Implicit Interactive Fleet Learning from Heterogeneous Human Supervisors

Gaurav Datta*[1], Anrui Gu*[1], Ryan Hoque*[1], Cesar Salcedo[1,2], Ken Goldberg[1]

***Abstract*— Learning from human demonstrations has been successfully applied to the automation of a range of robotic tasks, but can struggle when (1) robots sometimes encounter edge cases that are not represented in the training data or (2) the human demonstrations are multimodal (e.g., choosing different paths around an obstacle). Interactive fleet learning (IFL) increases reliability by allowing robots to fall back on remote human teleoperators during task execution and learn from them over time. While IFL mitigates the issue of edge cases, IFL often encounters multimodality as the humans may provide demonstrations in different ways. Human control policies are heterogeneous, noisy, multimodal, mixed quality, and non-Markovian. Recent work proposes Implicit Behavior Cloning, which models the robot policy *implicitly* rather than *explicitly* to represent multimodal demonstrations using energy functions. In this work, we propose a new algorithm and study how implicit control policies can mitigate the adverse effects of multimodality in IFL. We present Implicit Interactive Fleet Learning, the first extension of implicit behavior cloning to interactive imitation learning (including the single-robot, single-human setting). We also propose a novel metric for uncertainty quantification in energy-based models using Jeffreys divergence. Results suggest todo: insert results. See todo: website link for code and supplemental materials.**

I. Introduction

Imitation learning (IL), the paradigm of learning from human demonstrations and feedback, is a leading technique for efficiently automating diverse tasks such as autonomous driving [8, 31, 33], robot-assisted surgery [23, 32], and deformable object manipulation [4, 19, 36]. The most common for IL is behavior cloning (BC) [33], where the robot policy is derived via supervised machine learning on an offline set of human task demonstrations. Since BC can suffer from distribution shift between the states visited by the robot and those visited by the human, interactive IL algorithms such as DAgger [34] and variants [16, 22, 28] iteratively improve the robot policy with corrective human interventions during robot task execution. However, all of these IL algorithms are only reliable if the data is generated by a consistent human control policy.

In reality, human demonstrations and interventions are noisy, multimodal, suboptimal, mixed quality, non-Markovian, and nonstationary [27, 29]. A human providing demonstrations of a task may make mistakes, become more or less proficient at the task over time, or execute one of multiple equally valid actions when encountering the same state at different times (e.g., translating the end effector in an arbitrary trajectory through free space on the way to a target object). This issue is exacerbated when learning from a pool of multiple humans (e.g., for interactively training a fleet of robots [18]), since the humans have varying proficiency and may demonstrate the same task in different ways.

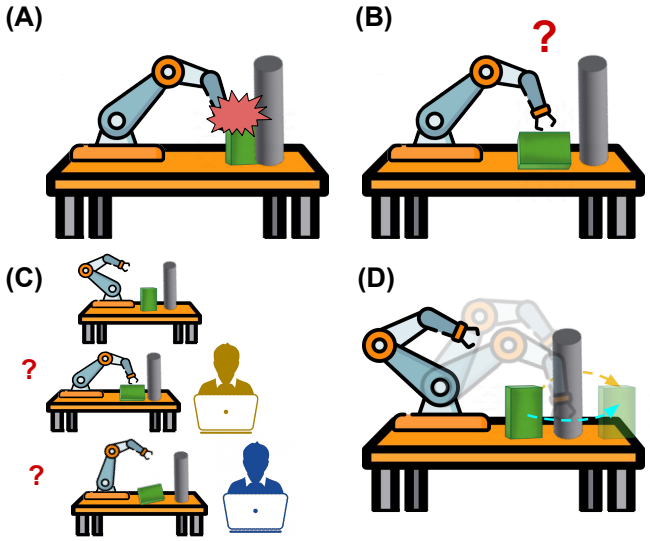

Fig. 1: Consider an autonomous robot pushing a block around a pillar. **(A) Multimodality:** Traditional imitation learning averages the two paths around the obstacle, leading to collision. **(B) Distribution Shift:** Compounding errors can lead the robot to visit unfamiliar states (e.g., knocking the block over on its side). **(C) Interactive Fleet Learning (IFL):** Prior work proposes a paradigm for correcting distribution shift in a robot fleet, but humans may operate the robots in different ways. **(D) Implicit IFL:** We propose a new algorithm that uses "implicit" energy-based models (EBMs) both for representing multiple modes and estimating uncertainty.

