

Spatial Intention Maps for Multi-Agent Mobile Manipulation

Jimmy Wu^{1,2}, Xingyuan Sun¹, Andy Zeng², Shuran Song³, Szymon Rusinkiewicz¹, Thomas Funkhouser^{1,2}

Abstract—The ability to communicate intention enables decentralized multi-agent robots to collaborate while performing physical tasks. In this work, we present spatial intention maps, a new intention representation for multi-agent vision-based deep reinforcement learning that improves coordination between decentralized mobile manipulators. In this representation, each agent’s intention is provided to other agents, and rendered into an overhead 2D map aligned with visual observations. This synergizes with the recently proposed spatial action maps framework, in which state and action representations are spatially aligned, providing inductive biases that encourage emergent cooperative behaviors requiring spatial coordination, such as passing objects to each other or avoiding collisions. Experiments across a variety of multi-agent environments, including heterogeneous robot teams with different abilities (lifting, pushing, or throwing), show that incorporating spatial intention maps improves performance for different mobile manipulation tasks while significantly enhancing cooperative behaviors.

I. INTRODUCTION

Multi-agent systems require a mutual understanding of intention in order to coordinate amongst each other and perform collaborative tasks. This is particularly important for decentralized robots that share the same physical space while interacting in complex environments under partial observability and limited-bandwidth communication. Applications of such systems include foraging [1], hazardous waste cleanup [2], object transportation [3], [4], and search and rescue [5]. The capacity to coordinate by conveying intentions can reduce collisions between robots (which can be both costly and disabling) while improving the efficiency with which tasks are completed.

Intention-aware multi-agent learning methods have shown promising success—from hide and seek [6] to Starcraft [7]—albeit largely in simulated domains such as video games. These systems often communicate intentions via sharing high-level state information, or by condensing intentions into a low-dimensional embedding such as the coordinates of a destination. Both of these approaches, however, exclude spatial structure and are thus a poor fit for reinforcement learning (RL) with convolutional neural networks, which has become a dominant strategy for learning with visual input.

In this work we propose *spatial intention maps* (Fig. 1), a new intention representation for multi-agent deep RL. In our approach, each agent’s *intention* (i.e., its most recently selected action) is rendered into a 2D map (first inset in Fig. 1) aligned with a state map derived from visual observations. The intention and state maps are input into a fully convolutional network to predict Q-values for a dense

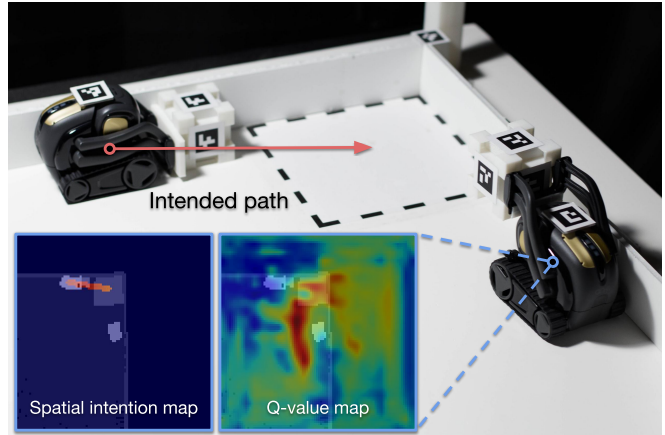


Fig. 1. Spatial intention maps allow agents to choose actions with knowledge of the actions being performed by other agents. In the figure, the robot on the left is moving towards the upper right corner. The robot on the right sees this in the spatial intention map, and instead of moving towards the goal straight ahead, it chooses to move to the left (dark red in the Q-value map). This avoids a potential collision with the other robot.

set of spatial actions (second inset in Fig. 1). At each step, the best action (e.g., navigate to a location) is selected for execution. Since spatial intention maps are encoded as images pixelwise aligned with state and action maps [8], we expect that fully convolutional networks can leverage them more efficiently than other intention encodings. Indeed, we find during experiments that encoding intentions spatially in this way provides a signal that encourages Q-functions to learn emergent collaborative behaviors, ranging from collision avoidance to spatial coordination and task specialization.

We demonstrate the efficacy of spatial intention maps across a variety of multi-agent environments, including robots with different abilities (lifting, pushing, or throwing) that are trained on different tasks (*foraging* and *search and rescue*). We observe that using spatial intention maps leads to better performance and significant gains in collaborative behavior, including fewer collisions, coordination around bottlenecks such as doorways, and improved distribution of agents throughout an environment. We also show that our learned policies (trained in simulation) are able to generalize to a real world setting. For qualitative video results and code, please see the supplementary material at <https://spatial-intention-maps.cs.princeton.edu>.

II. RELATED WORK

Multi-robot systems. The earliest works in multi-robot systems date back to the 1980s [9]. These works have studied a wide variety of challenges in multi-robot systems, such as architecture, communication, team heterogeneity, and learning. Typical applications areas include foraging [1],

¹Princeton University, ²Google, ³Columbia University

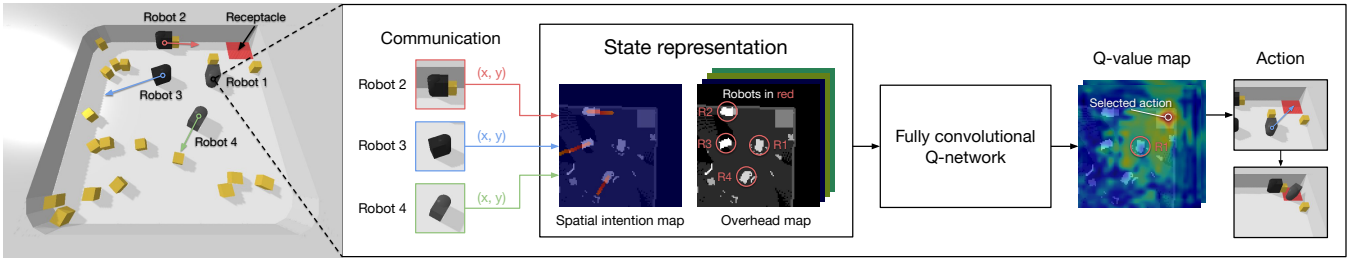


Fig. 2. **Overview.** Our system uses decentralized, asynchronous execution with communicating agents. In the figure, robot 1 is currently choosing a new action. It receives the intentions (represented as path waypoints) of the other agents, which are then encoded into a spatial map and used as input to the fully convolutional DQN policy network.

hazardous waste cleanup [2], object transportation [3], [4], search and rescue [5], and soccer [10]. Many of these early systems used reactive or behavior-based approaches, which require hand-crafting policies for the agents. In contrast to these works, we use reinforcement learning to enable our robots to automatically learn behaviors through experience.

