2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) | 979-8-3503-7770-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/IROS58592.2024.10802187

Beyond Success: Quantifying Demonstration Quality in Learning from Demonstration

Muhammad Bilal[a], Nir Lipovetzky[a], Denny Oetomo[b], and Wafa Johal[a]

***Abstract*— Learning from Demonstration (LfD) empowers novice users to teach robots daily life tasks without writing sophisticated code, thereby promoting the democratization of robotics. However, novice users often provide sub-optimal demonstrations, which can potentially impact the robot's ability to efficiently learn and execute the tasks. Prior research has assessed the quality of demonstrations by evaluating the robot's task performance; however, the approach remains insufficient to qualify individual demonstrations, leaving the reason for classifying demonstrations as high- or low-quality unknown. Therefore, this simulation-based study aims to quantify the quality of individual demonstration at each step by incorporating motion-related quality features such as manipulability and joint-space jerk. To assess the efficacy of these features, we initially evaluated the given demonstrations—taking into account each quality feature—to rank them from high- to low-quality. Subsequently, we investigated the impact of demonstration's quality on task performance and the quality of task execution. In this pursuit, we trained a series of LfD models for distinct manipulation tasks: cube lifting and pick-and-place of soda can. Our results illustrate a strong correlation between ranked demonstrations and the quality of task execution. Interestingly, we observed that the quality features have a significant impact on task performance, particularly when the provided demonstrations exhibit diversity in terms of quality. Overall, this analysis enables quantifying the quality of individual demonstrations based on motion-related quality features, thus improving learning from demonstration.**

***Index Terms*—Learning from demonstration, quality of demonstration, quality of task execution, task performance**

I. Introduction

To overcome the obstacle of using robots for daily life tasks without programming skills, the 'Learning from Demonstration' (LfD) approach offers a promising solution. This strategy, inspired by human interactions, enables novice users to instruct robots through demonstrations instead of navigating the complexities of coding [1]. Through LfD, users can seamlessly impart tasks to robots, making the integration of robotic assistance more accessible and user-friendly for individuals without programming expertise.

One of the key goals in LfD is to enable a robotic system to efficiently learn and execute tasks with a minimal set of demonstrations. Achieving this goal is heavily contingent upon the quality of the provided demonstrations as well as modality of demonstration [2], as it directly shapes the overall competency of the robotic system. According to Laskey et al. [3], LfD algorithms fall into two broad categories: robot-centric and human-centric. In the former type, the robotic system initiates its learning process from a given set of demonstrations and then actively learns from its surroundings, incorporating feedback provided by the demonstrator [4]. On the other side, the human-centric approach involves learning an LfD model from a set of demonstrations provided by novice users; an example of this approach is imitation learning [5]. In the literature, researchers have devoted considerable attention to the robot-centric side, with comparatively less emphasis on the human-centric aspect for enhancing LfD [6]. In the domain of LfD, the *quality of demonstration* encompasses the dexterity and smoothness exhibited by novice users [2]. Meanwhile, the *quality of task execution* gauges the robot's aptitude to carry out the intended tasks, taking into consideration features such as manipulability [7] and joint-space jerk [8]. This dual focus on both the quality of demonstration and the quality of task execution significantly contributes to the enhanced performance and adaptability of the robotic system within the LfD paradigm.

Existing evaluation of demonstration's quality mainly focuses on appraising the robot's task performance, classifying demonstrations as either high-quality or low-quality based on the success or failure of the task [9], [10]. While quality assessment of task performance is crucial for human-centric LfD, previous approaches do not allow for assessment of individual demonstration. Instead, the task performance's evaluation is typically applied to a set of demonstrations utilized for training an LfD model—a machine learning model that reproduces actions observed during demonstrations—employing task performance as the primary metric. The inclusion of a learning phase to train the LfD model makes this process potentially time-intensive. Furthermore, the underlying reason for the given demonstrations being of high or low quality remains unclear to the user. In response to these limitations, it becomes crucial to systematically measure and assess each demonstration, taking into consideration various quality features. Feature-based methods [11], [12] present a promising approach to address the mentioned limitations, specifically the issues of being time-intensive and lacking clarity on why a demonstration is classified as high or low quality. Therefore, this study aims to investigate a set of quality features by quantifying individual demonstrations at each step based on quality features and subsequently assessing the impact of the demonstration's quality on task performance and the quality of task execution.

This study makes three primary contributions: firstly, it advocates for the utilisation of motion-based features to quantify the quality of individual demonstrations at an early stage and

[a]Muhammad Bilal, Nir Lipovetzky, and Wafa Johal are with the School of Computing and Information Systems, The University of Melbourne, Parkville, Victoria 3010, Australia (email: m.bilal; nir.lipovetzky; wafa.johal@unimelb.edu.au)

[b]Denny Oetomo is with the Department of Mechanical Engineering, The University of Melbourne, Parkville, Victoria 3010, Australia (email: doetomo@unimelb.edu.au)

This work was partially supported by the Australian Research Council (Grant No. DE210100858)