One recently proposed approach for IL with multimodal human data is Implicit BC [11]. Florence *et al.* propose training an energy-based model (EBM) [24] that represents state-action mappings *implicitly* rather than explicitly. While this complicates model training and inference, implicit models can flexibly represent multiple action modes for each state. However, Implicit BC inherits the distribution shift problem from Explicit BC.

We extend Implicit BC to the interactive fleet learning (IFL) setting [18], a generalization of interactive imitation learning to multiple robots and multiple humans. Since existing state-of-the-art IFL algorithms rely on estimates of epistemic uncertainty like the output variance among an ensemble of networks, which are incompatible with implicit

[1]AUTOLab at the University of California, Berkeley
[2]University of Engineering and Technology (UTEC), Peru
*Equal contribution

models (Section IV-C), we propose a new measure of uncertainty in energy-based models using Jeffreys divergence [20].

This paper makes the following contributions: (1) Implicit Interactive Fleet Learning (Implicit IFL), the first application of implicit policies to interactive imitation learning, (2) a novel uncertainty metric and supervisor allocation algorithm using Jeffreys divergence, (3) simulation experiments with a fleet of 10 robots and 3 heterogeneous algorithmic supervisors, (4) physical experiments with a fleet of 4 robots and 2 heterogeneous human supervisors.

II. Preliminaries and Related Work

A. Imitation Learning

Learning from an offline set of human task demonstrations is an intuitive and effective way to train a robot control policy [2, 3]. Popular approaches include behavior cloning (i.e., supervised learning) [31, 33, 36] and inverse reinforcement learning [1, 3, 6], which first infers a reward function from demonstrations and then trains a reinforcement learning agent with this reward. These demonstrations can be augmented with additional offline information such as pairwise preferences [7] and natural language [43]. Ho *et al.* [14] propose an alternative to inverse reinforcement learning using techniques from training generative adversarial networks [13], and Torabi *et al.* [42] extend this to imitation from observation, where states are available but action labels are not. However, imitation learning from offline demonstration data can suffer from distributional shift [34], as compounding approximation error leads the robot to visit states that were not visited by the human.

B. Interactive Imitation Learning

To mitigate distribution shift, Ross *et al.* [34] propose dataset aggregation (DAgger), which collects online action labels on states visited by the robot during task execution and iteratively improves the robot policy. Since DAgger can request excessive queries to a human supervisor, interactive IL [16, 22, 47] is a variant of DAgger that seeks to reduce human burden by intermittently ceding control to the human during robot execution based on some switching criteria. Human-gated interactive IL [22, 25, 40] has the human decide when to take and cede control, while robot-gated interactive IL [16, 17, 28, 47] has the robot autonomously decide. Hoque *et al.* [18] propose Interactive Fleet Learning (IFL), which generalizes robot-gated interactive IL to multiple robots supervised by multiple humans. In this work, we consider the IFL setting.

Sun *et al.* [41] propose a method for interactive imitation learning from heterogeneous experts, but their method is designed for autonomous driving applications. Gandhi *et al.* [12] also interactively learn from multiple experts and propose actively soliciting the human supervisors to provide demonstrations that are compatible with the current data. However, this prevents the robot from learning alternative modes and requires the human supervisors to cooperate with the suggested actions, which may not be the case due to human suboptimality, fatigue, or obstinacy [10].

C. Robot Learning from Multimodal Data

Learning from multimodal examples is an active challenge in machine learning and robotics. A mixture density network [5] is a popular approach that fits a (typically Gaussian) mixture model to the data, but it requires setting a parameter for how many modes to fit, which may not be known a priori. When actions can be represented as pixels in an image (e.g., pick points), a Fully Convolutional Network [38] can be applied to learning pixelwise multimodality [19, 46]. In a very recent paper, Chi *et al.* [9] introduce diffusion policies, an application of diffusion models [15] to imitation learning from multimodal data. Shafiullah *et al.* [37] propose Behavior Transformers, a technique that applies the multi-token prediction of Transformer neural networks [44] to imitation learning. Other Transformer-based policies report similar benefits for multimodal data [21, 39]; however, these approaches require action discretization to cast behavior prediction as next-token prediction.

Florence *et al.* [11] propose Implicit BC, a technique that trains a conditional energy-based model [24] that is found to outperform explicit BC and mixture density networks in their experiments. As opposed to explicit models that take the form $\pi : S \rightarrow A$, implicit models take the form of a multimodal function $E : S \times A \rightarrow \mathbb{R}$; the action is an input rather than an output of the model. To sample an action from the policy, instead of evaluating the explicit model $\hat{a} = \pi(s)$, the implicit model must perform optimization over E conditioned on state s:

$$\hat{a} = \arg\min_{a} E(s, a) \tag{1}$$

In this work, we extend Implicit BC to interactive fleet learning in order to both mitigate the distributional shift problem that BC faces and the multimodality in human teleoperation that IFL faces. To our knowledge, we are the first to extend implicit policies to interactive imitation learning.

