Efficient Bimanual Handover and Rearrangement via Symmetry-Aware Actor-Critic Learning

Yunfei Li[1*], Chaoyi Pan[2*], Huazhe Xu[1,4,5], Xiaolong Wang[3] and Yi Wu[1,5]

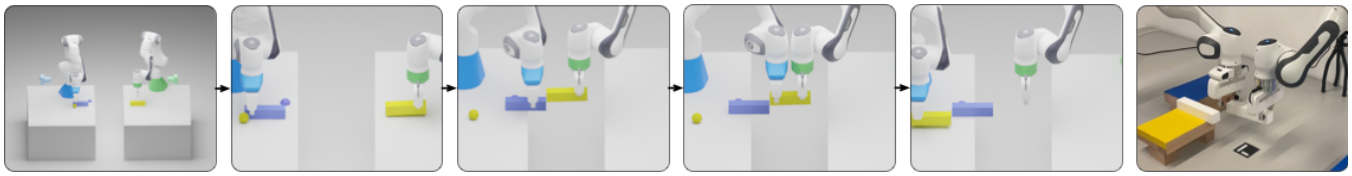


Fig. 1: Illustration of a bimanual handover and rearrangement task. Two Franka Panda arms mounted on separate tables aim to transport multiple objects to goal positions on either side of the gap.

***Abstract*— Bimanual manipulation is important for building intelligent robots that unlock richer skills than single arms. We consider a multi-object bimanual rearrangement task, where a reinforcement learning (RL) agent aims to jointly control two arms to rearrange these objects as fast as possible. Solving this task efficiently is challenging for an RL agent due to the requirement of discovering precise intra-arm coordination in an exponentially large control space. We develop a symmetry-aware actor-critic framework that leverages the interchangeable roles of the two manipulators in the bimanual control setting to reduce the policy search space. To handle the compositionality over multiple objects, we augment training data with an object-centric relabeling technique. The overall approach produces an RL policy that can rearrange up to 8 objects with a success rate of over 70% in simulation. We deploy the policy to two Franka Panda arms and further show a successful demo on human-robot collaboration. Videos can be found at `https://sites.google.com/view/bimanual`.**

I. INTRODUCTION

Bimanual manipulation is an important component for building intelligent robots [1]. With two controllable manipulators, an agent can solve a richer set of tasks compared with the setting of single-arm control [2] and can be even further applied to the setting of cooperation with humans by substituting a controller with a human at test time [3]. We focus on developing a reinforcement learning agent to tackle a sparse-reward bimanual manipulation task shown in Fig. 1, where two robotic arms on separate tables are required to rearrange each object to its goal position. A success reward is given only when a goal is reached. Some objects have goals on the same table that can be reached via a single arm, while other objects should be transported over the gap to the other table, which can only be achieved via a cooperative handover between two arms.

Learning efficient bimanual policies poses unique challenges. The agent must explore an enlarged control space with two arms for precise coordination between manipulators, such as cooperative handover and workload balance. In our rearrangement problem, the task space also grows exponentially with the number of objects to rearrange, which poses substantial challenges for policy learning. Existing approaches typically require predefined policy abstractions [4], [5] or expert demonstrations [6], [7] to learn bimanual tasks.

To enable efficient RL on this challenging bimanual rearrangement problem, we propose a novel symmetry-aware actor-critic framework by exploiting the interchangeable roles of two manipulators. As shown in Fig. 2, each scene can be viewed from two actor-centric frames as a pair of task instances s and s^M. The instances differ in the identities of manipulators. The optimal solution to one instance can be directly transferred to the mirrored instance by flipping actions across manipulators, i.e., the top manipulator in s must share the same optimal action with the bottom one in s^M since they face the same situation, and the same holds for the other two manipulators. Accordingly, the optimal values should be equivalent between the mirrored pair. We propose to incorporate the symmetric structure into actor-critic architectures so as to reduce the search space of RL. In critic networks, we take the average value of mirrored state-action pairs as the final value prediction. For policy learning, we use a shared actor network with mirrored input states to obtain actions for two manipulators.

Furthermore, to tackle the sparse-reward challenge in the multi-object rearrangement problem, we propose *object-centric relabeling* for data augmentation, which extends hindsight experience relay (HER) [8] from the single-goal setting to the multi-goal setting. In the early stage of training, only a few objects will be moved in a trajectory. Relabeling the goals for those untouched objects to achieved states will make the augmented data largely biased. We instead only relabel the goals of moved objects while sampling random goals for other objects, which enables more stable

*Equal contribution.

[1]Yunfei Li, Huazhe Xu and Yi Wu are with Institute of Interdisciplinary Information Sciences, Tsinghua University, Beijing, China. liyf20@mail.tsinghua.edu.cn

[2]Chaoyi Pan is with Department of Electronic Engineering, Tsinghua University, Beijing, China. pcy19@mails.tsinghua.edu.cn

[3]Xiaolong Wang is with Department of Electrical and Computer Engineering, UC San Diego, CA, USA.

[4] Shanghai Artificial Intelligence Lab, Shanghai, China.

[5] Shanghai Qi Zhi Institute, Shanghai, China.