Multi-agent reinforcement learning. There is a wide body of work in multi-agent reinforcement learning (MARL) [11], [12], [13]. One of the earliest works in this area was [14], in which multiple independent Q-learning agents were trained in a cooperative setting. More recently, with the rising popularity of deep reinforcement learning approaches such as DQN [15], many works have proposed ways to extend these methods to multi-agent settings [16]. These include addressing nonstationarity through modifications to experience replay [17], [18] or training with a centralized critic [19], [20], or improving credit assignment through decomposing the value function [21], [22], [23]. Most of these approaches assume access to the full state of the environment, though a few use raw visual data [24], [25]. In our approach, we learn directly from visual data reconstructed from partial observations of the environment.

Learning-based multi-robot systems. While there is a vast amount of literature on multi-agent reinforcement learning, most works focus on games and simulation environments. There is comparatively less work applying multi-agent learning to robotic systems. Some earlier works have applied Q-learning to applications such as box pushing [26], foraging [27], soccer [28], [29], and multi-target observation [30]. More recently, [31] and [32] have used macro-actions [33] with DQN to enable multiple robots to asynchronously reason and execute actions. Other works have investigated multi-agent tasks such as navigation [34] or cooperative manipulation [35]. As with the MARL works previously described, these typically assume access to high-level state information such as the positions of relevant objects. In contrast, we learn directly from visual data, allowing agents to automatically learn to detect relevant visual features.

Multi-robot communication. Many works in multi-robot systems have studied the effect of communication on improving team performance [36], [37], [38]. These comparisons typically span a continuum from no communication (implicit), to passive observation of teammates’ states, to direct communication (explicit). The same continuum of commu-

nication has been studied for multi-agent learning [14], [39], [40]. More recent works in multi-agent learning have studied other ways to communicate, such as learning what to communicate [41], [42], [43] or learning to model the intentions of other agents [44], [45], [46], [47], [34]. We similarly explore this communication continuum, but in our approach, communicated intentions are spatially encoded and aligned with the state and action representations. This allows agents to reason about their teammates’ intentions in the same domain as their own observations and actions.

III. METHOD

We study spatial intention maps in the context of two tasks. The first is a *foraging* task, in which a team of robots work together to move all objects in an environment to a target receptacle in the upper right corner (see Fig. 2 for an example environment). The second is a *search and rescue* task, in which a team of robots need to find and “rescue” all objects in an environment. We train our policies in a PyBullet [48] simulation environment and run them on real robots using a sim-to-real mirroring setup (see Sec. IV-D for more details).

A. Reinforcement Learning Formulation

We model our task as a Markov decision process (MDP) from the perspective of each individual agent, using single-agent state and action spaces. We train our policies using double deep Q-learning (DQN) [15], [49], giving individual rewards to agents during training. This formulation of multi-agent RL is similar to independent Q-learning [14], but we share policies between agents of the same type during training. Execution is decentralized—trained policies are run independently and asynchronously on each agent.

As widely discussed in the multi-agent literature, a common challenge with training multiple agents concurrently in the same environment is nonstationarity caused by other agents having policies that are constantly evolving. In our system, we use spatial intention maps (see Fig. 2) to directly encode the intentions of other agents into the state representation given to DQN, which helps to alleviate the uncertainty in other agents’ behaviors. Please see the supplementary material for more training details.

B. State Representation

Our state representation consists of a local overhead map, along with a series of auxiliary local maps with additional information useful to the agent when making its decisions [8].

To construct these local maps, each agent independently builds up its own global map of the environment using online mapping. Each agent has a simulated forward-facing RGB-D camera and captures partial observations of the environment, which are integrated with previous views to gradually build up a global overhead map. Agents must therefore learn policies that can deal with outdated information and seek out unexplored parts of the environment.

Each time an agent chooses a new action, it crops out a local map from its global map to generate a new state representation. These local maps are oriented such that the agent itself is in the center and facing up (see Fig. 2 for an example). The state representation consists of the following local maps, each of which is represented as an overhead image: (i) an environment map, (ii) an agent map encoding the agent’s own state and the observed states of other agents, (iii-iv) maps encoding shortest path distances to the receptacle and from the agent, and (v) the spatial intention map, which we describe in Sec. III-D. The agent map spatially encodes the state of every agent, which includes each agent’s pose, as well as whether they are carrying an object (if applicable).

C. Action Representation

We use spatial action maps [8] as our action representation. The action space is represented as a pixel map, spatially aligned with the state representation, in which every pixel represents the action that navigates the agent to the corresponding location in the environment (see Q-value map in Fig. 2). For agents that can execute an end effector action (such as lifting or throwing), we augment the action space with a second spatial channel, which represents navigating to the corresponding location and then attempting to perform the end effector action. (Note that figures in this paper only show one relevant action channel for clarity.) We execute the action corresponding to the argmax across all channels in the action space, employing high-level motion primitives implemented with low-level control. For movement, the primitive attempts to proceed to the specified location along the shortest path, computed using the agent’s own occupancy map. For end effector actions, the primitives attempt to lock onto an object before operating the end effector.

D. Spatial Intention Maps

Spatial intention maps provide an encoding of other agents’ intentions in the form of a map aligned with the state and action representations. This map-based encoding of intentions is key to our approach, as it allows us to train a fully convolutional deep residual Q-network [50], [51] that maps pixels representing states (and intentions) to a Q-value map that is pixelwise aligned with the state. This type of computation (dense pixelwise prediction using fully convolutional networks) has been shown to be effective for many visual tasks (e.g., semantic segmentation). In our case, it is a good fit because encoding intentions into the spatial domain allows deep Q-networks to reason about intentions in the same domain as the state representation.