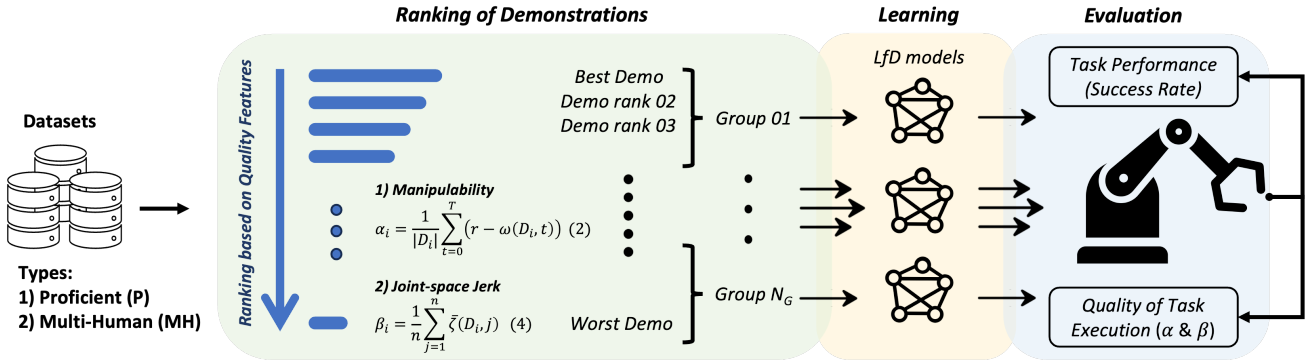


Fig. 1: **Overview** The given datasets have been evaluated based on the quality features, including manipulability and joint-space jerk, and ranked them from high to low quality. At next step, we trained an LfD model for each group, where N_G denotes the total number of groups, followed by rollouts the learned LfD model in order to compute the task performance and assess the quality of task execution based on the quality features.

regardless of learning an LfD model. Secondly, it highlights the transferability of quality features from provided demonstrations to the corresponding LfD model generated motions. Finally, it posits that our quality features can serve as an indicator of task performance when used in a diverse set of demonstrations.

II. Background and Related Work

A. Assessing Novice Users' Input in LfD

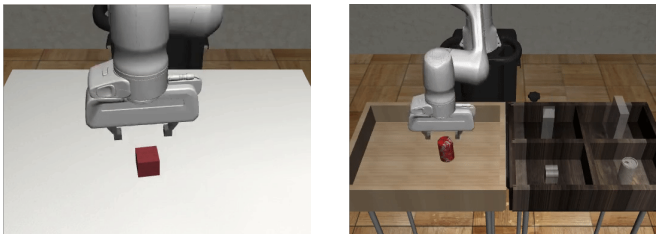
Several LfD algorithms assume that users are proficient and can provide optimal demonstrations [2]. However, this assumption often neglects the heterogeneous nature of human demonstrators, leading to sub-optimal demonstrations [13]. The quality of demonstrations can vary from user to user, influenced significantly by their understanding of the task and their capabilities to provide effective demonstrations [14].

Pais et al. [15] proposed three metrics to evaluate the user's ability, including maneuvering the tool, consistency in teaching, and arm coordination. The metric such as consistency in teaching may not inherently depict the quality of a demonstration, as users could potentially exhibit similar errors across different demonstrations. Fischer et al. [16] compared teaching modalities in LfD, and used three common issues associated with novice operators: 1) self-collision, 2) singularity, and 3) excessive force on the end-effector. However, multiple ways of controlling the robot have been studied but did not evaluate the impact of these issues on the learned LfD model.

To evaluate the diversity of demonstrations across task space, Sena et al. [17] proposed two metrics: 1) teaching efficacy and 2) teaching efficiency. These metrics alone are insufficient for assessing the quality of individual demonstrations, as they primarily address the issue of data sparsity. Furthermore, the concept of entropy has been utilized to identify areas where additional demonstrations are required for effective robot learning [18]. The entropy approach is an interesting but focuses on diversity of the demonstrations instead of quantification of individual demonstrations at each step.

B. Assessing Quality of Task Execution

To quantify the quality of task execution, previous research collected six core quality features: *manipulability*, *jerk* at Cartesian and joint-space, *trajectory length* in Cartesian and joint-space, and



(a) Lift task (b) Pick-and-place of soda can

Fig. 2: Manipulation tasks using Robosuite platform [19]

robot's joint limits [9]. Nevertheless, these features have not been utilized for quantifying the quality of demonstrations. Among these quality features, Chen et al. [8] investigated the importance of smoothness in robot learning, where they proposed an approach to identify and eliminate noises by providing alternative optimal control commands to the robot. Secondly, the manipulability is an important feature to consider because it directly impacts the maneuverability of the robotic system [7]. Furthermore, Jaquier et al. [20] proposed an interesting approach to learn and replicate manipulability ellipsoids from provided demonstrations. However, we aim to investigate the impact of quality features irrespective of learning particular feature from given demonstrations.

Our work introduces innovation by explicitly assessing individual demonstration based on quality features and evaluating their influence on both task performance and the quality of task execution. We selected two quality features—manipulability and joint-space jerk—based on the relationships established by [2]. To clarify, high-quality demonstrations are those with the highest measures of these features, while low-quality demonstrations have the lowest measures. We hypothesize:
H1 The quality of demonstration correlates with task performance.
H2 The quality of demonstration correlates with the quality of task execution.

III. Features Formulation

In this section, we present the mathematical formulation of motion-related quality features, including manipulability and joint-space jerk. Additionally, we provide a detailed description of the ranking criteria applied to the given demonstrations.

A. Manipulability

Manipulability measure quantifies how far the current configuration of a robotic system is from a singular configuration, where the robotic system cannot move in a specific direction [21]. We selected the manipulability measure because it directly represents the robot's dexterity. In assessing the demonstrations, higher measures of manipulability correspond to high-quality demonstrations, while the opposite holds for lower measures of manipulability.

Consider a robot with n degrees of freedom whose joint variables are denoted by $\vec{q} = \langle q_0, q_1, ..., q_n \rangle$, where $q_j \in \mathbb{R}$. A demonstration (D) can be defined as a joint-space trajectory $D = \langle \vec{q}_0, ..., \vec{q}_T \rangle$ over the time period $0 \leq t \leq T$, where $\vec{q}_t$ and T represent the joint configuration of the robot at time step t and the total duration of the demonstration, respectively.

