Dynamic Object Masks as Goal Representations for Visual Goal-Conditioned Reinforcement Learning

Fahim Shahriar, Cheryl Wang, Alireza Azimi, Gautham Vasan, Hany Hamed, Abhishek Naik, A. Rupam Mahmood, Colin Bellinger

Keywords: reinforcement learning, robotics, goal conditioning

Summary

Goal-conditioned reinforcement learning (GCRL) offers a unified way to pursue diverse tasks, yet most existing methods rely on state- or position-based goal representations that are unavailable in real-world robotics. Robots operating in warehouses, agriculture, or laboratory environments rarely have access to privileged goal states, object positions, or future observations, limiting the practicality of current GCRL approaches. We propose a dynamic mask-based goal representation that provides simple, object-agnostic visual cues for vision-based navigation and manipulation. At each timestep, an image-based goal detector produces a goal mask using standard image processing, task-specific object recognizers, or pretrained detectors such as Detic or Grounding DINO. These masks specify the spatial target without requiring privileged goal states, enabling broad applicability and strong generalization to unseen objects. Our method improves stability and sample efficiency in GCRL, achieving a 99% success rate in reaching both training and novel objects with Franka and UR10e robotic arms and faster learning in simulated navigation. We further demonstrate learning from scratch and sim-to-real transfer on both robotic arms, highlighting the effectiveness of our approach for real-world RL tasks. Our code is available at https://github.com/fahimfss/GCRL.

Contribution(s)

1. This paper introduces object masking as a goal representation strategy for visual GCRL in vision-based reaching and manipulation tasks.
   Context: For visual GCRL tasks, prior approaches face issues such as poor generalization to unseen objects (Nair et al., 2018; Cong et al., 2022; Yenamandra et al., 2023), slow convergence, and the need for specialized cameras (Dong et al., 2022). Our approach is significantly simpler, yet more efficient in learning visual manipulation and reaching tasks.
2. We introduce an effective mask-based reward signal that minimizes the need for privileged information.
   Context: Distance-to-target is a common reward strategy in visual GCRL, used to address the sample inefficiency of sparse rewards (Nair et al., 2018; Plappert et al., 2018; Cong et al., 2022). However, it requires privileged object positions that are difficult to obtain in the real world (Dong et al., 2022). Our mask-based reward eliminates distance-to-target computation, significantly simplifying the learning system.
3. We demonstrate that using our approach, visual GCRL agents can learn to pick up objects in a cluttered environment without requiring prior training data.
   Context: Visual pick-up requires precise end-effector alignment across diverse object shapes and sizes. Our region-of-interest masks guide the agent to good grasp positions.
4. We showcase the effectiveness of our proposed method on sim-to-real and real-world learning-from-scratch tasks.
   Context: Real-world deployment is harder than simulation due to detection errors and time constraints. Mask-based GCRL with pretrained detectors is effective in both tasks.

Dynamic Object Masks as Goal Representations for Visual Goal-Conditioned Reinforcement Learning

Fahim Shahriar[1,4], **Cheryl Wang**[2], **Alireza Azimi**[1,4], **Gautham Vasan**[1,4], **Hany Hamed**[1,4,8], **Abhishek Naik**[5], **A. Rupam Mahmood**[1,4,7], **Colin Bellinger**[3,6,*]
{fshahri1, sazimi, vasan, hhamed1, armahmood}@ualberta.ca,
huiyi.wang@mail.mcgill.ca, abhishek.naik@nrc-cnrc.gc.ca,
colin.bellinger@uottawa.ca

[1]**University of Alberta**
[2]**McGill University**
[3]**University of Ottawa**
[4]**Alberta Machine Intelligence Institute (Amii)**
[5]**National Research Council Canada**
[6]**Vector Institute**
[7]**CIFAR Canada AI Chair**
[8]**Generative Bionics**
*Work done while at the National Research Council Canada

Abstract

Goal-conditioned reinforcement learning (GCRL) offers a unified way to pursue diverse tasks, yet most existing methods rely on state- or position-based goal representations that are unavailable in real-world robotics. Robots operating in warehouses, agriculture, or laboratory environments rarely have access to privileged goal states, object positions, or future observations, limiting the practicality of current GCRL approaches. We propose a dynamic mask-based goal representation that provides simple, object-agnostic visual cues for vision-based navigation and manipulation. At each timestep, an image-based goal detector produces a goal mask using standard image processing, task-specific object recognizers, or pretrained detectors such as Detic or Grounding DINO. These masks specify the spatial target without requiring privileged goal states, enabling broad applicability and strong generalization to unseen objects. Our method improves stability and sample efficiency in GCRL, achieving a 99% success rate in reaching both training and novel objects with Franka and UR10e robotic arms and faster learning in simulated navigation. We further demonstrate learning from scratch and sim-to-real transfer on both robotic arms, highlighting the effectiveness of our approach for real-world RL tasks. Our code is available at https://github.com/fahimfss/GCRL.

1 Introduction

Many real-world robotic tasks, such as sorting objects in e-commerce warehouses, fruit picking, or visual navigation to a target site, involve solving a multi-goal problem. These tasks require the agent to act in a particular way among numerous options to achieve various desired outcomes. Goal Conditioned Reinforcement Learning (GCRL) enables multi-goal learning by allowing the agent to acquire the necessary skills to solve a range of objectives, using a unified policy.

Goal representation plays a key role in GCRL, affecting essential components of robot learning such as convergence speed, learned behavior, and generalization capabilities (Schaul et al., 2015;

Liu et al., 2022). Much of the GCRL literature relies on representations that assume access to privileged state information or desired future observations, such as vectorized position or orientation and target state images (Andrychowicz et al., 2017; Plappert et al., 2018; Nair et al., 2018; Liu et al., 2022). In many real-world robotic tasks, however, such information is not readily available. Researchers adapting GCRL to real-world manipulation have used representations such as 3D target position (Plappert et al., 2018), VAE-encoded goal images (Nair et al., 2018; Cong et al., 2022), and text-based embeddings (Yenamandra et al., 2023). While target position provides information-rich goal conditions that enable efficient learning, obtaining it requires specialized depth cameras and processing techniques such as feature matching and outlier rejection (Dong et al., 2022). The accuracy of these estimates is significantly impacted by the distance, reflectivity, and transparency of the target (Thalhammer et al., 2024), as well as unpredictable lighting conditions (Al-Selwi et al., 2025). Image- and text-based representations are easier to apply, but we find they limit the agent's ability to generalize to new targets (Section 7).

