Consistency Matters: Defining Demonstration Data Quality Metrics in Robot Learning from Demonstration

MARAM SAKR, University of British Columbia, Canada
JUYAN ZHANG, Monash University, Australia
H.F. MACHIEL VAN DER LOOS, University of British Columbia, Canada
DANA KULIĆ, Monash University, Australia
ELIZABETH CROFT, University of Victoria, Canada

Learning from Demonstration (LfD) empowers robots to acquire new skills through human demonstrations, making it feasible for everyday users to teach robots. However, the success of learning and generalization heavily depends on the quality of these demonstrations. Consistency is often used to indicate quality in LfD, yet the factors that define this consistency remain underexplored. In this paper, we evaluate a comprehensive set of motion data characteristics to determine which consistency measures best predict learning performance. By ensuring demonstration consistency prior to training, we enhance models' predictive accuracy and generalization to novel scenarios. We validate our approach with two user studies involving participants with diverse levels of robotics expertise. In the first study (N = 24), users taught a PR2 robot to perform a button-pressing task in a constrained environment, while in the second study (N = 30), participants trained a UR5 robot on a pick-and-place task. Results show that demonstration consistency significantly impacts success rates in both learning and generalization, with 70% and 89% of task success rates in the two studies predicted using our consistency metrics. Moreover, our metrics estimate generalized performance success rates with 76% and 91% accuracy. These findings suggest that our proposed measures provide an intuitive, practical way to assess demonstration data quality before training, without requiring expert data or algorithm-specific modifications. Our approach offers a systematic way to evaluate demonstration quality, addressing a critical gap in LfD by formalizing consistency metrics that enhance the reliability of robot learning from human demonstrations.

CCS Concepts: • **Computing methodologies** → **Learning from demonstrations**; • **Human-centered computing** → *Empirical studies in interaction design*; • **Computer systems organization** → Robotics.

ACM Reference Format:
Maram Sakr, Juyan Zhang, H.F. Machiel Van der Loos, Dana Kulić, and Elizabeth Croft. 2024. Consistency Matters: Defining Demonstration Data Quality Metrics in Robot Learning from Demonstration. In *Proceedings of Transactions on Human-Robot Interaction (THRI).* ACM, New York, NY, USA, 31 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Learning from Demonstration (LfD), also known as imitation learning (IL), empowers robots to acquire new skills and behaviors by observing human demonstrations. This approach opens possibilities for everyday users, even those without robotics or programming expertise, to teach robots new tasks [5]. However, while LfD holds significant promise, most algorithms implicitly assume that demonstrators are experts providing near-perfect demonstrations [42]. This

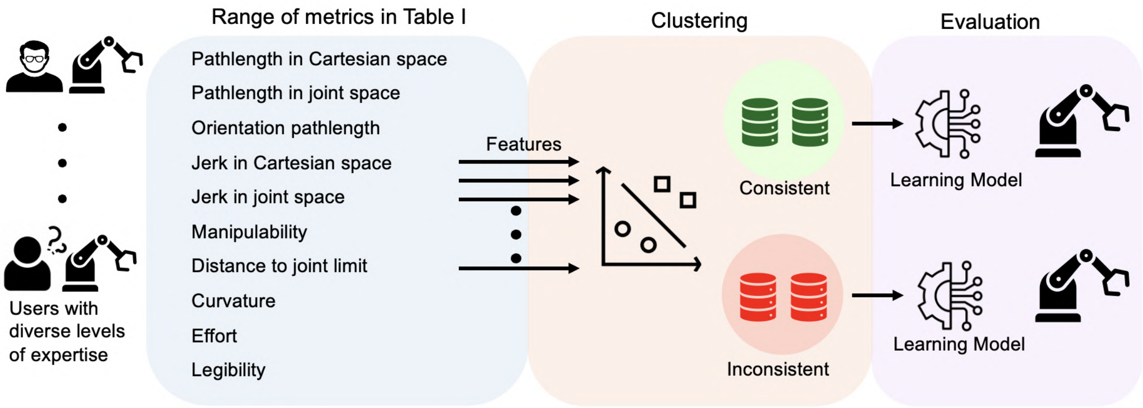


Fig. 1. Overview of the proposed approach. Data was collected from users with varying levels of expertise in robotics. The range of metrics listed in Table 1 was calculated to serve as feature inputs for the clustering algorithm. The data was then clustered into two groups: consistent and inconsistent. A learning model was subsequently trained using consistent and inconsistent data. Finally, performance was evaluated by calculating the success rate of the trained models.

idealized expectation often leads to a gap between real-world data quality and algorithmic assumptions, as practical demonstrations frequently include variability due to factors such as fatigue, lack of experience, or environmental constraints [4, 13].

Prior work has sought ways to address this gap by refining algorithms to cope with imperfect data [51, 57]. Some studies use heuristic quality indicators, such as undesired motions [38], varying lengths and amplitudes [58], distribution shift [7], or ambiguous demonstrations [49]. However, these notions of quality are often vague and incomplete, lacking well-defined metrics to identify what makes a demonstration "better" or to systematically detect imperfections. This lack of systematic and coherent metrics creates challenges for researchers and practitioners alike, as inconsistent or suboptimal demonstrations can degrade learning performance and limit generalization [5, 38].

Data quality can be assessed across various dimensions, each representing a measurable property of the data that influences its overall usefulness. This aligns with the broader definition of data quality in the ISO/IEC 25012 standard [27], which is defined as *"the degree to which a set of characteristics of the data meets the requirements"*. Examples of such characteristics in the machine learning literature include accuracy, completeness, consistency, and timeliness [17, 31]. Nevertheless, these general characteristics need to be adapted with domain-specific metrics and methods for measurement, particularly for LfD tasks. Questions arise, such as: *How is accuracy defined in demonstrations? What constitutes consistency of the demonstrations?*

Our approach, as shown in Fig. 1, directly addresses this need by introducing a comprehensive set of metrics for defining and quantifying consistency in demonstration data, across both Cartesian and joint-space characteristics. By rigorously quantifying data quality in this way, we can ensure that only high-quality demonstrations feed into learning models, enhancing the predictive accuracy of both learning and generalization performance. This approach is designed to be applicable across diverse LfD scenarios, making it a practical tool for both researchers and practitioners. By validating demonstration consistency as a key data quality component, we propose a new approach for evaluating and curating data before model training, bridging a significant gap in LfD research and offering a scalable solution for improving robot learning outcomes.

2 RELATED WORK

Obtaining optimal demonstrations for all tasks that a robot must learn is a significant challenge [42]. Identifying experts proficient in both the task domain and in robotics to ensure high-quality demonstrations tailored to task goals and the robot's constraints is difficult [38]. Consequently, mixed-quality demonstrations are commonly used in learning. Knowing the quality of demonstrations beforehand allows lower-quality data to be filtered out, while multimodal learning approaches can be employed to focus on the highest quality demonstrations [24, 50, 55].

2.1 Learning from Mixed-Quality Demonstrations

Examining the effect of diverse-quality demonstrations, Mandlekar et al. [35] found that IL models trained exclusively on high-quality demonstrations from a single human achieved higher success rates than models trained on larger datasets from multiple demonstrators with varied levels of proficiency. Similarly, in prior work [46], we showed that models trained on mixed-quality data perform worse than those trained on a smaller dataset of consistently high-quality demonstrations, underscoring the importance of ensuring high-quality data from each demonstrator before aggregation. Zhang et al. [61] proposed a confidence-aware IL framework that automatically estimates the optimality of demonstrations and dynamically weights them during policy training, yielding improved performance when suboptimal examples are present. Our work builds on these foundations by providing a lightweight, pre-training assessment of demonstration consistency across task-irrelevant metrics, enabling the selection (or weighting) of high-quality data before any policy learning takes place and offering a complementary approach to confidence-based methods for handling mixed-quality datasets.

2.2 Metrics for Evaluating Demonstration Quality

2.2.1 Task-Independent Quality Metrics. The literature presents multiple approaches to defining and assessing demonstration quality, although few papers propose standardized or comprehensive metrics. Chernova and Thomaz [15] suggested evaluating human input quality based on factors such as task skill, time taken, and consistency; however, they did not provide specific metrics to measure these aspects. Kaiser et al. [30] identified three primary sources of sub-optimality in human demonstrations: unnecessary actions that do not aid the goal, incorrect actions that reduce demonstration utility, and unmotivated actions influenced by human sensory information that is unavailable to the robot. While they proposed methods to measure unnecessary actions via motion ratios and discontinuity checks, identifying incorrect actions remains challenging, and distinguishing motivated from unmotivated actions often depends on the demonstration method, such as teleoperation or kinesthetic teaching [9].

Other studies focused on teaching efficacy and error detection within demonstrations. Fischer et al. [20] identified common user errors, such as excessive pressure on the gripper, proximity to singularities, and self-collisions, although they did not evaluate these errors' impact on robot learning and generalization. Sena and Howard [49] addressed data sparsity by defining two quality metrics: teaching efficacy, which measures task-space generalization, and teaching efficiency, which normalizes efficacy by the number of demonstrations. These metrics focus on optimizing demonstration quantity rather than assessing individual demonstration quality, which remains critical for ensuring generalizable and reliable learning.

Recent research has emphasized the importance of motion metrics in assessing demonstration quality and motion performance. Bilal et al. [8] highlighted the impact of manipulability and joint-space jerk on task performance, especially with varying demonstration quality. In the context of delicate tasks, such as surgical manipulation, Aghazadeh et al. [1]

demonstrated that motion smoothness metrics, particularly those related to jerk and trajectory length, are strong indicators of surgical expertise, underscoring the importance of selecting appropriate kinematic features when evaluating performance in precision tasks. Additionally, Meixner et al. [36] developed a unified human-likeness metric combining path length and joint-space smoothness to assess motion retargeting. Jaquier et al. [29] focused on manipulability, enabling robots to adapt configurations for enhanced dexterity. Together, these studies highlight the critical role of motion metrics in improving the reliability and interpretability of robot learning from demonstration.

2.2.2 Task-Dependent Quality Metrics. Researchers have developed task-specific metrics for assessing demonstration consistency. Ureche and Billard [39] proposed metrics for bimanual tasks, such as evaluating tool maneuverability, consistency across trials, and coordination between arms. They measured consistency by tracking constraint changes across each dimension, although performance tends to vary over time, leading to fluctuations in quality. This limitation points to the need for trial-independent evaluations to capture high-quality demonstrations more effectively. Additionally, our work proposes an extensive set of motion features to determine which measures best predict learning and generalization performance.

Despite the recognized need for quality assessments, there is a notable lack of metrics and methods that comprehensively evaluate multiple dimensions of demonstration quality. This paper addresses that gap by introducing task-agnostic metrics to assess consistency in demonstration data, focusing on motion characteristics in both Cartesian and joint-space dimensions. To validate these metrics, we conducted two user studies, demonstrating that our consistency metrics reliably predict both task success and generalization performance. The results highlight the robustness of the metrics in different tasks and robotic platforms. By applying these measures to evaluate demonstrations before learning, we can ensure the inclusion of only high-quality data, ultimately enhancing the efficiency and effectiveness of the Learning from Demonstration (LfD) process.

3 PROPOSED APPROACH

We begin by proposing a broad list of metrics to evaluate the consistency of demonstration data. These metrics evaluate various aspects of robot motion and investigate their influence on learning and generalization performance. These metrics consider both Cartesian and joint spaces for the robot. Table 1 shows the proposed metrics for evaluating the consistency of the demonstration data. Path length metrics are selected to detect unnecessary motions [38] that do not contribute to the task's goal. For instance, if a user struggles to manoeuvre a robot toward the goal, the provided trajectories will be longer than those of a proficient user who manoeuvres the robot directly toward the goal. Additionally, trajectory smoothness improves motion stability and smoothness in the learned and generalized trajectories [14]. Fluent trajectories that avoid abrupt changes are also important for efficient navigation, reducing wear and tear on robotic components, and minimizing energy consumption [44]. Hence, we propose jerk, curvature, and effort as potential metrics to quantify the quality and consistency of the demonstrations.

The robot's posture directly impacts reaching movements and manipulation tasks (e.g., pushing, pulling, reaching) [29]. In this context, manipulability [59] serves as a kinematic descriptor indicating the ability to arbitrarily move in different directions of the task in a given joint configuration. Since the ability to manoeuvre the robot in the workspace degrades at singular configurations, this quality metric is important for allowing the robot to generalize the learned task across the workspace [28, 52]. In addition to singularities, joint limits significantly affect the manoeuvrability of the robot in the workspace [21].

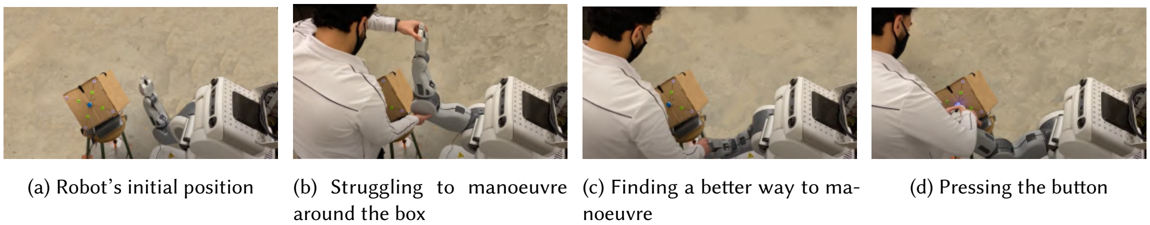
Table 1. Proposed quality metrics. The variables denoted here are: x for end-effector position, q for joint angle, R for rotation matrix, $\ddot{x}$ for end-effector acceleration, $\dddot{x}$ for end-effector jerk, $\dddot{q}$ for joint jerk, τ for torque, J for the Jacobian, n_d for the total number of degrees of freedom (DOFs), q_d^- for the lower limit of the d^{th} joint, q_d^+ for the upper limit of the d^{th} joint. $P(G_i|\xi_{1:t})$ is the posterior probability of goal G_i given the trajectory segment $\xi_{1:t}$.

