Shaping Imbalance into Balance: Active Robot Guidance of Human Teachers for Better Learning from Demonstrations

Muhan Hou
Department of Computer Science
Vrije Universiteit Amsterdam
Amsterdam, the Netherlands
m.hou@vu.nl

Koen Hindriks
Department of Computer Science
Vrije Universiteit Amsterdam
Amsterdam, the Netherlands
k.v.hindriks@vu.nl

A.E. Eiben
Department of Computer Science
Vrije Universiteit Amsterdam
Amsterdam, the Netherlands
a.e.eiben@vu.nl

Kim Baraka
Department of Computer Science
Vrije Universiteit Amsterdam
Amsterdam, the Netherlands
k.baraka@vu.nl

***Abstract*— Learning from Demonstrations (LfD) transfers skills from human teachers to robots. However, data imbalance in demonstrations can bias policies towards majority situations. Previous work attempted to solve this problem after data collection, but few efforts were made to maintain a balanced distribution from the phase of data acquisition. Our method accounts for the influence of robots on human teachers and enables robots to actively guide interaction to approximate demonstration distributions to target distributions. Simulated and real-world experiments validated the method's efficacy in shaping demonstration distribution into various target distributions and robustness to various levels of uncertainties. Also, our method significantly improved the generalization ability of robot learning when LfD policies were trained with data collected by our method compared to natural data collection.**

***Index Terms*— learning from demonstration, data imbalance, data collection, human-robot interaction**

I. Introduction

In recent years, Learning from Demonstration (LfD) has become one of the most popular methods to equip robots with various skills [1], [2]. Previous work explored different avenues to provide demonstrations from human teachers (e.g., kinesthetic teaching [3]) for various types of tasks. However, the performance of such a data-driven method is intrinsically influenced by the quality of demonstration data. Among the challenges brought by data quality, data imbalance is one of the most prominent issues. It can easily result in a biased policy against minor situations when these demonstrations are imbalanced [4], [5]. What is even worse is that balanced demonstration data can hardly be naturally obtained in real-world scenarios [6]. In real-world applications, robots are often confronted with highly dynamic environments where it is infeasible to capture all possible situations in data collection with equal abundance. In the case of engagement recognition [7], [8], for example, disengaged situations are less often to be naturally collected, leading to an imbalanced dataset that is heavily biased towards high-engagement classes. Such an imbalance will bias the learning process and demand properly addressed for an unbiased policy.

To tackle data imbalance, previous work mainly solved it from the side of algorithm design, adding extra consideration in the cost function to unbias policy learning [4], [5], [8]. Some work [9]–[11] also alleviated the imbalance issue by implementing different re-sampling methods (e.g., undersampling [12] and oversampling [13]) on the original dataset. However, these efforts only attempted to curate the existing datasets after they are already collected. Few of them paid careful attention to data abundance during the data acquisition process and put efforts to maintain a balanced data distribution from the early phase. Instead of solving the real bottleneck brought from the data side, as indicated in [14], prior work tended to make the opposite efforts only from the algorithm side. Therefore, our work aimed to solve the data imbalance in the early phase of data acquisition and attempted to answer the following research questions:

- **Q1**: How to actively shape the distribution of demonstration data to maintain data balance during the data acquisition process?
- **Q2**: How does such active data collection benefit robot learning performance?

To answer question **Q1**, we explicitly took into account the influence of robots on human teacher behaviors and enabled the robot to actively guide its interaction with humans to shape the distribution of collected data. More specifically, we formalized such an active data collection process into a discrete finite-horizon Markov Decision Process (MDP) to maintain data balance against uncertainties during the data collection process. Results for the experiments of simulated data collection verified our method's generalization capability to actively shape the resulting distribution into various target distributions, along with its robustness to different

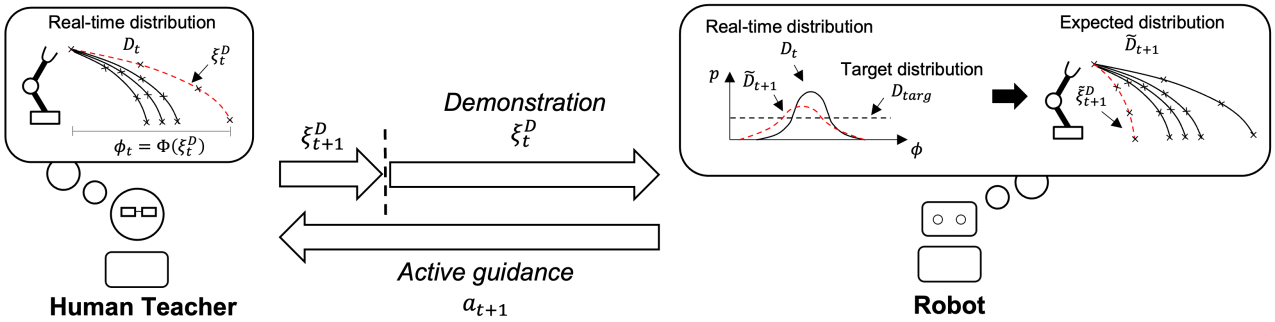


Fig. 1: Overview of our active data collection method. Given the latest demonstration from the human teacher, our method enables the robot to actively provide guidance to teachers (e.g., base movement) for the next demonstration, aiming to shape the distribution of collected demonstrations towards a given target distribution after all demonstrations are collected.

levels of uncertainties during the data collection process. To answer question **Q2**, we applied our method to real-world robot tasks (i.e., forward ball-throwing and backward ball-throwing) and trained LfD models with demonstration data that were actively gathered using our data collection method. Results validated the efficacy of our method to better maintain a more balanced distribution of demonstration data and indicated that it could improve robot generalization ability in unseen situations with a significant effect. For real-world tasks, we selected uniform distribution as the target distribution and showed that it benefited robot learning performance. However, this may not always be the optimal choice for target distributions to achieve improved learning performance. How to select the best target distribution for different tasks is out of the scope of this work.