Let $\mathcal{D} = \langle D_1, D_2, ... , D_K \rangle$ depict the set of all given demonstrations, where D_i indicates the i^{th} demonstration. The total number of demonstrations is represented by K, where $K = |\mathcal{D}|$. For a task with m dimensions, the Jacobian matrix $J(\vec{q}) : \mathbb{R}^n \rightarrow \mathbb{R}^{m \times n}$ is computed for the given joint configuration $(\vec{q})$ of the robot. The manipulability measure (ω) for demonstration D_i at time step t is then expressed as:

$$\omega(D_i,t) = \sqrt{det\ J(\vec{q}_t) J'(\vec{q}_t)} \tag{1}$$

Finally, we computed the manipulability based metric $\alpha_i \in \mathbb{R}$ for demonstration D_i, by finding the average difference between the manipulability measure at each time step and the global maximum value $r = \{max \langle \vec{\omega}(D_1), \vec{\omega}(D_2), ..., \vec{\omega}(D_K) \rangle\} \in \mathbb{R}$, where $\vec{\omega}(D_i) = \langle \omega(D_i,0), ..., \omega(D_i,T) \rangle$.

$$\alpha_i = \frac{1}{|D_i|} \sum_{t=0}^{T} (r - \omega(D_i,t)) \tag{2}$$

A higher α value indicates a poorer manipulability measure, representing a low quality demonstration, and vice versa.

B. Joint-space Jerk

For a robot joint index j in a given demonstration D_i, the jerk $\zeta(D_i,j,t)$ is computed by taking the triple derivative of the robot's joint position for the given period $0 \leq t \leq T$.

Next, the mean jerk value $\overline{\zeta}(D_i,j) \in \mathbb{R}$ is calculated followed by computing a jerk based metric $\beta_i \in \mathbb{R}$ for demonstration D_i, indicating how much demonstration D_i contains jerky motion at joint-space level.

$$\overline{\zeta}(D_i,j) = \frac{1}{|D_i|} \sum_{t=0}^{T} \zeta(D_i,j,t) \tag{3}$$

$$\beta_i = \frac{1}{n} \sum_{j=1}^{n} \overline{\zeta}(D_i,j) \tag{4}$$

Based on the β value, ranking is determined with lower values indicating smoother demonstrations, while higher values correspond to a higher magnitude of jerks, indicating low quality demonstrations.

IV. Analyses Setup

In this section, we discuss the dataset and the procedure employed for computing task performance. Additionally, we elaborate on the criteria for selecting an algorithm for training LfD models, focusing on task performance while utilizing a minimum number of demonstrations.

A. Dataset

In this study, we used the robomimic v0.1 dataset for the analyses [22], [23]. The demonstrations in the dataset were collected using the RoboTurk tool [24]. Among multiple tasks, we considered the lift and pick and place tasks because they encapsulate fundamental manipulation skills such as grasping, lifting, and placement as shown in Fig. 2. To investigate the influence of diverse-quality demonstrations, we considered both a "proficient" (P) and a "multi-human" (MH) dataset, with a total of 200 and 300 successful demonstrations for each manipulation task, respectively. In the multi-human dataset, the demonstrations were gathered by six operators with varying expertise, comprising two proficient, two normal, and two worst performers. Each individual contributed a total of 50 demonstrations, resulting in a dataset with mixed quality. On the other side, the proficient dataset consists of demonstrations provided by expert user.

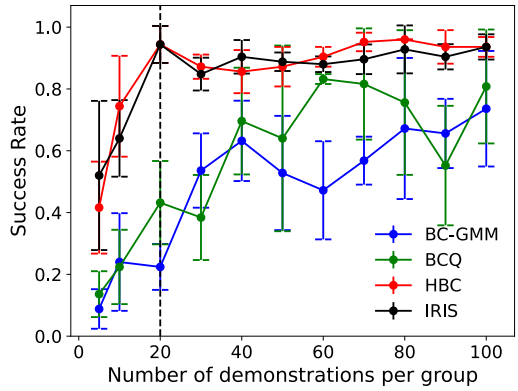
B. Performance Criteria

To compute the success rates, we utilized the learned LfD models—obtained using one of the LfD algorithms—to execute the same task as observed during the demonstrations. In both the demonstration and execution phases, the initial pose of the object was set randomly. For the lift task, the learned model requires to successfully lift a cube to a specific height of $0.06m$, while for the pick-and-place of soda can, the robot needs to pick up a soda can and accurately place it in its designated location (see Figure 2b).

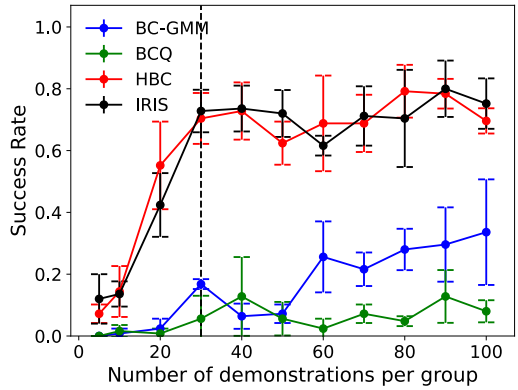
C. Learning Algorithms

In real-world scenarios, it is often impractical for human users to provide a large number of demonstrations; therefore, one of the objectives in LfD is to train a model with a minimum number of demonstrations [25]. In this pursuit, we propose to identify which of the algorithm provided by Mandlekar et al. [22] exhibits the best performance with a minimal number of demonstrations. We test: Behavioral Cloning (BC) [26], Hierarchical Behavioural Cloning (HBC) [27], Batch-Constrained Q-Learning (BCQ) [28], and Implicit Reinforcement without Interaction at Scale (IRIS) [29].