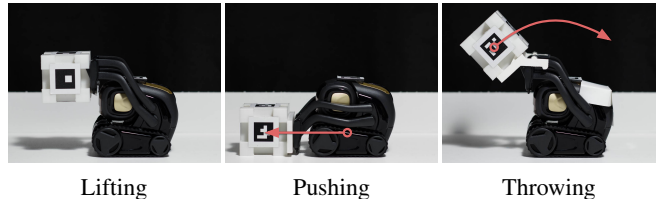


Fig. 3. We experiment with multiple robot types, each with a unique ability.

Agents in our setup use decentralized, asynchronous execution, which means that whenever an agent is choosing a new action, all of the other agents in the environment will be in motion, executing their most recently selected action. We encode these in-progress actions spatially as rasterized paths in a spatial intention map. Intended paths are encoded using a linear ramp function, with a value of 1 at the executing agent’s current location, and dropping off linearly along the path (see Fig. 2 for an example). A lower value at a point on a path indicates a longer period of time before the executing agent will reach that point. This information enables more fine-grained reasoning about time and distance (e.g., the agent over there intends to come here, but is still quite a distance away).

The bandwidth requirements for spatial intention maps during decentralized execution are low, as agents do not communicate maps (images) directly. Instead, they broadcast intentions as lists of (x, y) coordinates (waypoints of intended path). Then, whenever an agent is selecting a new action, it locally renders the most recently received intentions (paths) into an up-to-date spatial intention map.

IV. EXPERIMENTS

We evaluate spatial intention maps with decentralized multi-robot teams on both *foraging* and *search and rescue* tasks across a wide range of simulation environments. The goals of these experiments are to investigate (i) whether learning with spatial intention maps is better than without, and (ii) how learning with spatial intention maps compares with alternative representations of intention. We also test our policies, which were trained in simulation, on real robots.

A. Experimental Setup

We investigate two tasks: (i) *foraging*: where a team of robots must collaborate to move a set of objects into a receptacle (objects are removed from the environment when they enter the receptacle), and (ii) *search and rescue*: where a team of robots must navigate to objects scattered in an environment and “rescue” them (objects are removed after contact by a robot). The two tasks are trained in the same way for the most part. Robots receive a success reward (+1.0) whenever an object is removed from the environment, and receive penalties for collisions with obstacles (-0.25) or other agents (-1.0). In the foraging task, they receive a distance-based partial reward/penalty for moving an object closer to or further from the receptacle. The lifting robot is also penalized (-0.25) for dropping objects outside the receptacle.

We experiment with four types of robots (Fig. 3): (i) a lifting robot that can pick up objects and carry them, (ii) a

pushing robot that can push objects, (iii) a throwing robot that can throw objects backwards, and (iv) a rescue robot that can mark an object as “rescued” upon contact. Teams can be homogeneous (4 of the same) or heterogeneous combinations (2 robot types with 2 per type). For all robots, an action consists of moving to a target location and optionally executing an end effector action (lifting or throwing).

We test the robot teams in six different environments (see Fig. 4). In empty environments with no obstacles, robots need to coordinate to avoid colliding while going towards the receptacle. In environments with obstacles, there are doorways and tunnels wide enough for only one robot to pass through, requiring coordination to avoid obstructing one another. We initialize robots, objects, and obstacles (dividers and walls) in random configurations at the start of each episode. Small environments contain 10 objects, and large ones contain 20. For the SmallDivider, LargeDoors, and LargeTunnels environments, robots and objects are initialized in opposite sides of the room, requiring the robots to bring all objects through doorways or tunnels to reach the receptacle.

For each episode, our evaluation metric is the total number of objects gathered after a fixed time cutoff, which measures the efficiency of a team (higher is better). The cutoff is the time at which the most efficient policy places the last object into the receptacle, and is different for every unique combination of robot team and environment, but is kept consistent between methods. We evaluate a trained policy by averaging its performance over 20 test episodes. For each method, we train five policies and report the mean and standard deviation of those averages across all five. Between evaluation runs, randomly initialized environments are kept consistent using random seeds (training is not seeded).

B. Experimental Results

In this section, we investigate whether spatial intention maps are useful by comparing the performance of multi-robot teams trained with and without them.

Foraging task. We first investigate performance on the foraging task with teams of 4 lifting robots (4L) or 4 pushing robots (4P). The corresponding rows in Tab. I show a comparison of teams trained with and without spatial intention maps. The results suggest that teams trained with spatial intention maps (Ours column) are able to learn the foraging task better. The biggest quantitative differences occur in the larger and more complex environments—*e.g.*, the teams of lifting robots pick up an average of over 19 objects in the large environments with spatial intention maps, and only 11 to 17 without. However, the differences are significant in all settings.

We conjecture one reason for the differences is that robots without intention maps tend to move conservatively back and forth to minimize collisions with nearby robots. This is particularly evident when multiple robots are trying to approach the same object, or trying to put objects into the receptacle at the same time. For example, they may all approach the receptacle, observe that others are also moving

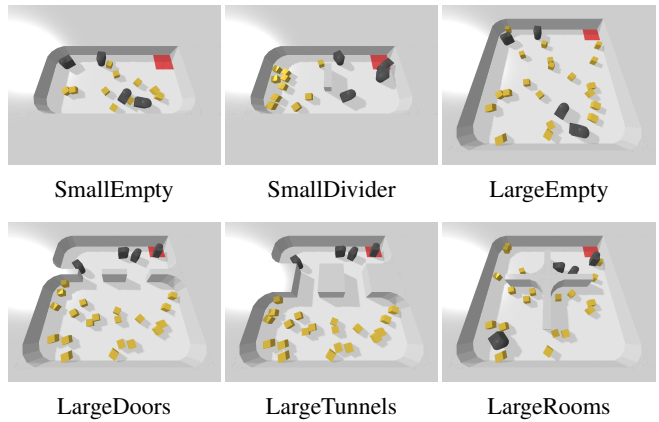


Fig. 4. **Environments.** We ran experiments in six different environments with a variety of different obstacle configurations. In each environment, a team of robots (dark gray) is tasked with moving all objects (yellow) to the receptacle in the corner (red).