To summarize, our work made contributions in the following aspects:

1) We presented an MDP-based active data collection method that could produce policies to actively shape data distribution in the phase of data acquisition.
2) We demonstrated via simulations that our trained data collection policies were of good generalization ability to actively shape collected demonstrations into various target distributions and of strong robustness to different levels of uncertainties during the data collection process.
3) We verified our method's efficacy in real-world tasks and demonstrated improved robot learning performance in unseen situations when models were trained with demonstrations of more balanced distributions shaped by our active data collection method.

II. Related Work

A. Data Imbalance in Data-Driven Learning

Data imbalance is a common problem for data-driven learning. It usually refers to the scenarios in classification problems where the ratio of different labels is highly uneven, leading to biased learning towards the majority class [7]. Prior work provided various methods, including data-centered methods and algorithm-centered methods [7], [15], [16]. For data-centered ones, common practices include different kinds of sampling strategies such as under-sampling [12] to decrease majority instances and over-sampling [13] to increase minority examples. Some works [17] also utilized feature selection techniques to solve the imbalance problem. By contrast, algorithm-centered methods attempted to solve the data imbalance by reconsidering the design of the learning objective. For instance, thresholding-based methods [18] solved the problem by adjusting the threshold for the classifier to distinguish between majority and minority classes. Cost-sensitive learning [19] reweighted the misclassification cost for different classes to balance the learning process. Some other methods also turned the multi-class problem into a one-class learning process and took the minorities as positive (or outlier) instances to overcome the imbalance among different labels [20]. Recent work also extended the scope from classification problems to regression problems [6]. However, all these methods attempted to solve the data imbalance after training data are already collected and seldomly tried to tackle it right from the data acquisition process. By contrast, our work actively monitors and shapes the distribution of training data during the data collection process to better benefit learning performance.

B. Data Collection for Learning from Demonstration

Previous works in LfD have utilized various avenues to collect demonstration data. In general, three categories of approaches are commonly employed [21]: *kinesthetic teaching*, *teleoperation*, and *passive observation*. Kinesthetic teaching [3] refers to the method where human teachers physically guide robots to demonstrate desired motions. It has been widely used especially for manipulation tasks to provide demonstrations in a more intuitive manner. Teleoperation [22] refers to the method where human teachers provide demonstrations via remote controlling (e.g., joysticks). In these cases, demonstrators do not have to be spatially present with the robot at the same time, much benefiting large-scale demonstration collection [21]. Passive observation [23] refers to the method where human teachers demonstrate how they complete the tasks without robots being involved in the execution process. Such a method makes it much easier for human teachers to provide demonstrations, but requires extra efforts in data curation and solving correspondence problems.

Combined with these three demonstration approaches, more advanced interfaces, either from the side of hardware [24] or software [25], can even further facilitate the human teaching experience and improve the quality of demonstration data. However, these previous works tend to place the robot in a passive position to receive demonstrations from human teachers and seldomly took the distribution of demonstration data into account. Furthermore, they barely explicitly investigated the potential benefits it might further bring for robot learning if proper guidance can be actively provided from robots to human teachers in the data acquisition process.

III. Methodology

The distribution of demonstration data tends to be imbalanced if they are naturally collected. Therefore, our method aims to solve the data imbalance issue in the phase of the data acquisition process. By modeling and utilizing the influence of robots on human teacher behaviors, our method produces a data collection policy for the robot to actively guide its interaction with humans, aiming to shape the distribution of collected demonstration data into any given target distribution as closely as possible.

A. Concepts

1) Demonstration Data: For a given task, we defined demonstration data $\Gamma = \{\xi_1^D, \xi_2^D, \ldots, \xi_{N_D}^D\}$, which includes N_D trajectories ξ_i^D. Each demonstration trajectory $\xi_i^D = \{(s_t^D, a_t^D)\}_{t=1}^L$ is a finite-horizon sequence of state-action pairs, where s_t^D and a_t^D are the state and action for the given task at step t, and L is the horizon length of the trajectory.

2) Feature Function: Given a high-dimensional trajectory ξ, we defined the low-dimensional feature variable $\phi = \Phi(\xi)$, where $\Phi(\cdot)$ is a predefined trajectory-based feature function. For example, in the task of robot ball-throwing, the feature function can be the distance between the landing position of the ball and its initial position given the robot's throwing movements. To simplify the problem, in this work we made several assumptions about feature function design:

- We assume that the feature variable ϕ is *representative* and *interpretable*. By *representative*, we mean that the selected feature is able to capture one important aspect of task dynamics and its distribution will closely impact that of demonstration data Γ. By *interpretable*, we mean that such a variable should have practical meaning (e.g., landing distance in the ball-throwing task) as opposed to some unexplainable latent variable.
- Without loss of generality, we only consider the case where ϕ is a one-dimensional continuous variable (i.e., $\phi \in \mathbb{R}$) and present a data collection method to actively shape the univariate distribution of demonstration data Γ along this feature dimension. However, the formulation of our method can be easily scaled to multi-dimensional cases.

