On-Robot Reinforcement Learning with Goal-Contrastive Rewards

Ondrej Biza[1,2], Thomas Weng[1,*], Lingfeng Sun[1,*], Karl Schmeckpeper[1,*], Tarik Kelestemur[1,*], Yecheng Jason Ma[3,†], Robert Platt[1,2,†], Jan-Willem van de Meent[4,†], Lawson L. S. Wong[2,†]

***Abstract*— Reinforcement Learning (RL) has the potential to enable robots to learn from their own actions in the real world. Unfortunately, RL can be prohibitively expensive, in terms of on-robot runtime, due to inefficient exploration when learning from a sparse reward signal. Designing dense reward functions is labour-intensive and requires domain expertise. In our work, we propose Goal-Contrastive Rewards (GCR), a dense reward function learning method that can be trained on passive video demonstrations. By using videos without actions, our method is easier to scale, as we can use arbitrary videos. GCR combines two loss functions, an implicit value loss function that models how the reward increases when traversing a successful trajectory, and a goal-contrastive loss that discriminates between successful and failed trajectories. We perform experiments in simulated manipulation environments across RoboMimic and MimicGen tasks, as well as in the real world using a Franka arm and a Spot quadruped. We find that GCR leads to a more-sample efficient RL, enabling model-free RL to solve about twice as many tasks as our baseline reward learning methods. We also demonstrate positive cross-embodiment transfer from videos of people and of other robots performing a task. Website: `https://gcr-robot.github.io/`.**

I. INTRODUCTION

On-robot Reinforcement Learning (RL) of manipulation policies is highly challenging because of the exploration problem: a robot can spend hours randomly exploring its environment before chancing upon a goal state. Prior approaches to improve sample-efficiency for on-robot RL relied on hand-designing behavior priors [1], primitives [2] and equivariant policies [3], or on bootstrapping RL with teleoperated demonstrations [4], [5]. Foundation models are effective at detecting goal completion [6], [7], [8], [9], but their application to fine-grained reward prediction is more challenging [10], [11]. Can we design a scalable way to guide RL exploration across many manipulation tasks?

The robotics community has had a growing interest in learning state similarities from passive videos (demonstrations without actions) [12], [13], [14], [15], [16], [17]. These works learn an **implicit state value function** that takes a query image from a video and a goal image and estimates the similarity of their underlying states. Implicit state value functions can be trained using diverse and relatively abundant video data, including videos of a robot or human performing desired "in-domain" tasks, as well as videos of "out-of-domain" tasks performed by humans [18], [19] and other robot morphologies [20], [21]. Therefore, implicit state value functions can scale much better than approaches requiring robot action data. While implicit value pre-training has been primarily used as a representation learning method [13], [14], [15], [16], [17], its application to online RL is less explored [22], [12]. In the context of RL, implicit state value functions could be used to scale reward learning by serving as an intrinsic reward for an RL agent; the agent is rewarded as its observed state progresses towards the desired goal state. Can implicit value functions effectively guide online RL?

In our work, we explore this question through the application of Value-Implicit Pre-training (VIP) [14] to reward shaping for on-robot RL. The primary challenge we face is that throughout training the RL agent observes new states that might be out of distribution for the implicit value function. Consequently, it is possible (and often easy) for

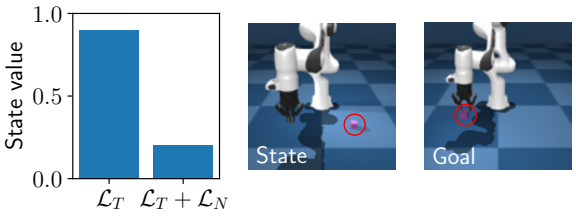

Fig. 1: A model trained with only a temporal loss $\mathcal{L}_T$ assigns a high state value (left) to the pair of state-goal images (right) due to the arm being in the same pose, whereas a model trained with a combination of $\mathcal{L}_T$ and a contrastive loss $\mathcal{L}_N$ (ours) learns to distinguish the position of the block, assigning a low state value.

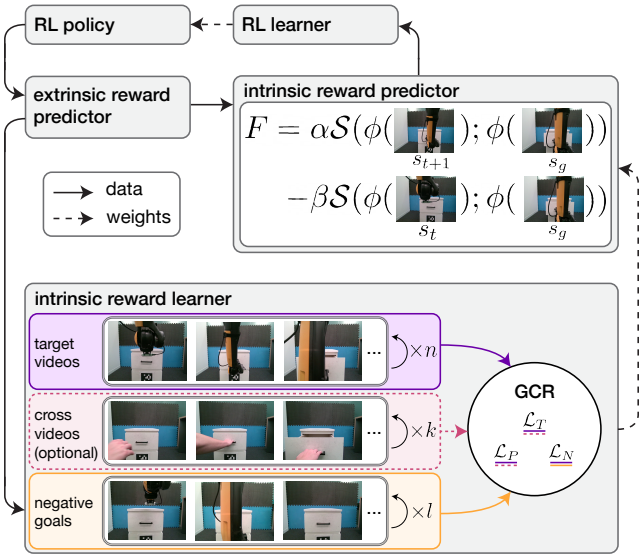

Fig. 2: On-robot reinforcement learning system overview.

*Equal contribution: implementation and experiments. †Equal contribution: technical advising. [1] Robotics and AI Institute, [2] Northeastern University, Khoury College of Computer Sciences, [3] University of Pennsylvania, [4] University of Amsterdam. Correspondence to `biza.o@northeastern.edu`.

the agent to "hack" the value function by reaching high-value states that do not satisfy the goal condition. Figure 1 shows one such example. We find that simply fine-tuning VIP on newly collected data is insufficient to learning an accurate value function.

To solve this problem, we propose a framework, Goal-Contrastive Rewards (GCR), that combines implicit value learning with a pair of contrastive losses over demonstrated and experienced goal states. The key intuition is that we should penalize recently visited states that achieve a high similarity to the goal (high state value) but do not satisfy the goal condition. Unlike [23], [24], which specifically focus on adversarial learning of goals, we leverage both our contrastive goal learning objective as well as the temporal-difference-like objective of VIP to propagate corrected state value estimates across time steps.

We perform experiments in simulated tasks from SERL [25], RoboMimic [26] and MimicGen [27], a drawer opening with a real Boston Dynamics Spot, and picking and button pressing with a real tabletop Franka arm. Our comparison shows that GCR enables efficient RL compared to learning with only sparse rewards and we further show a favorable comparison of GCR to prior Inverse RL methods. GCR enables real-world RL in around 1 to 2 hours, whereas learning with sparse rewards does not succeed. We also demonstrate that GCR can be combined with other RL bootstrapping approaches and that we can benefit from cross-embodiment learning. Specifically, we show positive transfer between two different robot embodiments in simulation and between human and robot videos in the real world. On the systems side, our approach demonstrates an effective combination of intrinsic reward prediction and extrinsic reward prediction with foundation models as goal classifiers. We extend the asynchronous learning framework of [25] to include implicit reward learning and intrinsic + extrinsic reward prediction that runs in parallel to RL control and learning.