We propose a mask-based goal representation that is well-suited for multi-goal and visual robotic reaching, manipulation, and navigation tasks. Our method adds a binary channel to the agent's visual input specifying the target. The mask is updated at each time step to provide the agent with object-agnostic visual cues and progress toward the target. Our experiments demonstrate that this representation matches or outperforms existing methods on visual reaching tasks with both seen and unseen targets, showing good generalization. Complementary to prior work on exploration in GCRL, we focus on goal representation and reward design for real-world robotic tasks, aiming to reduce reliance on privileged 3D information and enable faster, more reliable learning.

Beyond goal representation, reward design is an equally important challenge in GCRL. Sparse rewards are particularly difficult, since the agent may need to execute a long sequence of correct actions before receiving positive feedback (Vasan et al., 2024). This leads to poor sample efficiency and the failure to learn complex tasks. Dense reward, particularly distance-to-goal-based reward, is commonly used to mitigate this problem (Trott et al., 2019). However, it often introduces new local optima that strongly depend on the environment and task definition, ultimately preventing agents from learning the optimal behavior (Trott et al., 2019).

In this work, we show that mask size can be used as an effective dense reward signal, since the target mask changes size and location based on the position of the target throughout the episode (Figure 4). Using mask-based reward removes the dependency on computing the target's 3D world position, significantly simplifying real-world implementation. Mask-based goal conditioning and reward signals form a novel combination that, to our knowledge, has not been explored in prior work. In contrast to using masks only as a fixed goal representation, our mask is updated over time to track the target and provide a dense progress signal via its size. Our approach focuses on egocentric tasks, where the robot perceives the environment from its onboard camera as it physically moves toward the target. The mask-based goal representation works with any camera setup, but the mask-size reward assumes an egocentric view, where the mask grows as the robot approaches the target. For other camera configurations, the mask can still be incorporated into the reward signal in a task-specific way.

We demonstrate that pretrained object detection models such as Detic (Zhou et al., 2022) and Grounding DINO (Liu et al., 2024) can be used for mask generation in real-world tasks. We show the feasibility of using these pretrained models on two real-world applications: a) UR10e arm-based sim-to-real transfer, and b) learning from scratch using the Franka Panda arm. For the sim-to-real setup, we additionally use a novel image interpolation technique in training to mitigate visual domain shift during real-world policy transfer.

Our main contributions include: (1) the introduction of object masking as a goal representation strategy in vision-based GCRL; (2) an effective mask-based reward signal that minimizes the need for privileged information; (3) performing visual pick-up utilizing region-of-interest-based masks; (4) demonstrating the effectiveness of the proposed method in sim-to-real and real-world learning from scratch tasks.

2 Related Work

Goal Representations

Vectorized position or orientation of an object or the robot is a common goal representation system used in the GCRL literature (Plappert et al., 2018; Florensa et al., 2018; Tang & Kucukelbir, 2021). Other vectorized goal representations include one-hot encoding (Kalashnikov et al., 2022; Wu et al., 2021), and object text description to CLIP embeddings (Yenamandra et al., 2023). For visual tasks, a common approach is to use variational autoencoders (VAE) to encode visual inputs such as RGB images into a latent space (Nair et al., 2018; Cong et al., 2022). However, traditional image-based goal conditioning tasks require a predefined goal state, which is unfeasible for tasks where such conditions are unknown and require exploration by the agent. In contrast, masks are simpler to generate and provide direct state-dependent feedback to the agent.

Mask-based Goal Representation: Object masking has been applied by researchers to represent target objects and goal positions under stationary camera configurations (Wang et al., 2021; Cong et al., 2022). These approaches train VAEs to encode images into latent vectors prior to GCRL training and use a complex latent state-based reward. On the other hand, our approach learns directly from images, uses masks to track progression towards the target, and uses mask-size-based rewards, which greatly simplifies the learning system.

Yenamandra et al. (2023) used segmentation masks and depth data for target object finding and reaching tasks. Stone et al. (2023) used masks with single activated pixels to denote target objects, allowing good generalization to objects of any shape and unseen objects. Mendonca & Pathak (2024) used object masks to obtain the target object's point-cloud information to perform mobile manipulation tasks. Our proposed system uses dynamic masks to achieve faster convergence and utilizes mask-based dense reward signals, which none of the approaches mentioned above employ.

Sparse Reward Challenge for GCRL

Sparse reward-based GCRL tasks often face an efficiency issue, as it takes a very long time for the reward signal to propagate from the goal states to the starting states (Andrychowicz et al., 2017; Liu et al., 2022). One approach to tackle this challenge is the use of distance-based dense rewards in either Euclidean space or latent space (Nair et al., 2018; Plappert et al., 2018; Cong et al., 2022). Nevertheless, this simple reward shaping is vulnerable to local optima that stem from a poor exploration of the environment (Vasan et al., 2024). Alternatively, hindsight experience replay (HER) addresses the sparse reward issue by re-assigning failed states as alternative goals and providing denser reward signals to speed up learning (Andrychowicz et al., 2017). However, the HER-based approach still suffers from sample efficiency in complex environments (Luo et al., 2023). For vision-based tasks, imagined sub-goals between the current and final goal state can simplify learning for long-horizon tasks (Nair et al., 2018; Chane-Sane et al., 2021), but these predictions sometimes result in infeasible robotic states or unrealistic images.

Object Detection Models in Robotic Tasks

For vision-based robotic tasks, a common practice is to use object detection models to locate objects of interest. YOLO models are a common choice for researchers, due to their fast inference times and good accuracy (Uppal et al., 2024; Zhang et al., 2023). However, YOLO models support a fixed number of predefined object detections. As a result, users are required to retrain or fine-tune YOLO models to support novel objects (Chen et al., 2023; Song et al., 2021). Recently, open vocabulary-based object detection models have gained popularity in the robotics community due to zero-shot detection capabilities. Open vocabulary object detectors, such as Detic and Grounding DINO, have been used in various robotic manipulation tasks (Mendonca & Pathak, 2024; Stone et al., 2023; Yenamandra et al., 2023). In this work, we utilize pretrained open-vocabulary detection models to generate masks for goal conditioning and reward computation and to assess their feasibility in real-world tasks.

3 Background

In standard RL, the objective is to train a policy to maximize the expected return. GCRL policies have a similar objective, but the expected return in each episode is conditioned on a predetermined goal $g \in \mathcal{G}$. GCRL can be formally represented by the goal-augmented Markov decision process (GA-MDP). At each time-step t, the agent samples an action $A_t \in \mathcal{A}$ based on the current state $S_t \in \mathcal{S}$, and the current goal g, using a policy π_ϕ, $A_t \sim \pi_\phi(\cdot \mid S_t, g)$. The agent receives a reward R_{t+1} and the next state S_{t+1} at the next time-step $t+1$ based on a transition function P, $S_{t+1}, R_{t+1} \sim P(\cdot, \cdot \mid S_t, A_t, g)$. The episode ends when the agent reaches the goal, exceeds the allowed time limit, or leaves the allowed state space.