III. Problem Statement

We consider the interactive fleet learning (IFL) problem setting proposed by Hoque *et al.* [18]. A fleet of N robots operate in parallel independent Markov Decision Processes (MDPs) that are identical apart from their initial state distributions. The robots can query a set of $M < N$ human supervisors with action space $\mathcal{A}_H = \mathcal{A} \cup \{R\}$, where $a \in \mathcal{A}$ is teleoperation in the action space of the robots and R is a "hard reset" that physically resets a robot in a failure state (e.g., a delivery robot tipped over on its side). As in [18], we assume that (1) the robots share policy $\pi_{\theta_t} : \mathcal{S} \rightarrow \mathcal{A}$, (2) the MDP timesteps are synchronous across robots, and (3) each human can only help one robot at a time. However, unlike the original IFL formulation [18], in this work we do *not* assume that the human supervisors are homogeneous; each human i has a unique policy $\pi_H^i : \mathcal{S} \rightarrow \mathcal{A}_H$. Furthermore,

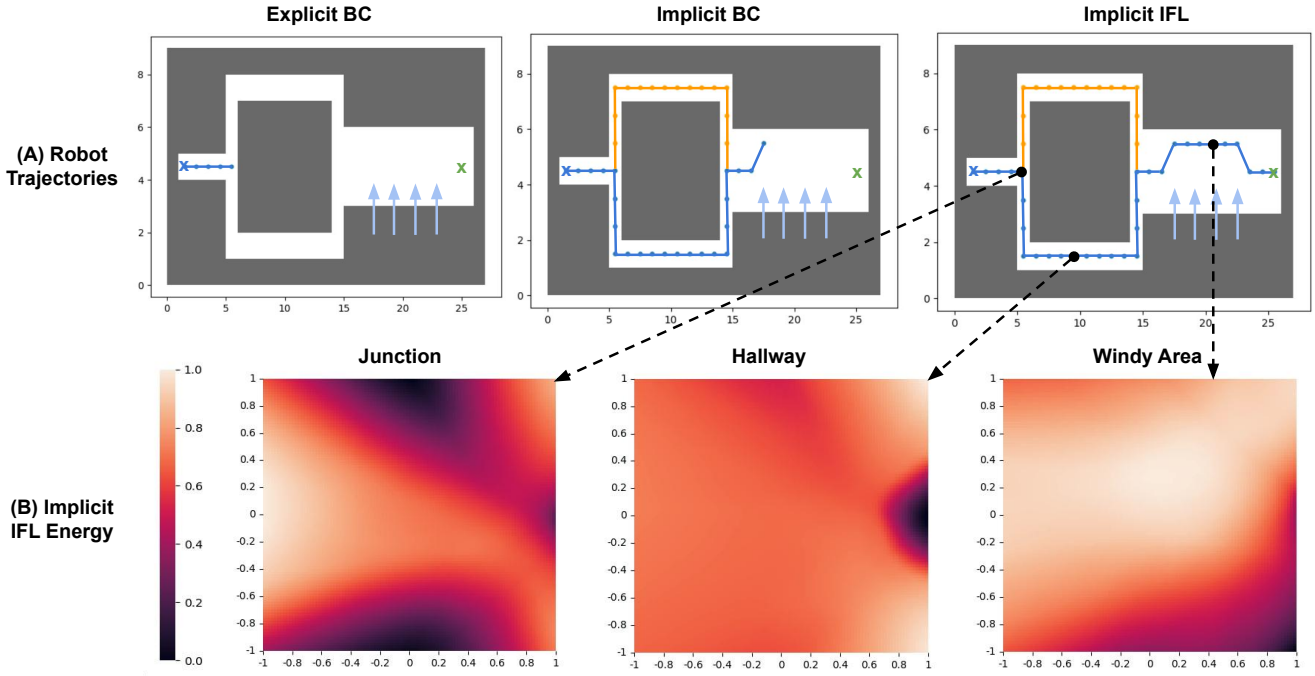

Fig. 2: In the 2D navigation experiments from Section V-A, the robot must navigate from the blue X marker to the green X marker. **(A) Robot Trajectories:** Explicit BC cannot make progress past the fork due to multimodal demonstrations, while Implicit BC cannot overcome the distribution shift due to wind in the $+y$ direction at execution time (denoted in light blue). Implicit IFL reaches the goal by handling both multimodality and distribution shift. **(B) Implicit IFL Energy:** We display normalized Implicit IFL energy distributions from representative states in the trajectory. Lower energy indicates a more optimal action, and the x and y axes are the 2D action deltas $\hat{a}$ that the robot can execute (which can be mapped directly onto the corresponding 1×1 cell in the maze). At the junction point, both upward and downward actions attain low energy; in a straight hallway, the rightmost actions attain low energy; in the windy area, actions toward the lower right corner (making progress toward the goal while fighting the wind) attain low energy.

each π_H^i may itself be nondeterministic and multimodal, but is assumed to be optimal or nearly optimal.

