1: System overview of our proposed learning-based legible motion planning frameworks. The policy which can be trained using imitation learning or reinforcement learning generates legible motions, enabling human collaborators to clearly understand the intent of the robot's movements. At each time step, the policy receives input including the state of the robot, the goal, the potential distraction that could confuse human collaborators, and the viewpoint of the human collaborators and predicts an action in the form of joint torques for the robot to execute. Since our focus is on the planning side, all observations except the robot state are specified manually. In real world, these observations can come from a perception module to enable full shared autonomy.

adaptation and IL offering the intuitiveness of human-like motion patterns.

Both RL and IL offer unique approaches to enhancing legible motion planning, yet the influence of observer viewpoints has been largely overlooked in these strategies. This limitation impacts human-robot collaboration as a robot's intent is often only clear from specific angles. Our research aims to integrate viewpoint conditioning into these learning-based algorithms to better align robotic behavior with human perceptions.

RL enables robots to autonomously identify optimal behaviors through rewards and penalties, ideal for environments lacking explicit behavior models and requiring adaptability. In contrast, IL uses human demonstrations to teach robots, aligning robotic actions with human norms and expectations, and is effective in contexts where human-like motion is essential. By separately investigating RL and IL, we gain insights into each method's capabilities and constraints in promoting legible motion. This comparative analysis helps identify the most effective techniques for various scenarios, enhancing future research and practical applications.

This paper delves into legible motion planning from the perspective of learning-based methods, specifically focusing on the promising applications of RL and IL (Fig. 1). We discuss methodologies for achieving legibility in robotic motions through each learning paradigm and evaluate the advantages of these approaches compared to previous work. Moreover, our novel idea of incorporating the user viewpoint aims to enhance the ways in which machines communicate their intentions to humans, thereby improving the interaction and cooperation between human and robotic agents.

Our main contributions include:

- A universal planning architecture for learned legibility using reinforcement learning and imitation learning.
- A novel legible motion planning model that considers human's viewpoint for effective human-robot collaboration.
- Evaluations with humans demonstrating the effectiveness of the proposed frameworks in real-world settings.

## II. Related Work

### A. Legible Motion Planning (LMP)

The field of legible motion planning (LMP) emphasizes refining robots' movements to enhance predictability and understandability, as pioneered by Dragan et al. [1]. They distinguished legibility, the ease of inferring a robot's goal from early trajectory cues, from predictability, which anticipates the robot's path given its goal. This work underlined the challenges of applying these concepts universally across varied tasks and user groups. A follow-up on these ideas was to take into account the viewpoint of observers in legibility. Nikolaidis et al. [3] proposed computing legibility by first projecting trajectory and any goals onto the plane aligned with the observers' point of view and then Dragan legibility is computed in the resulting space. This allows the robot to take into account the human's perspective and also allows it to account for occlusion from the perspective of the observer. Simplifying complex trajectories, Zhao et al. [6] demonstrated that straightforward paths often outperform in predictability, indicating the task context's significant role in legibility strategies. The literature explores various approaches to optimize robot motion for clear goal communication, from heuristic methods enhancing intent clarity to reinforcement learning techniques adapting based on user feedback. These studies underscore the critical balance between making robot actions both predictable and legible.

### B. From Reinforcement Learning to LMP

Reinforcement learning has been extensively explored for the motion planning of autonomous systems, with the incorporation of legibility achievable through engineered rewards or direct human feedback. Zhao et al. [7] utilize Deep Reinforcement Learning (DRL) to formulate a policy for robot motion, complemented by a Seq2Seq predictor to evaluate the legibility of said motion. This data-centric approach obviates the need for real human data by employing joint learning techniques. Nonetheless, this method might not accurately capture human anticipation of the robot's movements. Busch et al. [2] employ model-free RL coupled with a unique reward mechanism that integrates the joint execution time of both

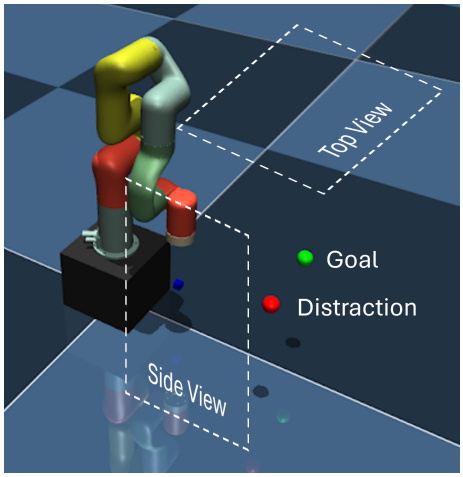



Fig. 2: Simulation setup for RL with the xArm robot. We use Mujoco as the physics engine and interface it through Gym for learning algorithms. The blue, green and red dots represent the positions of the end effector, goal and distraction, respectively. In human evaluation, we create videos of the same setup with the side and the top view.