In this work, we evaluate the mask-based goal representation using on and off-policy methods. A brief overview of each method is provided below. Readers are directed to the original papers (Haarnoja et al., 2018; Schulman et al., 2017), for more details.

Soft Actor-Critic (SAC) is widely used for vision-based robotic tasks due to its higher sample efficiency. In GCRL, SAC learns the parameters ϕ of a policy π_ϕ to maximize a trade-off between the policy's entropy and expected return. SAC uses a soft-Q function (critic) $Q_\theta(S_t, A_t, g)$, parameterized by θ, to estimate expected returns and the entropy bonus based on the current state, action, and goal.

Proximal Policy Optimization (PPO) is another frequently used RL algorithm for robotics tasks, particularly for sim-to-real applications. PPO updates policy parameters ϕ by maximizing the PPO-Clip objective and constraining changes to the new policy, improving the learning stability. PPO uses an advantage estimate $\hat{A}_t$ that denotes how good the selected action a_t is in the current state S_t, and learns to increase the probabilities of selecting good actions. The critic function $V_t(S_t, g)$ predicts the expected return for a given state and goal, and it is used to generate accurate advantage estimates.

For vision-based GCRL tasks, the state S_t generally consists of an image I_t and a vector v_t, $S_t = \langle I_t, v_t \rangle$. An image encoder $f_\psi : \mathbb{R}^{H \times W \times C} \to \mathbb{R}^d$ is used to process the image into a lower dimensional latent vector; here ψ represents the learnable parameters, d is the latent dimension, and H, W, and C denote the height, width, and the number of channels of the input image respectively. Both the policy π_ϕ and the critic Q_θ use f_ψ to process the image, but the parameter ψ is only updated during the critic update.

Reward Systems: In GCRL, sparse rewards are only provided to the agent upon task completion:

$$R^{sparse}(s_t, a_t, g) = R^{\text{term}} \quad \{\text{goal reached}\} \tag{1}$$

Distance-to-goal-based dense reward systems are often used as an alternative to the sparse reward system:

$$R^{dist}(s_t, a_t, g) = \begin{cases} -d(S_t, g) & \text{if } d(S_t, g) > \epsilon, \\ -d(S_t, g) + R^{\text{term}} & \text{if } d(S_t, g) \leq \epsilon \end{cases} \tag{2}$$

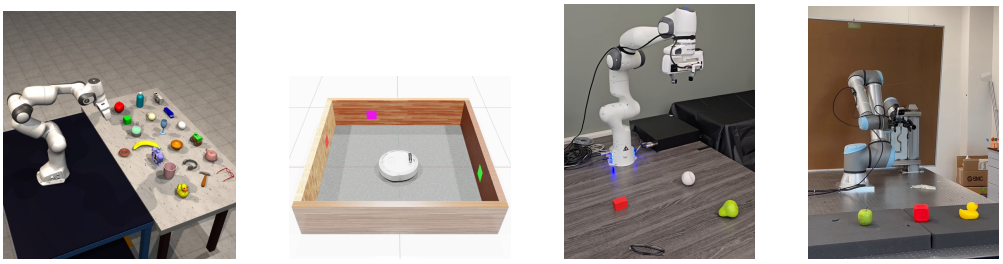
Here, $d(S_t, g)$ denotes the distance to the goal for the state S_t, ϵ is the minimum distance to the goal, and R^{term} denotes a terminal reward. In our simulated experiments, the distance is computed between the midpoint of the gripper fingers and the reference frame of the target object.

4 Mask-based Visual GCRL

Our proposed mask-based GCRL system includes a novel mask-based goal representation and reward formulation, along with targeted algorithmic enhancements for stable visual learning.

4.1 Mask-based Goal Representation

Masks are single-channel, binary images with the goal represented by a region of activated pixels. At each time step t, a mask m_t is created using the image from a camera sensor I_t^{cam}. The ground



(a) Franka-MuJoCo environment. Example object names: chocolate donut, banana, and red apple.

(b) Create3-Isaac Sim environment.

(c) Learning from scratch setup with the Franka arm.

(d) Sim-to-real setup with the UR10e arm.

Figure 1: Simulated and real-world environments.

truth mask can be generated using privileged positional information of the target in simulation. Alternatively, object color-based or shape-based segmentation or pretrained object detectors can be used to generate masks. The mask is then appended to I_t^{cam} on the channel axis, to form $I'_t = [I_t^{cam}\ m_t]_C$. Three of the most recent such images are stacked on the channel axis to form I_t, where $I_t = [I'_t\ I'_{t-1}\ I'_{t-2}]_C$, and have the shape $90 \times 160 \times 12$, representing height, width, and channel.

Mask-based goal representation offers several advantages (demonstrated in Section 7): (1) the mask is calculated at each time step, providing richer signals and enabling faster training. (2) It is object-agnostic and facilitates better generalization to novel objects. (3) Masks can be generated in multiple ways, and we demonstrate three such approaches: (a) using privileged object position, (b) using object detection models, and (c) using color filtering.

4.2 Mask-based Reward Formulation

The mask-based reward system serves as an alternative to the distance-based reward. The mask-based reward system replaces the distance to the goal term, $-d(S_t, g)$, in Equation (2) with R^{mask}:

$$R' = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \mathbb{1}\{m_t^{ij} > 0\} \tag{3}$$

$$R^{mask} = \frac{2}{1 + e^{-10\,R'}} - 1, \quad R^{mask} \in [0, 1] \tag{4}$$

Here, H and W are the height and width of the mask m_t, respectively. Equation (3) computes a normalized reward, R', based on the activated pixels in m_t. Equation (4) applies a scaled sigmoid function to R', which increases its value for small inputs, specifically when $R' < 0.5$. This scaling provides stronger feedback to the agent, particularly when the goal is further away and the changes in the mask sizes are small. Although the reward can be used without this sigmoid shaping (i.e., directly from R'), applying it amplifies small gains for $R' < 0.5$ and improves training speed.

4.3 Visual Learning Architectures

SAC: Our SAC implementation utilizes asynchronous network updates and replay buffer sampling to enable efficient learning in simulation and the real world (Yuan & Mahmood, 2022). To stabilize learning from pixels and to achieve better learning performance, we incorporated random shift (Yarats et al., 2021), used an ensemble of five Q-Networks (Chen et al., 2021), and weight clipping (Elsayed et al., 2024) with $\kappa = 2$ and $s_l = \sqrt{2}$ in our SAC implementation. All experiments in our work, except sim-to-real on UR10e (see Section 8.1), utilize SAC-based GCRL agents.

