# WISH YOU WERE HERE: HINDSIGHT GOAL SELECTION FOR LONG-HORIZON DEXTEROUS MANIPULATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Complex sequential tasks in continuous-control settings often require agents to successfully traverse a set of "narrow passages" in their state space. Solving such tasks with a sparse reward in a sample-efficient manner poses a challenge to modern reinforcement learning (RL) due to the associated long-horizon nature of the problem and the lack of sufficient positive signal during learning. Various tools have been applied to address this challenge. When available, large sets of demonstrations can guide agent exploration. Hindsight relabelling on the other hand does not require additional sources of information. However, existing strategies explore based on task-agnostic goal distributions, which can render the solution of long-horizon tasks impractical. In this work, we extend hindsight relabelling mechanisms to guide exploration along task-specific distributions implied by a small set of successful demonstrations. We evaluate the approach on four complex, single and dual arm, robotics manipulation tasks against strong suitable baselines. The method requires far fewer demonstrations to solve all tasks and achieves a significantly higher overall performance as task complexity increases. Finally, we investigate the robustness of the proposed solution with respect to the quality of input representations and the number of demonstrations.

## 1 INTRODUCTION

Recent advances in model-free reinforcement learning (RL) have enabled successful applications in a variety of practical, real-world tasks (Gu et al., 2017; Levine et al., 2018; Riedmiller et al., 2018; Akkaya et al., 2019; Sharma et al., 2020; Gupta et al., 2021). Given the complexity of these tasks, additional information such as demonstrations often plays an important role (Vecerik et al., 2019; Zuo et al., 2020; Davchev et al., 2020; Sutanto et al., 2020; Luo et al., 2021). These methods are particularly useful to robotics as they enable efficient learning with only sparse rewards, which are easier to define. However, scaling such solutions to more complex long-horizon sequential settings with limited number of demonstrations remains an open challenge (Gupta et al., 2019).

Solving long-horizon dexterous manipulation tasks is an important problem as it enables a wide range of useful robotic applications ranging from object-handling tasks, common in collaborative settings, through to contact-rich dual arm manipulations, often seen in manufacturing. Consider the bi-manual 3.5mm jack cable insertion problem as illustrated in Figure 1. Solving such tasks with vanilla *demo-driven RL* often fails due to the difficulty of assigning credit over long horizons in the presence of noisy exploration and sparse rewards.

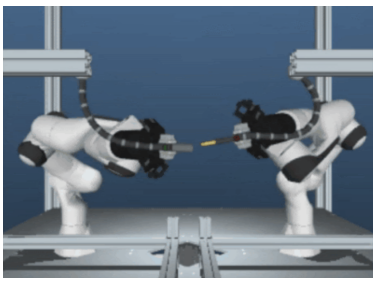


Figure 1: HinDRL solving a 3.5mm jack cable insertion task.

Self-supervision mechanisms like Hindsight Experience Replay (HER) (Andrychowicz et al., 2017) offer an alternative to ease the sparse reward problem by providing a stronger learning signal through additional goal-reaching tasks which are generated from the agent's trajectories. While the more frequent reward can be beneficial, the agent's trajectory distribution is often considerably more complex and non-stationary in comparison to the target-task distribution. When the final task is very complex and the main task reward is not perceived, this can lead to learning a capable, goal-reaching policy for states close to the agent's initial states while unable to complete the actual task.

Our contribution is based on the insight that demonstration states can be viewed as samples of the target task distribution, and can therefore be used to constrain the self-supervision process to task-relevant goals. We introduce a framework for *task-constrained* goal-conditioned RL that flexibly combines demonstrations with hindsight relabelling. Unlike HER, which learns a general goal-conditioned agent, we train a goal-conditioned agent specialized at achieving goals which directly lead to the task solution. The approach further allows to smoothly vary the task relevance of the relabelling process. We refer to our method as a Hindsight Goal Selection for Demo-Driven RL or HinDRL for short. We demonstrate that the proposed solution can solve tasks where both demonstration-driven RL and its self-supervised version with HER struggle. Specifically, we show that HinDRL can dramatically reduce the number of required demonstration by an order of magnitude on most considered tasks.

## 2 Related Literature

Using demonstrations in RL is a popular tool for robot learning. Classic approaches, such as (Atkeson & Schaal, 1997; Peters & Schaal, 2008; Tamosiunaite et al., 2011), use expert demonstrations to extract good initial policies before fine-tuning with RL. Kim et al. (2013); Chemali & Lazaric (2015) use demonstrations to learn an imitation loss function to bootstrap learning. However, those solutions are not applied to complex sequential tasks. Follow up work that use deep neural networks enable the application of demo-driven RL to more complex tasks such as Atari(Aytar et al., 2018; Pohlen et al., 2018; Hester et al., 2018), or robotics (Vecerik et al., 2017; 2019; Paine et al., 2018). We build upon DPGfD(Vecerik et al., 2019), which we explain in more details in Section 3.

The Atari literature often requires large number of environment steps (Salimans & Chen, 2018; Bacon et al., 2017). This can result in hardware wear and tear if applied on a physical robotics system. Alternatively, a number of robotics solutions assume the existence of a manually defined structure, e.g. through specifying an explicit set of primitive skills (Stulp et al., 2012; Xie et al., 2020), or a manually defined curricula (Davchev et al., 2020; Luo et al., 2021). However, hand-crafting structure can be sub-optimal in practice as it varies across tasks. In contrast, restricting the search space by greedily solving for specific goals (Foster & Dayan, 2002; Schaul et al., 2015) can be combined with implicitly defined curricula through retroactive goal selection (Kaelbling, 1993; Andrychowicz et al., 2017). Hindsight goal assignment has been successfully applied to multi-task RL (Li et al., 2020; Eysenbach et al., 2020), reset-free RL (Sharma et al., 2021) but also for batch RL (Kalashnikov et al., 2021; Chebotar et al., 2021) and hierarchical RL (Wulfmeier et al., 2019; 2021). However, these approaches explore based on task-agnostic goal distributions and compensate with additional policy structure such as allowing for longer training to concurrently learn a hierarchy, or using large amounts of offline data. In this work, we extend hindsight relabelling mechanisms to guide exploration along task-specific distributions implied from a limited number of demonstrations and study its performance against strong task-agnostic goal distributions.