In summary, our contributions are:

1) We propose a reward learning framework, Goal-Contrastive Rewards, that learns a shaped reward function from passive videos and adapts to the recent behavior of an RL agent. GCR significantly improves the sample-efficiency of on-robot RL both in sim and real experiments across two robots.
2) We show that GCR benefits from demonstrations of the tasks with different embodiments (robot and human).
3) We demonstrate an asynchronous, on-robot RL system that combines intrinsic reward learning (GCR) with extrinsic reward prediction (foundation models).

II. RELATED WORK

A. Inverse Reinforcement Learning

Inverse Reinforcement Learning (IRL) aims to recover a reward function that explains a set of expert demonstrations. It primarily deals with the ambiguity of learning the right reward function given a limited coverage of the state space. Early IRL works represent rewards as a linear combination of state features, $r(s;\theta) = \theta^T\phi(s)$, [28], [29], [30], [31]. Later work has employed Gaussian Processes [32] and neural networks [33] to approximate the reward function. These approaches rely on an outer loop that learns $r(s;\theta)$ and an inner RL loop that learns a policy given the latest θ, which might be too expensive for on-robot RL.

A second line of work frames IRL as an adversarial learning problem of distinguishing between state-action pairs (s,a) or trajectories $(s_t,a_t)_{t=0}^T$ that either come from demonstrations or from the current policy [34], [35], [36], [23], [24], [37]. These works combine policy learning and reward function (or discriminator) learning in the same loop. Their general focus is on learning a reward function that matches the expert behavior in an unbiased way [36]; conversely, our focus is on learning a shaped reward that guides the agent towards goal states classified by an extrinsic reward function. We pursue learning from passive videos without actions due to their wide availability. Rank2Reward [38] explores this setting by adapting prior Inverse RL methods to state-only trajectories. We use them as baselines.

B. Learning from human videos

A number of approaches attempt to translate videos of people into manipulation policies either directly or indirectly. Some methods detect the pose of the human hand and model the distribution of affordances and motion trajectories as a prior for robot policies [39], [40], [41], [42], [43], [44], [45], [46], [47]. MimicPlay [48] learns "latent plans" based on human hand trajectories as a conditioning for robot policies. Other approaches use in-painting to remove the human hand from videos [49], transform human videos into robot videos [50], or align the latent representations of the videos [51]. [52], [53] trained a world model over both human and robot videos and [54] reconstructed 3D models of objects found in people's hands in internet videos. We instead focus on using human videos to improve reward function learning for RL, making fewer assumptions about detecting and translating hand trajectories and detecting and understanding object interactions.

III. PROBLEM FORMULATION

We model an environment as a Markov Decision Process $M = \langle S, A, P, R, \gamma\rangle$ [55], where the reward function $R : S\times A\times S \to \mathbb{R}$ is sparse (zero for most states). We assume access to a dataset of n passive (action-less) videos $P = \{(s_1^{(i)}, s_2^{(i)}, ...)\}_{i=1}^n$ and optionally m demonstrations $D = \{(s_1^{(i)}, a_1^{(i)}, r_1^{(i)}, s_2^{(i)}, a_2^{(i)}, r_2^{(i)}, ...)\}_{i=1}^m$ with actions and sparse rewards. We do not assume these datasets contain optimal or even successful behaviors, but we filter the data to only contain successful demonstrations in our experiments.

Our objective is to learn a policy π_θ that maximizes the expected return $\mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^t r_t\right]$ while learning from a limited number of interactions with the environment. We use dataset P to train a reward shaping network $F_\phi : S\times S \to \mathbb{R}$ that induces an MDP with a reward function $R'(s,a,s') = R(s,a,s') + F_\phi(s,s')$, where policy learning is faster.

IV. STATE SIMILARITY LEARNING

State similarity (or implicit state value) learning methods use a dataset of videos $P = \{(s_1^{(i)}, s_2^{(i)}, ...)\}_{i=1}^n$ to learn a state similarity function $\mathcal{S}_\phi(s_1; s_2)$. The function can be implemented as a scaled dot product of encoded states, $\phi(s_1)^T\phi(s_2)$, [15], an L_2 distance between the encodings, $||\phi(s_1) - \phi(s_2)||_2$, [14], [12] or as an additional neural prediction head [38]. Common loss functions for learning the state similarity function include time-contrastive learning [12], learning to rank video frames [38] and action-less offline Reinforcement Learning [15], [14], [16].

In our work, we use Value-implicit Pre-training (VIP) [14]. VIP achieves action-less state value learning by optimizing a temporal-difference-like objective derived as a dual to a KL-regularized offline RL objective [56]:

$$\begin{aligned}\mathcal{L}_{\text{VIP}} = \mathbb{E}_{p(g)}\left[(1-\gamma)\,\mathbb{E}_{\mu_0(o;g)}[-\mathcal{S}_\phi(o;g)]\right] \\ + \log \mathbb{E}_{(o,o';g)\sim P}\left[\exp\left\{\mathcal{S}_\phi(o;g) - \delta(o) - \gamma\mathcal{S}_\phi(o';g)\right\}\right].\end{aligned} \tag{1}$$

$p(g)$ is a distribution of goal images, μ_0 is a distribution of initial states conditioned on a goal state and $\delta(o)$ is a goal indicator. In the absence of a reward function, goal states are sampled from the last few frames of each video.

V. GOAL-CONTRASTIVE REWARDS

We propose Goal-Contrastive Rewards (GCR), a framework for learning dense reward functions for RL agents based on passive videos. The key contribution of our method is that it leverages both pre-training on arbitrary videos and online adaptation by fine-tuning as the RL agent reaches incorrect goals. First, we propose a combination of three losses that lead to a discriminative value function (Section V-A). Second, we describe the conversion from a state value function to a dense reward function based on the theory of reward shaping (Section V-B). Third, we discuss the specifics of transferring values across embodiments, e.g. between two robots or between human and robot videos (Section V-C). Fourth, we describe our extension of the SERL asynchronous RL library [25] that includes intrinsic and extrinsic reward functions (Section V-D).

A. Goal-Contrastive Rewards (GCR)

We first make the observation that methods which learn to embed images based on their temporal distance in videos (e.g. R3M [13], VIP [14], LIV [15] and ICVF [16]) might not be suitable reward functions in online RL. These methods model the progress of successful behaviors towards the goal, but are not discriminative enough to penalize incorrect states that look similar to the goal state. For example, in Figure 1, we show that a model trained with a temporal loss $\mathcal{L}_T$ (VIP in this case) assigns a high value to a state where the robot matches the goal joint configuration without picking up a cube, thus failing the task. We note that RL agents tend to discover these gaps in the learned reward function, because they are easier to reach than the correct goal, leading to incorrect behaviors.

As a solution, we propose to use a contrastive loss consisting of a positive term $\mathcal{L}_P$ and a negative term $\mathcal{L}_N$ to learn the representation of goal images.

$$\mathcal{L} = \mathcal{L}_T + \mathbb{E}_{(g,g',g_{\text{neg}})}\left[\omega_1\mathcal{L}_P(g,g') + \omega_2\mathcal{L}_N(g,g_{\text{neg}})\right] \tag{2}$$

The positive term $\mathcal{L}_P$ pulls together the representation of similar goals (g, g'), whereas the negative term $\mathcal{L}_N$ decreases the similarity of a positive and negative goal image pair (g, g_{neg}). The effectiveness of this framework is highly dependent on the sampling of g_{neg}. In particular, the learning algorithm benefits from g_{neg} that are close to the goal image in the representation space of F_ϕ but far from the goal in the ground-truth MDP (in other words, states that have a high predicted value but a low ground-truth value).