TABLE I
FORAGING PERFORMANCE

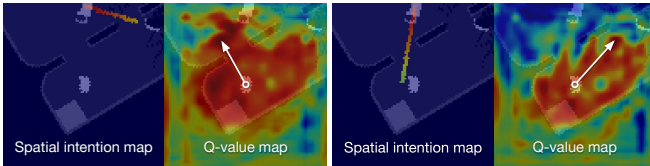
Robots	Environment	Ours	No intention maps
4L	SmallEmpty	9.54 ± 0.16	7.92 ± 0.86
	SmallDivider	9.44 ± 0.46	8.07 ± 0.75
	LargeEmpty	18.86 ± 0.85	15.58 ± 3.80
	LargeDoors	19.57 ± 0.25	13.96 ± 2.32
	LargeTunnels	19.00 ± 0.47	11.89 ± 5.96
	LargeRooms	19.52 ± 0.38	16.56 ± 1.53
4P	SmallEmpty	9.51 ± 0.20	8.73 ± 0.54
	SmallDivider	9.50 ± 0.24	8.40 ± 0.78
	LargeEmpty	19.51 ± 0.53	18.86 ± 0.72
2L+2P	LargeEmpty	19.52 ± 0.17	16.51 ± 4.27
	LargeDoors	19.55 ± 0.18	17.44 ± 0.63
	LargeRooms	19.51 ± 0.24	18.51 ± 0.75
2L+2T	LargeEmpty	19.51 ± 0.67	12.46 ± 4.34
	LargeDoors	19.50 ± 0.45	6.21 ± 4.12

L = lifting, P = pushing, T = throwing

forward, then all preemptively back away. In contrast, with intention maps, each robot is able to consider the intentions of other robots in order to choose actions that do not conflict.

By inspecting output Q-value maps, we can validate that the spatial intention maps are being used in this way. Fig. 5 examines a pair of scenarios requiring coordination between robots to pass through a shared doorway. We find that the predicted Q-value maps for each robot assign higher Q-values to actions that are compatible with what the other robot intends to do.

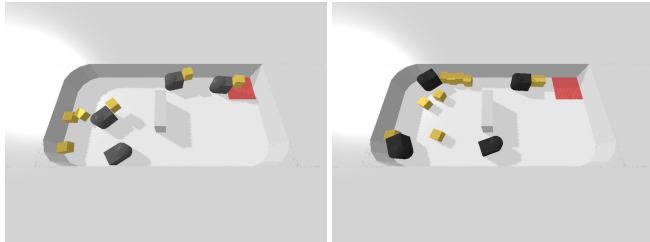
We observe several interesting emergent strategies when training with intention maps. For example, in the Small-Divider environment, both lifting and pushing robots consistently learn to move single file in a clockwise circle around the center divider (Fig. 6). In this emergent pattern, the robots move leftward through the bottom opening to find more objects, and rightward through the top opening to carry objects towards the receptacle. By maintaining one-way traffic in each opening, the robots avoid having to pause and coordinate passing through an opening from opposite directions. Without intention maps, we observe that this pattern still occasionally emerges, but is difficult to sustain.



Scenario 1

Scenario 2

Fig. 5. **Coordinating to go through doorways.** In these scenarios, the current robot (center) is choosing a new action, while the other robot (top) is already moving. The other robot is headed towards the further (right) doorway in scenario 1 and the closer (left) doorway in scenario 2. In both scenarios, the Q-value map suggests the current robot should go through the unoccupied doorway.



Lifting

Pushing

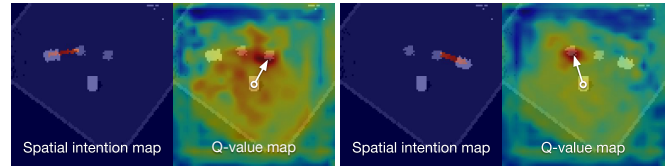
Fig. 6. **Emergent foraging strategy.** With intention maps, both lifting and pushing teams learn an effective and efficient strategy in which they move single file in a clockwise circle around the center divider.

TABLE II
SEARCH AND RESCUE PERFORMANCE

Environment	Ours	No intention maps
SmallEmpty	9.56 ± 0.28	9.08 ± 0.45
LargeEmpty	19.52 ± 0.21	18.49 ± 0.72

Search and rescue task. We then investigate whether spatial intention maps are also helpful for the search and rescue task. Tab. II shows results for teams of 4 rescue robots in the SmallEmpty and LargeEmpty environments. Both quantitative and qualitative results indicate that robots without intention maps are less efficient. They tend to move back and forth unproductively as multiple robots try to rescue the same object, especially towards the end of an episode (when there are only a few objects left). Since the search and rescue task by its nature finishes much more quickly than the foraging task, we are able to see this behavior very clearly in the movement trajectories (Fig. 8). By contrast, robots that use intention maps know the intentions of other robots, and can choose actions that avoid overlapping with others, leading to better distribution of robots throughout the environment. This emerging behavior can be seen clearly in visualizations of Q-value maps of trained policies. For example, in Fig. 7, we see that the locations corresponding to objects another robot already intends to rescue are generally assigned lower Q-values. These results indicate that spatial intention maps can lead to more efficient execution of the rescue task by helping agents avoid overlapping in their efforts.

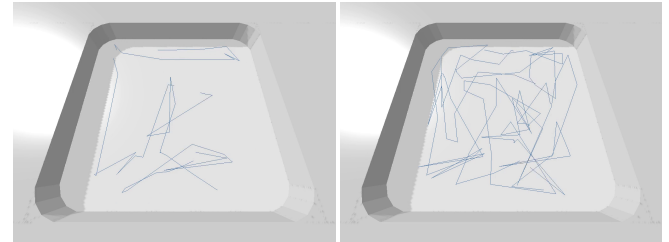
Heterogeneous teams. We also investigate whether spatial intention maps are useful for heterogeneous teams of robots. The bottom five rows of Tab. I show results on the foraging task for teams of 2 lifting and 2 pushing robots (2L+2P), and



Scenario 1

Scenario 2

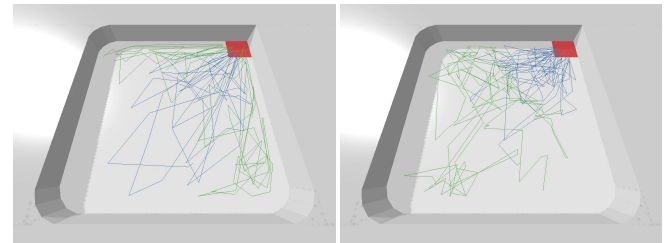
Fig. 7. **Coordinating to rescue objects.** The current robot (center) is choosing a new action, while the other robot is already moving. They are trying to rescue the two objects (small squares on top). The other robot intends to rescue the left object in scenario 1, and the right object in scenario 2. In both scenarios, the Q-value map suggests the current robot should rescue the opposite object, to avoid overlapping of efforts.



