# A Dual Representation Framework for Robot Learning with Human Guidance

**Ruohan Zhang*[1], Dhruva Bansal*[1], Yilun Hao*[1], Ayano Hiranaka[2], Jialu Gao[3], Chen Wang[1], Roberto Martín-Martín[4], Li Fei-Fei[1,5], Jiajun Wu[1,5]**
[1]Department of Computer Science, Stanford University
[2]Department of Mechnical Engineering, Stanford University
[3]Tsinghua University
[4]Department of Computer Science, The University of Texas at Austin
[5]Institute for Human-Centered AI (HAI), Stanford University

**Abstract:** The ability to interactively learn skills from human guidance and adjust behavior according to human preference is crucial to accelerating robot learning. But human guidance is an expensive resource, calling for methods that can learn efficiently. In this work, we argue that learning is more efficient if the agent is equipped with a high-level, symbolic representation. We propose a dual representation framework for robot learning from human guidance. The dual representation used by the robotic agent includes one for learning a sensorimotor control policy, and the other, in the form of a symbolic scene graph, for encoding the task-relevant information that motivates human input. We propose two novel learning algorithms based on this framework for learning from human evaluative feedback and from preference. In five continuous control tasks in simulation and in the real world, we demonstrate that our algorithms lead to significant improvement in task performance and learning speed. Additionally, these algorithms require less human effort and are qualitatively preferred by users. Project website: `https://sites.google.com/view/dr-hrl`.

**Keywords:** Human Guidance, Evaluative Feedback, Preference Learning

## 1 Introduction

Human guidance refers to a diverse set of human training signals provided to a learning agent [1–5]. These alternative forms of human input can be combined with the conventional reward signal in reinforcement learning (RL) [6] or demonstrations in imitation learning (IL) [7–9] to facilitate learning. Recently, the robot learning community has increased its attention to human guidance as a mechanism to overcome two critical challenges: 1) the low sample efficiency of learning algorithms, and 2) the effort in manually specifying the objectives for learning. Human guidance is helpful for these challenges because 1) guidance like evaluative feedback [10–27] can be used as domain knowledge to speed up learning, and 2) through their guidance, humans can define the learning objectives for robots so that the learning algorithm better infers and aligns to the underlying human goals and values, such as their preferences [28–46].

Despite its benefits, human guidance is a scarce and valuable resource, and human-in-the-loop mechanisms strive to find ways to reduce the amount of guidance required from humans to make them more broadly and practically applicable. This is only possible if the human guidance is properly represented and interpreted [47–49]. The first step is to choose a representation that allows the learning agent to query humans more efficiently and learn with less human guidance [47–49].

Robots typically use a fine-grained state and action space for continuous control. However, when humans observe, evaluate, and guide robot behaviors, their representation is likely different. For example, in a continuous control task such as placing an object, the robotic agent typically has access to low-level states including proprioceptive information and other objects' poses. But the guidance

*indicates equal contribution; correspondence to zharu@stanford.edu

6th Conference on Robot Learning (CoRL 2022), Auckland, New Zealand.

provided by humans is typically based on higher-level abstract information, such as *is the robot's end-effector to the left or the right of the goal?*, or *is the robot grasping the object?* This observation leads to the "dual representation hypothesis" proposed in this work, inspired by cognitive science studies [50–53]. This is analogical to the "fast and slow" systems proposed by Kahneman [52], in which the fast system manages intuitive, automatic, unconscious behaviors while the slow system manages logical, calculating, conscious thoughts [52]. We hypothesize that the "slow" system and its representation are useful in guiding learning agents.

In this work, we propose a *dual representation* framework for robot learning from human guidance: the robotic learning agent uses a low-level state representation for learning control policies, but keeps a *symbolic scene graph* [54–57] as a high-level representation of human internal states. We show that this framework enables novel learning algorithms: Dual Representation–based Evaluative Feedback (DREF) and Dual Representation–based Preference Learning (DRPL). DREF is based on the idea of *uncertainty-aware active learning* that allows the agent to estimate the uncertainty of human feedback in unseen states. Such generalization ability is achieved by using scene graph representation to group low-level states into abstract states. DRPL builds upon *scene graph–based trajectory segmentation and selection*, allows efficient reward learning from chosen trajectory segments.

In three simulation tasks and two real-robot tasks, we demonstrate that our proposed approaches lead to significant improvements in learning speed and performance. For challenging long-horizon real-world robot manipulation tasks, we show that DREF can learn to solve the task efficiently, while end-to-end RL algorithms fail to solve because of the high-dimensional continuous state and action space. Critically, we observe a significant reduction in the amount of human guidance required for learning, and an improvement in overall user experience, making learning from human guidance methods more practical and appealing for real robot learning.

## 2 Related Work

Among the multiple forms of human-in-the-loop robot learning [58–60], in recent years, the robotics community has paid increased attention to human guidance [3] because of being powerful and easy to collect, and complements standard training signals such as rewards or demonstrations.

**Learning from human evaluative feedback.** This is an approach in which human trainers monitor the learning process of an agents, and provide a scalar signal to indicate whether the observed behavior is desirable [10–27]. The agent then learns a policy to maximize positive feedback from humans. This approach has the advantage of placing minimum demand on both the human trainer's expertise and the ability to provide guidance, compared to learning from demonstrations. Significantly, human evaluation is often interpreted as a value function [14, 16] or an advantage function [15, 17], not as the reward itself. Nonetheless, human evaluation can be naturally combined with environment rewards so the agent learns simultaneously from both sources [18–20]. Evaluative feedback often targets individual state-action pairs. Hence one outstanding challenge is to generalize observed human feedback to unseen state-action pairs.

**Learning from human preference.** In this framework, the learning agent queries human trainers for their preferences over a set of exhibited behavior trajectories [28–46]. Preferences can be used to directly learn policies [28, 29], a preference model [30], or a reward function [31, 32]. Recent works often learn a hypothesized latent human reward function from the preference signals, and combine preference learning with deep RL to learn a policy afterwards [34–37]. Notably, a recent work attempts to align agent representation with human representation in preference learning [61]. While evaluative feedback targets a state-action pair, preference learning targets trajectories. Selecting optimal query trajectory length is a challenging problem in preference learning [3], we will show that the dual representation learning framework provides an intuitive way to segment lengthy trajectories.