B. Active Collection for Demonstration Data

Aiming to actively shape the distribution of demonstration data into any given target distribution, we formalized the task of active demonstration collection as a discrete-time Markov Decision Process (MDP) with a finite horizon.

1) States: We defined the state $s_t \in S$ as the discrete distribution result D_t for normalized feature values $\tilde{\phi}$ of already collected demonstrations (i.e., $\{\xi_1^D, \xi_2^D, \ldots, \xi_{t-1}^D\}$) before step t starts. Each normalized feature value $\tilde{\phi}$ was obtained by normalizing over the interval bin length l_{bin}, i.e., $\tilde{\phi} = \phi \,/\, l_{bin}$. More specifically,

$$s_t = D_t = \left\{ n_t^1,\, n_t^2,\, \ldots,\, n_t^{N_{intv}} \right\} \tag{1}$$

where n_t^i represents the number of already collected demonstrations before step t whose feature values fall into the range of the i-th interval. N_{intv} refers to the number of predefined intervals with which to characterize the resulting discrete distribution.

2) Actions: We defined the action a_t as the i_t-th predefined interval, which corresponds to the target range of the expected normalized feature value $\tilde{\phi}$ for the next demonstration ξ_t^D the robot aims to collect. Once the target interval is chosen, the robot will take corresponding low-level actions (e.g., verbal communication, base movement) to guide its interaction with human teachers, indicating its intention and consequently influencing human behaviors to increase the likelihood of obtaining the next target demonstration. Mathematically, the high-level action a_t can be expressed as:

$$a_t = i_t \;\in\; \{i_t \mid i_t \in N,\; 1 \le i_t \le N_{intv}\} \tag{2}$$

3) Transition model: After the robot observes the discrete distribution result (i.e., s_t) of collected demonstration data and selects the interval (i.e., a_t) for the normalized feature value of the next target data demonstration, s_t will transit to s_{t+1} by:

$$\begin{aligned} s_{t+1} &\sim P\left(s_{t+1} \,|\, s_t,\, a_t\right) \\ &= P\left(i_t^{real} \,\middle|\, s_t = D_t,\, a_t = i_t\right) \\ &= \frac{P\left(i_t^{real} \,\middle|\, a_t = i_t\right)}{\sum_{j=1}^{N_{intv}} P\left(i_t^{real} = j \,\middle|\, a_t = i_t\right)} \end{aligned} \tag{3}$$

where i_t^{real} represents the interval within the range of which the normalized feature value of the new demonstration actually falls. $P\left(i_t^{real} \,|\, a_t = i_t\right)$ was defined as:

$$\begin{aligned} &P\left(i_t^{real} = k \,\middle|\, a_t = i_t\right) \\ &= \int_{l_k}^{r_k} \mathcal{N}\left(\tilde{\phi}_t \,\middle|\, \frac{l_{i_t} + r_{i_t}}{2},\, \sigma_d^2\right) d\tilde{\phi}_t \end{aligned} \tag{4}$$

where we assumed the probability of the normalized feature value $\tilde{\phi}_t$ of the actually collected demonstration ξ_t^D follows a Gaussian distribution centered in the middle of the i_t-th interval with a predefined constant standard deviation σ_d. For convenience, we refer to σ_d as *transition standard deviation* thereafter. l_k and r_k represent the left and right boundary of the k-th interval. Similarly, l_{i_t} and r_{i_t} represent the left and right boundary of the i_t-th interval.

4) Reward: We defined the reward function $r(s_t)$ as:

$$r(s_t) = \begin{cases} \sum_{i=1}^{N_{intv}} \min\left(0,\, n^i_{targ} - n^i_t\right), \text{ if } t < T \\ \sum_{i=1}^{N_{intv}} -\left|n^i_{targ} - n^i_t\right|, \text{ otherwise.} \end{cases} \quad (5)$$

where $n^i_{targ} \in D_{targ}$ represents the target number of demonstrations whose normalized feature values fall within the range of i-th interval given a target discrete distribution D_{targ} for the whole demonstration dataset. T refers to the horizon length of one complete episode (i.e., the total number of demonstrations to be collected).

5) Policy: Given the reward function $r(s_t)$ and transition model $P\left(s_{t+1} \,|\, s_t,\, a_t\right)$, the goal is to obtain the optimal policy π^* that is able to achieve the maximum finite-horizon discounted reward.

IV. Experiments

To prove the efficacy of our method in shaping demonstration distribution and investigate its benefits for robot learning, we conducted experiments for both simulated and real-world data collection.

A. Experiments of Simulated Data Collection

1) Task Settings:

To simulate the data collection process, we generated a synthetic dataset Γ^{ϕ}_{syn} of 40 feature values ϕ_i, i.e., $\Gamma^{\phi}_{syn} = \{\phi_i \,|\, \phi_i \in \mathbb{R},\, 0 \leq \phi_i \leq 10\}^{40}_{i=1}$, whose distribution was discretized with 10 intervals of bin length 1. We used such a dataset as a pool of potential demonstrations available for collecting. We assumed the process of active data collection followed the transition dynamics we defined in (3).

2) Baselines:

Two methods were used to subsample from Γ^{ϕ}_{syn}: *natural data collection* and *active data collection*. For natural data collection, we randomly subsampled from Γ^{ϕ}_{syn} to simulate the natural data collection process. For our active data collection, we used pre-trained policies π^* for different target distributions to subsample from Γ^{ϕ}_{syn}. Given a target distribution, we set $\sigma_d = 0.5$ and obtained π^* via Proximal Policy Optimization (PPO) with a mini-batch size of 64, a learning rate η of 3×10^{-4}, a discount factor γ of 1, and trained for 4×10^3 episodes.