An IFL supervisor allocation algorithm is a policy ω that determines the assignment of humans to robots:

$$\omega : (\mathbf{s}^t, \pi_{\theta_t}, \cdot) \mapsto \boldsymbol{\alpha}^t \in \{0,1\}^{N\times M}$$
$$\text{s.t.} \quad \sum_{j=1}^{M} \boldsymbol{\alpha}_{ij}^t \leq 1 \text{ and } \sum_{i=1}^{N} \boldsymbol{\alpha}_{ij}^t \leq 1 \quad \forall i, j. \tag{2}$$

Here, $\mathbf{s}^t$ are the current states of each of the robots, $\boldsymbol{\alpha}^t$ is an $N \times M$ binary matrix that indicates which robot will receive assistance from which human at the current timestep t, and π_{θ_t} is the shared robot control policy at time t.

The IFL objective is to find an ω that maximizes the return on human effort (ROHE):

$$\max_{\omega\in\Omega} \mathbb{E}_{\tau\sim p_{\omega,\theta_0}(\tau)} \left[\frac{M}{N} \cdot \frac{\sum_{t=0}^{T} \bar{r}(\mathbf{s}^t, \mathbf{a}^t)}{1 + \sum_{t=0}^{T} \|\omega(\mathbf{s}^t, \pi_{\theta_t}, \boldsymbol{\alpha}^{t-1}, \mathbf{x}^t)\|_F^2} \right]$$

where $\|\cdot\|_F$ is the Frobenius norm, T is the amount of time the fleet operates (rather than an individual episode horizon), and θ_0 are the initial parameters of π_{θ_t}. The ROHE measures the average performance of the robot fleet normalized by the amount of human effort required to help the robots [18].

IV. Approach

A. Preliminaries: Implicit Models

We build on Implicit Behavior Cloning [11] in this work. Implicit BC seeks to learn a conditional energy-based model $E : S \times A \to \mathbb{R}$, where $E(s, a)$ is the scalar "energy" for action a conditioned on state s. Lower energy indicates a higher correspondence between s and a. The energy function defines a multimodal probability distribution π of action a conditioned on state s:

$$\pi(a|s) = \frac{e^{-E(s,a)}}{Z(s)} \tag{3}$$

where $Z(s)$ is a normalization factor known as the "partition function." Equivalently, the energy is the unnormalized negative log-probability of the action given the state. Intuitively, such a model handles multimodality by assigning low energy to multiple actions $a_1, a_2, \ldots$ for the same state s, rather than explicitly mapping s to a single action a. In practice, we estimate E with a learned neural network function approximator E_θ parameterized by θ and train E_θ on samples $\{s_i, a_i\}$ collected from the expert policies π_H. Given a set of

counter-examples $\{\tilde{a}_i^j\}$ for each s_i, Implicit BC minimizes the following InfoNCE [30] loss function:

$$\mathcal{L} = \sum_{i=1}^{N} -\log \hat{p}_\theta(a_i|s_i, \{\tilde{a}_i^j\}) \tag{4}$$

where

$$\hat{p}_\theta(a_i|s_i, \{\tilde{a}_i^j\}) = \frac{e^{-E_\theta(s_i,a_i)}}{e^{-E_\theta(s_i,a_i)} + \sum_j e^{-E_\theta(s_i,\tilde{a}_i^j)}}$$

Florence *et al.* [11] propose three techniques for generating these counter-examples $\{\tilde{a}_i^j\}$ and performing inference over the learned model E_θ; we choose gradient-based Langevin sampling [45] in this work as Florence *et al.* demonstrate that it scales with action dimensionality better than the alternate methods. This is a Markov Chain Monte Carlo (MCMC) method with stochastic gradient Langevin dynamics. More details are available in Appendix B.3 of [11].