The BC algorithm executes simple regression over a sequence of state-action pairs in the given dataset. Several variants of BC algorithm, including BC based on Gaussian mixture model (GMM), have been investigated for different types of manipulation tasks [22]. The HBC and IRIS algorithms are identical, except the latter includes the addition of a value function in a high-level mechanism to locate a state with the maximum expectation. In this way, the IRIS algorithm is suitable for identifying near optimal state-action pair. Lastly, the BCQ algorithm is an offline batch reinforcement learning [28]. In this study, we considered an offline learning, where the algorithms are not permitted to consider additional samples.



(a) Lift task



(b) Pick-and-place of soda can

Fig. 3: This figure illustrates the performance of algorithms for the given different number of demonstrations per group, considering both the (a) lift task and (b) pick-and-place of soda can.

For each algorithm, we computed the success rates — representing the number of successful task executions divided by the total number of rollouts [22] — for each set of demonstrations. The term "rollout" represents the deployment of the learned LfD model on a robotic system to evaluate its performance on the given task. To determine the minimal number of demonstrations to achieve an acceptable success rate, we devised 10 groups of varying sizes ranging from 10 to 100 demonstrations, with an incremental factor of 10 demonstrations. Additionally, we devised one group which consist of 5 demonstrations. Overall, this resulted in training 44 models per task, with 11 models for each algorithm. To validate the effectiveness of the models' performance, we employed the k-fold cross-validation technique for each group.

D. Experiment Parameters

Observing the learning curves [22], a suitable number of epochs for both the lift and pick-and-place tasks is 250, utilizing a batch size of 100. Next, the learned model was subjected to 25 rollouts after every 50 epochs, allowing the computation of success rates. To ensure the adaptability of the learned model to an actual robotic system, the rollouts were performed on the same robotic system using the Robosuite platform [19]. Throughout the experiments, the demonstration data was partitioned into training and validation sets with a split ratio of 80% and 20%, respectively.

V. Selection of Learning Algorithm

In our preliminary analyses, the algorithm with the minimum number of demonstrations required to achieve an acceptable success rate of 90% and 70%, values observed from the learning curves where the algorithm converged [27], for the lift and pick-and-place tasks, respectively. For each algorithm, the models were trained on the same set of demonstrations using 5-fold cross-validation with same performance criteria and experimental parameters. Additionally, the multi-human dataset was used for both manipulation tasks: lift task and pick-and-place of soda can.

Regarding BC-GMM algorithm, the success rates are significantly low for a small number of demonstrations; however, it improves with the increasing number of demonstrations as shown in Fig. 3a. Each algorithm, as illustrated in Fig. 3b, exhibited lower success rates as the task complexity increased from the simple lift task to the relatively complex pick-and-place of soda can. For example, the success rates of BC-GMM algorithm are almost halved for the pick-and-place task as compared to the lift task. In addition, the BCQ algorithm also exhibited poor task performance, less than 0.2 for all the given sets of demonstrations as depicted in Fig. 3b. In the case of HBC and IRIS algorithms, the success rates showed satisfactory outcomes for both manipulation tasks as shown in Fig. 3a and 3b.

A notable drop in task performance was observed, particularly in the case of the BCQ algorithm for the pick-and-place of soda can. One possible explanation is that the BCQ algorithm extrapolates to unseen states to find a solution, making it challenging to accomplish complex manipulation tasks. Similar issue has been addressed for grid world problems as well [29]. From the results, we noticed that the IRIS algorithm requires more data as the task complexity increases to yield satisfactory results. One possible reason is that the IRIS algorithm, due to its limitation [29], demands every possible state of the object in the training dataset. Therefore, for complex manipulation tasks, more data is needed to cover as many states as possible in the case of IRIS algorithm.

Overall, according to the criterion of achieving the maximum success rate, the minimum number of demonstrations is 20 per group for the lift task, as shown in Fig. 3a, while the minimum number is 30 per group for the relatively complex pick-and-place of soda can, as illustrated in Fig. 3b. Both algorithms, HBC and IRIS, exhibited satisfactory performance. Given the application of the IRIS algorithm in handling diverse and suboptimal demonstrations [29], we selected the IRIS algorithm for our analyses.

VI. Quality Features and Ranking of Demonstrations

In this section, we present the method followed for a series of experiments aiming at testing our hypotheses — high quality demonstrations will correspond to higher task performance (H1) and will attain an efficient task execution (H2). As a first step, we discuss here the ranking of demonstrations based on the quality features. For each demonstration, considering both manipulation tasks provided by proficient and multi-human, we computed the metrics based on manipulability and jerk features using (2) and (4), respectively. Finally, we ranked all demonstrations from high to lower value, creating groups of 20 and 30 consecutively for both the lift

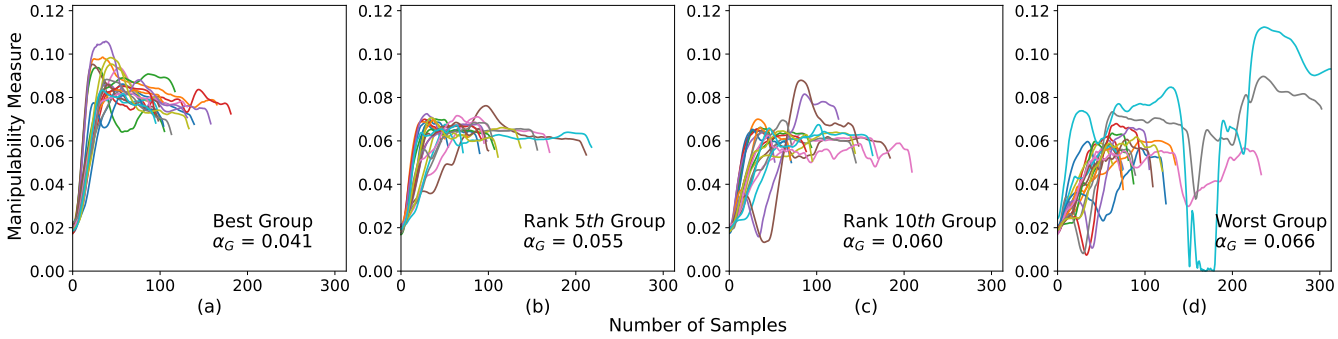


Fig. 4: Each sub-figure illustrates ranked demonstrations based on the manipulability feature, where α_G represents the average of manipulability based metrics (α) of the corresponding group. In Fig. 4d, the worst demonstration (cyan line) in the dataset falls to zero values for some interval, indicating a singular configuration.

