Graph Inverse Reinforcement Learning from Diverse Videos

Sateesh Kumar **Jonathan Zamora*** **Nicklas Hansen***
Rishabh Jangir **Xiaolong Wang**
UC San Diego

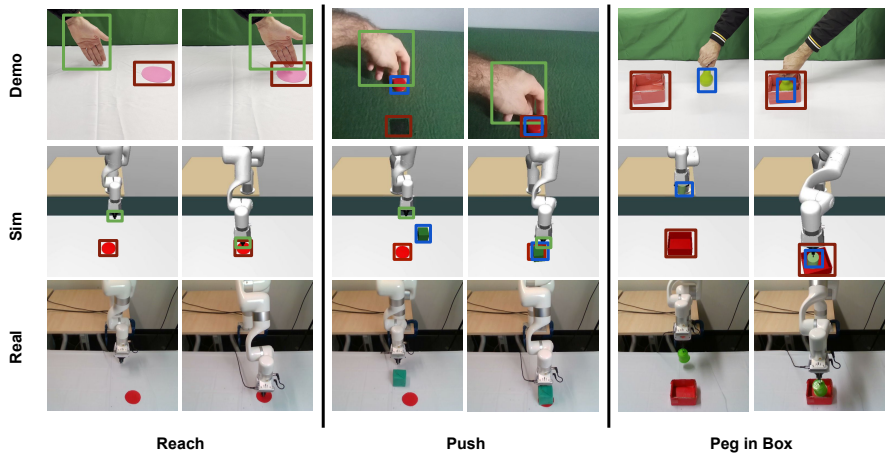

Figure 1: **GraphIRL.** We propose an approach for performing inverse reinforcement learning from diverse third-person videos via graph abstraction. Based on our learned reward functions, we successfully train image-based policies in simulation and deploy them on a real robot.

Abstract: Research on Inverse Reinforcement Learning (IRL) from third-person videos has shown encouraging results on removing the need for manual reward design for robotic tasks. However, most prior works are still limited by training from a relatively restricted domain of videos. In this paper, we argue that the true potential of third-person IRL lies in increasing the diversity of videos for better scaling. To learn a reward function from diverse videos, we propose to perform graph abstraction on the videos followed by temporal matching in the graph space to measure the task progress. Our insight is that a task can be described by entity interactions that form a graph, and this graph abstraction can help remove irrelevant information such as textures, resulting in more robust reward functions. We evaluate our approach, *GraphIRL*, on cross-embodiment learning in X-MAGICAL and learning from human demonstrations for real-robot manipulation. We show significant improvements in robustness to diverse video demonstrations over previous approaches, and even achieve better results than manual reward design on a real robot pushing task. Videos are available at https://sateeshkumar21.github.io/GraphIRL/.

Keywords: Inverse Reinforcement Learning, Third-Person Video, Graph Network

1 Introduction

Deep Reinforcement Learning (RL) is a powerful general-purpose framework for learning behavior policies from high-dimensional interaction data, and has led to a multitude of impressive feats in application areas such as game-playing [1] and robotics [2, 3]. Through interaction with an unknown environment, RL agents iteratively improve their policy by learning to maximize a reward signal, which has the potential to be used in lieu of hand-crafted control policies. However, the performance of policies learned by RL is found to be highly dependent on the careful specification of task-specific

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

reward functions and, as a result, crafting a good reward function may require significant domain knowledge and technical expertise.

As an alternative to manual design of reward functions, *inverse RL* (IRL) has emerged as a promising paradigm for policy learning. By framing the reward specification as a learning problem, operators can specify a reward function based on video examples. While *imitation learning* typically requires demonstrations from a first-person perspective, IRL can in principle learn a reward function, *i.e.*, a measure of task progression, from *any* perspective, including third-person videos of humans performing a task. This has positive implications for data collection, since it is often far easier for humans to capture demonstrations in third-person.

Although IRL from third-person videos is appealing because of its perceived flexibility, learning a good reward function from raw video data comes with a variety of challenges. This is perhaps unsurprising, considering the visual and functional diversity that such data contains. For example, the task of pushing an object across a table may require different motions depending on the embodiment of the agent. A recent method for cross-embodiment IRL, dubbed XIRL [4], learns to capture task progression from videos in a self-supervised manner by enforcing temporal cycle-consistency constraints. While XIRL can in principle consume any video demonstration, we observe that its ability to learn task progression degrades substantially when the visual appearance of the video demonstrations do not match that of the target environment for RL. Therefore, it is natural to ask the question: *can we learn to imitate others from (a limited number of) diverse third-person videos?*

In this work, we demonstrate that it is indeed possible. Our key insight is that, while videos may be of great visual diversity, their underlying scene structure and agent-object interactions can be abstracted via a graph representation. Specifically, instead of directly using images, we extract object bounding boxes from each frame using an off-the-shelf detector, and construct a graph abstraction where each object is represented as a node in the graph. Often – in robotics tasks – the spatial location of an object by itself may not convey the full picture of the task at hand. For instance, to understand a task like *Peg in Box* (shown in Figure 1), we need to also take into account how the agent *interacts* with the object. Therefore, we propose to employ *Interaction Networks* [5] on our graph representation to explicitly model interactions between entities. To train our model, we follow [4, 6] and apply a temporal cycle consistency loss, which (in our framework) yields task-specific yet embodiment- and domain-agnostic feature representations.

We validate our method empirically on a set of simulated cross-domain cross-embodiment tasks from X-MAGICAL [4], as well as three vision-based robotic manipulation tasks. To do so, we collect a diverse set of demonstrations that vary in visual appearance, embodiment, object categories, and scene configuration; X-MAGICAL demonstrations are collected in simulation, whereas our manipulation demonstrations consist of real-world videos of humans performing tasks. We find our method to outperform a set of strong baselines when learning from visually diverse demonstrations, while simultaneously matching their performance in absence of diversity. Further, we demonstrate that vision-based policies trained with our learned reward perform tasks with greater precision than human-designed reward functions, and successfully transfer to a real robot setup with only approximate correspondence to the simulation environment. Thus, our proposed framework completes the cycle of learning rewards from real-world human demonstrations, learning a policy in simulation using learned rewards, and finally deployment of the learned policy on physical hardware.

2 Related Work

Learning from demonstration. Conventional imitation learning methods require access to expert demonstrations comprised of observations and corresponding ground-truth actions for every time step [7, 8, 9, 10], for which kinesthetic teaching or teleoperation are the primary modes of data collection in robotics. To scale up learning, video demonstrations are recorded with human operating the same gripper that the robot used, which also allows direct behaviro cloning [11, 12]. More recently, researchers have developed methods that instead infer actions from data via a learned forward [13] or inverse [14, 15] dynamics model. However, this approach still makes the implicit assumption that imitator and demonstrator share a common observation and action space, and are therefore not directly applicable to the cross-domain cross-embodiment problem setting that we consider.

Inverse RL. To address the aforementioned limitations, inverse RL has been proposed [16, 17, 18, 19, 20, 21] and it has recently emerged as a promising paradigm for cross-embodiment imitation in particular [22, 23, 24, 25, 26, 27, 28, 4, 29]. For example, Schmeckpeper et al. [22] proposes

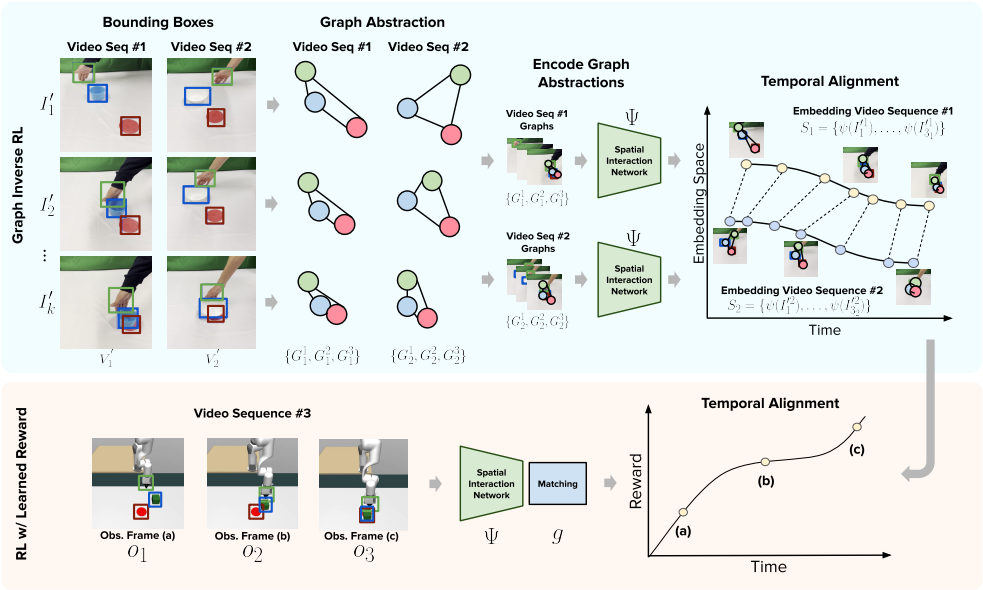

Figure 2: **Overview.** We extract object bounding boxes from video sequences using an off-the-shelf detector, and construct a graph abstraction of the scene. We model graph-abstracted object interactions using *Interaction Networks* [5], and learn a reward function by aligning video embeddings temporally. We then train image-based RL policies using our learned reward function, and deploy on a real robot.