Crucially, we sample g_{neg} from a small replay buffer that is updated with online data from the RL agent. Specifically, we sample the last β frames in each failed episode (zero return) expressed as the fraction of the length of the episode. Our learned reward function always adapts to the recent, incorrect, behaviors of the RL agent. This leads to a positive feedback loop, where the RL agent discovers gaps in the learned reward function and the collected data are immediately used to improve the reward model.

In our experiments, we instantiate Equation 2 with the VIP loss as L_T and with a simple contrastive (SC) loss as L_P and L_N. In the following equation, S_ϕ is the cosine similarity of images encoded by ϕ.

$$\begin{aligned}&\mathcal{L}_{\text{GCR(SC)}} = \\ &\quad\mathcal{L}_{\text{VIP}} + \mathbb{E}_{(g,g',g_{\text{neg}})}\left[-\omega_1\mathcal{S}_\phi(g;g') + \omega_2\mathcal{S}_\phi(g;g_{\text{neg}})\right].\end{aligned} \tag{3}$$

We also experiment with an InfoNCE-like contrastive (IC) loss [57]. $\mathcal{L}_{\text{GCR(SC)}}$ is an upper bound on $\mathcal{L}_{\text{GCR(IC)}}$.

$$\begin{aligned}&\mathcal{L}_{\text{GCR(IC)}} = \\ &\quad\mathcal{L}_{\text{VIP}} + \mathbb{E}_{(g,g')}\left[-\log\frac{\exp\{\omega_1\mathcal{S}_\phi(g;g')\}}{\mathbb{E}_{g_{\text{neg}}}\left[\exp\{\omega_2\mathcal{S}_\phi(g;g_{\text{neg}})\}\right]}\right].\end{aligned} \tag{4}$$

In summary, the proposed loss function $\mathcal{L}$ trains a state similarity function $\mathcal{S}_\phi(s_1, s_2)$ based on passive videos P as well as replay buffer data from online RL. Next, we describe turning $\mathcal{S}_\phi(s_1, s_2)$ into a reward function.

B. Dense reward functions

We formulate our reward function as the sum of the sparse reward R and a reward shaping term F. Based on the theoretical analysis of Ng, Harada and Russel [58], F is formulated as the difference of the potential Φ of the next state s' and the current state s.

$$\begin{aligned}R'(s,a,s') &= R(s,a,s') + F(s,s') \\ &= R(s,a,s') + \alpha\Phi(s';g) - \beta\Phi(s;g)\end{aligned} \tag{5}$$

Depending on the setting of α and β, we find various reward shaping terms from prior works:

1) $\alpha = 1, \beta = 1$: a simple difference between the potential of the current and next state [14], [59], [60]. It is a version of the potential-based reward function from [58] with a slight bias of $(1-\gamma)\Phi(s';g)$.

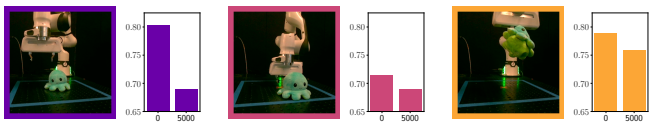

Fig. 3: We show three example states and their associated GCR state values (scaled zero to one) at 0 and 5000 online GCR training steps. The 0 step version of GCR is fine-tuned on demonstrations, but not on any online data. 5000 steps correspond to approximately one hour of on-robot training.

Fig. 4: Four different drawer opening policies learned by GCR. First: handle grasp, second: top grasp, third: top finger through handle, fourth: bottom grasp. We find that GCR improves exploration but does not prescribe a specific way of opening the drawer, whereas RLPD always converges to the same policy (handle grasp).

2) $\alpha = 1, \beta = 0$: prior inverse RL works commonly use only the potential of the next state as a reward function [38], [23], [35]; this reward might be susceptible to the agent getting stuck in endless loops [58], [61].

We use the first option in our work. We implement the potential Φ as a re-scaled cosine similarity $\mathcal{S}_\phi$ of images encoded with ϕ, learned with the loss functions described in Section V-A. This potential can be interpreted as the discount factor γ to the power of the number of steps to go before reaching a goal under some behavioral policy π_b:

$$\Phi(s; g) = (\mathcal{S}_\phi(s; g) + 1)/2 \approx \mathbb{E}_{\pi_b}\left[\sum_{t=0}^{\infty} \gamma^t \text{is\_goal}(s_t)\right] \quad (6)$$

C. Cross-embodiment learning

We have n passive video demonstrations with the target embodiment and further k in-domain videos with a different embodiment (e.g. different robotic arm or videos of people). Usually, k is many times larger than n (e.g. it is much easier to collect videos of people). First, we oversample the target embodiment videos so that they have the same parity as the other embodiment videos. Second, we find that the positive loss $\mathcal{L}_P$ is sufficient for positive transfer across embodiments, as it pulls together the encoded images of goal states across episodes invariant to the appearance of the manipulator. During inference, we only use goal states from the target embodiment videos. Third, we also benefit from a "negative" dataset sampling step in VIP (Appendix D1 in [14]) that applies TD backups to transitions with goal images randomly sampled from the dataset. This leads to backups that combine images from both embodiments.

D. Parallel reward and RL training

We base our system implementation on the SERL library [25], which implements the RL loop as two parallel processes that perform action execution and RL. ZeroMQ [62] is used for performant messaging between processes. Our system implements five processes that run in parallel: RL actor, RL learner, intrinsic reward learner, intrinsic reward predictor and extrinsic reward predictor (Figure 2). The RL actor and learner come from the original SERL implementation. The intrinsic reward learner runs GCR learning and the intrinsic reward predictor evaluates GCR rewards using the latest checkpoint from the learner. The extrinsic reward predictor runs various foundation models to evaluate if a goal condition was reached; it is only used in real-world experiments. This distributed system provides a significant benefit in terms of real-world experiment runtime, as we can commit extra computation to reward learning and prediction without limiting the speed of policy execution.

VI. EXPERIMENTS

We perform online Reinforcement Learning (RL) experiments with SERL, RoboMimic and MimicGen simulated tasks and two real-world robot platforms to answer the following questions:

1) Does GCR lead to efficient online RL in domains with sparse rewards (Section VI-A)?
2) Can we combine GCR with reinforcement learning from demonstrations (Section VI-B)?
3) Does GCR benefit from videos of other embodiments (Section VI-C)?

Simulation: We adapt SERL Pick cube [25], RoboMimic Lift, Can and Square [26] and MimicGen Stack D0 and Coffee D0 [27] to a reinforcement learning setting with sparse rewards. The observation space consists of two RGB images (shoulder and wrist camera) and robot proprioception. The action space represents the delta movements of the end-effector and a gripper open/close command. All environments use a Franka arm; we also collect cross-embodiment data (Section V-C) with a Kuka arm.

