Mirroring without Overimitation: Learning Functionally Equivalent Manipulation Actions

Hangxin Liu[1] and **Chi Zhang**[1,2] and **Yixin Zhu**[1,2] and **Chenfanfu Jiang**[3] and **Song-Chun Zhu**[1,2]
{hx.liu,chi.zhang,yixin.zhu}@ucla.edu, cffjiang@seas.upenn.edu, sczhu@stat.ucla.edu
[1] UCLA Center for Vision, Cognition, Learning and Autonomy
[2] International Center for AI and Robot Autonomy (CARA)
[3] UPenn Computer and Information Science Department

Abstract

This paper presents a *mirroring* approach, inspired by the neuroscience discovery of the mirror neurons, to transfer demonstrated manipulation actions to robots. Designed to address the different embodiments between a human (demonstrator) and a robot, this approach extends the classic robot Learning from Demonstration (LfD) in the following aspects: i) It incorporates fine-grained hand forces collected by a tactile glove in demonstration to learn robot's fine manipulative actions; ii) Through model-free reinforcement learning and grammar induction, the demonstration is represented by a goal-oriented grammar consisting of goal states and the corresponding forces to reach the states, independent of robot embodiments; iii) A physics-based simulation engine is applied to emulate various robot actions and mirrors the actions that are *functionally equivalent* to the human's in the sense of causing the same state changes by exerting similar forces. Through this approach, a robot *reasons* about which forces to exert and what goals to achieve to generate actions (*i.e.*, mirroring), rather than strictly mimicking demonstration (*i.e.*, overimitation). Thus the embodiment difference between a human and a robot is naturally overcome. In the experiment, we demonstrate the proposed approach by teaching a real Baxter robot with a complex manipulation task involving haptic feedback—opening medicine bottles.

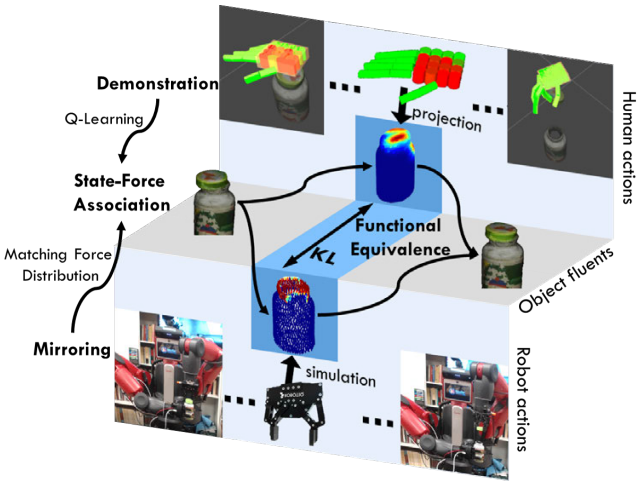


Figure 1: A robot mirrors human demonstrations with functional equivalence by inferring the action that produces similar force, resulting in similar changes of the physical states. Q-Learning is applied to associate types of forces with the categories of the object state changes to produce human-object-interaction (*hoi*) units.

1 Introduction

A hallmark of machine intelligence is the capability to adapt to new tasks rapidly and "achieve goals in a wide range of environments" (Legg and Hutter 2007). In comparison, a human can quickly learn new skills by observing other individuals, expanding their repertoire swiftly to adapt to the ever-changing environment. To emulate the similar learning process, the robotics community has been developing the framework of *Learning from Demonstration* (LfD) (Argall et al. 2009), some of which has shown promising results.

However, the "correspondence problem" (Dautenhahn and Nehaniv 2002), *i.e.*, the difference of embodiments between a human and a robot, is rarely addressed in the prior work of LfD. As a result, a one-to-one mapping is usually handcrafted between the human demonstration and the robot execution, restricting the LfD only to mimic the demonstrator's low-level motor controls and replicate the (almost) identical procedure to achieve the goal. Such behavior is analogous to a phenomenon called "overimitation" (Lyons, Young, and Keil 2007) observed in human children. Therefore, the acquired skills can hardly be adapted to new robots or new situations, demanding better solutions.

Meanwhile, the neuroscience discovery of mirror neurons in primate (Gallese et al. 1996) showed that the mirror neuron system facilitates imitation learning in macaque monkey and human (Rizzolatti, Fogassi, and Gallese 2001): a mirror neuron fires when a primate performs a *goal-directed* action or sees others performing the same action. Further findings verify that human mirror neurons are activated during the observation of robot task performance (Gazzola et al. 2007; Oberman et al. 2007) even with different embodiments.

Inspired by the mirror neurons, we propose a *mirroring* approach that extends the current LfD, through the physics-based simulation, to address the correspondence problem. Rather than overimitating the motion controls from the demonstration, it is advantageous for the robot to seek *func-*

tionally equivalent but possibly visually different actions that can produce the same effect and achieve the same goal as those in the demonstration. In particular, our approach has three characteristics compared to the standard LfD.

- *Force-based*: We deploy a low-cost tactile glove to collect human demonstration with fine-grained manipulation forces. Beyond visually observable space, these tactile-enabled demonstrations capture a deeper understanding of the physical world that a robot interacts with, providing an extra dimension to address the correspondence problem.
- *Goal-oriented*: A "goal" is defined as the desired state of the target object and is encoded in a grammar model. The terminal node of the grammar model is the state changes caused by the forces, independent of the embodiments.
- *Mirroring without overimitation*: Different from the classic LfD, a robot does not necessarily mimic every action in the human demonstration. Instead, the robot reasons about the action to achieve the goal states based on the learned grammar and the simulated forces.

To validate the proposed approach, we *mirror* the human manipulation actions of opening medicine bottles with a child-safety lock to a real Baxter robot. The challenge in this task lies in the fact that opening such bottles requires to push or squeeze various parts, which is visually similar to opening one without a child-safe lock. Figure 1 outlines the *mirroring* approach with *functional equivalence*. Specifically, we explicitly model the forces on the object exerted by the hand in the demonstration with a pose and force sensing tactile glove. The collected distribution of the forces on the object is compared to a set of the force distributions exerted by the robot gripper on the same object in a physics-based simulator. Simulated actions with sufficiently small KullbackLeibler (KL) divergence with respect to the demonstration are considered *functionally equivalent*, thus hinting this action would be the best robot action to accomplish the task.