Criterion	**Reference name**	**Formula**
Path Length in Cartesian Space	Q_x	$\sum_{t=1}^{T} \lVert x_t - x_{t-1} \rVert^2$
Path Orientation Length	Q_{rot}	$\sum_{t=1}^{T} \arccos\left(\frac{\mathrm{trace}(R_t \cdot R_{t-1}^T)}{2}\right)^2$
Path Length in Joint Space	Q_q	$\sum_{t=1}^{T} \lVert q_t - q_{t-1} \rVert^2$
Jerk in Cartesian Space	$Q_{\dddot{x}}$	$\sum_{t=1}^{T} \dddot{x}_t^2$
Jerk in Joint Space	$Q_{\dddot{q}}$	$\sum_{t=1}^{T} \dddot{q}_t^2$
Manipulability	Q_M	$\sum_{t=1}^{T} \sqrt{\det(J_t J_t^T)}$
Distance to Joint Limit	$Q_{q_{lim}}$	$\sum_{t=1}^{T} \prod_{d}^{n_d} \frac{4(q_d - q_d^-)(q_d^+ - q_d)}{(q_d^+ - q_d^-)^2}$
Joint Effort	Q_τ	$\sum_{t=1}^{T} \sum_{d}^{n_d} \tau_{d,t}^2$
Cartesian Curvature	Q_κ	$\sum_{t=1}^{T} \frac{\lVert \dot{x}(t) \times \ddot{x}(t) \rVert}{\lVert \dot{x}(t) \rVert^3}$
Legibility	Q_L	$-\sum_{i=1}^{n} P(G_i|\xi_{1:t}) \log P(G_i|\xi_{1:t})$

If the robot operates in the same space as humans, it is important to consider legibility as a quality metric [19]. Legibility measures how easily an observer can infer a robot's goal from its motion. It is defined based on the probability assigned to the actual goal across the trajectory. This probability is computed using the cost function for each potential goal, which combines the early differentiation of the trajectory and the progress towards the goal [54]. This cost function is not a ground-truth task reward but a heuristic approximation of goal predictability, and its use does not assume that the task can be solved analytically. Trajectories with high legibility will allow an observer to infer the goal early, given a small segment of the trajectory, and all points in the trajectory will be directed toward that goal. Details on the legibility calculation are in Appendix A.1.

It is worth noting that task goals are the primary objectives of the demonstrator, with quality being a secondary concern. For instance, when manoeuvring a robot in a constrained space, avoiding collisions takes precedence over the robot's manipulability or proximity to joint limits. Thus, it is essential to consider task and workspace constraints when applying these metrics. To reduce the dependency on task-specific constraints, we focus on the *consistency* of demonstrations in terms of these metrics rather than their absolute values.

Consistency provides a way to account for variations in task difficulty and environmental constraints. Demonstrations that are consistent across motion metrics often indicate a proficient demonstrator — one who can perform the task repeatedly, efficiently, and effectively [39]. This consistency offers a more reliable signal for learning than absolute performance metrics, which can be heavily influenced by outliers such as a single exceptionally good or poor demonstration [6].


(a) Robot's initial position

(b) Struggling to manoeuvre around the box

(c) Finding a better way to manoeuvre

(d) Pressing the button

Fig. 2. Button-pressing task demonstration overview from (a) the initial position of the robot, (b) an example of a user's struggle to get the robot around the box, (c) an example of a better manoeuvre of the robot around the box, and (d) the robot pressing the button.

Unlike success labels that rely on explicit checks for task completion, consistency derived from motion characteristics can be computed in real-time and reflects how stable and refined a user's demonstration strategy is. Prior research has shown that unsuccessful demonstrations are frequently accompanied by high variability, particularly during early attempts when users are still exploring how best to complete the task [3, 56]. By quantifying consistency in this way, researchers can prioritize high-quality demonstrations for training, thereby improving both learning outcomes and generalization performance.

4 EXPERIMENTAL DESIGN

To validate that the consistency of the proposed metrics is predictive of demonstration quality and improves learning performance (as in Fig. 1), we collected demonstrations from users with diverse robotics experiences using kinesthetic teaching. The demonstrations were then clustered into consistent and inconsistent groups based on the proposed metrics. Finally, we evaluated the learning and generalization performance of both groups to examine the correlation between the metrics and the robot's performance. Kinesthetic teaching allows users to physically guide the robot to achieve a task, with the robot's state recorded via its onboard sensors (e.g., joint angles, torques) [5]. Recording demonstrations directly on the robot using its integrated sensors eliminates the correspondence problem, which arises from the mismatch between the human teacher and the robot learner due to differences in sensing ability, body structure, and mechanics [15]. We experimentally validated the proposed approach across two different tasks and robots to demonstrate the generalizability of the proposed metrics.

4.1 Experiment 1: Button Pressing

4.1.1 Task Definition. Our first exemplar task is pressing a button, a generic task that models real-world tasks such as pressing a doorbell, an elevator call button, or a pedestrian crossing button. This task was selected because it does not require domain expertise but does require practice with the robot to provide high-quality demonstrations. The task involves both a constrained reaching task and a fine control motion for pressing the button.

The robot platform used in this work is the PR2 (Willow Garage, Personal Robot 2), a mobile manipulator with two 7-DoF arms operating under ROS Melodic. The passive spring counterbalance system in PR2's arms provides gravity compensation, giving users the ability to kinesthetically move the robot's arms within their kinematic range. Each arm has a 1-DoF under-actuated gripper. In this experiment, we only used the right arm in gravity compensation mode with the gripper closed.

Fig. 2 shows the experimental setup used for data collection. A cardboard box was fixed on one of its vertices so that all buttons were reachable by the robot gripper. Only one face of the box was used in the data collection. Buttons were placed in the center (large blue button), corners (purple foam markers), and midway between the corners and the center of each face (green foam markers). A total of nine-goal positions were used. This setup represents a reaching task in a constrained space, requiring the participant to manoeuvre the robot arm around the box to reach the goal positions while avoiding self-collisions and collisions with the box.

4.1.2 User Study. We recruited participants through advertisements on the university campus and social media. A total of 24 participants (18 male, 6 female) with an average age of 22 years and varying levels of robotics expertise (ranging from no experience to six years or more) participated in the study. Approximately 73% of the participants self-reported having no prior experience with robotics. The University Behavioral Research Ethics Board approved the research (application ID H20-03740), and informed consent was obtained from each participant before the experiment began.

The experiment was conducted in two sessions on different days, acknowledging the importance of practice in motor skill learning [48]. The two sessions were scheduled 1–3 days apart, depending on participant availability, to allow for overnight skill consolidation while minimizing external variability between sessions. We aimed to explore which consistency metrics improved from the first to the second session and how this would be reflected in robot learning and generalization. The experimenter briefly explained the long-term objective of the research and instructed the participants that their task was to program new skills into the PR2 robot. Participants were told to imagine a scenario in which they brought a robot to their home and wanted to teach it a task. The robot would imitate and learn from their demonstrations, so they should provide as natural demonstrations as possible.

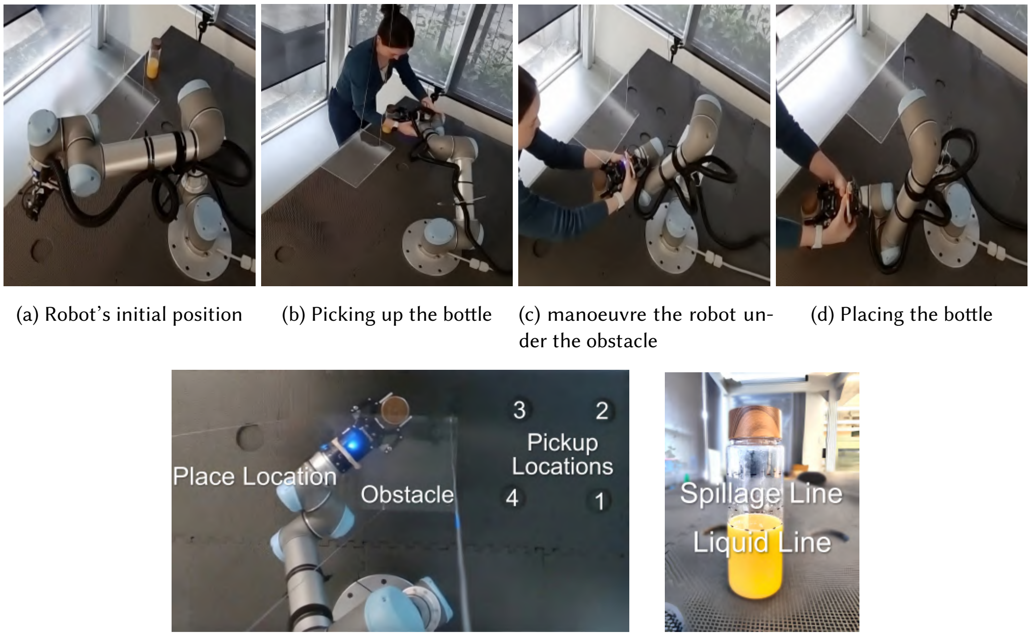
The robot was set in gravity compensation mode, and participants were asked to hold the robot's right arm and physically guide it to press a target button (kinesthetic teaching). The right arm always started in the same position, with the elbow at 90° and the gripper pointed up (untucked position), as in Fig. 2a. After each demonstration, the experimenter teleoperated the right arm via a joystick to return it to the initial position. The participants demonstrated the task for three trials in each session. The robot's joint angles were recorded during each demonstration and saved as ROSbag files for offline analysis.

Each participant provided 27 demonstrations per session[1]. After the first session, the participants reflected on their learning and practice and scheduled the second session on a different day. This procedure was motivated by Walker et al. [53], who found that a night of sleep after motor skill training significantly improves skill levels in subsequent repetitions. In the second session, participants repeated the same procedure as in the first session.

4.2 Experiment 2: Pick and Place

4.2.1 Task Definition. Our second exemplar experimental task represents picking and placing a liquid-filled container while avoiding collisions and spillage. This task does not require any domain expertise, only some practice to provide smooth demonstrations from which the robot can learn. Fig. 3 shows the experimental setup with a Universal Robots (UR5) equipped with a Robotiq 2F-85 two-finger gripper as its end effector and operating under ROS Melodic. The task is to pick up the bottle from any of the four positions in the pickup area and place it in the designated "place" location, as shown in Fig. 3e, without "spilling". An obstacle is positioned above the table, midway between the pick and place locations. In the first user study, participants provided nine demonstrations per trial across nine different goal locations to broadly sample the workspace. Based on exploratory analyses and pilot testing, we found that the four corner locations were most critical for task performance. Thus, in the second user study, participants provided four

[1] Each trial involved the participant demonstrating the task across 9 goal positions, resulting in 27 demonstrations per session (3 trials × 9 goals)



(a) Robot's initial position (b) Picking up the bottle (c) manoeuvre the robot under the obstacle (d) Placing the bottle

(e) Top view showing the pickup and place locations (f) The experimental bottle with spillage line

Fig. 3. Task demonstration overview: (a) Robot's initial position, (b) Picking up the bottle, (c) manoeuvring under the obstacle, (d) Placing the bottle, (e) Top view of pickup, obstacle and place locations, and (f) Experimental bottle with spillage line.

demonstrations per stage corresponding to these key corner positions, maintaining spatial coverage while streamlining the data collection process.

To simplify spillage detection, the robot picks up a sealed bottle with landmarks (dotted lines above the liquid level) to define the "spillage" threshold, as shown in Fig. 3f. A 2 cm margin was set between the top surface of the liquid at rest and the spillage landmark. If the user significantly changes the end effector orientation, provides jerky trajectories, or moves the robot too quickly, the liquid may exceed the spillage mark.

4.2.2 User Study. Participants were recruited through social media and word of mouth. A total of 30 participants (24 male, 5 female, 1 non-binary) with an average age of 27 years took part in the study. The participants had varying expertise in robotics, with 33.3% never having prior physical interaction with robots. The study was approved by the Monash University Human Ethics Committee (Project ID: 37303), and informed consent was obtained from each participant before the experiment began.

Participants were introduced to the experimental setup and the task. The experimenter demonstrated physically guiding the robot to perform the task and emphasized that the robot would imitate and learn from the provided demonstrations. Participants were instructed to provide smooth and natural demonstrations for the robot to learn from. They were also instructed to avoid collisions with the obstacle, and in case of a collision, they were told to continue and attempt to avoid any further collisions. Additionally, participants were asked to keep the liquid as level as possible. If the liquid touched the landmark line, it would be counted as a spill. Once they gripped the bottle, they were not permitted to release it until reaching the placement location.

The participants started with the bottle in the first location (shown in Fig. 3f), and the robot was set in gravity compensation mode, allowing them to physically guide the robot to pick up the bottle and place it in the designated location. To control the gripper, we used a Wizard-of-Oz technique: participants gave voice commands such as “open“ or “close“, and the experimenter executed these commands on the computer to control the gripper. Once the bottle was placed, the data was saved as ROSbag files for offline analysis. The robot was then reset to the initial position, as in Fig. 3a, and the bottle was moved to the second location. This process was repeated until participants provided four demonstrations from each pickup location. The order of the pickup locations (as shown in Fig. 3e) was designed to progress from easy to hard; the closer the pickup location is to the robot, the more challenging it becomes to position the robot to pick up the bottle. This phase was referred to as the baseline performance, as it captured participants' demonstrations during their first interaction with the task.