Physical robots: We perform real-world experiments with a table-top Franka and a Boston Dynamics Spot with a back-mounted arm and a Fin Ray style gripper. The Franka tasks are picking up a plushie and opening a kettle by pushing a button. The Spot task is to open a drawer. We use a pair of front and wrist cameras in all experiments. We use impedance and admittance control on the Franka and the Spot respectively to limit the forces on the end-effector. The extrinsic rewards (which predict if we have reached a goal state) are implemented with foundation models: we use Grounding DINO [63] with Segment Anything [64] to detect and mask the plushie and the drawer. We predict success heuristically based on the mask position and size. For kettle opening, we query GPT-4 vision [65] with a yes/no question.

Baselines and ablations: We compare our method to previous Inverse Reinforcement Learning methods. Specifically, Rank2Reward (R2R) [38] and a version of Generative Adversarial Imitation Learning (GAIL) [35], Variational Inverse Control with Events (VICE) [36] and Adversarial

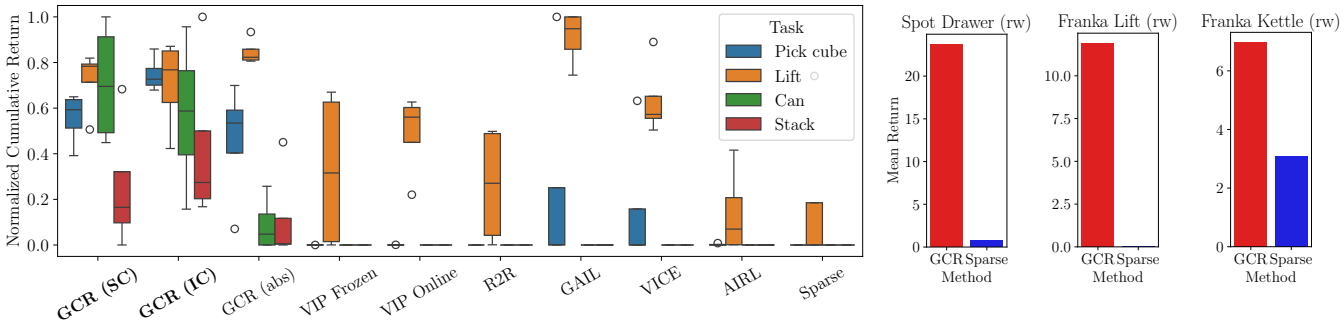

Fig. 5: (Left) Cumulative returns of DrQ trained with different reward functions (x axis). We normalize the cumulative returns by the highest achieved value across all methods and runs. We report four random seeds across three simulated tasks, SERL Pick cube (20 *passive* demos), RoboMimic Lift (20) and Can (20), and MimicGen Stack D0 (100). (Right) We also add results for real-world Franka Lift, Kettle and Spot drawer opening tasks (Figure 6). We run these tasks with only GCR and Sparse rewards.

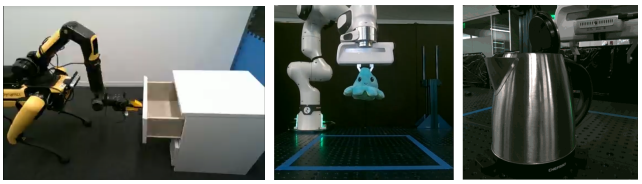

Fig. 6: Real-world reinforcement learning task: Spot drawer opening, Franka plushie picking and Franka kettle opening.

Inverse Reinforcement Learning (AIRL) [36] adapted to passive videos. We further compare GCR to only using sparse rewards (Sparse) and to using sparse rewards with RLPD [4] (Sparse + RLPD). RLPD requires demonstrations with actions. We consider five ablations of our method. VIP Frozen: a reward function derived from VIP fine-tuned on passive demonstrations and frozen during RL training; VIP Online: VIP fine-tuned on both passive demonstrations and online data from the RL agent; GCR (abs) uses the absolute reward (Section V-B, $\alpha = 0, \beta = 1$) instead of a delta reward; GCR (IC) uses an InfoNCE contrastive loss instead of the Simple Contrastive (GCR (SC)) loss. We use an unfrozen ResNet-50 LIV backbone [15] pre-trained on EPIC-KITCHENS [19] in all our experiments unless stated otherwise. Further details are in our appendix[1].

A. Model-free RL with learned reward shaping

We first measure the impact of reward shaping using GCR, our baselines, and ablations on model-free reinforcement learning. Figure 5 shows that GCR learns all four tasks (SERL Pick cube, RoboMimic Lift and Can, MimicGen Stack D0) with 4/4 seeds eventually converging in Pick cube, Lift and Can, and 1/4 seeds converging in Stack. In contrast, R2R, GAIL, VICE, AIRL and Sparse rewards generally only lead to success in Lift. While Rank2Reward (R2R) performs well on MetaWorld [66] tasks in [38], we did not see the same benefit in RoboMimic and MimicGen tasks. To investigate this further, we ran R2R with four backbones – R3M [13] frozen/unfrozen and LIV [15] frozen/unfrozen – and we report the best result. VIP Frozen and VIP Online, which use only the $\mathcal{L}_T$ component of our loss function, only learn Lift, underscoring the importance of goal-contrastive learning. Pick cube and Lift both involve picking up a cube, with a different camera angle and environment appearance. Since the cube looks smaller in Pick cube, we hypothesize that the baselines tend to ignore the cube in favour of overfitting on the robot movements.

The InfoNCE variant of our method (GCR (IC)) performs similarly to the Simple Contrastive variant (GCR (SC)). Using absolute rewards (GCR (abs)) leads to comparable performance on the simple tasks (Pick cube and Lift), but delta rewards (used in GCR (SC) and (IC)) significantly outperform on the difficult tasks (Can and Stack). Finally, we show a comparison between Sparse and GCR rewards in three real-world experiments in Figure 5, right. We find that GCR successfully converges in one to two hours in all three tasks: drawer opening, lifting a plushie, and opening a kettle.

B. Bootstrapped RL

In the previous section, we focused on using action-less videos to learn a shaped reward function. But, there are other approaches to bootstrapping RL exploration. We chose RLPD, a simple method that combines model-free RL and demonstrations with actions. Here, we use the same videos with actions for RLPD and without actions for GCR ($n = m$). Future work could explore a setting where we have passive videos with a few videos with actions ($n > m$). We examine the following question: can GCR reward function learning synergize with RLPD replay buffer bootstrapping?

In Table I, we list the mean returns of Sparse rewards + RLPD or GCR + RLPD on the following tasks: SERL Pick cube, RoboMimic Lift, Can and Square and MimicGen Stack D0 and Coffee D0. Note that we added RoboMimic Square and MimicGen Coffee D0, which were too difficult to solve while learning from action-less demonstrations only.