a method for integrating video demonstrations without corresponding actions into off-policy RL algorithms via a latent inverse dynamics model and heuristic reward assignment, and Zakka et al. [4] (XIRL) learns a reward function from video demonstrations using temporal cycle-consistency and trains an RL agent to maximize the learned rewards. In practice, however, inverse RL methods such as XIRL are found to require limited visual diversity in demonstrations. Our work extends XIRL to the setting of diverse videos by introducing a graph abstraction that models agent-object and object-object interactions while still enforcing temporal cycle-consistency.

Object-centric representations. have been proposed in many forms at the intersection of computer vision and robotics. For example, object-centric scene graphs can be constructed for integrated task and motion planning [30, 31, 32], navigation [33, 34], relational inference [35, 36], dynamics modeling [5, 37, 38, 39, 40, 41], model predictive control [42, 43, 44], grasping [45, 46] or visual imitation learning [47, 48, 49]. Similar to our work, Sieb et al. [49] propose to abstract video demonstrations as object-centric graphs for the problem of single-video cross-embodiment imitation, and act by minimizing the difference between the demonstration graph and a graph constructed from observations captured at each step. As such, their method is limited to same-domain visual trajectory following, whereas we learn a general alignment function for cross-domain cross-embodiment imitation and leverage *Interaction Networks* [5] for modeling graph-abstracted spatial interactions rather than relying on heuristics.

3 Our Approach

In this section, we describe our main contribution, which is a self-supervised method for learning a visually invariant reward function directly from a set of diverse third-person video demonstrations via a graph abstraction. Our Graph Inverse Reinforcement Learning (GraphIRL) framework, shown in Figure 2, consists of building an object-centric graph abstraction of the video demonstrations and then learn an embedding space that captures task progression by exploiting the temporal cue in the videos. This embedding space is then used to construct a *domain invariant* and *embodiment invariant* reward function which can be used to train any standard reinforcement learning algorithm.

Problem Formulation. Given a task T, our approach takes a dataset of video demonstrations $D = \{V_1, V_2, \ldots, V_n\}$. Each video consists of image frames $\{I_1^i, I_2^i, \ldots, I_k^i\}$ where i denotes the video frame index and k denotes the total number of frames in V_i. Given D, our goal is to learn a reward function that can be used to solve the task T for any robotic environment. Notably, we do *not*

assume access to any action information of the expert demonstrations, and our approach does *not* require objects or embodiments in the target environment to share appearance with demonstrations.

3.1 Representation Learning

To learn task-specific representations in a self-supervised manner, we take inspiration from Dwibedi et al. [6] and employ a temporal cycle consistency loss. However, instead of directly using images, we propose a novel object-centric graph representation, which allows us to learn an embedding space that not only captures task-specific features, but depends *solely* on the spatial configuration of objects and their interactions. We here detail each component of our approach to representation learning.

Object-Centric Representation. Given video frames $\{I_1^i, I_2^i, \ldots, I_k^i\}$, we first extract object bounding boxes from each frame using an off-the-shelf detector. Given N bounding boxes for an image, we represent each bounding box as a $4+m$ dimensional vector $o_j = \{x_1, y_1, x_2, y_2, d_1, d_2, \ldots, d_m\}$, where the first 4 dimensions represent the leftmost and rightmost corners of the bounding box, and the remaining m dimensions encode distances between the centroids of the objects. For each frame I_j^i we extract an object-centric representation $I_j'^i = \{o_1, o_2, \ldots, o_m\}$ such that we can represent our dataset of demonstrations as $D' = \{V_1', V_2', \ldots, V_n'\}$ where V_i' is the sequence of bounding boxes corresponding to video V_i. Subsequent sections describe how we learn representations given D'.

Spatial Interaction Encoder. Taking inspiration from recent approaches on modeling physical object dynamics [5, 37], we propose a *Spatial Interaction Encoder Network* to explicitly model object-object interactions. Specifically, given a sequence V_i' from D', we model each element I' as a graph, $G = (O, R)$, where O is the set of objects $\{o_1, o_2, \ldots, o_m\}$, m is the total number of objects in I', and R denotes the relationship between objects (*i.e.*, whether two objects interact with each other). For simplicity, all objects are connected with all other objects in the graph such that $R = \{(i,j) \mid i \neq j \wedge i \leq m \wedge j \leq m\}$. We use a fully-connected graph because this makes the least assumption about the problem and task specific object interaction structure could then be learned directly from the data. We compose an object embedding for each of $o_i \in O$ by combining *self* and *interactional* representations as follows:

$$f_o(o_i) = \phi_{\text{agg}}(f_{\text{s}} + f_{\text{in}}) \quad \text{with} \quad f_s(o_i) = \phi_s(o)\,, \qquad f_{\text{in}}(o_i) = \sum_{j=1}^{m} \phi_{\text{in}}(o_i, o_j) \mid (i,j) \in \mathbb{R}\,, \quad (1)$$

where $f_s(o_i)$ represents the *self* or independent representation of an object, f_{in} represents the *interactional* representation, *i.e.*, how it interacts with other objects in the scene, f_o is the final object embedding, and $(,)$ represents concatenation. Here, the encoders ϕ_{s}, ϕ_{in} and ϕ_{agg} denote Multi layer Perceptron (MLP) networks respectively. We emphasize that the expression for $f_{\text{in}}(\cdot)$ implies that the object embedding $f_o(.)$ depends on *all* other objects in the scene; this term allows us to model relationships of an object with the others. The final output from the spatial interaction encoder $\psi(\cdot)$ for object representation I' is the mean of all object encodings:

$$\psi(I') = \frac{1}{m} \sum_{i}^{m} f(o_i)\,. \quad (2)$$

The spatial interaction encoder is then optimized using the temporal alignment loss introduced next.

Temporal Alignment Loss. Taking inspiration from prior works on video representatoin learning [6, 50, 51], we use temporal alignment as a proxy for video representation learning. Given a pair of videos, the task of self-supervised alignment implicitly assumes that there exists true semantic correspondence between the two sequences, *i.e.*, both videos share a common semantic space. These works have shown that optimizing for alignment leads to representations that could be used for tasks that require understanding task progression such as action-classification. This is because in order to solve for alignment, a learning model has to learn features that are (1) common across most videos and (2) exhibit temporal ordering. For a sufficiently large dataset with single task, the most common visual features would be distinct phases of a task that appear in all videos and if the task has small permutations, these distinct features would also exhibit temporal order. In such scenarios, the representations learned by optimizing for alignment are *task-specific* and invariant to changes in viewpoints, motions and embodiments.

In this work, we employ Temporal Cycle Consistency (TCC) [6] loss to learn temporal alignment. TCC optimizes for alignment by learning an embedding space that maximizes one-to-one nearest neighbour mappings between sequences. This is achieved through a loss that maximizes for cycle-consistent nearest neighbours given a pair of video sequences. In our case, the cycle consistency

is applied on the *graph abstraction* instead of image features as done in the aforementioned video alignment methods. Specifically, given D', we sample a pair of bounding box sequences $V'_i = \{I'^i_1, \ldots, I'^i_{m_i}\}$ and $V'_j = \{I'j_1, \ldots, I'^j_{m_j}\}$, we extract embeddings by applying our spatial interaction encoder defined in Equation 2. Thus, we obtain the encoded features $S_i = \{\psi(I'^i_1), \ldots, \psi(I'^i_{m_i})\}$ and $S_j = \{\psi(I'^j_1), \ldots, \psi(I'^j_{m_j})\}$. For the nth element in S_i, we first compute its nearest neighbour, v^n_{ij}, in S_j and then compute the probability that it cycles-back to n, β^k_{ijn} as

$$v^n_{ij} = \sum_k^{m_j} \alpha_k S^k_j \,, \alpha_k = \frac{e^{-||S^n_i - S^k_j||^2}}{\sum_k^{m_j} e^{-||S^n_i - S^k_j||^2}} \,, \beta^k_{ijn} = \frac{e^{-||v^n_{ij} - S^k_i||^2}}{\sum_k^{m_j} e^{-||v^n_{ij} - S^k_i||^2}} \,. \tag{3}$$

The cycle consistency loss for nth element can be computed as $L^{ij}_n = (\mu^n_{ij} - n)^2$, where $\mu^n_{ij} = \sum_k^{mi} \beta^k_{ijn} k$ is the expected value of frame index n as we cycle back. The overall TCC loss is then defined by summing over all pairs of sequence embeddings (S_i, S_j) in the data, *i.e.*, $L^n_{ij} = \sum_{ijn} L^n_{ij}$.