**Scene graph as abstract representation.** In robot learning, abstract representation has been an active research topic, exemplified by recent works in neuro-symbolic robot learning [62–65]. Human-robot interaction tasks, such as shared autonomy [66, 67], highlight the importance of abstract representation since humans often communicate high-level goals to the agent in this setting. Low-dimensional, abstract representations could potentially be learned [68–72] but sample efficiency is a major limiting factor when applying these methods to physical robot learning from real humans.

A scene graph, often specified by humans, is a form of abstract representation for state information in which objects are represented as nodes and relations between objects are represented as edges [54–

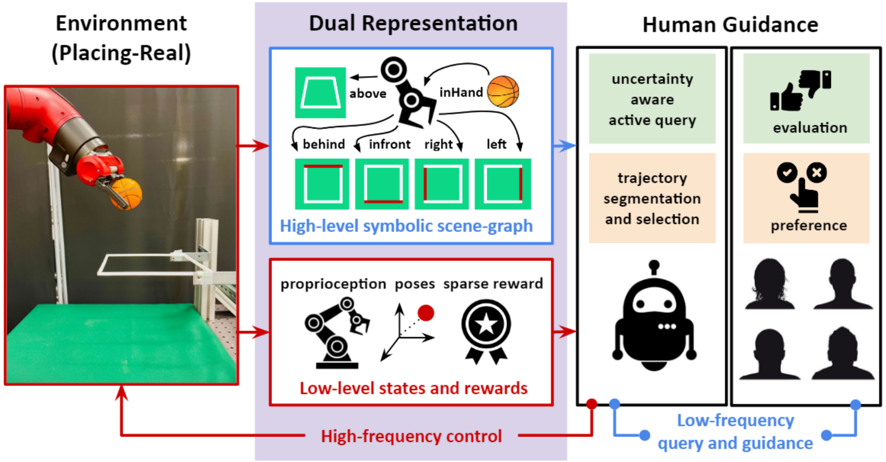


Figure 1: Overview of the proposed dual representation framework. The robot maintains two state representations. The first one is for learning a fine-grained, low-level continuous control policy. The second representation is specified by an expert human in the format of a symbolic scene graph. The robot uses this representation to actively query human trainers for evaluation or preference during the training process.

57]. Scene graphs can store explicitly and compactly information about object geometry, placement, semantics, and relationships, making them suitable for tasks that require sophisticated reasoning about these types of information. Recently, scene graphs have started to be used in robotics [73–75] as a state representation that facilitates planning and reasoning. We explore how a symbolic scene graph could be a useful additional representation for human-in-the-loop robot learning. We hypothesize that it allows the learning agent to query humans more efficiently and learn with less human guidance than using the low-level, raw state representation alone.

## 3 Method

Our method is designed to overcome a significant challenge in human-in-the-loop robot learning: human feedback is expensive, and frequently asking for guidance is infeasible in real-world robotic systems. To address this challenge, we propose a novel dual representation to facilitate human-in-the-loop policy learning. Below we introduce the general human guidance learning problem and our proposed dual representation setups (Sec. 3.1). Then, we propose an algorithm to learn from evaluation feedback that decreases the amount of human feedback needed during policy learning (Sec. 3.2), and a query generation algorithm to efficiently query humans for their preference (Sec. 3.3).

### 3.1 Dual Representation for Learning from Human Guidance

We represent the robot learning problem as a Markov Decision Process denoted by the tuple $\langle \mathcal{S}, \mathcal{A}, P, R, \gamma \rangle$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $R : S \times A \rightarrow \mathbb{R}$ is the reward function, $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$ is the transition function, and $\gamma$ is the discount factor. A policy $\pi : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$ is a mapping from $\mathcal{S}$ to probability distribution over $\mathcal{A}$.

We propose a dual representation framework for human-in-the-loop robot learning that consists of the common low-level state information from the environment and an additional high-level symbolic scene graph $\mathcal{G}$. $\mathcal{G}$ is represented as a binary vector of $dim(\mathcal{G})$, where each dimension represents a unary state of an object or a pairwise semantic relation between the objects and the robot. Let $g$ be an instance of $\mathcal{G}$, i.e., an abstract state which contains infinite low-level, continuous states. The objects and relations in $\mathcal{G}$ are specified by humans based on task knowledge. The proposed dual representation framework is shown in Fig. 1.

### 3.2 Learning from Evaluative Feedback with Dual Representation

Human evaluative feedback contains rich task-level knowledge that could be used to assist robot learning. We adopt an *active* learning setting, in which an RL agent *asks* humans for their evaluative feedback. Hypothetically, asking and receiving feedback in the right state could lead to better learning results with less human effort, but it is unclear *when* to ask for feedback. Our key insight

**Algorithm 1** Dual Representation–based Evaluative Feedback (DREF)

```
Initialize UCB1($g$), $\forall g \in \mathcal{G}$, as 0
Initialize $\mathcal{D}$ as the replay buffer
Initialize network weights for actor $\theta_A$, environment critic $\theta_E$, and human feedback critic $\theta_H$
for t = 1, T do
    Select $a_t = \pi(\cdot|s_t; \theta_A)$
    Infer current scene graph state $g_t$ from $s_t$
    if $rank(\texttt{UCB1}(g_t)) \leq k$, across $g \in \mathcal{G}$ sorted by UCB1 score then
        Query for evaluative feedback $H_t(s_t, a_t)$
    end if
    Execute $a$, observe reward $r$, and next state $s_{t+1}$,
    Store transition ($s_t$, $a_t$, $r_t$, $H_t$, $s_{t+1}$) in $\mathcal{D}$
    Sample random minibatch of transitions ($s_t$, $a_t$, $r_t$, $H_t$, $s_{t+1}$) from $\mathcal{D}$
    Perform a gradient step according to Equations 3 4, and 5 for $\theta_H$, $\theta_E$, $\theta_A$.
    Update UCB1($g$), $\forall g \in \mathcal{G}$ using Equations 1,2
end for
```

is that querying for human feedback can be formulated as a multi-armed bandit problem with symbolic scene graph representation. This formulation brings the opportunities to use formal methods designed for discrete state and action spaces, such as Upper Confidence Bound (UCB1). Integrating UCB1 with continuous control enables efficient *uncertainty-aware active learning* from humans.