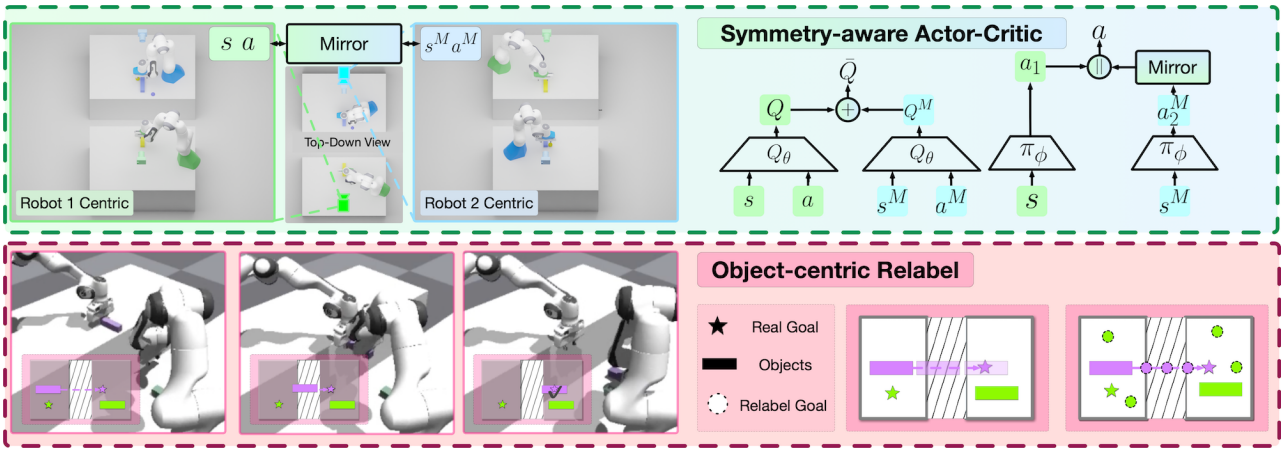


Fig. 2: Overview of the proposed symmetry-aware actor-critic framework for efficient multi-object handover and rearrangement. *Top*: symmetry-aware actor-critic network architecture leveraging the policy and value equivalence in pairs of symmetric tasks. *Bottom*: object-centric hindsight goal relabeling to capture compositionality in sparse-reward multi-object scenarios.

and faster training. We implement the proposed framework upon SAC [9] and conduct simulated experiments in Isaac gym [10]. Experiment results show that the proposed framework can efficiently learn to rearrange 8 objects to either side of the gap. We also deploy the learned policy to two Franka Panda arms and demonstrate an initial attempt at human-robot cooperative handover and rearrangement.

II. RELATED WORK

Object rearrangement is a popular test bench for developing robots with embodied intelligence [11]. A considerable number of works study planning-based methods to compute either prehensile [12], [13], [14] or non-prehensile [15], [16], [17] robot motions for object rearrangement. Some recent works integrate planning-based methods with data-driven models when the environment is not perfectly known [18]. As for end-to-end learning of object rearrangement tasks, most works study single-arm manipulation with emphasis on visual learning [19], [20], [21] or automatic goal discovery [22]. We instead learn *bimanual* object rearrangement with deep reinforcement learning.

Task and motion planning is studied for multi-robot collaboration problems in many works [23], [24], [25]. Most existing works that apply RL to bimanual cooperation rely on different levels of policy abstractions to reduce exploration space [4], [5] or learn from expert demonstrations to solve compositional tasks [6], [7]. In contrast, we do not assume pre-defined skills or pre-collected demonstrations and directly learn the low-level motion for each step. Many learning-based bimanual cooperation works focus on short-horizon tasks such as insertion [26], connection [27], and cloth folding [28], while we tackle a long-horizon problem with a relatively large number of objects. Zhang *et al.* [29] propose an RL method for dual-arm collaboration with a different focus on disentanglement to avoid conflict.

Our method leverages a symmetric structure between manipulators to improve RL training. The idea of exploiting structures in decision-making problems can be traced back to model minimization [30], [31], [32], which leverages homomorphism and symmetries to reduce redundancies in MDP. Many recent works try to incorporate structures into deep RL by encoding them into the architecture of neural networks [33], [34], [35] or discovering symmetries from data [36]. Our symmetry-aware actor-critic is conceptually similar to value decomposition methods [37], [38] in multi-agent RL, but they are motivated to address the issue of partial observability in decentralized multi-agent learning. We model our object rearrangement task as a sparse reward goal-conditioned RL problem [39], [40], [41]. Hindsight experience replay (HER) [8] is one of the most popular techniques for sparse-reward goal-conditioned problems. In the original HER, failed trajectories are relabeled into complete success using goals achieved later in the trajectory. Our object-centric relabel extend HER to multi-object sparse reward problems and converts failed data into partially successful data. Other relabeling methods are proposed to improve the efficiency of HER by model-based prediction [42], encouraging diversity of goals [43]. Generalized HER [44], [45] formulate hindsight relabeling with inverse RL and extend goal relabeling to arbitrary multi-task scenarios.

III. PRELIMINARY

We focus on bimanual handover and rearrangement of multiple objects. Two identical Franka Panda arms are mounted on two tables with a gap between them. There are multiple cuboids initialized on the tables, and the task is to control two arms to move each object to its desired goal position, which can be at either side of the gap.

We model the bimanual rearrangement task as a goal-conditioned Markov Decision Process defined by $(\mathcal{S}, \mathcal{A}, \mathcal{G}, P(s'|s,a), r(s,a,g), \rho_0, \gamma)$, where $\mathcal{S}, \mathcal{A}, \mathcal{G}$ represents state space, action space, and goal space respectively, $P(s'|s,a)$ denotes the probability for the environment to transit to state s' when taking action a at state s, $r(s,a,g)$ is the goal-conditioned reward function, ρ_0 is a distribution from which to sample the initial state and goal of each trajectory. Each state s_t is a concatenation of end effectors' states of two robots $s^r_{n,t}, n = 1, 2$ and m objects states

$s^{obj}_{n,t}, n = 1, 2, \cdots, m$. The RL agent jointly controls the two robots, i.e., $a = [a_1, a_2]$, by commanding their desired end effector displacement and finger widths. The desired orientations of the end effectors are kept fixed throughout the trajectories. The goal g specifies the desired 3-D positions for all the objects. The agent receives an object-level sparse reward after each step $r(s_t, a_t, g) = -\sum_{i=1}^{m} \mathbb{I}(||s^{obj}_{i,t} - g_i||_2 > d_\epsilon)/m$, where d_ϵ is a distance threshold.