Combining hindsight relabelling with demonstrations is not a novel concept. Ding et al. (2019) relabels demonstrations in hindsight to improve generative adversarial imitation learning. The proposed solution is used to learn a goal-conditioned reward function that can be applied on demo-driven RL for hindsight relabelling. Gupta et al. (2019) propose a hierarchical data relabelling mechanism that uses demonstrations to extract goal-conditioned hierarchical imitation-based policies and apply them to multi task RL. Nair et al. (2018a) fuses demo-driven RL with HER's final-goal sampling mechanism and combines it with structured resets to specific demo states to overcome the difficulties of exploration for long-horizon tasks, which would be difficult to achieve in a continuous robotics task in practice. Zuo et al. (2020) uses demonstrations to bootstrap TD3 and combines it with HER. However, in all those works, hindsight goals were always chosen directly from the agent's own trajectory. Instead, we *consistently select goals directly from a task distribution implied from a collection of successful demonstrations*. Hindsight goal selection has previously been generalized via adopting representation learning. Florensa et al. (2019) applied HER to a pixels-to-torque reaching task where sample efficiency was not a direct constraint. Nair et al. (2018b) used a $\beta-$balanced variational auto encoder ($\beta$-VAE) to enable learning from visual inputs with a special reward. While $\beta$-VAE is broadly a powerful tool for unsupervised representation learning, it does not take into account the temporal nature of sequential robotics tasks. This makes them sub-optimal in the context of our work (Chen et al., 2021). In contrast, contrastive-learning based techniques such as temporal cycle consistency (TCC) (Dwibedi et al., 2019) can provide more informative representations as

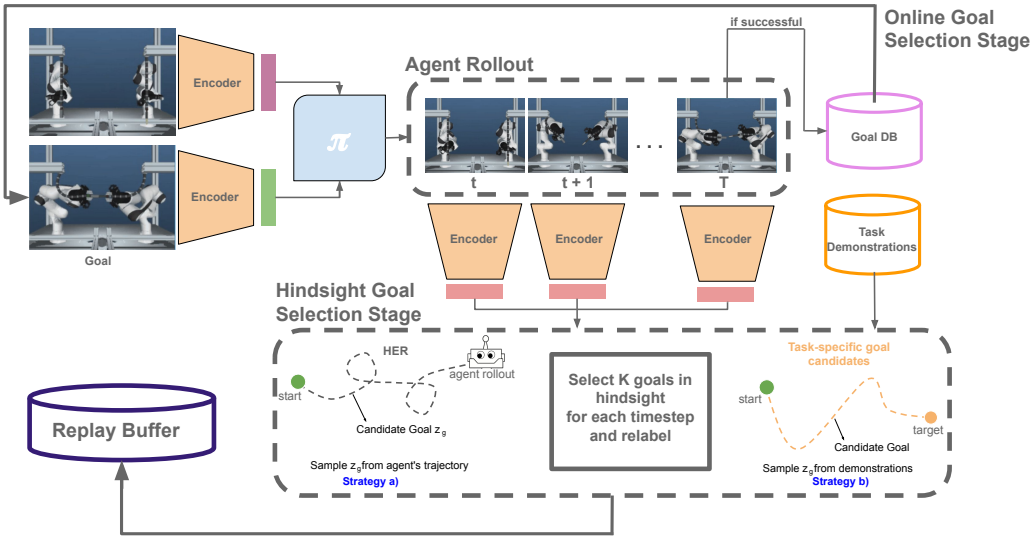



Figure 2: The HinDRL pipeline performing a 3.5mm jack cable insertion. The input to the policy are encoded target goal and the current robot state. At train time, the produced episodes are relabelled in hindsight using a choice of sampling strategies. We consider two main strategies for sampling goals in hindsight. Strategy a) shows the standard hindsight goal selection strategies that select goals from states within the trajectory being relabelled; and strategy b) shows a mechanism that focuses on sampling goals directly from some collection of successful trajectories. The resulted relabelled data is fed into the replay buffer.

previously discussed in concurrent work (Zakka et al., 2021). In this work we combine HinDRL, a framework for goal-conditioned demo-driven RL with both learnt and engineered representations and provide a comprehensive sensitivity analysis to the quality of the encoder and the dependency against the number of demonstrations used.

## 3 Preliminaries

Consider a finite-horizon discounted Markov decision process (MDP), $\mathscr{M} = (S, A, P, r, \gamma, T)$ with transition probability $P : S \times A \times S \mapsto [0, 1]$. Let the current state and goal $s, g \in S \subseteq \mathbb{R}^{n_s}$ be elements in the state space $S$, and $a \in A \subseteq \mathbb{R}^{n_a}$ denote the desired robot action. We define a sparse environmental reward $r(s)$ that assigns 0/1 reward only at the final state of an episode. Let $\psi(\cdot)$ be an encoding function that embeds a given state and goal to a latent state $z = \psi(s)$ and a latent goal $g = \psi(g)$. Let $\zeta = (\mathbf{x}_{t_0}, \ldots, \mathbf{x}_T)$ be a trajectory with a discrete horizon $T$, a discount function $\gamma(\cdot)$, a state-action tuple $x_t = (s_t, a_t, g)$ and a trajectory return $R(\zeta) = \sum_{t=t_0}^{T} \gamma r(s_t)$. In this setting, a transition $(s_t, a_t, s_{t+1}, g, r = 0)$ can be 'relabelled' as $(s_t, a_t, s_{t+1}, \hat{g} \approx s_{t+1}, r = 1)$ and both the original and relabelled transitions can be used for training. A separate, goal conditioned reward $r(z_{t+1}, \hat{z}_g) = \mathbb{1}[z_{t+1} = \hat{z}_g]$ uses the latent state and goal and assigns reward during relabelling. Finally, we also assume access to D demonstration trajectories $\mathscr{D} = \{\zeta_j\}_{j=0}^{D}$ that reach the goals $\{\hat{g}_j\}_{j=1}^{D}$ retrieved from the final state of the demonstration. Then, a goal-conditioned deterministic control policy parameterised by $\theta$, $\pi_\theta(a|z, z_g)$ selects an action $a$ given a latent state $z$ and goal $z_g$. In this context, we define an optimal policy $\pi^* = \arg\max_{\pi \in \bar{\pi}} J(\pi)$, where $J(\pi) = \mathbb{E}_{s_0 \sim p(s_0), g \sim \pi(g)}[R(\zeta)]$.