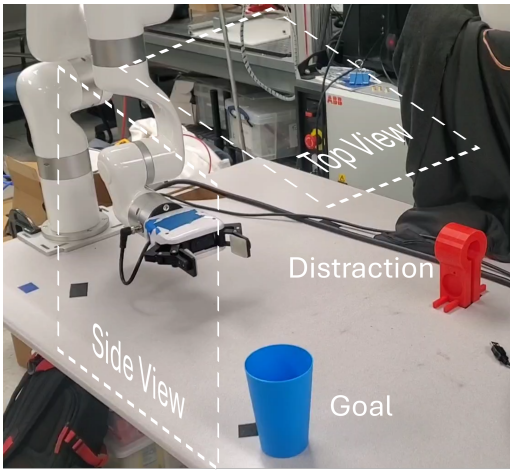



Fig. 3: Toy experimental setup for testing IL model with the xArm robot. The setup shows a reaching task towards the object as goal while legibly avoid the red object as distraction. The model also accommodate to different viewpoint condition as we have the same set up with the top and the side view in human evaluation.

humans and robots. However, this methodology necessitates a specialized setup, potentially costly or inaccessible for various tasks across different domains.

### C. From Imitation Learning to LMP

Imitation learning is a widely adopted approach for representing complex behaviors that are best acquired through expert demonstrations. Recent research by Wallkotter et al. [8] tackles this challenge by focusing on learning an observation model rather than the policy itself to devise motion plans. The observation model learns user preferences from labeled trajectories, enabling generalization, and subsequently, motion plans are sampled and passed to the observation model for legibility assessment. Lamb et al. [9] address a similar issue by introducing a bio-inspired behavioral dynamic model for cooperative pick-and-place behaviors. This model, based on low-dimensional dynamical movement primitives and nonlinear action selection functions, was effectively implemented as an artificial agent control architecture to produce human-like behavior during interactions with agents. The commonality among these approaches lies in learning a policy from data and utilizing it to generate legible behavior. This methodology is advantageous as it allows the learning of crucial aspects of legibility directly from the observer, potentially yielding more accurate results than manually crafting such a policy.

## III. Technical Approaches

We propose a universal learning-based legible motion planning architecture which is shared across both RL and IL setups, with the difference being the policy learning pipeline (Fig. 1). We focus on goal-reaching manipulation tasks where the robot is required to reach one out of two objects, referred as goal and distraction. To achieve legibility, it is necessary for the robot to move in a way that humans can easily understand and predict which object it wants to reach. Additionally, the robot's actions adjust based on the human's perspective of the task environment, such as observing from a top or side view.

### A. Reinforcement Learning-Based Approach

*1) Simulation Setup:* We developed a simulation of our xArm6 robot based on [10], using Mujoco [11] as the physics engine backend. Furthermore, we wrapped the simulation using OpenAI Gym [12] as a RL environment. Fig. 2 illustrates the simulation setup. The environment provides the robot configurations at each update, and the agent can control the torque applied at all six motors that correspond to each joint on the xArm. Noted that while we are using joint control in our simulation, lots of benchmark goal-reaching environments use end effector control (such as the Fetch environments in Gymnasium-Robotics). We found that since xArm6 has only 6 degrees of freedom, its joint space trajectory with end effector control is much more unstable compared to robots with higher degree of freedom, making RL training a lot harder (see our Phase 1 report).

*2) Proposed Model and Algorithm:* We represent the RL policy using a multi-layer perceptron (MLP) network with two layers, each consisting of 128 neurons. The simulation environment provides information about the robot configurations such as joint angles, joint velocities and end-effector/gripper pose. We randomly sample goal, distraction positions and viewpoint (top view or side view). These become observations for the RL agent to generate actions in the form of joint torques. In RL, the reward is key in producing required behaviors and needs proper hyper-parameter tuning. Our legibility reward is defined as follows

$$r = -\alpha_1 d_{\text{goal}} + \alpha_2 g(d_{\text{distract}}, \beta_1) - \alpha_3 e_{\text{rot}} + \alpha_4 g(z, \beta_2). \quad (1)$$

$$g(x, y) = -1 \text{ if } x \leq y, \text{ otherwise } 0. \quad (2)$$

| Hyper-parameter | Value |
|---|---|
| Reward | |
| $\alpha_1$ | 6.0 |
| $\alpha_2$ | 3.0 |
| $\alpha_3$ | 0.5 |
| $\alpha_4$ | 1.0 |
| $\beta_1$ | 0.1 |
| $\beta_2$ | 0.0 |
| SAC | |
| `buffer_size` | 1000000 |
| `batch_size` | 256 |
| `gamma` | 0.95 |
| `learning_rate` | 0.003 |
| `MLP` | [256, 256] |