Our contributions are three-fold. First, we extend the classical LfD to a *mirroring* approach represented by a goal-oriented grammar to overcome the differences between embodiments. Second, we allow a robot itself to reason about *functionally equivalent* actions, instead of overimitating demonstrations. Third, we show that the proposed system performs well in a complex manipulation task of opening medicine bottles.

1.1 Related Work

Mirror Neurons or the mirror neuron system (MNS) has been found and proven to play an essential role in human action recognition, understanding, and imitations (Rizzolatti and Craighero 2004; Thill et al. 2013). These findings motivate several neural-network-based computational models, which primarily study grasping actions and are validated by hand stimuli (Oztop and Arbib 2002; Bonaiuto, Rosta, and Arbib 2007). In these cases, the studies were strictly confined by hands' relative position to an object, and little embodiment difference was presented. In parallel work, Ito *et al*. (Ito and Tani 2004) adopted a Recurrent Neural Network to model MNS. Although it extends hand-object relation to the body movements, it is still restricted to similar embodiments. Further, these models lack a deeper understanding (*e.g*., the goal) of the demonstration. In contrast, the proposed *mirroring* approach emphasizes the intent of the demonstration as changing the target object to desired states regardless of the embodiment.

LfD contains a vast amount of literature; we refer the readers to two surveys (Argall et al. 2009; Osa et al. 2018). Here, we only highlight some exemplary work that is closely related, specifically on how to address the correspondence problem. Defining a set of task-level actions on robots (Konidaris et al. 2012; Niekum et al. 2015; Edmonds et al. 2017) omits the correspondence as a robot only learns action scheduling from the set based on the demonstration. By manually defining the keypoints between the demonstrator and the robot, keypoint-based methods (Koenemann, Burget, and Bennewitz 2014; Shu et al. 2017) are capable of mapping between different embodiments but with limited flexibility. Trajectory-based methods are more favorable since the robot's motion planner handles the embodiments difference. Specifically, the robot end-effector's trajectory is either directly mapped to demonstrator's trajectory (Pastor et al. 2009; Yang et al. 2015) or indirectly mapped using trajectory optimization methods (Maeda et al. 2016). However, these approaches fall short of complex manipulation due to the lack of haptic information.

Force-based Demonstration exists in prior work of LfD. Lin et al. (2012) utilized fingertip force for grasping but did not address the correspondence problem. Although using kinesthetic teaching methods (Kormushev, Calinon, and Caldwell 2011; Montebelli, Steinmetz, and Kyrki 2015; Manschitz et al. 2016; Racca et al. 2016) was capable of incorporating forces into the demonstrations for in-contact tasks, it is still difficult to transfer the demonstration to a different embodiment. Furthermore, kinesthetic teaching is hard to design for fine interactive tasks, *e.g*., opening medicine bottles with a child-safety lock.

Policy Search Methods use human demonstrations as the initial policy to constrain the search space (Kober and Peters 2009), and reinforcement learning is usually applied to derive a control policy, circumventing the correspondence problem. These methods have succeeded in robot's constrained reaching (Guenter et al. 2007), locomotion (Theodorou, Buchli, and Schaal 2010), grasping (Prieur, Perdereau, and Bernardino 2012) and soft hand controlling (Gupta et al. 2016). To avoid being confined by the human demonstration, Levine and Abbeel (2014; 2015) uses guided policy search for robot manipulations. However, policy search methods have not yet demonstrated successful applications in very complex tasks, especially those where unobservable/latent information (*e.g*., force) plays a vital role.

Inverse Reinforcement Learning (IRL) or inverse optimal control (Ng, Russell, and others 2000; Abbeel and Ng 2004; Ramachandran and Amir 2007; Ziebart et al. 2008) gains increasing interests in robotics community. Although it alleviates the need for reward engineering by inferring the reward/objective function from demonstrations, IRL has not been shown to scale to the same complexity of tasks as direct imitation learning, since there may exist many optimal policies that can explain a set of given demonstra-

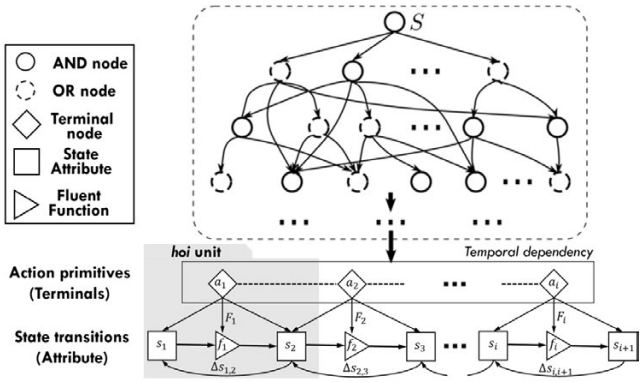


Figure 2: Illustration of a T-AOG. The T-AOG is a temporal grammar in which the terminal nodes are the *hoi* units. An *hoi* unit (shown in the grey area) contains a single action a_i that transits the state from the pre-condition s_i to the post-condition s_{i+1}. The fluents function f_i represents the changes of the physical state s_i on object caused by the forces F_i exerted by the action a_i: $s_{i+1} = f_i(s_i, a_i; F_i)$.

tions (Ng, Harada, and Russell 1999). This challenge is often magnified by task complexity, making it computationally highly expensive (MacGlashan and Littman 2015). Our approach is partially similar to IRL in the sense that it recovers the action-state relations from the demonstration. Instead of learning only from the demonstration, we deploy a physics-based simulation to generate feasible motions.

2 Representation

We represent the action sequence to execute a task by a structural grammar model *Temporal And-Or Graph (T-AOG)* (Zhu and Mumford 2007) (see Figure 2). A T-AOG is a directed graph which describes a stochastic context-free grammar (SCFG), encoding both a hierarchical and a compositional representation. Formally, a T-AOG is defined as a five-tuple $G = (S, V, R, P, \Sigma)$. Specifically,

- S is the start symbol that represents an event category (*e.g.*, opening a bottle).
- V is a set of nodes including non-terminal nodes V^{NT} and terminal nodes V^T: $V = V^{NT} \cup V^T$.
- The **non-terminal** nodes can be divided into And-nodes and Or-nodes: $V^{NT} = V^{AND} \cup V^{OR}$. And-nodes V^{AND} represent the compositional relations: a node v is an And-node if the entity represented by v can be decomposed into multiple parts represented by its child nodes. Or-nodes V^{OR} indicate the alternative configuration among its child nodes: a node v is an Or-node if the entity represented by v has multiple mutually exclusive configurations represented by its child nodes.
- The **terminal** nodes V^T are the entities that cannot be further decomposed or do not have different configurations. For a T-AOG, the terminal nodes represent the *human-object-interaction (hoi)* units (Johnson-Frey et al. 2003). An *hoi* unit encodes actions a_i that an agent can perform (*e.g.* grasp, twist), the spatiotemporal relations between the object and the agent's hand, and how the force F_i produced by such primitive causes the changes of physical states on the object.
- $R = \{r : \alpha \to \beta\}$ is a set of production rules that represent the top-down sampling process from a parent node α to its child nodes β.
- $P : p(r) = p(\beta|\alpha)$ is the probability associated with each production rule.
- Σ is the language defined by the grammar, *i.e.*, the set of all valid sentences that can be generated by the grammar.