Our results show that GCR greatly increases the sample-efficiency of RLPD, allowing it to learn Pick cube and Lift with a single demonstration (used by both RLPD and GCR). RLPD cannot solve RoboMimic Can, Square, Stack and Coffee with 100+ demonstrations due to the long-horizon nature of the tasks, but GCR enables RLPD to solve Can in 4/4 seeds, Stack in 4/4 and Square in 2/4. GCR occasionally

[1]Appendix: `https://tinyurl.com/gcr-appendix-2`

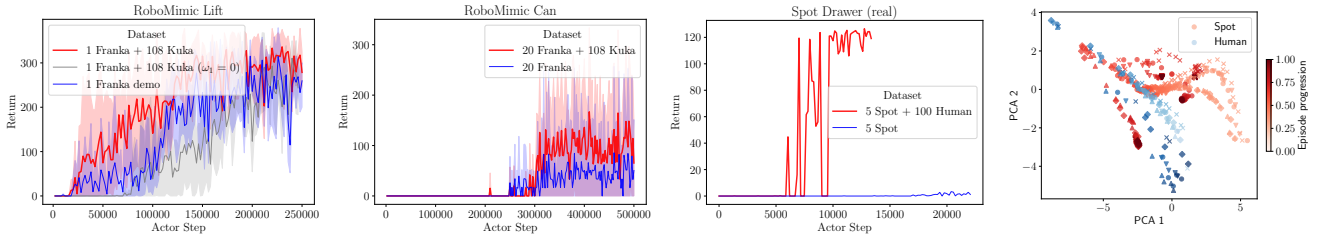

Fig. 7: Episode returns for DrQ trained with GCR rewards. GCR is trained with only target domain demonstrations (blue) or with additional cross-embodiment demonstrations (red). The fourth panel show a latent space of a GCR model with encoded images from human (blue) and robot (red) demonstrations. Each symbol marks a different episode and the brightness of the point represent the episode time step.

Task	Num. Demos	Sparse + RLPD	**GCR + RLPD**
Pick cube	1	0.0 ± 0.0	**28.3 ± 7.3**
Pick cube	5	**54.6 ± 6.2**	**53.4 ± 2.1**
Lift	1	28.2 ± 42.0	**306.8 ± 11.1**
Lift	5	**271.7 ± 16.1**	**287.1 ± 21.7**
Can	20	0.0 ± 0.0	**89.3 ± 28.9**
Can	100	0.0 ± 0.0	**91.2 ± 22.9**
Square	110	0.0 ± 0.0	**22.6 ± 27.0**
Stack D0	20	0.0 ± 0.0	**107.6 ± 21.7**
Stack D0	100	0.0 ± 0.0	**100.1 ± 28.9**
Coffee D0	100	**0.6 ± 0.8**	**7.7 ± 8.4**

TABLE I: Mean returns of GCR (ours) + RLPD compared to Sparse rewards + RLPD. Mean and standard deviations over four random seeds are reported. Bold numbers are within one standard deviation of the best result. Figure 8 shows an example learning curve.

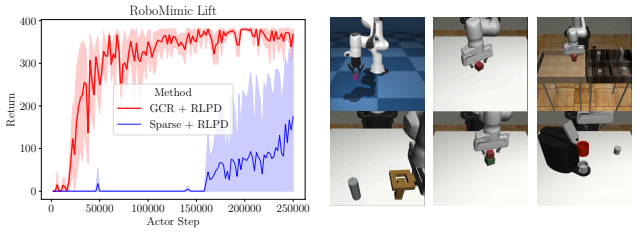

Fig. 8: (Left) Example learning curve of RLPD with learned rewards (GCP) and sparse rewards (Sparse). RoboMimic Lift with one demonstration. Table I lists quantitative results for RLPD. (Right) Simulated tasks: SERL Pick cube, RoboMimic Lift, Can and Square, MimicGen Stack D0 and Coffee D0.

reaches the goal of Coffee D0 but does not converge in 500k environment steps.

We further visualize an interesting qualitative difference between GCR and RLPD policies in Figure 4. GCR + DrQ (without RLPD) learns four different ways of opening the drawer across separate runs. Conversely, RLPD always converges to the demonstrated policy, which involves opening the drawer by grasping its handle. We believe that reward shaping (with GCR) is less descriptive in terms of what policy should be learned, compared to RLPD, which directly feeds demonstrations with actions into the model-free RL replay buffer. This can be a drawback if we want the RL agent to closely follow the demonstration, but also a benefit if we want to learn a more diverse set of policies or possibly find a better way of performing a task.

C. Cross-embodiment learning

An additional benefit of GCR is the ability to learn from videos from other embodiments. We collect 108 videos of a Kuka robot in RoboMimic Lift and Can, and 100 videos of real-world drawer opening with three human participants. We find that cross-embodiment learning benefits all three tasks (Figure 7). The difference is more pronounced in the drawer opening experiment. We suspect that the real-world data contains more noise due to lighting changes, object movements, camera pose changes, etc., which make it so that GCR greatly benefits from a larger dataset of human drawer opening. We also ablate GCR by setting the positive loss weight to zero ($\omega_1 = 0$), which negates the benefit of cross-embodiment data (Figure 7, first panel). Finally, we visualize the GCR latent space of demonstrations across human and Spot episodes in Figure 7, fourth panel. We find that the embeddings especially overlap in the middle of the episode, when the drawer is being pulled open.

VII. CONCLUSION

Goal-Contrastive Rewards provide an effective way of learning a shaped reward from passive videos. The learned dense reward function can then guide the Reinforcement Learning of visuomotor policies in both simulation and the real world, greatly increasing its sample efficiency.

Limitations and future work: Our experiments are limited to learning a reward function from a single camera that is fixed in all experiments except for the Spot drawer opening. It is crucial to develop approaches that can effectively predict rewards from arbitrary views; we note prior works on view-invariant representation learning: [12], [67], [68]. For our RL agent, we use implementations of DrQ [69] and RLPD [4] from [25] that have small convolutional networks as backbones, which might struggle to generalize across objects and environments. Future works could leverage improvements in RL architectures and hyper-parameter tuning [5] and in the pre-training of visual backbones [70], [71], [72].

ACKNOWLEDGMENT

This work is supported in part by NSF 1750649, NSF 2107256, NSF 2314182, NSF 2409351 and NASA 80NSSC19K1474. Yecheng Jason Ma is supported by the Apple Scholars in AI/ML PhD Fellowship. We would like to thank Osman Dogan Yirmibesoglu for designing and manufacturing a flexible Fin-ray style griper for the Spot that we use in our drawer opening experiment. We would also like to thank Andy Park for creating the teleoperation system that we use for data collection and Kuan Fang for helpful discussion.

REFERENCES

[1] Russell Mendonca, Bernadette Bucher, Jiuguang Wang, Deepak Pathak. Continuously Improving Mobile Manipulation with Autonomous Real-World RL. 2024. https://spot-rl-manip.github.io.

[2] S. Nasiriany, H. Liu, and Y. Zhu, "Augmenting reinforcement learning with behavior primitives for diverse manipulation tasks," in *2022 International Conference on Robotics and Automation, ICRA 2022, Philadelphia, PA, USA, May 23-27, 2022*. IEEE, 2022, pp. 7477–7484.