TABLE I: Hyper-parameters for reinforcement learning.

| Hyper-parameter | Value |
|---|---|
| input Layer neurons | 19.0 |
| hidden Layer neurons | 40.0 |
| no. of Hidden layers | 2.0 |
| output Layer neurons | 6.0 |
| activation function | tanh |
| no. of Epochs | 1000 |
| batch size | 1708 |
| learning rate | 0.001 |

TABLE II: Hyper-parameters for imitation learning.

where $d_{\text{goal}}$ and $d_{\text{distract}}$ are the Euclidean distances from the gripper to the goal and distraction positions; $e_{\text{rot}}$ is the absolute rotation error between the gripper's and goal's poses. The first term encourages the robot to reach the goal; the second term encourages the robot to stay away from the distraction region of radius $\beta_1$; the third term encourages the robot to fix it gripper orientation for consistency; the last term encourages the robot to move above the table at least by $\beta_2$ m. This reward is computed using projected (vertical or horizontal) 2D coordinates given the side or top viewpoint (illustrated in Fig. 2 and 3). The task is considered successful when the robot manages to arrive at the goal within a small tolerance.

We then utilize the soft actor-critic algorithm (SAC) [13] to train the RL policy based on observation, action, and reward data. Our choice was based on SAC's effectiveness as an off-policy algorithm that seeks to find a balance between maximizing expected return and entropy, thereby enhancing the policy's robustness. This connection is closely related to the trade-off between exploration and exploitation: By raising entropy, we promote greater exploration, which could accelerate the learning process. Moreover, it serves to avoid the policy from prematurely settling on suboptimal local solutions. The hyper-parameters for the reward function and SAC algorithm are shown in Table I.

### *B. Imitation Learning-Based Approach*

*1) Expert Demonstrations:* In its simplest form, imitation learning, also known as behavior cloning, involves physically manipulating the robot to generate the dataset. The robot is manually operated in a way considered legible, and these observation-action pairs are recorded as expert demonstrations. These movements' trajectories are then saved based on the robot's joint states. This setup, depicted in Fig. 3, mirrors the setup for RL. For straightforward tasks like reaching a goal object amidst distractions, past attempts at imitation learning suggest that a few dozen trajectories could suffice. To balance accuracy with cost-effectiveness in gathering demonstrations, we opted for 30 trajectories as a starting point for our pilot study. In this phase, we conducted 30 expert demonstrations each for two view points top and side ways, wherein the robot was guided manually towards the target object, positioned 50 cm away, to demonstrate the desired motion. While recording joint state trajectories the position of distraction and goal were varied widely to cover the area in front of the robot.

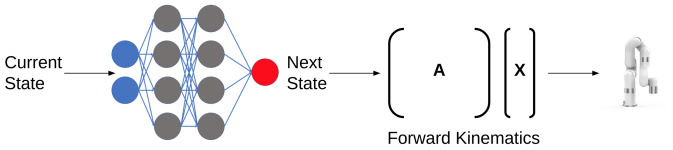



Fig. 4: Imitation learning detailed pipeline. We represent the IL policy using a multi-layer perceptron (MLP) network. It receives the state of the robot as input and output a prediction of the next state in order to achieve legible motion. The model operates in joint space, hence we need to use forward kinematics to transform it to end effector position for robot controller.

*2) Proposed Model and Algorithm:* We represent the IL policy using a multi-layer perceptron (MLP) network with three hidden layers (40 neurons for each), and `tanh` activation function. The input to the policy includes goal and distraction positions, robot state and viewpoint. Here, the state comprises of a 6-dimensional vector containing joint angles, each corresponding to a joint on the xArm6. The model then predicts the subsequent state of the robot and feeds it to a forward kinematics block, which furnishes an end-effector position to the position controllers of the robot. This process is illustrated in Fig. 4. We conducted supervised learning on the expert data to obtain a mapping from observation to action. Hyper-parameters for imitation learning are shown in Fig. II.