A **parse tree** *pt* is an instance of the T-AOG, where one of the child nodes is selected for each Or-node. The terminal nodes of a *pt* form a valid sentence; in this case, terminal nodes are a set of *hoi* units consisting of the actions for an agent to execute in a fixed order, as well as the state changes after performing such an action sequence.

3 Force-based Goal-oriented Mirroring

3.1 Learning Force and State Associations as *hoi*

To transfer across different embodiment, we need to know the effect of a particular type of forces so that the desired action can be planned, requiring to investigate the state changes caused by the forces. We cast this problem in a reinforcement learning framework to learn a policy that associates forces and state changes. The state space and the action (force) space from human demonstrations are discretized and quantized, and an iterative Q-Learning scheme is applied. We believe the proposed learning framework does not lose generality since one can scale up the process to continuous state space or action space by using DQN (Mnih et al. 2015) or other advanced policy gradient methods.

Categorize Force The pose and force data of human demonstrations were collected using a tactile glove. See Sec. 4.1 for details. The forces exerted by a human hand, together with the poses, are projected onto the mesh of the object. Formally,

$$F_t^o = g(a_t^h(F_t^h, p_t^h)), \qquad t \in \{1, 2, \ldots, n\} \tag{1}$$

where t is the frame index, and n is the total number of frames. g is an implicit projection function that maps a human action a_t^h, parameterized by the force exerted F_t^h and the pose p_t^h, to F_t^o the force projected on the object mesh.

Each element in the resulting force F_t^o is a 4-dimensional vector, where the first three dimensions represent the position of one object surface vertex and the fourth dimension the force magnitude on this vertex.

K-means clustering (Kanungo et al. 2002) is adopted to categorize the force F_i^o into N types, *i.e.*,

$$l_k = c(F_t^o), \qquad t \in \{1, 2, \ldots, n\}, k \in \{1, \ldots, N\} \tag{2}$$

where $c(\cdot)$ denotes the clustering function and l_k is the label of the k-th cluster type. After assigning labels to each frame, we aggregate the frames with the same label into a segment and take the average,

$$F_k = \text{avg}(F_t^o), \qquad \forall t, c(F_t^o) = l_k. \tag{3}$$

The segments form a discretized action (force) sequence (Figure 3c) to complete the given task.

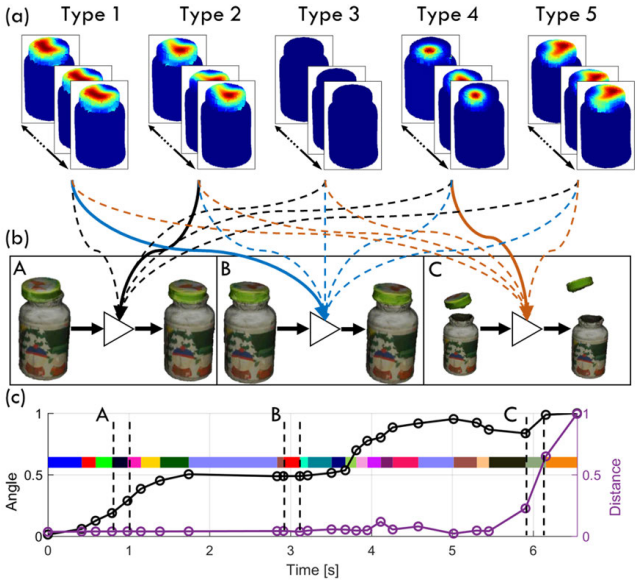


Figure 3: Force and state associations as *hoi* units. The manipulation force is clustered into 21 types. (a) Five examples of force types, in which Type 3 has no force. (c) Given the categorized force and quantized states based on the forces, (b) the Q-learning algorithm associates a force to a specific state change (A: lid is twisted; B: initiate contact; C: pull off the lid) shown by the solid lines. The dash lines indicate the forces that are incompatible to the given fluents functions, represented by the triangles.

Quantize State The relative poses can describe the states of a rigid target object under manipulation actions among object's parts, *e.g.*, bottle and lid, multiple Lego blocks, *etc.*. Without loss of generality, we use the relative distance and relative rotation angle between the lid and the bottle, which are derived from their relative poses, as our state space. As shown in Figure 3b, within each segment of the force (shown in color bars), we take the average of the corresponding angle and distance and normalize their magnitude to unit size,

$$s_i = \langle d_i, \theta_i \rangle \in [0,1]^2, \forall i \in \{1, \dots, M\} \tag{4}$$

where M is the total number of states, and d_i and θ_i denotes the relative distance and angle, respectively.

Associate Force and State as *hoi* Units by Q-Learning By replacing the actions in Q-learning with the labels of the force l_k, we adopt the tabular Q-Learning that associates the current state s_i to a force type using the iterative Q-Learning update rule in a temporal difference fashion,

$$Q(s_i, l_k) = (1-\alpha) \cdot Q(s_i, l_k) + \alpha \cdot \left[r(s_i, l_k) + \gamma \cdot \max_k Q(s_{i+1}, l_k) \right], \tag{5}$$

where r denotes the reward, Q the Q-function, α the learning rate, and γ the discount factor. Here, we assume the system dynamics is deterministic.

Inference We pick the best action according to the Q-function $l_* = \arg\max_k Q(s_i, l_k)$. The association among s_i, s_{i+1} and corresponding F_k naturally forms an *hoi* unit (see Figure 2) and will be used for learning a goal-oriented grammar discussed in the next section.