TABLE I: Statistics of dataset based on quality features along with correlation coefficients and the corresponding p-values.

			Lift task		Pick-and-Place Soda Can	
			Proficient (P)	Multi-Human (MH)	Proficient (P)	Multi-Human (MH)
Stats	Manip	mean ± std	0.032±0.004	0.057±0.007	0.035±0.009	0.042±0.011
	Jerk	mean ± std	685±212	**1401±1256**	1940±513	**3031±1868**
H1	Manip	corr	0.17	−0.11	−0.28	0.11
		p-value	0.635	0.680	0.590	0.750
	Jerk	corr	0.38	**−0.60**	0.31	**−0.66**
		p-value	0.270	**0.019**	0.55	**0.036**
H2	Manip	corr	0.68	0.78	0.77	0.60
		p-value	9.59×10^{-24}	9.62×10^{-30}	2.35×10^{-15}	1.63×10^{-12}
	Jerk	corr	0.75	0.84	0.86	0.73
		p-value	2.06×10^{-32}	4.38×10^{-66}	1.68×10^{-29}	3.39×10^{-16}

and the pick-and-place of soda can tasks, respectively. The best group comprises top-ranked demonstrations while the worst group contains low-ranked demonstrations. To illustrate the quality diversity among the provided demonstrations, the mean and standard deviation based on the quality features are summarized in Tab. I.

A. *Manipulability*

Considering the multi-human dataset, the manipulability curve $\vec{\omega}(D)$ of each demonstration for the lift task can be represented as shown in Fig. 4. The best group, as illustrated in Fig. 4a, comprises demonstrations with high manipulability measures, with average group α_G value 0.041. The term α_G represents the average of manipulability based metrics (α) of the corresponding group. As the value of α_G increases, the corresponding demonstrations in the group gradually fall to the lower manipulability measures as shown in Fig. 4. The worst group consists of demonstrations with average group α_G value 0.066, including demonstration that contain zero values for some interval, representing a singular configuration, as shown in Fig. 4d.

B. *Joint-space Jerk*

For the pick-and-place of soda can using proficient dataset, the jerky motion of multiple demonstrations with varying magnitude are depicted in Fig. 5. The best demonstration showcases the

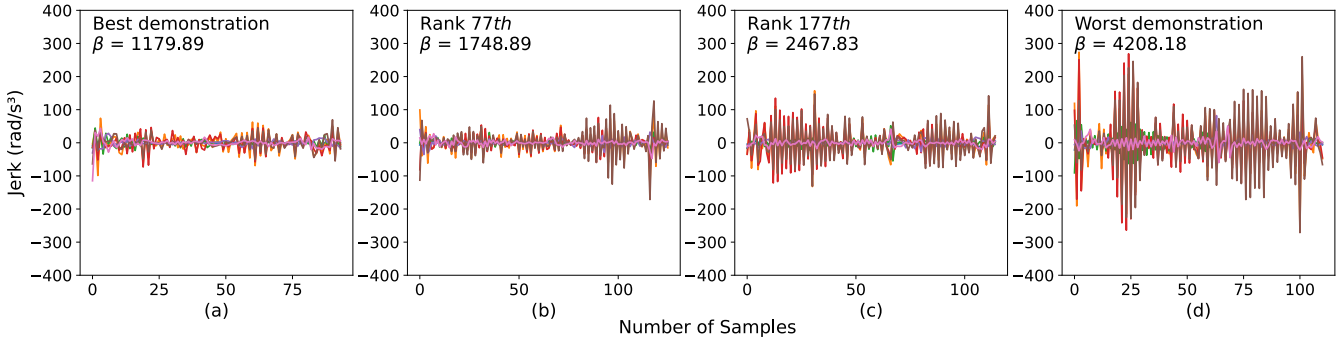


Fig. 5: Each sub-figure depicts a ranked demonstration based on the jerk feature. The worst demonstration exhibits jerky motion throughout the entire length of the demonstration, while the best demonstration indicates smoother motion profiles.

lowest jerk magnitude for all joints of the robotic system, as illustrated in Fig. 5a, with a β value 1179.89. Furthermore, the value of β increases with the jerk magnitude, where the worst demonstrations exhibiting high magnitude with β equals to 4208.18, as illustrated in Fig. 5d.

VII. Results

As per the results on the performance of various LfD algorithms on the lift and pick-and-place tasks (see Section V), we choose to use IRIS in the following experiments. We trained a series of IRIS-LfD models for both manipulation tasks and analyzed the correlation between input ranked groups and corresponding task performance (H1), as well as the quality of task execution (H2). The findings of each study are presented in detail in the following subsections.

A. Study 01 (H1)

To evaluate the first hypothesis, we computed the correlation coefficients between ranked groups and the corresponding success rates (i.e. task performance). As depicted in Fig. 6, there is no apparent impact of the manipulability feature on task performance in both manipulation tasks as well as both datasets (P and MH). Moreover, the p-values exceed 0.05 in all cases. Regarding the jerk feature, the correlation coefficients also show no effect on task performance in the case of the proficient (P) dataset as shown in Fig. 7a and 7c. However, we found strong correlations for both manipulation tasks, lift ($r = -0.60$, $p = 0.019$) and pick-and-place of soda can ($r = -0.66$, $p = 0.036$), in the dataset provided by multi-human (MH), as illustrated in Fig. 7b and 7d. Overall, the impact of demonstration's quality on task performance is significant when the demonstrations are of diverse-quality.