TAMER+RL [18–20] is a widely used framework for learning from evaluative feedback. For the RL part, we use Soft Actor-Critic (SAC) [76]. In addition to the environment reward, human trainers provide a scalar signal $H_t(s, a) \in \{-1, 0, +1\}$ to indicate whether the observed state-action pair is desirable or not. Our goal is to reduce the amount of total feedback while achieving the same or better task performance.

Inspired by the standard UCB1 for value function in bandit problems, we use the following equation to estimate the upper confidence bound of human feedback prediction error (FPE), the type of uncertainty we care about, in an abstract state $g_t$:

$$\texttt{UCB1}(g_t) = \texttt{FPE}(g_t) + c\sqrt{\frac{2\log N_t}{N_t(g_t)}}, \tag{1}$$

where $N_t$ is the total number of human feedback received at time $t$, and $N_t(g_t)$ is the number of feedback given to the abstract state $g_t$. The constant $c$ weighs the exploitation and exploration terms. $\texttt{FPE}(g_t)$ is the average feedback prediction error for all the low-level states encountered so far that belongs to $g_t$:

$$\texttt{FPE}(g_t) = \frac{1}{N_t(g_t)} \sum_{i=0}^{t} \mathbb{1}(s_i \in g_t) \|\hat{H}(s_i, a_i) - H(s_i, a_i)\|_2^2. \tag{2}$$

We can estimate $\hat{H}(s, a)$ using an additional critic head in SAC. Assume $\theta_H$, $\theta_E$, $\theta_A$, parameterize the human feedback critic, the environment critic, and the actor, respectively, the learning objectives are:

$$\mathcal{L}(\theta_H) = \mathbb{E}_{(s,a,H)\sim\mathcal{D}} \|\hat{H}(s, a; \theta_H) - H(s, a)\|_2^2 \tag{3}$$

$$\mathcal{L}_{\texttt{SAC}}(\theta_E) = \mathbb{E}_{(s,a)\sim\mathcal{D}} \|Q(s, a; \theta_E) - \hat{Q}(s, a)\|_2^2 \tag{4}$$

$$\mathcal{L}(\theta_A) = \mathbb{E}_{s\sim\mathcal{D}} \Big[ \mathbb{E}_{a\sim\pi(\cdot|s;\theta_A)} \big[ \alpha \log(\pi(a|s; \theta_A)) - (Q(s, a; \theta_E) + \lambda\hat{H}(s, a; \theta_H)) \big] \Big] \tag{5}$$

In Eq. 4, $\mathcal{L}_{\texttt{SAC}}(\theta_E)$ is the standard soft Bellman residual in SAC [76, 77]. In Eq. 5, $\alpha$ is the temperature parameter in SAC [76, 77]. Note that the actor updates the policy distribution in the direction suggested by the weighted average of both critics. The agent learns a policy to maximize expected positive feedback from humans and environment reward simultaneously.

The full algorithm, **DREF** or Dual Representation–based Evaluative Feedback, is shown in Algorithm 1. We first calculate the running mean of UCB value for each abstract state and rank them.

**Algorithm 2** Dual Representation–based Preference Learning (DRPL)

```
Collect a set of random trajectories $\mathcal{T}$
Set prior of reward weights $p^0(\theta)$ randomly
Segment all trajectories in $\mathcal{T}$ based on abstract state changes: $\mathcal{T}_{seg} = \{\xi_1, \xi_2, ...\}$
for $i = 0, dim(\mathcal{G})$ do
    Select two segments $\xi, \xi' \in \mathcal{T}_{seg}$, with ties broken arbitrarily
    s.t. their associated abstract states $g[i] \neq g'[i]$
    and $d(g, g') = 1$ or $d(g, g') = \min_{g,g'} d(g, g')$
    Query for preference $q(\xi, \xi')$
    Update posterior $p^{(i+1)}(\theta) \propto P(q|\theta)p^i(\theta)$, where $P(q|\theta)$ is the probability of preference
    response $q$ given $\theta$
end for
```

Then we query for feedback at a specific state $s_t$ if its abstract state $g_t$ has a high rank (determined by a hyperparameter $k$). As a result, the agent actively asks for feedback in abstract states with uncertainty in the human evaluation of actions, e.g., query for feedback when the gripper and the ball move into the hoop for the first time, as shown in Fig. 1. With the low-level state and action space alone, this cannot be easily done because the number of states is infinite. The abstract state representation effectively groups these states together, and estimates the uncertainty in predicting feedback of a new state using the average FPE of other states in its group. We hypothesize that in this way, the dual representation could lead to efficient *uncertainty-aware active learning* from human feedback, which we will demonstrate in our experiments.

### 3.3 Preference Learning with Dual Representation

Preference learning is an important method to define the objective for learning agents. While evaluative feedback targets a state-action pair, in preference learning human trainers indicate their preference over a pair of trajectories, from which the agent learns a reward function $R$. For simplicity, we assume that reward function $R$ is a linear combination of state-action features [78]: $r(s_t, a_t) = \theta^T \phi(s_t, a_t)$, hence the goal is to infer $\theta$. Our goal is to reduce the amount of human guidance: the algorithm should accurately estimate $\theta$ with a minimum number of queries.

Here we address two issues in preference learning: *generating* and *selecting* meaningful queries. For query generation, selecting trajectory length is challenging [3]. Indicating preference over longer trajectories requires cognitive effort (e.g., summing all the rewards in each trajectory). Short trajectories (e.g., random 1-2 seconds clips [34]) allow humans to provide feedback of high granularity, but these clips may not be meaningful or comparable. After trajectories are segmented, selecting meaningful pairs to query humans is yet another challenge: apples-to-apples comparisons are more meaningful.