DPGfD(Vecerik et al., 2019) is a powerful algorithm that bootstraps learning through a BC loss applied directly to the actor. This loss is typically defined as the L2 distance between the predicted by the policy actions and the true actions of a demonstration and is decayed proportionally to the number of environmental steps. Gradual conversion from using a BC loss to using the actor-critic loss is typically done using Q filtering or gradually decaying the BC loss. We employ the latter. The algorithm uses an L2 regularised actor-critic loss with a distributional critic, introduced in (Bellemare et al., 2017). Using BC loss with DDPG was recently applied to a goal-conditioned setting (Ding et al., 2019) but goal-conditioned policy learning was never used with demo-driven distributional RL.

# 4 METHODOLOGY

This section describes our method, HinDRL. First, we extend DPGfD to its goal-conditioned interpretation and introduce a mechanism that deals with intermediate goals. Next, we introduce two demonstration-driven goal sampling strategies and discuss how to use them for goal-conditioned policy learning. Finally, we discuss how to extract temporally consistent representations and define a distance-based goal-conditioned reward function. We summarise our framework in Figure 2.

## 4.1 GOAL-CONDITIONED DEMO-DRIVEN REINFORCEMENT LEARNING

Similar to DPGfD we combine reinforcement and imitation learning. We utilise a data set of demonstrations, $\mathscr{D} = \{(s_0, a_0, s_g, \ldots)_t^k)\}_{k=1}^D$ which are used to seed a replay buffer for RL, and for supervised training of the policy. We used both 6D Cartesian velocity and a binary open/close actions, which we modelled using a standard regression for the predicted velocity and a classification for the binary prediction,

$$\mathscr{L}_{BC} = ||\pi(z, z_g)_c - a_c||_2^2 - \sum_{i=1}^{2} a_b^o log(\pi(z, z_g)_b^o). \tag{1}$$

The BC loss is applied only to transitions from successful trajectories.

For RL we utilized the standard deterministic policy-gradient (DPG, Silver et al. (2014)) loss for the deterministic actions (velocities), and stochastic value-gradient (SVG, Heess et al. (2015)) for the stochastic actions (gripper positions). We used the Gumbel-Softmax trick (Jang et al., 2016; Maddison et al., 2016) to reparameterize the binary actions for the SVG loss. Following Vecerik et al. (2019) we applied both the BC and RL losses to train the policy, and annealed the BC loss throughout training to allow the agent to outperform the expert.

We use a distributional critic (Bellemare et al., 2017), modeling the Q-function with a categorical distribution between 0 and 1 (with 60 evenly spaced bins). The learning loss is the KL-divergence between our current estimate and a projection of the 1-step return on the current bins. Therefore, the loss becomes,

$$\mathscr{L}_{TD} = KL(Q(z_t, a_t, z_g)||\Phi(r_t + \gamma(t) * Q_{target}(z_t, \pi(z_t, z_g), z_g))), \tag{2}$$

where $\Phi$ is an operator that projects a distribution on the set of bins.

## 4.2 GOAL SELECTION FROM DEMONSTRATIONS

Typically, goal-conditioned policy learning has two separate stages for goal selection (Andrychowicz et al., 2017): an online stage where a policy is conditioned on a specific goal during a policy rollout, and a hindsight goal selection stage where the produced trajectory is being relabelled in hindsight (See Figure 2). While Andrychowicz et al. (2017) only consider the original goal in the online stage, we describe below how this perspective can be extended to increase robustness.

**Online goal selection** A target goal of a sequential task, such as inserting a plug into a socket, is in principle more abstract than the physical skill of reaching a specific configuration. That is, being an $\varepsilon$ distance away from a target configuration might still result in a failed insertion. We mitigate this issue by maintaining a separate goal database, $G_{db}$, comprised of final states over successful executions. We bootstrap this database with the final states from $\mathscr{D}$ but we also continuously grow $G_{db}$ as learning progresses. Conceptually, the more examples of final goal states we collect, the better understanding of the target goal of the current task we will have. In this context, for every episode we sample a new $\hat{z}_g$ from a distribution over the set of all possible goals in $G_{db}$. We denote this as $R_G = \{z_g \in G_{db} : p(z_g) > 0\}$ and we refer to this process as online goal-conditioning. In this work, we always store the resulted $\hat{z}_g$-conditioned trajectory, $\zeta$, in the replay buffer. As in classic RL, the last step of $\zeta$ is rewarded using the environment reward $r(s_t)$.

**Hindsight goal selection** In contrast, hindsight goal selection is the process of sampling candidate goal states from some goal distribution $p(z_g)$, typically future states in rollout trajectories.Here, $z_g$ are not conceptually related to the abstract formulation of a target task as $\hat{z}_g$ is. Instead, we sample goals for each time step from trajectory $\zeta$ using some candidate goal distribution $p(z_g)$. Then, we re-evaluate the given transition with the newly sampled goal and assign a new reward. This is achieved with the help of a goal conditioned reward $r(z_t, z_g)$ and not the environment reward from

**Algorithm 1** HinDRL

Initial $\theta$, $\phi$, $\mathscr{D} = \{\tau^k\}_{k=1}^D$, $RB = \emptyset$
Policy $\pi_\theta$, encoder $\psi_\phi$, $p(s_0)$
// if trainable, learn encoder offline here
// encode demos, add to replay buffer
$RB \leftarrow RB \bigcup \mathscr{D}$
// relabel trajectories and add to RB
$RB \leftarrow RB \bigcup$ {generate_samples($\tau^k, \mathscr{D}$)}$_{k=1}^D$
// add final state from demos to goal db
$G_{db} = G_{db} \bigcup \{\tau_T^k\}_{k=1}^D$
**for** iter $i \in iters_{max}$ **do**
  $s_0 \leftarrow p(s_0)$ // initial state
  $\hat{g} \sim G_{db}$ // target goal
  // trajectory $\zeta = \{(z_t, a_t, z_{t+1}, r_t, z_g)\}_{t=1}^T$
  $\zeta \leftarrow$ rollout($s_0, g, \pi(\cdot), \psi(\cdot)$)
  **if** *success* **then**
    // if successful, store last state in goal db
    $G_{db} \leftarrow G_{db} \bigcup z_T$ // $z_T \in \zeta$
  **end if**
  $RB \leftarrow RB \bigcup \zeta$ // add rollout to RB
  // relabel $\zeta$ and add to RB
  $RB \leftarrow RB \bigcup$ generate_samples($\zeta, \mathscr{D}$)
  // optionally, retrain $\phi$ here
  $\theta \leftarrow train_\pi(R)$
**end for**