3.2 Reinforcement Learning

We learn a task-specific embedding space by optimizing for temporal alignment. In this section, we define how to go from this embedding space to a reward function that measures task progression. For constructing the reward function, we leverage the insight from Zakka et al. [4] that in a task-specific embedding space, we can use euclidean distance as a notion of task progression, *i.e.*, frames far apart in the embedding space will be far apart in terms of task progression and vice versa. We therefore choose to define our reward function as $r(o) = -\frac{1}{c}||\psi(o) - g||^2$, with $g = \sum_{i=1}^{n} \psi(I'^i_{m_i})$, where o is the current observation, ψ is our Interaction Networks-based encoder from Section 3, g is the representative goal frame, m_i is the length of sequence V'^i and c is a scaling factor. This definition gives us a dense reward because as the observed state gets closer and closer to the goal, the reward starts going down and approaches zero when the goal and current observation are close in embedding space; refer to supplementary material for qualitative analysis of the learned reward function. After constructing the learned reward, we can use it to train any standard RL algorithm. We note that, unlike previous approaches [22, 4], our method does not use *any* environment reward to improve performance, and instead relies *solely* on the learned reward, which our experiments demonstrate is sufficient for solving diverse robotic manipulation tasks.

4 Experiments

In this section, we demonstrate how our approach uses diverse video demonstrations to learn a reward function that generalizes to unseen domains. In particular, we are interested in answering the questions: (1) How do vision-based methods for IRL perform when learning from demonstrations that exhibit *domain shift*? (2) Can our approach learn a stronger reward signal under this challenging setting? To that end, we first conduct experiments on the X-MAGICAL benchmark [4]. We then conduct experiments on multiple robot manipulation tasks using a diverse set of demonstrations.

Implementation Details. All MLPs have 2 hidden layers, and the embedding layer outputs features of size 128 in all experiments. We use Soft Actor-Critic (SAC) [52] as backbone RL algorithm for all methods. For experiments on X-MAGICAL, we follow Zakka et al. [4] and learn a state-based policy. For robotic manipulation experiments, we learn a multi-view image-based SAC policy [53]. For fair comparison, we only change the learned reward function across methods and keep the RL setup identical. Refer to the supplementary material for further implementation details.

Baselines. We compare against multiple vision-based approaches that learn rewards in a self-supervised manner: **(1) XIRL** [4] that learns a reward function by applying the TCC [6] loss on demonstration video sequences, **(2) TCN** [54] which is a self-supervised contrastive method for video representation learning that optimizes for temporally disentangled representations, and **(3) LIFS** [55] that learns an invariant feature space using a dynamic time warping-based contrastive loss. Lastly, we also compare against the manually designed **(4) Environment Reward** from Jangir et al. [53]. The environment reward baseline is an oracle method since it is a dense reward and is carefully designed for the task under consideration. For learning-based baselines, we use a ResNet-18 encoder pretrained on ImageNet [56] classification. See supplementary material for details.

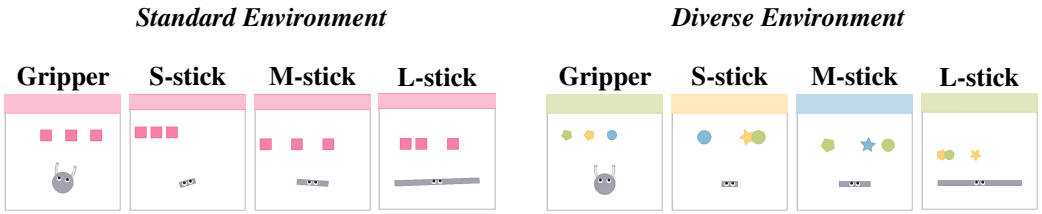

Figure 3: **Overview of X-MAGICAL task variants.** We consider two environment variants and four embodiments for our simulated sweeping task experiments. We evaluate IRL algorithms in both the *Diverse* and *Standard* environments across four embodiments in the *Cross-Embodiment settings*.

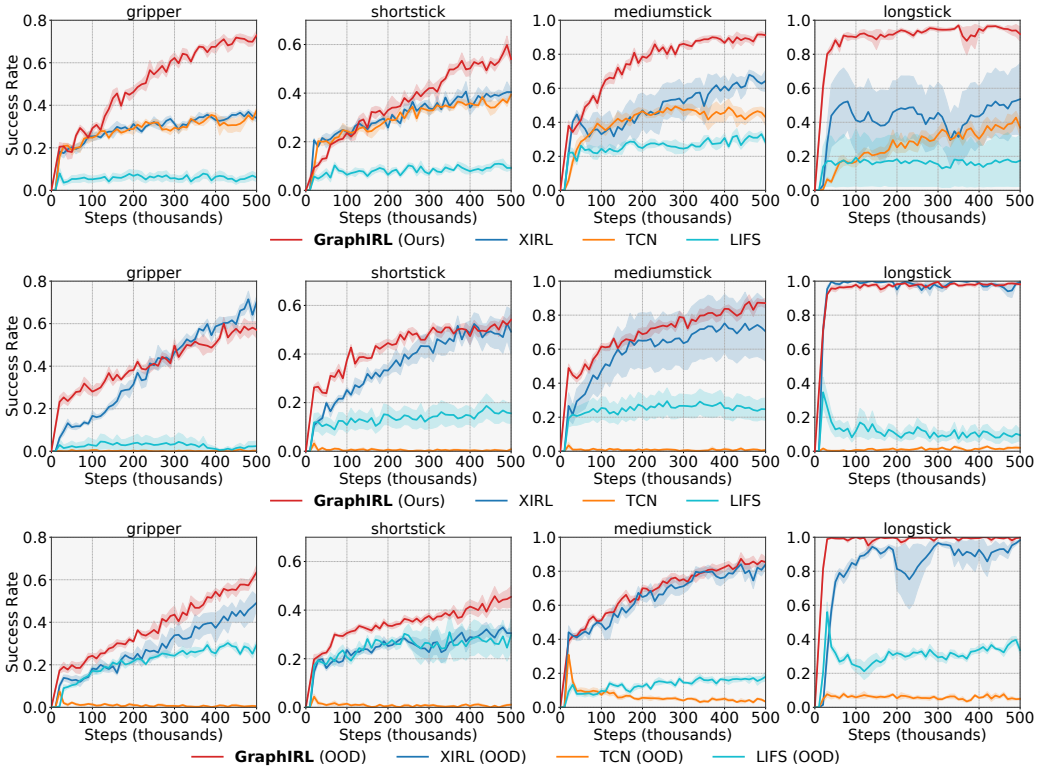

Figure 4: **Cross-Embodiment Cross-Environment.** Success rates of our method *GraphIRL* and baselines on *(top)* Standard Environment Pretraining $\rightarrow$ Diverse Environment RL, *(middle)* Diverse Environment Pretraining $\rightarrow$ Standard Environment RL and *(bottom)* Diverse (OOD) Environment Pretraining $\rightarrow$ Standard Environment RL. All reported numbers are averaged over 5 seeds. Our approach performs favorably when compared to other baselines on all three settings.

4.1 Experimental Setup

We conduct experiments under two settings: the *Sweep-to-Goal* task from X-MAGICAL [4], and robotic manipulation tasks with an xArm robot both in simulation and on a real robot setup. We describe our experimental setup under these two settings in the following.