B. Interactive Dataset Aggregation

Behavior cloning is prone to distribution shift due to compounding approximation errors [34], and any data-driven robot policy may encounter edge cases at execution time that are not represented in the training data [18]. We extend Implicit BC to interactive imitation learning using dataset aggregation of online human data, as in DAgger [34] and variants [18, 22]:

$$D^{t+1} \leftarrow D^t \cup D_H^t$$

$$D_H^t := \{(s_i^t, \pi_H^j(s_i^t)) : \pi_H^j(s_i^t) \neq R \text{ and } \sum_{j=1}^{M} \boldsymbol{\alpha}_{ij}^t = 1\}$$

where $\pi_H^j(s_i^t)$ is the teleoperation action from human j for robot i at time t, and α_{ij}^t is the binary assignment of human j to robot i at time t, as in Equation 2. The shared robot policy is iteratively updated with the aggregate dataset at a fixed interval $1 \leq \hat{t} \leq T$ via supervised learning:

$$\pi_{\theta_t} \leftarrow \arg\min_\theta \mathcal{L}(\pi_\theta, D^t)$$

Ross *et al.* [34] show that such a policy incurs approximation error that is linear in the time horizon rather than quadratic, as in behavior cloning.

C. Energy-Based Allocation

IFL requires specification of a metric for autonomously determining the assignment of available human supervisors to robots that require assistance. Several prior methods [16, 18, 28] use the variance of the predictions of an ensemble of typically 5-10 (unimodal) explicit BC policies bootstrapped on subsets of the training data as an estimate for the policy's epistemic uncertainty at a given state. This approach is not applicable to implicit policies because multimodality results in a false positive: different ensemble members can select equally good actions from different modes, leading to high variance even when there should be low uncertainty. Additionally, training and inference in EBMs are much more computationally expensive than in explicit models due to the InfoNCE loss (Equation 4) and implicit optimization (Equation 1), making ensembles of 5+ models impractical. Finally, inference is inherently nondeterministic, creating an additional source of variance that is not due to uncertainty.

The notion of disagreement between models can still be applicable to implicit policies by considering their *distributions* at a given state rather than the single predicted actions. Accordingly, we consider bootstrapping 2 implicit policies and calculate the Jeffreys divergence D_J [20] between them to measure their disagreement. Jeffreys divergence, the sum of two pairwise KL divergences D_{KL} of each distribution from the other, offers two compelling properties: (1) it is symmetric, consistent with both policies having been trained the same way, and (2) it is possible to estimate Jeffreys divergence for EBMs without knowing the partition function $Z(s)$, which is intractable to compute in high dimensional spaces. To show (2), we derive the following identity:

Identity 1: Let E_1 and E_2 be two energy-based models that respectively define distributions π_1 and π_2 according to Equation 3. Then,

$$D_J\left(\pi_1(\cdot|s) \| \pi_2(\cdot|s)\right) = \mathbb{E}_{a\sim\pi_1(\cdot|s)}\left[E_2(s,a) - E_1(s,a)\right] + \mathbb{E}_{a\sim\pi_2(\cdot|s)}\left[E_1(s,a) - E_2(s,a)\right].$$

Proof:

$$\begin{aligned}
&D_J\left(\pi_1(\cdot|s) \| \pi_2(\cdot|s)\right) \\
&\quad \triangleq D_{KL}\left(\pi_1(\cdot|s) \| \pi_2(\cdot|s)\right) + D_{KL}\left(\pi_2(\cdot|s) \| \pi_1(\cdot|s)\right) \\
&\quad \triangleq \mathbb{E}_{a\sim\pi_1(\cdot|s)}\left[\log \frac{\pi_1(a|s)}{\pi_2(a|s)}\right] + \mathbb{E}_{a\sim\pi_2(\cdot|s)}\left[\log \frac{\pi_2(a|s)}{\pi_1(a|s)}\right] \\
&\quad = \mathbb{E}_{a\sim\pi_1(\cdot|s)}\left[E_2(s,a) - E_1(s,a)\right] - \log Z_1(s) + \log Z_2(s) \\
&\qquad + \mathbb{E}_{a\sim\pi_2(\cdot|s)}\left[E_1(s,a) - E_2(s,a)\right] - \log Z_2(s) + \log Z_1(s) \\
&\quad = \mathbb{E}_{a\sim\pi_1(\cdot|s)}\left[E_2(s,a) - E_1(s,a)\right] \\
&\qquad + \mathbb{E}_{a\sim\pi_2(\cdot|s)}\left[E_1(s,a) - E_2(s,a)\right]
\end{aligned}$$

■

todo: move to appendix if short on space. From an initial search this appears to be novel, but we will more thoroughly check if anyone has done this before Crucially, the intractable log partition functions are cancelled out due to the symmetry of the Jeffreys divergence. We estimate the expectations in Identity 1 using Langevin sampling. Intuitively, if both policies are the same, the Jeffreys divergence will be zero; otherwise, it will be positive and independent of the magnitude of the energy functions. Since the Jeffreys divergence provides a measure of the robot's epistemic uncertainty, we use it with Fleet-DAgger [18] for energy-based allocation in Implicit IFL. As in Fleet-EnsembleDAgger [18], we prioritize robots that have uncertainty above a given threshold value followed by robots which are violating constraints.