3.2 Learning Goal-Oriented Grammar

Grammar Induction Each successful demonstration contributes a sequence of *hoi* units that encode the types of forces and the state evolvement. We induce a T-AOG $\mathcal{G}$ from multiple demonstrations using a modified version of Automatic Distillation of Structure (ADIOS) algorithm presented in (Qi et al. 2017). The objective function is the posterior probability of the grammar given the training data X,

$$p(G|X) \propto p(G)p(X|G) = \frac{1}{Z} e^{-\alpha||G||} \prod_{pt_i \in X} p(pt_i|G), \tag{6}$$

where $pt_i = (hoi_1, hoi_2, \dots, hoi_m) \in X$ represents a valid parse graph of *hoi* units with length m.

Action Sequence Sampling To generate a valid sentence, *i.e.*, a parse tree $pt = (hoi_0, \dots, hoi_K)$, we sample T-AOG $\mathcal{G}$ by decomposing all the And-nodes and selecting one branch at each Or-node. This pt is *goal-oriented* in the sense that its terminal nodes $hoi_k \in pt$ encode the forces of reaching sub-goal states that are invariant across embodiments for the given task. Note that this process is non-Markovian, while the force-state association using Q-Learning is Markovian.

3.3 Mirroring to Robot without Overimitation

Simulation-based Action Synthesis Discrete robot action primitives are given by a dictionary $\Omega_{a^r} = \{a^r_1, \dots, a^r_M\}, M = 10$, parameterized by the change of end-effector poses, including moves in all six canonical directions, rotations in both clockwise and counter-clockwise directions, and opening/closing the gripper. The task of opening a medicine bottle can be accomplished by the combinatorics of the actions. Given a pt, we seek to generate a sequence of robot actions $\{a^r_i, i = 1, \dots, m\}$ that produce forces sufficient to cause the same changes of states as encoded in the sampled pt. In this sense, we say the robot action a^r_i is *functionally equivalent* to the demonstration action sequence a^h_i. Additionally, since the goal of the generated action sequences is to achieve the same effects, such generated action sequences can be different from the observed demonstrations and will not overimitate the observed ones.

A physics-based simulator (see Figure 4) is introduced to estimate the force exerted by the robot gripper on the bottle. We denote the force obtained from the simulator as F^{sim}_m, where m is the index of the robot primitives, and compare it to the corresponding F_k, the average force exerted by human demonstrations with label l_k. Formally, F_k and F^{sim}_m are formalized as distributions,

$$P(F_k) = \frac{1}{Z_k} F_k, \quad \text{and} \quad P(F^{sim}_m) = \frac{1}{Z^{sim}_m} F^{sim}_m, \tag{7}$$

where Z_k and Z^{sim}_m are the normalization factors, obtained by summing over the force magnitudes on all vertices of the object. The similarity of the two forces can be measured by the KL divergence, and the robot action is selected by

$$\begin{aligned} F^{sim}_* &= \arg\min_m \mathrm{KL}\left(P(F_k) \,\|\, P(F^{sim}_m)\right) \\ &= \arg\min_m \sum_v \left[P_{F_k}(v) \log \frac{P_{F_k}(v)}{P_{F^{sim}_m}(v)} \right], \end{aligned} \tag{8}$$

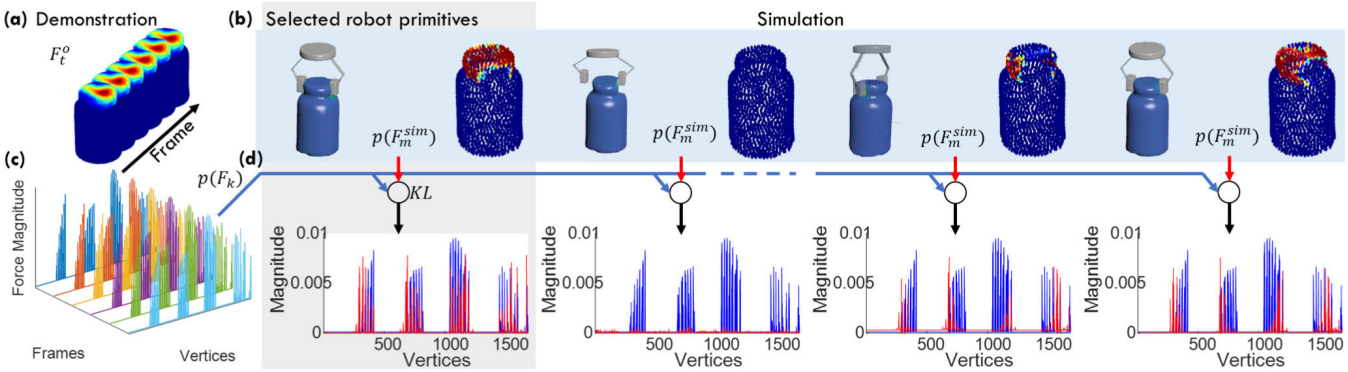


Figure 4: Based on the demonstrations, the force in the same cluster l_k produces a force distribution on the object F_t^o, and the average is the distribution of the force category F_k. Among the simulated force responses F_m^{sim} obtained from a physics-based simulator, the corresponding primitive of the most similar force, measured by the KL distance, is selected for the robot execution. (a) The forces in the same cluster. (b) The simulated robot primitives (downward, no contact, contact, and twist) and their force responses. (c) The force distributions of the same cluster in each frame. (d) The distributions of F_k against each simulated force distribution F_m^{sim}, denoted by blue and red, respectively.

where v is the vertex index on the object mesh. Once F_*^{sim} is selected, the robot would choose the corresponding primitive a_*^r that produces F_*^{sim}.

Physics-based Simulation The physics-based simulation needs to be able to capture intricate frictional contact between the robot gripper and the bottle. The total force applied at each point located at the surface of the bottle consists of several terms: the normal component of squeezing force from the gripper, the tangential component of static friction force from the gripper, the internal elastic force from the rest of the continuous bottle material and gravity.

The key to achieving such a force balance in the simulator is to model the deformation of the bottle. Various physical constitutive models and stress-strain relationships exist for polymers, and it is impractical for us to find the exact material parameters through mechanical tension or compression tests. Thus, we assume the deformation of the bottle is sufficiently far away from the plastic regime, and adopt a standard hyperelastic model: the Neo-Hookean model (Macosko 1994) to describe the mechanical stress under deformation

$$\mathbf{P} = \mu(\mathbf{F} - \mathbf{F}^{-T}) + \lambda \log(\det(\mathbf{F}))\mathbf{F}^{-T}, \tag{9}$$

where $\mathbf{F}$ is the deformation gradient tensor encoding the strain at each point, $\mathbf{P}$ is the first Piola-Kirchoff stress tensor describing its elastic mechanical stress, and μ, λ are material parameters describing the stiffness and incompressibility of the bottle, respectively. The governing equation describing the force balance of the bottle is given by

$$\nabla \cdot \mathbf{P} = \mathbf{f}^{ext}, \tag{10}$$

where $\mathbf{f}^{ext}$ denotes the total external force on the bottle.