We propose a simple solution to these challenges: scene graph–based trajectory segmentation and selection. The abstract state dimensions in the scene graphs naturally overlap with the reward features, since these scene graphs are designed to contain information that is critical to task success. The key observation is that in long-horizon tasks, a trajectory consists of multiple abstract state transitions, and *two consecutive abstract state transitions* naturally define the starting and ending points of a segment, i.e., a segment corresponds to only *one* abstract state $g$. Then, for query selection, we select a pair of segments with two abstract states $g$ and $g'$, such that $d(g, g') = 1$ where $d$ denotes the hamming distance (recall that $g$ is a binary vector). In other words, $g$ and $g'$ are two abstract states that only differ in one dimension. Hence the remaining dimensions are held constant, leading to a controlled comparison. If we cannot find a pair of segments such that $d(g, g') = 1$, we choose the pair that has the minimum hamming distance. The algorithm, **DRPL**, or Dual Representation based Preference Learning, is shown in Algorithm 2. We use a standard reward weight estimation algorithm (a Bayesian inverse reinforcement learning algorithm) that is based on the low-level representation [41, 78], the only difference is how we segment and select trajectories for a query.

## 4 Experiments

We test our algorithms in five continuous control tasks (Fig. 2) for both evaluative feedback and preference learning. The following describes the dual representation (low and high level) of each task.

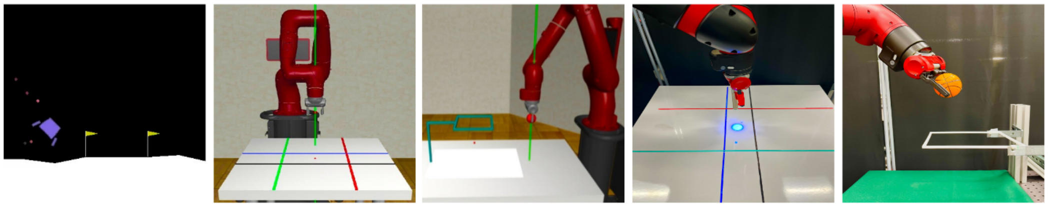


Figure 2: Five continuous control tasks in our experiments: `Lunar-Lander`, `Reaching-Sim`, `Placing-Sim`, `Reaching-Real`, and `Placing-Real`.

**Low-level representation.** In `Lunar-Lander-Continuous-v2` [79], we control a spaceship to land in the middle of two flags without crashing. The state, $s \in \mathbb{R}^8$, includes position, velocity, and leg-contact information, and the action signals, $a \in \mathbb{R}^2$ control the engines. In `Reaching-Sim` and `Reaching-Real`, the robot's goal is to move its gripper, to the center area marked on the table from a random starting location. The state, $s \in \mathbb{R}^2$, specifies the position of the gripper in the $xy$-plane and the actions, $a \in \mathbb{R}^2$, are delta movements on that plane. In `Placing-Sim` and `Placing-Real`, the robot must move an initially grasped ball from a random location into a hoop and drop it there. The state, $s \in \mathbb{R}^4$, includes the gripper position in 3D space and the gripper state (close/open). The actions, $a \in \mathbb{R}^4$, are delta movements in 3D Cartesian space and the gripper control. The simulation tasks are implemented in the Robosuite [80]. Further details can be found in Appendix 1.

**High-level representation.** Fig. 1 depicts the symbolic scene graph for `Placing-` tasks, where nodes correspond to task-relevant objects, and edges encode binary relations between them. Appendix 1 includes further details about the tasks and their symbolic scene graphs.

### 4.1 Results: Evaluative Feedback

We use synthetic humans in simulated environments and real humans in real-world environments. For synthetic humans (a fully trained SAC agent, more details in Appendix 2), the learning agent chooses an action $a$ in state $s$, and the oracle chooses an action $a^*$. The oracle SAC computes the Q values for these actions: $Q(s,a)$ and $Q(s,a^*)$. If the learning agent chooses an action that has a Q-value close enough to $Q(s,a^*)$, it is a good action and the agent should receive positive feedback. Otherwise, it should receive negative feedback: $H(s,a) = +1,$ if $Q(s,a) \geq \alpha Q(s,a^*)$; $-1$ otherwise. The $\alpha$ increases over time (see Appendix 2) to encourage the agent to learn to choose better actions during training.

We compare our method to the following baselines in simulated environments: (a) the SAC baseline without any human feedback (b) EF-100%, EF-50%, and EF-25%: the agent asks for feedback with a probability of 100%, 50%, or 25% at every timestep. Hyperparameters of all the algorithms can be found in Appendix 2.

Results for simulation (averaged across 5 seeds) are shown in Fig. 3. Our algorithm DREF (shown in blue) achieves comparable performance to EF-100% with only 6.2%, 1.6%, and 8.4% feedback in `Lunar-Lander`, `Reaching-Sim`, and `Placing-Sim`, respectively, corresponding to approximately 16×, 62×, and 11× improvement in feedback sample efficiency.

For real robot experiments with six humans per task, we omit the EF-100% baseline since it requires step-by-step feedback from humans for a very long period of time. Humans are also allowed to choose "no feedback" if they prefer. The order of the three methods is randomized to counterbalance the ordering effect (see Appendix 4). Fig. 4 shows the results. In `Reaching-Real`, DREF achieves comparable performance to EF-50% with 13.8% feedback, leading to a 3× improvement in human feedback sample efficiency. `Placing-Real` is extremely challenging due to the long task horizon and sparsity in reward signals. DREF achieves a better performance than all baselines, by a large margin, with only 17% human feedback.

The human experiments are followed by a survey (details in Appendix 4.1) asking humans about their overall experience training the robots (E), perceived intelligence level (I), and cognitive ease (C) (see Fig. 4). It is evident that our algorithm leads to a better user experience and reported effort (additional analyses are in Appendix 4.2). These results are supported by the observation that the total training time and no feedback count are lower for DREF. This may also explain why EF-50% performs poorly in the challenging `Placing-Real` task: the training is laborious and cognitively demanding with this amount of guidance, and humans are prone to errors in this process.