We use soft actor-critic (SAC) [9] as the backbone RL algorithm. Denote the Q-network as Q_θ and the policy network as π_ϕ. The learning objective for Q-values is

$$L_Q(\theta) = \mathbb{E}_{(s,a,g,s',r)\sim\mathcal{D}}[(Q_\theta(s, a, g) - y(r, s', Q))^2], \quad (1)$$

where $y(r, s', Q) = r + \gamma(Q_{\theta'}(s', \tilde{a}') - \alpha \log \pi_\phi(\tilde{a}'|s')), \tilde{a}' \sim \pi_\phi(\cdot|s')$. $Q_{\theta'}$ is the target Q-network, and α is the entropy coefficient. The policy is optimized w.r.t.

$$L_\pi(\phi) = -\mathbb{E}_{(s,g)\sim\mathcal{D}}[Q_\theta(s, \tilde{a}_\phi, g) - \alpha \log \pi_\phi(\tilde{a}_\phi|s)]. \quad (2)$$

IV. METHOD

In this section, we introduce our symmetry-aware actor-critic framework for efficient bimanual rearrangement. We first describe how we incorporate a symmetric structure in bimanual tasks into actor-critic architectures (Sec. IV-A). Then, we present object-centric relabeling that extends hindsight experience replay [8] to the setting of multiple goals (Sec. IV-B). The overall algorithm and implementation details are summarized in Sec. IV-C.

A. Symmetry-aware Actor-Critic Architecture

Training an RL agent to control two cooperative arms is typically less sample efficient than controlling one arm due to increased task space and degrees of freedom. There exists symmetry in the bimanual task with two interchangeable manipulators that can reduce its intrinsic dimension.

Consider a pair of mirrored bimanual task instances where only the identity of two arms are swapped. Assuming that the roles of two manipulators are interchangeable, arm 1 in one instance should optimally behave the same as arm 2 in the mirrored instance since they are in the same situation, and the same relationship applies to the other two arms. Since the two instances have a direct mapping in their optimal policies, they must have equivalent critic values. Such mirrored pair can be generated by viewing one scene from two actor-centric frames. Formally, define a bijection M that maps task instances between two frames. Each state $s = [s^r_1, s^r_2, s^{obj}_{1:m}]$ in the frame of arm 1 is flipped into $s^M = [(s^r_2)^M, (s^r_1)^M, (s^{obj}_{1:m})^M]$ in the frame of arm 2. Each action $a = [a_1, a_2]$ in one frame becomes $a^M = [a^M_2, a^M_1]$ in the other frame. The optimal policy $\Pi^\star$ and value function evaluated at (s, g) and (s^M, g^M) must follow a one-on-one correspondence to each other: $\Pi^\star(s^M, g^M) = (\Pi^\star(s, g))^M$, $Q(s^M, \Pi^\star(s^M, g^M), g^M) = Q(s, \Pi^\star(s, g), g)$.

We incorporate the symmetric structure into the design of actor-critic architectures. Our symmetry-aware critic takes the average of Q values evaluated at pairs of instances,

$$\bar{Q}_\theta(s, a, g) = \frac{Q_\theta(s, a, g) + Q_\theta(s^M, a^M, g^M)}{2}. \quad (3)$$

Algorithm 1: Symmetry-aware Actor-Critic

Input: Q-network Q_θ, value target $Q_{\theta'}$ and policy network π_ϕ, bijection M, empty replay buffer $\mathcal{D}$, total number of objects m, symmetry-aware actor-critic $\bar{Q}_\theta(s, a, g)$ and $\Pi_\phi(s, g)$ (Eqn. 3, 4)
for $e = 0 :$ *num_epoch* **do**
  for $i = 0 :$ *num_traj* **do**
    Rollout a trajectory $\tau = (s_{0:T}, a_{0:T}, g, r_{0:T})$, where $s_0, g \sim \rho_0$, $a_t \sim \Pi_\phi(s_t, g)$
    $\mathcal{D} \leftarrow \mathcal{D} \cup \tau^i$
  for $i = 0 :$ *num_update* **do**
    for (s_t, a_t, g, r_t) ***in*** $\mathcal{D}$ **do**
      for $n = 0 : m$ **do**
        $g'_n \leftarrow s^{obj}_{n,l}, l > t$ **if** $s^{obj}_{n,t} \neq s^{obj}_{n,t+1}$ **else** $g'_n \leftarrow$ Uniform$(\mathcal{G})$
      Augment $\mathcal{D}$ with (s_t, a_t, g', r'), where $g' \leftarrow g'_{0:m}$, $r' \leftarrow R(s_t, a_t, g')$
    Optimize θ and ϕ w.r.t. $L_{\bar{Q}}(\theta)$ and $L_\Pi(\phi)$ (Eqn. 1, 2), soft update $Q_{\theta'}$

Symmetry-aware critic architecture encourages value representation to be invariant to the order/identity of the two manipulators. We can also embed the structure in the actor similar to the formulation of a shared policy in multi-agent RL. We train a single shared policy to control both arms. The action distribution is $\pi_\phi(s, g)$ for one arm, and $(\pi_\phi(s^M, g^M))^M$ for the other arm. The joint policy prediction of the symmetry-aware actor becomes

$$\Pi_\phi(s, g) = \pi_\phi(s, g) \parallel (\pi_\phi(s^M, g^M))^M, \quad (4)$$

which is, in practice, a concatenation of sufficient statistics of two single-arm action distributions.