X-MAGICAL. We choose to extend X-MAGICAL [4], a 2D simulation environment for cross-embodiment imitation learning. On this benchmark, we consider a multi-object sweeping task, where the agent must push three objects towards a static goal region. We utilize two variants of the X-MAGICAL benchmark, which we denote as ***Standard*** (original) and ***Diverse*** (ours) environments, shown in Figure 3. ***Standard*** only randomizes the position of objects, whereas ***Diverse*** also randomizes visual appearance. We consider a set of four unique embodiments {*gripper, short-stick, medium-stick, long-stick*}. In particular, we conduct experiments in the *cross-environment* and *cross-embodiment setting* where we learn a reward function in the ***Standard*** environment on 3 held-out embodiments and do RL in the ***Diverse*** environment on 1 target embodiment, or vice-versa. This provides an additional layer of difficulty for the RL agent as visual randomizations show the brittleness of vision-based IRL methods.

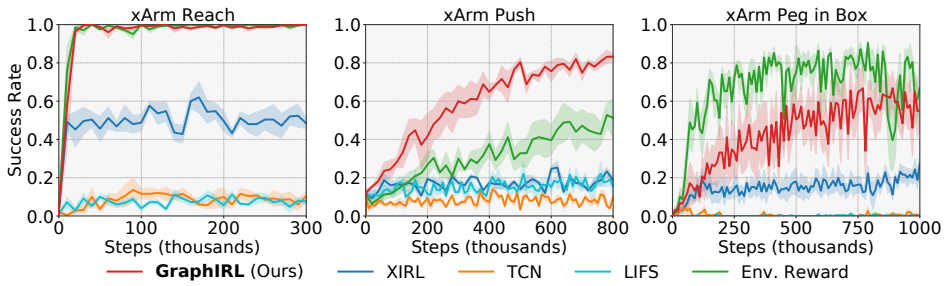

Figure 5: **Robotic Manipulation.** Success rates of our method *GraphIRL* and baselines on the tasks of *Reach*, *Push* and *Peg in Box*. All results are averaged over 5 seeds. We observe significant gains in performance specially over vision-based baselines due to large *domain-gap*

Real	XIRL	Env. Reward	GraphIRL (Ours)
Push	0.27	0.47	0.60
Reach	0.26	0.93	0.86
Peg in Box	0.06	0.60	0.53

Table 1: **Real robot experiments.** Success rate on robot manipulation tasks on physical hardware. We evaluate each method for 15 trials using a fixed set of goal and start state configurations.

Robotic Manipulation. Figure 1 shows initial and success configurations for each of the three task that we consider: **(1) Reach** in which the agent needs to reach a goal (red disc) with its end-effector, **(2) Push** in which the goal is to push a cube to a goal position, and **(3) Peg in Box** where the goal is to put a peg tied to the robot's end-effector inside a box. The last task is particularly difficult because it requires geometric 3D understanding of the objects. We collect a total of 256 and 162 video demonstrations for *Reach* and *Peg in Box*, respectively, and use 198 videos provided from Schmeckpeper et al. [22] for *Push*. The videos consist of human actors performing the same tasks but with a number of diverse objects and goal markers, as well as varied positions of objects. Unlike the data collected by Schmeckpeper et al. [22], we do not fix the goal position in our demonstrations. Additionally, we do not require the demonstrations to resemble the robotic environment in terms of appearance or distribution of goal location. In order to detect objects in our training demonstrations, we use a trained model from Shan et al. [57]. Refer to supplementary for details on robot platform.

4.2 Results

X-MAGICAL. Results for the *cross-embodiment and cross-environment* setting are shown in Figure 4. When trained on *Standard*, our method performs significantly better than vision-based baselines (*e.g.*, 0.58 GraphIRL for gripper vs 0.35 for XIRL and 0.99 GraphIRL for longstick vs 0.56 XIRL). We conjecture that vision-based baselines struggle with visual variations in the environment, which our method is unaffected by due to its graph abstraction. Additionally, we note that XIRL performs strongly when trained on *Diverse*. This could be because the *Diverse* environment contains some examples that are visually the same as *Standard* environment. To verify this, we conduct an experiment where we remove the shape (Square) and color (Red) used in *Standard* environment from the random configurations for *Diverse* environment and then construct a new set of demonstration. We refer to this environment as *Diverse (Out Of Distribution)*. In this setting, XIRL's performance drops significantly for 3 out of 4 embodiments.

Robotic manipulation in simulation. In this section, we answer the core question of our work: *can we learn to imitate others from diverse third-person videos?* In particular, we collect human demonstrations for manipulation tasks as explained in Section 4.1 and learn a reward function as explained in Section 3. This is a challenging setting because as shown in Figure 1, the collected data and robotic environments belong to different domains and do not share any appearance characteristics. Further, unlike previous works [22, 4], we do not use any environment reward as an additional supervision to the reinforcement learning agent. Figure 5 presents our results. For the **Reach** task, GraphIRL and environment reward achieve a success rate of 1.0, while other baseline methods are substantially behind GraphIRL (e.g. 0.477 XIRL and 0.155 TCN). In the **Push** setting, vision-based baseline methods still perform poorly. Similar to **Reach**, XIRL performed the best out of the vision-based baselines with a success rate of 0.187, and GraphIRL performed better than environment

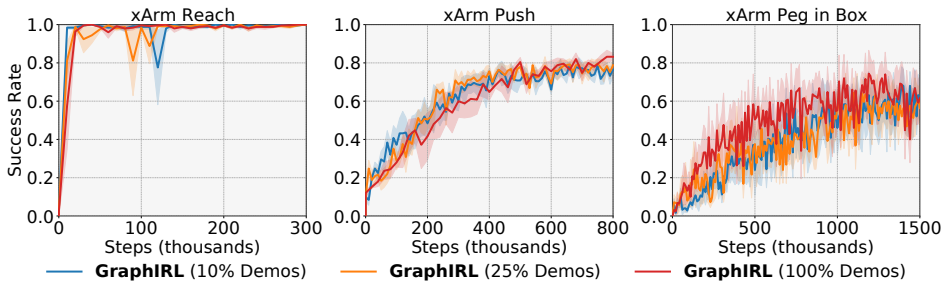

Figure 6: **Ablation Study on Number of Demonstrations.** Results averaged over 5 seeds. We use a total of 256, 198 and 162 videos for *Reach*, *Push* and *Peg in Box* respectively.

reward with a success rate of 0.832. The **Peg in Box** task is rigorous to solve since it requires 3-d reasoning and a precise reward function. Here, while all vision-based methods fail, GraphIRL solves the task with 55.2% success rate. Overall, our GraphIRL method is able to solve 2D and 3D reasoning tasks with a real-robot without a hand-designed reward function or access to 3D scene information.

Real robot experiments. Finally, we deploy the learned policies on a real robot. For each experiment, we conduct 15 trials per method and report the average success rate. Results are shown in Table 1. Interestingly, we find that GraphIRL outperforms XIRL in all three tasks on the real robot setup, and on *Push*, GraphIRL performs better than the environment reward specifically designed for the task (e.g. 0.93 Environment Reward vs 0.86 GraphIRL) which is in line with our findings in simulation.

4.3 Ablations

Impact of Modelling Spatial Interactions. We study the impact of modeling object-object spatial interactions using Spatial Interaction Encoder Network (IN) described in Section 3.1. Specifically, we replace our proposed encoder component with an Multi-Layer Perceptron (MLP) by concatenating representations of all objects into a single vector and then feeding it to a 3-layer MLP network. Results in Table 2. We observe that modeling object interactions leads to a 20% improvement in the RL success rate (i.e. 0.61 for MLP vs 0.804 for IN).

Variant	Success Rate
MLP	$0.61_{\pm 0.116}$
IN	$0.804_{\pm 0.054}$

Table 2: Impact of modelling object-object interaction on *Push* task. **MLP:** Multi-layer perceptron and **IN:** Spatial Interaction Network Encoder. Results over 5 seeds.

Impact of Decreasing Number of Demonstration Videos. Results in Figure 6. We find that our approach is very data efficient and can learn meaningful rewards even from a small number of videos. It achieves decent RL success rate with only 10% of total videos used (e.g. 72% for Push).

5 Conclusions and Limitations

Conclusion. We demonstrate the effectiveness of our proposed method, *GraphIRL*, in a number of IRL settings with diverse third-person demonstrations. In particular, we show that our method successfully learns reward functions from human demonstrations with diverse objects and scene configurations, that we are able to train image-based policies in simulation using our learned rewards, and that policies trained with our learned rewards are more successful than both prior work and manually designed reward functions on a real robot. We also hereby commit to release our complete code and data to the public.

