Graph-Based Reward Learning from Demonstrations for Long-Horizon Manipulation

Andrea Protopapa[1], Davide Buoso[1], Francesca Pistilli[1] and Giuseppe Averta[1]

***Abstract*— Learning skills from visual demonstrations remains challenging because reward design is difficult and pixel-based representations does not take advantage of the spatial and relational structure that makes manipulation tasks interpretable and robust. In this work, we study whether structured scene representations can improve reward learning for long-horizon manipulation. Our approach converts each observation into an object-centric graph whose nodes encode task entities and whose edges encode pairwise spatial relations. A graph neural network maps this representation into a latent space, and a weighted pooling mechanism emphasizes task-relevant object dynamics while suppressing robot-dominated motion. The learned embedding defines a dense reward through latent distance to the goal. Across two manipulation tasks, this structured reward is both more effective and more interpretable than less structured alternatives: on the multi-object manipulation task *Match Regions*, it improves downstream reinforcement learning performance over pixel and graph baselines and exceeds the simulator reward, while on the long-horizon task *Shoes in Box* it exhibits clear stage-wise transitions aligned with semantic task phases. These results suggest that lightweight geometric structure over objects and relations provides a useful inductive bias for modern data-driven robot learning.**

I. INTRODUCTION

Robotics has long relied on geometric structure to model interaction with the physical world. However, many modern learning-based pipelines operate directly on raw observations, discarding explicit structure over objects, relations, and task progression. This drawback is especially pronounced in long-horizon manipulation, where success depends more on semantically meaningful changes in spatial configuration.

Reward learning from demonstrations offers a natural bridge between Imitation Learning (IL) and Reinforcement Learning (RL), since demonstrations are used to provide supervision for task progress, while Reinforcement Learning uses the learned reward to optimize behavior through interaction. In particular, Inverse Reinforcement Learning (IRL) learns rewards directly from videos removing the need for action annotations and enables scalable data collection [1]. Nevertheless, pixel-based reward learning is often brittle, since visual changes caused by viewpoint, texture, or robot motion can dominate the representation even when they are only weakly related to actual task completion. In complex manipulation tasks and long-horizon settings, this makes it difficult to obtain rewards that are both informative and semantically aligned.

In this work, we investigate a structured alternative. Rather than learning rewards directly from pixels, we represent each scene as an object-centric graph whose nodes encode entities and whose edges encode pairwise spatial relations. A graph neural network (GNN) then learns a latent representation from action-free video demonstrations. The central hypothesis is that for manipulation, progress is better described by the evolution of object relations than by raw visual similarity. To encourage this behavior, we introduce a weighted pooling mechanism that emphasizes active task objects while suppressing robot nodes, reducing the influence of manipulator motion that may dominate frame-to-frame change without reflecting actual semantic progress.

Our experiments support this perspective. On *Match Regions*, a multi-object manipulation task, the learned object-centric reward improves downstream reinforcement learning performance over pixel-based and prior graph-based baselines, and even outperforms the handcrafted environmental reward defined by the simulator. On the longer-horizon *Shoes in Box* task, the learned reward evolves through a clear sequence of step-like transitions aligned with meaningful interaction phases, suggesting a natural choice to subtask discovery. Taken together, these results indicate that explicit relational structure remains a valuable inductive bias within modern data-driven robot learning.

II. METHOD

A. Object-Centric Structured Representation

We consider the problem of learning reward functions for long-horizon manipulation from action-free video demonstrations. Let $\mathcal{D} = \{V_k\}_{k=1}^{K}$ denote a set of demonstration videos for a task $\mathcal{T}$, where each video $V_k = \{I_k^1, I_k^2, \dots, I_k^{T_k}\}$ is a sequence of image frames.