After this baseline phase, all participants underwent a practice period using one of the pickup locations (location 4), selected because it was the most challenging position. Based on findings from prior work [46], where training on simpler tasks yielded relatively small improvements due to high baseline performance, we chose to focus practice on the most difficult location to maximize skill acquisition and promote better generalization to the easier pickup locations. Each participant was assigned to one of three training groups as in [45], with equal practice time allotted for each group. They were encouraged to experiment with different strategies to develop a smooth approach to teaching the robot the task. Following the practice session, a retention test identical to the baseline was conducted to capture participants' performance after practice. For this test, participants provided four demonstrations from each location, picking up the bottle and placing it in the designated spot, to assess their performance post-practice.

4.3 Task Learning

Several factors contribute to the success rate of a learning policy, such as the number of provided demonstrations, their distribution over the task space, the learning algorithm, and demonstration quality. Here, we focus on the quality of the provided demonstrations while keeping all other factors fixed (i.e., the number of demonstrations, the distribution, and the learning model). For the learning algorithm, we used a Task Parameterized Gaussian Mixture Model (TP-GMM) combined with Gaussian Mixture Regression (GMR) for task learning. TP-GMMs have been extensively used in the LfD literature [38, 40], providing good generalization with a limited set of demonstrations. In this approach, demonstrations are encoded as sequences of end-effector positions and orientations relative to task-parameterized frames. Joint angles are also recorded to assist with kinematic redundancy resolution. During generalization, the task frames are updated for new goal configurations, and the learned trajectory distributions are transformed accordingly, without the need for handcrafted observation engineering.

TP-GMM models a task using task parameters defined by a sequence of coordinate frames, each represented by a matrix for orientation and a vector for origin relative to a global frame [12]. A Gaussian Mixture Model (GMM) is fitted to the data in each local frame, with the parameters estimated using an expectation-maximization algorithm. To generate a trajectory, the local models are transformed back into the global frame and combined into a global model through a product of Gaussians. The resulting trajectory is then used to derive a feasible joint space trajectory. To ensure smooth and accurate task execution, both the singularity-robust inverse [37] and a closed-loop inverse kinematics (CLIK) algorithm [18] are employed. Given the redundancy in a 7-DOF manipulator, the null space is utilized to control joint space motion without affecting task-space motion, using the closest human demonstration as a policy [33]. This comprehensive approach allows for accurate task reproduction while avoiding kinematic singularities and minimizing errors.

4.4 Evaluation of Learning and Generalization Performance

4.4.1 Experiment 1: Button-pressing task evaluation. Model performance was evaluated on the same set of demonstrated tasks (task performance), as in Fig. 4a, and on new target positions (generalized performance), as in Fig. 4b. Generalization was evaluated by discretizing the face of the box in Fig. 2a into a 7x7 grid, resulting in a total of 49 new target positions. The grid size was determined based on the dimensions of the PR2 arm's tip and the box to avoid overlap between targets. The arm tip's dimensions are ($W = 2.1$ cm, $L = 2.2$ cm, $H = 3.5$ cm), and the box is a cube with edges of 26 cm. We specified the target point as a sphere with a 3 cm diameter centered on the button.

The learned trajectory was considered successful if it reached the goal position (within the goal sphere) while avoiding self-collisions and collisions with the box. To account for the non-zero size of the end-effector tip, we considered the robot to have reached the goal if any point of the tip touched the goal sphere. The success rate for each trial was calculated by dividing the number of successful trajectories by the total number of tested points (nine in the task performance test and 49 in the generalization test).

4.4.2 Experiment 2: Pick-and-place task evaluation. Model performance was evaluated by generating trajectories to pick and place the bottle from the same demonstrated locations (task performance), as in Fig. 4d, and from new pickup locations (generalized performance), as in Fig. 4e. The generalized pickup locations was created by discretizing the pickup space into a 4x4 grid, resulting in 16 new pickup locations. The grid size was chosen based on the UR5 robot's reachability space and the diameter of the experimental bottle (7 cm).

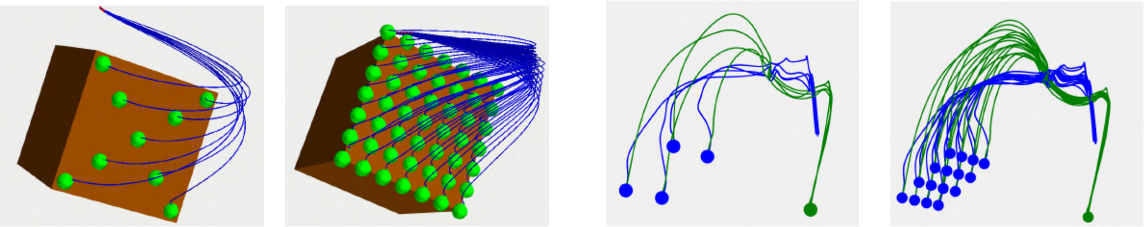
The learned trajectory was considered successful if the bottle was picked up and placed without “spillage“. Spillage was detected by monitoring deviations in upright orientation, acceleration, and jerk, which serve as proxies for potential liquid movement. These proxies were validated using the video recordings of the demonstrations before being applied to evaluate the learned trajectories. To assess spillage risk, we normalized the orientation deviation, jerk, and acceleration values to assess spillage risk. As the jerk, acceleration, and orientation deviation values approach zero, the success score approaches 1, provided the bottle was picked up and placed correctly. However, as jerk, acceleration, and/or orientation deviation increase, the success score is penalized accordingly. The Overall Success was computed using a weighted sum of the pick-and-place success and the normalized spillage proxies, as shown in Equation 1.

$$\text{Overall Success} = \begin{cases} 0, & \text{if the bottle was not picked up} \\ 0.5 \times (S_p + S_l) + 0.5 \times \frac{(1-J_n)+(1-A_n)+(1-O_n)}{3}, & \text{if the bottle was picked up successfully} \end{cases} \tag{1}$$

where S_p denotes the pickup success, with $S_p = 0.5$ if the bottle was successfully picked up and $S_p = 0$ otherwise. S_l represents the place success, with $S_l = 0.5$ if the bottle was placed correctly and $S_l = 0$ otherwise. The values J_n, A_n, and O_n correspond to the normalized metrics for jerk, acceleration, and orientation deviation, respectively,

The Phase Success score was then calculated by averaging the overall success rates across all tested locations, as defined in Equation 2. In the task performance test, this phase success was calculated across four locations, while in the generalization test, it was calculated across 16 locations.

$$\text{Phase\_Success} = \frac{\sum_{i=1}^{N} \text{Overall\_Success}_i}{N} \tag{2}$$



(a) Same task states (b) Generalized states (d) Same task states (e) Generalized states

(I) button pressing task with PR2 (II) pick and place task with UR5

Fig. 4. The tasks simulation includes both the same task states and the generalized states used to evaluate model performance. (I) In the button-pressing task, the buttons are represented as green spheres. (II) In the pick-and-place task, the pickup and place locations are also represented by spheres. The blue segment of the trajectory represents the movement from the initial position to the pickup location, while the green segment represents the movement from the pickup to the place location.

where Phase_Success represents the average success score across all tested locations within the phase. Overall_Success$_i$is the individual success score for each location i, calculated as in Equation 1. N denotes the total number of tested locations.

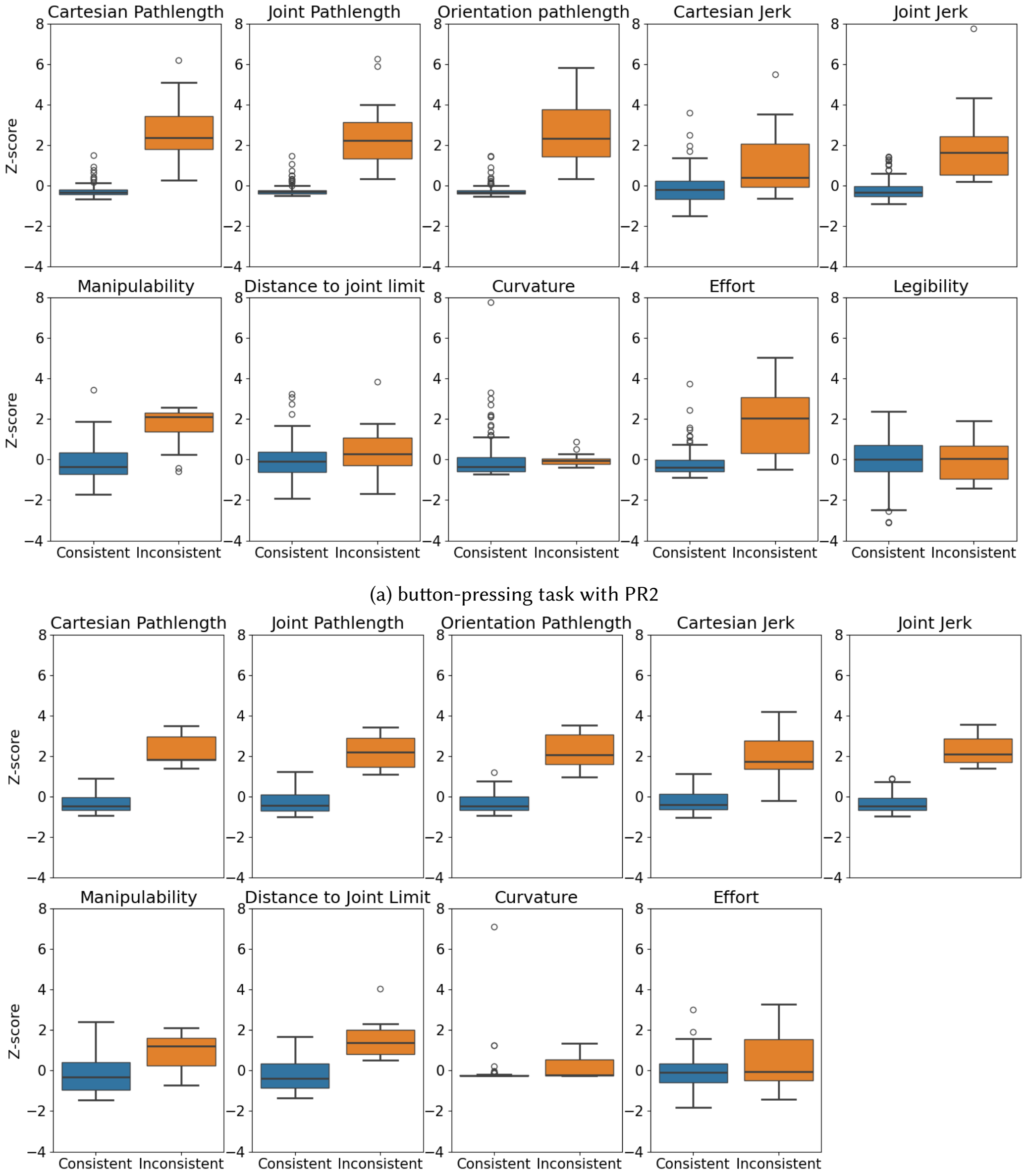
5 HYPOTHESES

We expect that the demonstrations' consistency in terms of the proposed measures is a key identifier of the quality, and consequently, of task performance and generalized performance. Moreover, we believe that practice is crucial for improving the quality of demonstrations, which in turn enhances task performance and generalized performance. Based on these expectations, we formulate the following hypotheses:

- **H1-a**: The consistency of demonstrations in terms of the proposed metrics significantly contributes to the task performance success rate.
- **H1-b**: The consistency of demonstrations in terms of the proposed metrics significantly contributes to the generalized performance success rate.
- **H2**: Participants' practice significantly improves the consistency of demonstrations.
- **H3-a**: Participants' practice significantly improves the task performance success rate.
- **H3-b**: Participants' practice significantly improves the generalized performance success rate.

6 RESULTS

To evaluate our hypotheses, we assessed the consistency of the demonstrations provided by each participant. We employed a K-means clustering algorithm [25] to categorize each set of demonstrations into one of two clusters: a *consistent* cluster and an *inconsistent* cluster, as shown in Fig. 1. For each trial in the button-pressing experiment, the range (maximum value minus minimum value) of the ten metrics listed in Table 1 was calculated and used as input for the clustering algorithm (multi-dimensional clustering). The rationale for using the range is that consistent demonstrations are expected to exhibit smaller variations in metric values, while inconsistent demonstrations are expected to show larger variations. Prior to clustering, all metric range values were standardized using z-score normalization across participants. This preprocessing step ensured that all features were unit-invariant and contributed equally to the clustering process, regardless of their original scale (e.g., centimetres, radians, torque units). Since ranges are positive



(a) button-pressing task with PR2

(b) pick-and-place task with UR5

Fig. 5. The standardized feature values in the resulting two clusters of demonstrations. The first cluster (blue) includes consistent demonstrations and the second cluster (orange) includes inconsistent demonstrations. Legibility is not reported for the pick-and-place experiment because the task involved a single known goal, making goal discrimination-based legibility undefined.

measures of variability, the orientation (positive or negative association with success) of each metric does not impact clustering outcomes. K-means clustering was then performed using squared Euclidean distance, which is the standard distance metric in K-means.

Fig. 5a shows the standardized metric values for both the consistent and inconsistent clusters in the button-pressing task. The inconsistent cluster displayed a larger range across most metrics, while the consistent cluster exhibited a smaller range. Specifically, the inconsistent cluster had a broader distribution in all pathlength, jerk, and effort metrics. In contrast, the distribution for distance to joint limits, curvature, and legibility was similar between the two clusters. Additionally, the inconsistent cluster showed higher median values for pathlength, jerk, manipulability, and effort metrics, indicating that demonstrations in the inconsistent trials involved more variability, longer paths, greater jerk, more singular poses, and higher energy expenditure. Conversely, the consistent trials, with lower median values for these metrics, demonstrated smoother, more conservative, and less energy-intensive movements.