B. Object-centric Relabeling

HER is a commonly adopted data augmentation technique to improve sample efficiency in sparse reward goal-conditioned problems. But it is not sufficient to scale up to scenarios with a large number of objects. For example, naively applying HER to multi-object tasks would relabel goals of *all* the objects as their later achieved states without considering the compositionality among different objects. Such relabel technique can lead to detrimental data bias towards trivial tasks with most of the objects already in place, especially in the early stage of training when the agent cannot manipulate all the objects. To better deal with the sparse reward issue in multi-object scenarios, we propose an object-centric goal-relabel technique leveraging the compositional structure in this problem. We propose to relabel goals per object: if an object is perturbed during the episode, we relabel its goal to its later achieved state; if an object stays still, we relabel its goal to a random position. In this way, we create partially successful but more diverse trajectories so that the agent can capture the multi-object structure better.

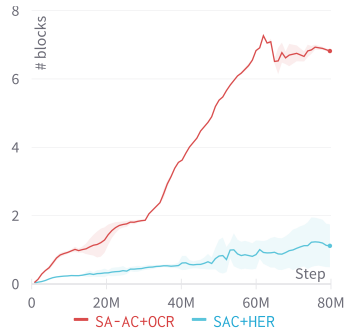


Fig. 3: The number of objects successfully rearranged by two robots over the last 1024 episodes vs. environment steps.

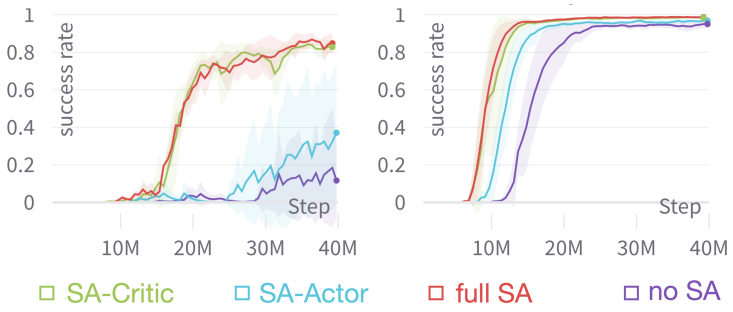


Fig. 4: Performances of the symmetry-aware architecture applied to different RL networks trained in settings with at most 2 objects. The success rate is evaluated in "fully-cooperative" (left) and "local" tasks (right).

TABLE I: Average episode lengths over 100 successful trajectories in different settings using our method and a phasic baseline. Our method is more efficient in settings with mixed local and cooperative sub-tasks.

# handovers / # blocks	0/8	1/8	2/8	3/8	4/8	5/8	6/8	7/8	8/8
Ours	50.76	**74.76**	**93.11**	**110.96**	**136.35**	**150.80**	**163.57**	**187.11**	**188.5**
Phasic (first local, then handover)	49.57	122.76	151.68	161.20	202.60	210.34	211.30	211.38	193.60

C. Symmetry-aware SAC for Bimanual Rearrangement

Combining the symmetry-aware actor-critic architecture with a popular backbone SAC, and further enhancing the training data with object-centric relabeling, we get the algorithm for efficient bimanual handover and rearrangement. The pseudocode is summarized in Alg. 1.

We adopt a Transformer-based [46] network architecture to extract object-centric features with a stack of self-attention layers. For the actor network, we aggregate features with max-pooling to extract local information for decision making; for the critic network, we fuse features with mean-pooling to ensure a consistent value range over different numbers of objects. To learn the key behavior "handover" for object rearrangement more efficiently, we progressively enlarge the gap between the tables and increase the probability to sample goals on the opposite side as the scurriculum for the agent. The table gap is set to 10cm (half of the cuboid length) in the beginning to help robots discover cooperative operations and gradually extend to 30cm. The probability of sampling goals in the other table is set to 0.2 in the beginning and gradually increases to 0.8. We also adopt another training curriculum in terms of object amount that starts from single-object tasks to tasks with more objects.

D. Real Robot Deployment

We build up a physical experiment platform including two Franka Panda arms, two RealSense D455 cameras, and cuboid blocks with size 4cm×4cm×20cm. We use Aruco [47] markers attached to the blocks to track their poses. To obtain physically feasible and stable strategies that can easily transfer to real robots, we fine-tune the trained policy in a simulated environment with more constraints. For safety reasons, we bound the moving range of the end effector and penalize the agent if a large contact force is experienced by the robot fingers in the z-axis. In human-robot cooperation, we mask the opposite robot state during training, since the robot can only observe its own state and all object states during deployment.

V. EXPERIMENTS

In this section, we present experiment results to address the following questions: 1) Can the symmetry-aware actor-critic framework effectively solve the bimanual handover and rearrangement tasks with a high success rate and scale up to a large number of objects? 2) How does each component in our framework contribute to the overall performance? 3) Can learned strategies be deployed to the real world? Our simulated environment is built with Isaac gym [10]. All the experiments are run over 3 seeds with a 3080Ti GPU.