PPO: For the sim-to-real application, we used the Stable Baselines 3 implementation of PPO (Raffin et al., 2021). We chose PPO because it is widely used in sim-to-real tasks (Zhao et al., 2020). We used a novel image interpolation technique for training the PPO agent, which was essential for addressing visual domain shift during real-world policy transfer. More details are in Section 8.1.

5 Environments

We evaluate our method across three simulated environments covering both manipulation and navigation tasks. Applications to physical robots, including sim-to-real transfer and learning from scratch, are discussed in Section 8.

Franka-MuJoCo Environment

We created a tabletop setup with a Franka Panda robotic arm collected from MuJoCo Menagerie (Zakka et al., 2022), and 20 everyday objects using RoboHive, a MuJoCo-based robot learning framework. Figure 1a shows the simulated environment with all the objects placed on the tabletop. We attached a camera near the end-effector of the arm to facilitate vision-based reach, grasp, and lift. During training, three to five objects are arbitrarily selected from a pool of fifteen objects in each episode. The remaining five objects are held out as test objects. The selected objects are placed randomly on the table in front of the arm, and one object is randomly selected to be the target.

The goal of the GCRL agent for this task is to control individual joints of the Franka arm using low-level velocity commands to reach the target object within 150 time steps. The common state space across experiments contains a stack of the three most recent camera images of the shape $90 \times 160 \times 9$, the joint positions, and the last action.

UR10e-MuJoCo Environment

For this environment, we replaced the Franka arm with the UR10e arm (MuJoCo Menagerie). All other aspects of the environment remained the same. The joint lengths, angular velocities (range and speed), and orientation of the joints for the UR10e arm are substantially different compared to the Franka arm. Furthermore, the UR10e arm is more sensitive to actions compared to the Franka arm, making the learning task slightly more difficult.

Create 3-Isaac Sim Environment

We created a square arena in the Isaac Sim simulator (NVIDIA, 2025) and used an iRobot Create 3 mobile robot to test the visual navigation performance of the five goal conditioning methods. In each episode, we placed randomly chosen colored stickers (magenta, green, red, or blue) on each wall of the arena, and assigned a random pose to the Create 3 (Figure 1b). The GCRL agent needs to control the linear and angular velocities of the Create 3 to reach the selected target sticker in 200 time steps. We used three stacked images and a vectorized history, v_t, of the last 15 actions made by the agent. This allows the agent to have access to detailed information about the past motions of the robot. For this environment, we used color-filtering techniques to generate the masks.

In addition to these environments, in Section 8, we demonstrate sim-to-real transfer and learning from scratch on the UR10e and Franka Panda robots.

6 Experimental Setup

We aim to answer the following research questions through our experiments:

1. *Does the mask-based system enable (a) faster training, and (b) better generalization to unseen objects?*
2. *Is mask-based reward a viable alternative to distance-based dense reward?*
3. *Is it possible to use mask-based GCRL and reward systems for performing complex tasks, such as picking up?*

To address these research questions more specifically, we evaluate our approach against existing goal representations. The **target state** goal representation adds a pre-existing close-up RGB image

of the target with size $90 \times 160 \times 12$ to the stacked image states, I_t. **One-hot** encoding, **3D position**, and **CLIP embedding** are vector-based goal representations and appended to the vector part of the state space, v_t. One-hot encoding uses a 20D zero vector with a single activated index that identifies the target. CLIP embedding uses a 512D vector generated using a pretrained CLIP encoder based on a text description of the target object (Yenamandra et al., 2023). 3D position uses the privileged world position of the target collected from the simulator.

We design a sequence of experiments that address each of the research questions above.

Experiment 1 (a)

To demonstrate the learning efficiency of the mask-based system, we trained mask-based GCRL agents in *Franka-MuJoCo*, *UR10e-MuJoCo*, and *Create3-Isaac Sim* environments and compared the learning performance against one-hot encoding, 3D position, CLIP embeddings, and target state-based goal representation systems. Note that the mask-based rewards are not used in these experiments. We used the distance-based dense reward defined in Equation (2), with ϵ set to 10 cm, and $R^{term} = 5$. Each experiment used eleven seeds, and we trained each agent for 300,000 steps. We used a replay buffer of size 300,000, and the initial 5,000 steps were used to pre-fill it.

Experiment 1 (b)

We tested the reaching capabilities of the agents trained in *Experiment 1 (a)* for both seen and unseen objects for the *Franka-MuJoCo* and *UR10e-MuJoCo* environments. We ensured similarity across the experiments by selecting the five test objects deterministically based on the seed. To collect the reaching accuracy, for each seed, we carried out twenty-five trials for each object, resulting in a total of 5,500 trials for each goal representation system. A trial is considered successful when the end-effector reaches within ϵ (10 cm) of the target object before the episode ends.

Experiment 2

In this experiment, we show that mask-based reward is a viable alternative to the distance-based reward. We trained mask-based GCRL agents with mask-based rewards on the *Franka-MuJoCo* and *UR10e-MuJoCo* environments and compared their performance against the mask-based GCRL agents with distance-based rewards from *Experiment 1 (a)*. The mask-based reward is calculated according to Equation (4), with an added -1 reward per step to facilitate faster training. We used the same settings defined in *Experiment 1 (a)* for the two environments.

Experiment 3

In this experiment, we demonstrate that mask-based GCRL agents can learn a complex visual pick-up task from scratch. The objective of the GCRL agent is to grab a target object placed randomly on a table and lift it up 30 cm above the table's surface. We used a modified version of the *UR10e-MuJoCo* environment for this experiment. We used six objects out of 20 for the pick-up targets, which are: apple, green block, donut, toy duck, banana, and white egg. These objects were slightly scaled down in size to facilitate easier grasping. The other 14 objects were used as distractors during training, and in each episode, 2 to 4 objects were randomly sampled and placed on the table.

We compared the performance of mask-based goal conditioning (GC) and mask-based reward system with mask-based GC and distance-based reward system, and 3D position-based GC and distance-based reward system. For the mask-based GC, we selected the part of the image between the gripper fingers as the region of interest (ROI). This is where the object is expected to be during a successful grasp (Figure 4). The agent received the mask-size reward only if the target mask was inside the ROI. This approach allowed the agent to better align the robotic arm's fingers with the target for good grasping, resulting in more successful pick-ups. For all three systems, we included the touch information of the target object and the arm's fingers using a binary vector of size 2. For the mask-based system, once both fingers touched the target object, the mask activation was removed,

and only a blank mask was provided. We trained the agents for one million steps, and the first 5000 steps were used to pre-fill the replay buffer, which had a size of 300,000.