3) Procedures:

We conducted experiments for three different target distributions (i.e., uniform distribution, normal distribution, and bi-modal normal distribution). For each of them, we conducted 30 independent trials of simulated data collection using two baselines. In each trial, we used each method to subsample 20 datapoints from Γ^{ϕ}_{syn}. We compared data collection performance using *distribution deviation* defined as:

$$e_{dev} = \frac{1}{2T} \sum_{i=1}^{N_{intv}} \left|n^i_{targ} - n^i_T\right| \quad (6)$$

where n^i_T refers to the number of collected data points in the i-th interval after the whole episode of data collection completes (i.e., finishing the final step T).

To investigate our method's robustness to different levels of uncertainty, we also conducted experiments for sensitivity analysis to investigate the effects of transition standard deviation σ_d defined in (4) on the resultant data collection performance. After training policies π^* with different σ_d for the three target distributions, we conducted 30 independent trials of simulated data collection. In each trial, we used each of these policies to subsample 20 datapoints from Γ^{ϕ}_{syn}, and compared the data collection performance using disbribution deviation e_{dev}.

B. Experiments of Real-World Data Collection

1) Task Settings:

We designed two real-world tasks: *forward ball-throwing* and *backward ball-throwing*, as shown in Fig. 2. In these tasks, human teachers will kinesthetically teach the robot how to throw a ball towards given target areas that are either in front of it (i.e., *forward ball-throwing*) or behind it (i.e., *backward ball-throwing*). Given these demonstrations, the robot aims to learn a low-level control policy to throw the ball onto any given target area as closely as possible.

We used Behavioural Cloning (BC) [26] for control policies. For both tasks, we defined the low-level state s^D_t and action a^D_t as $(x^{rh}_t, y^{rh}_t, z^{rh}_t, x^{targ})$ and $(q^{rsp}_t, q^{rsr}_t, q^{rey}_t, q^{rer}_t, q^{rwy}_t)$ respectively. x^{rh}_t, y^{rh}_t, and z^{rh}_t are the 3D position of the robot's right hand in the robot base frame shown in Fig. 2. x^{targ} is the absolute distance between the given target area and the robot standing position along the x-axis. $q^{rsp}_t, q^{rsr}_t, q^{rey}_t, q^{rer}_t$, and q^{rwy}_t are the target joint values of the robot's right shoulder pitch, right shoulder roll, right elbow yaw, right elbow roll, and right wrist yaw. We set the episode length L as 30, i.e., $\xi^D_i = \{(s^D_t, a^D_t)\}^{30}_{t=1}$.

For the high-level data collection tasks, we chose the feature function $\Phi(\cdot)$ as the absolute distance between the landing position and the robot's standing position along the x-axis. The real-time distribution D_t was discretized with 6 intervals of bin length 0.2m. We labeled on the floor a valid landing area to limit ϕ to the range of $[0.3\text{m}, 1.5\text{m}]$.

2) Baselines:

We employed two different methods to collect demonstrations: *natural data collection* and *active data collection*. For natural data collection, the robot stayed at a fixed position and human subjects physically guided the robot's right arm to provide demonstrations to throw the ball into the *valid landing area* (shown in Fig. 2).

For our active data collection, the robot actively moved its base back or forth relative to a fixed target landing area (shown in Fig. 2) after each demonstration was provided by human subjects who physically guided the robot arm to throw the ball aiming at the *target landing area*. The robot moving distances corresponded to a_t and were determined by the policy π^* trained as that in section IV-A, with uniform distribution as the target distribution and $\sigma_d = 0.5$.

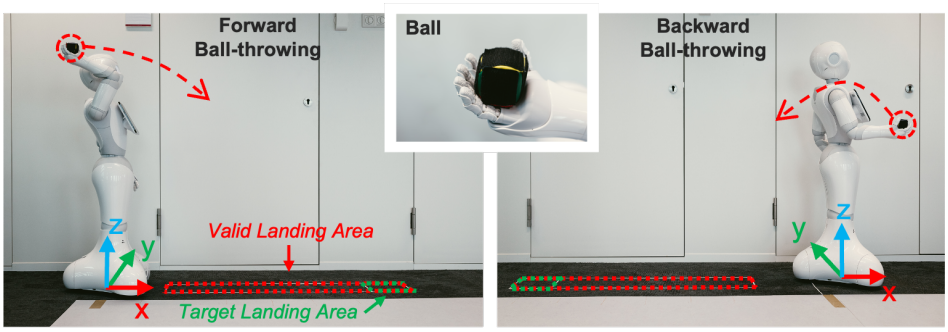


Fig. 2: Hardware and task settings for real-world experiments.

3) Hardware:

We used the humanoid robot Pepper of the SoftBank Robotics company for experiments. The ball was with a diameter of 5 cm, a weight of 270g, and covered with hook-and-loop fasteners (shown in Fig. 2) to easily attach itself to the fabric floor against further slipping once landing. We manually measured the ball landing distance after each demonstration finished.

4) Participants:

Following the ethical guidelines provided by our faculty's research ethics board, we recruited 17 human subjects (9 male, 7 female, 1 other gender) from campus via poster advertisement. The ages of participants were distributed as follows: 13 between 18-29, 3 between 30-39, and 1 between 40-49. Coding experience was categorized as none (2), some (4), and extensive (11). The ball-throwing experience was categorized as none (1), some (13), and extensive (3). Experience in movement activities was categorized as none (2), some (11), and extensive (4). Participants were compensated with a €10 digital gift card after the experiments finished.