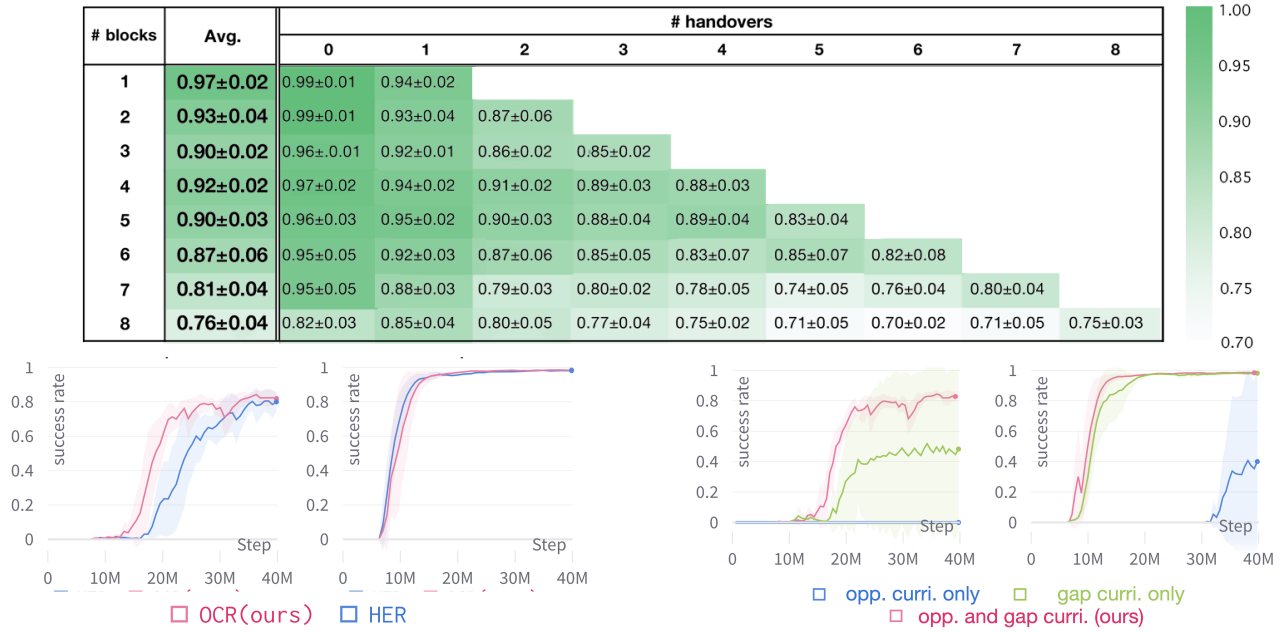
A. Main Results

We first report the performances for rearranging different numbers of objects. The positions and orientations of two robot bases are randomly initialized in each episode. Our framework uses symmetry-aware actor-critic networks, object-centric relabel, and the proposed curriculum throughout training. We compare against standard SAC+HER baseline trained with the same curriculum. The average number of successful rearranged blocks throughout the training process is illustrated in Fig. 3. Our symmetry-aware framework (red) is significantly more efficient than the SAC+HER baseline (blue). Moreover, the number of rearranged objects of our framework scales up almost linearly w.r.t. environment steps after it masters how to rearrange two objects, indicating its good compositional generalization ability over exponentially growing task configurations. We set the maximum number of objects to 8 due to the limited workspace on the table.

To show the efficiency of the learned policy, we also implement a phasic baseline by properly setting the attention mask in our learned policy so that the agent will not perform handovers until all the local sub-tasks are finished. The comparisons are shown in Table I. Our method discovers non-trivial rearrangement strategies with higher efficiency.

TABLE II: Detailed success rate of multi-object handover and rearrangement problem. The second column shows average success rates in settings with a different total number of objects. Other columns show success rates in all sub-tasks, further categorized by the required times of bimanual handover.

# blocks	Avg.	# handovers								
		0	1	2	3	4	5	6	7	8
1	**0.97±0.02**	0.99±0.01	0.94±0.02							
2	**0.93±0.04**	0.99±0.01	0.93±0.04	0.87±0.06						
3	**0.90±0.02**	0.96±.0.01	0.92±0.01	0.86±0.02	0.85±0.02					
4	**0.92±0.02**	0.97±0.02	0.94±0.02	0.91±0.02	0.89±0.03	0.88±0.03				
5	**0.90±0.03**	0.96±0.03	0.95±0.02	0.90±0.03	0.88±0.04	0.89±0.04	0.83±0.04			
6	**0.87±0.06**	0.95±0.05	0.92±0.03	0.87±0.06	0.85±0.05	0.83±0.07	0.85±0.07	0.82±0.08		
7	**0.81±0.04**	0.95±0.05	0.88±0.03	0.79±0.03	0.80±0.02	0.78±0.05	0.74±0.05	0.76±0.04	0.80±0.04	
8	**0.76±0.04**	0.82±0.03	0.85±0.04	0.80±0.05	0.77±0.04	0.75±0.02	0.71±0.05	0.70±0.02	0.71±0.05	0.75±0.03



(a) Effectiveness of object-centric relabeling.

(b) Ablation studies on adaptive curricula.

Fig. 5: The left figure compares object-centric relabeling and hindsight experience replay. In the right figure, we ablate the curriculum on the probability of sampling goals on the opposite side and the curriculum on the table gap. In each subplot, the learning curves on fully cooperative (left) and local (right) scenarios with at most 2 objects are reported.

Since each goal can be sampled from either side, the settings with the same number of objects still contain sub-tasks with various levels of difficulty. To better analyze the agent's performance, we categorize the whole task space according to both the total number of objects and the number of opposite goals and report the detailed success rates in Table II. Our framework achieves a success rate of over 0.7 in every sub-task that requires different times of handover.

B. Ablation Studies

For a fair comparison, all the ablations are conducted over settings with at most two objects. Due to the multi-task nature of our setting, we present success rates in two specific tasks: "local", where all goals are in the same workspace of the objects; "fully cooperative" where all goals are sampled from the opposite side and requires the most cooperation.