[3] X. Zhu, D. Wang, O. Biza, G. Su, R. Walters, and R. Platt, "Sample efficient grasp learning using equivariant models," in *Robotics: Science and Systems XVIII, New York City, NY, USA, June 27 - July 1, 2022*, K. Hauser, D. A. Shell, and S. Huang, Eds., 2022.

[4] P. J. Ball, L. M. Smith, I. Kostrikov, and S. Levine, "Efficient online reinforcement learning with offline data," in *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 2023, pp. 1577–1594.

[5] H. Hu, S. Mirchandani, and D. Sadigh, "Imitation bootstrapped reinforcement learning," *CoRR*, vol. abs/2311.02198, 2023.

[6] J. Yang, M. S. Mark, B. Vu, A. Sharma, J. Bohg, and C. Finn, "Robot fine-tuning made easy: Pre-training rewards and policies for autonomous real-world reinforcement learning," in *IEEE International Conference on Robotics and Automation, ICRA 2024, Yokohama, Japan, May 13-17, 2024*. IEEE, 2024, pp. 4804–4811.

[7] N. Kumar, T. Silver, W. McClinton, L. Zhao, S. Proulx, T. Lozano-Pérez, L. P. Kaelbling, and J. L. Barry, "Practice makes perfect: Planning to learn skill parameter policies," *CoRR*, vol. abs/2402.15025, 2024.

[8] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, P. Florence, C. Fu, M. G. Arenas, K. Gopalakrishnan, K. Han, K. Hausman, A. Herzog, J. Hsu, B. Ichter, A. Irpan, N. J. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, L. Lee, T. E. Lee, S. Levine, Y. Lu, H. Michalewski, I. Mordatch, K. Pertsch, K. Rao, K. Reymann, M. S. Ryoo, G. Salazar, P. Sanketi, P. Sermanet, J. Singh, A. Singh, R. Soricut, H. T. Tran, V. Vanhoucke, Q. Vuong, A. Wahid, S. Welker, P. Wohlhart, J. Wu, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and B. Zitkovich, "RT-2: vision-language-action models transfer web knowledge to robotic control," *CoRR*, vol. abs/2307.15818, 2023.

[9] Y. Du, K. Konyushkova, M. Denil, A. Raju, J. Landon, F. Hill, N. de Freitas, and S. Cabi, "Vision-language models as success detectors," in *Conference on Lifelong Learning Agents, 22-25 August 2023, McGill University, Montréal, Québec, Canada*, ser. Proceedings of Machine Learning Research, S. Chandar, R. Pascanu, H. Sedghi, and D. Precup, Eds., vol. 232. PMLR, 2023, pp. 120–136.

[10] W. Huang, C. Wang, R. Zhang, Y. Li, J. Wu, and L. Fei-Fei, "Voxposer: Composable 3d value maps for robotic manipulation with language models," in *Conference on Robot Learning, CoRL 2023, 6-9 November 2023, Atlanta, GA, USA*, ser. Proceedings of Machine Learning Research, J. Tan, M. Toussaint, and K. Darvish, Eds., vol. 229. PMLR, 2023, pp. 540–562.

[11] W. Yu, N. Gileadi, C. Fu, S. Kirmani, K. Lee, M. G. Arenas, H. L. Chiang, T. Erez, L. Hasenclever, J. Humplik, B. Ichter, T. Xiao, P. Xu, A. Zeng, T. Zhang, N. Heess, D. Sadigh, J. Tan, Y. Tassa, and F. Xia, "Language to rewards for robotic skill synthesis," in *Conference on Robot Learning, CoRL 2023, 6-9 November 2023, Atlanta, GA, USA*, ser. Proceedings of Machine Learning Research, J. Tan, M. Toussaint, and K. Darvish, Eds., vol. 229. PMLR, 2023, pp. 374–404.

[12] P. Sermanet, C. Lynch, Y. Chebotar, J. Hsu, E. Jang, S. Schaal, and S. Levine, "Time-contrastive networks: Self-supervised learning from video," in *2018 IEEE International Conference on Robotics and Automation, ICRA 2018, Brisbane, Australia, May 21-25, 2018*. IEEE, 2018, pp. 1134–1141.

[13] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta, "R3M: A universal visual representation for robot manipulation," in *Conference on Robot Learning, CoRL 2022, 14-18 December 2022, Auckland, New Zealand*, ser. Proceedings of Machine Learning Research, K. Liu, D. Kulic, and J. Ichnowski, Eds., vol. 205. PMLR, 2022, pp. 892–909.

[14] Y. J. Ma, S. Sodhani, D. Jayaraman, O. Bastani, V. Kumar, and A. Zhang, "VIP: towards universal visual reward and representation via value-implicit pre-training," in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[15] Y. J. Ma, V. Kumar, A. Zhang, O. Bastani, and D. Jayaraman, "LIV: language-image representations and rewards for robotic control," in *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 2023, pp. 23 301–23 320.

[16] D. Ghosh, C. A. Bhateja, and S. Levine, "Reinforcement learning from passive data via latent intentions," in *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 2023, pp. 11 321–11 339.

[17] C. Bhateja, D. Guo, D. Ghosh, A. Singh, M. Tomar, Q. Vuong, Y. Chebotar, S. Levine, and A. Kumar, "Robotic offline RL from internet videos via value-function pre-training," *CoRR*, vol. abs/2309.13041, 2023.

[18] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, M. Martin, T. Nagarajan, I. Radosavovic, S. K. Ramakrishnan, F. Ryan, J. Sharma, M. Wray, M. Xu, E. Z. Xu, C. Zhao, S. Bansal, D. Batra, V. Cartillier, S. Crane, T. Do, M. Doulaty, A. Erapalli, C. Feichtenhofer, A. Fragomeni, Q. Fu, A. Gebreselasie, C. González, J. Hillis, X. Huang, Y. Huang, W. Jia, W. Khoo, J. Kolár, S. Kottur, A. Kumar, F. Landini, C. Li, Y. Li, Z. Li, K. Mangalam, R. Modhugu, J. Munro, T. Murrell, T. Nishiyasu, W. Price, P. R. Puentes, M. Ramazanova, L. Sari, K. Somasundaram, A. Southerland, Y. Sugano, R. Tao, M. Vo, Y. Wang, X. Wu, T. Yagi, Z. Zhao, Y. Zhu, P. Arbeláez, D. Crandall, D. Damen, G. M. Farinella, C. Fuegen, B. Ghanem, V. K. Ithapu, C. V. Jawahar, H. Joo, K. Kitani, H. Li, R. A. Newcombe, A. Oliva, H. S. Park, J. M. Rehg, Y. Sato, J. Shi, M. Z. Shou, A. Torralba, L. Torresani, M. Yan, and J. Malik, "Ego4d: Around the world in 3, 000 hours of egocentric video," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 2022, pp. 18 973–18 990.

[19] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray, "Scaling egocentric vision: The epic-kitchens dataset," in *European Conference on Computer Vision (ECCV)*, 2018.