We used the following reward function for this experiment:

$$R^{\text{pick}} = -1.1 + R^{\text{reach}} + R^{\text{contact}} + R^{\text{lift}} + R^{\text{goal}} \tag{5}$$

Here, R^{reach} is set to distance-based reward (Equation 2) or mask-based reward (Equation 4), and only applied before the agent touches the target with both fingers. R^{contact} is 10 for the first time both fingers touch the target and 0.1 for subsequent touches with both fingers. R^{lift} is set to $1 - 3.3 \cdot (D)$, where D is the current distance to the target height. R^{lift} is used after both fingers touch the target object. R^{goal} is a one-time reward of 10 upon reaching the target height.

After the training was completed, we used the trained policy to perform 50 pick-up trials for each of the six objects we used during training.

7 Results

Experiment 1 (a)

Figure 2a, 2b, and 2c show the results for *Experiment 1 (a)*. The bold lines show the learning performance of the median run selected based on the episodic return, and the faint lines represent the remaining seeds following Tanaka & Mahmood (2026). In all three experiments, the mask-based system achieved the fastest learning performance. For the *Franka-MuJoCo* and the *UR10e-MuJoCo* runs, the mask-based GCRL agent learned the optimal policy similar to the 3D position system. For the *Create3-Isaac Sim* experiments, the overall performance of the mask-based approach was slightly lower than the 3D-position method. This is because, for the 3D position, the agent always had access to the privileged target position, allowing it to learn to reach the target directly, even when not in view. On the other hand, for the mask-based method, the agent learns two skills: first, to find the target by rotating, and second, to move close to the target. Approaches such as one-hot encoding, CLIP embeddings, and target state showed poor learning performance, revealing the importance of selecting the appropriate goal conditioning method to achieve fast policy convergence.

Experiment 1 (b)

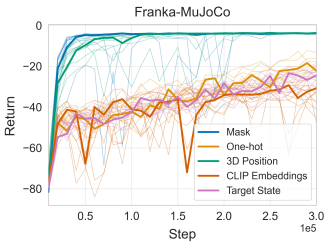
Figure 2d shows the results for *Experiment 1 (b)*. The bars in the plot represent the median reaching accuracy, and the whiskers represent the first and third quartiles. The mask-based GCRL agents achieved the highest performance for train and test objects, with 99.9% average accuracy for both. 3D position, on the other hand, achieved 98.9% and 98.8% average accuracy for train and test, respectively. This result shows that the agents learned to utilize the masks effectively for both train and test objects.

Experiment 2

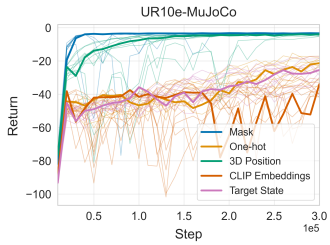
In Figure 2e, we used average episode steps instead of return for comparison, as the distance-based and mask-based rewards are not directly comparable. The median runs are selected based on average episodic steps. The mask-based reward system performed similarly to the distance-based reward system while having higher stability during training. This result shows that the mask-based reward system is a viable alternative to the distance-based reward.

Experiment 3

This experiment examines whether mask-based goal conditioning supports learning a complex pick-up task across diverse object shapes and sizes. Figure 3a shows the pick-up accuracy during training, and Figure 3b shows the pick-up accuracy of the trained agents. Nine out of eleven seeds for the mask-based GC and rewards system achieved over 90% pick-up success. On the other hand, only

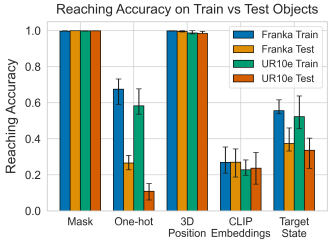


(a) Franka-MuJoCo experiments: The mask-based GCRL agent shows faster learning and similar final performance compared to the 3D position-based agent.

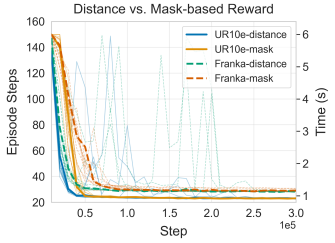


(b) UR10e-MuJoCo experiments: The mask-based GCRL agent shows similar learning and performance behavior as observed in 2a.

(c) Create3-Isaac-Sim: The median mask-based agent shows faster learning, but the median 3D position-based agent has better final performance due to having access to the privileged target position.

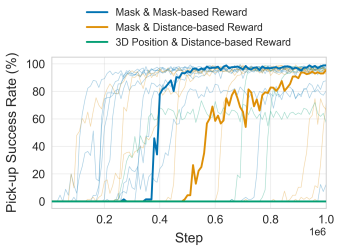


(d) Reaching accuracy of trained agents (Franka & UR10e MuJoCo environments). The mask-based agents have the best reaching accuracy, closely followed by the 3D position-based agents.

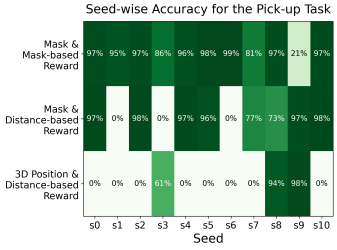


(e) Comparison of distance-based vs. mask-based reward systems. The final performance between these systems is similar.

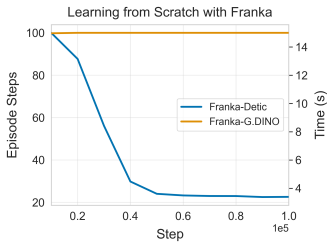
Figure 2: Simulated experiment results. Bold lines show the median run (episodic return), and faint lines indicate other seeds.



(a) Pick-up success rate during training using the UR10e arm. The mask-based systems performed better than the 3D-position-based system.



(b) Success rate of the trained agents in the post-training trials.



(c) Real-world experiments with Detic and Grounding DINO. For our real-world setup (1c), Detic-based agent performed much better than Grounding DINO.

Figure 3: (a) Pick-up success rate during training (UR10e MuJoCo), where bold lines show the median runs. (b) Post-training success per seed (50 trials per object). (c) Learning curves for the real-world learning from scratch task.

two seeds for the 3D position-based GC and distance-based reward system achieved over 90% success. This result shows that the ROI based masks guided the arm to a good grasp position (Figure 4), which resulted in higher pick-up accuracy. The dense distance-based reward (Equation 2) provides no incentive for the agent to align its end-effector with the target for grasping. In the failed runs, the agent hovered near the target object but did not learn to grasp or lift it. This experiment demonstrates that the mask-based system can be trained to perform complex visual pick-up tasks from scratch and shows the robustness of our proposed system.