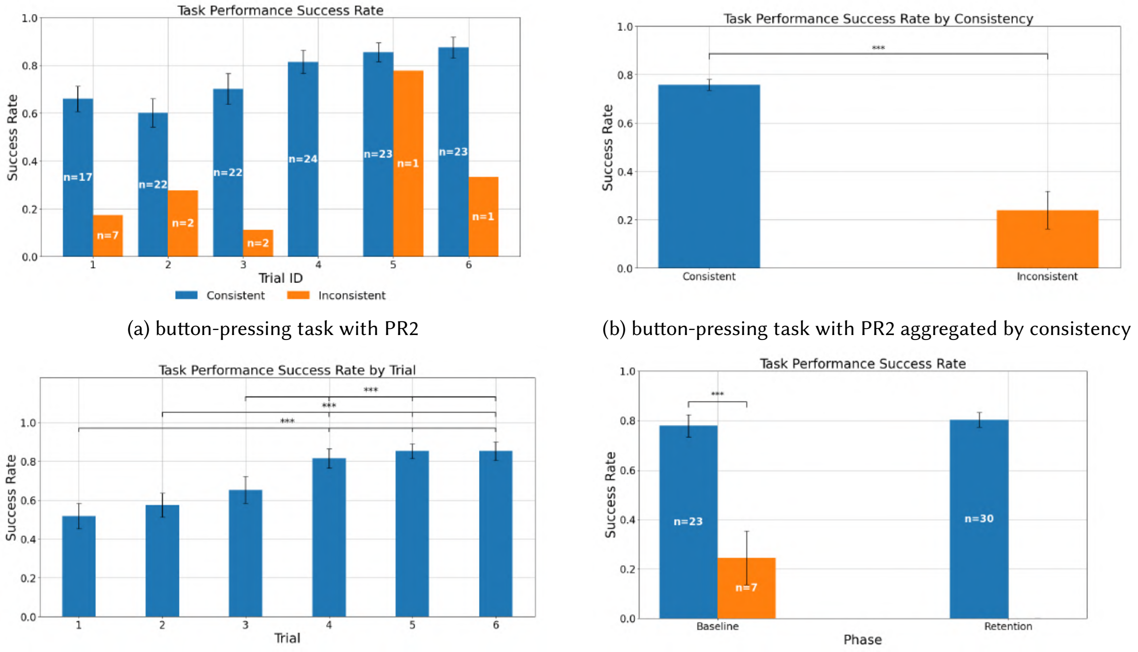
In the pick-and-place experiment, we applied the same approach by using the range of the proposed metrics from each participant's baseline and retention phases as input features for the K-means clustering algorithm. Unlike the button-pressing experiment, legibility was excluded here since there was only one goal (placement) location. Fig. 5b shows the standardized metric values for both clusters. Similar to the button-pressing experiment, the inconsistent cluster had a higher range for most metrics compared to the consistent cluster, which had a smaller range. Moreover, the distribution of the range for pathlength, jerk, curvature, and effort metrics differed significantly between the two clusters, while the distribution for manipulability and distance to joint limits was more similar between them. The average range values for each metric within the consistent and inconsistent clusters for both experiments are summarized in Appendix A.2, offering additional guidance for future applications of this clustering approach.

6.1 Consistency impact on the task and generalized success rate

After categorizing the consistent and inconsistent clusters, we investigated the task performance success rate for both clusters. As shown in Fig. 6, the task performance success rate when training the model with consistent trials is higher than when training with inconsistent ones. For the button-pressing experiment results, a one-way ANOVA was performed to investigate the impact of consistency on the success rate. Since the number of inconsistent sets in each trial is low(as shown in Fig. 6a), we aggregated the inconsistent and consistent sets across all trials, as shown in Fig. 6b. This analysis yields the relationship between the two vectors: one for the task success rate and the other for the consistency label (0 or 1). We found that the consistent group had a significantly higher success rate than the inconsistent group ($F(1, 142) = 45.46, p < 0.001$), with an average success rate of (0.76 ± 0.02)[2] in the consistent group and (0.24 ± 0.08) in the inconsistent group. Similarly, a one-way ANOVA was performed on the data from the pick-and-place experiment. We found that the consistent group had a significantly higher success rate than the inconsistent group ($F(1, 58) = 44.01, p < 0.001$) with an average success rate of (0.79 ± 0.03) in the consistent group and (0.25 ± 0.11) in the inconsistent group. These results support **H1-a**.

We conducted the same ANOVA analysis on the generalized performance success rate and found a similar result, as shown in Fig. 7. In the button-pressing experiment, the consistent group had a significantly higher success rate than the inconsistent group ($F(1, 143) = 36.57, p < 0.001$), with an average success rate of (0.63 ± 0.02) in the consistent group and (0.22 ± 0.06) in the inconsistent group, as shown in Fig. 7b. Similarly, in the pick-and-place experiment, the consistent group had a significantly higher success rate than the inconsistent group ($F(1, 58) = 15.50, p < 0.001$), with an average success rate of (0.81 ± 0.03) in the consistent group and (0.49 ± 0.10) in the inconsistent group, supporting **H1-b**. We further evaluated **H1-a** and **H1-b** using a different learning model (diffusion model) on the pick-and-place task, as detailed in A.3.

[2] $(mean \pm SE)$



(a) button-pressing task with PR2

(b) button-pressing task with PR2 aggregated by consistency

(c) button-pressing task with PR2 aggregated by trial

(d) pick and place task with UR5

Fig. 6. Average task performance success rate across trials/phases for both consistent and inconsistent groups. (a) Success rate across six trials in the button-pressing task with PR2, split by consistency. (b) Success rate aggregated by consistency across all trials for the PR2 task. (c) Success rate across trials for the PR2 task. (d) Success rate in the pick-and-place task with UR5, grouped by phase (baseline and retention) and consistency.

6.2 Impact of practice on demonstration consistency

To assess **H2** and investigate the impact of practice on the consistency of the demonstrations, a MANOVA (Multivariate Analysis of Variance) was performed to investigate the impact of practice on the range of the ten metrics, as detailed in Table 2. In the button-pressing experiment, the multivariate tests revealed a significant effect of practice (trials) on the consistency of the demonstrations ($Roy'sLargestRoot = 0.559, F(10, 110) = 6.144, p < 0.001, \eta_p^2 = 0.358$), with 35.8% of the variance in the range of metrics explained by the effect of the trial. Moreover, Fig. 6a shows that the number of inconsistent trials that are also first trials is higher compared to later trials, indicating that consistency improves with practice. Similarly, we conducted MANOVA on the range of nine metrics of the pick-and-place experiment data to investigate the practice/training impact on the consistency of these metrics. The test revealed a significant effect of practice/training on the consistency of the demonstrations ($Roy'sLargestRoot = 0.559, F(9, 21) = 4.35, p < 0.005, \eta_p^2 = 0.651$), with 65.1% of the variance in the range of metrics explained by the effect of practice, supporting **H2**.

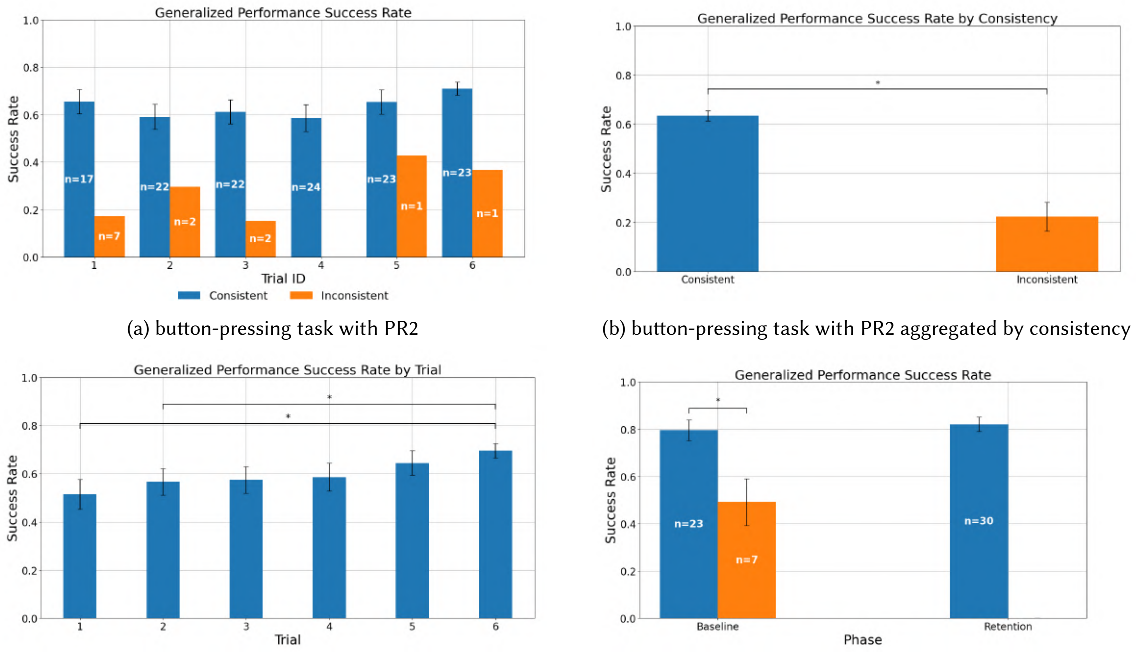
Univariate tests on the button-pressing experiment showed significant effects of the trials on the following range of metrics: jerk in Cartesian space, path orientation length, and effort ($p < 0.05$). They also revealed marginally significant effects of the trials on the range of metrics for pathlength in Cartesian space, pathlength in joint space, and manipulability ($p = 0.07$). Post hoc pairwise comparisons revealed that participants may need more time to achieve significantly better values for some metrics than others. For example, the participants showed a significant improvement after one trial in terms of pathlength in Cartesian and joint space, manipulability, and curvature. For metrics like jerk in Cartesian and joint space and effort, two trials were required for participants to better demonstrate the task and achieve significantly lower jerk values. Additionally, for legibility, only the first and last trials were significantly different, indicating that participants may need more time to generate legible demonstrations. Similarly in the pick-and-place experiment, univariate tests showed a significant effect of the practice on the range of the following metrics: pathlength in Cartesian and joint space, orientation pathlength and jerk in Cartesian and joint space $p < 0.01$.

Table 2. Results of MANOVA and Univariate Tests for Button-Pressing and Pick-and-Place Experiments

Experiment	**Roy's Largest Root**	**F-value**	**p-value**	**Partial Eta Squared (η_p^2)**
		MANOVA Results		
Button-Pressing	0.559	6.144	< 0.001	0.358
Pick-and-Place	0.559	4.350	0.003	0.651
		Univariate Test Results		
Metric	**Experiment**	**F-value**	**p-value**	**Partial Eta Squared (η_p^2)**
Path Orientation Length	Button-Pressing	2.40	0.04	0.10
Jerk in Cartesian Space	Button-Pressing	4.12	0.002	0.15
Effort	Button-Pressing	3.07	0.012	0.12
Pathlength in Cartesian Space	Button-Pressing	2.10	0.07	0.08
Pathlength in Joint Space	Button-Pressing	2.10	0.07	0.08
Manipulability	Button-Pressing	2.12	0.07	0.08
Pathlength in Cartesian Space	Pick-and-Place	13.23	0.001	0.31
Pathlength in Joint Space	Pick-and-Place	7.90	0.009	0.21
Path Orientation Length	Pick-and-Place	9.52	0.004	0.25
Jerk in Cartesian Space	Pick-and-Place	12.51	0.001	0.30
Jerk in Joint Space	Pick-and-Place	7.87	0.009	0.21

6.3 Practice impact on the task and generalized success rate

We hypothesized that practice improves the overall success rate for both task performance and generalized performance. This is because the consistency (as shown earlier) and other factors improve over time. To evaluate this hypothesis (**H3-a** and **H3-b**), a repeated ANOVA with a Greenhouse-Geisser correction was performed to account for the violation of the sphericity assumption, exploring the impact of practice (trials) on the success rate of task performance, as shown in Table 3. We found that practice has a significant impact on the success rate ($F(1, 23) = 23.47, p < 0.001, \eta_p^2 = 0.51$). The success rate improved from (0.52 ± 0.07) in the first trial to (0.85 ± 0.05) in the last trial. In addition, the post hoc pairwise comparison of the trials shows that trials 1, 2, and 3 (corresponding to the first session) have significantly lower success rates ($p < 0.001$) than trials 4, 5, and (corresponding to the second session) while there is no significant difference within the first three trials or within the last three trials. This highlights that the significant improvement is revealed in the second session on later days, which agrees with Walker et al. [53]. The same analysis was applied to the pick-and-place experiment task success rate. The repeated ANOVA showed task success rate in retention (after training) is significantly higher than the success rate in the baseline (before training) ($F(1, 29) = 9.09, p < 0.01, \eta_p^2 = 0.24$) with an average success rate of (0.66 ± 0.06) in the baseline and (0.80 ± 0.03) in the retention, supporting **H3-a**.



(a) button-pressing task with PR2

(b) button-pressing task with PR2 aggregated by consistency

(c) button-pressing task with PR2 aggregated by trial

(d) pick and place task with UR5

Fig. 7. Average generalized performance success rate across trials/phases for both consistent and inconsistent groups. (a) Generalized success rate across six trials in the button-pressing task with PR2. (b) Generalized success rate aggregated by consistency across all trials for the PR2 task. (c) Generalized success rate aggregated by trial for the PR2 task. (d) Generalized success rate in the pick-and-place task with UR5, grouped by phase (baseline and retention) and consistency.