[20] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis, P. D. Fagan, J. Hejna, M. Itkina, M. Lepert, Y. J. Ma, P. T. Miller, J. Wu, S. Belkhale, S. Dass, H. Ha, A. Jain, A. Lee, Y. Lee, M. Memmel, S. Park, I. Radosavovic, K. Wang, A. Zhan, K. Black, C. Chi, K. B. Hatch, S. Lin, J. Lu, J. Mercat, A. Rehman, P. R. Sanketi, A. Sharma, C. Simpson, Q. Vuong, H. R. Walke, B. Wulfe, T. Xiao, J. H. Yang, A. Yavary, T. Z. Zhao, C. Agia, R. Baijal, M. G. Castro, D. Chen, Q. Chen, T. Chung, J. Drake, E. P. Foster, and et al., "DROID: A large-scale in-the-wild robot manipulation dataset," *CoRR*, vol. abs/2403.12945, 2024.

[21] O. X. Collaboration, A. Padalkar, A. Pooley, A. Jain, A. Bewley, A. Herzog, A. Irpan, A. Khazatsky, A. Raj, A. Singh, A. Brohan, A. Raffin, A. Wahid, B. Burgess-Limerick, B. Kim, B. Schölkopf, B. Ichter, C. Lu, C. Xu, C. Finn, C. Xu, C. Chi, C. Huang, C. Chan, C. Pan, C. Fu, C. Devin, D. Driess, D. Pathak, D. Shah, D. Büchler, D. Kalashnikov, D. Sadigh, E. Johns, F. Ceola, F. Xia, F. Stulp, G. Zhou, G. S. Sukhatme, G. Salhotra, G. Yan, G. Schiavi, G. Kahn, H. Su, H. Fang, H. Shi, H. B. Amor, H. I. Christensen, H. Furuta, H. Walke, H. Fang, I. Mordatch, I. Radosavovic, and et al., "Open x-embodiment: Robotic learning datasets and RT-X models," *CoRR*, vol. abs/2310.08864, 2023.

[22] P. Sermanet, K. Xu, and S. Levine, "Unsupervised perceptual rewards for imitation learning," in *Robotics: Science and Systems XIII, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA, July 12-16, 2017*, N. M. Amato, S. S. Srinivasa, N. Ayanian, and S. Kuindersma, Eds., 2017.

[23] J. Fu, A. Singh, D. Ghosh, L. Yang, and S. Levine, "Variational inverse control with events: A general framework for data-driven reward definition," in *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., 2018, pp. 8547–8556.

[24] A. Singh, L. Yang, C. Finn, and S. Levine, "End-to-end robotic reinforcement learning without reward engineering," in *Robotics: Science*

and Systems XV, University of Freiburg, Freiburg im Breisgau, Germany, June 22-26, 2019, A. Bicchi, H. Kress-Gazit, and S. Hutchinson, Eds., 2019.

[25] J. Luo, Z. Hu, C. Xu, Y. L. Tan, J. Berg, A. Sharma, S. Schaal, C. Finn, A. Gupta, and S. Levine, "SERL: A software suite for sample-efficient robotic reinforcement learning," *CoRR*, vol. abs/2401.16013, 2024.

[26] A. Mandlekar, D. Xu, J. Wong, S. Nasiriany, C. Wang, R. Kulkarni, L. Fei-Fei, S. Savarese, Y. Zhu, and R. Martín-Martín, "What matters in learning from offline human demonstrations for robot manipulation," in *Conference on Robot Learning, 8-11 November 2021, London, UK*, ser. Proceedings of Machine Learning Research, A. Faust, D. Hsu, and G. Neumann, Eds., vol. 164. PMLR, 2021, pp. 1678–1690.

[27] A. Mandlekar, S. Nasiriany, B. Wen, I. Akinola, Y. S. Narang, L. Fan, Y. Zhu, and D. Fox, "Mimicgen: A data generation system for scalable robot learning using human demonstrations," in *Conference on Robot Learning, CoRL 2023, 6-9 November 2023, Atlanta, GA, USA*, ser. Proceedings of Machine Learning Research, J. Tan, M. Toussaint, and K. Darvish, Eds., vol. 229. PMLR, 2023, pp. 1820–1864.

[28] B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey, "Maximum entropy inverse reinforcement learning," in *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, AAAI 2008, Chicago, Illinois, USA, July 13-17, 2008*, D. Fox and C. P. Gomes, Eds. AAAI Press, 2008, pp. 1433–1438.

[29] P. Abbeel and A. Y. Ng, "Apprenticeship learning via inverse reinforcement learning," in *Machine Learning, Proceedings of the Twenty-first International Conference (ICML 2004), Banff, Alberta, Canada, July 4-8, 2004*, ser. ACM International Conference Proceeding Series, C. E. Brodley, Ed., vol. 69. ACM, 2004.

[30] A. Y. Ng and S. Russell, "Algorithms for inverse reinforcement learning," in *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000), Stanford University, Stanford, CA, USA, June 29 - July 2, 2000*, P. Langley, Ed. Morgan Kaufmann, 2000, pp. 663–670.

[31] S. Russell, "Learning agents for uncertain environments (extended abstract)," in *Proceedings of the Eleventh Annual Conference on Computational Learning Theory, COLT 1998, Madison, Wisconsin, USA, July 24-26, 1998*, P. L. Bartlett and Y. Mansour, Eds. ACM, 1998, pp. 101–103.

[32] S. Levine, Z. Popovic, and V. Koltun, "Nonlinear inverse reinforcement learning with gaussian processes," in *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. C. N. Pereira, and K. Q. Weinberger, Eds., 2011, pp. 19–27.

[33] M. Wulfmeier, P. Ondruska, and I. Posner, "Maximum entropy deep inverse reinforcement learning," *CoRR*, vol. abs/1507.04888, 2015.

[34] C. Finn, S. Levine, and P. Abbeel, "Guided cost learning: Deep inverse optimal control via policy optimization," in *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, ser. JMLR Workshop and Conference Proceedings, M. Balcan and K. Q. Weinberger, Eds., vol. 48. JMLR.org, 2016, pp. 49–58.

[35] J. Ho and S. Ermon, "Generative adversarial imitation learning," in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, Eds., 2016, pp. 4565–4573.

[36] J. Fu, K. Luo, and S. Levine, "Learning robust rewards with adversarial inverse reinforcement learning," *CoRR*, vol. abs/1710.11248, 2017.

[37] S. Reddy, A. D. Dragan, and S. Levine, "SQIL: imitation learning via reinforcement learning with sparse rewards," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[38] D. Yang, D. Tjia, J. Berg, D. Damen, P. Agrawal, and A. Gupta, "Rank2reward: Learning shaped reward functions from passive video," *CoRR*, vol. abs/2404.14735, 2024.

[39] K. Fang, T. Wu, D. Yang, S. Savarese, and J. J. Lim, "Demo2vec: Reasoning object affordances from online videos," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 2139–2147.

[40] H. Xiong, Q. Li, Y. Chen, H. Bharadhwaj, S. Sinha, and A. Garg, "Learning by watching: Physical imitation of manipulation skills from human videos," in *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2021, Prague, Czech Republic, September 27 - Oct. 1, 2021*. IEEE, 2021, pp. 7827–7834.