The data collected from the xArm robot is sampled at 250 Hz. In our post-processing routine, we downsample these trajectories by a factor of 20. This step is pivotal for the efficiency of our network in imitation learning contexts, where error propagation through consecutive time steps can significantly degrade performance. By reducing the number of predictions the model needs to make for each trajectory, we enhance the success rates during task execution, a finding that is substantiated by our research. Given our requirement for the model to predict a trajectory's next state based solely on the current state—without leveraging a history of previous states—we opt not to utilize Long Short-Term Memory (LSTM) networks. This choice aligns with our objective to minimize dependency on sequential past data, ensuring our model's focus remains on predicting future states from immediate, singular inputs.

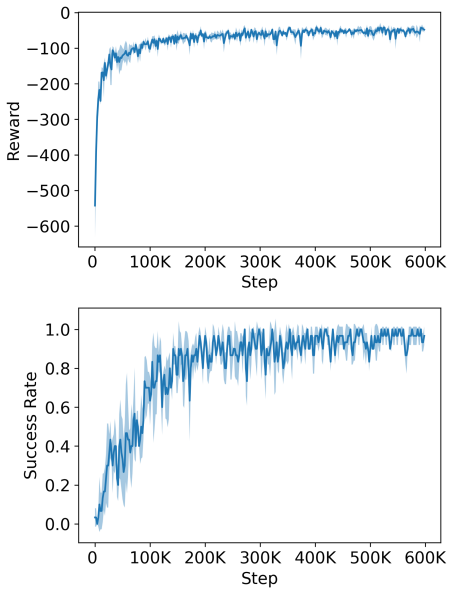



Fig. 5: Learning progress of the reinforcement learning policy across 3 different seeds. Top, the cumulative rewards increased consistently and converged quickly to around -50 at the end. Bottom, the success rates reached almost 100% after 200K iterations.

## IV. Results and Discussion

### A. Policy Training Results

*1) Reinforcement Learning:* The reward and success rate progression of the SAC policy are illustrated in Fig. 5, we ran the same settings with 3 different seeds and plot the mean and standard deviation to see the consistency of the training. The policy demonstrated effective training, converging to 100% success rate and $-50$ reward after roughly 200K steps. The training is stable as the standard deviation for reward is quite small while success rate has a bigger variation due to the discrete nature of thresholding and infeasible randomization.

*2) Imitation Learning:* Our training efforts resulted in a successful imitation learning (IL) policy, achieving data overfitting within 1000 iterations. However, upon deployment, the model's performance was subpar. The robot arm exhibited an initial drastic movement towards the target followed by a sudden halt. Further analysis revealed that the model was predicting excessively large time steps despite the data being recorded at 250 Hz. To address this issue, we down-sampled the training data to 20 Hz and retrained the model for 1000 iterations. This time, our robot arm behavior aligned with expectations, faithfully replicating the expert demonstration.

*3) Unified Behaviors:* Our results for both RL and IL provide compelling evidence on the critical role of the observer's viewpoint in the legibility of robotic trajectories (Fig. 6). When observed from a side view, it is necessary for the robot to execute a trajectory that ascends vertically to circumvent a distraction object colored in red. This path ensures the trajectory remains clear and purposeful. Conversely, when the same policy is applied from a top view, the trajectory's legibility diminishes; the robot appears to traverse directly towards the distraction, following a straight path. To maintain legibility from this viewpoint, the robot must instead navigate around the object on the plane of the table.

This disparity illustrates that even under identical conditions, the perception of a trajectory as legible can vary dramatically depending on the viewer's perspective. These findings underscore the importance of viewpoint conditioning in designing trajectories that are universally understood across different observational perspectives. Moreover, our model can be easily generalized to continuous viewpoint characterization with the use of perspective projections.

While RL is able to search for the best behaviors given a carefully engineered reward, IM implicitly learns the hard-to-specified objectives but being limited to the dataset diversity. Our paper serves as a foundational framework for learning-based legible motion planning where both methods are extensively evaluated. Our architecture is universal and can be used with optimization-based techniques as well.

