

A Appendix

A.1 Environment Description

A.1.1 Cooperative Transport

The agents should cooperate to push a cylinder to a target location. The local positions between the agents, cylinder, and target are initialized randomly. The cylinder is set in the middle of the map, and agents are initialized in the four quadrants of the cylinder respectively.

Observation The agents receive two sets of observations: (i) a set of "proprioceptive" states which would be given to the lower layer controller. This part of observation is exactly the same in all agents in different tasks because all tasks use the same pre-trained lower layer policy, including their self-observed information, such as its local position, local velocity, angular velocity, orientation, body upright angle, joint force and its limitations, etc. (ii) a set of "exteroceptive" features containing task-relevant information which would be given to upper layer controller. In this task, the observation includes the locations of the nearest agent, the current position of the cylinder, and the destination.

Rewards We set three types of rewards: (i) Cylinder distance reward: a term which is the reciprocal L2 distance of the agent to the target. This is the main reward of this task. It will increase when the cylinder becomes closer to the target. The increasing trend is the same as the inverse proportional function. (ii) Agent cylinder distance reward: this term is to encourage agents firstly come closer to the cylinder so that they can learn how to push the object. (iii) Agent distance reward: a term used to guide the agent to walk towards the target, thereby obtaining a large cylinder distance reward.

A.1.2 Corridor Crossing

The agents are expected to reach a designated line through a narrow slit. Two walls block most of the way forward, leaving only a narrow corridor that only allows one agent to pass at a time. Yet all agents are at the same distance to the entrance of the slit, which means that agents would get stuck at the entrance of the slit together if nobody is slow or stop to wait for others to pass first. We expected agents can acquire a strategy so that all the agents could go through the corridor as fast as possible.

Observation The lower layer controllers' observation is the same as the last task. In this task, agents' upper layer controllers' observation includes the locations of itself and the nearest neighbour, the position, translation, and orientation of the walls.

Rewards Two rewards are set: (i) Destination reward: if an agent has passed the finish line, all agents will receive a positive reward. It stimulates the strategy allowing as many agents as possible to pass through the corridor to the finish line. (ii) Distance reward to enter: a term to encourage agents first to come close to the corridor.

A.1.3 Ravine Bridging

The agents are expected to reach a designated line where they must cross over a ravine first. There is a ravine blocking the way of the agents. Agents cannot jump over or climb out of the ravine. Agents need colleagues' help to come over the ravine. At this point, we place a stone of the right size and depth in the ravine to serve as a "bridge" to help the agents cross it. We pre-set two agents on either side of the stone to push the stone to help their colleagues cross the ravine and reach the destination. So this is a complex task that requires teamwork among heterogeneous agents.

Observation The observation of the lower layer policy is similar to the previous tasks. Agents' upper layer controllers' observation includes the locations of itself and the nearest agents, the position of the stone, and the position and depth of the ravine.

Rewards Two rewards are set: (i) Destination reward: if any agent has passed the finish line, all agents will receive a positive reward. It stimulates the strategy allowing as many agents as possible to go through the ravine and get to the finish line. (ii) Bridging reward: a term to encourage agents in the ravine to carry the stone to the agent above which is the closest to the ravine.

A.2 Curriculum

We utilize sparse rewards to direct agents towards proximity with the object, subsequently acquiring the target’s location in the Cooperative Transport task. In the Corridor Crossing task, we train individual agents to navigate to the finish line by passing through the slit and then refine the pre-trained policy for application in a multi-agent environment. The learning sequence for the Ravine Bridging task is more intricate which means we have to give more guidance; we instruct agents to wait at the ravine’s edge at the opportune moment to prevent falling, while other agents within the ravine initially learn to convey ”bridge” to an appropriate location.

A.3 Scalability

The distributed HRL framework allows the collaborative tasks to be explicitly represented as a ”group”, thereby enabling tasks to be extended to large numbers of agents team. We have tested the framework with different populations of Ant in various tasks (There are only two Ants for pushing bridge in the task Ravine Bridging), and the success rates of each task with different populations are shown in the Tab. 2. The results show the potential of our framework for handling tasks with more agents.

Table 2: Each percentage results from 50 trials—10 attempts per three independently trained models.

Success rate	Two Ants	Three Ants	Four Ants	Five Ants
Cooperative Transport	$8 \pm 2\%$	$42 \pm 3\%$	$74 \pm 7\%$	$82 \pm 5\%$
Corridor Crossing	$98 \pm 2\%$	$95 \pm 2\%$	$93 \pm 3\%$	$79 \pm 2\%$
Ravine Bridging	$48 \pm 5\%$	$36 \pm 8\%$	$30 \pm 2\%$	$12 \pm 5\%$

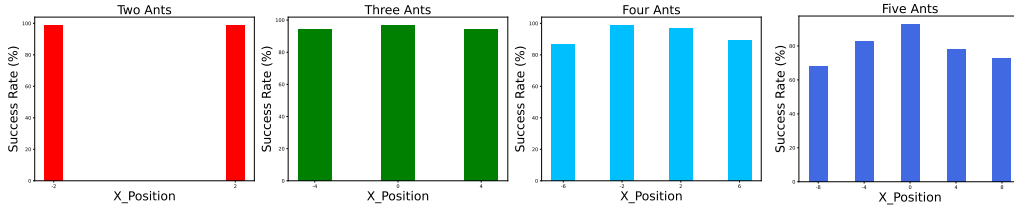
Table 3: Success Rate Comparison between average moving distances in task Ravine Bridging.

Distance from Ravine	Two Ants	Three Ants	Four Ants	Five Ants
2 m	49%	45%	38%	35%
4 m	38%	33%	20%	11%
6 m	30%	19%	5%	3%

A.4 Typical Result

A.4.1 Corridor Crossing

We record the success rates of each Ant at its initial position separately, as illustrated in Fig. A.4.1. We have a higher success rate when the initial position is closer to the center, and a higher failure rate when the initial position is closer to the edge. This is also explainable, as the Ant at the centre is closer to the slit and does not need to adjust velocity.



A.4.2 Ravine Bridging

The bridging and coordination in the ravine is the key to completing the task. We set up three difficulty levels - easy, medium, and hard - based on the distance from the ravine’s edge. The easy level corresponds to a distance of 2 m, the medium level to 4 m, and the hard level to 6 cm. The Tab. 3 shows the success rates of the three difficulty levels with different numbers of Ants. We can

480 observe that it becomes more difficult for any number of Ants to complete the task as the distance
481 increases. Because cooperation on the timing of stopping and bridging becomes more important
482 when moving length increases.