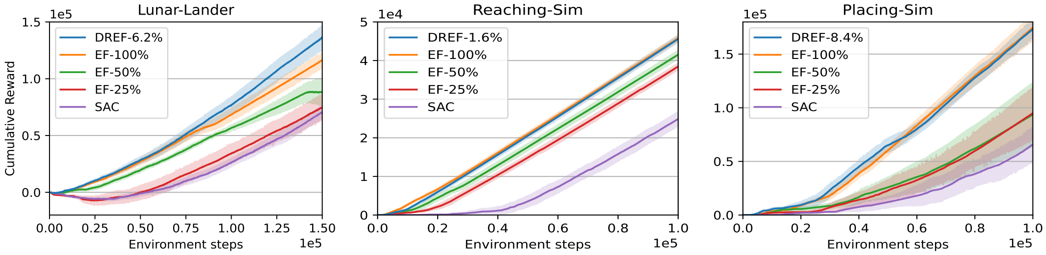
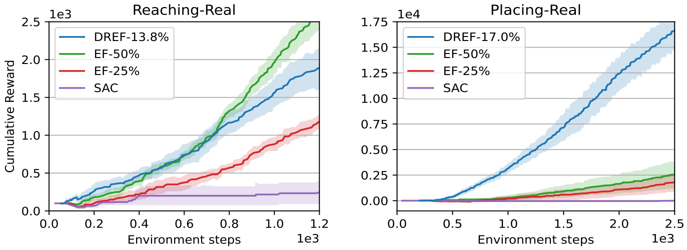


Figure 3: Cumulative rewards gained during training for `Lunar-Lander`, `Reaching-Sim`, and `Placing-Sim`. The percentage corresponds to the percentage of feedback provided by the oracle during training. The proposed algorithm, DREF, achieves comparable performance with EF-100% with much less feedback and outperforms all other baselines. Error bars indicate the standard error of the means ($n = 5$).



| Reaching | EF-25% | EF-50% | DREF |
|---|---|---|---|
| T↓ | 20.2±1.6 | 26.6±1.0 | **16.1±0.7** |
| NFC↓ | 54.5±13.5 | 89.8±26.8 | **22.2±8.8** |
| E↑ | 2.7±0.3 | 2.2±0.3 | **4.2±0.3** |
| I↑ | 2.3±0.4 | 3.0±0.6 | **4.2±0.5** |
| C↑ | 3.0±0.3 | 3.6±0.5 | **4.0±0.4** |
| Placing | EF-25% | EF-50% | DREF |
| T↓ | 36.9±4.0 | 46.1±7.5 | **25.0±10.2** |
| NFC↓ | 156.7±43.6 | 421.5±93.0 | **119.7±25.4** |
| E↑ | 2.7±0.4 | 2.5±0.3 | **4.2±0.2** |
| I↑ | 2.7±0.3 | 2.7±0.5 | **4.7±0.2** |
| C↑ | 2.7±0.6 | 2.7±0.7 | **4.2±0.3** |

Figure 4: Cumulative rewards gained during training for `Reaching-Real`, and `Placing-Real`. The proposed algorithm, DREF, achieves comparable or better performance with 50% feedback with much less feedback and outperforms all other baselines. Error bars indicate the standard error of the means ($n = 6$). Post-completion user survey (5-point Likert scale) results indicate that DREF leads to better user experience, perceived intelligence, and less reported effort ($n = 6$). T: total training time, NFC: number of "no feedback" responses, E: overall experience, I: perceived intelligence, C: cognitive ease. See Appendix 4.1 and 4.2 for survey design and additional analyses.

To conclude, the results strongly support our hypothesis about evaluative feedback: asking for human feedback sensibly leads to a better learning outcome. DREF achieves this by implementing uncertainty-aware active learning within the dual representation framework.

## 4.2 Results: Preference Learning

We now present the results of preference learning experiments. Similar to evaluative feedback, we use synthetic humans in simulated environments and real humans in real-world environments. The synthetic human has access to the true reward weight $\theta$ (more details in Appendix 3). For every pair of queries($\xi, \xi'$), the oracle calculates the true reward $r$ and $r'$. The oracle returns $q(\xi, \xi') = 0$ if $r(\xi) > r(\xi')$, and $q(\xi, \xi') = 1$ otherwise. The trajectories are generated by agents starting at random positions and taking random actions. We record such trajectories, store them, and select trajectories or segments to query synthetic or real humans. Preference learning algorithms update the posterior belief of the reward function after each query. We use cosine similarity as the alignment score to measure the distance between the estimated reward weights $\hat{\theta}$ and true weights $\theta$ [35, 41].

We compare DRPL with three baselines: (a) full trajectory query: randomly select two full trajectories to query; (b) random fragment query: randomly select two full trajectories, and cut one fragment out of each with a length equal to the average length of DRPL queries; (c) DRPL-SS: a version of our algorithm that selects a pair of trajectories associated with the *same* abstract state, instead of two abstract states that differ in one dimension. Further details can be found in Appendix 3.

Results for simulation (averaged across 5 seeds with 5 different true rewards) are shown in Fig. 5. DRPL (blue) converges to an alignment score around or greater than 0.8 for all tasks and outperforms all other algorithms. DRPL-SS performs better than the other baselines but is worse than DRPL, indicating that both components of our DRPL (segmentation and selection) are important.

Results for real robot experiments with humans (six for each task) are shown in Fig. 6. We omit the full trajectory baseline due to the time required to compare long trajectories generated by the random

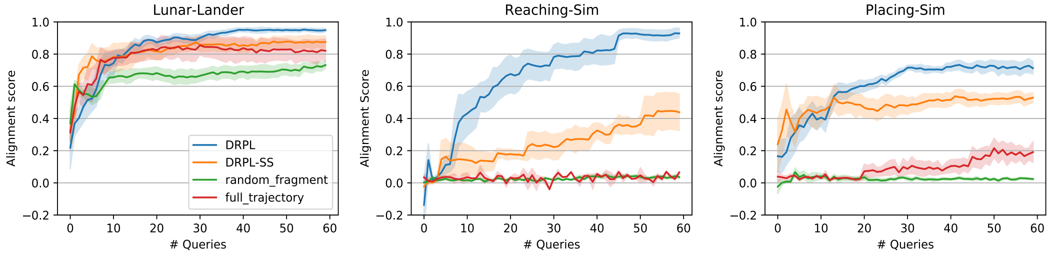
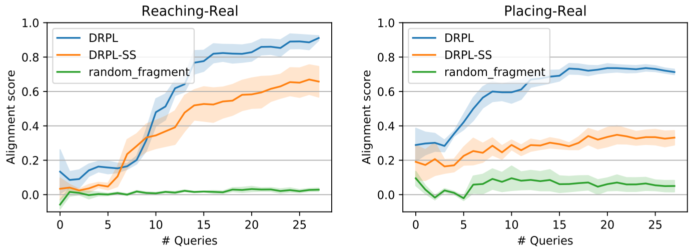