We solve Equation 10 using the Finite Element Method (Bonet and Wood 1997). The input bottle geometry is first converted from a triangulated surface to a tetrahedralized volume using TetGen (Si 2015). The robot gripper mesh is converted into a watertight level set represented by OpenVDB (Museth et al. 2013), which allows natural treatment of frictional contact under arbitrary kinematic rigid motion. The additional parameters including friction coefficient, μ, and λ are set empirically. Once the discretized equation system is solved to convergence, we evaluate the force magnitude at each discrete point of the object surface mesh and store them in F_m^{sim}.

Execution Ideally, a robot can accomplish the task using the primitives obtained from the simulator (see Figure 5a). However, we may encounter two types of discontinuity issues in robot execution.

1. *Discontinuity in object state space.* The post-condition s_{i+1} of action a_i and the pre-condition $s_{i+1}{}'$ of the next action a_{i+1} are not necessarily the same from the sampled *pt*; *e.g.*, $s_{i+1} \neq s_j$ shown in Figure 5b. A discontinuity exists between two consecutive object states, thus an

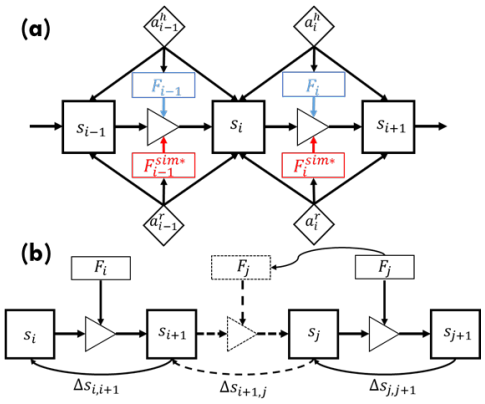


Figure 5: (a) A fragment of a *pt*. The forces F_{i-1} and F_i produced by human (in blue) change the bottle states from s_{i-1} to s_{i+1}. The robot action that produces the closest force distribution obtained by simulation (in red) is chosen and causes the same changes of object states. (b) If $s_{i+1} \neq s_j$, and $\Delta s_{i+1,j}$ is closer to $\Delta s_{j,j+1}$ than any other fluents, we assign the force F_j to change the state from s_{i+1} to s_j.

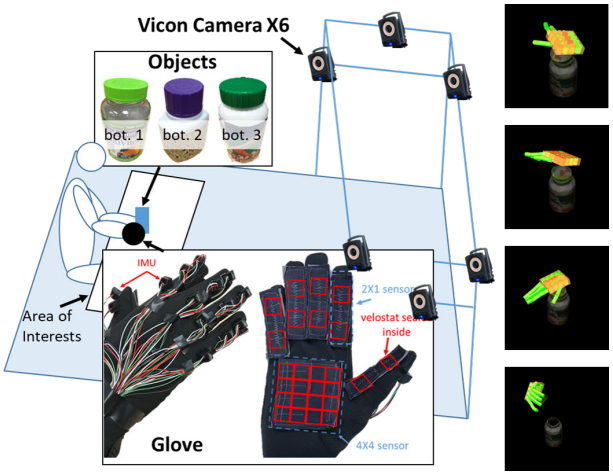


Figure 6: Data collection environment. A tactile glove is utilized to collect hand poses and forces, and the Vicon MoCap system for relative poses of hand and objects.

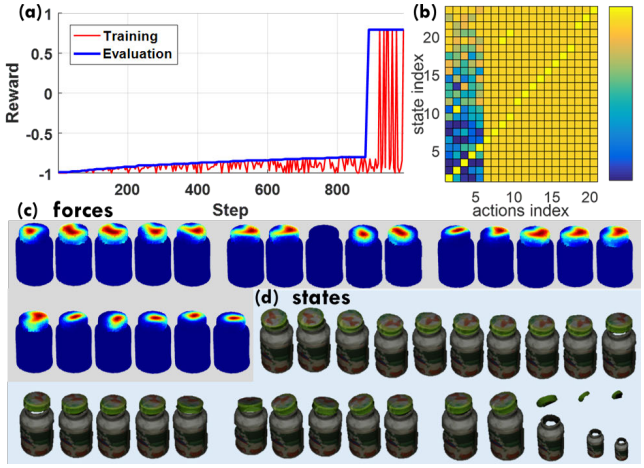


Figure 7: (a) The cumulative rewards during training and evaluation. (b) The landscape of the learned Q table, where yellow indicates high values and blue low. (c) The 21 types of actions (forces) by clustering in one exemplary demonstration. (d) The 25 discretized states based on the forces (some force types appear more than once).

additional primitive is needed. We define a subtraction operator, $\nabla_-(\cdot,\cdot)$ for two consecutive states such that

$$\Delta s_{i,j} = \nabla_-(s_i, s_j) = \langle d_j - d_i, \theta_j - \theta_i \rangle. \quad (11)$$

As shown in Figure 5b, by comparing the discontinuity $\Delta s_{i+1,j}$ with any other changes of states, we choose the most similar one in terms of L_2 norm. The corresponding type of force is assigned to interpolate the discontinuity.

2. *Discontinuity in robot action space.* We use B-Spline to fill in two discontinuous primitives assuming no obstacles, and the robot is able to follow the trajectory specified by the spline. Once the end-effector reaches the joint limit, it is set to restore to the initial position. For instance, if the generated primitive is rotating in the clockwise direction, reaching robot's joint limit, the robot will first rotate back to its natural pose before the execution.

4 Experiment

4.1 Preliminary

Robot Platform We exercise the proposed framework in a robot platform with a dual-armed 7-DoF Baxter robot mounted on a DataSpeed mobility base. The robot is equipped with a ReFlex TakkTile gripper on the right wrist and a Robotiq S85 parallel gripper on the left. The entire system runs on ROS, and the arm motion is planned by *MoveIt!*.

Dataset The hand pose and force data is collected using an open-sourced tactile glove (Liu et al. 2017) that is equipped with i) a network of 15 IMUs to measure the rotations between individual phalanxes, and ii) 6 customized force sensors using Velostat, a piezoresistive material, to record the force in two regions (proximal and distal) on each phalange and a 4×4 regions on palm. Figure 6 depicts the tactile glove and the data collection environment. The relative poses between the wrist of hand and object parts (*i.e.*, bottle, and lid) are obtained from Vicon. The data of 10 human manipulation sequences is collected, processed, and visualized using ROS.