Ours

No intention maps

Fig. 8. **Search and rescue team efficiency.** Movement trajectories (blue) over an episode show that rescue robots finish their task more efficiently when intention maps are used. Without intention maps, the robots are unable to coordinate as well since they do not know the intentions of other robots.



Lifting + Pushing

Lifting + Throwing

Fig. 9. **Emergent division of labor for heterogeneous teams.** When we train heterogeneous teams with spatial intention maps, we see from movement trajectories that a natural division of labor emerges (lifting trajectories are blue, pushing/throwing trajectories are green). Notice that there is almost no overlap between blue and green trajectories in either image. We see that pushing robots focus on objects along the wall since those can be pushed much more reliably, while throwing robots focus on faraway objects since they can throw them backwards a long distance towards the receptacle.

for teams of 2 lifting and 2 throwing robots (2L+2T). While the lifting robot is generally versatile, the pushing robot is particularly adept at pushing objects along walls, while the throwing robot can throw objects long distances backwards. The unique capabilities of the different robot types are complementary to each other, and indeed, we find that when we train these heterogeneous teams, a natural division of labor emerges, as shown in Fig. 9. This specialization generally happens with or without spatial intention maps, but we find that when training without intention maps, some of the robots in the team may learn to perpetually wander around without doing anything useful, or the team may never learn to fully complete the task. They likely develop this conservative behavior since they are unable to coordinate and avoid collisions without knowing each other's intentions. By contrast, when training with intention maps, heterogeneous teams are more efficient and productive.

TABLE III
COMPARISONS AND ABLATIONS

Environment	Explicit communication				Implicit communication					
	Ours	Intention maps			Intention channels		Baselines		Predicted intention	
		Binary	Line	Circle	Spatial	Nonspatial	No intention	History maps	No history	With history
SmallEmpty	9.54 ± 0.16	9.25 ± 0.27	9.56 ± 0.15	9.19 ± 0.33	9.33 ± 0.43	8.38 ± 0.52	7.92 ± 0.86	9.29 ± 0.16	8.95 ± 0.32	9.05 ± 0.30
SmallDivider	9.44 ± 0.46	9.28 ± 0.49	8.98 ± 0.89	9.55 ± 0.16	9.47 ± 0.37	8.73 ± 0.85	8.07 ± 0.75	9.20 ± 0.61	8.69 ± 0.90	9.11 ± 0.43
LargeEmpty	18.86 ± 0.85	19.51 ± 0.47	19.43 ± 0.17	17.41 ± 3.75	18.36 ± 0.94	18.15 ± 0.54	15.58 ± 3.80	17.88 ± 1.56	18.18 ± 1.32	18.29 ± 1.45
LargeDoors	19.57 ± 0.25	18.38 ± 1.98	17.84 ± 1.16	17.89 ± 1.43	18.43 ± 0.52	14.07 ± 1.89	13.96 ± 2.32	16.14 ± 2.15	17.84 ± 1.55	18.81 ± 0.94
LargeTunnels	19.00 ± 0.47	18.95 ± 0.75	18.11 ± 1.96	19.51 ± 0.42	18.65 ± 0.87	12.43 ± 1.73	11.89 ± 5.96	18.08 ± 1.35	18.74 ± 0.81	18.07 ± 1.89
LargeRooms	19.52 ± 0.38	18.59 ± 0.99	18.84 ± 0.96	19.51 ± 0.31	19.15 ± 0.57	17.55 ± 0.30	16.56 ± 1.53	17.84 ± 0.58	18.97 ± 0.34	19.35 ± 0.19

C. Comparisons and Ablations

In this section, we perform comparisons and ablations across a wide range of communication variants. For all variants, we run experiments on the foraging task using homogeneous teams of 4 lifting robots in all six environments.

Comparison to nonspatial intentions. We compare spatial intention maps to a nonspatial encoding of intentions, in which we add two channels per robot (sorted closest robot to furthest) and tile them with the x and y values of each robot’s intended target location. This nonspatial encoding conveys to the network the (x, y) target of other robots’ intended destination in the best way we can imagine without a spatially varying map. As shown in Tab. III, we find that providing the intentions encoded in a nonspatial format results in much worse performance, suggesting that it is important to use a spatial encoding for agent intentions.

Encoding of spatial intentions. Our method encodes intended paths (with linearly ramped values) in a 2D map. Here we investigate four alternative encodings: (i) a binary encoding of the intention map where the linear ramp of values is replaced with a binary-on-path off-path map, (ii) a line encoding that replaces the path encoding with a simple straight line between each agent to its intended target location (which may not necessarily reflect the agent’s true trajectory), (iii) a circle encoding, where a circle marks each agent’s target location (this variant does not associate each agent with its target location), and (iv) an encoding that decomposes the circle intention map into multiple channels (“spatial intention channels” in Tab. III), one per robot, sorted in order from closest robot to furthest (thus allowing a way to associate agents with intentions). Results in Tab. III show that the binary and line variants generally perform on par with our method, while both circle variants are worse overall (but still better than without intention maps). These results suggest that providing any spatial encoding of intention is highly beneficial to multi-agent coordination, as long as there is a clear visual association between agents and their intentions.