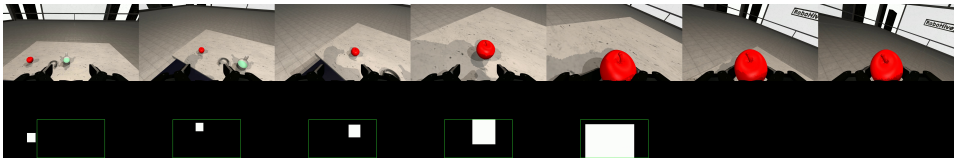


Figure 4: A trained mask-based GC and reward agent successfully picks up the apple object. The green rectangles in the masks (bottom images) represent the ROI (shown only for illustration). In the last two frames, the agent lifts up the apple object after a successful grasp.

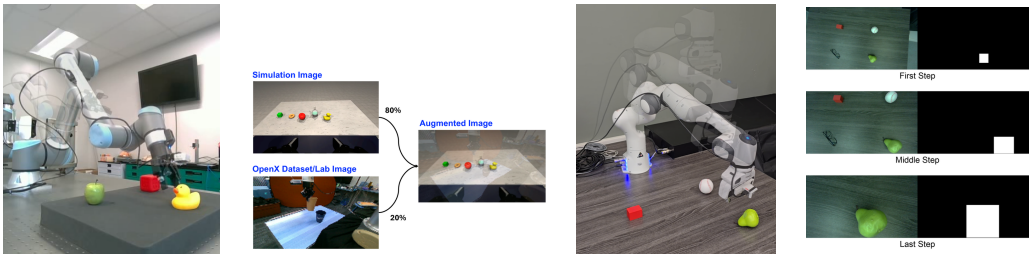
8 Real-world Applications

We demonstrate the real-world applicability of our system through two applications: sim-to-real transfer and learning from scratch on physical robots.

8.1 Sim-to-Real Application with UR10e

To assess the real-world applicability of our mask-based GCRL system, we performed a sim-to-real transfer of the proposed algorithm onto a physical UR10e robot (Figure 5a). The policy was trained using PPO in the *UR10e-MuJoCo* environment with the addition of domain randomization. We trained PPO with a custom multi-input policy using three-frame stacking and four parallel environments. We used ground-truth masks in simulation, where the privileged target information is readily available, and utilized Grounding DINO for mask generation in the real world. Grounding DINO uses (480×848) resolution images from a wrist-mount RealSense D435 camera. These are rescaled to 120×212 as input to the PPO CNN network. The best-performing policy from 10 seeds was selected for zero-shot transfer. The agent operates at a wall-clock time step of about 120 ms, dominated by Grounding DINO inference. A background control thread sends `servoJ` commands at 500 Hz, and a successful reach takes under 30 seconds.

Domain randomization techniques are used to mitigate the discrepancies between simulated environments and the real world. The UR10e robot's joint readings received Gaussian noise $\sim \mathcal{N}(0, 0.05)$ while its initial qpos varied by ± 0.2 for the first 5 DoFs. RGB images were augmented with contrast, saturation, brightness, and Gaussian blur (σ), all randomized within [0.8, 1.2] using Kornia (Riba et al., 2020). As shown in Figure 5b, we introduce a novel image augmentation process to reduce the image domain shift by linearly interpolating simulation images with randomly selected images from diverse real-world robotics tasks during training. This image augmentation was found to be essential for policy transfer.



(a) UR10e arm reaching for the rubber duck, controlled using a zero-shot sim-to-real GCRL agent.

(b) Image augmentation during policy training mixed 80% of a simulation image with 20% of a random real-world image.

(c) Franka arm reaching for a pear in the learning from scratch task using Detic.

(d) Franka camera images and masks created using Detic at different stages of the episode to reach the pear.

Figure 5: Sim-to-real using UR10e and real-world learning from scratch using Franka.

We evaluate our system using in-distribution objects (green apple, red cube, yellow rubber duck) and two out-of-distribution test sets (Demo 1: toy robot, frog, water bottle; Demo 2: toy robot, plush dinosaur, water bottle). These objects were selected to provide variation in mass, size, color,

shape, and surface texture, testing whether the reaching behavior learned in simulation generalizes to visually diverse real-world objects. The objects were placed at pre-selected positions whose coordinates in the robot's base frame were measured in advance. The distance in each trial is computed automatically from the final gripper midpoint to the recorded object position.

Table 1: Performance Comparison Between In-Distribution and Out-of-Distribution (OOD) Objects. Results obtained with a frame skip of 4 on the UR10e, 25 trials per object set. We classify a trial as successful if the end effector reaches within 15 cm of the target object.

Metric	**In-Distribution**	**OOD (Demo1)**	**OOD (Demo2)**
Success Rate (%)	96	92	92
Distance (m)	0.076 ± 0.035	0.13 ± 0.07	0.092 ± 0.050

Table 1 demonstrates the successful real-world transfer of the policy, achieving success rates comparable to simulation with ground-truth masks (89% in-distribution, 90% OOD), indicating no degradation from sim-to-real transfer. The end-effector can reach within 15 cm of the target objects, regardless of their size and location.

We performed two additional tasks in the real world using the pretrained policy. We conducted pick-and-place by coupling a heuristic grab, lift, and place mechanism with the reaching policy. We also performed a tracking task by placing a target object in front of the arm and moving it gradually. The agent executed both tasks successfully, showing the robustness of the pretrained policy and Grounding DINO's ability to generate masks for GCRL in the real world.

8.2 Learning from Scratch with Franka

We show the application of the mask-based system (goal representation and reward) in a real-world learning from scratch task using the Franka Panda arm (Figure 1c). We simplified the reaching task in the real world by using four stationary objects placed on a table in front of the arm: a red block, a baseball, a pear (artificial), and eyeglasses. We used Detic and Grounding DINO models for object detection and used slightly lower resolution images (450×800) to facilitate faster inference times. We used 100 steps per episode and trained for 100,000 steps. The termination distance, ϵ, was set to 13 cm from the target. We used an Intel RealSense D405 camera attached near the end-effector of the arm to capture images. We performed the training on a computer equipped with a Ryzen Threadripper 3970X processor and an Nvidia RTX 3090 GPU. We used the same SAC GCRL system that we used in the simulated experiments. The GCRL agent had to learn to reach the randomly selected target object by controlling the joint velocities of the arm. Average step times were 157 ms with Detic (131 ms inference) and 168 ms with G.DINO (158 ms inference). Total run durations were 449 and 338 minutes, respectively. The Detic agent completed more episodes (3,328 vs. 1,000) and therefore spent more time on arm resets, which take about 3 seconds each.