**Algorithm 2** Generate samples

Input: trajectory, $\zeta$, task demos $\mathscr{D}$
$\chi \leftarrow \emptyset$ // relabelled transitions
**for** sampler $\in$ sampling strategies list **do**
  data $\leftarrow \emptyset$ // support data
  **if** sampler.name is Rollout-conditioned **then**
    data = $\zeta$ // HER
  **else if** sampler.name is Task-conditioned **then**
    data = $\mathscr{D}$ // Task
  **end if**
  // compose support for goal distribution
  $R_G = \{z_g \in data : p(z_g) > 0\}$
  // sample new goals, $T = |\zeta|$
  $\{z_g \sim p(z_g)\}_{t=1}^T$
  // store relabelled transitions using Eq. 3
  $\chi \leftarrow \chi \bigcup \{(z_t, a_t, z_{t+1}, r(z_{t+1}, z_g), z_g)\}_{t=1}^T$
**end for**
Returns: $\chi$

Figure 3: Algorithm for Hindsight goal selection for Demo-driven Reinforcement learning (HinDRL)

the previous paragraph. Thanks to the goal-conditioned formulation, fitting a Q function from both of these sources of reward is well-posed. In this section we propose a way for composing a hindsight goal distribution $p(z_g)$, that is targeted on the specific task at hand.

Candidate goals can be acquired through self-supervision (as in HER), where a goal distribution is comprised of states from the agent's trajectory $\zeta$. The agent retrospectively samples feasible goals to 'explain' its own behaviour under the current goal-conditioned policy $\pi$. However, similar to concurrent work (Gupta et al., 2019; Pertsch et al., 2020) we observed that choosing goals from $p(z_g|\zeta)$ does not work well for long-horizon tasks.

However, hindsight goal selection does not need to be constrained to the same trajectory. Goals can be sampled directly from a task distribution, implied from a set of successful trajectories, such as demonstrations. Therefore, selecting task-specific goals can come from $p(z_g)$ with support $R_G = \{z_g \in \mathscr{D} : p(z_g) > 0\}$. In practice, the support of the task-distribution, $R_G$, can be enriched over time with additional positive task completions from the training process too. However, we do not do it in this work.

Finally, these two support distributions do not need to be disjoint and can be mixed together. We provide an ablation against mixing samplers in Appendix A.4, including using the union, $\zeta \bigcup \mathscr{D}$, and intersection $\zeta \bigcap \mathscr{D}$ of the two sets of goals.

### 4.3 Time-consistent Representations

Synthesising goal-conditioned policies from raw state observations can be impractical in high dimensional spaces. Using raw states can not only deteriorate the speed of learning (Hessel et al., 2018), but it also makes it difficult to define a comprehensive goal-conditioned reward for relabelling. Instead, we encode our state observations into a lower dimensional latent space $\mathbb{R}^{n_z}$, using an encoder $\psi$ to obtain a latent state $z_t = \psi(s_t)$ and a latent goal $z_g = \psi(g)$. Next, we propose both expert-engineered and learnt time-consistent representations.

**Engineered representations** In the presence of an expert, low dimensional representations can be programmed to include features that bear a notion of progress, e.g. through measuring the distance

from a target goal, as well as the notion of contact, such as whether an insertion was successful. These notions vary across each task and we provide additional details for each of the engineered encoders in Appendix A.9.

**Learnt representations** We propose as an alternative to the engineered encoding, a self-supervised learnt representation that can be directly used with the provided demonstrations and refined over time. We ensure consistent and aligned notion of episode progress in our learnt representation without the need of labels by employing a differentiable cycle-consistency loss (TCC)(Dwibedi et al., 2019). This allows us to learn a representation by finding correspondences across time between trajectories of differing length, effectively preserving the notion of progress.

Training of TCC involves encoding an observation $o_t$ from trajectory $U$ to a latent $z_t^U$, and checking for cycle-consistency with another trajectory $V$ that is not necessarily of the same length as $U$. Cycle consistency is defined via nearest-neighbor – if the nearest embedding in $V$ to $z_t^U$ is $z_m^V$, i.e. $nearest(z_t^U, V) = z_m^V$, then $z_t^U$ and $z_m^V$ are cycle-consistent if and only if $nearest(z_m^V, U) = z_t^U$. Learning this requires a differentiable version of cycle-consistency measure, such as regression-based or classification-based one. In this work, we employ the latter.

We additionally evaluate a non-temporal embedding using a $\beta$-VAE, similar to (Nair et al., 2018b) (see appendix).

### 4.4 Distance-based Reward Specification

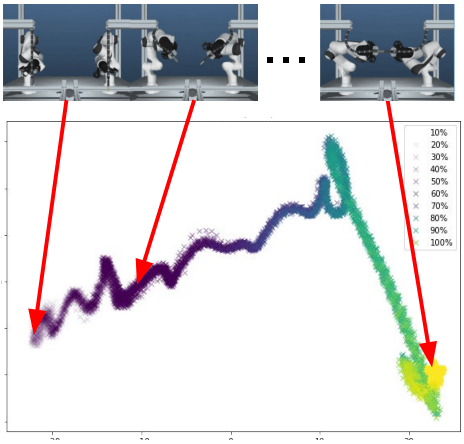



Figure 4: t-SNE of latent representations obtained with TCC (Dwibedi et al., 2019) against novel successful trajectories. Temporal color-coding, purple is start and yellow is goal.

Considering local smoothness in the learnt goal representation space, we use a distance-based goal-conditioned reward function. This allows us to increase the number of positive rewards and therefore ease the hard exploration problem. We particularly use,

$$r(z, z_g) = \mathbb{1}[||z - z_g|| < \varepsilon], \tag{3}$$

for some threshold $\varepsilon$, where latent state $z$ and latent goal $z_g$ are considered similar if the $l_2$ distance between them is below $\varepsilon$. We provide more details on how we obtain the threshold in Appendix A.8. Figure 4 illustrates the t-SNE visualisation of a temporally aligned representation space learnt with TCC. The notion of progress in TCC is consistent across all 100 plotted trajectories. We provide additional details in Appendix A.6. TCC is inherently capable of temporally aligning different in length and motion trajectories that pass through similar stages. This makes it particularly suitable to use with distance-based metrics like Eq. 3.