5) Procedures:

We conducted within-subject experiments of data collection for the two ball-throwing tasks. Under each condition of the data collection methods, human subjects provided 12 demonstrations for each task, resulting in two sets of demonstrations Γ^k_{nat} and Γ^k_{act} for human subject k. The condition order was counter-balanced. Before the experiments of active data collection, we informed human participants of the potential movements of the robot base after each of their demonstrations was provided, aiming to alleviate the impact of unexpected surprise on human demonstrations. Prior to providing demonstrations for each task for the first time, human subjects went through a training session to get familiar with physically controlling the robot. It finished once they succeeded to throw the ball into the valid landing area successively for 3 times, or it reached the time limit of 5 minutes. To evaluate data collection performance, we respectively calculated the *distribution deviation* e_{dev} for the two datasets (i.e., Γ^k_{nat} and Γ^k_{act}) of each human subject. All demonstrations and video footage are available at https://github.com/MH-Hou/active-data-collection.git.

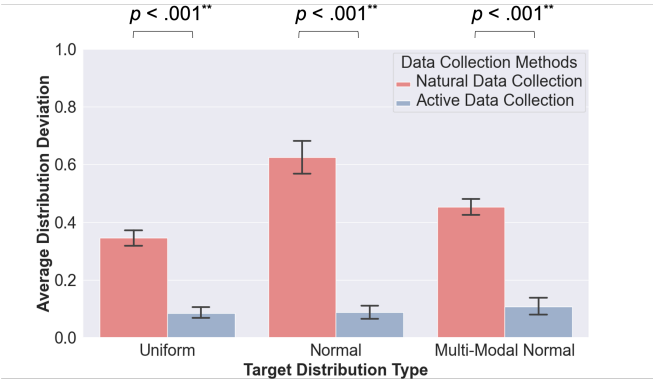
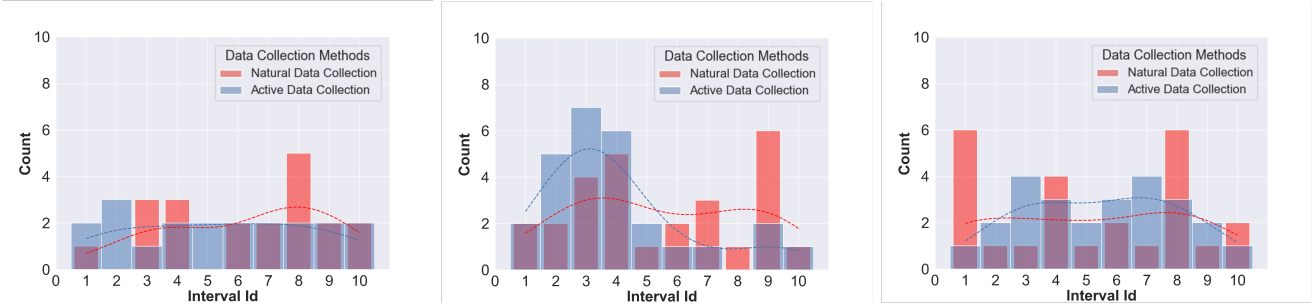


Fig. 3: Results for average distribution deviation in simulated data collection processes.

After the experiments under the first condition finished, human subjects filled out a questionnaire to evaluate the importance of the *diversity of landing distances* in demonstration data for robots to master ball-throwing. The scores were given via a 5-point Likert Scale where 1 represents "not important at all" and 5 represents "extremely important". To avoid bias, we also asked about the importance of the *robot arm average velocity* and *robot hand final stopping position* and randomly set the question order.

For each ball-throwing task, we trained two BC models with demonstrations collected respectively using two baseline methods. Each BC model consisted of a 4-dimensional input layer, two fully connected layers with 64 units each, and a 4-dimensional output layer. We used Mean Square Error (MSE) as the loss function and trained the BC model for 300 epochs with a learning rate of 0.001 and a mini-batch size of 64.

We tested the performance of trained BC models on two sets of distinct target areas, whose distances to the fixed robot position were uniformly distributed across the range of $[0.3\text{m}, 2.0\text{m}]$ and $[1.55\text{m}, 2.0\text{m}]$. The first set consisted of 18 areas and was used to evaluate the overall performance. The second one consisted of 19 areas and was used to evaluate the generalization ability of BC models in completely unseen situations. We evaluated the model performance with landing error e_{dist} defined as $|x^{targ} - x^{real}|$, where x^{real} is the real landing distance from the robot position along the x-axis.



(a) uniform target distribution (b) normal target distribution (c) bi-modal normal target distribution

Fig. 4: Distribution plots for simulated data collection in one random trial using various types of target distributions.