Limitations. While our method relaxes the requirements for human demonstrations, collecting the demonstrations still requires human labor; and although our results indicate that we can learn from relatively few videos, eliminating human labor entirely remains an open problem. Moreover, our approach assumes access to object bounding boxes. This implies that our method's performance is dependent on the performance of the object detector. Fortunately, 2d object detectors have become very reliable as we show in our experiments, we are able to use an off-the-shelf object detector to extract the bounding boxes without having to perform any manual labeling. Finally, the proposed graph abstraction allows us to solve tasks despite the large domain gap but it has some potential disadvantages too. In particular, we lose fine grained information such as object poses and precise object interactions which could be useful for complex tasks such as medical procedures. However, we conjecture that 2D images might also not be sufficient for such a task since inferring accurate object pose, 3D geometric information from a 2D image itself is a challenging problem.

Acknowledgments

This work was supported, in part, by grants from NSF CCF-2112665 (TILOS), NSF 1730158 CI-New: Cognitive Hardware and Software Ecosystem Community Infrastructure (CHASE-CI), NSF ACI-1541349 CC*DNI Pacific Research Platform, and gifts from Meta, Google, Qualcomm.

The authors would like to thank Sanjay Haresh, Kevin Zakka, Jianglong Ye and Mohit Jain for helpful discussions.

References

[1] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. A. Riedmiller, A. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015.

[2] S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, and D. Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International journal of robotics research*, 37(4-5):421–436, 2018.

[3] O. M. Andrychowicz, B. Baker, M. Chociej, R. Jozefowicz, B. McGrew, J. Pachocki, A. Petron, M. Plappert, G. Powell, A. Ray, et al. Learning dexterous in-hand manipulation. *The International Journal of Robotics Research*, 39(1):3–20, 2020.

[4] K. Zakka, A. Zeng, P. Florence, J. Tompson, J. Bohg, and D. Dwibedi. Xirl: Cross-embodiment inverse reinforcement learning. In *Conference on Robot Learning*, pages 537–546. PMLR, 2022.

[5] P. Battaglia, R. Pascanu, M. Lai, D. Jimenez Rezende, et al. Interaction networks for learning about objects, relations and physics. *Advances in neural information processing systems*, 29, 2016.

[6] D. Dwibedi, Y. Aytar, J. Tompson, P. Sermanet, and A. Zisserman. Temporal cycle-consistency learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1801–1810, 2019.

[7] D. A. Pomerleau. Alvinn: An autonomous land vehicle in a neural network. *Advances in neural information processing systems*, 1, 1988.

[8] C. G. Atkeson and S. Schaal. Robot learning from demonstration. In *ICML*, volume 97, pages 12–20, 1997.

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

[10] H. Ravichandar, A. S. Polydoros, S. Chernova, and A. Billard. Recent advances in robot learning from demonstration. *Annual review of control, robotics, and autonomous systems*, 3:297–330, 2020.

[11] S. Song, A. Zeng, J. Lee, and T. Funkhouser. Grasping in the wild: Learning 6dof closed-loop grasping from low-cost demonstrations. *IEEE Robotics and Automation Letters*, 5(3): 4978–4985, 2020.

[12] S. Young, D. Gandhi, S. Tulsiani, A. Gupta, P. Abbeel, and L. Pinto. Visual imitation made easy. In *Conference on Robot Learning*, pages 1992–2005. PMLR, 2021.

[13] D. Pathak, P. Mahmoudieh, G. Luo, P. Agrawal, D. Chen, Y. Shentu, E. Shelhamer, J. Malik, A. A. Efros, and T. Darrell. Zero-shot visual imitation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2131–21313, 2018.

[14] F. Torabi, G. Warnell, and P. Stone. Behavioral cloning from observation. *ArXiv*, abs/1805.01954, 2018.

[15] I. Radosavovic, X. Wang, L. Pinto, and J. Malik. State-only imitation learning for dexterous manipulation. *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7865–7871, 2021.

[16] A. Y. Ng, S. Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, page 2, 2000.

[17] P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1, 2004.

[18] J. Ho and S. Ermon. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29, 2016.

[19] J. Fu, K. Luo, and S. Levine. Learning robust rewards with adversarial inverse reinforcement learning. *arXiv preprint arXiv:1710.11248*, 2017.

[20] Y. Aytar, T. Pfaff, D. Budden, T. Paine, Z. Wang, and N. De Freitas. Playing hard exploration games by watching youtube. *Advances in neural information processing systems*, 31, 2018.

[21] F. Torabi, G. Warnell, and P. Stone. Behavioral cloning from observation. *arXiv preprint arXiv:1805.01954*, 2018.

[22] 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.

[23] J. Jin, L. Petrich, Z. Zhang, M. Dehghan, and M. Jagersand. Visual geometric skill inference by watching human demonstration. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8985–8991. IEEE, 2020.

[24] H. Xiong, Q. Li, Y.-C. Chen, H. Bharadhwaj, S. Sinha, and A. Garg. Learning by watching: Physical imitation of manipulation skills from human videos. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7827–7834. IEEE, 2021.

[25] Y. Lee, A. Szot, S.-H. Sun, and J. J. Lim. Generalizable imitation learning from observation via inferring goal proximity. *Advances in Neural Information Processing Systems*, 34, 2021.

[26] A. S. Chen, S. Nair, and C. Finn. Learning generalizable robotic reward functions from "in-the-wild" human videos. *arXiv preprint arXiv:2103.16817*, 2021.

[27] A. Fickinger, S. Cohen, S. Russell, and B. Amos. Cross-domain imitation learning via optimal transport. *arXiv preprint arXiv:2110.03684*, 2021.

[28] Y. Qin, Y.-H. Wu, S. Liu, H. Jiang, R. Yang, Y. Fu, and X. Wang. Dexmv: Imitation learning for dexterous manipulation from human videos. In *European Conference on Computer Vision*, pages 570–587. Springer, 2022.

[29] S. P. Arunachalam, S. Silwal, B. Evans, and L. Pinto. Dexterous imitation made easy: A learning-based framework for efficient dexterous manipulation. *arXiv preprint arXiv:2203.13251*, 2022.

[30] G. E. Fainekos, A. Girard, H. Kress-Gazit, and G. J. Pappas. Temporal logic motion planning for dynamic robots. *Automatica*, 45(2):343–352, 2009.

[31] S. Srivastava, E. Fang, L. Riano, R. Chitnis, S. Russell, and P. Abbeel. Combined task and motion planning through an extensible planner-independent interface layer. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 639–646, 2014.

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

[33] S. Gupta, V. Tolani, J. Davidson, S. Levine, R. Sukthankar, and J. Malik. Cognitive mapping and planning for visual navigation. *International Journal of Computer Vision*, 128:1311–1330, 2019.

[34] W. Yang, X. Wang, A. Farhadi, A. Gupta, and R. Mottaghi. Visual semantic navigation using scene priors. *arXiv preprint arXiv:1810.06543*, 2018.

[35] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei. Scene graph generation by iterative message passing. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3097–3106, 2017.

[36] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang. Scene graph generation from objects, phrases and region captions. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1270–1279, 2017.

[37] N. Watters, D. Zoran, T. Weber, P. Battaglia, R. Pascanu, and A. Tacchetti. Visual interaction networks: Learning a physics simulator from video. *Advances in neural information processing systems*, 30, 2017.

[38] J. Materzynska, T. Xiao, R. Herzig, H. Xu, X. Wang, and T. Darrell. Something-else: Compositional action recognition with spatial-temporal interaction networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1049–1059, 2020.

[39] Y. Ye, M. Singh, A. Gupta, and S. Tulsiani. Compositional video prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10353–10362, 2019.

[40] H. Qi, X. Wang, D. Pathak, Y. Ma, and J. Malik. Learning long-term visual dynamics with region proposal interaction networks. *arXiv preprint arXiv:2008.02265*, 2020.

[41] L. Manuelli, Y. Li, P. Florence, and R. Tedrake. Keypoints into the future: Self-supervised correspondence in model-based reinforcement learning. *arXiv preprint arXiv:2009.05085*, 2020.

[42] A. Sanchez-Gonzalez, N. Heess, J. T. Springenberg, J. Merel, M. Riedmiller, R. Hadsell, and P. Battaglia. Graph networks as learnable physics engines for inference and control. In *International Conference on Machine Learning*, pages 4470–4479. PMLR, 2018.

[43] Y. Li, H. He, J. Wu, D. Katabi, and A. Torralba. Learning compositional koopman operators for model-based control. *arXiv preprint arXiv:1910.08264*, 2019.

[44] Y. Ye, D. Gandhi, A. Gupta, and S. Tulsiani. Object-centric forward modeling for model predictive control. In *Conference on Robot Learning*, pages 100–109. PMLR, 2020.