For each frame I_t, we extract a set of N objects using an off-the-shelf detector and represent the scene as a object-centric fully-connected graph $G_t = (V_t, E_t)$. Each node $v_i \in V_t$ corresponds to a detected object and is described by the semantic class and bounding-box geometry as node features. Edges $e_{ij} \in E_t$ encode pairwise spatial relations between objects, namely the relative distances. This representation removes much of the irrelevant appearance variation present in raw pixels while preserving the entities and interactions that matter for manipulation.

Given a sequence of graphs $\{G_t\}$, we learn an embedding function $\phi(G_t)$ that produces a compact representation of the

This study was carried out within the project FAIR - Future Artificial Intelligence Research - and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 – D.D. 1555 11/10/2022, PE00000013). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

[1]Politecnico di Torino, Turin, Italy, [firstname].[lastname]@polito.it

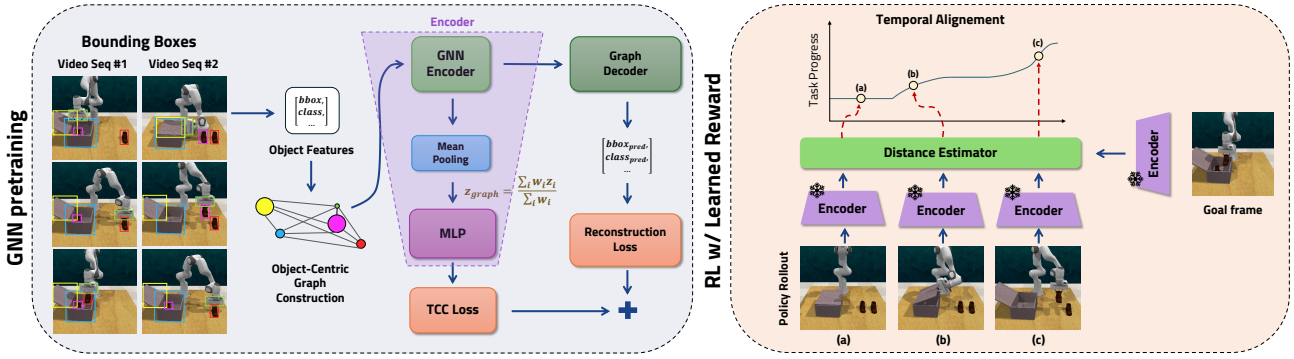

Fig. 1. Overview of the proposed framework. Demonstration videos are converted into object-centric graphs and encoded with a graph neural network trained using temporal cycle-consistency (TCC) [2] and reconstruction losses. The learned encoder is then frozen and used to define a reward for reinforcement learning by measuring the latent-space distance between the current observation and the goal. The resulting reward captures semantic task progress and can be used directly for policy learning; in longer-horizon tasks, its stage-wise temporal evolution can additionally be exploited to reveal subtasks structure.

task state encoded in the graph. A GNN encoder processes node and edge features and produces node embeddings $\{z_i\}_{i=1}^{N}$. These are aggregated into a graph-level representation $z_t = \phi(G_t)$ that serves as the latent task state.

B. Weighted Pooling for Task-Relevant Progress

A key challenge in visual manipulation is that robot motion often dominates observation changes without being the most informative signal for task progress. We therefore aggregate node embeddings with a weighted pooling rule:

$$z_g = \frac{\sum_i w_i z_i}{\sum_i w_i}, \tag{1}$$

where

$$w_i = (1 + (\alpha - 1) \cdot \text{active}_i)(1 - \text{robot}_i), \tag{2}$$

Here, $\text{active}_i \in \{0, 1\}$ indicates whether object i is currently active, and $\text{robot}_i \in \{0, 1\}$ indicates whether the node corresponds to the robot. This construction upweights task-relevant moving objects while suppressing robot nodes entirely. Intuitively, it biases the representation toward changes in object-centric spatial configuration rather than manipulator-dominated visual motion.