Symmetry-aware architecture: We apply the symmetry-aware architecture to the following settings: (1) both actor and critic (red), (2) critic only (green), (3) actor only (blue), and compare their results with SAC that does not incorporate any symmetry (purple). The learning curves are reported in Fig. 4. SAC without symmetric representation performs the worst; it only solves local tasks while struggling in cooperative tasks. SAC with symmetric actor performs slightly better, while symmetric critic can significantly improve the sample efficiency and final success rate, especially when evaluated in fully cooperative tasks. Combining symmetric actor and critic performs similarly well as symmetric critic only. In our main result, we adopt symmetric representation for both networks since they can both accelerate training.

Object-centric relabeling: We compare object-centric relabeling with the "future" strategy in HER [8], a popular relabeling method that replaces goals of *all* objects with their future achieved states. As shown in Fig. 5a, object-centric relabeling performs on par with HER in local tasks, while outperforms HER with a clear margin in cooperative tasks.

Curriculum learning: We compare with variants that remove different adaptive curricula in Fig. 5b. The green curve is replacing the adaptive ratio of opposite goals with a fixed ratio of 0.8, and the blue curve is using a fixed distance of 30cm between tables. It is unstable to learn cooperation without the adaptive opposite side ratio, as indicated by the large variance. Without the curriculum on the table gap, the agent completely fails to discover cooperative behaviors, and even cannot learn local tasks efficiently.

C. Learned Strategies and Failure Cases

We then visualize how our agent completes an 8-object rearrangement to both sides of the table gap. As shown in Fig. 6, the two arms first complete the local rearrangement tasks. One arm then passes some object to the opposite arm and waits for its partner to put the object in place or move to pick up the next object to start another handover.

Two typical failure cases are shown in Fig. 7. When the workspace is cluttered with a large number of objects, the agent may accidentally knock a block off the table when

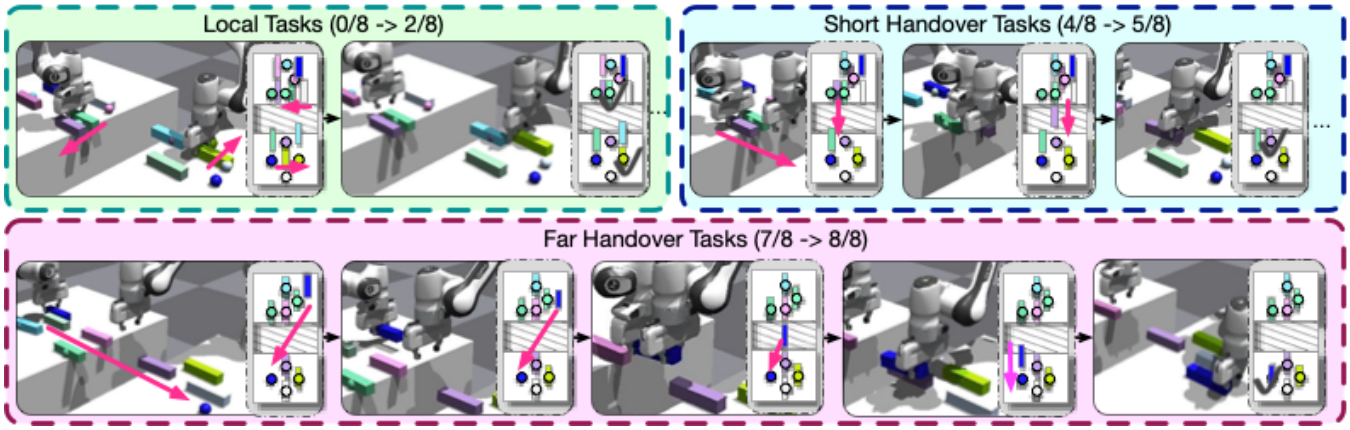


Fig. 6: Visualization of learned strategy. Two robots first conduct local tasks and prepare for the handover. They then turned to objects at the edge of the desktop that could be moved with a few steps across the table. During the cooperative handover process, two arms learn the strategy to minimize the total working time to move objects while moving back.

(a) Accidentally knock an object off the table.

(b) Fail to grasp due to collision with a nearby object.

Fig. 7: Visualization of two failure cases.

Fig. 8: Rearrange 8 objects using real panda arms.

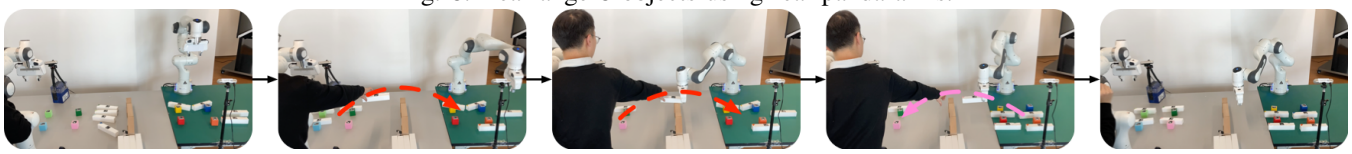


Fig. 9: Human-robot cooperative rearrangement. The arm needs to complete one local rearrangement by itself and one cooperative handover with the human.

moving other blocks. The agent may also fail to grasp an object due to collision with other blocks.

D. Deployment on Real Robots

From the results of the 21 consecutive experiments carried out on the real machine using randomly generated object locations, 14 were successful in completing the transfer of objects. On average, each task was able to transfer 7 objects with a success rate of 66.7% of rearranging all 8 objects. We showcase a successful deployment of the learned policy on two Panda arms. In Fig. 8, the arms aim to exchange the two objects on blue and yellow platforms. They each pick up one object within their reach in the beginning, then start moving the objects towards the opposite platforms. The left arm decides to put down its grasped object since the right arm has already reached out and is about to pass over the object. The left arm then takes over the object from the right arm in time, and picks up the temporarily dropped object to perform another handover.

We also extend our method to accomplish human-robot cooperation. The deployed policy is shown in Fig. 9, in which the Panda arm first accomplishes a local task by pushing the block on the table to its goal, then takes another block from a human and transports it to the goal spot.