B. Study 02 (H2)

Our second hypothesis aimed at assessing if quality features were implicitly learned by our LfD algorithm. To calculate the outcomes of task execution's quality, we assessed each rollout using the same metrics as above, α and β. We then tried to see if there was any correlation between the quality ranking of each group of demonstrations used as input and the task execution quality produced after rollout. Furthermore, we used the actual values to provide a more realistic interpretation of the outcomes. The correlation analyses were performed for each feature across both types of datasets: proficient and multi-human.

Regarding the manipulability feature, there were strong correlations between input ranked groups and the corresponding rollouts as presented in Tab. I. For the jerk feature, the coefficients also showed strong correlations between input-ranked groups and the corresponding rollouts as shown in Tab. I. As compared to the manipulability feature, the correlation coefficients concerning the jerk feature are higher in each scenario. The overall correlation coefficients and the corresponding p-values are summarised in Tab. I.

VIII. Discussion

A. Role of Diversity (H1)

Using the two robotmimic v0.1 datasets (P and MH), we observed no distinct correlation between input ranked groups and the success rates; however, two cases showed strong correlation with respect to the jerk feature as shown in Fig. 7b and 7d.

Joint-Space Jerk: Regarding the above two cases, one of the possible reasons is that the multi-human dataset includes a diverse-quality of demonstrations. For instance, in the proficient-human dataset, the coefficient of variation (CV) for the lift and pick-and-place tasks is **30.94%** and **26.44%**, respectively. Conversely, in the multi-human dataset, the CV values for these manipulation tasks are substantially higher at **89.65%** and **61.62%**. Furthermore, the corresponding p-values are below 0.05, specifically **0.019** and **0.036** as shown in Tab. I, respectively.

Manipulability: The CV values for the lift and pick-and-place tasks using the proficient-human dataset are **12.5%** and **25.71%**, while in the multi-human dataset, the corresponding CV values are **12.8%** and **26.19%**, respectively. Because the datasets were collected using a simulator, the data exhibits a skewed distribution, particularly concerning the manipulability feature.

Overall, we hypothesize that H1 may hold if we have a diverse quality of demonstrations. This result somewhat aligns with [10], although their study considered a set of demonstrations rather than individual ones. For simple tasks, such as cube lifting, we expect minimal variations in manipulability on real robots, while noticeable joint-space jerk may occur due to the potential lack of smoothness from novice users. Due to dataset limitations, we plan to further investigate this hypothesis (H1) in our future work.

B. Learning Quality Features (H2)

Based on findings in study 02, the results support our second hypothesis (H2) as shown in Tab. I. For each quality feature, the coefficients show strong correlations between input-ranked groups and the corresponding rollouts. All of the correlation coefficients are higher than **0.60** in both manipulation tasks and types of datasets. While the p-value for each case is below 0.05 as shown in Tab. I. In the multi-human dataset, as depicted in Fig. 9b and 9d, outliers with input group (β_G) values around 4500 and 7500 slightly impact the results for both manipulation tasks. Additionally, the figures illustrate consistency among ranked groups, except for the last ones.

From the results, we posit that quality features can be transferred from given demonstrations to the corresponding LfD model. Alternatively, a set of demonstrations with high manipulability measure and low jerk values will yield an LfD model that executes tasks with nearly the same metrics as the given demonstrations.

IX. Conclusion

In this work, we investigated the role of quality of individual demonstration in LfD. We quantified each demonstration's quality based on the quality features and then evaluated the impact of demonstration's quality on both task performance and task execution's quality. This exploration involved the training of a set of LfD models using the robomimic v0.1 dataset. In the preliminary analysis, we selected the IRIS algorithm based on its performance, requiring 20 demonstrations for the lift task and 30 for the pick-and-place of soda can. In study 01, we computed the correlation coefficients between input-ranked groups and their corresponding success rates. The findings revealed a significant correlation in two instances, specifically when the demonstrations exhibited diverse qualities. Conversely, no evident correlations emerged in

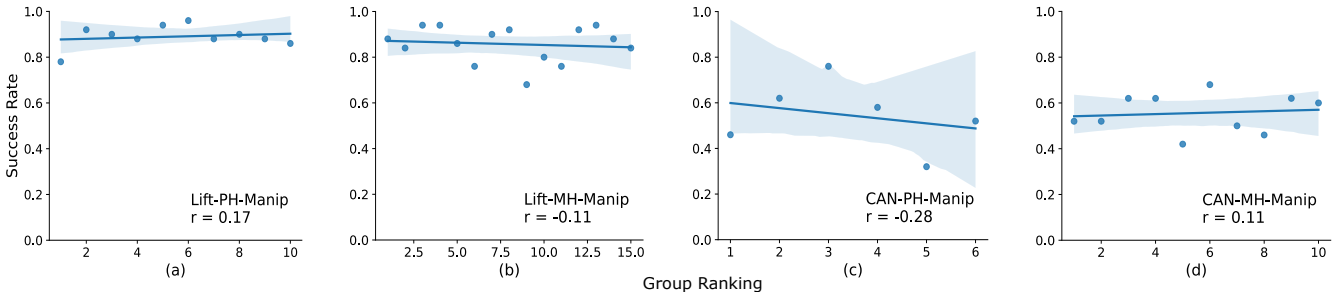


Fig. 6: Correlation between input ranked groups based on the manipulability feature and the corresponding success rates