We estimate activity from temporal displacement of bounding-box centers, normalized by object size. Once activated, an object remains active for the rest of the episode. This persistence encourages a monotonic notion of progress that is especially useful in long-horizon tasks.

C. Learning the Reward from Demonstrations

The encoder is pretrained self-supervised from action-free demonstration sequences using two complementary objectives: temporal alignment across demonstrations and reconstruction of object-level structure. Temporal alignment encourages embeddings from similar task stages to lie nearby even when demonstrations differ in timing [2], while reconstruction preserves enough scene information to support reward learning.

TABLE I

FINAL SUCCESS RATE ON *Match Regions*.

Method	Success Rate (%) ↑
Environmental reward	85.0
XIRL [1]	0.0
GraphIRL [4] (w/ robot)	0.0
GraphIRL [4] (w/o robot)	33.7
Ours (w/ robot)	33.3
Ours (w/o robot)	**94.6**

After pretraining, we freeze the encoder and define reward through latent distance between the current observation graph G_t and a goal graph G_g:

$$r_t = -\|\phi(G_t) - \phi(G_g)\|_2. \tag{3}$$

This yields a dense reward for reinforcement learning. Because the latent space is built from objects and relations rather than pixels, reward progression is better aligned with semantic task completion.

An overview of the proposed framework is shown in Fig. 1.

III. EXPERIMENTS

We evaluate the approach on two manipulation tasks that highlight complementary benefits of structured reward learning: improved downstream RL on a shorter task, and interpretable long-horizon task progression on a more complex one.

A. Match Regions: Improved Downstream RL

Match Regions, drawn from the MAGICAL benchmark [3], evaluates whether the learned reward is effective for downstream policy optimization. In this task, the robot must place objects into designated target regions. Success therefore depends on correctly tracking object placement and scene configuration.

We compare our method against a pixel-based IRL baseline, namely XIRL [1], and a graph-based variant – i.e., GraphIRL [4] – with different pooling strategies. Moreover,

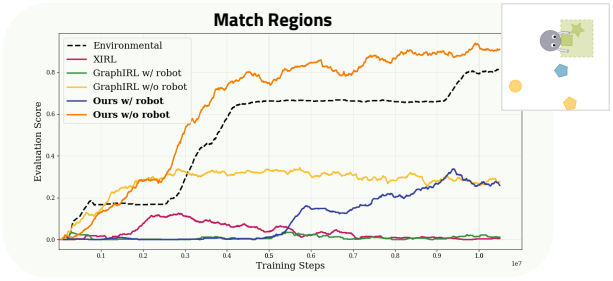

Fig. 2. Results on *Match Regions*. Our object-centric GNN-based reward learning approach achieves the best final performance and the fastest convergence among all compared methods.

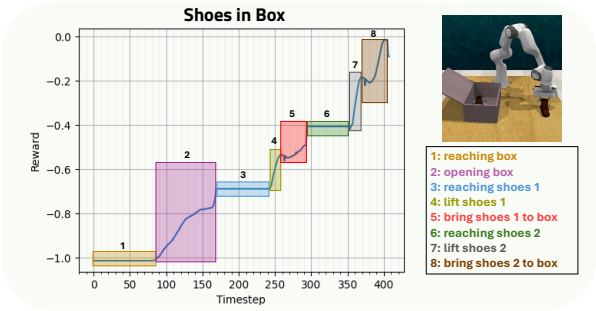

Fig. 3. Learned reward profile on *Shoes in Box*. The reward evolves through a sequence of step-like increases aligned with meaningful interaction phases, such as opening the box and placing each shoe.

we also compare it against the hand-crafted environmental reward provided by the simulator.

Our structured reward achieves the strongest final performance among compared learned-reward methods. In particular, the version that suppresses robot nodes during pooling reaches a final success rate of 94.6%, outperforming the handcrafted environmental reward at 85.0%, as shown also in Table I. This result is notable because the simulator reward is manually engineered using privileged task information, whereas our method learns reward directly from demonstrations. As shown in Fig. 2, it also converges faster, suggesting that the object-centric latent reward provides a denser and more informative optimization signal. Overall, this experiment indicates that even a lightweight structured prior over objects and relations can substantially improve reward quality for manipulation.