4.2 Learning

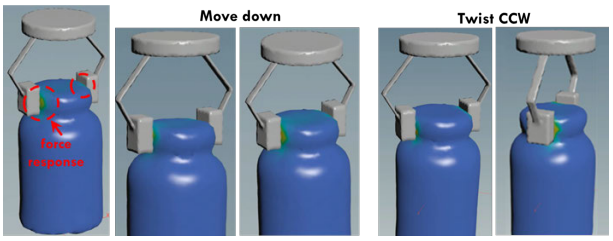
Figure 7a-b shows the Q-learning results, with a discount factor 0.99, reward for success $+1$, reward for failure -1, and reward for all others 0. We use ϵ-greedy exploration with exponential decay to obtain the state-force associations.

Figure 7a shows the cumulative reward during each training episode in red, and the average cumulative reward during evaluation in blue. During training, the cumulative reward generally increases until finding a path that leads to the maximum reward and begins fluctuating. This fluctuation happens due to the marginal probability of a non-optimal action being chosen at each step in ϵ-greedy exploration policy, even though an optimal path has been found. The evaluation is performed every ten episodes during training with a policy induced by the Q-table. During the evaluation, the reward monotonically increases slowly at first and jumps to the maximum, due to the optimal path found during training and the learning signals propagated into the Q-table. The policy induced from the Q-table converges to the optimum after approximately 900 episodes in training. The resulting Q-table is shown in Figure 7b.

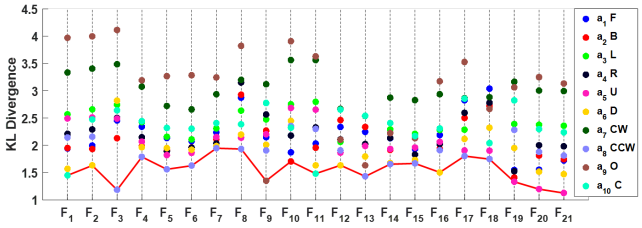
4.3 Robot Execution with Functional Equivalence

A *pt* is first sampled from the T-AOG induced from the learned policy to obtain a sequence of force types the robot should imitate in order to cause the same changes of object states. Our physics-based simulation then emulates a set of robot actions to obtain their force responses; some examples are shown in Figure 8a.

Figure 8b shows an example of a *pt* consisting of 21 *hoi* units (x-axis). The force responses of the ten robot primitives are simulated, and the similarities (y-axis) to the corresponding F_k are measured in each stage. The primitives with the lowest KL divergence (connected by the red line)



(a) Force simulation



(b) KL divergence for action primitives

Figure 8: (a) Simulations of the robot actions' force responses. (b) The KL divergence for all action primitives in a *pt*. In this case, the primitives are a_1 *move forward*, a_2 *move backward*, a_3 *move left*, a_4 *move right*, a_5 *move up*, a_6 *move down*, a_7 *rotate clockwise*, a_8 *rotate counter-clockwise*, a_9 *open gripper*, and a_{10} *close gripper*. The solid red line is the sequence of actions for a robot to execute.

are selected for robot execution.

The execution of a Baxter robot is shown in Figure 9a. It starts from an initial position and sequentially performs the corresponding primitives indicated in the grey area in the lower right corner. The a_6 *downward* primitive indeed generates forces which are captured by the force sensor (top left) in the robot wrist, which demonstrate that the *mirroring* approach indeed allows the robot to fulfill the challenging task of opening medicine bottles with a set of actions that are different from demonstrations.

The result also shows that the similarity between forces can be adequately measured by KL divergence to determine whether two actions are *functionally equivalent*. For instance, the primitive *opening the gripper* has the largest divergence in most of the cases as it produces no force to the object, except in F_9 when the demonstrator releases the lid after one rotation. The pressing force critical to our task is also captured and mirrored to robot well (see F_2 and F_{16} where a *downward* primitive is planned). Finally, *upward* primitives are selected to finish the task by pulling the lid.

4.4 Ablative Analysis

We further present a quantitative analysis of the proposed *mirroring* approach. A baseline experiment is designed to only account for the trajectories of the hand in demonstration, which are directly mapped to robot end-effectors. We also test the proposed approach, without new training, in similar tasks of opening two additional medicine bottles– Bottle 2 with the same press-open mechanism but in different shape and size, and Bottle 3 with no lock.

Table 1 shows the success rate. Frames of the executions are shown in Figure 9b-9c. The success rates of the baseline (B) are significantly lower than those of our approach (M) for Bottle 1 and 2, showing the necessity of capturing the hidden force information and the effectiveness of our approach in transferring the manipulative actions. As pressing the lid is not required for opening Bottle 3, the success rates do not vary a lot between both methods.

Without mimicking the observed demonstrations, the learned the grammar model T-AOG is capable of sampling a *pt* that can be different from demonstration to alleviate overimitation. In the particular *pt* shown in Figure 9c, an a_6 *move down* action is not planned and no significant downward force is applied when twisting. The ablative analysis performed here shows that our approach can be generalized to similar but different tasks, and it is possible to avoid overimitation in new scenarios using the proposed approach.

5 Discussion

Overimitation in Infants, Children, and Animals. A phenomenon termed overimitation (Lyons, Young, and Keil 2007) illustrates a seeming cost of our imitative prowess. Children have been observed to overimitate, *i.e.*, to reproduce an adults obviously irrelevant actions (Tomasello et al. 2005; McGuigan et al. 2007), even in situations where chimpanzees correctly ignored the unnecessary steps (Horner and Whiten 2005; Want and Harris 2002; Whiten et al. 1996). However, such overimitation only occurs when the children are trying to discover the hidden structure of complex problems given the demonstrations (Lyons, Young, and Keil 2007). Younger infants, on the contrary, often do not exhibit the overimitation and only imitate rationally (Skerry, Carey, and Spelke 2013); but will still try to explore the world by discovering the hidden causes to explain the unexpected observations (Stahl and Feigenson 2015).