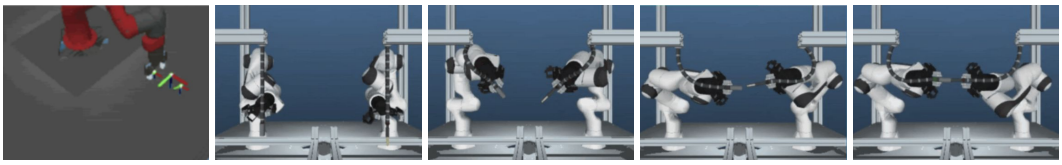
The final proposed algorithm, together with the relabelling technique, is detailed in Algorithm 1.

## 5 Experiments

We are interested in answering the following questions: i) is task-conditioned hindsight goal selection useful for long-horizon sequential tasks; ii) does it improve over DPGfD and HER's sample efficiency; iii) is HinDRL a demonstration efficient solution and how does it perform in a few-shot setting; and iv) how useful are learnt time-consistent representations.

### 5.1 Tasks

We answer these questions based on four tasks with increasing complexity, implemented with MuJoCo (Todorov et al., 2012). We consider an attempt successful if the agent gets under a threshold distance from its target. The metrics include the running accuracy of completing the tasks and the overall accumulated reward. All policies are executed in 10Hz with 12 random seeds. At test time we sample



(a) (b) (c) (d) (e)

Figure 5: **Tasks description** a) is the parameterised reach task. Visiting the small yellow goals must happen before reaching the middle goal. b) is dual arm reaching, we combine this with c) lifting, d) aligning in both position and orientation and e) is the dual arm insertion.

| **Environment** | BC<br>Mean (Std) | DPGfD<br>Mean (Std) | HER (final)<br>Mean (Std) | HER (future)<br>Mean (Std) | HinDRL (Our)<br>Mean (Std) |
|---|---|---|---|---|---|
| Parameterised Reach (2wp) | 82.05% (0.0) | 88.62% (7.7) | 84.92% (28.6) | 86.92 (5.2) | **92.61% (4.7)** |
| Bring Near | 88.90% (0.0) | **99.74% (1.0)** | 32.03% (41.8) | 97.75% (2.9) | 98.79% (4.1) |
| Bring Near + Orient | 57.89% (0.0) | 83.52% (18.0) | 8.33% (27.6) | 23.28% (33.5) | **90.77% (10.0)** |
| Bimanual Insertion | 16.06% (0.0) | 4.28% (4.4) | 3.38% (3.3) | 18.39% (12.3) | **78.92% (10.3)** |
| Average | 61.23% (0.0) | 69.04% (7.8) | 32.17% (25.32) | 56.59% (12.9) | **90.27% (7.3)** |

Table 1: Records performance in terms of accuracy.

a random goal from all successful goal states and for the agent. We use $z = (s, e(s))_t$ and $z_g = e(g)$, for all encoders $e(\cdot)$ and always train with a sparse binary environment reward. The number $K$ of goal samples is dependent on the complexity of the task with further details in the next paragraphs.

**Parameterised Reach:** In this environment, we use a single 7DoF Sawyer robot to perform parameterised reaching. Parameterised reaching is the task of visiting two randomly located in free space way-points before reaching for its target goal, Figure 5a. The continuous action space is comprised of the Cartesian 6 DoF pose of the robot and a 1 DoF representing the gripper open/close motion. The goal space is equivalent to the state space. We use a total of two hindsight goal samples, $z_g$, to relabel a single time step.

**Bring Near:** This is a dual-arm task situated in a cell with 2 Panda Franka Emika arms and hanging audio cables, Figure 5b-c. The agent has to use both robot arms to reach for two cables and bring both tips within an $\varepsilon$ distance from each other. The action space is 7DoF for each arm. The state and goal spaces are twice as large due to the dual nature of the task. Here we use a total of four hindsight goal samples, $z_g$, to relabel a single time step.

**Bring Near and Orient:** This is a dual-arm task using the same action, state and goal space as above. Similar to above, the agent needs to reach and grasp the two cables. However, this time it is also required to align both tips, that is position and reorient them within a certain threshold. This task is significantly more complicated than the one above since it requires an additional planning component. The two cables have rigid components at their tips, a 3.5mm jack and a socket respectively. In order to be manipulated to a specific orientation those cables have to be grasped at the rigid parts as grasping the flexible part of the cable prevents from aligning the two tips (Figure 5d). We used a total of six hindsight goal samples, $z_g$, to relabel a single time step.

**Bimanual Insertion:** This is the most complex task we evaluate against. It has the same dual-arm set up as above but requires the complete insertion of the 3.5mm jack as well as the reaching, grasping and alignment stages from the previous two bimanual tasks (Figure 5e). This task requires not only careful planning but also very precise manipulation in order to successfully complete insertion. We used a total of 12 hindsight goal samples, $z_g$, to relabel a single time step for this task.

### 5.2 Hindsight Goal selection for demo-driven RL: HinDRL

In this section we evaluate the utility of using different strategies for goal sampling in hindsight in the context of demo-driven RL. We compare the performance of HinDRL against several strategies for hindsight goal sampling, as well as two non-goal-conditioned baselines – vanilla DPGfD and Behavious Cloning (BC).

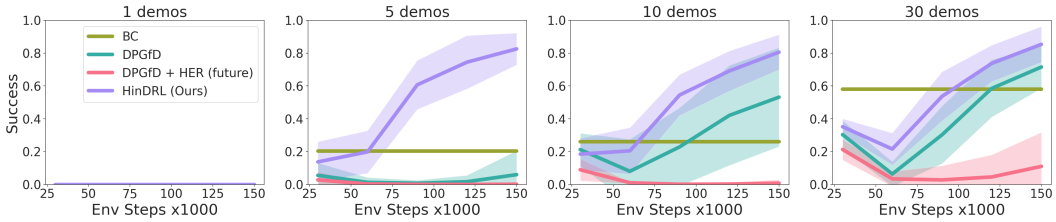



Figure 7: Achieved accuracy over different number of demonstrations. Overall performance on the Bring Near + Orient task. Each plot shows the overall achieved accuracy during training. HinDRL consistently outperforms the alternatives.