Figure 5: Alignment scores [35, 41] for `Lunar-Lander`, `Reaching-Sim`, and `Placing-Sim`. DRPL performs the best upon convergence. Error bars indicate the standard error of the means ($n = 5$).



| Reaching | RF | DRPL-SS | DRPL |
|---|---|---|---|
| E↑ | 3.2±0.2 | 3.2±0.2 | **3.3±0.2** |
| I↑ | 1.5±0.2 | 3.8±0.3 | **4.5±0.3** |
| C↑ | 3.3±0.3 | 3.2±0.2 | **4.0±0.0** |
| Placing | RF | DRPL-SS | DRPL |
| E↑ | 3.0±0.0 | 3.2±0.2 | **3.3±0.2** |
| I↑ | 1.5±0.3 | 2.7±0.4 | **3.8±0.2** |
| C↑ | 3.2±0.3 | **3.3±0.2** | **3.3±0.2** |

Figure 6: Alignment scores [35, 41] for `Reaching-Real` and `Placing-Real`. DRPL achieves the best performance upon convergence. Error bars indicate the standard error of the means ($n = 6$). Post-completion survey (5-point Likert scale) results indicate that DRPL leads to better perceived intelligence. E: overall experience, I: perceived intelligence, C: cognitive ease. See Appendix 4.3 and 4.4 for survey design and additional analyses.

agents. Our algorithm DRPL (shown in blue) achieves better performance than both baselines. Similar to evaluative feedback, the user experience survey (Fig. 6) shows that DRPL leads to better perceived intelligence (see Appendix 4.3 and 4.4). To conclude, the results support our key insight about preference learning: abstract state–based trajectory segmentation and query lead to a better learning outcome. DRPL is an instantiation of this idea within the dual representation framework.

## 5 Discussion

To summarize, we have proposed a dual representation framework for robot learning from human guidance. In the context of robot learning, abstract representation has long been a research topic. However, adopting it in decision learning often comes with a price: the critical information needed to learn a good policy may be lost in the process of abstraction. Researchers typically avoid this problem by having a hierarchical representation. The key difference is that we only use abstract representation as an auxiliary representation.

We show that the abstract scene graph representation allows us to utilize important heuristics that facilitate training in two popular forms of human guidance: evaluation and preference, both of which utilize human evaluations for observed agent policies. Evaluative feedback targets state-action pairs that are fine-grained, while preference learning targets trajectories that could be too coarse. Hence the former needs a grouping mechanism for generalization, and the latter needs a segmentation mechanism for efficient queries. Our proposed framework is a unified approach to provide both.

We demonstrate the effectiveness of our approaches in five challenging continuous control tasks. Our algorithms show significant improvements in performance and reduction in human effort, compared to algorithms without an auxiliary scene graph representation. These improvements make learning from human guidance methods significantly more appealing for real-robot learning.

**Limitations.** Currently, our proposed approaches leverage expert-defined scene graphs. Although this representation should be closer to the human representation, the actual representations are likely to be different across human trainers. Adapting to different abstract representations may further improve learning outcomes and user experience.

**Acknowledgement**

The work is in part supported by the Stanford Institute for Human-Centered AI (HAI), ONR MURI N00014-22-1-2740, ONR MURI N00014-21-1-2801, Amazon, Analog, IBM, JPMC, Meta, Salesforce, and Samsung. Ruohan Zhang is supported by Wu Tsai Human Performance Alliance Fellowship.

## References

[1] A. L. Thomaz, C. Breazeal, et al. Reinforcement learning with human teachers: Evidence of feedback and guidance with implications for learning performance. In *Aaai*, volume 6, pages 1000–1005. Boston, MA, 2006.

[2] R. Zhang, F. Torabi, L. Guan, D. H. Ballard, and P. Stone. Leveraging human guidance for deep reinforcement learning tasks. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 6339–6346. AAAI Press, 2019.

[3] R. Zhang, F. Torabi, G. Warnell, and P. Stone. Recent advances in leveraging human guidance for sequential decision-making tasks. *Autonomous Agents and Multi-Agent Systems*, 35(2): 1–39, 2021.

[4] H. B. Suay and S. Chernova. Effect of human guidance and state space size on interactive reinforcement learning. In *2011 Ro-Man*, pages 1–6. IEEE, 2011.

[5] C. Basu, M. Singhal, and A. D. Dragan. Learning from richer human guidance: Augmenting comparison-based learning with feature queries. In *2018 13th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 132–140. IEEE, 2018.

[6] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[7] S. Schaal. Is imitation learning the route to humanoid robots? *Trends in cognitive sciences*, 3 (6):233–242, 1999.

[8] B. D. Argall, S. Chernova, M. Veloso, and B. Browning. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5):469–483, 2009.

[9] T. Osa, J. Pajarinen, G. Neumann, J. A. Bagnell, P. Abbeel, J. Peters, et al. An algorithmic perspective on imitation learning. *Foundations and Trends® in Robotics*, 7(1-2):1–179, 2018.

[10] C. Isbell, C. R. Shelton, M. Kearns, S. Singh, and P. Stone. A social reinforcement learning agent. In *Proceedings of the fifth international conference on Autonomous agents*, pages 377–384. ACM, 2001.

[11] A. C. Tenorio-Gonzalez, E. F. Morales, and L. Villaseñor-Pineda. Dynamic reward shaping: training a robot by voice. In *Ibero-American conference on artificial intelligence*, pages 483–492. Springer, 2010.

[12] S. Griffith, K. Subramanian, J. Scholz, C. L. Isbell, and A. L. Thomaz. Policy shaping: Integrating human feedback with reinforcement learning. In *Advances in neural information processing systems*, pages 2625–2633, 2013.

[13] T. Cederborg, I. Grover, C. L. Isbell, and A. L. Thomaz. Policy shaping with human teachers. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pages 3366–3372. AAAI Press, 2015.