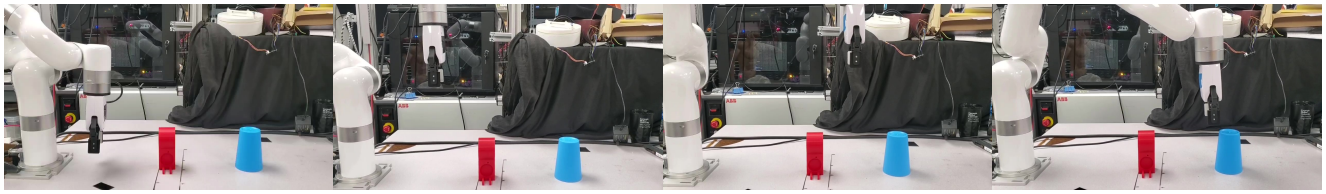
### B. Human Evaluation Results

As legibility measure the ability to produce motion sequences that are obvious to human, it is hard to create a metric that accurately follows the human preference. Hence, it is imperative to conduct human evaluation in addition to the metrics we set for our RL and IL training pipelines. We created two videos each for expert, RL and IL policies. Additionally we intentionally created an ablated scenario with RL (called RL-Mix in Fig. 7) to assess the importance of viewpoint conditioning. The model in RL-Mix is conditioned with side view for the top view video and is conditioned with top for the side view video. In total, 10 people are asked to watch all the 8 videos (4 are shown in Fig. 6). They were asked to pause the video as soon as they think they recognize the robot intention and then report the object they think the robot is going for. We constructed a score system to balance these factors as follows:

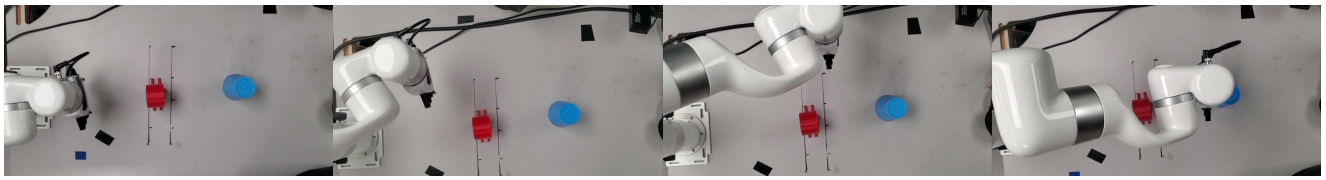
$$\text{score} = \frac{t_{\text{video}} - t_{\text{pause}}}{t_{\text{video}}} \times \texttt{check}(\text{guess}) \tag{3}$$

$$\texttt{check}(\text{guess}) = \begin{cases} 1 & \text{if guess correct} \\ 0 & \text{if guess wrong} \end{cases} \tag{4}$$

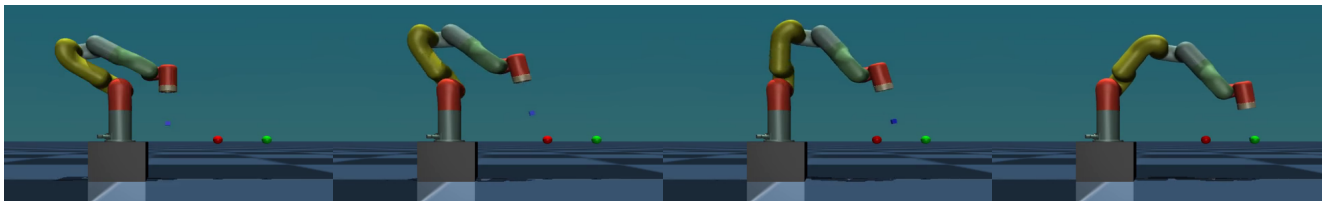
Comparing the scores in Fig. 7, the IL agent achieved almost the same average score as expert demonstration with a better performance in top view. As for RL, it achieved a roughly 15% higher score compared to the IL agent and expert. We speculate that it's due to the faster motion in RL videos that generally makes recognizing the robot intention easier. Besides, RL agent tries to maximize the legibility reward by interacting with the environment, hence being more effective. On the other hand, for RL-Mix, top view video has a expected poor result of 0.281 as the robot reaching the goal but not avoiding the distraction in the top view. This is ambiguous to viewers, not only causing longer pause times but also leading to some incorrect answers. However, side



(a) Side view of imitation learning legible trajectory.



(b) Top view of imitation learning legible trajectory.



(c) Side view of reinforcement learning trajectory.

(d) Top view of reinforcement learning trajectory.

Fig. 6: Snapshots of viewpoint-conditioned learned policy for imitation learning and reinforcement learning. Our policy is able to execute legible motion plans for two views, side view (a) and (c), and top view (b) and (d).