V. Results

A. Simulated Data Collection

1) Results for approximating various types of target distributions:

For the experiments of each target distribution, we conducted a paired samples t-test to determine the effect of the type of data collection methods on distribution deviation e_{dev}, the results of which are shown in Fig. 3. With uniform distribution as the target distribution, there was a significant difference in e_{dev} between natural data collection ($M = 0.345, SD = 0.061$) and active data collection ($M = 0.085, SD = 0.042$); $t(19) = 14.75,\ p < .001$ with a large effect size (Cohen's $d = 4.83$). With the normal distribution as the target distribution, there was also a significant difference in e_{dev} between natural data collection ($M = 0.625, SD = 0.126$) and active data collection ($M = 0.0875, SD = 0.052$); $t(19) = 17.24,\ p < .001$ with a large effect size (Cohen's $d = 5.43$). With the bi-modal normal distribution as the target distribution, we also observed a significant difference in e_{dev} between natural data collection ($M = 0.452, SD = 0.064$) and active data collection ($M = 0.1075, SD = 0.064$); $t(19) = 15.06,\ p < .001$ with a large effect size (Cohen's $d = 5.25$). We also visualized the distribution result of a randomly selected trial during the experiment, shown in Fig. 4. Consistent with the results of t-tests, our active data collection method yielded a dataset of normalized feature values that more closely approximated various target distributions, in contrast to that collected by the natural data collection method.

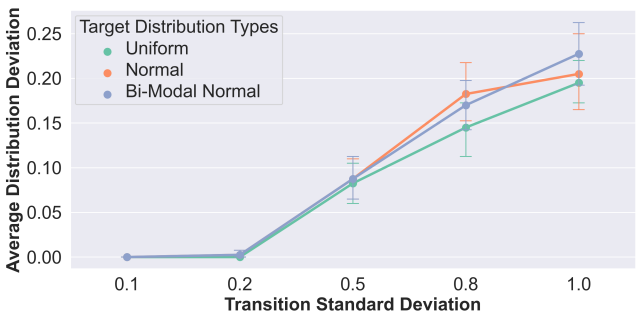


Fig. 5: Results for sensitivity analysis under various transition uncertainties in simulated data collection processes.

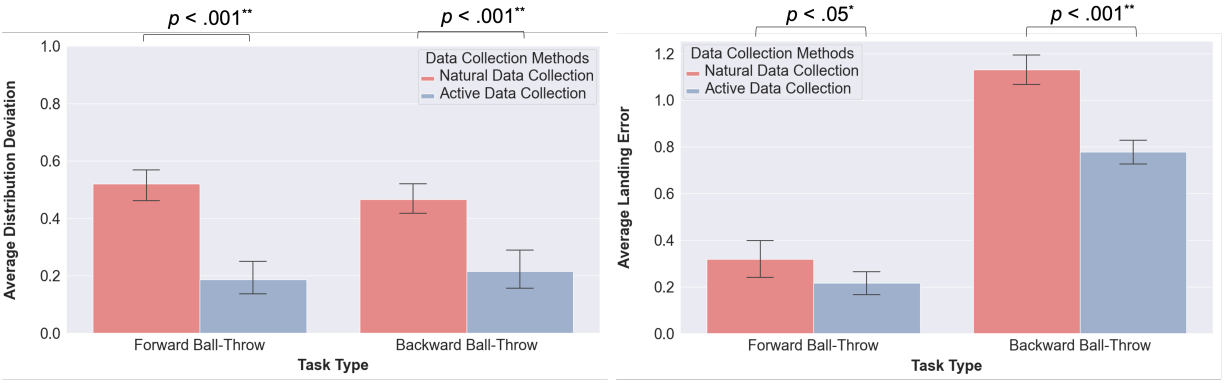
2) Results for sensitivity analysis:

As shown in Fig. 5, with various target distributions, the average distribution deviation all tended to increase as the transition standard deviation became larger. More specifically, such an increase remained relatively minimal when the transition standard deviation σ_d stepped from 0.1 to 0.2, where the average distribution deviation was quite close to 0. However, the average distribution deviation surged to much higher values when σ_d increased beyond 0.5, arriving at around 0.08. When σ_d increased to around 1.0, the average distribution deviation also reached its highest value of around 0.2 for each type of target distributions. Nevertheless, compared with the average distribution deviations of natural data collection shown in Fig. 3, active data collection still achieved much smaller average distribution deviations, even when σ_d arrived at 1.0 under each condition of target distributions.

B. Real-World Data Collection

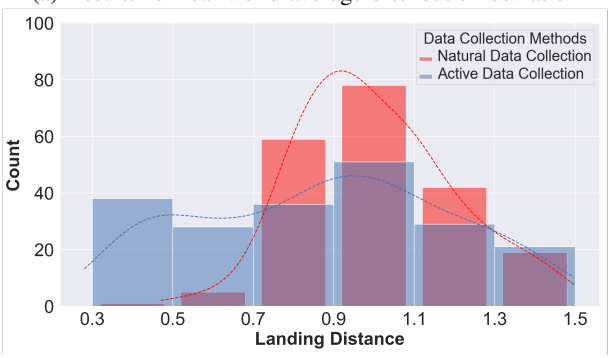
1) Results for Balancing Data Distribution:

For both ball-throwing tasks, we performed a paired samples t-test to determine the effect of the type of data collection methods on distribution deviation e_{dev}, the results of which are shown in Fig. 6a. For the task of forward ball-throwing, we observed a significant difference in distribution deviation between natural data collection ($M = 0.520, SD = 0.116$) and active data collection ($M = 0.186, SD = 0.120$); $t(16) = 10.43,\ p < .001$ with a large effect size (Cohen's $d = 2.74$). For the task of backward ball-throwing, there was also a significant difference in distribution deviation between natural data collection ($M = 0.466, SD = 0.115$) and active data collection ($M = 0.216, SD = 0.131$); $t(16) = 5.46,\ p < .001$ with a large effect size (Cohen's $d = 1.97$). We also visualized the distributions for feature values (i.e., landing distance) of all demonstrations collected respectively by active data collection method and natural data collection method, shown in Fig. 6c and Fig. 6d. Consistent with the results of the t-test, feature values of demonstrations collected by our active data collection method more closely approximated the target uniform distribution, resulting in a more balanced set of