[14] W. B. Knox and P. Stone. Interactively shaping agents via human reinforcement: The tamer framework. In *Proceedings of the fifth international conference on Knowledge capture*, pages 9–16. ACM, 2009.

[15] J. MacGlashan, M. K. Ho, R. Loftin, B. Peng, G. Wang, D. L. Roberts, M. E. Taylor, and M. L. Littman. Interactive learning from policy-dependent human feedback. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2285–2294. JMLR. org, 2017.

[16] G. Warnell, N. Waytowich, V. Lawhern, and P. Stone. Deep tamer: Interactive agent shaping in high-dimensional state spaces. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

[17] D. Arumugam, J. K. Lee, S. Saskin, and M. L. Littman. Deep reinforcement learning from policy-dependent human feedback. *arXiv preprint arXiv:1902.04257*, 2019.

[18] W. B. Knox and P. Stone. Combining manual feedback with subsequent mdp reward signals for reinforcement learning. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1*, pages 5–12. International Foundation for Autonomous Agents and Multiagent Systems, 2010.

[19] W. B. Knox and P. Stone. Reinforcement learning from simultaneous human and mdp reward. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 475–482. International Foundation for Autonomous Agents and Multiagent Systems, 2012.

[20] R. Arakawa, S. Kobayashi, Y. Unno, Y. Tsuboi, and S.-i. Maeda. Dqn-tamer: Human-in-the-loop reinforcement learning with intractable feedback. *arXiv preprint arXiv:1810.11748*, 2018.

[21] Y. Cui, Q. Zhang, A. Allievi, P. Stone, S. Niekum, and W. B. Knox. The empathic framework for task learning from implicit human feedback. *arXiv preprint arXiv:2009.13649*, 2020.

[22] T. Kessler Faulkner, R. A. Gutierrez, E. S. Short, G. Hoffman, and A. L. Thomaz. Active attention-modified policy shaping: Socially interactive agents track. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '19, pages 728–736, Richland, SC, 2019. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 978-1-4503-6309-9. URL http://dl.acm.org/citation.cfm?id=3306127.3331762.

[23] G. Li, S. Whiteson, W. B. Knox, and H. Hung. Using informative behavior to increase engagement while learning from human reward. *Autonomous agents and multi-agent systems*, 30(5): 826–848, 2016.

[24] J. Grizou, I. Iturrate, L. Montesano, P.-Y. Oudeyer, and M. Lopes. Interactive learning from unlabeled instructions. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, pages 290–299, 2014.

[25] R. Loftin, B. Peng, J. MacGlashan, M. L. Littman, M. E. Taylor, J. Huang, and D. L. Roberts. Learning behaviors via human-delivered discrete feedback: modeling implicit feedback strategies to speed up learning. *Autonomous agents and multi-agent systems*, 30(1):30–59, 2016.

[26] L. Guan, M. Verma, S. S. Guo, R. Zhang, and S. Kambhampati. Widening the pipeline in human-guided reinforcement learning with explanation and context-aware data augmentation. *Advances in Neural Information Processing Systems*, 34:21885–21897, 2021.

[27] A. Najar, O. Sigaud, and M. Chetouani. Interactively shaping robot behaviour with unlabeled human instructions. *Auton. Agents Multi Agent Syst.*, 34(2):35, 2020.

[28] A. Wilson, A. Fern, and P. Tadepalli. A bayesian approach for policy learning from trajectory preference queries. In *Advances in neural information processing systems*, pages 1133–1141, 2012.

[29] R. Busa-Fekete, B. Szörényi, P. Weng, W. Cheng, and E. Hüllermeier. Preference-based evolutionary direct policy search. In *ICRA Workshop on Autonomous Learning*, 2013.

[30] J. Fürnkranz, E. Hüllermeier, W. Cheng, and S.-H. Park. Preference-based reinforcement learning: a formal framework and a policy iteration algorithm. *Machine learning*, 89(1-2):123–156, 2012.

[31] C. Wirth, J. Fürnkranz, and G. Neumann. Model-free preference-based reinforcement learning. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 2222–2228, 2016.

[32] R. Akrour, M. Schoenauer, M. Sebag, and J.-C. Souplet. Programming by feedback. In *International Conference on Machine Learning (ICML)*, volume 32, pages 1503–1511. JMLR. org, 2014.

[33] C. Wirth, R. Akrour, G. Neumann, and J. Fürnkranz. A survey of preference-based reinforcement learning methods. *The Journal of Machine Learning Research*, 18(1):4945–4990, 2017.

[34] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, pages 4299–4307, 2017.

[35] D. Sadigh, A. D. Dragan, S. Sastry, and S. A. Seshia. Active preference-based learning of reward functions. In *Robotics: Science and Systems*, 2017.

[36] A. Bestick, R. Pandya, R. Bajcsy, and A. D. Dragan. Learning human ergonomic preferences for handovers. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–9. IEEE, 2018.

[37] Y. Cui and S. Niekum. Active reward learning from critiques. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6907–6914. IEEE, 2018.

[38] M. Palan, N. C. Landolfi, G. Shevchuk, and D. Sadigh. Learning reward functions by integrating human demonstrations and preferences. In *Proceedings of Robotics: Science and Systems (RSS)*, June 2019. doi:10.15607/rss.2019.xv.023.

[39] E. Bıyık, D. P. Losey, M. Palan, N. C. Landolfi, G. Shevchuk, and D. Sadigh. Learning reward functions from diverse sources of human feedback: Optimally integrating demonstrations and preferences. *arXiv preprint arXiv:2006.14091*, 2020.

[40] L. M. Zintgraf, D. M. Roijers, S. Linders, C. M. Jonker, and A. Nowé. Ordered preference elicitation strategies for supporting multi-objective decision making. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1477–1485. International Foundation for Autonomous Agents and Multiagent Systems, 2018.

[41] E. Biyik and D. Sadigh. Batch active preference-based learning of reward functions. In *Conference on Robot Learning*, pages 519–528, 2018.

[42] E. Bıyık, D. A. Lazar, D. Sadigh, and R. Pedarsani. The green choice: Learning and influencing human decisions on shared roads. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 347–354. IEEE, 2019.