In addition, we combine DPGfD with hindsight relabelling using only the final reached state by the agent, similar to (Nair et al., 2018a) although without a curriculum over start configurations. We refer to this as HER (final). We consider using samples from the future trajectory to relabel both the demonstrations and the agent rollouts, similar to (Ding et al., 2019), although we do not consider using a stronger BC component. we refer to this baseline as HER (future). We report our findings in Table 6. In all cases HinDRL performs best. Notably, assuming a sufficient number of demonstrations, the DPGfD agent is able to solve both the parameterised reach and the bring near tasks very well, but struggles on the full cable-insertion task. Both HER-based samplers struggle on this task as well, but HER (future) performs better. In summary, the ability of HinDRL to specialize at achieving goals on track to the specific task solution through the proposed task-constrained self-supervision results in a superior performance across all considered long-horizon tasks.

### 5.3 Performance on few-shot tasks

| **Environment** | BC<br>Mean (Std) | DPGfD<br>Mean (Std) | HER (future)<br>Mean (Std) | HinDRL (Our)<br>Mean (Std) |
|---|---|---|---|---|
| 55 demos | 16.06% (0.0) | 4.28% (4.4) | 18.39% | **78.92% (10.3)** |
| 10 demos | 2.94% (0.0) | 0.0% (0.0) | 0.0% (0.0) | **32.05% (23.3)** |
| 5 demos | 0.52% (0.0) | 0.0% (0.0) | 0.0% (0.0) | **17.78% (18.9)** |
| 1 demos | 0.13% (0.0) | 0.0% (0.0) | 0.0% (0.0) | **12.58% (18.3)** |

Figure 6: Records performance against using different number of demonstrations.

Using demonstrations can significantly speed up the training time while still successfully outperforming the expert eventually, (Vecerik et al., 2019). However, the final achieved performance in sparse reward settings depends on the complexity of the task and the number of provided demonstrations. In this section we show how using task-conditioned hindsight goal selection can significantly reduce the demonstration requirements and relax the dependency of the RL agent to the performance of the BC loss. We evaluate the ability of HinDRL to work in a few-shot setting. We train with 1, 5, 10 and 30 demonstrations and report our findings against the Bring Near + Orient task in Figure 7. The figure shows the overall achieved accuracy across different number of demonstrations. There, all solutions fail to solve the task with a single demonstration. This shows that all considered DPGfD-based solutions were dependent on the ability of the BC component to achieve more than 0% accuracy. HinDRL is able to outperform the BC policy rather quickly on the 5-and-10-shot tasks while both DPGfD and HER took significantly longer. Furthermore, HinDRL manages to maintain its final achieved accuracy across the 5, 10 and 30-shot cases for the dedicated training budget while both DPGfD and BC didn't. Table 6 shows the final achieved accuracy on the Bi-manual Insertion task. There, only HinDRL was able to learn in the 10-5-1-shot scenarios. This indicates the ability of task-constrained hindsight relabelling to bootstrap learning for long-horizon dexterous manipulation tasks, particularly with high dimensional continuous action spaces.

### 5.4 Studying the speed of execution

We study the dependency of HinDRL on the number of demonstrations and the achieved accuracy of the BC component. The results show a definite improvement over these two factors. However, this does not necessarily mean that HinDRL is strictly better than the expert's performance. In practice, a quick and nimble expert can still produce higher total output and be more valuable on average than an RL solution even if it is less successful. Therefore, in this section we study the overall speed of performing a task. We measure the speed of solving a task as a function of the per-step reward accumulated over the course of the entire training. Therefore, an increased per-step reward means that an episode takes less environment steps to complete. We report our findings in Figure 8. We

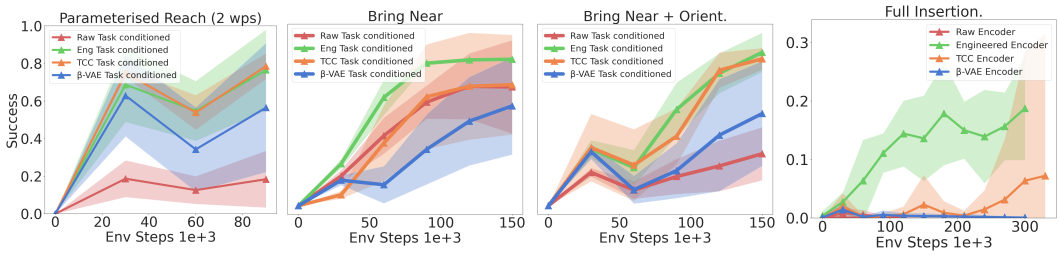



Figure 9: Robustness to the quality of the encoder. Comparing the engineered encoder with raw state observations and learnt $\beta$-VAE and TCC based encoders. The engineered encoder performs best with TCC achieving close to commensurate performance.

provide an additional ablation over using different numbers of demonstrations in Appendix A.5. Our findings indicate that HinDRL was always able to outperform the speed of execution obtained from the BC policy while using only DPGfD or DPGfD with HER did not. HER is beneficial for the easier tasks with insufficient number of demonstrations for DPGfD. However, HER alone is unable to solve the longer-horizon complex sequential tasks.

### 5.5 Robustness to the quality of the encoder

In the above sections we study performance when using different hindsight goal samplers, HinDRL sensitivity to the number of demonstrations and its ability to learn more efficient policies than the provided expert. However, in all those cases we assumed access to a hand-engineered state encoder. Informative representations is of central important to HinDRL as the algorithm uses the state representations both as input to the RL agent but also during the relabelling stage to assign reward, as discussed in Section 4. However, access to near perfect representations is not always feasible. Therefore, we study the sensitivity of our solution to the quality of the encoder. We evaluate the achieved performance against using learnt representations and raw input too. We compare learnt representations. One is extracted with a $\beta$-VAE, for $\beta = 0.5$, and one with TCC. We hypothesise that the notion of episode progress is of crucial importance to HinDRL. We report our findings in Figure 9. Representations that preserve temporal consistency were able to achieve higher results on all tasks. Using a learnt encoding obtained with TCC was capable of achieving near equivalent performance to using an engineered encoding. We notice that using a support, $R_G$ for the goal distribution that is comprised of the union between the task-conditioned distribution and HER help improve the $\beta$-VAE performance, indicating that the agent's rollout was able to partially substitute the lack of notion of progress. However, we provide this more detailed ablation in Appendix A.6. Unlike the engineered encoding, the TCC-based representation does not contain any manipulation-specific information.