Predicting spatial intention maps. We further explore whether agents can learn to predict useful intention maps, either with heuristics or deep networks. We explore three possible methods: (i) predicted intention, which involves training an additional fully convolutional network to predict intention maps from the state representation so that they can be used during execution with no communication, (ii) history maps, where instead of encoding intention, the recent trajec-

tory history of other agents are encoded into a map (assumes robots can track each other’s poses without communicating), and (iii) a combination of (i) and (ii), where an additional network is trained to predict the intention map from a state representation augmented with the history map. Results in Tab. III show that in isolation, (i) and (ii) do not improve performance as much as spatial intention maps do. However, we find that the last variant (iii) which combines history maps with predicted intention achieves performance (Tab. III) almost on par with explicit communication using spatial intention maps. This result is significant in that it provides a method for robot teams to leverage the benefits of spatial intention maps to coordinate without explicit communication, albeit at the cost of extra computation.

D. Real Robot Experiments

As a final test, we take the best policies from our simulation experiments and run them directly on the real robots by building upon the physical setup used in [8]. We use fiducial markers to estimate the poses of robots and objects, which allows us to mirror the real setup inside the simulation. This means we can execute the trained policies (without fine-tuning) on the real robots as if they were moving around in the simulation. During tests with 4 lifting robots in a mirror of the SmallEmpty environment, the robots are able to gather all 10 objects within 1 minute and 56 seconds (on average over 5 episodes). For video results across a wide range of environments, please see the supplementary material at <https://spatial-intention-maps.cs.princeton.edu>.

V. CONCLUSION

In this work, we propose *spatial intention maps* as a new way of communicating intention for improved collaboration in multi-agent RL. In our framework, intentions are spatially encoded in a 2D map, allowing vision-based RL agents to spatially reason about intentions in the same domain as the states and actions. During experiments, we observe that spatial intention maps provide significant gains in performance across a broad range of multi-agent environments, and help robot teams learn emergent collaborative behaviors such as avoiding collisions, coordinating passage through bottlenecks, and distributing throughout an environment.

ACKNOWLEDGMENTS

We thank Naomi Leonard, Anirudha Majumdar, Naveen Verma, and Anish Singhani for fruitful technical discussions. This work was supported by Princeton School of Engineering, as well as NSF under IIS-1815070 and DGE-1656466.