VI. CONCLUSION

We tackle a bimanual multi-object handover and rearrangement task with a symmetry-aware deep reinforcement learning framework. We embed interchangeable roles of two manipulators into the design of actor-critic networks, which significantly improves the sample efficiency of RL. Combined with object-centric relabeling and adaptive curricula, the whole framework solves 8-object rearrangement tasks efficiently. It is interesting to extend this framework to more complex bimanual tasks such as assembly in the future, in which object rearrangement is an important ingredient.

REFERENCES

[1] D. Rakita, B. Mutlu, M. Gleicher, and L. M. Hiatt, "Shared control–based bimanual robot manipulation," *Science Robotics*, vol. 4, no. 30, p. eaaw0955, 2019.

[2] A. Edsinger and C. C. Kemp, "Two arms are better than one: A behavior based control system for assistive bimanual manipulation," in *Recent progress in robotics: Viable robotic service to human*. Springer, 2007, pp. 345–355.

[3] C. Smith, Y. Karayiannidis, L. Nalpantidis, X. Gratal, P. Qi, D. V. Dimarogonas, and D. Kragic, "Dual arm manipulation—a survey," *Robotics and Autonomous systems*, vol. 60, no. 10, pp. 1340–1353, 2012.

[4] R. Chitnis, S. Tulsiani, S. Gupta, and A. Gupta, "Efficient bimanual manipulation using learned task schemas," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 1149–1155.

[5] A. Colomé and C. Torras, *Reinforcement Learning of Bimanual Robot Skills*. Springer, 2020.

[6] F. Xie, A. Chowdhury, M. C. D. P. Kaluza, L. Zhao, L. L. S. Wong, and R. Yu, "Deep imitation learning for bimanual robotic manipulation," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020.

[7] A. Tung, J. Wong, A. Mandlekar, R. Martín-Martín, Y. Zhu, L. Fei-Fei, and S. Savarese, "Learning multi-arm manipulation through collaborative teleoperation," in *IEEE International Conference on Robotics and Automation, ICRA 2021, Xi'an, China, May 30 - June 5, 2021*. IEEE, 2021, pp. 9212–9219. [Online]. Available: https://doi.org/10.1109/ICRA48506.2021.9561491

[8] M. Andrychowicz, F. Wolski, A. Ray, J. Schneider, R. Fong, P. Welinder, B. McGrew, J. Tobin, O. Pieter Abbeel, and W. Zaremba, "Hindsight experience replay," *Advances in neural information processing systems*, vol. 30, 2017.

[9] 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*. PMLR, 2018, pp. 1861–1870.

[10] V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa *et al.*, "Isaac gym: High performance gpu-based physics simulation for robot learning," *arXiv preprint arXiv:2108.10470*, 2021.

[11] D. Batra, A. X. Chang, S. Chernova, A. J. Davison, J. Deng, V. Koltun, S. Levine, J. Malik, I. Mordatch, R. Mottaghi *et al.*, "Rearrangement: A challenge for embodied ai," *arXiv preprint arXiv:2011.01975*, 2020.

[12] A. Krontiris, R. Shome, A. Dobson, A. Kimmel, and K. Bekris, "Rearranging similar objects with a manipulator using pebble graphs," in *2014 IEEE-RAS International Conference on Humanoid Robots*. IEEE, 2014, pp. 1081–1087.

[13] J. E. King, M. Cognetti, and S. S. Srinivasa, "Rearrangement planning using object-centric and robot-centric action spaces," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 3940–3947.

[14] J. Lee, Y. Cho, C. Nam, J. Park, and C. Kim, "Efficient obstacle rearrangement for object manipulation tasks in cluttered environments," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 183–189.

[15] O. Ben-Shahar and E. Rivlin, "Practical pushing planning for rearrangement tasks," *IEEE Transactions on Robotics and Automation*, vol. 14, no. 4, pp. 549–565, 1998.

[16] A. Cosgun, T. Hermans, V. Emeli, and M. Stilman, "Push planning for object placement on cluttered table surfaces," in *2011 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 2011, pp. 4627–4632.

[17] E. Huang, Z. Jia, and M. T. Mason, "Large-scale multi-object rearrangement," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 211–218.

[18] M. Danielczuk, A. Mousavian, C. Eppner, and D. Fox, "Object rearrangement using learned implicit collision functions," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 6010–6017.

[19] F. Ebert, C. Finn, S. Dasari, A. Xie, A. Lee, and S. Levine, "Visual foresight: Model-based deep reinforcement learning for vision-based robotic control," *arXiv preprint arXiv:1812.00568*, 2018.

[20] W. Yuan, J. A. Stork, D. Kragic, M. Y. Wang, and K. Hang, "Rearrangement with nonprehensile manipulation using deep reinforcement learning," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 270–277.

[21] A. Zeng, P. Florence, J. Tompson, S. Welker, J. Chien, M. Attarian, T. Armstrong, I. Krasin, D. Duong, V. Sindhwani *et al.*, "Transporter networks: Rearranging the visual world for robotic manipulation," in *Conference on Robot Learning*. PMLR, 2021, pp. 726–747.

[22] O. OpenAI, M. Plappert, R. Sampedro, T. Xu, I. Akkaya, V. Kosaraju, P. Welinder, R. D'Sa, A. Petron, H. P. d. O. Pinto *et al.*, "Asymmetric self-play for automatic goal discovery in robotic manipulation," *arXiv preprint arXiv:2101.04882*, 2021.

[23] J. Motes, R. Sandström, H. Lee, S. Thomas, and N. M. Amato, "Multi-robot task and motion planning with subtask dependencies," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3338–3345, 2020.