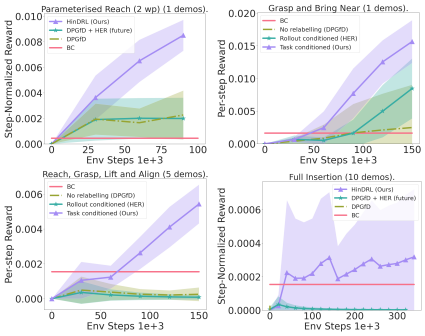



Figure 8: Measuring per-step reward using the smallest number of demonstrations that resulted in learning.

## 6 Conclusion

We proposed an efficient goal-conditioned, demonstration-driven solution to solving complex sequential tasks in sparse reward settings. We introduce a novel sampling heuristic for hindsight relabelling using goals directly from the demonstrations and combine it with both engineered and learnt encoders that consistently preserve the notion of progress. We show that HinDRL outperforms competitive baselines. Moreover, the method requires a considerably lower number of demonstrations. In the future, we would like to extend this work to multi-task settings, employ vision and move towards the context of batch RL, e.g. through employing distance learning techniques (Hartikainen et al., 2019) to build on more informative goal spaces.

## REFERENCES

Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, et al. Solving rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019.

Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. *arXiv preprint arXiv:1707.01495*, 2017.

Christopher G Atkeson and Stefan Schaal. Robot learning from demonstration. In *ICML*, volume 97, pp. 12–20. Citeseer, 1997.

Yusuf Aytar, Tobias Pfaff, David Budden, Tom Le Paine, Ziyu Wang, and Nando de Freitas. Playing hard exploration games by watching youtube. *arXiv preprint arXiv:1805.11592*, 2018.

Pierre-Luc Bacon, Jean Harb, and Doina Precup. The option-critic architecture. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International Conference on Machine Learning*, pp. 449–458. PMLR, 2017.

Yevgen Chebotar, Karol Hausman, Yao Lu, Ted Xiao, Dmitry Kalashnikov, Jake Varley, Alex Irpan, Benjamin Eysenbach, Ryan Julian, Chelsea Finn, et al. Actionable models: Unsupervised offline reinforcement learning of robotic skills. *arXiv preprint arXiv:2104.07749*, 2021.

Jessica Chemali and Alessandro Lazaric. Direct policy iteration with demonstrations. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

Xin Chen, Sam Toyer, Cody Wild, Scott Emmons, Ian Fischer, Kuang-Huei Lee, Neel Alex, Steven H Wang, Ping Luo, Stuart Russell, et al. An empirical investigation of representation learning for imitation. 2021.

Todor Davchev, Kevin Sebastian Luck, Michael Burke, Franziska Meier, Stefan Schaal, and Subramanian Ramamoorthy. Residual learning from demonstration. *arXiv preprint arXiv:2008.07682*, 2020.

Yiming Ding, Carlos Florensa, Mariano Phielipp, and Pieter Abbeel. Goal-conditioned imitation learning. *arXiv preprint arXiv:1906.05838*, 2019.

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal cycle-consistency learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1801–1810, 2019.

Benjamin Eysenbach, Xinyang Geng, Sergey Levine, and Ruslan Salakhutdinov. Rewriting history with inverse rl: Hindsight inference for policy improvement. *arXiv preprint arXiv:2002.11089*, 2020.

Carlos Florensa, Jonas Degrave, Nicolas Heess, Jost Tobias Springenberg, and Martin Riedmiller. Self-supervised learning of image embedding for continuous control. *arXiv preprint arXiv:1901.00943*, 2019.

David Foster and Peter Dayan. Structure in the space of value functions. *Machine Learning*, 49(2): 325–346, 2002.

Shixiang Gu, Ethan Holly, Timothy Lillicrap, and Sergey Levine. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *2017 IEEE international conference on robotics and automation (ICRA)*, pp. 3389–3396. IEEE, 2017.

Abhishek Gupta, Vikash Kumar, Corey Lynch, Sergey Levine, and Karol Hausman. Relay policy learning: Solving long-horizon tasks via imitation and reinforcement learning. *arXiv preprint arXiv:1910.11956*, 2019.

Abhishek Gupta, Justin Yu, Tony Z Zhao, Vikash Kumar, Aaron Rovinsky, Kelvin Xu, Thomas Devlin, and Sergey Levine. Reset-free reinforcement learning via multi-task learning: Learning dexterous manipulation behaviors without human intervention. *arXiv preprint arXiv:2104.11203*, 2021.

Kristian Hartikainen, Xinyang Geng, Tuomas Haarnoja, and Sergey Levine. Dynamical distance learning for semi-supervised and unsupervised skill discovery. *arXiv preprint arXiv:1907.08225*, 2019.

Nicolas Heess, Greg Wayne, David Silver, Timothy Lillicrap, Yuval Tassa, and Tom Erez. Learning continuous control policies by stochastic value gradients. *arXiv preprint arXiv:1510.09142*, 2015.

Matteo Hessel, Joseph Modayil, Hado Van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver. Rainbow: Combining improvements in deep reinforcement learning. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

Todd Hester, Matej Vecerik, Olivier Pietquin, Marc Lanctot, Tom Schaul, Bilal Piot, Dan Horgan, John Quan, Andrew Sendonaris, Ian Osband, et al. Deep q-learning from demonstrations. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

Leslie Pack Kaelbling. Learning to achieve goals. In *IJCAI*, pp. 1094–1099. Citeseer, 1993.

Dmitry Kalashnikov, Jacob Varley, Yevgen Chebotar, Benjamin Swanson, Rico Jonschkowski, Chelsea Finn, Sergey Levine, and Karol Hausman. Mt-opt: Continuous multi-task robotic reinforcement learning at scale. *arXiv preprint arXiv:2104.08212*, 2021.

Beomjoon Kim, Amir-massoud Farahmand, Joelle Pineau, and Doina Precup. Learning from limited demonstrations. In *NIPS*, pp. 2859–2867. Citeseer, 2013.