REFERENCES

- [1] M. J. Mataric, "Issues and approaches in the design of collective autonomous agents," *Robotics and autonomous systems*, vol. 16, no. 2-4, pp. 321–331, 1995.
- [2] L. E. Parker, "Alliance: An architecture for fault tolerant multirobot cooperation," *IEEE transactions on robotics and automation*, vol. 14, no. 2, pp. 220–240, 1998.
- [3] D. Rus, B. Donald, and J. Jennings, "Moving furniture with teams of autonomous robots," in *Proceedings 1995 IEEE/RSJ International Conference on Intelligent Robots and Systems. Human Robot Interaction and Cooperative Robots*, vol. 1. IEEE, 1995, pp. 235–242.
- [4] E. Tuci, M. H. Alkilabi, and O. Akanyeti, "Cooperative object transport in multi-robot systems: A review of the state-of-the-art," *Frontiers in Robotics and AI*, vol. 5, p. 59, 2018.
- [5] R. R. Murphy, "Marsupial and shape-shifting robots for urban search and rescue," *IEEE Intelligent Systems and their applications*, vol. 15, no. 2, pp. 14–19, 2000.
- [6] B. Baker, I. Kanitscheider, T. Markov, Y. Wu, G. Powell, B. McGrew, and I. Mordatch, "Emergent tool use from multi-agent autotutorials," in *International Conference on Learning Representations*, 2019.
- [7] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, et al., "Grandmaster level in starcraft ii using multi-agent reinforcement learning," *Nature*, vol. 575, no. 7782, pp. 350–354, 2019.
- [8] J. Wu, X. Sun, A. Zeng, S. Song, J. Lee, S. Rusinkiewicz, and T. Funkhouser, "Spatial action maps for mobile manipulation," in *Proceedings of Robotics: Science and Systems (RSS)*, 2020.
- [9] L. E. Parker, D. Rus, and G. S. Sukhatme, "Multiple mobile robot systems," in *Springer Handbook of Robotics*. Springer, 2016, pp. 1335–1384.
- [10] H. Kitano, M. Asada, Y. Kuniyoshi, I. Noda, E. Osawa, and H. Matsubara, "Robocup: A challenge problem for ai," *AI magazine*, vol. 18, no. 1, pp. 73–73, 1997.
- [11] G. Weiss, "Multiagent systems: a modern approach to distributed artificial intelligence," 1999.
- [12] P. Stone and M. Veloso, "Multiagent systems: A survey from a machine learning perspective," *Autonomous Robots*, vol. 8, no. 3, pp. 345–383, 2000.
- [13] L. Busoniu, R. Babuska, and B. De Schutter, "A comprehensive survey of multiagent reinforcement learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 38, no. 2, pp. 156–172, 2008.
- [14] M. Tan, "Multi-agent reinforcement learning: Independent versus cooperative agents," in *ICML*, 1993.
- [15] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, p. 529, 2015.
- [16] P. Hernandez-Leal, B. Kartal, and M. E. Taylor, "A survey and critique of multiagent deep reinforcement learning," *Autonomous Agents and Multi-Agent Systems*, vol. 33, no. 6, pp. 750–797, 2019.
- [17] J. Foerster, N. Nardelli, G. Farquhar, T. Afouras, P. H. Torr, P. Kohli, and S. Whiteson, "Stabilising experience replay for deep multi-agent reinforcement learning," *arXiv preprint arXiv:1702.08887*, 2017.
- [18] S. Omidshafiei, J. Papis, C. Amato, J. P. How, and J. Vian, "Deep decentralized multi-task multi-agent reinforcement learning under partial observability," in *International Conference on Machine Learning*, 2017, pp. 2681–2690.
- [19] R. Lowe, Y. Wu, A. Tamar, J. Harb, O. P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Advances in neural information processing systems*, 2017, pp. 6379–6390.
- [20] J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, "Counterfactual multi-agent policy gradients," *arXiv preprint arXiv:1705.08926*, 2017.
- [21] P. Sunehag, G. Lever, A. Gruslys, W. M. Czarnecki, V. Zambaldi, M. Jaderberg, M. Lanctot, N. Sonnerat, J. Z. Leibo, K. Tuyls, et al., "Value-decomposition networks for cooperative multi-agent learning," *arXiv preprint arXiv:1706.05296*, 2017.
- [22] T. Rashid, M. Samvelyan, C. Schroeder, G. Farquhar, J. Foerster, and S. Whiteson, "Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning," in *International Conference on Machine Learning*, 2018, pp. 4295–4304.
- [23] T. Zhang, H. Xu, X. Wang, Y. Wu, K. Keutzer, J. E. Gonzalez, and Y. Tian, "Multi-agent collaboration via reward attribution decomposition," *arXiv preprint arXiv:2010.08531*, 2020.
- [24] A. Tampuu, T. Matiisen, D. Kodelja, I. Kuzovkin, K. Korjus, J. Aru, J. Aru, and R. Vicente, "Multiagent cooperation and competition with deep reinforcement learning," *PLoS one*, vol. 12, no. 4, p. e0172395, 2017.
- [25] M. Jaderberg, W. M. Czarnecki, I. Dunning, L. Marris, G. Lever, A. G. Castaneda, C. Beattie, N. C. Rabinowitz, A. S. Morcos, A. Ruderman, et al., "Human-level performance in 3d multiplayer games with population-based reinforcement learning," *Science*, vol. 364, no. 6443, pp. 859–865, 2019.
- [26] S. Mahadevan and J. Connell, "Automatic programming of behavior-based robots using reinforcement learning," *Artificial intelligence*, vol. 55, no. 2-3, pp. 311–365, 1992.
- [27] M. J. Mataric, "Reinforcement learning in the multi-robot domain," in *Robot colonies*. Springer, 1997, pp. 73–83.
- [28] M. Asada, E. Uchibe, S. Noda, S. Tawaratsumida, and K. Hosoda, "Coordination of multiple behaviors acquired by a vision-based reinforcement learning," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'94)*, vol. 2. IEEE, 1994, pp. 917–924.
- [29] P. Stone and M. Veloso, "Towards collaborative and adversarial learning: A case study in robotic soccer," *International Journal of Human-Computer Studies*, vol. 48, no. 1, pp. 83–104, 1998.
- [30] F. FERNÁNDEZ, D. BORRAJO, and L. E. PARKER, "A reinforcement learning algorithm in cooperative multi-robot domains," *Journal of Intelligent and Robotic Systems*, vol. 43, pp. 161–174, 2005.
- [31] Y. Xiao, J. Hoffman, and C. Amato, "Macro-action-based deep multi-agent reinforcement learning," in *Conference on Robot Learning*, 2020, pp. 1146–1161.
- [32] Y. Xiao, J. Hoffman, T. Xia, and C. Amato, "Learning multi-robot decentralized macro-action-based policies via a centralized q-net," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 10695–10701.
- [33] C. Amato, G. Konidaris, L. P. Kaelbling, and J. P. How, "Modeling and planning with macro-actions in decentralized pomdps," *Journal of Artificial Intelligence Research*, vol. 64, pp. 817–859, 2019.
- [34] R. E. Wang, J. C. Kew, D. Lee, T.-W. E. Lee, T. Zhang, B. Ichter, J. Tan, and A. Faust, "Cooperation without coordination: Hierarchical predictive planning for decentralized multiagent navigation," *arXiv preprint arXiv:2003.06906*, 2020.
- [35] O. Nachum, M. Ahn, H. Ponte, S. S. Gu, and V. Kumar, "Multi-agent manipulation via locomotion using hierarchical sim2real," in *Conference on Robot Learning*, 2020, pp. 110–121.
- [36] B. J. MacLennan and G. M. Burghardt, "Synthetic ethology and the evolution of cooperative communication," *Adaptive behavior*, vol. 2, no. 2, pp. 161–188, 1993.
- [37] T. Balch and R. C. Arkin, "Communication in reactive multiagent robotic systems," *Autonomous robots*, vol. 1, no. 1, pp. 27–52, 1994.
- [38] L. E. Parker, "The effect of action recognition and robot awareness in cooperative robotic teams," in *Proceedings 1995 IEEE/RSJ International Conference on Intelligent Robots and Systems. Human Robot Interaction and Cooperative Robots*, vol. 1. IEEE, 1995, pp. 212–219.
- [39] C. Zhang and V. Lesser, "Coordinating multi-agent reinforcement learning with limited communication," in *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*, 2013, pp. 1101–1108.
- [40] L. Panait and S. Luke, "Cooperative multi-agent learning: The state of the art," *Autonomous agents and multi-agent systems*, vol. 11, no. 3, pp. 387–434, 2005.
- [41] J. Foerster, I. A. Assael, N. De Freitas, and S. Whiteson, "Learning to communicate with deep multi-agent reinforcement learning," in *Advances in neural information processing systems*, 2016, pp. 2137–2145.
- [42] S. Sukhbaatar, R. Fergus, et al., "Learning multiagent communication with backpropagation," in *Advances in neural information processing systems*, 2016, pp. 2244–2252.
- [43] I. Mordatch and P. Abbeel, "Emergence of grounded compositional language in multi-agent populations," *arXiv preprint arXiv:1703.04908*, 2017.
- [44] F. Stulp, M. Isik, and M. Beetz, "Implicit coordination in robotic teams using learned prediction models," in *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006*. IEEE, 2006, pp. 1330–1335.