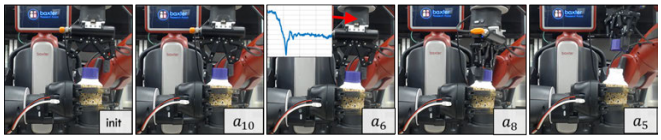
What are advantages and disadvantages of using simulation? Physics-based simulation can be difficult to tune and has a reality gap that does not match the real world perfectly. However, it still affords a powerful tool for a robot to explore the action space to reach the desired goal of a task rather than mimicking the demonstrations. On the other hand, if the accurate tactile sensing gripper and good object state perception are available, the simulation engine could be replaced in the proposed approach.

What is unique about *Goal-oriented* and *Mirroring* compared with other LfD methods? Prior work that learns action-state policy in using IRL (Osa et al. 2018) is similar to our *Goal-oriented* method that associates force types to state changes as *hoi* units. The IRL assumes a Markov Decision Process; however, the use of T-AOG that is composed

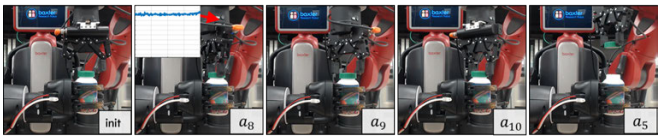
Table 1: The success rate for opening 3 bottles using the baseline model (B) and the proposed approach (M).

	Bottle 1	Bottle 2	Bottle 3
B	38.5%	30.8%	76.9%
M	69.2%	53.8%	73.1%

(a) Robot execution to open Bottle 1



(b) Robot execution to open Bottle 2



(c) Robot execution to open Bottle 3

Figure 9: Starting from the initial pose, the primitives (in grey) are performed sequentially. The robot "pushes" by a_6 (*downward*) (see force plot) and opens the medicine bottle by a_5 (*upward*).

of *hoi* units yields a non-Markovian model. *Mirroring* actions allows to generate an unseen action sequence for the robots that goes beyond the observed demonstration.

6 Conclusion

We present a *force-based, goal-oriented mirroring* approach for a robot to learn a manipulation task by imitating the forces from the demonstrations and the changes of the object's physical states caused by the forces. This approach differs from prior work that either mimics the demonstrator's trajectory or matches the keypoints, providing a deeper understanding of the physical world for a robot.

In the experiment, we use a tactile glove to collect human demonstrations of opening medicine bottles with safety locks. The discretized force types and object state changes are associated by a policy learned using the Q-Learning. A T-AOG is further induced to provide a robot the task planner on how to change the physical states by exerting a particular type of force. Our physics-based simulation engine is capable of emulating the forces produced by a set of robot actions. Finally, human demonstrations are successful "mirrored" to robot's actions with *functional equivalence* as they both produce similar forces and cause similar changes in object states, which is validated by an actual Baxter robot opening various medicine bottles.

Acknowledgments: The work reported herein was supported by DARPA XAI grant N66001-17-2-4029, ONR MURI grant N00014-16-1-2007, ARO grant W911NF-18-1-0296, and an NVIDIA GPU donation grant. We thank Prof. Tao Gao from the UCLA Statistics Department for useful discussions on the motivation of this work.

References

Abbeel, P., and Ng, A. Y. 2004. Apprenticeship learning via inverse reinforcement learning. In *ICML*.

Argall, B. D.; Chernova, S.; Veloso, M.; and Browning, B. 2009. A survey of robot learning from demonstration. *Robotics and Autonomous Systems* 57(5):469–483.

Bonaiuto, J.; Rosta, E.; and Arbib, M. 2007. Extending the mirror neuron system model, i. *Biological Cybernetics* 96(1):9–38.

Bonet, J., and Wood, R. D. 1997. *Nonlinear continuum mechanics for finite element analysis*. Cambridge university press.

Dautenhahn, K., and Nehaniv, C. L. 2002. *Imitation in Animals and Artifacts*. MIT Press Cambridge, MA.

Edmonds, M.; Gao, F.; Xie, X.; Liu, H.; Qi, S.; Zhu, Y.; Rothrock, B.; and Zhu, S.-C. 2017. Feeling the force: Integrating force and pose for fluent discovery through imitation learning to open medicine bottles. In *IROS*.

Gallese, V.; Fadiga, L.; Fogassi, L.; and Rizzolatti, G. 1996. Action recognition in the premotor cortex. *Brain* 119(2):593–609.

Gazzola, V.; Rizzolatti, G.; Wicker, B.; and Keysers, C. 2007. The anthropomorphic brain: the mirror neuron system responds to human and robotic actions. *Neuroimage* 35(4):1674–1684.

Guenter, F.; Hersch, M.; Calinon, S.; and Billard, A. 2007. Reinforcement learning for imitating constrained reaching movements. *Advanced Robotics* 21(13):1521–1544.

Gupta, A.; Eppner, C.; Levine, S.; and Abbeel, P. 2016. Learning dexterous manipulation for a soft robotic hand from human demonstrations. In *IROS*.

Horner, V., and Whiten, A. 2005. Causal knowledge and imitation/emulation switching in chimpanzees (pan troglodytes) and children (homo sapiens). *Animal cognition* 8(3):164–181.

Ito, M., and Tani, J. 2004. On-line imitative interaction with a humanoid robot using a dynamic neural network model of a mirror system. *Adaptive Behavior* 12(2):93–115.

Johnson-Frey, S. H.; Maloof, F. R.; Newman-Norlund, R.; Farrer, C.; Inati, S.; and Grafton, S. T. 2003. Actions or hand-object interactions? human inferior frontal cortex and action observation. *Neuron* 39(6):1053–1058.

Kanungo, T.; Mount, D. M.; Netanyahu, N. S.; Piatko, C. D.; Silverman, R.; and Wu, A. Y. 2002. An efficient k-means clustering algorithm: Analysis and implementation. *PAMI* 24(7):881–892.

Kober, J., and Peters, J. R. 2009. Policy search for motor primitives in robotics. In *NIPS*, 849–856.

Koenemann, J.; Burget, F.; and Bennewitz, M. 2014. Real-time imitation of human whole-body motions by humanoids. In *ICRA*.

Konidaris, G.; Kuindersma, S.; Grupen, R.; and Barto, A. 2012.

Robot learning from demonstration by constructing skill trees. *IJRR* 31(3):360–375.

Kormushev, P.; Calinon, S.; and Caldwell, D. G. 2011. Imitation learning of positional and force skills demonstrated via kinesthetic teaching and haptic input. *Advanced Robotics* 25(5):581–603.

Legg, S., and Hutter, M. 2007. Universal intelligence: A definition of machine intelligence. *Minds and Machines* 17(4):391–444.