Sergey Levine, Peter Pastor, Alex Krizhevsky, Julian Ibarz, and Deirdre Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International Journal of Robotics Research*, 37(4-5):421–436, 2018.

Alexander C Li, Lerrel Pinto, and Pieter Abbeel. Generalized hindsight for reinforcement learning. *arXiv preprint arXiv:2002.11708*, 2020.

Jianlan Luo, Oleg Sushkov, Rugile Pevceviciute, Wenzhao Lian, Chang Su, Mel Vecerik, Ning Ye, Stefan Schaal, and Jon Scholz. Robust multi-modal policies for industrial assembly via reinforcement learning and demonstrations: A large-scale study. *arXiv preprint arXiv:2103.11512*, 2021.

Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

Ashvin Nair, Bob McGrew, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Overcoming exploration in reinforcement learning with demonstrations. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6292–6299. IEEE, 2018a.

Ashvin Nair, Vitchyr Pong, Murtaza Dalal, Shikhar Bahl, Steven Lin, and Sergey Levine. Visual reinforcement learning with imagined goals. *arXiv preprint arXiv:1807.04742*, 2018b.

Tom Le Paine, Sergio Gómez Colmenarejo, Ziyu Wang, Scott Reed, Yusuf Aytar, Tobias Pfaff, Matt W Hoffman, Gabriel Barth-Maron, Serkan Cabi, David Budden, et al. One-shot high-fidelity imitation: Training large-scale deep nets with rl. *arXiv preprint arXiv:1810.05017*, 2018.

Karl Pertsch, Youngwoon Lee, and Joseph J Lim. Accelerating reinforcement learning with learned skill priors. *arXiv preprint arXiv:2010.11944*, 2020.

Jan Peters and Stefan Schaal. Natural actor-critic. *Neurocomputing*, 71(7-9):1180–1190, 2008.

Tobias Pohlen, Bilal Piot, Todd Hester, Mohammad Gheshlaghi Azar, Dan Horgan, David Budden, Gabriel Barth-Maron, Hado Van Hasselt, John Quan, Mel Večerík, et al. Observe and look further: Achieving consistent performance on atari. *arXiv preprint arXiv:1805.11593*, 2018.

Martin A. Riedmiller, Roland Hafner, Thomas Lampe, Michael Neunert, Jonas Degrave, Tom Van de Wiele, Volodymyr Mnih, Nicolas Heess, and Jost Tobias Springenberg. Learning by playing - solving sparse reward tasks from scratch. *CoRR*, abs/1802.10567, 2018. URL `http://arxiv.org/abs/1802.10567`.

Tim Salimans and Richard Chen. Learning montezuma's revenge from a single demonstration. *arXiv preprint arXiv:1812.03381*, 2018.

Tom Schaul, Daniel Horgan, Karol Gregor, and David Silver. Universal value function approximators. In *International conference on machine learning*, pp. 1312–1320. PMLR, 2015.

Archit Sharma, Michael Ahn, Sergey Levine, Vikash Kumar, Karol Hausman, and Shixiang Gu. Emergent real-world robotic skills via unsupervised off-policy reinforcement learning. *arXiv preprint arXiv:2004.12974*, 2020.

Archit Sharma, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Persistent reinforcement learning via subgoal curricula. *arXiv preprint arXiv:2107.12931*, 2021.

David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. In *International conference on machine learning*, pp. 387–395. PMLR, 2014.

Freek Stulp, Evangelos A Theodorou, and Stefan Schaal. Reinforcement learning with sequences of motion primitives for robust manipulation. *IEEE Transactions on robotics*, 28(6):1360–1370, 2012.

Giovanni Sutanto, Katharina Rombach, Yevgen Chebotar, Zhe Su, Stefan Schaal, Gaurav S Sukhatme, and Franziska Meier. Supervised learning and reinforcement learning of feedback models for reactive behaviors: Tactile feedback testbed. *arXiv preprint arXiv:2007.00450*, 2020.

Minija Tamosiunaite, Bojan Nemec, Aleš Ude, and Florentin Wörgötter. Learning to pour with a robot arm combining goal and shape learning for dynamic movement primitives. *Robotics and Autonomous Systems*, 59(11):910–922, 2011.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033. IEEE, 2012.

Mel Vecerik, Todd Hester, Jonathan Scholz, Fumin Wang, Olivier Pietquin, Bilal Piot, Nicolas Heess, Thomas Rothörl, Thomas Lampe, and Martin Riedmiller. Leveraging demonstrations for deep reinforcement learning on robotics problems with sparse rewards. *arXiv preprint arXiv:1707.08817*, 2017.

Mel Vecerik, Oleg Sushkov, David Barker, Thomas Rothörl, Todd Hester, and Jon Scholz. A practical approach to insertion with variable socket position using deep reinforcement learning. In *2019 international conference on robotics and automation (ICRA)*, pp. 754–760. IEEE, 2019.

Markus Wulfmeier, Abbas Abdolmaleki, Roland Hafner, Jost Tobias Springenberg, Michael Neunert, Tim Hertweck, Thomas Lampe, Noah Siegel, Nicolas Heess, and Martin Riedmiller. Compositional transfer in hierarchical reinforcement learning. *arXiv preprint arXiv:1906.11228*, 2019.

Markus Wulfmeier, Dushyant Rao, Roland Hafner, Thomas Lampe, Abbas Abdolmaleki, Tim Hertweck, Michael Neunert, Dhruva Tirumala, Noah Siegel, Nicolas Heess, et al. Data-efficient hindsight off-policy option learning. In *International Conference on Machine Learning*, pp. 11340–11350. PMLR, 2021.

Fan Xie, Alexander Chowdhury, M Kaluza, Linfeng Zhao, Lawson LS Wong, and Rose Yu. Deep imitation learning for bimanual robotic manipulation. *arXiv preprint arXiv:2010.05134*, 2020.

Kevin Zakka, Andy Zeng, Pete Florence, Jonathan Tompson, Jeannette Bohg, and Debidatta Dwibedi. Xirl: Cross-embodiment inverse reinforcement learning. *arXiv preprint arXiv:2106.03911*, 2021.

Guoyu Zuo, Qishen Zhao, Jiahao Lu, and Jiangeng Li. Efficient hindsight reinforcement learning using demonstrations for robotic tasks with sparse rewards. *International Journal of Advanced Robotic Systems*, 17(1):1729881419898342, 2020.