B. *Shoes in Box: Interpretable Long-Horizon Progress*

We next evaluate the method on *Shoes in Box*, a significantly more challenging long-horizon task from RLBench [5] involving multiple sequential interaction phases, including reaching and opening the box, manipulating the first shoe, and repeating the procedure for the second shoe.

For this experiment, depicted in Figure 3, the most interesting effect is the structure of the learned reward by our method. The learned reward evolves through a sequence of clear step-like transitions over time, aligned with semantically meaningful events such as opening the box and placing each shoe. This indicates that the learned representation semantically captures the discrete progress in object-centric scene configuration.

TABLE II

REWARD DISCRIMINATION ON *Shoes in Box*, MEASURED BY THE SUCCESS/FAILURE CUMULATIVE REWARD RATIO. LOWER IS BETTER.

Method	Positive/Negative Ratio $\rho \downarrow$
XIRL [1]	0.713
GraphIRL [4] w/ robot	0.844
GraphIRL [4] w/o robot	1.098
Ours w/ robot	**0.673**
Ours w/o robot	**0.684**

To quantify reward quality, we compute cumulative learned reward on validation trajectories and compare successful and unsuccessful executions through

$$\rho = \frac{\bar{R}^{+}}{\bar{R}^{-}}, \tag{4}$$

where $\bar{R}^{+}$ and $\bar{R}^{-}$ are the mean cumulative rewards for successful and unsuccessful trajectories. Because rewards are always negative, lower ρ indicates better separation.

As shown in Table II, our method achieves stronger separation than the compared baselines, indicating that unsuccessful executions accumulate substantially worse reward. Combined with the observed step-like temporal profile, this suggests that the learned reward provides a useful signal for identifying candidate subtask boundaries without manual annotation.

IV. DISCUSSION

Our results suggest that incorporating object-centric spatial structure into reward learning improves both performance and interpretability. In manipulation, progress is often defined by changes in spatial relations among task objects, i.e., whether an object has been grasped, moved into a container, aligned with a target region, or placed in the correct location. Pixel-space embeddings are only an indirect proxy for this structure and are easily confounded by irrelevant variation. By contrast, an object-centric relational representation provides a simple geometric prior for learning.

Overall, our results support combining learned representations with explicit structure over objects and spatial relations, rather than relying on raw visual input alone.

V. CONCLUSION

We presented a geometry-aware graph-based reward learning approach for long-horizon manipulation that combines object-centric relational structure with data-driven representation learning. Across two tasks, the resulting reward is both more effective for reinforcement learning and more interpretable than less structured alternatives. These findings suggest that explicit spatial and relational structure remains highly valuable in modern robot learning, not as a substitute for data-driven methods, but as a useful geometric prior within them.

REFERENCES

[1] K. Zakka, A. Zeng, P. Florence, J. Tompson, J. Bohg, and D. Dwibedi, "XIRL: Cross-embodiment inverse reinforcement learning," in *Proceedings of the Conference on Robot Learning*, pp. 537–546, 2022.

[2] 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 (CVPR)*, 2019, pp. 1801–1810.

[3] S. Toyer, R. Shah, A. Critch, and S. Russell, "The MAGICAL benchmark for robust imitation," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 18284–18295, 2020.

[4] S. Kumar, J. Zamora, N. Hansen, R. Jangir, and X. Wang, "Graph inverse reinforcement learning from diverse videos," in *Proceedings of the Conference on Robot Learning*, pp. 55–66, 2023.

[5] S. James, Z. Ma, D. R. Arrojo, and A. J. Davison, "RLBench: The robot learning benchmark & learning environment," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3019–3026, 2020.