Table 3. Repeated ANOVA Results for Task Performance Success Rates

Experiment	**Phases/Trials**	**F-value**	**p-value**	**Partial Eta Squared (η_p^2)**
Button-Pressing	6 Trials	23.47	< 0.001	0.51
Pick-and-Place	2 Phases	9.09	< 0.01	0.24

Trial/Phase	**Button-Pressing**	**Pick-and-Place**
Initial (First Trial/Baseline)	0.52 ± 0.07	0.66 ± 0.06
Final (Last Trial/Retention)	0.85 ± 0.05	0.80 ± 0.03

A repeated ANOVA was conducted on the generalized success rate (as detailed in Table 4), and we found a similar result. In the button-pressing experiment, practice has a significant impact on the generalized success rate ($F(1, 23) = 6.16, p < 0.05, \eta_p^2 = 0.21$) as shown in Fig. 7c. The success rate improved from (0.51 ± 0.06) in the first trial to (0.70 ± 0.03) in the last trial. In the pick-and-place experiment, training has a significant impact on the generalized success rate ($F(1, 29) = 5.87, p < 0.05, \eta_p^2 = 0.17$) as shown in Fig. 7d. The success rate improved from (0.73 ± 0.05) in the baseline(before training) to (0.82 ± 0.03) in the retention(after training), supporting **H3-b**.

The post hoc pairwise comparison of the trials in the button-pressing experiment shows that trials 1 and 2 are significantly less successful ($p < 0.05$) than trial 6, while there is no significant difference among the intermediate trials, as shown in Fig. 7c. This indicates that improvement in the generalized success rate is slower than in the task performance success rate, and more time is needed for the participants to achieve a good quality of demonstrations that attain a better generalized success rate.

Table 4. Repeated ANOVA Results for Generalized Success Rates

Experiment	**Phases/Trials**	**F-value**	**p-value**	**Partial Eta Squared (η_p^2)**
Button-Pressing	6 Trials	6.16	< 0.05	0.21
Pick-and-Place	2 Phases	5.87	< 0.05	0.17

Trial/Phase	**Button-Pressing**	**Pick-and-Place**
Initial (First Trial/Baseline)	0.51 ± 0.06	0.73 ± 0.05
Final (Last Trial/Retention)	0.70 ± 0.03	0.82 ± 0.03

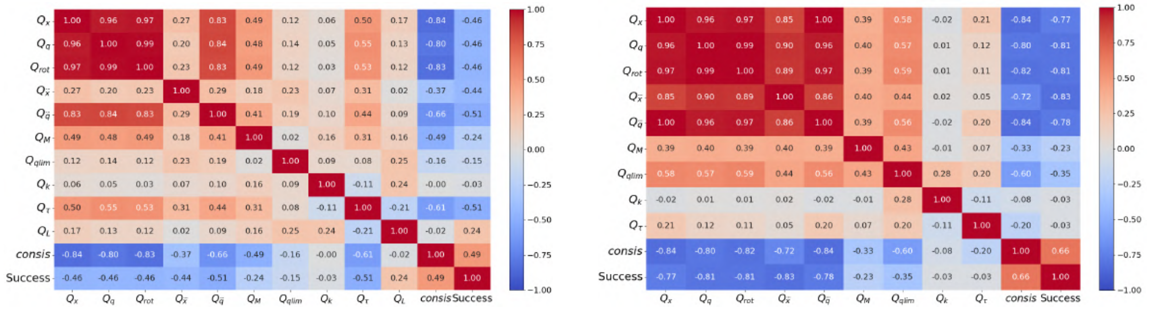
6.4 The importance of each metric to the task success rate

Thus far, we have demonstrated that consistency is critical for achieving higher success rates and that practice improves demonstration quality, which, in turn, enhances success rates. We have also observed that the proposed metrics show different patterns across trials and clusters. In this section, we explore the importance of each metric in determining success rates by investigating the question of *What are the key aspects to evaluate in demonstrations before learning?*

To address this, we conducted a correlation analysis between the task success rate and the range of ten proposed metrics. Additionally, we included a binary parameter, consistency, to examine its contribution to the success rate. Fig. 8 illustrates a heatmap showing the correlations between the metrics and the success rate in both experiments. In both experiments, pathlength, jerk, and consistency exhibit the highest correlations with task success rate. Specifically, pathlength and jerk have negative correlations with success rate, indicating that greater variability reduces success, while consistency shows a positive correlation, suggesting that more consistent demonstrations improve performance. The consistency label is computed based on clustering over the proposed metrics, and thus its relationship with success should be interpreted as associative rather than causal.

Other terms vary between experiments. Effort has a high correlation with success rate in the button-pressing experiment but a lower correlation in the pick-and-place. Conversely, the distance to joint limits correlates more strongly with success rate in the pick-and-place experiment. We also found that some metrics, like pathlength, are highly correlated with each other, and consistency correlates strongly with pathlength, jerk in joint space, and effort. This pattern aligns with Fig. 5, where pathlength, jerk, and effort show the most distinct patterns between clusters.

Motivated by these correlations, we performed multiple regression analysis, including interaction terms, to examine how changes in correlated metrics affect the success rate. The range of the proposed metrics collectively predicts 70.2% and 88.6% of the task success rate in the two experiments, respectively. Notably, the interactions among the ranges of pathlength and jerk metrics are the most significant factors with high coefficients for estimating the success rate in both experiments. Details of the regression models and the analysis are provided in Appendix A.4.



(a) button-pressing task with PR2

(b) pick and place task with UR5

Fig. 8. Heatmap showing the correlations between the task success rate and the range of the proposed metrics and consistency label. The abbreviated metric labels (e.g., Q_x–Q_L) correspond to the metrics listed in Table 1

6.5 The importance of each metric to the generalized success rate

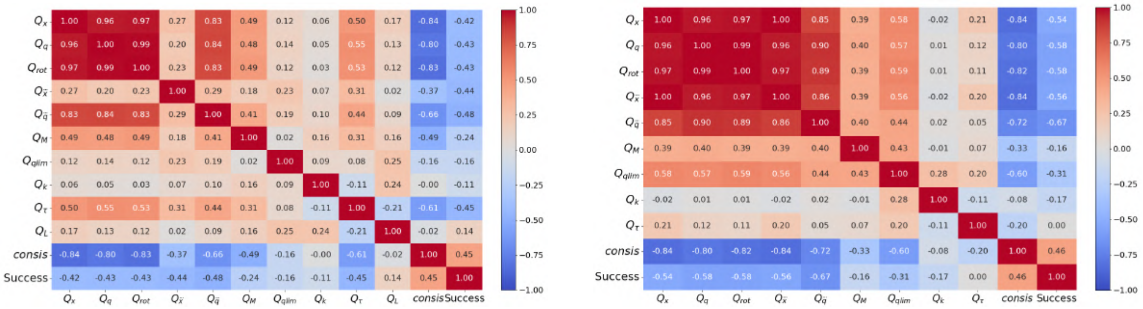
In this section, we examine the relationship between the generalized success rate and demonstration consistency using the proposed measures. Similar to the analysis of the task success rate, we first conducted a correlation analysis. Fig. 9 presents a heatmap showing the correlation between the generalized success rate, the range of the proposed measures, and the binary consistency variable across the two experiments. Pathlength, jerk, and consistency measures exhibit the highest correlation with the generalized success rate, mirroring the pattern observed for task success. Similarly, effort shows a high correlation with the generalized success rate in the button-pressing experiment but a weaker correlation in the pick-and-place experiment. Conversely, distance to joint limits has a stronger correlation with the generalized success rate in the pick-and-place experiment than in the button-pressing. Several measures are also strongly correlated with one another, which motivated the use of regression analysis with interaction terms to better understand their combined impact on success.

The regression models indicate that the range of the proposed metrics collectively predicts 75.6% and 90.7% of the generalized success rate in the two experiments, respectively. Interactions among the pathlength and jerk metrics are the most significant factors for estimating the generalized success rate in both experiments, as observed in the task success rate models. Details of the regression models and the analysis are provided in Appendix A.5.

7 DISCUSSION

In this paper, we propose a set of metrics to evaluate the quality of demonstrations prior to learning from demonstration (LfD). Specifically, we focused on consistency as a key measure of data quality and explored various motion characteristics to identify which metrics best predict learning performance. Unlike previous studies that artificially inject noise into data [10], we evaluated our approach using real demonstration data collected from users with varying levels of robotics expertise. The experimental results clearly demonstrate that consistency in demonstrations significantly impacts both the success rate of learning and its generalization in different robot tasks.

Our clustering analysis revealed that inconsistent demonstrations are characterized by larger ranges and higher means in path length, jerk, and effort. In contrast, consistent demonstrations exhibit smaller ranges in these metrics, indicating smoother, more conservative, and less energy-intensive movements. This finding supports the notion that



(a) button-pressing task with PR2

(b) pick and place task with UR5

Fig. 9. Heatmap showing the correlations between the generalized success rate and the range of the proposed metrics and consistency label. Metric definitions are provided in Table 1

smoother and more controlled demonstrations provide better input for learning algorithms, reducing the likelihood of overfitting to noisy or suboptimal data.

Statistical analyses further highlight the significant contribution of consistency to both task success rates and generalization performance. These findings align with Pomerleau's work [41], where consistent data provided clearer, more repeatable patterns for learning algorithms, leading to improved performance. Importantly, we observed that practice leads to significant improvements in consistency. However, the rate of improvement varied across metrics. For example, path length (in both Cartesian and joint spaces), manipulability, and curvature improved significantly after just one trial, whereas metrics like jerk, legibility, and effort required more practice to show meaningful improvement. This suggests that while some aspects of demonstrations can be refined quickly, others take time and experience to reach optimal levels. This opens up an exciting direction for future work, such as developing personalized training frameworks that track and adjust for individual user progress across specific metrics. By monitoring these metrics, we could avoid overtraining, which might otherwise diminish user performance, and ensure the highest quality demonstrations from each participant. Researchers have attempted to implement training curricula for users [2, 11, 45], but they did not provide a way to evaluate performance over trials, which the proposed metrics would help in achieving.

It is important to note that our focus throughout the analysis is on the consistency of each metric across demonstrations, rather than on the absolute values of the metrics themselves. While certain qualities, such as improving legibility, might lead to increased pathlengths or altered trajectories, our methodology considers demonstrations to be high-quality if such characteristics are expressed consistently across repetitions. Thus, even demonstrations that involve longer paths (to improve goal inference) can achieve high consistency scores if they are performed reliably and with low variability across trials.

All of the proposed metrics can be computed during data collection, prior to running the learning algorithm, with the exception of legibility, which requires prior knowledge of task goals. Legibility becomes especially important in human-robot interaction scenarios, where the robot's actions need to be easily understood by humans [19]. Thus, while legibility may not be necessary in all learning tasks, it plays a critical role in ensuring intuitive and seamless interaction in shared environments. It does so by evaluating goal-predictive motion rather than requiring any knowledge of the task's reward structure.

The correlation and regression analyses provided further insights into the metrics that are most influential in achieving high task success and generalization rates. They revealed that all metrics significantly impact the task learning and generalization success rate, either directly or through interaction with each other. However, the nature of the task influences the relative importance of each metric. For example, in the first experiment (button-pressing task in a constrained space), minimizing pathlength in joint space was essential to avoid self-collision and collision with the experimental setup. Orientation also played a key role in the button-pressing task, as the angle of approach was critical to successful execution. Interestingly, effort and jerk were highly correlated with pathlength and orientation, as shown in Fig. 8. Thus, their interactions have a high impact on the success rate, as shown in Table A2. Notably, even though these metrics had the highest impact, the influence of the other metrics cannot be ignored. When we reran the regression using only these four metrics and ignored the others, only 45% of the variance in the task success rate was captured, highlighting the importance of a comprehensive evaluation. These findings suggest that researchers should prioritize the selection of the most consistent demonstrations across all metrics to achieve high performance in both learning and generalization. While our correlation analysis revealed strong associations between certain metrics and task success, future work could employ causal inference techniques or controlled interventions to better understand the directional effects of demonstration quality on learning outcomes.

In the second experiment (pick-and-place task), pathlength and jerk again emerged as critical metrics for both task and generalization success. The consistency of these findings across experiments highlights the robustness of these metrics in predicting learning performance. However, we also observed differences in the role of other metrics. For instance, the range of effort was more highly negatively correlated with success in the first experiment than in the second one. Conversely, the range of distance to joint limits had a stronger negative correlation with success in the second experiment. These differences in metric influence likely stem from the task-specific constraints. In the button-pressing task, the robot was required to perform precise, localized motions in a constrained space. In this context, high joint effort may reflect unnecessary user-applied forces or inefficient joint configurations that compromise control precision. In contrast, the pick-and-place task involved a ceiling-mounted obstacle that constrained the overhead space. Successful trajectories required navigating around this obstacle using configurations away from joint limits. Demonstrations that pushed the robot close to its limits were more likely to result in unstable or inefficient motion, particularly when carrying a dynamic object like a liquid-filled bottle. This likely explains the stronger correlation between distance to joint limits and success in the pick-and-place task. These findings suggest that the specific nature of the task and its constraints influence which metrics are most important for optimizing learning and generalization.

Our proposed approach enables the evaluation of demonstration quality and consistency without the need to train the learning model, which sets it apart from prior work [22, 46]. Moreover, the proposed metrics implicitly capture inconsistencies in the solution strategy. For example, in the first experiment, users initially pushed the robot's shoulder joint to its limit to avoid collisions with a box. With practice, users discovered smoother and more efficient ways of manoeuvring around the box, as seen in Fig.2c. This pattern places the first trial in the inconsistent group, while subsequent trials fall into the consistent group. This suggests a promising direction for future work in active learning, where users could be guided to provide more consistent and higher-quality demonstrations, as explored in [22]. While their approach requires retraining the policy and evaluating performance to measure compatibility, our method reveals inconsistencies and offers insights into demonstration quality without the need for training, providing a more efficient and practical approach.

While consistency in demonstrations is crucial for reducing noise and ensuring reliable learning, diversity plays an equally important role in enabling generalization across varied task scenarios. Our paper emphasizes consistency to mitigate the adverse effects of inconsistent or noisy demonstrations, which can hinder learning performance. However, we acknowledge that a certain degree of diversity, particularly task-relevant variation, is essential for robust generalization. This perspective aligns with the findings of Mandlekar et al. [34], who show that policies trained on diverse demonstrations composed of overlapping substructures can generalize to unseen goal configurations. Similarly, Levy et al. [32] demonstrated that selecting diverse demonstrations enhances compositional generalization in semantic parsing tasks, a result that aligns with our earlier findings in [47].