Figure 3c shows the results of the real-world learning from scratch experiments. The Franka-Detic agent learned an optimal reaching policy in 60,000 steps. Figure 5c shows the trained agent reaching for a pear, and Figure 5d shows the associated camera images and masks generated using Detic. On the other hand, the Grounding DINO model had severe false-positive detection issues. The agent failed to reach the target objects as it tried to exploit rewards from false detection masks. This experiment shows the possibilities and difficulties of using pretrained object detection models in real-world learning from scratch tasks.

We further evaluated the trained Detic-based GCRL agent by performing a tracking task with an unseen object. We placed a plastic banana on top of the table and moved it using a stick while the GCRL agent was controlling the arm. The agent managed to follow the banana easily, even though it was trained on reaching stationary objects. The successful tracking behavior demonstrates the robustness and the generalization capabilities of our proposed system.

9 Limitations and Future Work

Our results demonstrate that dynamic object masks provide a simple and effective goal representation and dense reward signal for visual GCRL. We outline two directions for future work that would extend the scope of this paper.

Mask quality and robustness. In this work, we used bounding-box-based masks as the goal representation. We experimented with segmentation masks as a finer alternative, but observed a slight performance degradation, likely due to reduced generalization across objects, but this requires further investigation. We note that the mask by itself enables an object-agnostic *reaching* skill rather than a fully goal-agnostic manipulation skill. In addition, the bounding box is agnostic to the nuances of the object. Downstream tasks (after reaching) involving grasping or picking may require reaching for, say, the handle of the cup instead of the middle of the cup. However, the agent also has access to information apart from the mask, such as the RGB camera images, which can help the agent learn to perform downstream tasks. The performance of our system is upper-bounded by the quality of the object detector, as observed in our real-world experiments (Figure 3c). However, the detector is a replaceable component, and our approach benefits directly from improved detection models. A useful next step is to study how sensitive the proposed system is to mask errors, such as missing regions, flickering detections, partial occlusions, noise, and false detections, and measure the impact on performance. We note that improving mask quality is more closely related to image processing than GCRL, and we leave a thorough investigation to future work.

Sim-to-real and learning from scratch in more challenging environments. We primarily evaluated indoor tabletop manipulation and navigation scenarios, where our image interpolation technique using indoor real-world data helped bridge the simulation-to-real domain gap, and learning from scratch was feasible due to the controlled nature of the setup. An important next step is to extend mask-conditioned GCRL to more challenging settings, including outdoor environments and multi-task settings such as navigation followed by manipulation, sequential pick-and-place, or outdoor mobile manipulation tasks (e.g., weeding or fruit harvesting, such as strawberry picking). For sim-to-real, this direction requires building high-quality outdoor simulation setups and collecting diverse outdoor real-world data for image interpolation, which is substantially more challenging than our current indoor setup. For learning from scratch, unstructured outdoor environments introduce additional challenges for the GCRL agent, such as a more complex state space, movement limitations, and safety concerns, which we aim to address in future work.

10 Conclusion

We introduced an object-agnostic mask-based GCRL and reward system. Experimental results suggest that our proposed system performs similarly to or better than other approaches on visual manipulation and reaching tasks. The mask-based systems showed superior reaching performance on novel objects utilizing RGB images only, achieving 99% accuracy. Moreover, we showed that dense reward based on mask size is an effective alternative to distance-to-target-based reward, diminishing the necessity for distance calculation methods. The mask-based agents learned successful pick-up behavior, with nine of eleven seeds exceeding 90% success. Finally, we showed that our proposed method is effective in real robot learning, and open-vocabulary object detection models are effective when transferring trained policies from simulation to real robots.

Acknowledgments

We thank the reviewers for their valuable feedback. We gratefully acknowledge the National Research Council of Canada's (NRC) AI for Design Challenge Program and the Canada CIFAR Chairs program for their financial support. We also thank the Digital Research Alliance of Canada for providing computational resources.

References

Metwalli Al-Selwi, Huang Ning, Yin Gao, Yan Chao, Qiming Li, and Jun Li. Enhancing object pose estimation for rgb images in cluttered scenes. *Scientific Reports*, 15(1):8745, 2025.

Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, OpenAI Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. *Advances in neural information processing systems*, 30, 2017.

Elliot Chane-Sane, Cordelia Schmid, and Ivan Laptev. Goal-conditioned reinforcement learning with imagined subgoals. In *International conference on machine learning*, pp. 1430–1440. PMLR, 2021.

Xinyue Chen, Che Wang, Zijian Zhou, and Keith W. Ross. Randomized ensembled double q-learning: Learning fast without a model. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=AY8zfZm0tDd.

Ya-Ling Chen, Yan-Rou Cai, and Ming-Yang Cheng. Vision-based robotic object grasping—a deep reinforcement learning approach. *Machines*, 11(2):275, 2023.

Lin Cong, Hongzhuo Liang, Philipp Ruppel, Yunlei Shi, Michael Görner, Norman Hendrich, and Jianwei Zhang. Reinforcement learning with vision-proprioception model for robot planar pushing. *Frontiers in Neurorobotics*, 16:829437, 2022.

Xingshuai Dong, Matthew A Garratt, Sreenatha G Anavatti, and Hussein A Abbass. Towards real-time monocular depth estimation for robotics: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(10):16940–16961, 2022.

Mohamed Elsayed, Qingfeng Lan, Clare Lyle, and A. Rupam Mahmood. Weight clipping for deep continual and reinforcement learning. *RLJ*, 5:2198–2217, 2024.

Carlos Florensa, David Held, Xinyang Geng, and Pieter Abbeel. Automatic goal generation for reinforcement learning agents. In *International conference on machine learning*, pp. 1515–1528. PMLR, 2018.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pp. 1861–1870. Pmlr, 2018.

Dmitry Kalashnikov, Jake Varley, Yevgen Chebotar, Benjamin Swanson, Rico Jonschkowski, Chelsea Finn, Sergey Levine, and Karol Hausman. Scaling up multi-task robotic reinforcement learning. In Aleksandra Faust, David Hsu, and Gerhard Neumann (eds.), *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pp. 557–575. PMLR, 08–11 Nov 2022. URL https://proceedings.mlr.press/v164/kalashnikov22a.html.