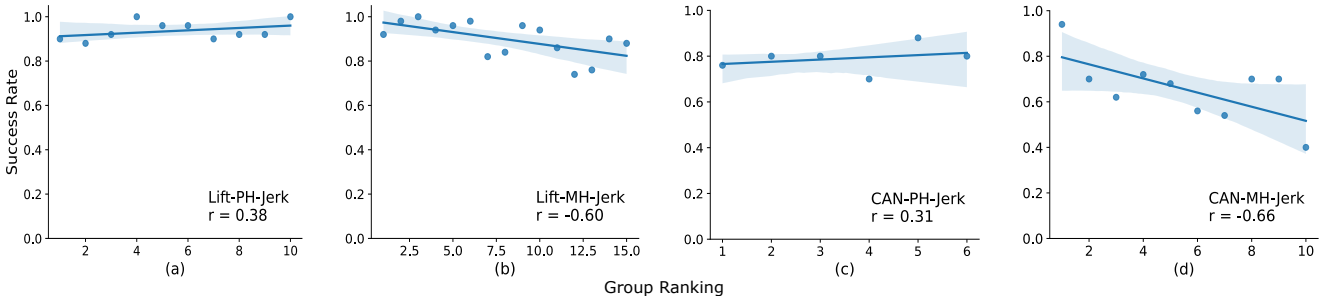


Fig. 7: Correlation between input ranked groups based on the jerk feature and the corresponding success rates

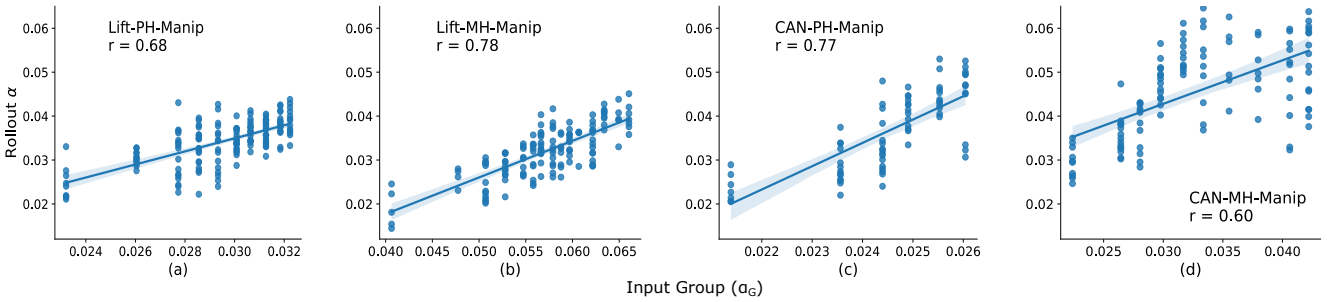


Fig. 8: Correlation between input ranked groups based on the manipulability feature and the corresponding rollouts.

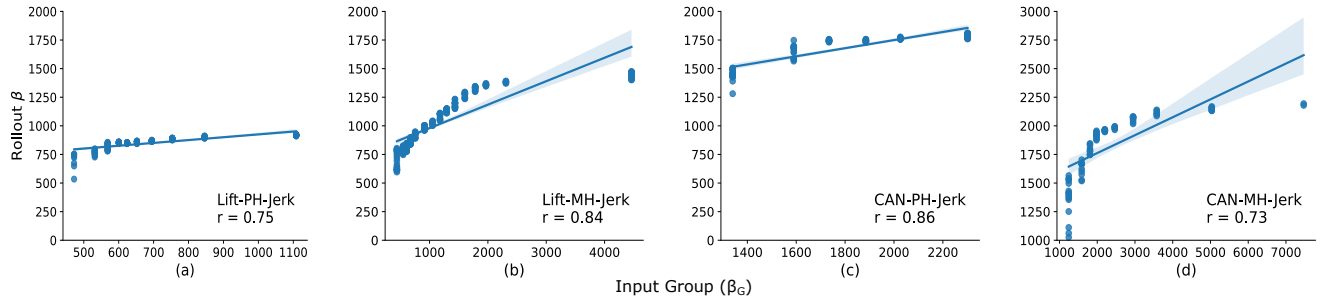


Fig. 9: Correlation between input ranked groups based on the jerk feature and the corresponding rollouts.

the cases where the demonstrations displayed less diverse quality. Regarding study 02, the results exhibited a strong correlation between the quality of input-ranked group of demonstrations and the corresponding rollouts for each quality feature.

This signify that not only the system can learn to perform the task (i.e. typically assessed through success rate) but also how to perform the task (through quality features). We hope that this work can lead to develop new ways to assess LfD models optimizing quality features rather than only looking at success rates. Furthermore, the proposed methodology for assessing the quality of demonstrations can be applied to a diverse range of tasks.

Although this study focused on two potential quality features

using the IRIS algorithm, we aim to assess additional relevant features and algorithms in future research. We plan to develop a feedback system to make capable novice users to provide demonstrations of desired quality. This study enables to identify the cut-off regions beyond which the quality of task execution becomes unsatisfactory as shown in Fig. 8 and 9. Eventually, a better execution function can be developed to guide novice users to provide effective demonstrations.

REFERENCES

[1] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robotics and Autonomous Systems*, vol. 57, pp. 469–483, 2009.

[2] H. Ravichandar, A. S. Polydoros, S. Chernova, and A. Billard, "Recent advances in robot learning from demonstration," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 3, pp. 297–330, 2020. [Online]. Available: https://www.annualreviews.org/doi/10.1146/annurev-control-100819-063206

[3] M. Laskey, C. Chuck, J. Lee, J. Mahler, S. Krishnan, K. Jamieson, A. Dragan, and K. Goldberg, "Comparing human-centric and robot-centric sampling for robot deep learning from demonstrations," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 358–365. [Online]. Available: http://ieeexplore.ieee.org/document/7989046/

[4] S. Ross, G. J. Gordon, and J. A. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," *Journal of Machine Learning Research*, vol. 15, pp. 627–635, 2011.

[5] S. Chernova and A. L. Thomaz, *Robot learning from human teachers*. Morgan & Claypool Publishers, 2014.