[24] R. Shome and K. E. Bekris, "Synchronized multi-arm rearrangement guided by mode graphs with capacity constraints," in *International Workshop on the Algorithmic Foundations of Robotics*. Springer, 2020, pp. 243–260.

[25] V. N. Hartmann, A. Orthey, D. Driess, O. S. Oguz, and M. Toussaint, "Long-horizon multi-robot rearrangement planning for construction assembly," *arXiv preprint arXiv:2106.02489*, 2021.

[26] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1334–1373, 2016.

[27] S. Kataoka, S. K. S. Ghasemipour, D. Freeman, and I. Mordatch, "Bi-manual manipulation and attachment via sim-to-real reinforcement learning," *arXiv preprint arXiv:2203.08277*, 2022.

[28] Y. Wu, W. Yan, T. Kurutach, L. Pinto, and P. Abbeel, "Learning to manipulate deformable objects without demonstrations," in *Robotics: Science and Systems XVI, Virtual Event / Corvalis, Oregon, USA, July 12-16, 2020*, M. Toussaint, A. Bicchi, and T. Hermans, Eds., 2020. [Online]. Available: https://doi.org/10.15607/RSS.2020.XVI.065

[29] M. Zhang, P. Jian, Y. Wu, H. Xu, and X. Wang, "Dair: Disentangled attention intrinsic regularization for safe and efficient bimanual manipulation," *arXiv preprint arXiv:2106.05907*, 2021.

[30] T. Dean and R. Givan, "Model minimization in markov decision processes," in *AAAI/IAAI*, 1997, pp. 106–111.

[31] B. Ravindran and A. G. Barto, "Symmetries and model minimization in markov decision processes," 2001.

[32] B. Ravindran, *An algebraic approach to abstraction in reinforcement learning*. University of Massachusetts Amherst, 2004.

[33] A. Tamar, Y. Wu, G. Thomas, S. Levine, and P. Abbeel, "Value iteration networks," *Advances in neural information processing systems*, vol. 29, 2016.

[34] E. van der Pol, D. Worrall, H. van Hoof, F. Oliehoek, and M. Welling, "Mdp homomorphic networks: Group symmetries in reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 4199–4210, 2020.

[35] L. Zhao, L. Kong, R. Walters, and L. L. Wong, "Toward compositional generalization in object-oriented world modeling," *arXiv preprint arXiv:2204.13661*, 2022.

[36] A. Mahajan and T. Tulabandhula, "Symmetry learning for function approximation in reinforcement learning," *arXiv preprint arXiv:1706.02999*, 2017.

[37] P. Sunehag, G. Lever, A. Gruslys, W. M. Czarnecki, V. F. Zambaldi, M. Jaderberg, M. Lanctot, N. Sonnerat, J. Z. Leibo, K. Tuyls, and T. Graepel, "Value-decomposition networks for cooperative multi-agent learning based on team reward," in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2018, Stockholm, Sweden, July 10-15, 2018*, E. André, S. Koenig, M. Dastani, and G. Sukthankar, Eds. International Foundation for Autonomous Agents and Multiagent Systems Richland, SC, USA / ACM, 2018, pp. 2085–2087. [Online]. Available: http://dl.acm.org/citation.cfm?id=3238080

[38] T. Rashid, M. Samvelyan, C. S. de Witt, G. Farquhar, J. N. Foerster, and S. Whiteson, "QMIX: monotonic value function factorisation for deep multi-agent reinforcement learning," in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, ser. Proceedings of Machine Learning Research, J. G. Dy and A. Krause, Eds., vol. 80. PMLR, 2018, pp. 4292–4301. [Online]. Available: http://proceedings.mlr.press/v80/rashid18a.html

[39] L. P. Kaelbling, "Learning to achieve goals," in *Proceedings of the 13th International Joint Conference on Artificial Intelligence. Chambéry,*

France, August 28 - September 3, 1993, R. Bajcsy, Ed. Morgan Kaufmann, 1993, pp. 1094–1099.

[40] T. Schaul, D. Horgan, K. Gregor, and D. Silver, "Universal value function approximators," in *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, ser. JMLR Workshop and Conference Proceedings, F. R. Bach and D. M. Blei, Eds., vol. 37. JMLR.org, 2015, pp. 1312–1320. [Online]. Available: http://proceedings.mlr.press/v37/schaul15.html

[41] V. Pong, S. Gu, M. Dalal, and S. Levine, "Temporal difference models: Model-free deep RL for model-based control," in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. [Online]. Available: https://openreview.net/forum?id=Skw0n-W0Z

[42] R. Yang, M. Fang, L. Han, Y. Du, F. Luo, and X. Li, "Mher: Model-based hindsight experience replay," in *Deep RL Workshop NeurIPS 2021*, 2021.

[43] M. Fang, T. Zhou, Y. Du, L. Han, and Z. Zhang, "Curriculum-guided hindsight experience replay," *Advances in neural information processing systems*, vol. 32, 2019.

[44] B. Eysenbach, X. Geng, S. Levine, and R. R. Salakhutdinov, "Rewriting history with inverse rl: Hindsight inference for policy improvement," *Advances in neural information processing systems*, vol. 33, pp. 14 783–14 795, 2020.

[45] A. Li, L. Pinto, and P. Abbeel, "Generalized hindsight for reinforcement learning," *Advances in neural information processing systems*, vol. 33, pp. 7754–7767, 2020.

[46] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[47] S. Garrido-Jurado, R. Muñoz-Salinas, F. J. Madrid-Cuevas, and M. J. Marín-Jiménez, "Automatic generation and detection of highly reliable fiducial markers under occlusion," *Pattern Recognition*, vol. 47, no. 6, pp. 2280–2292, 2014.