[45] E. Jang, C. Devin, V. Vanhoucke, and S. Levine. Grasp2vec: Learning object representations from self-supervised grasping. *arXiv preprint arXiv:1811.06964*, 2018.

[46] Y. Yang, Y. Liu, H. Liang, X. Lou, and C. Choi. Attribute-based robotic grasping with one-grasp adaptation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6357–6363. IEEE, 2021.

[47] P. Florence, L. Manuelli, and R. Tedrake. Self-supervised correspondence in visuomotor policy learning. *IEEE Robotics and Automation Letters*, 5(2):492–499, 2019.

[48] M. Vecerik, J.-B. Regli, O. Sushkov, D. Barker, R. Pevceviciute, T. Rothörl, C. Schuster, R. Hadsell, L. Agapito, and J. Scholz. S3k: Self-supervised semantic keypoints for robotic manipulation via multi-view consistency. *arXiv preprint arXiv:2009.14711*, 2020.

[49] M. Sieb, Z. Xian, A. Huang, O. Kroemer, and K. Fragkiadaki. Graph-structured visual imitation. In *Conference on Robot Learning*, pages 979–989. PMLR, 2020.

[50] S. Haresh, S. Kumar, H. Coskun, S. N. Syed, A. Konin, Z. Zia, and Q.-H. Tran. Learning by aligning videos in time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5548–5558, 2021.

[51] W. Liu, B. Tekin, H. Coskun, V. Vineet, P. Fua, and M. Pollefeys. Learning to align sequential actions in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2181–2191, 2022.

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

[53] R. Jangir, N. Hansen, S. Ghosal, M. Jain, and X. Wang. Look closer: Bridging egocentric and third-person views with transformers for robotic manipulation. *IEEE Robotics and Automation Letters*, 7(2):3046–3053, 2022.

[54] P. Sermanet, C. Lynch, Y. Chebotar, J. Hsu, E. Jang, S. Schaal, S. Levine, and G. Brain. Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 1134–1141. IEEE, 2018.

[55] A. Gupta, C. Devin, Y. Liu, P. Abbeel, and S. Levine. Learning invariant feature spaces to transfer skills with reinforcement learning. *arXiv preprint arXiv:1703.02949*, 2017.

[56] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[57] D. Shan, J. Geng, M. Shu, and D. F. Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9869–9878, 2020.

[58] D. Yarats, I. Kostrikov, and R. Fergus. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. In *International Conference on Learning Representations*, 2020.

[59] N. Hansen, H. Su, and X. Wang. Stabilizing deep q-learning with convnets and vision transformers under data augmentation. *Advances in Neural Information Processing Systems*, 34: 3680–3693, 2021.

Appendix A Qualitative Analysis of Learned Reward

In this section, we present qualitative analysis of the reward learned using *GraphIRL*. We plot the reward for *GraphIRL* and two baseline IRL methods for three test examples across three tasks. The tasks we evaluate with are *Peg in Box*, *Push*, and *Reach*. For each task, we use show two successful episodes and one unsuccessful episode. The length of each episode is 50, and for each figure we have included, we provide images that align with critical points in the completion of the task.

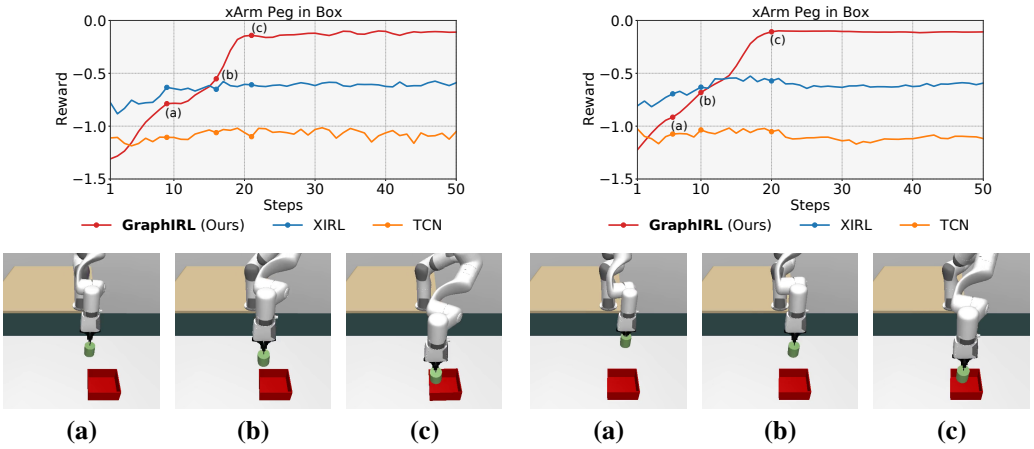

Figure 7: **Peg in Box Task Progress: Success**. For the *Peg in Box* task setting, we find that *GraphIRL* provides an accurate measurement of task progress. Pictured are video frames **(a)**, **(b)**, **(c)** which denote critical points of task progress. Task progress is measured using video frames from a 50-step evaluation episode.

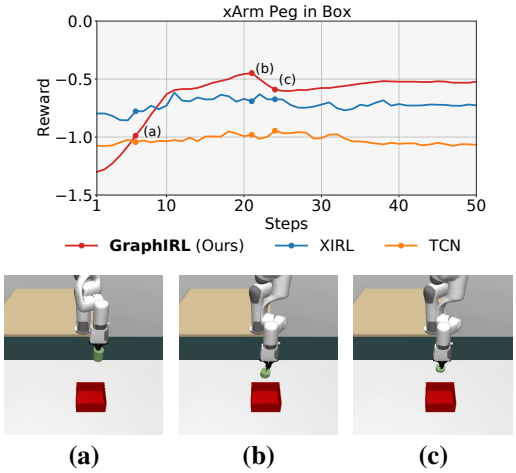

Figure 8: **Peg in Box Task Progress: Failure**. *GraphIRL* measures positive task progress until the peg goes into the table, a critical failure point for the task. The physical interaction between the peg and table is unnatural, and our method succeeds in recognizing this.

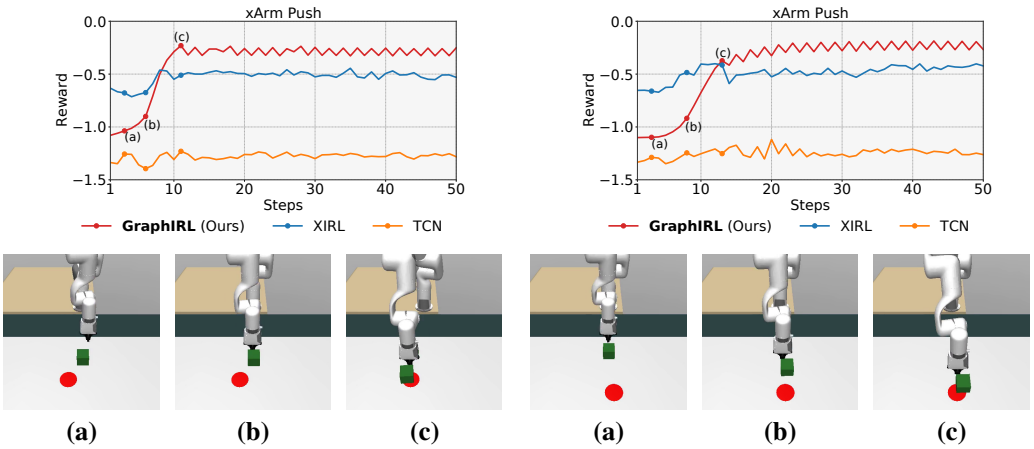

Figure 9: **Push Task Progress: Success**. The *Push* task setting is often completed within the first 10 steps of the evaluation episode, and as shown between Steps 1 through 10 in both success examples, *GraphIRL* measures high task progress. XIRL and TCN on the other hand, incorrectly show much lower task progress.

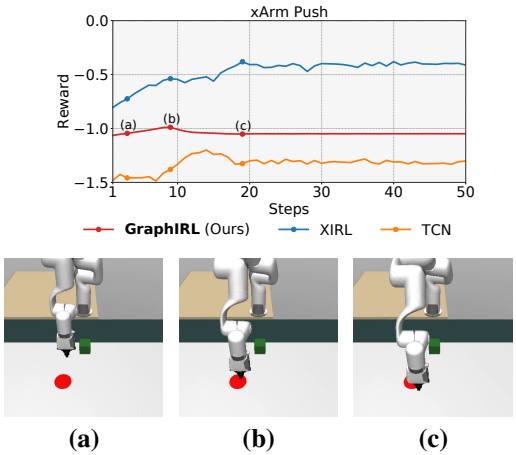

Figure 10: **Push Task Progress: Failure**. *GraphIRL*'s understanding of object relationships is made clear in this *Push* task failure, since without any forward movement of the box object towards the goal, no positive task progress is made. Other baselines rely on direct visual input of the task, and because of this, they inaccurately align visual states **(a)**, **(b)**, **(c)** of the task with positive task progress.