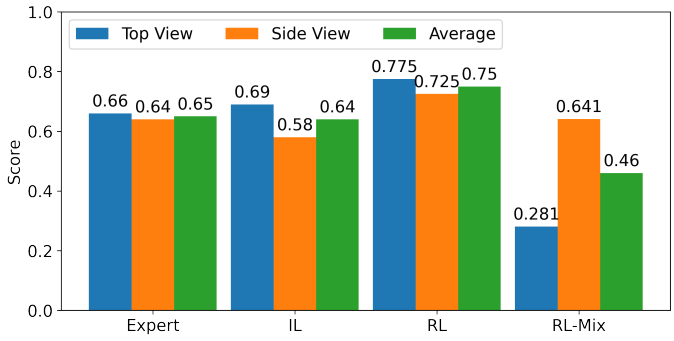



Fig. 7: Human-based evaluation results on the recorded legible motions. While our IL-trained agent performed competitively well compared to the expert demonstrations, our RL-trained one outperformed the others by 15%. Without proper viewpoint conditioning, RL-trained agent obtained a low score.

view in RL-Mix surprising still performed comparably well to the normal settings. This means that the robot actually went around the distraction both from the top view and the side view at the same time. This suggests that there are still rooms for improvement for the RL legibility algorithm, but overall, the human evaluation demonstrates the ability of RL and IL agents to produce compelling legible motion plans that are aligned with human preferences.

## V. Conclusions and Future Work

In this paper, we first introduce a universal planning architecture for generating legible motions, realized and evaluated using two learning-based approaches. While the reinforcement learning (RL) agent attempts to maximize a legibility reward by interacting with the simulated environment, the imitation learning (IL) agent performs supervised learning on legible expert demonstrations collected in real world. Furthermore, we propose a novel planning model that incorporates human's viewpoint to produce adaptive motion plans. We demonstrated the proposed legibility frameworks in goal-reaching manipulation tasks on an xArm6 robot. Human evaluations show that our IL-trained agent can perform comparatively well while our RL-trained agent outperforms the expert baseline by 15%. Moreover, viewpoint conditioning is shown to be essential in achieving robot legible motion.

Future work will evaluate the proposed frameworks in more complex tasks including multiple dynamic objects and compare with more diverse baselines. We also plan to bridge the gap between simulation and real world for RL and incorporate richer sensory inputs such as perception to generate more adaptive legible motions and enable full shared autonomy.

## REFERENCES


[1] A. D. Dragan, K. C. Lee, and S. S. Srinivasa, "Legibility and predictability of robot motion," *Proceedings of the 8th ACM/IEEE International Conference on Human-robot Interaction*, pp. 301–308, 2013.

[2] B. Busch, J. Grizou, M. Lopes, and F. Stulp, "Learning legible motion from human–robot interactions," *International Journal of Social Robotics*, vol. 9, no. 5, pp. 765–779, 2017.

[3] S. Nikolaidis, A. Dragan, and S. Srinivasa, "Viewpoint-based legibility optimization," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2016, pp. 271–278.

[4] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," 2023.

[5] M. M. H. Shuvo, S. K. Islam, J. Cheng, and B. I. Morshed, "Efficient acceleration of deep learning inference on resource-constrained edge devices: A review," *Proceedings of the IEEE*, 2022.

[6] M. Zhao, R. Shome, I. Yochelson, K. Bekris, and E. Kowler, "An experimental study for identifying features of legible manipulator paths," in *International Symposium on Experimental Robotics*. Springer, 2014.

[7] X. Zhao, T. Fan, D. Wang, Z. Hu, T. Han, and J. Pan, "An actor-critic approach for legible robot motion planner," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 5949–5955.

[8] S. Wallkötter, M. Chetouani, and G. Castellano, "Slot-v: supervised learning of observer models for legible robot motion planning in manipulation," in *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2022, pp. 1421–1428.

[9] M. Lamb, P. Nalepka, R. W. Kallen, T. Lorenz, S. J. Harrison, A. A. Minai, and M. J. Richardson, "A hierarchical behavioral dynamic approach for naturally adaptive human-agent pick-and-place interactions," *Complexity*, vol. 2019, pp. 1–16, 2019.

[10] J. C. Ramirez, "xarm6-gym-env," https://github.com/julio-design/xArm6-Gym-Env, 2022.

[11] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012, pp. 5026–5033.

[12] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "Openai gym," 2016.

[13] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International conference on machine learning*. PMLR, 2018, pp. 1861–1870.