[41] Y. Qin, Y. Wu, S. Liu, H. Jiang, R. Yang, Y. Fu, and X. Wang, "Dexmv: Imitation learning for dexterous manipulation from human videos," in *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXIX*, ser. Lecture Notes in Computer Science, S. Avidan, G. J. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., vol. 13699. Springer, 2022, pp. 570–587.

[42] K. Shaw, S. Bahl, and D. Pathak, "Videodex: Learning dexterity from internet videos," in *Conference on Robot Learning, CoRL 2022, 14-18 December 2022, Auckland, New Zealand*, ser. Proceedings of Machine Learning Research, K. Liu, D. Kulic, and J. Ichnowski, Eds., vol. 205. PMLR, 2022, pp. 654–665.

[43] M. Goyal, S. Modi, R. Goyal, and S. Gupta, "Human hands as probes for interactive object understanding," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 2022, pp. 3283–3293.

[44] S. Bahl, A. Gupta, and D. Pathak, "Human-to-robot imitation in the wild," in *Robotics: Science and Systems XVIII, New York City, NY, USA, June 27 - July 1, 2022*, K. Hauser, D. A. Shell, and S. Huang, Eds., 2022.

[45] H. Bharadhwaj, A. Gupta, S. Tulsiani, and V. Kumar, "Zero-shot robot manipulation from passive human videos," *CoRR*, vol. abs/2302.02011, 2023.

[46] S. Bahl, R. Mendonca, L. Chen, U. Jain, and D. Pathak, "Affordances from human videos as a versatile representation for robotics," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 2023, pp. 1–13.

[47] G. Papagiannis, N. D. Palo, P. Vitiello, and E. Johns, "R+X: retrieval and execution from everyday human videos," *CoRR*, vol. abs/2407.12957, 2024.

[48] C. Wang, L. Fan, J. Sun, R. Zhang, L. Fei-Fei, D. Xu, Y. Zhu, and A. Anandkumar, "Mimicplay: Long-horizon imitation learning by watching human play," *CoRR*, vol. abs/2302.12422, 2023.

[49] M. Chang, A. Prakash, and S. Gupta, "Look ma, no hands! agent-environment factorization of egocentric videos," in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023.

[50] L. M. Smith, N. Dhawan, M. Zhang, P. Abbeel, and S. Levine, "AVID: learning multi-stage tasks via pixel-level translation of human videos," in *Robotics: Science and Systems XVI, Virtual Event / Corvalis, Oregon, USA, July 12-16, 2020*, M. Toussaint, A. Bicchi, and T. Hermans, Eds., 2020.

[51] K. Schmeckpeper, O. Rybkin, K. Daniilidis, S. Levine, and C. Finn, "Reinforcement learning with videos: Combining offline observations with interaction," *arXiv preprint arXiv:2011.06507*, 2020.

[52] K. Schmeckpeper, A. Xie, O. Rybkin, S. Tian, K. Daniilidis, S. Levine, and C. Finn, "Learning predictive models from observation and interaction," in *European Conference on Computer Vision*. Springer, 2020, pp. 708–725.

[53] Y. Seo, K. Lee, S. James, and P. Abbeel, "Reinforcement learning with action-free pre-training from videos," in *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 2022, pp. 19 561–19 579.

[54] A. Patel, A. Wang, I. Radosavovic, and J. Malik, "Learning to imitate object interactions from internet videos," *CoRR*, vol. abs/2211.13225, 2022.

[55] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: A Bradford Book, 2018.

[56] O. Nachum, B. Dai, I. Kostrikov, Y. Chow, L. Li, and D. Schuurmans, "Algaedice: Policy gradient from arbitrary experience," *CoRR*, vol. abs/1912.02074, 2019.

[57] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *CoRR*, vol. abs/1807.03748, 2018.

[58] A. Y. Ng, D. Harada, and S. Russell, "Policy invariance under reward transformations: Theory and application to reward shaping," in *Proceedings of the Sixteenth International Conference on Machine Learning (ICML 1999), Bled, Slovenia, June 27 - 30, 1999*, I. Bratko and S. Dzeroski, Eds. Morgan Kaufmann, 1999, pp. 278–287.

[59] Y. Li, T. Gao, J. Yang, H. Xu, and Y. Wu, "Phasic self-imitative reduction for sparse-reward goal-conditioned reinforcement learning," in *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 2022, pp. 12 765–12 781.

[60] Y. Lee, A. Szot, S. Sun, and J. J. Lim, "Generalizable imitation learning from observation via inferring goal proximity," in *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021, pp. 16 118–16 130.

[61] J. Randløv and P. Alstrøm, "Learning to drive a bicycle using reinforcement learning and shaping," in *Proceedings of the Fifteenth International Conference on Machine Learning (ICML 1998), Madison, Wisconsin, USA, July 24-27, 1998*, J. W. Shavlik, Ed. Morgan Kaufmann, 1998, pp. 463–471.

[62] "Zeromq: An open-source universal messaging library," https://zeromq.org/.

[63] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, and L. Zhang, "Grounding DINO: marrying DINO with grounded pre-training for open-set object detection," *CoRR*, vol. abs/2303.05499, 2023.

[64] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W. Lo, P. Dollár, and R. B. Girshick, "Segment anything," in *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*. IEEE, 2023, pp. 3992–4003.

[65] OpenAI, "GPT-4 technical report," *CoRR*, vol. abs/2303.08774, 2023.

[66] T. Yu, D. Quillen, Z. He, R. Julian, K. Hausman, C. Finn, and S. Levine, "Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning," in *3rd Annual Conference on Robot Learning, CoRL 2019, Osaka, Japan, October 30 - November 1, 2019, Proceedings*, ser. Proceedings of Machine Learning Research, L. P. Kaelbling, D. Kragic, and K. Sugiura, Eds., vol. 100. PMLR, 2019, pp. 1094–1100.

[67] Z. Xue and K. Grauman, "Learning fine-grained view-invariant representations from unpaired ego-exo videos via temporal alignment," in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023.

[68] Y. Huang, G. Chen, J. Xu, M. Zhang, L. Yang, B. Pei, H. Zhang, L. Dong, Y. Wang, L. Wang, and Y. Qiao, "Egoexolearn: A dataset for bridging asynchronous ego- and exo-centric view of procedural activities in real world," *CoRR*, vol. abs/2403.16182, 2024.

[69] D. Yarats, I. Kostrikov, and R. Fergus, "Image augmentation is all you need: Regularizing deep reinforcement learning from pixels," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[70] A. Majumdar, K. Yadav, S. Arnaud, Y. J. Ma, C. Chen, S. Silwal, A. Jain, V. Berges, T. Wu, J. Vakil, P. Abbeel, J. Malik, D. Batra, Y. Lin, O. Maksymets, A. Rajeswaran, and F. Meier, "Where are we in the search for an artificial visual cortex for embodied intelligence?" in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023.

[71] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P. Huang, S. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jégou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "Dinov2: Learning robust visual features without supervision," *Trans. Mach. Learn. Res.*, vol. 2024, 2024.

[72] J. Shang, K. Schmeckpeper, B. B. May, M. V. Minniti, T. Kelestemur, D. Watkins, and L. Herlant, "Theia: Distilling diverse vision foundation models for robot learning," *CoRR*, vol. abs/2407.20179, 2024.