In our experiments, we observed that consistent demonstrations, when encompassing a range of task parameters (e.g., varying start and goal configurations), can retain meaningful diversity while minimizing undesirable variability. This balance between consistency and diversity is pivotal; excessive uniformity may limit exposure to the breadth of task scenarios, whereas uncontrolled variability can introduce noise detrimental to learning. Future work could explore methodologies to quantify and incorporate structured diversity alongside consistency metrics. By doing so, we can better understand how these factors jointly influence generalization across different learning frameworks and task domains.

8 LIMITATIONS AND FUTURE WORK

While our proposed consistency metrics have shown promising results in structured manipulation tasks, several limitations should be acknowledged. First, although these metrics are well-suited to trajectory-based manipulation tasks, they may not directly generalize to other domains of robotics. Tasks involving dexterous hand manipulation, tactile feedback, locomotion, or high-dimensional perceptual inputs (e.g., vision-based navigation) may require additional or entirely different quality indicators. Our framework is general in its ability to evaluate consistency across various metrics, but the choice of metrics must be tailored to the specific task modality. Extending this approach to new task families will require adapting or learning domain-specific features that meaningfully capture demonstration quality.

Second, while our results indicate that consistency significantly impacts both task success and generalization performance, future studies with larger sample sizes could enable more fine-grained analysis of consistency dynamics over time. For instance, applying two-way ANOVA could help reveal how consistency evolves across repeated trials and interacts with task progression, offering deeper insight into the learning process and individual teaching strategies.

Third, although we employ TP-GMM in this study, our consistency metrics are algorithm-agnostic and not specific to any one imitation learning (IL) technique. TP-GMM was selected due to its interpretability, low sample complexity, and suitability for structured, trajectory-based tasks. However, the core principle behind our metrics—capturing variability in motion quality—can be applied to a broad range of IL methods. To this end, we provide a preliminary evaluation using a diffusion-based learning model in the appendix A.3. While the diffusion model replicated our main findings regarding consistency, it also exhibited different generalization behavior compared to TP-GMM, underscoring the importance of evaluating consistency effects across diverse modeling paradigms. Future work will expand this analysis to include additional algorithms, such as Behavioral Cloning and adversarial imitation learning, using larger and more diverse datasets. This direction is supported by prior work in both imitation and supervised learning, which has shown that the consistency and quality of training data significantly affect learning performance regardless of the model used [7, 23, 38].

Finally, our experiments used kinesthetic teaching to ensure direct user control of the robot and avoid confounds from teleoperation hardware. This allowed us to isolate natural human variability when studying the effect of demonstration consistency on robot learning. Future work could extend this analysis to teleoperation settings, where different control dynamics may influence consistency and its role in robot learning and generalization.

9 CONCLUSION

This paper presents an extensive set of metrics for evaluating the quality of demonstrations in Learning from Demonstration (LfD) tasks, with a particular focus on consistency as a key determinant of success. By rigorously analyzing real demonstration data from users of varying expertise, we have demonstrated that consistency in demonstrations significantly influences both the task success rate and generalization performance in robot learning. Our findings show that smoother, more controlled demonstrations, as indicated by smaller ranges in pathlength and jerk in both Cartesian and joint spaces, lead to better learning outcomes and enhanced generalization. Furthermore, our approach allows for real-time evaluation of demonstration quality, making it possible to assess data consistency prior to the learning process, thus ensuring that only high-quality data is used for robot training. The results of two user studies, involving distinct tasks and robot platforms, underline the robustness of the proposed metrics. Key metrics like pathlength and jerk in both Cartesian and joint spaces emerged as critical predictors of learning success across both experiments, highlighting their importance for a wide range of LfD tasks. Additionally, our work emphasizes the role of task-specific factors in shaping the contribution of each metric, suggesting that personalized training protocols could further improve the quality of demonstrations. Overall, this work fills an important gap in LfD research by offering a practical, data-driven method to evaluate and ensure demonstration quality before learning. It opens new avenues for optimizing robot learning and generalization in real-world scenarios, empowering non-expert users to teach robots effectively while reducing the risk of poor performance due to suboptimal data.

REFERENCES

[1] Farzad Aghazadeh, Bin Zheng, Mahdi Tavakoli, and Hossein Rouhani. 2023. Motion smoothness-based assessment of surgical expertise: The importance of selecting proper metrics. *Sensors* 23, 6 (2023), 3146.

[2] Gopika Ajaykumar, Gregory D Hager, and Chien-Ming Huang. 2023. Curricula for teaching end-users to kinesthetically program collaborative robots. *Plos one* 18, 12 (2023), e0294786.

[3] Pourya Aliasghari, Moojan Ghafurian, Chtystopher L Nehaniv, and Kerstin Dautenhahn. 2024. How Non-experts Kinesthetically Teach a Robot over Multiple Sessions: Diversity in Teaching Styles and Effects on Performance. *International Journal of Social Robotics* (2024), 1–27.

[4] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *AI magazine* 35, 4 (2014), 105–120.

[5] Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. 2009. A survey of robot learning from demonstration. *Robotics and autonomous systems* 57, 5 (2009), 469–483.

[6] Mark Beliaev, Andy Shih, Stefano Ermon, Dorsa Sadigh, and Ramtin Pedarsani. 2022. Imitation learning by estimating expertise of demonstrators. In *International Conference on Machine Learning*. PMLR, 1732–1748.

[7] Suneel Belkhale, Yuchen Cui, and Dorsa Sadigh. 2024. Data quality in imitation learning. *Advances in Neural Information Processing Systems* 36 (2024).

[8] Muhammad Bilal, Nir Lipovetzky, Denny Oetomo, and Wafa Johal. 2024. Beyond Success: Quantifying Demonstration Quality in Learning from Demonstration. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 5120–5127.

[9] Aude Billard, Sylvain Calinon, Ruediger Dillmann, and Stefan Schaal. 2008. Survey: Robot programming by demonstration. *Handbook of robotics* 59, BOOK_CHAP (2008).

[10] Daniel S Brown, Wonjoon Goo, and Scott Niekum. 2020. Better-than-demonstrator imitation learning via automatically-ranked demonstrations. In *Conference on robot learning*. PMLR, 330–359.

[11] Maya Cakmak and Leila Takayama. 2014. Teaching people how to teach robots: The effect of instructional materials and dialog design. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. 431–438.

[12] Sylvain Calinon. 2016. A tutorial on task-parameterized movement learning and retrieval. *Intelligent service robotics* 9, 1 (2016), 1–29.

[13] Sylvain Calinon and Aude G Billard. 2007. What is the teacher's role in robot programming by demonstration?: Toward benchmarks for improved learning. *Interaction Studies* 8, 3 (2007), 441–464.

[14] Jason Chen and Alex Zelinsky. 2003. Programing by demonstration: Coping with suboptimal teaching actions. *The International Journal of Robotics Research* 22, 5 (2003), 299–319.

[15] Sonia Chernova and Andrea L Thomaz. 2014. *Robot learning from human teachers*. Morgan & Claypool Publishers.

[16] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. 2024. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. *The International Journal of Robotics Research* (2024).

[17] Corinna Cichy and Stefan Rass. 2019. An overview of data quality frameworks. *Ieee Access* 7 (2019), 24634–24648.

[18] Adria Colomé and Carme Torras. 2014. Closed-loop inverse kinematics for redundant robots: Comparative assessment and two enhancements. *IEEE/ASME Transactions On Mechatronics* 20, 2 (2014), 944–955.

[19] Anca D Dragan, Kenton CT Lee, and Siddhartha S Srinivasa. 2013. Legibility and predictability of robot motion. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 301–308.

[20] Kerstin Fischer, Franziska Kirstein, Lars Christian Jensen, Norbert Krüger, Kamil Kukliński, Maria Vanessa aus der Wieschen, and Thiusius Rajeeth Savarimuthu. 2016. A comparison of types of robot control for programming by demonstration. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 213–220.

[21] Fabrizio Flacco, Alessandro De Luca, and Oussama Khatib. 2012. Prioritized multi-task motion control of redundant robots under hard joint constraints. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 3970–3977.

[22] Kanishk Gandhi, Siddharth Karamcheti, Madeline Liao, and Dorsa Sadigh. 2023. Eliciting compatible demonstrations for multi-human imitation learning. In *Conference on Robot Learning*. PMLR, 1981–1991.

[23] Venkat Gudivada, Amy Apon, and Junhua Ding. 2017. Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations. *International Journal on Advances in Software* 10, 1 (2017), 1–20.

[24] Jiawei Han, Micheline Kamber, and Jian Pei. 2012. Data Mining: Concepts and. *Techniques, Waltham: Morgan Kaufmann Publishers* (2012).

[25] John A Hartigan, Manchek A Wong, et al. 1979. A k-means clustering algorithm. *Applied statistics* 28, 1 (1979), 100–108.

[26] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.

[27] ISO/IEC. 2008. *Software engineering – Software product Quality Requirements and Evaluation (SQuaRE) – Data quality model*. Technical Report. International Standard under systematic review ISO/IEC 25012.

[28] Noémie Jaquier, Leonel Rozo, Darwin G Caldwell, and Sylvain Calinon. 2020. Geometry-aware manipulability learning, tracking, and transfer. *The International Journal of Robotics Research* (2020). https://doi.org/10.1177/0278364920946815

[29] Noémie Jaquier, Leonel Rozo, Darwin G Caldwell, and Sylvain Calinon. 2021. Geometry-aware manipulability learning, tracking, and transfer. *The International Journal of Robotics Research* 40, 2-3 (2021), 624–650.

[30] Michael Kaiser, Holger Friedrich, and Rudiger Dillmann. 1995. Obtaining good performance from a bad teacher. In *Programming by Demonstration vs. Learning from Examples Workshop at ML*, Vol. 95.

[31] Nuno Laranjeiro, Seyma Nur Soydemir, and Jorge Bernardino. 2015. A survey on data quality: classifying poor data. In *2015 IEEE 21st Pacific rim international symposium on dependable computing (PRDC)*. IEEE, 179–188.

[32] Itay Levy, Ben Bogin, and Jonathan Berant. 2022. Diverse demonstrations improve in-context compositional generalization. *arXiv preprint arXiv:2212.06800* (2022).

[33] Alain Liegeois et al. 1977. Automatic supervisory control of the configuration and behavior of multibody mechanisms. *IEEE transactions on systems, man, and cybernetics* 7, 12 (1977), 868–871.

[34] Ajay Mandlekar, Danfei Xu, Roberto Martín-Martín, Silvio Savarese, and Li Fei-Fei. 2020. Learning to generalize across long-horizon tasks from human demonstrations. *arXiv preprint arXiv:2003.06085* (2020).

[35] Ajay Mandlekar, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. 2021. What matters in learning from offline human demonstrations for robot manipulation. *arXiv preprint arXiv:2108.03298* (2021).

[36] Andre Meixner, Mischa Carl, Franziska Krebs, Noémie Jaquier, and Tamim Asfour. 2024. Towards unifying human likeness: Evaluating metrics for human-like motion retargeting on bimanual manipulation tasks. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 13015–13022.

[37] Yoshihiko Nakamura and Hideo Hanafusa. 1986. Inverse kinematic solutions with singularity robustness for robot manipulator control. *J. Dyn. Sys., Meas., Control.* 108, 3 (1986), 163–171.

[38] Takayuki Osa, Joni Pajarinen, Gerhard Neumann, J Andrew Bagnell, Pieter Abbeel, Jan Peters, et al. 2018. An algorithmic perspective on imitation learning. *Foundations and Trends® in Robotics* 7, 1-2 (2018), 1–179.

[39] Ana-Lucia Pais Ureche and Aude Billard. 2015. 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*. 37–38.

[40] Affan Pervez and Dongheui Lee. 2018. Learning task-parameterized dynamic movement primitives using mixture of GMMs. *Intelligent Service Robotics* 11, 1 (2018), 61–78.

[41] Dean A Pomerleau. 1988. Alvinn: An autonomous land vehicle in a neural network. *Advances in neural information processing systems* 1 (1988).

[42] Harish Ravichandar, Athanasios S Polydoros, Sonia Chernova, and Aude Billard. 2020. Recent advances in robot learning from demonstration. *Annual review of control, robotics, and autonomous systems* 3, 1 (2020), 297–330.

[43] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 234–241.

[44] Boem-Sahng Ryuh. 1989. *Robot trajectory planning using the curvature theory of ruled surfaces*. Ph.D. Dissertation. Purdue University.

[45] Maram Sakr, Martin Freeman, H F Machiel Van der Loos, and Elizabeth Croft. 2020. Training Human Teacher to Improve Robot Learning from Demonstration: A Pilot Study on Kinesthetic Teaching. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 800–806.

[46] Maram Sakr, Zexi Jesse Li, HF Machiel Van der Loos, Dana Kulić, and Elizabeth A Croft. 2022. Quantifying demonstration quality for robot learning and generalization. *IEEE Robotics and Automation Letters* 7, 4 (2022), 9659–9666.

[47] Maram Sakr, Zhikai Zhang, Benjamin Li, Haomiao Zhang, H. F. Machiel Van der Loos, Dana Kulić, and Elizabeth Croft. 2025. How Can Everyday Users Efficiently Teach Robots by Demonstration? *ACM Transactions on Human-Robot Interaction (THRI)* (May 2025). https://doi.org/10.1145/3737892

[48] Richard A Schmidt and Craig A Wrisberg. 2008. *Motor learning and performance: A situation-based learning approach.* Human kinetics.

[49] Aran Sena and Matthew Howard. 2020. Quantifying Teaching Behavior in Robot Learning from Demonstration. *The International Journal of Robotics Research* 39, 1 (2020), 54–72.