Levine, S., and Abbeel, P. 2014. Learning neural network policies with guided policy search under unknown dynamics. In *NIPS*.

Levine, S.; Wagener, N.; and Abbeel, P. 2015. Learning contact-rich manipulation skills with guided policy search. In *ICRA*.

Lin, Y.; Ren, S.; Clevenger, M.; and Sun, Y. 2012. Learning grasping force from demonstration. In *ICRA*, 1526–1531.

Liu, H.; Xie, X.; Millar, M.; Edmonds, M.; Gao, F.; Zhu, Y.; Santos, V. J.; Rothrock, B.; and Zhu, S.-C. 2017. A glove-based system for studying hand-object manipulation via joint pose and force sensing. In *IROS*.

Lyons, D. E.; Young, A. G.; and Keil, F. C. 2007. The hidden structure of overimitation. *PNAS* 104(50):19751–19756.

MacGlashan, J., and Littman, M. L. 2015. Between imitation and intention learning. In *IJCAI*.

Macosko, C. W. 1994. *Rheology: principles, measurements, and applications*. Wiley-vch.

Maeda, G.; Ewerton, M.; Koert, D.; and Peters, J. 2016. Acquiring and generalizing the embodiment mapping from human observations to robot skills. *IEEE RA-L* 1(2):784–791.

Manschitz, S.; Gienger, M.; Kober, J.; and Peters, J. 2016. Probabilistic decomposition of sequential force interaction tasks into movement primitives. In *IROS*.

McGuigan, N.; Whiten, A.; Flynn, E.; and Horner, V. 2007. Imitation of causally opaque versus causally transparent tool use by 3-and 5-year-old children. *Cognitive Development* 22(3):353–364.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518(7540):529–533.

Montebelli, A.; Steinmetz, F.; and Kyrki, V. 2015. On handing down our tools to robots: Single-phase kinesthetic teaching for dynamic in-contact tasks. In *ICRA*.

Museth, K.; Lait, J.; Johanson, J.; Budsberg, J.; Henderson, R.; Alden, M.; Cucka, P.; Hill, D.; and Pearce, A. 2013. Openvdb: an open-source data structure and toolkit for high-resolution volumes. In *ACM SIGGRAPH 2013 courses*, 19.

Ng, A. Y.; Harada, D.; and Russell, S. 1999. Policy invariance under reward transformations: Theory and application to reward shaping. In *ICML*.

Ng, A. Y.; Russell, S. J.; et al. 2000. Algorithms for inverse reinforcement learning. In *ICML*.

Niekum, S.; Osentoski, S.; Konidaris, G.; Chitta, S.; Marthi, B.; and Barto, A. G. 2015. Learning grounded finite-state representations from unstructured demonstrations. *IJRR* 34(2):131–157.

Oberman, L. M.; McCleery, J. P.; Ramachandran, V. S.; and Pineda, J. A. 2007. Eeg evidence for mirror neuron activity during the observation of human and robot actions: Toward an analysis of the human qualities of interactive robots. *Neurocomputing* 70(13-15):2194–2203.

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

Oztop, E., and Arbib, M. A. 2002. Schema design and implementation of the grasp-related mirror neuron system. *Biological Cybernetics* 87(2):116–140.

Pastor, P.; Hoffmann, H.; Asfour, T.; and Schaal, S. 2009. Learning and generalization of motor skills by learning from demonstration. In *ICRA*.

Prieur, U.; Perdereau, V.; and Bernardino, A. 2012. Modeling and planning high-level in-hand manipulation actions from human knowledge and active learning from demonstration. In *IROS*.

Qi, S.; Huang, S.; Wei, P.; and Zhu, S.-C. 2017. Predicting human activities using stochastic grammar. In *ICCV*.

Racca, M.; Pajarinen, J.; Montebelli, A.; and Kyrki, V. 2016. Learning in-contact control strategies from demonstration. In *IROS*, 688–695.

Ramachandran, D., and Amir, E. 2007. Bayesian inverse reinforcement learning. In *IJCAI*.

Rizzolatti, G., and Craighero, L. 2004. The mirror-neuron system. *Annual Reviews Neuroscience* 27:169–192.

Rizzolatti, G.; Fogassi, L.; and Gallese, V. 2001. Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews Neuroscience* 2(9):661–670.

Shu, T.; Gao, X.; Ryoo, M. S.; and Zhu, S.-C. 2017. Learning social affordance grammar from videos: Transferring human interactions to human-robot interactions. In *ICRA*.

Si, H. 2015. Tetgen, a delaunay-based quality tetrahedral mesh generator. *ACM Transactions on Mathematical Software* 41(2):11.

Skerry, A. E.; Carey, S. E.; and Spelke, E. S. 2013. First-person action experience reveals sensitivity to action efficiency in prereaching infants. *PNAS* 201312322.

Stahl, A. E., and Feigenson, L. 2015. Observing the unexpected enhances infants learning and exploration. *Science* 348(6230):91–94.

Theodorou, E.; Buchli, J.; and Schaal, S. 2010. Reinforcement learning of motor skills in high dimensions: A path integral approach. In *ICRA*.

Thill, S.; Caligiore, D.; Borghi, A. M.; Ziemke, T.; and Baldassarre, G. 2013. Theories and computational models of affordance and mirror systems: an integrative review. *Neuroscience & Biobehavioral Reviews* 37(3):491–521.

Tomasello, M.; Carpenter, M.; Call, J.; Behne, T.; and Moll, H. 2005. In search of the uniquely human. *Behavioral and brain sciences* 28(5):721–727.

Want, S. C., and Harris, P. L. 2002. How do children ape? applying concepts from the study of non-human primates to the developmental study of imitationin children. *Developmental Science* 5(1):1–14.

Whiten, A.; Custance, D. M.; Gomez, J.-C.; Teixidor, P.; and Bard, K. A. 1996. Imitative learning of artificial fruit processing in children (homo sapiens) and chimpanzees (pan troglodytes). *Journal of comparative psychology* 110(1):3.

Yang, Y.; Li, Y.; Fermüller, C.; and Aloimonos, Y. 2015. Robot learning manipulation action plans by" watching" unconstrained videos from the world wide web. In *AAAI*.

Zhu, S.-C., and Mumford, D. 2007. A stochastic grammar of images. *Foundations and Trends® in Computer Graphics and Vision* 2(4):259–362.

Ziebart, B. D.; Maas, A. L.; Bagnell, J. A.; and Dey, A. K. 2008. Maximum entropy inverse reinforcement learning. In *AAAI*.