[43] E. Biyik, N. Huynh, M. J. Kochenderfer, and D. Sadigh. Active preference-based gaussian process regression for reward learning. In *Proceedings of Robotics: Science and Systems (RSS)*, July 2020. doi:10.15607/rss.2020.xvi.041.

[44] F. Memarian, Z. Xu, B. Wu, M. Wen, and U. Topcu. Active task-inference-guided deep inverse reinforcement learning. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 1932–1938. IEEE, 2020.

[45] H. Sikchi, A. Saran, W. Goo, and S. Niekum. A ranking game for imitation learning. *arXiv preprint arXiv:2202.03481*, 2022.

[46] A. Jain, S. Sharma, T. Joachims, and A. Saxena. Learning preferences for manipulation tasks from online coactive feedback. *The International Journal of Robotics Research*, 34(10):1296–1313, 2015.

[47] A. L. Thomaz and C. Breazeal. Teachable robots: Understanding human teaching behavior to build more effective robot learners. *Artificial Intelligence*, 172(6-7):716–737, 2008.

[48] O. Amir, E. Kamar, A. Kolobov, and B. J. Grosz. Interactive teaching strategies for agent training. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 804–811. AAAI Press, 2016.

[49] M. K. Ho, M. Littman, J. MacGlashan, F. Cushman, and J. L. Austerweil. Showing versus doing: Teaching by demonstration. In *Advances in neural information processing systems*, pages 3027–3035, 2016.

[50] D. H. Uttal, K. O'Doherty, R. Newland, L. L. Hand, and J. DeLoache. Dual representation and the linking of concrete and symbolic representations. *Child Development Perspectives*, 3(3): 156–159, 2009.

[51] J. S. DeLoache. Dual representation and young children's use of scale models. *Child development*, 71(2):329–338, 2000.

[52] D. Kahneman. *Thinking, fast and slow*. Macmillan, 2011.

[53] M. K. Ho, D. Abel, C. G. Correa, M. L. Littman, J. D. Cohen, and T. L. Griffiths. People construct simplified mental representations to plan. *Nature*, 606(7912):129–136, 2022.

[54] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. Shamma, M. Bernstein, and L. Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3668–3678, 2015.

[55] I. Armeni, Z.-Y. He, J. Gwak, A. R. Zamir, M. Fischer, J. Malik, and S. Savarese. 3d scene graph: A structure for unified semantics, 3d space, and camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5664–5673, 2019.

[56] Z. Ravichandran, L. Peng, N. Hughes, J. D. Griffith, and L. Carlone. Hierarchical representations and explicit memory: Learning effective navigation policies on 3d scene graphs using graph neural networks. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2022.

[57] N. Hughes, Y. Chang, and L. Carlone. Hydra: A real-time spatial perception engine for 3d scene graph construction and optimization. *arXiv preprint arXiv:2201.13360*, 2022.

[58] S. Ross, G. J. 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*, pages 627–635, 2011.

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

[60] A. Mandlekar, D. Xu, R. Martín-Martín, Y. Zhu, L. Fei-Fei, and S. Savarese. Human-in-the-loop imitation learning using remote teleoperation. *arXiv preprint arXiv:2012.06733*, 2020.

[61] A. Bobu, M. Wiggert, C. Tomlin, and A. D. Dragan. Inducing structure in reward learning by learning features. *The International Journal of Robotics Research*, page 02783649221078031, 2022.

[62] T. Silver, R. Chitnis, J. Tenenbaum, L. P. Kaelbling, and T. Lozano-Pérez. Learning symbolic operators for task and motion planning. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3182–3189. IEEE.

[63] T. Silver, A. Athalye, J. B. Tenenbaum, T. Lozano-Perez, and L. P. Kaelbling. Learning neuro-symbolic skills for bilevel planning. *arXiv preprint arXiv:2206.10680*, 2022.

[64] R. Chitnis, T. Silver, J. B. Tenenbaum, T. Lozano-Perez, and L. P. Kaelbling. Learning neuro-symbolic relational transition models for bilevel planning. *arXiv preprint arXiv:2105.14074*, 2021.

[65] J. Achterhold, M. Krimmel, and J. Stueckler. Learning temporally extended skills in continuous domains as symbolic actions for planning. *arXiv preprint arXiv:2207.05018*, 2022.

[66] A. D. Dragan and S. S. Srinivasa. *Formalizing assistive teleoperation*. MIT Press, July, 2012.

[67] N. Amirshirzad, A. Kumru, and E. Oztop. Human adaptation to human–robot shared control. *IEEE Transactions on Human-Machine Systems*, 49(2):126–136, 2019.

[68] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, 29, 2016.

[69] R. T. Chen, X. Li, R. B. Grosse, and D. K. Duvenaud. Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 31, 2018.

[70] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016.

[71] D. Brown, R. Coleman, R. Srinivasan, and S. Niekum. Safe imitation learning via fast bayesian reward inference from preferences. In *International Conference on Machine Learning*, pages 1165–1177. PMLR, 2020.

[72] D. Abel, D. Arumugam, K. Asadi, Y. Jinnai, M. L. Littman, and L. L. Wong. State abstraction as compression in apprenticeship learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3134–3142, 2019.

[73] C. Agia, K. M. Jatavallabhula, M. Khodeir, O. Miksik, V. Vineet, M. Mukadam, L. Paull, and F. Shkurti. Taskography: Evaluating robot task planning over large 3d scene graphs. In *Conference on Robot Learning*, pages 46–58. PMLR, 2022.

[74] Y. Zhu, J. Tremblay, S. Birchfield, and Y. Zhu. Hierarchical planning for long-horizon manipulation with geometric and symbolic scene graphs. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6541–6548. IEEE, 2021.

[75] S. Nguyen, O. S. Oguz, V. N. Hartmann, and M. Toussaint. Self-supervised learning of scene-graph representations for robotic sequential manipulation planning. In *CoRL*, pages 2104–2119, 2020.

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

[77] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel, et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018.

[78] E. Bıyık, A. Talati, and D. Sadigh. Aprel: A library for active preference-based reward learning algorithms. *arXiv preprint arXiv:2108.07259*, 2021.

[79] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

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