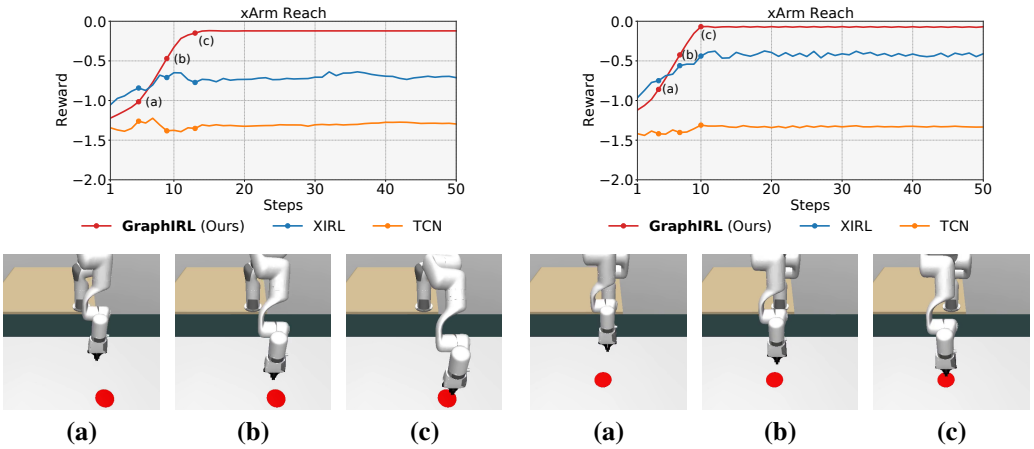

Figure 11: **Reach Task Progress: Success**. In the *Reach* task setting, positive task progress is measured by *GraphIRL* with forward movement of the end-effector gripper towards the goal location. The image frames **(a)**, **(b)**, **(c)** reflect the alignment between measured task progress and visual state of the task.

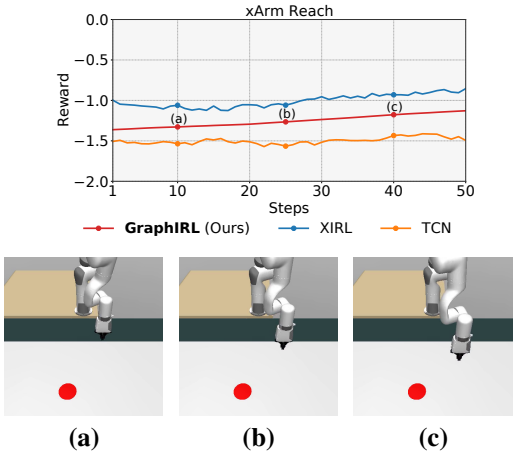

Figure 12: **Reach Task Progress: Failure**. Our *GraphIRL* method measures an approximately linear task progress in this failure example for *Reach*. The gripper's distance to the goal region is indeed minimized over time, though since it does not get within close-enough distance to the goal, the measured task progress is lower compared to success examples shown in Figure 11.

We find that our method provides a superior and accurate reward signal to the agent compared to the baseline visual IRL methods. We observe that if a task is being completed successfully or unsuccessfully in a video, our method can obtain a reward that accurately reflects how close the agent is to completing the task. Additionally, both *XIRL* and *TCN* yield low reward even for successful episodes due to large distance between the current observation and the representative goal observation in the embedding space which could be attributed to visual domain shift.

Appendix B Additional Implementation Details

Representation Learning. Each MLP in the Spatial Interaction Encoder Network defined in Equation 3.1 is implemented as a 2-layer network with a ReLU activation. The size of the final embedding $\psi(\cdot)$ is 128 in our experiments. Please see Table 3 for a detailed list of hyperparameters for representation learning. All the hyperparameters in Table 3 are kept fixed for all tasks considered in this work.

Reinforcement Learning. For X-MAGICAL, we follow Zakka et al. [4] and learn a state based policy. The state vector has dimensions of 16 and 17 for the *Standard* and *Diverse* environments

Hyperparameter	Value
# of sampled frames	90
Batch Size	2
Learning Rate	10^{-5}
Weight Decay	10^{-5}
# of training iterations	12000
Embedding Size	128
Softmax Temperature	0.1

Table 3: Hyperparameters for Representation Learning with GraphIRL.

respectively. The *Diverse* environment state has an additional dimension to represent the size of blocks. For xArm, we learn an image based policy. Specifically, we use first-person and third-person cameras to learn a policy from multi-view image data. We extract 84×84 image from both cameras and concatenate them channel-wise. We use the network architecture and attention mechanism proposed in Jangir et al. [53]. Additionally, we apply data augmentation techniques: random ± 4 pixel shift [58] and color jitter [59].

Extracting Reward. In order to compute the reward during Reinforcement Learning (RL) training, we use the locations of objects available in simulation to extract the bounding boxes corresponding to the current observation. The bounding boxes are used to construct the object representation which is then passed to the trained Spatial Interaction Encoder Network to get the reward.

Criterion for Success. We use distance threshold to determine the success of an episode. The thresholds are 5cms, 10cms and 8cms for *Reach*, *Push* and *Peg in Box* respectively. The distance refers to distance between goal position and end-effector for *Reach*, and goal position and object position for *Push* and *Peg in Box*.

Baseline Implementation Details. For all the vision-based baselines, we use the hyperparameters, data augmentation schemes and network architectures provided in Zakka et al. [4]. Readers are encouraged to read Zakka et al. [4] for more details on the vision-based baselines.

Appendix C X-MAGICAL Experiment Details

C.1 Demonstration Data

For collecting demonstration data in the X-MAGICAL *Diverse* environment, we trained 5 uniquely-seeded Soft Actor-Critic (SAC) RL policies for 2 million steps for each embodiment using the environment reward. We collect 1000 successful episode rollouts for each embodiment using the 5 trained policies. In particular, each policy is used to produce 200 episode rollouts for a given embodiment.

C.2 Diverse Environment

Below, we explain the randomization performed on the blocks in the diverse environment that we use in our experiments:

- **Color:** We randomly assign 1 out of 4 colors to each block.
- **Shape:** Each block is randomly assigned 1 out of 6 shapes.
- **Size:** The block sizes are also varied. In particular, we generate a number between 0.75 and 1.25 and multiply the default block size by that factor.
- **Initial Orientation:** The initial orientation of the blocks is also randomized. We randomly pick a value between 0 to 360 degrees.
- **Initial Location:** The initial location of the boxes is randomized by first randomly picking a position for the y-coordinate for all blocks and then randomly selecting x-coordinate separately for each block. This randomization is also performed in the standard environment.

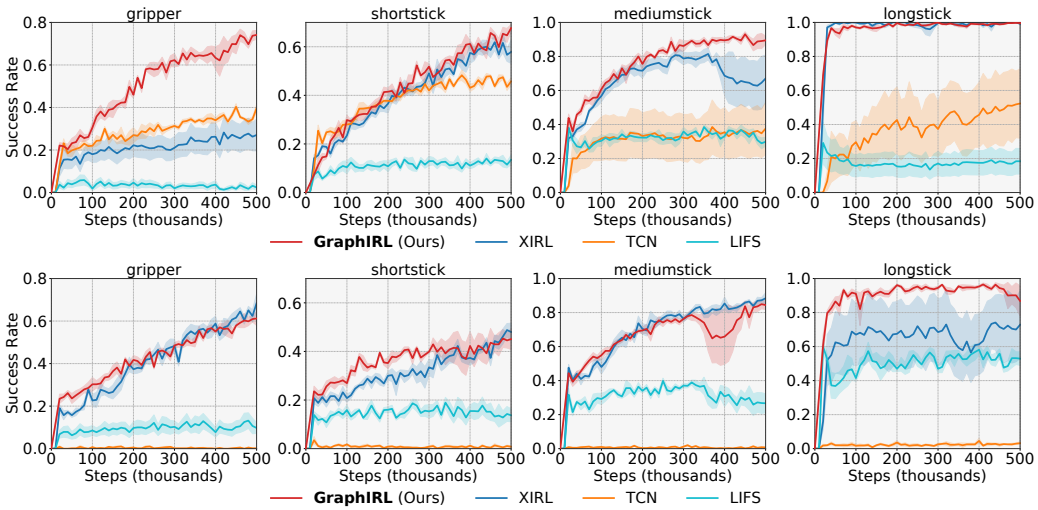

Figure 13: ***Cross-Embodiment Same-Environment:*** We further evaluate *GraphIRL* in the cross-embodiment same-environment setting *(top)* Standard Environment *(bottom)* Diverse Environment, and it continues to provide competitive success rates akin to those achieved by XIRL. These results confirm that *GraphIRL* is a consistent and reliable method for learning from video demonstrations in visually similar environments.