[6] T. Osa, J. Pajarinen, G. Neumann, J. A. Bagnell, P. Abbeel, J. Peters *et al.*, "An algorithmic perspective on imitation learning," *Foundations and Trends® in Robotics*, vol. 7, no. 1-2, pp. 1–179, 2018.

[7] N. Vahrenkamp and T. Asfour, "Representing the robot's workspace through constrained manipulability analysis," *Autonomous Robots*, vol. 38, pp. 17–30, 2015.

[8] J. Chen and A. Zelinsky, "Programing by demonstration: Coping with suboptimal teaching actions," *The International Journal of Robotics Research*, vol. 22, no. 5, pp. 299–319, 2003.

[9] M. Sakr, H. F. M. V. der Loos, D. Kulic, and E. Croft, "What makes a good demonstration for robot learning generalization?" in *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2021, pp. 607–609.

[10] M. Sakr, Z. J. Li, H. F. M. V. der Loos, D. Kulic, and E. A. Croft, "Quantifying demonstration quality for robot learning and generalization," *IEEE Robotics and Automation Letters*, vol. 7, pp. 9659–9666, 2022.

[11] P. Dhal and C. Azad, "A comprehensive survey on feature selection in the various fields of machine learning," *Applied Intelligence*, pp. 1–39, 2022.

[12] P. Duboue, *The art of feature engineering: essentials for machine learning*. Cambridge University Press, 2020.

[13] S. Amershi, M. Cakmak, W. B. Knox, and T. Kulesza, "Power to the people: The role of humans in interactive machine learning," *AI Magazine*, vol. 35, pp. 105–120, 2014. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1609/aimag.v35i4.2513

[14] S. Calinon, F. Guenter, and A. Billard, "On learning, representing, and generalizing a task in a humanoid robot," *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*, vol. 37, pp. 286–298, 2007. [Online]. Available: http://ieeexplore.ieee.org/document/4126276/

[15] A. L. P. Ureche and A. Billard, "Metrics for assessing human skill when demonstrating a bimanual task to a robot," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*. ACM, 2015, pp. 37–38. [Online]. Available: https://dl.acm.org/doi/10.1145/2701973.2702017

[16] K. Fischer, F. Kirstein, L. C. Jensen, N. Kruger, K. Kuklinski, M. V. aus der Wieschen, and T. R. Savarimuthu, "A comparison of types of robot control for programming by demonstration," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2016, pp. 213–220. [Online]. Available: http://ieeexplore.ieee.org/document/7451754/

[17] A. Sena and M. Howard, "Quantifying teaching behavior in robot learning from demonstration," *The International Journal of Robotics Research*, vol. 39, pp. 54–72, 2020. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0278364919884623

[18] M. Sakr, Z. Zhang, B. Li, H. Zhang, V. der Loos, H. Machiel, D. Kulic, and E. Croft, "How can everyday users efficiently teach robots by demonstrations?" *arXiv preprint arXiv:2310.13083*, 2023.

[19] Y. Zhu, J. Wong, A. Mandlekar, R. Martín-Martín, A. Joshi, S. Nasiriany, and Y. Zhu, "robosuite: A modular simulation framework and benchmark for robot learning," in *arXiv preprint arXiv:2009.12293*, 2020.

[20] N. Jaquier, L. Rozo, D. G. Caldwell, and S. Calinon, "Geometry-aware manipulability learning, tracking, and transfer," *The International Journal of Robotics Research*, vol. 40, pp. 624–650, 2021. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0278364920946815

[21] T. Yoshikawa, "Manipulability of robotic mechanisms," *The international journal of Robotics Research*, vol. 4, no. 2, pp. 3–9, 1985.

[22] A. Mandlekar, D. Xu, J. Wong, S. Nasiriany, C. Wang, R. Kulkarni, L. Fei-Fei, S. Savarese, Y. Zhu, and R. Martín-Martín, "What matters in learning from offline human demonstrations for robot manipulation," in *Conference on Robot Learning*. PMLR, 2022, pp. 1678–1690.

[23] A. Mandlekar, D. Xu, J. Wong, S. Nasiriany, C. Wang, R. Kulkarni, L. Fei-Fei, S. Savarese, Y. Zhu, and R. Martín-Martín, "Robomimic v0.1 dataset," https://robomimic.github.io/docs/datasets/robomimic_v0.1.html, accessed: January 2, 2024.

[24] A. Mandlekar, Y. Zhu, A. Garg, J. Booher, M. Spero, A. Tung, J. Gao, J. Emmons, A. Gupta, E. Orbay *et al.*, "Roboturk: A crowdsourcing platform for robotic skill learning through imitation," in *Conference on Robot Learning*. PMLR, 2018, pp. 879–893.

[25] S.-A. Chen, V. Tangkaratt, H.-T. Lin, and M. Sugiyama, "Active deep q-learning with demonstration," *Machine Learning*, vol. 109, no. 9, pp. 1699–1725, 2020.

[26] D. A. Pomerleau, "Alvinn: An autonomous land vehicle in a neural network," *Advances in neural information processing systems*, vol. 1, 1988.

[27] A. Mandlekar, D. Xu, R. Martín-Martín, S. Savarese, and L. Fei-Fei, "Learning to generalize across long-horizon tasks from human demonstrations," *arXiv preprint arXiv:2003.06085*, 2020.

[28] S. Fujimoto, D. Meger, and D. Precup, "Off-policy deep reinforcement learning without exploration," in *International conference on machine learning*. PMLR, 2019, pp. 2052–2062.

[29] A. Mandlekar, F. Ramos, B. Boots, S. Savarese, L. Fei-Fei, A. Garg, and D. Fox, "Iris: Implicit reinforcement without interaction at scale for learning control from offline robot manipulation data," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 4414–4420.