Minghuan Liu, Menghui Zhu, and Weinan Zhang. Goal-conditioned reinforcement learning: Problems and solutions. In Lud De Raedt (ed.), *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pp. 5502–5511. International Joint Conferences on Artificial Intelligence Organization, 7 2022. DOI: 10.24963/ijcai.2022/770. URL https://doi.org/10.24963/ijcai.2022/770. Survey Track.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024.

Yongle Luo, Yuxin Wang, Kun Dong, Qiang Zhang, Erkang Cheng, Zhiyong Sun, and Bo Song. Relay hindsight experience replay: Self-guided continual reinforcement learning for sequential object manipulation tasks with sparse rewards. *Neurocomputing*, 557:126620, 2023.

Russell Mendonca and Deepak Pathak. Continuously improving mobile manipulation with autonomous real-world RL. In *RSS 2024 Workshop: Data Generation for Robotics*, 2024. URL https://openreview.net/forum?id=qASoq07bXh.

Ashvin V Nair, Vitchyr Pong, Murtaza Dalal, Shikhar Bahl, Steven Lin, and Sergey Levine. Visual reinforcement learning with imagined goals. *Advances in neural information processing systems*, 31, 2018.

NVIDIA. Isaac sim. https://developer.nvidia.com/isaac-sim, 2025. Version 4.5.0.

Matthias Plappert, Marcin Andrychowicz, Alex Ray, Bob McGrew, Bowen Baker, Glenn Powell, Jonas Schneider, Josh Tobin, Maciek Chociej, Peter Welinder, et al. Multi-goal reinforcement learning: Challenging robotics environments and request for research, 2018. URL https://arxiv.org/abs/1802.09464.

Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021. URL http://jmlr.org/papers/v22/20-1364.html.

Edgar Riba, Dmytro Mishkin, Daniel Ponsa, Ethan Rublee, and Gary R. Bradski. Kornia: an open source differentiable computer vision library for pytorch. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass Village, CO, USA, March 1-5, 2020*, pp. 3663–3672. IEEE, 2020. DOI: 10.1109/WACV45572.2020.9093363. URL https://arxiv.org/pdf/1910.02190.pdf.

RoboHive. Robohive – a unified framework for robot learning, 2020. URL https://sites.google.com/view/robohive.

Tom Schaul, Daniel Horgan, Karol Gregor, and David Silver. Universal value function approximators. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1312–1320, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/schaul15.html.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017.

Qisong Song, Shaobo Li, Qiang Bai, Jing Yang, Xingxing Zhang, Zhiang Li, and Zhongjing Duan. Object detection method for grasping robot based on improved yolov5. *Micromachines*, 12(11), 2021. ISSN 2072-666X. URL https://www.mdpi.com/2072-666X/12/11/1273.

Austin Stone, Ted Xiao, Yao Lu, Keerthana Gopalakrishnan, Kuang-Huei Lee, Quan Vuong, Paul Wohlhart, Sean Kirmani, Brianna Zitkovich, Fei Xia, Chelsea Finn, and Karol Hausman. Open-world object manipulation using pre-trained vision-language models, 2023. URL https://arxiv.org/abs/2303.00905.

Haruto Tanaka and A. Rupam Mahmood. Performance variation in deep reinforcement learning, 2026. URL https://arxiv.org/abs/2606.06746.

Yunhao Tang and Alp Kucukelbir. Hindsight expectation maximization for goal-conditioned reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 2863–2871. PMLR, 2021.

Stefan Thalhammer, Dominik Bauer, Peter Hönig, Jean-Baptiste Weibel, Jose Garcia-Rodriguez, and Markus Vincze. Challenges for monocular 6-d object pose estimation in robotics. *IEEE Transactions on Robotics*, 40:4065–4084, 2024.

Alexander Trott, Stephan Zheng, Caiming Xiong, and Richard Socher. Keeping your distance: Solving sparse reward tasks using self-balancing shaped rewards. *Advances in Neural Information Processing Systems*, 32, 2019.

Shagun Uppal, Ananye Agarwal, Haoyu Xiong, Kenneth Shaw, and Deepak Pathak. Spin: Simultaneous perception interaction and navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18133–18142, 2024.

Gautham Vasan, Yan Wang, Fahim Shahriar, James Bergstra, Martin Jägersand, and A. Rupam Mahmood. Revisiting sparse rewards for goal-reaching reinforcement learning. *Reinforcement Learning Journal*, 4:1841–1854, 2024.

Yufei Wang, Narasimhan Gautham, Xingyu Lin, Brian Okorn, and David Held. Roll: Visual self-supervised reinforcement learning with object reasoning. In *Conference on robot learning*, pp. 1030–1048. PMLR, 2021.

Qiaoyun Wu, Kai Xu, Jun Wang, Mingliang Xu, Xiaoxi Gong, and Dinesh Manocha. Reinforcement learning-based visual navigation with information-theoretic regularization. *IEEE Robotics and Automation Letters*, 6(2):731–738, 2021.

Denis Yarats, Rob Fergus, Alessandro Lazaric, and Lerrel Pinto. Mastering visual continuous control: Improved data-augmented reinforcement learning. In *Deep RL Workshop NeurIPS 2021*, 2021. URL https://openreview.net/forum?id=L5HKN-IsdSE.

Sriram Yenamandra, Arun Ramachandran, Karmesh Yadav, Austin Wang, Mukul Khanna, Theophile Gervet, Tsung-Yen Yang, Vidhi Jain, Alexander William Clegg, John Turner, Zsolt Kira, Manolis Savva, Angel Chang, Devendra Singh Chaplot, Dhruv Batra, Roozbeh Mottaghi, Yonatan Bisk, and Chris Paxton. Homerobot: Open vocabulary mobile manipulation, 2023. URL https://github.com/facebookresearch/home-robot.

Yufeng Yuan and A Rupam Mahmood. Asynchronous reinforcement learning for real-time control of physical robots. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 5546–5552. IEEE, 2022.

Kevin Zakka, Yuval Tassa, and MuJoCo Menagerie Contributors. MuJoCo Menagerie: A collection of high-quality simulation models for MuJoCo, 2022. URL http://github.com/google-deepmind/mujoco_menagerie.

Yunchu Zhang, Liyiming Ke, Abhay Deshpande, Abhishek Gupta, and Siddhartha S Srinivasa. Cherry-picking with reinforcement learning. In *Robotics: Science and Systems*, 2023.

Wenshuai Zhao, Jorge Peña Queralta, and Tomi Westerlund. Sim-to-real transfer in deep reinforcement learning for robotics: a survey. In *2020 IEEE symposium series on computational intelligence (SSCI)*, pp. 737–744. IEEE, 2020.

Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European conference on computer vision*, pp. 350–368. Springer, 2022.