Appendix D Additional Results on X-MAGICAL Benchmark

To complement our cross-embodiment cross-environment results from the main paper, we also report results for *X-MAGICAL* in the *cross-embodiment same-environment* setting. As shown in Figure 13, we outperform *TCN* and *LIFS* by significant margins and achieve comparable results to *XIRL*. These results reflect the effectiveness of *GraphIRL* when learning in a visually similar environment with visually different agents.

Appendix E Appendix E: xArm Experiment Details

E.1 Description of Environment Rewards

In this section, we define the environment rewards for xArm environments that were compared against GraphIRL in robot manipulation experiments under Section 4. We define p_g, p_o, and p_e as the positions of the goal, object and robot end-effector respectively. The reward for *Push* is defined as $||p_o - p_g||^2$, for reach it becomes $||p_e - p_g||^2$ and finally for *Peg in Box*, the reward is $||p_o - p_g||^2$. Note that the distances are computed using 2-d positions in the case of *Reach* and *Push* and 3-d positions in the case of *Peg in Box*.

E.2 Demonstration Data

We use data from [22] for *Push*. We collect 256 and 162 demonstrations respectively for *Reach* and *Peg in Box*. For *Reach*, we use 18 visually distinct goal position markers *i.e.* 3 different shapes and each shape with 6 different colors in order to ensure visual diversity. *Reach* demonstrations have minimum, average and maximum demonstration lengths of 1.76 seconds, 4.51 seconds and 9.23 seconds respectively. For *Peg in Box*, we use 4 visually distinct objects. In this case, the minimum, average and maximum demonstration lengths are 1.73 seconds, 4.74 seconds and 11.7 seconds respectively. For both *Reach* and *Peg in Box*, the goal and object positions are also varied across demonstrations to diversify trajectories. Please see `https://sateeshkumar21.github.io/GraphIRL/` for examples of collected demonstrations.

Appendix F Impact of Decreasing Number of Demonstration Videos for Baselines

In Section 4.3, we show how decreasing number of demonstrations impacts performance for *GraphIRL*. For completeness, we repeat the same experiment for the vision-based baslines con-

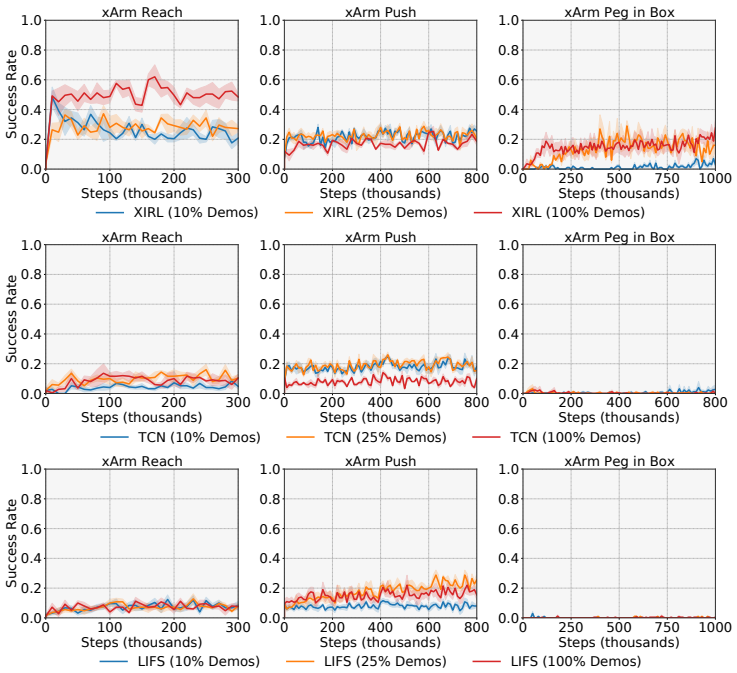

Figure 14: **Ablation Study on Number of Demonstrations.** Results averaged over 5 seeds. We use a total of 256, 198 and 162 videos for *Reach*, *Push* and *Peg in Box* respectively.

sidered in this work i.e. *XIRL*, *TCN* and *LIFS*. Results in Figure 14. We note that for *Reach*, *XIRL*'s performance degrades significantly with lesser number of demonstrations. However, in other cases where the success rates are lower, we notice an inconsistent trend, indicating that vision-based baseline fail to learn meaningful representations regardless of the number of demonstrations.

Appendix G Additional Results on Robot Manipulation in Simulation

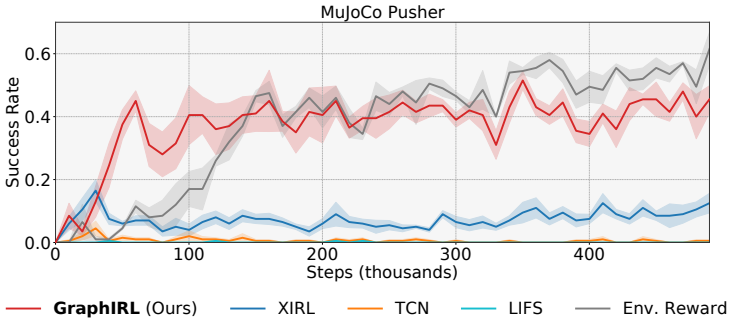

Figure 15: **Results on MuJoCo Pusher Task** Success rates of *GraphIRL* and baselines. All results are averaged over 5 seeds. GraphIRL outperforms vision-based baselines and achieves performance close to the environment reward.

We also experiment with the *MuJoCo State Pusher* environment used by Schmeckpeper et al. [22] and Zakka et al. [4]. However, we make two changes, (1) Instead of using a fixed goal position, we use a randomized goal position and learn a goal-conditioned policy and (2) we do not use the sparse environment reward and instead only use the learned rewards for GraphIRL and learning-based baselines. Figure 15 presents our results, we note that GraphIRL achieves slightly lower success rate than the task-specific environment reward (e.g. GraphIRL 0.455 vs Environment Reward 0.6133). Further, all vision-based baselines perform significantly lower than GraphIRL (e.g. GraphIRL 0.455 vs XIRL 0.125 and TCN 0.005). For all learning-based methods, we use the data from Schmeckpeper et al. [22] as training demonstrations similar to *Push* experiments conducted in Section 4.

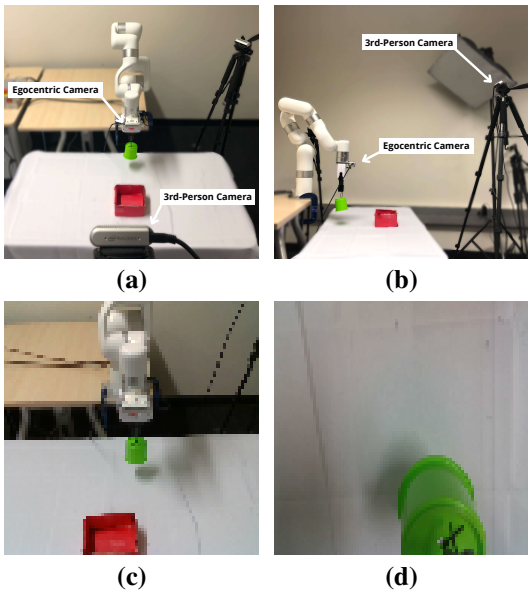

Figure 16: **Real Robot Setup**. In **(a)** and **(b)**, we provide images of our real-world environment for the *Peg in Box* task. We use a static third-person camera and an egocentric camera which moves with the arm while completing the task. Pictured in **(c)** and **(d)** are single image frames captured by our third-person and egocentric cameras.

Appendix H Robot Setup

We use a Ufactory xArm 7 robot for our real robot experiments. As shown in Figure 16, we use a fixed third-person camera and an egocentric camera that is attached above the robot's gripper. Example images of the egocentric and third-person camera feeds passed to the RL agent are shown in Figure 16 (c) and Figure 16 (d).