V. Experiments

A. 2D Navigation Simulation Experiments

To evaluate the correctness of our implementation and provide visual intuition, we first run experiments in a 2D

pointbot navigation environment. See Figure 2 for the maze environment, representative trajectories, and energy distribution plots. We consider discrete 2D states $s = (x, y) \in \mathbb{N}^2$ and continuous 2D actions $a = (\Delta x, \Delta y) \in [-1, 1]^2$. The maze has a fixed start and goal location and consists of a forked path around a large obstacle followed by a long corridor. An algorithmic supervisor provides 100 demonstrations of the task, randomly choosing to go upward or downward at the fork with 50% probability each. Since a model can simply overfit to the demonstrations in this low-dimensional environment, to induce distribution shift we add "wind" at execution time to a segment of the right corridor with magnitude 0.75 in the $+y$ direction.

In 100 trials, Explicit BC achieves a 0% success rate, Implicit BC achieves a 0% success rate, and Implicit IFL achieves a 100% autonomous success rate (i.e., robot-only trajectories without human interventions, after interactive training). In Figure 2 we observe that Explicit BC cannot pass the fork due to averaging the two modes to zero. Meanwhile, Implicit BC is not robust to the distribution shift: once the wind pushes the robot to the top of the corridor, it does not know how to return to the center. We also observe that the Implicit IFL energy distributions in Figure 2(B) reflect the desired behavior in accordance with intuition.

B. IFL Benchmark Simulation Experiments

Environments: We evaluate with the following 3 continuous control environments from Isaac Gym [26] and the Interactive Fleet Learning Benchmark [18]: Ball Balance, Ant, and Anymal. Isaac Gym enables efficient GPU-accelerated simulation of robot fleets. Environment details are available in the appendix.

Metrics: We measure the average episode reward across the fleet, the total number of hard resets, and the total idle time (how long robots that require a hard reset spend idle waiting for humans). For interactive algorithms, we also measure the return on human effort. We use the episode rewards provided by Isaac Gym to measure performance but do not require the task to define a reward in order to apply our method, as Implicit IFL does not perform reinforcement learning.

Baselines: We compare Implicit IFL to the following baselines: Explicit BC, Implicit BC, Explicit IFL (specifically, Fleet-EnsembleDAgger [18]), and Random Implicit IFL (which performs random human-to-robot allocation rather than the proposed energy-based allocation).

Experimental Setup: We run experiments with a fleet of $N = 10$ robots and $M = 3$ heterogeneous algorithmic supervisors. The supervisors are reinforcement learning agents trained with Isaac Gym's reference implementation of PPO [35]. To create heterogeneous supervisors, we train 3 PPO agents to convergence with different seeds: 10, 100, and 1000. This creates supervisors that attain similar rewards but execute different actions. For instance, when compared to the actions taken by the seed 10 Ball Balance expert on its own state trajectory, the seed 100 and seed 1000 expert actions for those states have a mean L2 distance of 1.49 ± 0.55 and 1.50 ± 0.54 respectively, which are as far apart as actions taken by a random agent (mean L2 distance 1.47 ± 0.55). All training runs have hard reset time $t_R = 5$ timesteps, minimum intervention time $t_T = 5$ timesteps, and fleet operation time $T = 10,000$ timesteps, and are averaged over 3 random seeds. The initial robot policy π_{θ_0} for all algorithms is initialized with behavior cloning on 5 full task demonstrations from each of the 3 experts (e.g., Ant with an episode length of 1000 gets $3 \times 5 \times 1000 = 15000$ state-action pairs); the offline algorithms of Explicit BC and Implicit BC receive double this amount since they do not receive online data.

Results: The results are shown in Figure todo: . We observe that ...