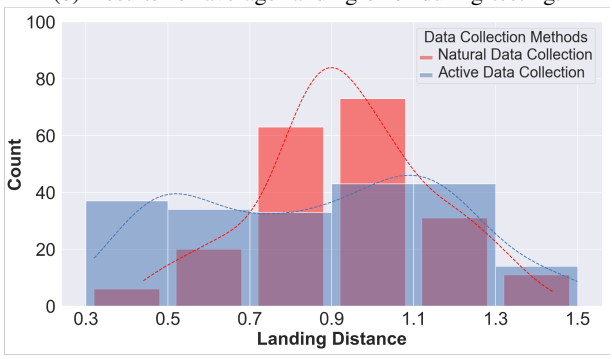


(a) Results for real-world average distribution deviation

(b) Results for average landing error during testing.



(c) Distribution plot for forward ball-throw



(d) Distribution plot for backward ball-throw

Fig. 6: Results for experiments of real-world data collection. (b) reveals the testing performance of BC models that were trained with various data collection methods on completely unseen situations. (c) and (d) visualize the distribution of demonstrations from all human subjects.

demonstration data as compared to the natural data collection method.

Regarding the results of the questionnaire, we selected 16 out of 17 questionnaires (8 for natural data collection and 8 for active data collection). We conducted an independent samples t-test to investigate the influence of the effect of the type of data collection methods on human subjective evaluation of the importance of the diversity in landing distance. We observed a significant difference in subjective importance scores between natural data collection ($M = 2.875, SD = 1.269$) and active data collection ($M = 4.125, SD = 0.781$); $t(14) = -2.220$, $p < .05$ with a large effect size (Cohen's $d = 1.11$).

2) Results for Robot Learning Performance:

For each ball-throwing task, we performed a paired samples t-test to determine the effect of the type of data collection methods on landing error e_{dist} when testing in completely unseen situations. The results of these analyses are shown in Fig. 6b. For the task of forward ball-throwing, there was a significant difference in landing error between natural data collection ($M = 0.319m, SD = 0.169m$) and active data collection ($M = 0.216m, SD = 0.103m$); $t(18) = 3.127$, $p < .05$ with a medium effect size (Cohen's $d = 0.72$). For the task of backward ball-throwing, we also observed a significant difference in landing error between natural data collection ($M = 1.131m, SD = 0.145m$) and active data collection ($M = 0.778m, SD = 0.111m$); $t(18) = 24.36$, $p < .001$ with a large effect size (Cohen's $d = 2.65$).

Similarly, we also conducted a paired samples t-test to investigate the effect of the type of data collection methods on landing error e_{dist} when testing on target areas ranging between $[0.3\text{m}, 2.0\text{m}]$. For the task of forward ball-throwing, we observed no significant difference in landing error between natural data collection ($M = 0.611m, SD = 0.199m$) and active data collection ($M = 0.556m, SD = 0.263m$); $t(17) = 1.199$, $p = 0.247$. By contrast, for the task of backward ball-throwing, we observed a significant difference in landing error between natural data collection ($M = 0.579m, SD = 0.406m$) and active data collection ($M = 0.445m, SD = 0.217m$); $t(17) = 2.430$, $p < .05$ with a small effect size (Cohen's $d = 0.40$).

VI. Discussion

The results of simulated data collection experiments confirm the efficacy of our method to significantly better shape demonstration distribution into various types of target distributions, as compared with natural data collection. Furthermore, the follow-up sensitivity analysis proves that our method significantly outperformed the baseline to approximate different types of target distributions under various levels of data collection uncertainties, suggesting its strong

robustness for application.

Results of real-world experiments suggest that our method better enabled the robot to obtain a more balanced distribution of demonstrations than natural data collection. Also, the results of the questionnaires reveal that our method was able to better convey robot intentions and expectations to human teachers. Furthermore, we observed benefits in robot learning, with the models trained on more balanced demonstration data collected by our method showing significantly better generalization ability to unseen situations. However, the improvements in overall robot learning performance were task-dependent and only confirmed in the task of backward ball-throwing.

Our approach does not come without limitations. First, the transition model of the data collection process was assumed to follow a normal distribution, with the uncertainty of data collection only reflected in the choice of the variance. In reality, the mean might also be shifted and the choice of variance could be different for each human teacher. Second, our work only considered the cases of the scalar feature. We leave it as future work to scale our approach to multivariate distributions. Finally, when applying our method to real-world tasks, we chose the feature function based on observation and heuristics. It might help to employ feature engineering techniques (e.g., feature selection) to produce better designs of feature functions and hopefully further improve the robot learning performance.

VII. Conclusions

We presented an active data collection method that shapes the distribution of demonstration data to given target distributions, taking into account the influence of robots on human teacher behaviors. Simulated experiments validated the method's efficacy in distribution shaping and robustness to different levels of data collection uncertainties. Results of real-world tasks (i.e., forward and backward ball-throwing) further showed significant improvements in balancing demonstration data distribution and conveying the robot's intention to human teachers. When trained with demonstrations collected by our method, the robot control policies significantly outperformed those trained with naturally collected demonstrations in both tasks. We leave it as future work to scale our methods to cases of multivariate distributions and integrate our method with advanced LfD algorithms to enhance robot learning performance.