- [45] S. Qi and S.-C. Zhu, "Intent-aware multi-agent reinforcement learning," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 7533–7540.
- [46] N. Jaques, A. Lazaridou, E. Hughes, C. Gulcehre, P. Ortega, D. Strouse, J. Z. Leibo, and N. De Freitas, "Social influence as intrinsic motivation for multi-agent deep reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2019, pp. 3040–3049.
- [47] K. Ndousse, D. Eck, S. Levine, and N. Jaques, "Multi-agent social reinforcement learning improves generalization," *arXiv preprint arXiv:2010.00581*, 2020.
- [48] E. Coumans and Y. Bai, "Pybullet, a python module for physics simulation for games, robotics and machine learning," *GitHub repository*, 2016.
- [49] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double q-learning," in *Thirtieth AAAI conference on artificial intelligence*, 2016.
- [50] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.
- [51] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

A. Additional Reinforcement Learning Details

We model our task as a Markov decision process (MDP) from the perspective of each individual agent. Specifically, given state s_t at time t , the agent takes an action a_t following policy $\pi(s_t)$, and arrives at a new state s_{t+1} while receiving reward r_t . In Q-learning, the goal is to find some optimal policy π^* that selects actions maximizing the Q-function $Q(s_t, a_t) = \sum_{i=t}^{\infty} \gamma^{i-t} r_i$, which represents the discounted sum of future rewards. We train our policies using deep Q-learning (DQN) [15], which approximates the Q-function using a neural network and uses a policy that greedily selects actions maximizing the Q-function: $\pi(s_t) = \arg \max_{a_t} Q_{\theta}(s_t, a_t)$, where θ refers to the parameters of the neural network. We train using the double DQN learning objective [49] with smooth L1 loss. Formally, at each training iteration i , we minimize $\mathcal{L}_i = |r_t + \gamma Q_{\theta^-}(s_{t+1}, \arg \max_{a_{t+1}} Q_{\theta_i}(s_{t+1}, a_{t+1})) - Q_{\theta_i}(s_t, a_t)|$ where (s_t, a_t, r_t, s_{t+1}) is a transition uniformly sampled from the replay buffer, and θ^- refers to the parameters of the DQN target network.

B. Additional Training Details

We train DQN using SGD for either 160k or 240k timesteps (depending on the robot types in the team), with batch size 32, learning rate 0.01, momentum 0.9, and weight decay 0.0001. Norms of gradients are clipped to 100 during training. We use a replay buffer of size 10k, constant discount factor $\gamma = 0.85$, train frequency of 4 (train the policy network every 4 timesteps), and we update the target network every 1,000 timesteps. Episodes end after all objects have been removed from the environment, or after 400 consecutive steps in which no object has been removed. Training takes approximately 6 hours on a single Nvidia Titan Xp GPU (for the lifting team).

Network architecture. We model our Q-function Q_{θ} with a ResNet-18 [51] backbone, which we transform into a fully convolutional network by removing the AvgPool and fully connected layers and adding three 1x1 convolution layers interleaved with bilinear upsampling layers. Unlike in [8], we apply BatchNorm after convolutional layers.

Exploration. We run a random policy for a small number of timesteps (1/40 of the total number of timesteps) in order to fill the replay buffer with some initial experience before we start training the policy network. We use ϵ -greedy exploration, in which we linearly anneal the exploration fraction from 1 to 0.01 during the first 1/10 of training.

Multi-agent. We use one buffer/policy per robot type during training. This means that homogeneous teams pool all transitions into one replay buffer and train one shared policy, while heterogeneous teams share a buffer/policy within each group of robots (of one robot type). For example, the lifting and pushing team trains with two buffers/policies. For decentralized execution, trained policies are independently and asynchronously run on each individual agent.

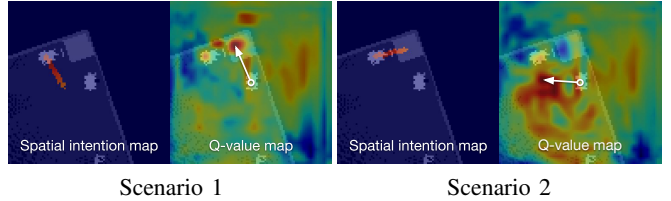


Fig. 10. **Coordinating to go towards the receptacle.** In these scenarios, the current robot (center) is choosing a new action, while the other robot (left) is already moving. The other robot is moving away from the receptacle in scenario 1, and towards the receptacle in scenario 2. The Q-value map suggests that the current robot should drop off the object in the receptacle only if it does not conflict with the other robot. Otherwise, it suggests that the current robot should move left (and wait for the other robot to finish).

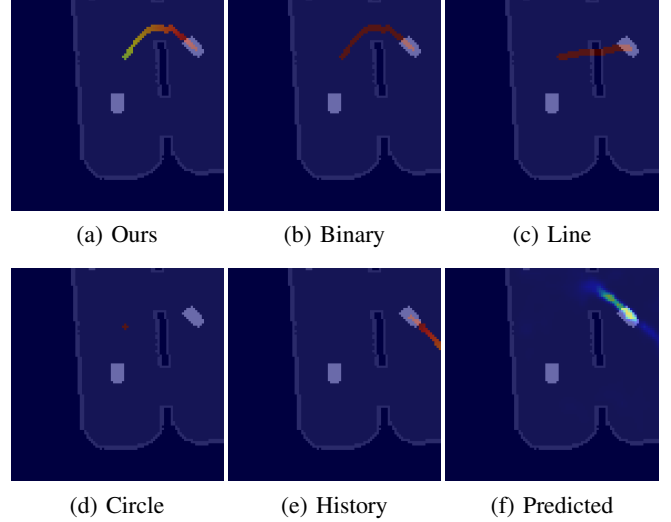


Fig. 11. **Communication variants.** Here we illustrate the different communication variants we experimented with. In each image, the current robot (center) is choosing a new action, while the other robot (right) is moving towards the left. The images show (a) our method, (b-d) variants of spatial intention maps, (e) history maps, and (f) predicted intention maps.

Search and rescue. We made a few small modifications to the foraging task for search and rescue. Since the receptacle is not relevant, we remove it for this task, and also remove the corresponding channel (shortest path distances to the receptacle) from the state representation. To reflect the shorter term nature of this simpler task, we lower the discount factor γ to 0.35 and reduce the number of training timesteps to 15k.

Predicted intention. To predict intention maps, we train an additional fully convolutional network in tandem with our policy. We use the same network architecture as our policy network (but with a sigmoid layer at the end) and train with binary cross entropy loss (supervised on our intention map). The policy network uses our intention map as input for the first 9/10 of training and switches to the predicted intention map for the last 1/10 of training (as well as execution).

C. Additional Figures

Fig. 10 shows output Q-value maps for another pair of scenarios (lifting robots coordinate to go towards the receptacle) to further validate that spatial intention maps are used by trained policies in a meaningful way. Fig. 11 shows the different communication variants we study in Sec. IV-C.