[50] Voot Tangkaratt, Bo Han, Mohammad Emtiyaz Khan, and Masashi Sugiyama. 2019. Vild: Variational imitation learning with diverse-quality demonstrations. *arXiv preprint arXiv:1909.06769* (2019).

[51] Voot Tangkaratt, Bo Han, Mohammad Emtiyaz Khan, and Masashi Sugiyama. 2020. Variational imitation learning with diverse-quality demonstrations. In *International Conference on Machine Learning*. PMLR, 9407–9417.

[52] Nikolaus Vahrenkamp and Tamim Asfour. 2015. Representing the robot's workspace through constrained manipulability analysis. *Autonomous Robots* 38, 1 (2015), 17–30.

[53] Matthew P Walker, Tiffany Brakefield, Joshua Seidman, Alexandra Morgan, J Allan Hobson, and Robert Stickgold. 2003. Sleep and the time course of motor skill learning. *Learning & memory* 10, 4 (2003), 275–284.

[54] Sebastian Wallkotter, Mohamed Chetouani, and Ginevra Castellano. 2022. A new approach to evaluating legibility: Comparing legibility frameworks using framework-independent robot motion trajectories. *arXiv preprint arXiv:2201.05765* (2022).

[55] Ziyu Wang, Josh S Merel, Scott E Reed, Nando de Freitas, Gregory Wayne, and Nicolas Heess. 2017. Robust imitation of diverse behaviors. *Advances in Neural Information Processing Systems* 30 (2017).

[56] Nils Wilde, Alexandru Blidaru, Stephen L Smith, and Dana Kulić. 2020. Improving user specifications for robot behavior through active preference learning: Framework and evaluation. *The International Journal of Robotics Research* 39, 6 (2020), 651–667.

[57] Yueh-Hua Wu, Nontawat Charoenphakdee, Han Bao, Voot Tangkaratt, and Masashi Sugiyama. 2019. Imitation learning from imperfect demonstration. In *International Conference on Machine Learning*. PMLR, 6818–6827.

[58] Shuqi Xu, Hao Zhang, and Zhuping Wang. 2024. Learning to Perform Trajectory Generation From Low-Quality Demonstrations. *IEEE Transactions on Neural Networks and Learning Systems* (2024).

[59] Tsuneo Yoshikawa. 1985. Manipulability of robotic mechanisms. *The international journal of Robotics Research* 4, 2 (1985), 3–9.

[60] Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 2024. 3D Diffusion Policy: Generalizable Visuomotor Policy Learning via Simple 3D Representations. In *Proceedings of Robotics: Science and Systems (RSS)*.

[61] Songyuan Zhang, Zhangjie Cao, Dorsa Sadigh, and Yanan Sui. 2021. Confidence-aware imitation learning from demonstrations with varying optimality. *Advances in Neural Information Processing Systems* 34 (2021), 12340–12350.

APPENDIX

A.1 LEGIBILITY CALCULATIONS

The formula for calculating the legibility score is shown in Table 1, where $P(G_i|\xi_{1:t})$ is the posterior probability of goal G_i given the trajectory segment $\xi_{1:t}$. The probability is computed using the cost function for each potential goal, which calculates the likelihood of a trajectory for a given goal. It combines early differentiation and progress towards the goal as follows:

$$C(\xi, G) = w_1 \left(\frac{1}{D_{\text{early}}}\right) + w_2 \left(\frac{1}{P_{\text{goal}}}\right) \tag{3}$$

where D_{early} is the early differentiation score, P_{goal} is the progress towards the goal, and w_1 and w_2 are the weights for each term. Early differentiation measures how distinguishable the trajectory is from other potential goals early in the motion as follows:

$$D_{\text{early}} = \frac{1}{n} \sum_{i=1}^{n} ||\xi_{1:t} - G|| \tag{4}$$

where n is the number of potential goals, $\xi_{1:t}$ is the early trajectory segment, and G is the goal. Progress towards the goal measures how directly the trajectory moves towards the goal as follows:

$$P_{\text{goal}} = \frac{\sum_{i=1}^{n} \max(\text{proj}(\vec{\xi}_i, \vec{G}), 0)}{\|G - \xi_0\|} \tag{5}$$

where $\vec{\xi}_i$ is the vector segment of the trajectory, $\vec{G}$ is the vector towards the goal, and ξ_0 is the starting point of the trajectory.

A.2 SUMMARY OF METRIC RANGES FOR CONSISTENCY CLUSTERS

To support future research and practice, Table A1 reports the average range values for each of the proposed consistency metrics, computed separately for consistent and inconsistent demonstration clusters in both user studies. While these values are task-specific and depend on factors such as robot kinematics, task constraints, and user proficiency, they offer practical reference points that may guide consistency evaluation in similar settings. Notably, the absolute range values differ between the button-pressing and pick-and-place tasks due to differences in task complexity and motion scale. Therefore, these values should be interpreted within the context of each task, rather than used for direct cross-task comparison.

Table A1. Average range values for each metric in consistent vs. inconsistent demonstrations for both experiments.

Task	Metric	Consistent	Inconsistent
Button-Pressing (PR2)	Pathlength in Cartesian Space	0.43	2.52
	Pathlength in Joint Space	1.33	10.71
	Path Orientation Length	1.13	9.58
	Jerk in Cartesian Space	1.50×10^{3}	2.43×10^{3}
	Jerk in Joint Space	4.40×10^{3}	1.39×10^{4}
	Manipulability	0.032	0.054
	Distance to Joint Limits	0.042	0.051
	Curvature	235.12	236.74
	Joint Effort	68.93	234.78
	Legibility	0.88	0.90
Pick-and-Place (UR5)	Pathlength in Cartesian Space	2.45	10.32
	Pathlength in Joint Space	7.31	30.67
	Path Orientation Length	6.05	29.08
	Jerk in Cartesian Space	1.13×10^{5}	4.33×10^{5}
	Jerk in Joint Space	1.38×10^{5}	5.68×10^{5}
	Manipulability	0.028	0.044
	Distance to Joint Limits	0.079	0.205
	Curvature	2.79×10^{24}	5.58×10^{24}
	Joint Effort	4.78	5.88

A.3 CONSISTENCY IMPACT ACROSS DIFFERENT LEARNING METHODS

To evaluate the robustness of our hypotheses **H1-a** and **H1-b** across different learning approaches, we trained a diffusion-based trajectory generation model on the pick-and-place task. Denoising Diffusion Probabilistic Models (DDPMs)[26] have gained popularity in robot policy and trajectory generation due to their strong generative performance and ease of adaptation[16, 60]. Unlike TP-GMM, which models trajectories as sequences over time, diffusion models treat the entire trajectory as a single sample. During training, Gaussian noise is incrementally added in a forward process, and a neural network, typically a U-Net [43], is trained to reverse this process by denoising. The model is supervised using a mean squared error (MSE) loss, which predicts the added noise from the ground truth trajectory. In our experiment, the model was conditioned on both the pickup and place poses, including end-effector positions, orientations (quaternions), and gripper states. These conditioning signals were concatenated and provided at each denoising step to guide the model toward task-relevant trajectories.

As shown in Fig.10a, task performance remained significantly higher in the consistent group compared to the inconsistent group, ($F(1, 58) = 8.07$, $p = 0.006$), with average success rates of (0.84 ± 0.02) and (0.70 ± 0.04), respectively. In the generalization performance (Fig.10b), the consistent group also outperformed the inconsistent group, though the difference was marginally significant, ($F(1, 58) = 3.50$, $p = 0.066$), with success rates of (0.26 ± 0.01) and (0.22 ± 0.03), respectively.

While both TP-GMM and the diffusion model benefited from consistent demonstrations, their absolute performance profiles diverged. The diffusion model slightly outperformed TP-GMM on the task performance for the consistent group (0.84 ± 0.02 vs. 0.79 ± 0.03), reflecting its strength in faithfully reproducing training data. However, it underperformed TP-GMM in generalization: with consistent demonstrations, the diffusion model achieved only a (0.26 ± 0.01) success rate on novel pickup and place combinations, compared to (0.81 ± 0.03) for TP-GMM.

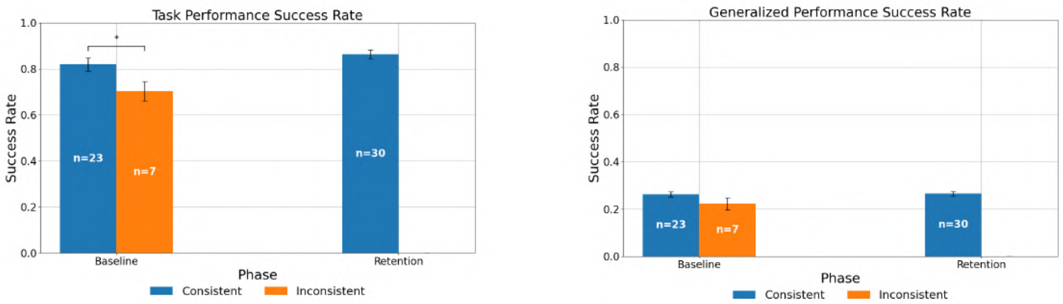
We attribute this difference largely to data efficiency between the two models. In our experiment, each diffusion policy was trained on only four demonstration trajectories per participant, an extremely low data regime in which the model likely overfitted to the training scenarios. Most diffusion-based robot learning studies utilize an order of magnitude more demonstrations to attain robust performance. For example, Diffusion Policy [16] was benchmarked with around 100–200 demonstrations per task, and even more data-efficient variants like the 3D Diffusion Policy (DP3) [60] demonstrated good performance with some tasks using 10 demonstrations in simulation and around 40 demonstrations per task on real robots. Additionally, DP3 incorporates a richer set of inputs—including RGB-D images, language instructions, and 3D voxel grids, whereas our model relies solely on kinematic features. These comparisons highlight the diffusion model's sensitivity to data scarcity and suggest that overfitting may have limited its generalization capability in our setting. In contrast, TP-GMM explicitly models temporal structure and can generalize from fewer examples by leveraging time-aligned Gaussian components.

Despite these differences, the consistency effect emerged in both approaches. In the diffusion model, as in TP-GMM, consistent demonstrations yielded superior performance metrics relative to inconsistent ones. This reinforces our central conclusion: enhancing the quality and consistency of demonstrations is essential for achieving robust and generalizable robot learning, regardless of the learning paradigm used.

A.4 REGRESSION ANALYSIS OF TASK SUCCESS RATE

Initially, linear regression was employed, but it captured only a small variation of the success rate using linear relationships with the predictors. Consequently, stepwise regression was applied to consider interactions among the predictors and their quadratic terms. Additionally, we standardized the predictors and outcome values to facilitate the interpretation of the main effects and interaction terms.

Table A2 shows the resulting regression model that represents the relationship between the range of the metrics and the task success rate. The model is significant and captures 70.2% of the variance in the success rate. The model shows a



(a) Task performance success rate

(b) Generalized performance success rate

Fig. 10. Average task and generalized performance success rates in the pick-and-place task using the diffusion model.

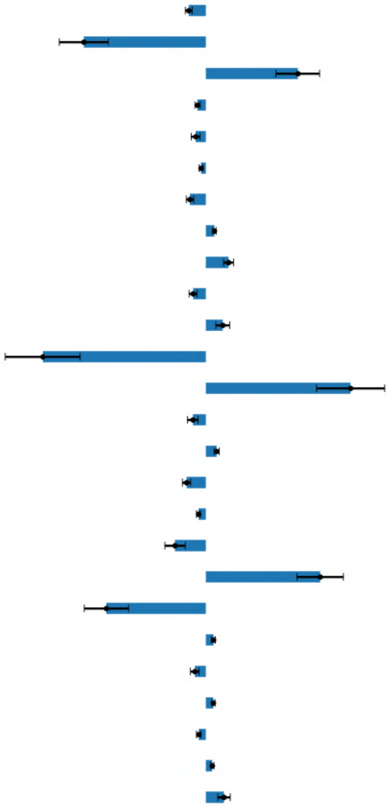
significant linear relationship with eight predictors: ran_path_joint, ran_orient, ran_jerk_cart, ran_jerk_joint, ran_joint_limit, ran_curv, ran_legibility, and consistency. While the other metrics exhibit a more complex relationship with the success rate through interaction and quadratic terms, this indicates the importance of evaluating the metrics together rather than focusing on any single metric.

The visualization in Table A2 highlights the coefficients and standard errors of the significant terms. Notably, ran_path_joint, ran_orient, and their interactions with ran_jerk_joint and ran_effort are crucial for determining success, given their higher coefficient values. The negative coefficient of ran_path_joint suggests that shorter, more consistent paths reduce unnecessary movements, thus lowering the likelihood of errors [38]. Variability in both ran_path_joint and ran_jerk_joint reflects inefficient, erratic movements, detrimental to precision-based tasks, which is why their interaction term negatively impacts success rate. Conversely, the positive interaction between ran_path_joint and ran_effort suggests that aligning effort with path variability improves control and adaptation, enhancing performance.

The range of ran_orient has a positive effect on the success rate, indicating the importance of flexible orientation adjustments to complete the task, especially in high-constraint environments. Additionally, its interaction with ran_jerk_joint positively influences success, suggesting that quick, even jerky, orientation adjustments are beneficial in such contexts. However, its interaction with ran_effort shows a negative coefficient, indicating that excessive variability in both orientation and effort reduces success. This suggests that high, consistent effort associated with orientation adjustments may impede performance.