References

[1] A. Billard, S. Calinon, R. Dillmann, and S. Schaal, "Robot programming by demonstration," in *Springer handbook of robotics*, pp. 1371–1394, Springer, 2008.

[2] S. R. Ahmadzadeh, R. Kaushik, and S. Chernova, "Trajectory learning from demonstration with canal surfaces: A parameter-free approach," in *2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)*, pp. 544–549, IEEE, 2016.

[3] S. Elliott, Z. Xu, and M. Cakmak, "Learning generalizable surface cleaning actions from demonstration," in *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pp. 993–999, IEEE, 2017.

[4] B. César-Tondreau, G. Warnell, K. Kochersberger, and N. R. Waytowich, "Towards fully autonomous negative obstacle traversal via imitation learning based control," *Robotics*, vol. 11, no. 4, p. 67, 2022.

[5] W. Zhou, S. Bohez, J. Humplik, N. Heess, A. Abdolmaleki, D. Rao, M. Wulfmeier, and T. Haarnoja, "Forgetting and imbalance in robot lifelong learning with off-policy data," in *Conference on Lifelong Learning Agents*, pp. 294–309, PMLR, 2022.

[6] Y. Yang, K. Zha, Y. Chen, H. Wang, and D. Katabi, "Delving into deep imbalanced regression," in *International Conference on Machine Learning*, pp. 11842–11851, PMLR, 2021.

[7] J. L. Leevy, T. M. Khoshgoftaar, R. A. Bauder, and N. Seliya, "A survey on addressing high-class imbalance in big data," *Journal of Big Data*, vol. 5, no. 1, pp. 1–30, 2018.

[8] D. Dresvyanskiy, W. Minker, and A. Karpov, "Deep learning based engagement recognition in highly imbalanced data," in *Speech and Computer: 23rd International Conference, SPECOM 2021, St. Petersburg, Russia, September 27–30, 2021, Proceedings 23*, pp. 166–178, Springer, 2021.

[9] G. Sawadwuthikul, T. Tothong, T. Lodkaew, P. Soisudarat, S. Nutanong, P. Manoonpong, and N. Dilokthanakul, "Visual goal human-robot communication framework with few-shot learning: a case study in robot waiter system," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 3, pp. 1883–1891, 2021.

[10] A. Amini, W. Schwarting, G. Rosman, B. Araki, S. Karaman, and D. Rus, "Variational autoencoder for end-to-end control of autonomous driving with novelty detection and training de-biasing," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 568–575, IEEE, 2018.

[11] M. J. Procopio, J. Mulligan, and G. Grudic, "Coping with imbalanced training data for improved terrain prediction in autonomous outdoor robot navigation," in *2010 IEEE International Conference on Robotics and Automation*, pp. 518–525, IEEE, 2010.

[12] S. Kotsiantis and P. Pintelas, "Mixture of expert agents for handling imbalanced data sets," *Annals of Mathematics, Computing & Teleinformatics*, vol. 1, no. 1, pp. 46–55, 2003.

[13] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[14] M. Stonebraker and E. K. Rezig, "Machine learning and big data: What is important?," *IEEE Data Eng. Bull.*, vol. 42, no. 4, pp. 3–7, 2019.

[15] S. Kotsiantis, D. Kanellopoulos, P. Pintelas, *et al.*, "Handling imbalanced datasets: A review," *GESTS international transactions on computer science and engineering*, vol. 30, no. 1, pp. 25–36, 2006.

[16] R. Longadge and S. Dongre, "Class imbalance problem in data mining review," *arXiv preprint arXiv:1305.1707*, 2013.

[17] Z. Zheng, X. Wu, and R. Srihari, "Feature selection for text categorization on imbalanced data," *ACM Sigkdd Explorations Newsletter*, vol. 6, no. 1, pp. 80–89, 2004.

[18] G. Weiss, "Mining with rarity: a unifying framework, sigkdd explorations 6 (1): 7–19," *Special issue on learning from imbalanced datasets*, 2004.

[19] P. Domingos, "Metacost: A general method for making classifiers cost-sensitive," in *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 155–164, 1999.

[20] A. Ali, S. M. Shamsuddin, and A. L. Ralescu, "Classification with class imbalance problem," *Int. J. Advance Soft Compu. Appl*, vol. 5, no. 3, 2013.

[21] H. Ravichandar, A. S. Polydoros, S. Chernova, and A. Billard, "Recent advances in robot learning from demonstration," *Annual review of control, robotics, and autonomous systems*, vol. 3, pp. 297–330, 2020.

[22] P. Abbeel, A. Coates, and A. Y. Ng, "Autonomous helicopter aerobatics through apprenticeship learning," *The International Journal of Robotics Research*, vol. 29, no. 13, pp. 1608–1639, 2010.

[23] B. Hayes and B. Scassellati, "Discovering task constraints through observation and active learning," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 4442–4449, IEEE, 2014.

[24] T. Cerqueira, F. M. Ribeiro, V. H. Pinto, J. Lima, and G. Gonçalves, "Glove prototype for feature extraction applied to learning by demonstration purposes," *Applied Sciences*, vol. 12, no. 21, p. 10752, 2022.

[25] C. Crick, S. Osentoski, G. Jay, and O. C. Jenkins, "Human and robot perception in large-scale learning from demonstration," in *Proceedings of the 6th international conference on Human-robot interaction*, pp. 339–346, 2011.

[26] M. Bain and C. Sammut, "A framework for behavioural cloning.," in *Machine Intelligence 15*, pp. 103–129, 1995.