C. Physical Experiments

Experimental Setup: Similar to Hoque *et al.* [18], we run an image-based block-pushing experiment with a fleet of $N = 4$ ABB YuMi robot arms operating simultaneously in 2 labs about 1 km apart. See Figure todo: for the physical setup. Each robot has an identical square workspace with a small blue cube and rectangular pusher as an end effector. Unlike the prior work, we add a large square obstacle (side length about $3\times$ the block's side length) to the center of each workspace, which the robot must avoid on the way to the goal. The task is reset-free in that the goal region is generated in the overhead image observation to be diametrically opposite the cube's initial location and across the center obstacle. However, if the cube collides with the obstacle or the boundaries of the workspace, a physical hard reset R is required. Furthermore, unlike the discrete actions in the prior work, we use a continuous 2D action space of $a = (\Delta x, \Delta y)$ that corresponds to the vector along which to push the block, starting from the block's center. We set $t_T = 3, t_R = 5$, and $T = 150$ for a total of $150 \times 4 = 600$ pushing actions for each of 4 algorithms: Explicit BC, Implicit BC, Explicit IFL (Fleet-EnsembleDAgger), and Implicit IFL.

Results: The results are shown in Table todo: . We observe that ...

VI. Limitations and Future Work

In this paper we present Implicit IFL, an algorithm for interactive fleet learning from multimodal human supervisors. Since we extend Implicit BC, we inherit some of its limitations: namely, model training and inference require more computation time than explicit methods, and performance falls when the action space dimensionality is very high ($|A| > 16$). Our energy-based switching also requires the training of two implicit models rather than one.

In future work, we hope to extend recently proposed alternative approaches for handling multimodality such as Behavior Transformers [37] and Diffusion Policies [9] to the IFL setting and compare them to Implicit IFL. We also hope to develop algorithms that effectively learn from human demonstrations that are not only multimodal but also suboptimal or of mixed quality.

REFERENCES

[1] P. Abbeel and A. Y. Ng, "Apprenticeship learning via inverse reinforcement learning," in *Proceedings of the twenty-first international conference on Machine learning*, 2004, p. 1.

[2] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robotics and autonomous systems*, vol. 57, no. 5, pp. 469–483, 2009.

[3] S. Arora and P. Doshi, "A survey of inverse reinforcement learning: Challenges, methods and progress," *arXiv preprint arXiv:1806.06877*, 2018.

[4] Y. Avigal, L. Berscheid, T. Asfour, T. Kroger, and K. Goldberg, "Speedfolding: Learning efficient bimanual folding of garments," *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1–8, 2022.

[5] C. M. Bishop, "Mixture density networks," *Neural Computing Research Group Report*, 1994.

[6] D. S. Brown, W. Goo, and S. Niekum, "Better-than-demonstrator imitation learning via automaticaly-ranked demonstrations," in *Conference on Robot Learning (CoRL)*, 2019.

[7] D. S. Brown, S. Niekum, R. Coleman, and R. Srinivasan, "Safe imitation learning via fast bayesian reward inference from preferences," in *International Conference on Machine Learning (ICML)*, 2020.

[8] J. Chen, B. Yuan, and M. Tomizuka, "Deep imitation learning for autonomous driving in generic urban scenarios with enhanced safety," *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2884–2890, 2019.

[9] C. Chi *et al.*, "Diffusion policy: Visuomotor policy learning via action diffusion," *arXiv preprint arXiv:2303.04137*, 2023.

[10] S. E. F. Chipman, *The Oxford Handbook of Cognitive Science*. Oxford University Press, Oct. 2017.

[11] P. R. Florence *et al.*, "Implicit behavioral cloning," in *Conference on Robot Learning (CoRL)*, 2021.

[12] K. Gandhi, S. Karamcheti, M. Liao, and D. Sadigh, "Eliciting compatible demonstrations for multi-human imitation learning," in *Conference on Robot Learning (CoRL)*, 2022.

[13] I. Goodfellow *et al.*, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., vol. 27, Curran Associates, Inc., 2014.

[14] J. Ho and S. Ermon, "Generative adversarial imitation learning," in *Neural Information Processing Systems (NeurIPS)*, 2016.

[15] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *arXiv preprint arXiv:2006.11239*, 2020.

[16] R. Hoque, A. Balakrishna, E. Novoseller, A. Wilcox, D. S. Brown, and K. Goldberg, "ThriftyDAgger: Budget-aware novelty and risk gating for interactive imitation learning," in *Conference on Robot Learning (CoRL)*, 2021.

[17] R. Hoque *et al.*, "LazyDAgger: Reducing context switching in interactive imitation learning," in *IEEE Conference on Automation Science and Engineering (CASE)*, 2021, pp. 502–509.

[18] R. Hoque *et al.*, "Fleet-dagger: Interactive robot fleet learning with scalable human supervision," in *Conference on Robot Learning (CoRL)*, 2022.

[19] R. Hoque *et al.*, "Learning to fold real garments with one arm: A case study in cloud-based robotics research," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022.

[20] H. Jeffreys, *The theory of probability*. OuP Oxford, 1998.