Similarly, a regression model was developed for the second experiment, examining the relationship between task performance success rate and the range of metrics. Table A3 illustrates the model and provides a visualization of the coefficients and standard errors of significant terms. The model is significant, capturing 88.6% of the variance in success rate. Key interactions involve ran_path_cart, ran_path_joint, ran_orient, ran_jerk_cart, and ran_jerk_joint, with higher coefficient values highlighting their importance in estimating success rate. The positive interaction between ran_path_joint and ran_orient suggests that flexible orientation adjustments, alongside variability in joint movements, enhance performance in dynamic tasks. However, the negative interaction between ran_path_cart and ran_jerk_cart indicates that excessive variability in both metrics can hinder task performance. Moreover, interactions involving ran_jerk_joint with both ran_orient and ran_jerk_cart exhibit contrasting effects: the former negatively impacts success, reflecting that erratic joint movements combined with orientation changes reduce performance, while

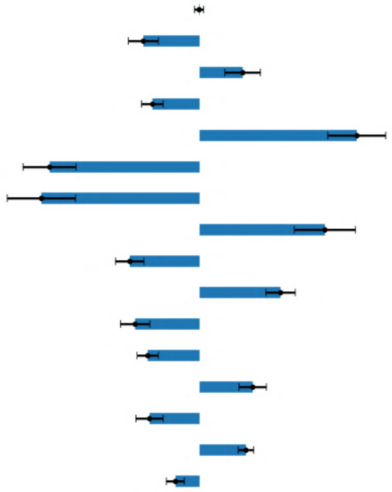
Table A2. Regression model detailing the relationship between the range of the proposed metrics and the task success rate in the button-pressing task. Variable abbreviations: `ran_path_cart` = range of Cartesian pathlength, `ran_path_joint` = range of joint space pathlength, `ran_orient` = range of orientation pathlength, `ran_jerk_cart` = range of Cartesian jerk, `ran_jerk_joint` = range of joint space jerk, `ran_manip` = range of manipulability, `ran_joint_limit` = range of distance to joint limits, `ran_curv` = range of curvature, `ran_effort` = range of joint effort, `ran_legibility` = range of legibility, and `consistency` = consistency label (binary).

Coefficient	**Estimate**	**SE**	**t-Statistic**	**Visualization of coefficients and their standard errors**
(Intercept)	-0.5499	0.1092	-5.0347	
ran_path_joint	**-3.8873**	0.7807	-4.9794	
ran_orient	**2.9233**	0.6930	4.2184	
ran_jerk_cart	-0.2674	0.0884	-3.0254	
ran_jerk_joint	-0.3258	0.1448	-2.2505	
ran_joint_limit	-0.1508	0.0612	-2.4649	
ran_curv	-0.5086	0.1198	-4.2468	
ran_legibility	0.2706	0.0595	4.5450	
consistency	0.7208	0.1574	4.5795	
ran_orient:ran_jerk_cart	-0.4071	0.1274	-3.1964	
ran_path_cart:ran_jerk_joint	0.5346	0.2232	2.3950	
ran_path_joint:ran_jerk_joint	**-5.1903**	1.1900	-4.3617	
ran_orient:ran_jerk_joint	**4.6017**	1.0848	4.2420	
ran_jerk_cart:ran_jerk_joint	-0.4149	0.1721	-2.4112	
ran_jerk_cart:ran_manip	0.3448	0.0848	4.0686	
ran_jerk_joint:ran_manip	-0.6147	0.1307	-4.7043	
ran_path_joint:ran_joint_limit	-0.2345	0.0641	-3.6568	
ran_path_cart:ran_curv	-0.9810	0.3257	-3.0124	
ran_path_joint:ran_effort	**3.6332**	0.7405	4.9061	
ran_orient:ran_effort	**-3.1695**	0.7050	-4.4955	
ran_manip:ran_effort	0.2374	0.0628	3.7817	
ran_curv:ran_effort	-0.3518	0.1390	-2.5321	
ran_jerk_cart:ran_legibility	0.2371	0.0509	4.6618	
ran_manip:ran_legibility	-0.2263	0.0754	-3.0016	
ran_jerk_cart^2	0.2007	0.0502	3.9971	
ran_jerk_joint^2	0.5752	0.1976	2.9117	

Number of observations:	144
Error degrees of freedom:	118
Root Mean Squared Error:	0.601
R-squared:	0.702
Adjusted R-Squared:	0.639
F-statistic vs. constant model:	11.1
p-value:	9.14e-21

the latter positively impacts success, suggesting that coordinated variability in jerk across joint and Cartesian spaces improves adaptability, leading to better task outcomes.

Table A3. Regression model detailing the relationship between the range of the proposed metrics and the task success rate in the pick-and-place task. Variable abbreviations: `ran_path_cart` = range of Cartesian pathlength, `ran_path_joint` = range of joint space pathlength, `ran_orient` = range of orientation pathlength, `ran_jerk_cart` = range of Cartesian jerk, `ran_jerk_joint` = range of joint space jerk, `ran_manip` = range of manipulability, `ran_joint_limit` = range of distance to joint limits, `ran_curv` = range of curvature, `ran_effort` = range of joint effort, `ran_legibility` = range of legibility, and `consistency` = consistency label (binary).

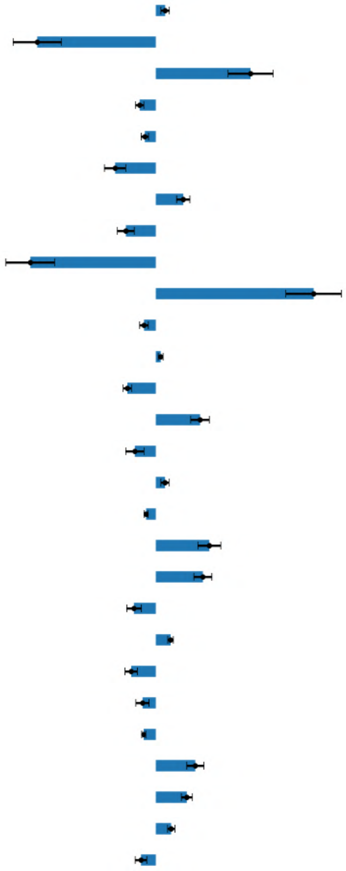
Coefficient	**Estimate**	**SE**	**t-Statistic**	**Visualization of coefficients and their standard errors**
Intercept	-0.011078	0.090473	-0.12244	
ran_path_cart	-1.1226	0.30021	-3.7392	
ran_path_joint	0.85762	0.35571	2.411	
ran_jerk_joint	-0.94164	0.2167	-4.3454	
ran_path_joint : ran_jerk_cart	**3.135**	0.57773	5.4264	
ran_path_cart : ran_jerk_cart	**-2.9919**	0.52756	-5.6711	
ran_jerk_cart : ran_jerk_joint	**-3.1553**	0.68183	-4.6277	
ran_jerk_cart : ran_jerk_joint	**2.4976**	0.61275	4.0761	
ran_path_cart : ran_manip	-1.3935	0.28438	-4.9002	
ran_path_joint : ran_manip	1.6134	0.29548	5.4603	
ran_jerk_cart : ran_curv	-1.2869	0.29361	-4.3828	
ran_manip : ran_curv	-1.0353	0.21373	-4.8443	
ran_path_cart : ran_effort	1.0596	0.27262	3.8867	
ran_path_joint : ran_effort	-0.99994	0.27296	-3.6633	
ran_curv : ran_effort	0.92439	0.15412	5.998	
ran_jerk_cart : consistency	-0.48341	0.18134	-2.6657	
Number of observations:	60			
Error degrees of freedom:	44			
Root Mean Squared Error:	0.392			
R-squared:	0.886			
Adjusted R-Squared:	0.847			
F-statistic vs. constant model:	22.7			
p-value:	6.95e-16			

A.5 REGRESSION ANALYSIS OF GENERALIZED SUCCESS RATE

Similar regression analysis was conducted to explore the relationship between the range of the metrics and the `consistency` label with the generalized success rate for the two experiments. Table A4 presents the regression model detailing the relationship between the range of the proposed metrics and the generalized success rate of the button-pressing experiment. The model is significant, explaining 75.6% of the variance in the generalized success rate. Notably, the same metrics that were important in predicting the task success rate also contribute the most to the generalized success rate. These key metrics include `ran_path_joint`, `ran_orient`, and their interactions with `ran_jerk_joint`. This underscores the critical role of these metrics in predicting both task and generalized success.

In the pick-and-place experiment, the regression model (Table A5) explains 90.7% of the variance in the generalized success rate. Key interactions involving `ran_path_cart`, `ran_path_joint`, `ran_jerk_cart`, and `ran_jerk_joint` are among the most important terms, similar to the task performance model. This underscores the critical role of these metrics in predicting both task and generalized success. `consistency` also plays a critical role, positively affecting

Table A4. Regression model detailing the relationship between the range of the proposed metrics and the generalized success rate in the button-pressing task. Variable abbreviations: `ran_path_cart` = range of Cartesian pathlength, `ran_path_joint` = range of joint space pathlength, `ran_orient` = range of orientation pathlength, `ran_jerk_cart` = range of Cartesian jerk, `ran_jerk_joint` = range of joint space jerk, `ran_manip` = range of manipulability, `ran_joint_limit` = range of distance to joint limits, `ran_curv` = range of curvature, `ran_effort` = range of joint effort, `ran_legibility` = range of legibility, and `consistency` = consistency label (binary).

Coefficient	**Estimate**	**SE**	**t-Statistic**	**Visualization of coefficients and their standard errors**
(Intercept)	0.27975	0.12896	2.1692	
ran_path_joint	**-3.5869**	0.7299	-4.9143	
ran_orient	**2.8648**	0.68134	4.2047	
ran_jerk_joint	-0.49044	0.12662	-3.8734	
ran_effort	-0.33265	0.11054	-3.0092	
consistency	-1.2349	0.32368	-3.8152	
ran_path_cart : ran_jerk_cart	0.82844	0.19886	4.1659	
ran_orient : ran_jerk_cart	-0.90945	0.25697	-3.5391	
ran_path_joint : ran_jerk_joint	**-3.7961**	0.73839	-5.141	
ran_orient : ran_jerk_joint	**4.7715**	0.83905	5.6868	
ran_jerk_cart : ran_jerk_joint	-0.35898	0.13353	-2.6883	
ran_jerk_cart : ran_manip	0.1458	0.073164	1.9927	
ran_jerk_joint : ran_manip	-0.86245	0.12318	-7.0017	
ran_path_cart : ran_joint_limit	1.3359	0.28081	4.7574	
ran_path_joint : ran_joint_limit	-0.63637	0.27687	-2.2984	
ran_jerk_joint : ran_joint_limit	0.27406	0.12249	2.2373	
ran_manip : ran_joint_limit	-0.29517	0.061987	-4.7617	
ran_path_joint : ran_curv	1.6195	0.34553	4.687	
ran_path_joint : ran_effort	1.4209	0.27381	5.1892	
ran_jerk_joint : ran_effort	-0.66221	0.22121	-2.9936	
ran_manip : ran_effort	0.44656	0.08497	5.2555	
ran_curv : ran_effort	-0.74627	0.18647	-4.0022	
ran_path_joint : ran_legibility	-0.40238	0.20063	-2.0056	
ran_joint_limit : ran_legibility	-0.36737	0.059606	-6.1633	
ran_path_joint : consistency	1.1933	0.25787	4.6274	
ran_joint_limit : consistency	0.93652	0.16325	5.7367	
ran_effort : consistency	0.46362	0.10807	4.2899	
ran_legibility : consistency	-0.44841	0.17632	-2.5432	
Number of observations:	144			
Error degrees of freedom:	116			
Root Mean Squared Error:	0.549			
R-squared:	0.756			
Adjusted R-Squared:	0.699			
F-statistic vs. constant model:	13.3			
p-value:	2.43e-24			

the generalized success rate, indicating that consistent demonstrations significantly improve the robot's ability to generalize.

Table A5. Regression model detailing the relationship between the range of the proposed metrics and the generalized success rate in the pick-and-place task. Variable abbreviations: `ran_path_cart` = range of Cartesian pathlength, `ran_path_joint` = range of joint space pathlength, `ran_orient` = range of orientation pathlength, `ran_jerk_cart` = range of Cartesian jerk, `ran_jerk_joint` = range of joint space jerk, `ran_manip` = range of manipulability, `ran_joint_limit` = range of distance to joint limits, `ran_curv` = range of curvature, `ran_effort` = range of joint effort, and `consistency` = consistency label (binary).

Coefficient	**Estimate**	**SE**	**t-Statistic**	**Visualization of coefficients and their standard errors**
Intercept	-2.1426	0.25463	-8.4146	
ran_path_joint	1.5602	0.2692	5.7956	
ran_jerk_joint	-1.1614	0.2206	-5.2647	
ran_manip	0.46128	0.090664	5.0878	
ran_effort	0.45899	0.085092	5.394	
consistency	**6.933**	0.83754	8.2778	
ran_path_cart : ran_path_joint	**11.712**	1.1528	10.159	
ran_path_cart : ran_jerk_cart	**-12.053**	1.1215	-10.748	
ran_path_joint : ran_jerk_joint	**-8.8461**	0.89576	-9.8755	
ran_jerk_cart : ran_jerk_joint	**8.4328**	0.85875	9.8198	
ran_path_cart : ran_manip	-1.649	0.27123	-6.0799	
ran_path_joint : ran_manip	2.1474	0.27391	7.8396	
ran_joint_limit : ran_curv	-0.22681	0.051082	-4.4401	
ran_path_joint : ran_effort	-4.2964	0.44981	-9.5516	
ran_jerk_cart : ran_effort	4.946	0.49788	9.9341	
ran_joint_limit : ran_effort	-0.28161	0.10056	-2.8003	
ran_jerk_cart : consistency	-4.5763	0.56098	-8.1576	
ran_jerk_joint : consistency	1.6174	0.3304	4.8954	
ran_manip : consistency	-0.69759	0.2299	-3.0343	
ran_effort : consistency	-0.74814	0.13176	-5.6778	

Number of observations:	60
Error degrees of freedom:	40
Root Mean Squared Error:	0.371
R-squared:	0.907
Adjusted R-Squared:	0.862
F-statistic vs. constant model:	20.5
p-value:	6.37e-15