[21] Y. Jiang *et al.*, "VIMA: General robot manipulation with multimodal prompts," in *NeurIPS 2022 Foundation Models for Decision Making Workshop*, 2022.

[22] M. Kelly, C. Sidrane, K. Driggs-Campbell, and M. J. Kochenderfer, "Hg-dagger: Interactive imitation learning with human experts," *2019 International Conference on Robotics and Automation (ICRA)*, pp. 8077–8083, 2018.

[23] J. W. Kim, P. Zhang, P. L. Gehlbach, I. I. Iordachita, and M. Kobilarov, "Towards autonomous eye surgery by combining deep imitation learning with optimal control," in *Conference on Robot Learning (CoRL)*, 2020.

[24] Y. LeCun, S. Chopra, R. Hadsell, A. Ranzato, and F. J. Huang, "A tutorial on energy-based learning," *Predicting Structured Data*, vol. 1, no. 0, 2006.

[25] H. Liu, S. Nasiriany, L. Zhang, Z. Bao, and Y. Zhu, "Robot learning on the job: Human-in-the-loop autonomy and learning during deployment," *arXiv*, vol. abs/2211.08416, 2022.

[26] V. Makoviychuk *et al.*, "Isaac gym: High performance gpu-based physics simulation for robot learning," *arXiv preprint arXiv:2108.10470*, 2021.

[27] A. Mandlekar *et al.*, "What matters in learning from offline human demonstrations for robot manipulation," in *Conference on Robot Learning (CoRL)*, 2021.

[28] K. Menda, K. Driggs-Campbell, and M. J. Kochenderfer, "EnsembleDAgger: A Bayesian Approach to Safe Imitation Learning," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019.

[29] C. G. Northcutt, A. Athalye, and J. Mueller, "Pervasive label errors in test sets destabilize machine learning benchmarks," in *Neural Information Processing Systems (NeurIPS)*, 2021.

[30] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *ArXiv preprint arXiv:1807.03748*, 2018.

[31] Y. Pan *et al.*, "Agile autonomous driving using end-to-end deep imitation learning," in *Robotics: Science and Systems (RSS)*, 2018.

[32] S. Paradis *et al.*, "Intermittent visual servoing: Efficiently learning policies robust to instrument changes for high-precision surgical manipulation," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 7166–7173.

[33] D. A. Pomerleau, "Alvinn: An autonomous land vehicle in a neural network," in *Neural Information Processing Systems (NeurIPS)*, D. Touretzky, Ed., vol. 1, Morgan-Kaufmann, 1988.

[34] S. Ross, G. Gordon, and D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011, pp. 627–635.

[35] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[36] D. Seita *et al.*, "Deep imitation learning of sequential fabric smoothing from an algorithmic supervisor," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 9651–9658.

[37] N. M. Shafiullah, Z. J. Cui, A. Altanzaya, and L. Pinto, "Behavior transformers: Cloning k modes with one stone," in *Neural Information Processing Systems (NeurIPS)*, 2022.

[38] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 640–651, 2017.

[39] M. Shridhar, L. Manuelli, and D. Fox, "Perceiver-actor: A multi-task transformer for robotic manipulation," in *Conference on Robot Learning (CoRL)*, 2022.

[40] J. Spencer *et al.*, "Learning from interventions: Human-robot interaction as both explicit and immplicit feedback," in *Robotics: Science and Systems (RSS)*, 2020.

[41] X. Sun, S. Yang, and R. Mangharam, "Mega-dagger: Imitation learning with multiple imperfect experts," *ArXiv*, vol. arXiv preprint arXiv:2303.00638, 2023.

[42] F. Torabi, G. Warnell, and P. Stone, "Generative adversarial imitation from observation," *ArXiv*, vol. abs/1807.06158, 2018.

[43] H.-Y. Tung, A. W. Harley, L.-K. Huang, and K. Fragkiadaki, "Reward learning from narrated demonstrations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7004–7013.

[44] A. Vaswani *et al.*, "Attention is all you need," in *Neural Information Processing Systems (NeurIPS)*, 2017.

[45] M. Welling and Y. W. Teh, "Bayesian learning via stochastic gradient langevin dynamics," in *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ser. ICML'11, Bellevue, Washington, USA: Omnipress, 2011, 681–688.

[46] A. Zeng *et al.*, "Transporter networks: Rearranging the visual world for robotic manipulation," *Conference on Robot Learning (CoRL)*, 2020.

[47] J. Zhang and K. Cho, "Query-efficient imitation learning for end-to-end autonomous driving," in *Association for the Advancement of Artificial Intelligence (AAAI)*, 2017.